applied sciences

MDPI

Article

HP-YOLO: A Lightweight Real-Time Human Pose Estimation Method

Haiyan Tu [1,2], Zhengkun Qiu [1,2], Kang Yang [1,2], Xiaoyue Tan [1,2] and Xiujuan Zheng [1,*]

[1] Department of Automation, College of Electrical Engineering, Sichuan University, Chengdu 610065, China; haiyantu@scu.edu.cn (H.T.); qzk1598185258@gmail.com (Z.Q.); sonyyangks@outlook.com (K.Y.); xytan1328592924@gmail.com (X.T.)
[2] Key Laboratory of Information and Automation Technology of Sichuan Province, Sichuan University, Chengdu 610065, China
* Correspondence: xiujuanzheng@scu.edu.cn

Abstract: Human Pose Estimation (HPE) plays a critical role in medical applications, particularly within nursing robotics for patient monitoring. Despite its importance, HPE faces several challenges, including high rates of false positives and negatives, stringent real-time requirements, and limited computational resources, especially in complex backgrounds. In response, we introduce the HP-YOLO model, developed using the YOLOv8 framework, to effectively address these issues. We designed an Enhanced Large Separated Kernel Attention (ELSKA) mechanism and integrated it into the backbone network, thereby improving the model's effective receptive field and feature separation capabilities, which enhances keypoint detection accuracy in challenging environments. Additionally, the Reparameterized Network with Cross-Stage Partial Connections and Efficient Layer Aggregation Network (RepNCSPELAN4) module was incorporated into the detection head, boosting accuracy in detecting small-sized targets through multi-scale convolution and reparameterization techniques while accelerating inference speed. On the COCO dataset, our HP-YOLO model outperformed existing lightweight methods by increasing average precision (AP) by 4.9%, while using 18% fewer parameters and achieving 1.4× higher inference speed. Our method significantly enhances the real-time performance and efficiency of human pose estimation while maintaining high accuracy, offering an optimal solution for applications in complex environments.

Keywords: HPE; attention mechanism; model pruning; lightweight structure

Received: 5 February 2025
Revised: 1 March 2025
Accepted: 5 March 2025
Published: 11 March 2025

Citation: Tu, H.; Qiu, Z.; Yang, K.; Tan, X.; Zheng, X. HP-YOLO: A Lightweight Real-Time Human Pose Estimation Method. *Appl. Sci.* **2025**, *15*, 3025. https://doi.org/10.3390/app15063025

1. Introduction

Envision a future where patient well-being is proactively safeguarded through continuous, intelligent monitoring. This vision is rapidly materializing, driven by remarkable progress in HPE, a transformative computer vision technology dedicated to automatically deciphering human body configurations from visual data. HPE, now a cornerstone of modern computer vision, is not merely enhancing virtual reality and sports analytics; it is fundamentally reshaping diverse sectors, from enabling nuanced human–computer interaction to revolutionizing movement-based rehabilitation [1–3]. Recent advancements have further highlighted the effectiveness of integrating attention mechanisms and multi-scale feature fusion strategies in medical image analysis, significantly improving the accuracy and robustness of automated segmentation and classification tasks within complex clinical environments [4–6]. Nevertheless, the deployment of HPE specifically for real-time bedside patient monitoring stands out as possessing unmatched potential to address pressing

healthcare imperatives [7]. Globally, healthcare facilities grapple with persistent challenges like patient falls and pressure injuries, often exacerbated by delayed or inadequate responses to subtle shifts in patient posture. Herein lies the power of bedside HPE: by providing clinicians with a stream of objective, real-time postural insights, it empowers proactive intervention, facilitating timely repositioning, personalized comfort optimization, and ultimately, a significant elevation in patient safety and overall care quality.

First, there are challenges in recognition due to large pose variations: Adding to the complexities of pose variability, another significant hurdle in bedside HPE arises from the intricacies of multi-scale feature representation and fusion. Within bedside images, the human body presents itself across a wide spectrum of scales, from minute details like fingers and toes to expansive regions such as the torso and limbs. Conventional HPE models, often employing single-pathway feature extraction, are ill-equipped to effectively capture and integrate information across this diverse scale range. These architectures typically lack the capacity to simultaneously process both fine-grained features, essential for precise localization of small body parts, and coarse-grained, contextual features, necessary for understanding the overall body structure. This inherent limitation in multi-scale feature handling directly impacts the model's ability to accurately discern keypoints, particularly in the visually cluttered and scale-variant scenes characteristic of bedside monitoring [8]. Therefore, advancements in multi-scale feature fusion techniques are crucial for enhancing the robustness and accuracy of HPE in this challenging domain. To address the challenge of large pose variations, we propose the ELSKA Attention Mechanism.

Second, there are the issues of insufficient large-scale information and inadequate feature fusion: Furthermore, the practical deployment of bedside HPE systems is critically contingent upon achieving high computational efficiency and minimal resource consumption. The clinical environment often presents stringent limitations in terms of available computing power, particularly at the point of care. Existing HPE methodologies [9–12], while demonstrating progress in accuracy, frequently rely on computationally intensive architectures, rendering them unsuitable for real-time bedside applications. This imperative for lightweight, efficient models necessitates a paradigm shift towards solutions that not only enhance pose estimation accuracy but also prioritize operational efficiency on resource-constrained devices. To overcome the limitations of insufficient multi-scale feature handling, we introduce the RepNCSPELAN4 Structure.

Third, there is the demand for lightweight models for resource-constrained devices: Despite the impressive strides made in Human Pose Estimation (HPE) across diverse applications, its translation to bedside patient monitoring encounters a unique set of obstacles [13]. Unlike open-environment scenarios where HPE models typically excel, the controlled yet complex environment of a hospital bedside presents significant challenges. Patients in these settings are rarely in unobstructed, standardized poses. Instead, their postures are often dictated by medical conditions, comfort needs, and the physical constraints inherent to the bedside environment: beds themselves, layers of bedding, and an array of medical equipment all contribute to occlusions, atypical viewpoints, and significant pose variations. Consequently, HPE algorithms trained on conventional datasets often struggle to generalize effectively to these uncharacteristic and challenging bedside poses [14]. This is particularly evident when considering pose variability. While HPE models are increasingly robust to pose changes within typical human movement ranges, the extreme and often unpredictable postures adopted by bedridden patients from fully supine to tightly curled push the boundaries of current HPE capabilities, resulting in reduced accuracy and reliability [15]. To tackle the challenge of computational constraints, we employ the OKS-Guided L1 Pruning Strategy.

To effectively address these three aforementioned challenges, large pose variations, multi-scale feature representation, and computational constraints, we introduce a novel

bedside HPE framework incorporating three key innovations in model design, with each innovation directly targeting one specific challenge:

1. ELSKA Attention Mechanism for Handling Pose Variations: We introduce the ELSKA mechanism, strategically integrated within the SPPF feature extraction pyramid of the YOLOv8 backbone. This innovation fundamentally optimizes the network's architecture, endowing it with superior spatial context awareness and an enhanced effective receptive field. ELSKA leverages spatial attention for selective focus, thus enhancing the effective receptive field. ELSKA facilitates enhanced extraction of fine-grained features, directly addressing the challenges posed by significant pose variations. As explained above, the ELSKA mechanism is specifically designed to enhance the model's robustness to diverse and unpredictable patient poses by improving spatial context awareness and enhancing the effective receptive field. It is important to note that ELSKA primarily focuses on broadening the effective receptive field in the region in the input image that truly influences the network's output rather than solely expanding the theoretical receptive field defined by the network architecture.
2. RepNCSPELAN4 Structure for Multi-Scale Feature Fusion: In the model's head structure, we replace the conventional C2f module with the RepNCSPELAN4 (Reparameterized Non-Cross Stage Partial Efficient Layer Aggregation Network v4) structure. This architectural modification significantly amplifies the model's feature extraction prowess and its adaptability to multi-scale scenarios. RepNCSPELAN4 effectively mitigates the issues of insufficient large-scale information capture and inadequate feature fusion. As mentioned earlier, the RepNCSPELAN4 structure is specifically implemented to enhance the model's ability to capture and fuse multi-scale features, enabling it to effectively handle the scale variations inherent in bedside images.
3. OKS-Guided L1 Pruning Strategy for Real-Time Performance on Resource-Constrained Devices: To ensure real-time operation and compatibility with resource-limited devices, we employ a novel L1-norm based pruning strategy, intelligently guided by the Object Keypoint Similarity (OKS) metric. This pruning approach judiciously reduces computational complexity and storage footprint while preserving high model accuracy and guaranteeing real-time inference speeds, crucial for practical bedside deployment. As previously stated, the OKS-Guided L1 Pruning strategy is specifically employed to reduce the computational complexity and storage requirements of the model, ensuring its efficient operation on resource-constrained devices commonly found in clinical settings.

The structure of this paper is as follows: Section 2 reviews the related research and technical background. Section 3 details the design and implementation of HP-YOLO. Section 4 presents experimental validation of the algorithm's performance and summarizes the experimental results. Section 5 discusses the limitations and outlines future research directions, and Section 6 concludes the paper.

2. Related Work

2.1. Classical HPE Methods

HPE can be broadly categorized into two main approaches: top–down and bottom–up. Top–down methods use object detection frameworks, such as Faster R-CNN [16] and YOLO [17], to first localize human bodies before estimating keypoints. These methods generally achieve high accuracy but struggle in crowded environments, where overlapping detection boxes degrade performance. Mask R-CNN [18] addresses this issue by incorporating a mask branch, significantly enhancing detection precision and segmentation quality in complex scenes.

The advent of Transformers has revolutionized HPE. ViTPose [19], for instance, leverages the self-attention mechanism to model long-range dependencies, improving robustness, especially in challenging backgrounds and under occlusions.

In contrast, bottom–up methods like OpenPose [11] and DeepCut [20] first detect keypoints and then assemble them into skeletal structures. These methods are computationally efficient for multi-person scenarios, as their complexity does not scale with the number of individuals. However, they often suffer from reduced keypoint matching accuracy in occluded or complex backgrounds. HigherHRNet [21] mitigates these issues by integrating high-resolution features with multi-scale fusion, achieving robust performance in densely populated environments.

2.2. Attention Mechanisms

Attention mechanisms are extensively used in HPE to enhance keypoint recognition and localization. Channel Attention (CA) adjusts channel-wise weights to focus on keypoint-relevant features, improving recognition accuracy [22]. The Squeeze-and-Excitation (SE) block recalibrates channel-wise responses, enhancing detection robustness, particularly in complex backgrounds [23].

The Convolutional Block Attention Module (CBAM) combines spatial and channel attention to optimize localization under occlusions and pose variations [24]. This dual-attention mechanism is particularly useful in challenging scenarios where body parts are occluded or poses vary significantly.

Large-Scale Kernel Attention (LSKA) reduces computational complexity by decomposing large convolutional kernels while maintaining performance. Research by Lau et al. [25] shows that LSKA improves long-range dependency capture, making it effective for multi-scale tasks involving large-scale datasets.

Overall, these attention mechanisms collectively enhance network optimization, leading to improved performance in multi-pose and multi-view scenarios, which is crucial for HPE robustness in real-world applications.

2.3. Lightweight Networks

Lightweight design is crucial in HPE to reduce computational complexity while maintaining accuracy. Lightweight OpenPose [26] achieves real-time keypoint detection on CPUs by reducing redundant parameters through pruning and quantization, using lightweight convolutional architectures. However, its accuracy is limited in complex scenes. The EfficientHR [21] series improves feature extraction efficiency via multi-resolution branch networks and depthwise separable convolutions, enabling efficient inference on mobile devices, making it suitable for low-resource environments.

LitePose [27] and MFite-HRNet [9] maintain high accuracy while reducing computational demands by utilizing deconvolutions and large convolutional kernels. Deconvolutions are used for upsampling, while large kernels integrate additional contextual information to optimize computational load. SRPose [28] redefines pose estimation as a super-resolution problem to enhance detection accuracy under low-resolution conditions. However, it struggles with prediction accuracy in the presence of occlusions or significant pose variations due to blurred input features.

2.4. Model Pruning

Model pruning aims to reduce the computational load and parameter count of models, making them more efficient without compromising performance. Wang et al. proposed a joint approach called APQ, which searches for an optimal combination of network architecture, pruning, and quantization policies. This method compresses model size effectively by reducing both computational demands and storage requirements while maintaining

accuracy, making it suitable for deployment in resource-constrained environments [29]. Similarly, Tang et al. introduced Manifold Regularized Dynamic Network Pruning, which dynamically adapts the model structure based on the characteristics of each input sample. By pruning unnecessary parts of the model during inference, this approach minimizes computational overhead while ensuring high model accuracy [10]. This adaptability is particularly valuable for scenarios involving variable data, where input characteristics can vary significantly.

3. HP-YOLO Method

The proposed HP-YOLO model incorporates three key innovations aimed at addressing challenges related to significant pose variations, limited feature extraction capabilities, and the need for high real-time performance. These innovations target different stages of the model pipeline, optimizing both accuracy and efficiency, and each innovation offers specific advantages in practical real-world scenarios, especially in tasks involving pose variation and multi-scale object detection.

First, the ELSKA mechanism is integrated into the backbone network to enhance the model's ability to capture complex spatial relationships, particularly in scenarios with significant pose variations. This mechanism proves crucial in applications such as human pose estimation and object detection under varying orientations, where adapting to different poses is essential for robust feature extraction. ELSKA leverages both separated and dilated convolutions to effectively differentiate critical features from complex backgrounds, which is particularly advantageous when dealing with small, occluded, or overlapping objects.

Second, the RepNCSPELAN4 structure is introduced in the detection head to enhance feature fusion across multiple scales. By combining multi-scale convolutions, CSP (Cross Stage Partial) [30] connections, and ELAN mechanisms [31], this structure efficiently aggregates features at different scales. This multi-scale feature fusion is particularly beneficial for tasks like autonomous driving or surveillance, where accurate detection of objects of varying sizes (e.g., vehicles, pedestrians) is required. By fusing features from multiple scales, RepNCSPELAN4 significantly improves detection accuracy, especially for small objects in environments with varying object sizes.

Finally, to improve real-time performance while preserving high accuracy, an L1 pruning strategy is employed in conjunction with Object Keypoint Similarity (OKS) [32]. This strategy identifies and prunes parameters that contribute minimally to model predictions, thereby enhancing computational efficiency without compromising predictive performance. In real-time applications, such as video surveillance or mobile device deployment, computational resources are often limited, making this pruning strategy critical for maintaining a balance between speed and accuracy.

The overall architecture of the HP-YOLO model is illustrated in Figure 1, which highlights the integration of these key modules and their respective contributions.

3.1. ELSKA Module

The ELSKA module is specifically designed to enhance the model's ability to capture complex spatial relationships, particularly in scenarios characterized by significant pose variations. This is especially useful in scenarios where objects are presented in varying orientations, such as human pose estimation or animal detection in the wild. The module incorporates two primary innovations:

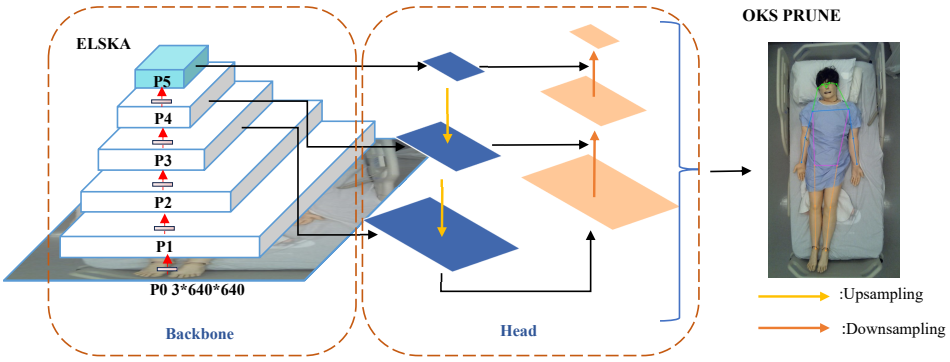


Figure 1. Network structure of HP-YOLO, illustrating the integration of the ELSKA attention mechanism, the RepNCSPELAN4 module, and the pruning strategy.

First, the ELSKA module employs horizontal and vertical separated convolutions [33] to process the input feature map. This approach enables the model to capture pose information from different orientations, reducing interference from complex backgrounds. The horizontal and vertical convolutions are formulated as follows:

$$F_h = X * W_h, \quad F_v = X * W_v \tag{1}$$

where $*$ denotes the convolution operation, and F_h and F_v are the feature maps extracted from the horizontal and vertical convolutions, respectively. This design significantly mitigates background noise, improving pose recognition, especially in scenarios with occlusions or cluttered backgrounds, such as crowded urban environments or industrial settings.

Next, dilated convolutions [34] with expansion coefficients are incorporated to enhance spatial contextual perception. Let r represent the expansion coefficient, and W_h^d and W_v^d denote the dilated convolution kernels in the horizontal and vertical directions. The dilated convolution operation is given by the following:

$$F_h^d = X * W_h^d, \quad F_v^d = X * W_v^d \tag{2}$$

The expansion coefficient r is adjusted based on the complexity of the pose variation, typically set to $r = 2$ or $r = 4$, allowing the model to capture spatial information across various scales. This flexibility enables the model to adapt to varying object sizes and poses, ensuring that both small and large objects are accurately represented in the feature maps.

For cases of occlusion or weak features, the ELSKA module combines the original feature map X with the enhanced feature map F^d through element-wise multiplication, yielding the final output feature map F_{out}:

$$F_{out} = X \odot F^d \tag{3}$$

where $\odot$ denotes element-wise multiplication. This fusion mechanism significantly enhances the model's robustness in complex backgrounds and improves pose recognition accuracy.

In summary, the feature maps produced by the ELSKA module, enriched with spatial and pose information, are passed to the RepNCSPELAN4 [31] module for multi-scale feature fusion, ensuring that the high-quality spatial information is fully utilized in the subsequent stages.

The network architecture of the ELSKA module is shown in Figure 2, highlighting how the module processes pose variations and background noise through separated and dilated convolutions.

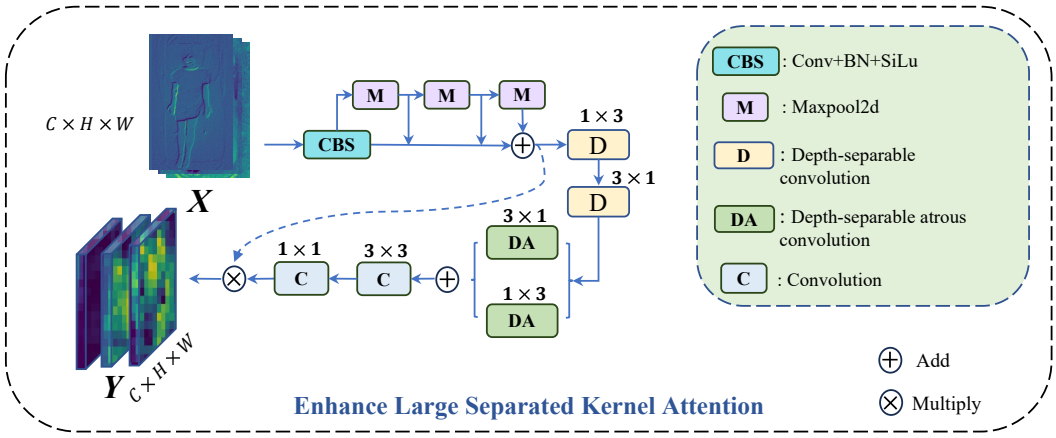


Figure 2. Network architecture of the ELSKA module, showing the separated and dilated convolutions used for spatial feature extraction. The figure highlights how the module processes pose variations and background noise.

3.2. RepNCSPELAN4

While the ELSKA module addresses challenges arising from pose variation in feature extraction, the detection head of HP-YOLO relies on the RepNCSPELAN4 structure to enhance multi-scale feature fusion. This structure proves particularly effective in complex scenes with varying object sizes, such as autonomous driving or security surveillance. By incorporating multi-scale convolutions, CSP (Cross Stage Partial) connections, and the ELAN (Efficient Layer Aggregation Networks) mechanism, RepNCSPELAN4 efficiently aggregates features at different scales and hierarchical levels.

First, multi-scale convolution is performed on the input feature map X using varying kernel sizes $\{W_k\}$, where k represents the kernel size:

$$F_k = X * W_k, \tag{4}$$

yielding feature maps $\{F_k\}$ that capture both fine and coarse details. This design is beneficial in scenes with objects of disparate sizes, such as vehicles, pedestrians, or wildlife.

Next, we introduce CSP connections to reduce redundant computation while retaining essential information. We split X into two parts, X_1 and X_2. Only X_1 is convolved:

$$F_1 = X_1 * W_{\text{CSP}}, \tag{5}$$

where W_{CSP} is the convolution kernel in the CSP path. Then, we concatenate as follows:

$$F_{\text{CSP}} = \text{Concat}(F_1, X_2), \tag{6}$$

ensuring that the resulting F_{CSP} integrates both newly extracted features and the original unaltered ones.

Subsequently, the ELAN mechanism is applied to further aggregate multi-scale features. For instance, if two additional convolution branches produce F_3 and F_4, we combine them via

$$F_{\text{out}} = W_{\text{final}} * \text{Concat}\big(F_{\text{CSP}}, F_3, F_4\big), \tag{7}$$

where W_{final} merges the concatenated features into a final enriched representation. This hierarchical aggregation enables the model to better adapt to small and irregular objects in complex environments.

To optimize runtime, *rep* blocks inside RepNCSPELAN4 can be merged into an equivalent single convolution kernel during inference. This offline transformation (often termed

"RepConv") reduces the deployment complexity and memory footprint without altering the learned weights' representational power.

In summary, RepNCSPELAN4 produces multi-scale and hierarchical feature maps that are fully enriched before the pruning step. By fusing multi-scale convolutions, CSP connections, and ELAN aggregation, our detection head ensures robust feature extraction across diverse object sizes. Figure 3 depicts the overall structure, highlighting the flexible multi-branch design and the final concatenation stage for efficient feature fusion.

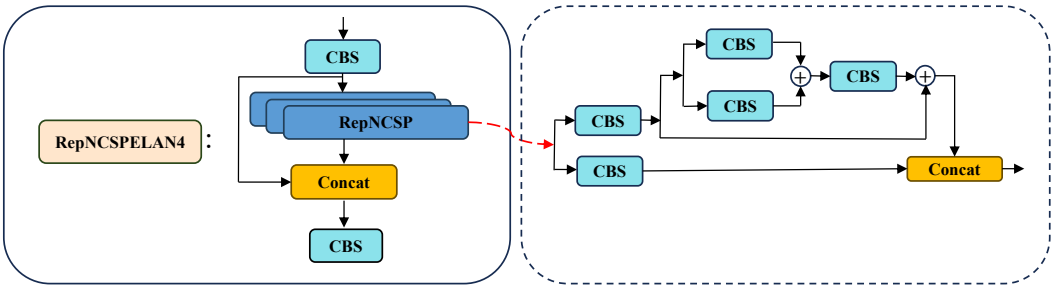


Figure 3. Network architecture of RepNCSPELAN4, illustrating multi-scale convolutions, CSP connections, and the ELAN mechanism for efficient feature fusion. This design captures features at various scales, essential for detecting objects of different sizes.

3.3. Model Pruning and OKS Integration

To further optimize HP-YOLO's performance on keypoint-based tasks, we develop an L1 pruning algorithm guided by the Object Keypoint Similarity (OKS) [32]. The objective is to prune less significant weights while retaining high accuracy in keypoint localization. Such a balance between model compactness and accuracy is crucial for real-time applications (e.g., robotics or video surveillance) where processing speed is as important as detection performance.

By combining OKS with L1 pruning, we ensure that critical weights for accurate keypoint localization are preserved [35]. Specifically, we compute the gradient of OKS with respect to each weight W_j, and define an importance score S_j as follows:

$$S_j = |W_j| \times \left| \frac{\partial \mathrm{OKS}}{\partial W_j} \right|. \tag{8}$$

A higher S_j indicates a stronger contribution to maintaining OKS-based accuracy.

Weights with S_j below a threshold θ are pruned, as outlined in Algorithm 1. After pruning, the model is fine-tuned on the training set to adjust the remaining weights [36], thus minimizing any accuracy drop due to weight removal. Finally, the pruned model is evaluated again to ensure that keypoint detection precision is not significantly degraded.

In Algorithm 1, if $S_j < \theta$, the weight W_j is set to zero and effectively removed from the network. This ensures that weights crucial for keypoint localization (i.e., with higher scores) are retained. Excessive pruning may cause a notable accuracy drop, whereas insufficient pruning yields minimal speed gains; thus, selecting θ is a key design choice. After fine-tuning, the model typically recovers or even surpasses its original accuracy, owing to the removal of redundant parameters.

Algorithm 1 Prune Weights Using OKS

Require: N (number of keypoints), $\{W_j\}$ (model weights), θ (pruning threshold)
Ensure: Pruned weights $\{W'_j\}$
Compute OKS (Equation (9)) between predictions and ground truth.
Compute gradients $G_j = \frac{\partial \text{OKS}}{\partial W_j}$ for all weights $\{W_j\}$.
Compute scores $S_j = |W_j| \times |G_j|$ (Equation (8)).
for each weight W_j **do**
    if $S_j < \theta$ **then**
        $W'_j \leftarrow 0$ // Prune weight
    else
        $W'_j \leftarrow W_j$
    end if
end for
Fine-tune the pruned model on training data.
Evaluate the final model performance (OKS, AP, etc.).

4. Experiment

4.1. Experimental Datasets

Two datasets were utilized for this experiment: the COCO [32] dataset and a smaller domain-specific bedside human pose estimation dataset. The first dataset is a standard benchmark widely used for object detection, segmentation, and pose estimation tasks. It includes 17 keypoints per human for pose estimation. The dataset was used for both training and validation, with a division of the data into training and validation sets, as shown in Table 1. In addition to COCO, we evaluated the model's generalization performance on a smaller bedside human pose estimation dataset [13]. This dataset, consisting of approximately 300 images, was specifically designed to assess the model's ability to handle real-world clinical settings, focusing on patient postures such as lying down or sitting up in bed.

Table 1. Overview of datasets used for training and evaluation.

Dataset	Number of Images
COCO (Train)	56,599
COCO (Val)	2367
Bedside Pose Estimation	300

4.2. Experimental Environments and Hyperparameters

The HP-YOLO model was trained using a high-performance workstation featuring an AMD EPYC 7B13 processor with 22 virtual CPUs (vCPUs) and an NVIDIA RTX 4090 GPU with 24GB of GDDR6 memory (Santa Clara, CA, USA). This hardware provided ample computational power for efficient processing of large batches and complex models. The operating system used was Ubuntu 20.04, and the deep learning framework employed was PyTorch 2.0.0 with CUDA 12.2 for GPU acceleration.

Hyperparameters were carefully tuned to optimize the model's performance for both object detection and human pose estimation. We utilized several data augmentation techniques inherent in the YOLOv8 framework, including Mosaic, MixUp, HSV jittering, random scaling and translation, random rotation, and random erasing. These augmentations effectively improved the model's robustness to variations in image quality, lighting conditions, object pose, and occlusion, crucial for reliable performance in practical scenarios. Specific augmentation parameters are detailed in Table 2.

We employed Stochastic Gradient Descent (SGD) with momentum = 0.9 and weight decay = 0.0005. The initial learning rate was set to 0.0005 and was warmed up linearly from 1×10^{-6} during the first 5 epochs, followed by a cosine annealing schedule reducing it to 1×10^{-5} over a total of 300 epochs. Table 2 summarizes the core training hyperparameters and detailed data augmentation settings.

Table 2. Training hyperparameters and data augmentation settings.

Hyperparameter	Value	Notes
Image Size	640 × 640	Input resolution
Batch Size	8	SGD optimizer
Epochs	300	Cosine LR decay
Initial LR	0.0005	5-epoch warm-up
Weight Decay	0.0005	momentum = 0.9
Data Augmentations		
Random Scale	±10%	
Random Rotation	±10°	
HSV (H, S, V)	(0.015, 0.7, 0.4)	Probability = 1.0 each batch
Mosaic Probability	0.4	Four-image mosaic
MixUp Probability	0.5	Secondary blending

4.3. Evaluation Metrics

This section introduces the evaluation metrics used to assess the performance of the HP-YOLO model in object detection and human pose estimation tasks. These metrics offer a comprehensive understanding of the model's accuracy, computational efficiency, and real-time performance.

OKS [32] is utilized for human pose estimation tasks. It measures the similarity between predicted keypoints and the ground truth keypoints by computing the normalized distance between them. The OKS metric is defined as follows:

$$\text{OKS} = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{9}$$

where d_i represents the Euclidean distance between the predicted keypoint and the corresponding ground truth keypoint, s is the object scale, k_i is a keypoint-specific constant adjusting for object size, and v_i is the visibility flag for keypoint i. OKS is similar to IoU but adapted for keypoint detection.

Average Precision (AP) is the area under the precision–recall curve, evaluating the detection performance at various recall levels. In this study, we report AP^{50} (IoU = 0.50) and $AP^{50:95}$, which averages AP over IoU thresholds from 0.50 to 0.95. The mean AP (mAP) is calculated as follows:

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n} \text{AP}^i \tag{10}$$

where n is the number of object categories, and AP^i represents the average precision for each category.

We also compute AP^M and AP^L, which represent the model's average precision for medium-sized and large-sized objects, respectively:

- AP^M: evaluates objects with an area between 32^2 and 96^2 pixels.
- AP^L: evaluates objects with an area greater than 96^2 pixels.

GMACs (Giga Multiply–Accumulate Operations) measures the computational complexity of the model by calculating the number of floating-point operations required to process a

single input image. A lower GMACs value indicates a more computationally efficient model, which is crucial for resource-constrained environments or real-time applications.

Latency refers to the average time taken by the model to process a single image during inference. It is an important metric for real-time systems, where lower latency translates to faster processing speeds. Latency is measured in milliseconds (ms).

Effective Receptive Field (ERF) quantifies the area of the input image that contributes to a specific output activation in the feature map. Larger ERFs allow the model to capture more global context, improving detection accuracy for larger objects and increasing robustness to occlusions. The ERF is calculated by tracing the model's forward pass and identifying the regions of the input image that influence each output. This method follows the approach from [37], where the HP-YOLO model uses large convolution kernels (e.g., 31×31) to expand its ERF without increasing network depth.

$\mathrm{T-Score}$ (t%) measures the percentage of the receptive field that effectively contributes to the output prediction. It quantifies how much of the input image's information is utilized by the model. A higher T-Score indicates more efficient use of the receptive field, which leads to more accurate predictions in complex scenes.

To mitigate the impact of random variations and ensure the stability of our experimental results, all performance metrics reported in this paper, including AP and AP^{50}, are the average values obtained from five independent experiments. For each model, we conducted five runs with identical settings (training and validation datasets, hyperparameters) but different random seeds. The reported values are the means of these five runs.

4.4. Experimental Results and Analysis

To comprehensively evaluate the performance of our HP-YOLO model, we conducted comparative experiments with several state-of-the-art human pose estimation methods, including lightweight OpenPose, the HRNet series, and the latest SRPose model. All models were evaluated on the COCO validation set under identical conditions, without the use of pre-trained weights. The experimental results are summarized in Table 3.

The results in Table 3 clearly demonstrate that while SRPose [28] achieves an AP score of 48.4 with a parameter size of 23.5 M, it incurs a significant computational overhead. In contrast, our HP-YOLO model, with only 4.08 M parameters, achieves a higher AP score of 53.4 and an AP^{50} score of 82.0, demonstrating both superior detection accuracy and computational efficiency. Specifically, compared to other lightweight models, such as lightweight OpenPose [26] and EfficientHR-H3 [21], HP-YOLO shows a notable improvement. Our HP-YOLO achieves an AP that is 10.6% higher than lightweight OpenPose and 8.6% higher than EfficientHR-H3, highlighting its superior performance in pose estimation tasks while maintaining a compact model size.

Furthermore, we evaluated the robustness of HP-YOLO in challenging scenarios from the COCO validation set, such as single-person, multi-person, partial occlusion, and crowded environments. As illustrated in Figure 4, we present a visual comparison of YOLOv8n and HP-YOLO's performance across increasingly complex scenarios from the COCO validation set. From left to right, the columns represent scenarios of increasing difficulty: single-person, single-person with complex background, multi-person, partial human body (occlusion), and multi-person with complex background. In each scenario, the traditional YOLOv8n algorithm exhibits limitations, such as inaccurate pose estimation or missed detections, while HP-YOLO consistently demonstrates improved accuracy and robustness by effectively reducing these errors. This visual evidence highlights HP-YOLO's enhanced ability to handle diverse and challenging real-world scenarios compared to the baseline model.

Finally, to assess the generalizability of HP-YOLO in real-world clinical settings, we conducted bedside pose estimation experiments involving complex postures and occlusions. Figure 5 visually compares the performance of the baseline YOLOv8 model (top row) and our proposed HP-YOLO model (bottom row) under identical challenging bedside scenarios. Figure 5 highlights HP-YOLO's practical advantages in clinical applications, demonstrating superior bedside pose estimation. Specifically, Figure 5a, illustrating 'Leg Joint Misalignment', shows the baseline model inaccurately localizing distal leg and foot joints, resulting in misalignment, while HP-YOLO accurately captures the posture. Figure 5b, depicting 'Partial Missed Detection' in a 'Cluttered Background', reveals the baseline model's failure to fully detect the person amidst clutter, leading to incomplete pose estimation. Figure 5c, highlighting 'Complex Posture Missed Detection + Misalignment', demonstrates the baseline model's struggle with complex poses and occlusion, resulting in both joint misdetection and misalignment, indicating limited robustness in such scenarios. Error Cause Analysis: Observed limitations of the baseline YOLOv8 in both COCO (Figure 4) and bedside scenarios (Figure 5) indicate its vulnerability to complex visual environments, clutter, occlusion, and unusual poses. HP-YOLO's enhanced architecture, particularly the ELSKA and REPNCSPELAN4 modules, effectively addresses these challenges. These modules likely improve feature extraction, contextual understanding, and robustness to visual variations, enabling HP-YOLO's superior performance in both general and application-specific challenging scenarios, thus demonstrating its practical value for robust human pose estimation. Across all cases, HP-YOLO demonstrates superior accuracy and robustness compared to the baseline model, making it well suited for clinical applications.

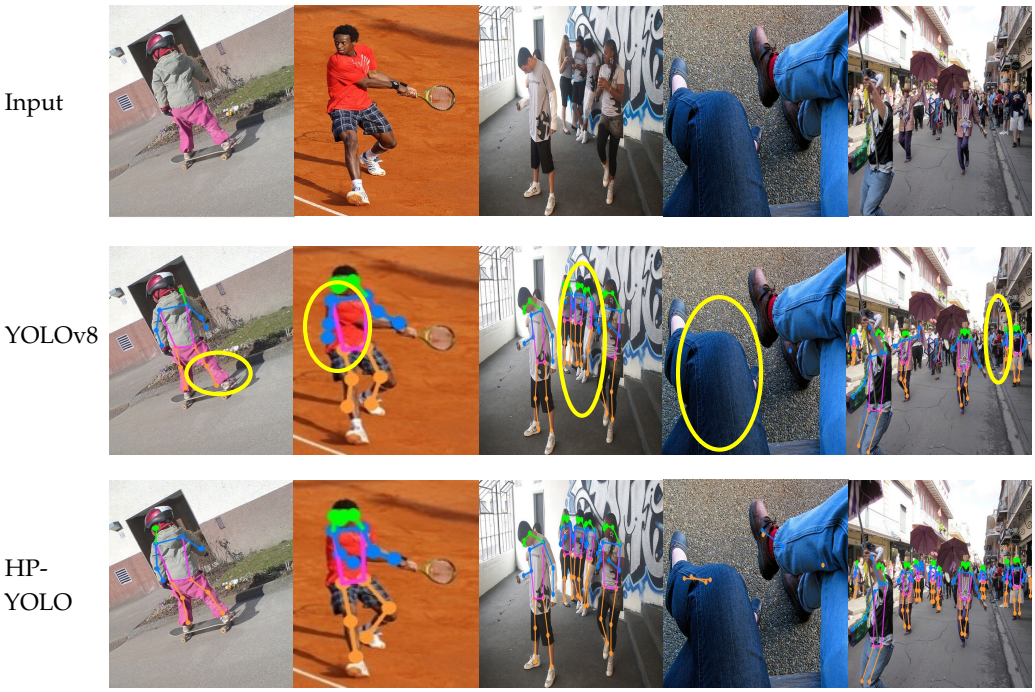


Figure 4. Visual comparison of pose estimation performance on the COCO validation set across scenarios of increasing complexity. Yellow circles are used to highlight areas of significant errors.

Table 3. COCO [32] Validation Set Results for Comparison (Average Over 5 Runs).

Method	Params (M) ↓	GMACs ↓	AP (%) ↑	AP^{50} (%) ↑
Lightweight OpenPose [26]	4.1	18	42.8	-
EfficientHR-H2 [21]	10.3	15.4	52.9	80.5
EfficientHR-H3 [21]	6.9	8.4	44.8	76.7
EfficientHR-H4 [21]	3.7	4.2	35.7	69.6
LitePose-XS [27]	**1.7**	**1.2**	40.6	-
MFite-HRNet [9]	1.8	2.43	41.4	-
SRPose [28]	23.5	30.86	48.4	-
YOLOv5 [17]	11.3	14.41	45.4	77.1
YOLOv8n	3.2	4.35	47.5	79.6
HP-YOLO (Ours)	4.08	5.7	**53.4**	**82.0**
P-HP-YOLO (Ours)	2.7	3.65	52.4	80.7

Note: Bold indicates the best performance in each column; "Ours" denotes our proposed methods. ↓ indicates smaller-is-better metrics, ↑ indicates larger-is-better metrics.

Figure 5. Visualcomparison of bedside pose estimation performance in challenging clinical scenarios. Black dashed boxes highlight error regions (misalignment or missed detections). (**a**) Leg Joint Misalignment; (**b**) Partial Missed Detection; and (**c**) Complex Posture (Missed Detection + Misalignment).

4.5. HP-YOLO Ablation Study

To rigorously assess the effectiveness of our proposed enhancements, we conducted comprehensive ablation experiments on the COCO dataset, utilizing both YOLOv8n and YOLOv5 architectures. This study systematically deconstructs the individual contributions of the ELSKA attention module and the REPNCSPELAN4 module, and critically evaluates

the efficacy of ELSKA by benchmarking its performance against well-established attention mechanisms: CBAM [24] and SE [23] blocks.

For the YOLOv5 model, integrating ELSKA yielded a 0.9% increment in AP and a 3.1% increment in APL, demonstrably enhancing large object detection. Substituting the C2f module with REPNCSPELAN4 independently improved AP by 3.5%. The synergistic combination of ELSKA and REPNCSPELAN4 culminated in a noteworthy AP of 50.5%. To contextualize ELSKA's performance, we benchmarked it against CBAM and SE, with results meticulously detailed in Table 4. ELSKA consistently surpassed these alternatives. Specifically, deploying ELSKA in isolation achieved 46.3% AP, a 0.5% advantage over CBAM (45.8%) and a substantial 1.8% advantage over SE (44.5%). The performance superiority of ELSKA was further accentuated when synergistically integrated with REPNCSPELAN4, attaining 50.5% AP, in contrast to 48.5% with CBAM and 47.3% with SE. Notably, isolated implementation of the SE block within YOLOv5 resulted in a marginal performance regression relative to the baseline, potentially indicating architectural incompatibility under such conditions. Conversely, ELSKA consistently delivered robust performance augmentations across all evaluated configurations within the YOLOv5 framework.

Similarly, for the YOLOv8n model, ELSKA in isolation improved AP by 0.4%, while REPNCSPELAN4 independently contributed a 3.4% AP improvement. Their combined deployment achieved a culminating AP of 53.4%. Comparative evaluations against CBAM and SE on YOLOv8n unveiled a congruent trend. ELSKA, utilized alone, attained 47.9% AP, marginally outperforming CBAM (47.2%) and SE (46.6%). Conjointly employed with REPNCSPELAN4, ELSKA again reached the apex AP of 53.4%, exceeding CBAM+REPNCSPELAN4 (50.4%) and SE+REPNCSPELAN4 (49.2%). These findings collectively reinforce that while CBAM and SE offer incremental enhancements over the 'None' baseline in YOLOv8n, their performance gains remain consistently less pronounced than those afforded by ELSKA. Across both YOLOv5 and YOLOv8n backbones, the consistently superior performance profile of ELSKA, both standalone and in conjunction with the REPNCSPELAN4 module, unequivocally substantiates its selection as the preferred attention mechanism for the HP-YOLO architecture, primarily due to its demonstrably enhanced feature extraction efficacy for object detection.

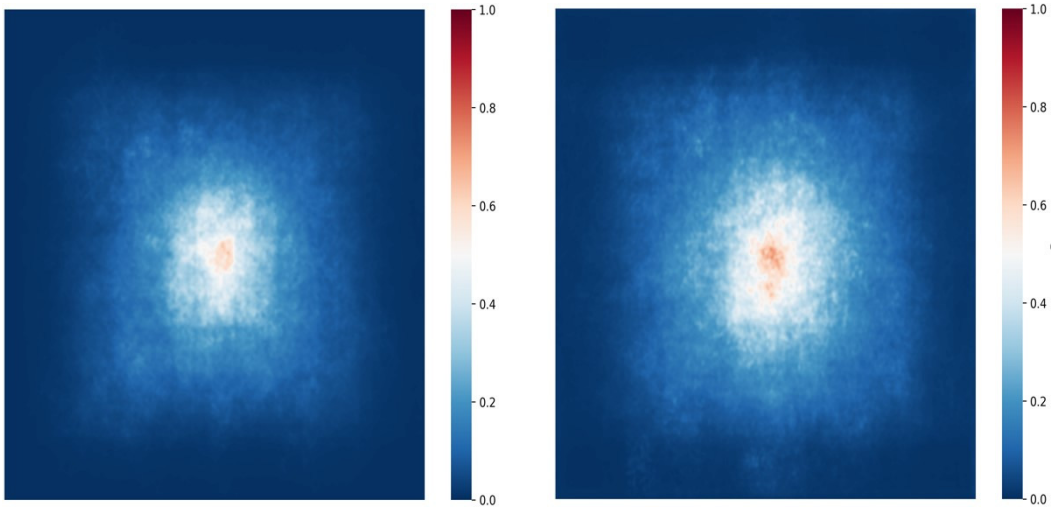
Comprehensive ablation results, encompassing the comparative analysis of attention mechanisms, are meticulously summarized in Table 4.

Table 4. Ablation study results comparing different attention mechanisms and the effectiveness of the REPNCSPELAN4 module (✓ indicates inclusion of the module).

Backbone	Attention	REPNCSPELAN4	AP (%)	AP^{50} (%)
YOLOv5	None		45.4	77.1
	CBAM		45.8	76.4
	SE		44.5	75.9
	ELSKA		46.3	77.9
	CBAM	✓	48.5	79.2
	SE	✓	47.3	78.5
	ELSKA	✓	50.5	81.1
YOLOv8n	None		47.5	79.6
	CBAM		47.2	79.2
	SE		46.6	78.4
	ELSKA		47.9	79.9
	CBAM	✓	50.4	80.6
	SE	✓	49.2	79.8
	ELSKA	✓	53.4	82.0

4.5.1. Effective Receptive Field

To further assess the impact of integrating the ELSKA attention module, we compare the Effective Receptive Field (ERF) [37] across different models, as shown in Table 5 and Figure 6. Integrating the ELSKA module within the SPPF block effectively broadens the network's ERF, enhancing spatial awareness and recognition capabilities. Table 5 presents ERF measurements at various thresholds. As shown, our model consistently achieves a larger ERF than YOLOv8n, highlighting its enhanced spatial feature extraction across different thresholds. Figure 6 visually contrasts the ERF of the baseline model with that of our HP-YOLO model (integrating the ELSKA module). The heatmaps clearly reveal that our model's ERF spans a broader area, significantly improving its capacity to capture meaningful spatial information and contributing to enhanced recognition performance in practical scenarios.



(**a**) Baseline Model (**b**) HP-YOLO Model

Figure 6. Effective Receptive Field heatmaps comparing the baseline model (**a**) and HP-YOLO model (**b**).

Table 5. Effective Receptive Field (ERF) quantitative comparison at different thresholds.

Method	t = 20%	t = 30%	t = 50%	t = 99%
YOLOv8n	3.45	5.7	12.5	76.2
Ours	4.31	7.3	15.4	92.3

4.5.2. Pruning Experiment

To assess the impact of our pruning strategy, Figures 7 and 8 illustrate the changes in channel count and inference speed, respectively. Figure 7 displays a stacked area chart depicting the number of channels before and after pruning, where the orange area represents the pruned channels and the gray area indicates the retained channels. The significant reductions observed across most layers underscore the effective removal of redundant channels, thereby reducing model complexity and improving computational efficiency. To determine an appropriate pruning level, we conducted experiments with varying pruning intensities, denoted as 'L1-OKS X.X' in Figure 8. Here, 'L1-OKS 1.0X' refers to pruning based on a base L1-norm threshold τ. Subsequent configurations, 'L1-OKS 1.5X', 'L1-OKS 1.75X', and 'L1-OKS 2.0X', represent pruning with thresholds scaled by factors of 1.5, 1.75, and 2.0, respectively, i.e., 1.5τ, 1.75τ, and 2.0τ. The base threshold τ was empirically determined by starting with a small value and iteratively increasing it while monitoring the validation set performance, selecting a value that initiated effective pruning

without significant accuracy loss. Figure 8 compares the model's performance metrics before and after pruning, with mean Average Precision (mAP@50) displayed on the primary y-axis (in red) and inference latency on the secondary y-axis (in blue). Observing Figure 8, the 'L1-OKS 1.5X' configuration achieves a 1.4× improvement in inference speed, with only a 1% reduction in mAP@50, demonstrating that the pruning strategy effectively accelerates inference while maintaining competitive accuracy. While 'L1-OKS 1.75X' and 'L1-OKS 2.0X' offer further speed improvements, they also lead to a more pronounced drop in mAP@50. Therefore, we selected the 'L1-OKS 1.5X' configuration as our final pruned model, as it provides an optimal balance between inference speed enhancement and accuracy preservation for real-time bedside HPE applications. This configuration effectively reduces model complexity while maintaining competitive performance, making it well suited for deployment in resource-constrained environments, such as mobile devices. In summary, our approach significantly reduces computational complexity and enhances inference speed, making it well suited for deployment in resource-constrained environments.

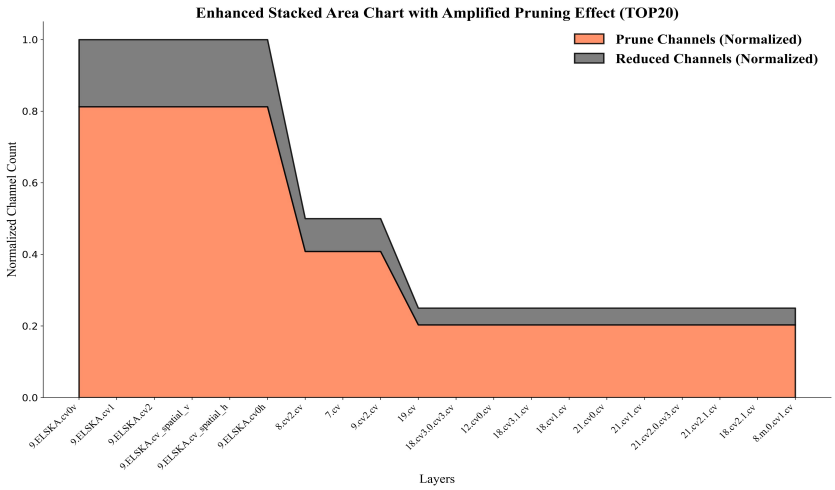


Figure 7. Channel count before and after pruning. The orange area represents pruned channels, and the gray area indicates remaining channels.

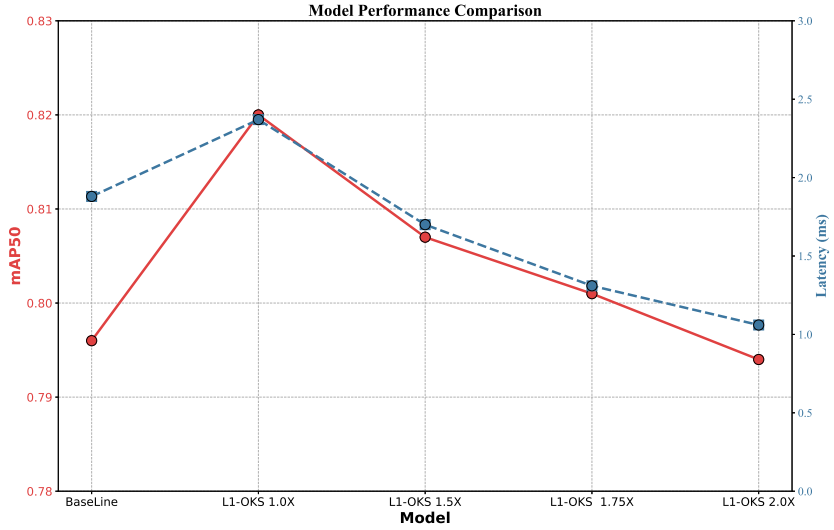


Figure 8. Model performance before and after pruning. Pruning improves inference speed (blue) with minimal impact on mAP@50 (red).

5. Discussion

Although HP-YOLO shows promising results across both the COCO and bedside pose estimation datasets, several limitations warrant further investigation. First, our bedside dataset, while domain-specific, remains relatively small and lacks the breadth needed to ensure robust generalization in diverse clinical settings. As discussed in Section 4, additional data collection and validation on larger, more varied patient cohorts would bolster the reliability of our approach.

Second, the proposed OKS-L1 pruning strategy relies on an empirically chosen threshold. Although we conducted ablation studies (Section 4) to identify a reasonably effective pruning ratio, automated threshold selection or adaptive pruning policies could further optimize accuracy-speed trade-offs, especially under different real-time constraints. Moreover, exploring structured or block-level pruning might maintain critical spatial features more effectively than channel-wise or weight-wise pruning.

Third, while the ELSKA and RepNCSPELAN4 modules successfully address the challenges of pose variation and multi-scale feature fusion, future research could investigate complementary methods, such as attention-based Transformers or graph convolutional networks (GCNs), to further enhance keypoint detection in highly cluttered or occluded scenarios. Likewise, hardware-aware optimizations (e.g., quantization, mixed-precision training) are essential for deployment in resource-constrained environments, aligning with the real-world need for minimal latency in bedside monitoring applications.

Finally, although preliminary bedside experiments (Figure 5) indicate that HP-YOLO can handle complex patient postures, rigorous clinical validation under institutional review is imperative to ascertain its efficacy in routine healthcare workflows. Such studies would examine how HP-YOLO's improved accuracy and speed translate to tangible improvements in patient care, for instance, by reducing the frequency of falls or enhancing intervention timing for pressure ulcer prevention. Ultimately, further expansions to both methodology and dataset diversity are crucial for establishing HP-YOLO as a robust, generalizable tool in real-world medical settings.

6. Conclusions

This study addresses the challenge of human pose estimation in complex backgrounds by proposing the HP-YOLO model, which aims to enhance detection accuracy and efficiency. The model achieves significant improvements through three key innovations: (1) the ELSKA module, leveraging large receptive fields and separable convolutions to mitigate false positives and false negatives in cluttered environments; (2) the RepNCSPELAN4 module, employing multi-scale convolutions and reparameterization techniques to improve detection accuracy for small-scale targets while maintaining high inference speed; and (3) an OKS-guided L1 pruning strategy, which reduces computational complexity and parameter count without sacrificing accuracy. Comprehensive experiments on COCO (Section 4) demonstrate that HP-YOLO achieves higher average precision and faster inference compared to competing lightweight methods. Moreover, bedside experiments indicate its potential in patient-monitoring scenarios, handling complex and often occluded poses more effectively than baseline YOLOv8n. In summary, HP-YOLO markedly enhances real-time performance and efficiency for human pose estimation, making it a strong candidate for deployment in demanding real-world applications, ranging from clinical settings to industrial surveillance.

Author Contributions: Methodology, Z.Q. and X.T.; Software, Z.Q. and K.Y.; Investigation, X.T.; Writing—original draft, Z.Q. and K.Y.; Writing—review & editing, H.T. and X.Z.; Visualization, Z.Q.; Supervision, H.T. All authors have read and approved the final version of the manuscript.

Funding: This work is supported by the Sichuan Science and Technology Program under Grant 2022YFS0032.

Institutional Review Board Statement: Not applicable. The study did not involve human or animal participants.

Informed Consent Statement: Not applicable. The study did not involve human participants.

Data Availability Statement: Data is available upon request, restricted by institutional privacy policies.

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

References

1. Lan, G.; Wu, Y.; Hu, F.; Hao, Q. Vision-based human pose estimation via deep learning: A survey. *IEEE Trans. Hum.-Mach. Syst.* **2022**, *53*, 253–268. [CrossRef]
2. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.* **2023**, *56*, 11. [CrossRef]
3. Liu, W.; Bao, Q.; Sun, Y.; Mei, T. Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Comput. Surv.* **2022**, *55*, 80. [CrossRef]
4. Ahmad, N.; Strand, R.; Sparresäter, B.; Tarai, S.; Lundström, E.; Bergström, G.; Ahlström, H.; Kullberg, J. Automatic segmentation of large-scale CT image datasets for detailed body composition analysis. *BMC Bioinform.* **2023**, *24*, 346. [CrossRef] [PubMed]
5. Hayat, M.; Aramvith, S.; Achakulvisut, T. SEGSRNet for stereo-endoscopic image super-resolution and surgical instrument segmentation. In Proceedings of the 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 15–19 July 2024; pp. 1–4.
6. Hayat, M. Squeeze & Excitation joint with Combined Channel and Spatial Attention for Pathology Image Super-Resolution. *Frankl. Open* **2024**, *8*, 100170.
7. Black, J.; Baharestani, M.M.; Cuddigan, J.; Dorner, B.; Edsberg, L.; Langemo, D.; Posthauer, M.E.; Ratliff, C.; Taler, G. National Pressure Ulcer Advisory Panel's updated pressure ulcer staging system. *Adv. Ski. Wound Care* **2007**, *20*, 269–274. [CrossRef] [PubMed]
8. Cao, D.; Liu, W.; Xing, W.; Wei, X. Human pose estimation based on feature enhancement and multi-scale feature fusion. *Signal Image Video Process.* **2023**, *17*, 643–650. [CrossRef]
9. Li, S.; Dai, J.; Chen, Z.; Pan, J. A lightweight pose estimation network with multi-scale receptive field. *Vis. Comput.* **2023**, *39*, 3429–3440. [CrossRef]
10. Tang, Y.; Wang, Y.; Xu, Y.; Deng, Y.; Xu, C.; Tao, D.; Xu, C. Manifold regularized dynamic network pruning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5018–5028.
11. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
12. Liu, J.; Lei, Q.; Qiao, Y.; Gui, G.; Li, X.; Jin, J.; Wang, W. A visual based robot trajectory teaching method for traditional chinese medical moxibustion therapy. In Proceedings of the 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 11–14 December 2020; pp. 2356–2363.
13. Chen, K.; Gabriel, P.; Alasfour, A.; Gong, C.; Doyle, W.K.; Devinsky, O.; Friedman, D.; Dugan, P.; Melloni, L.; Thesen, T.; et al. Patient-specific pose estimation in clinical environments. *IEEE J. Transl. Eng. Health Med.* **2018**, *6*, 2101111. [CrossRef] [PubMed]
14. Cao, T.; Armin, M.A.; Denman, S.; Petersson, L.; Ahmedt-Aristizabal, D. In-bed human pose estimation from unseen and privacy-preserving image domains. In Proceedings of the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), Kolkata, India, 28–31 March 2022; pp. 1–5.
15. Liu, S.; Yin, Y.; Ostadabbas, S. In-bed pose estimation: Deep learning with shallow dataset. *IEEE J. Transl. Eng. Health Med.* **2019**, *7*, 4900112. [CrossRef] [PubMed]
16. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
17. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
18. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
19. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. Vitpose: Simple vision transformer baselines for human pose estimation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 38571–38584.

20. Pishchulin, L.; Insafutdinov, E.; Tang, S.; Andres, B.; Andriluka, M.; Gehler, P.V.; Schiele, B. Deepcut: Joint subset partition and labeling for multi person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4929–4937.
21. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5386–5395.
22. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
24. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
25. Lau, K.W.; Po, L.M.; Rehman, Y.A.U. Large separable kernel attention: Rethinking the large kernel attention design in cnn. *Expert Syst. Appl.* **2024**, *236*, 121352. [CrossRef]
26. Osokin, D. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. *arXiv* **2018**, arXiv:1811.12004.
27. Wang, Y.; Li, M.; Cai, H.; Chen, W.M.; Han, S. Lite pose: Efficient architecture design for 2d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13126–13136.
28. Wang, H.; Liu, J.; Tang, J.; Wu, G. Lightweight Super-Resolution Head for Human Pose Estimation. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 2353–2361.
29. Wang, T.; Wang, K.; Cai, H.; Lin, J.; Liu, Z.; Wang, H.; Lin, Y.; Han, S. Apq: Joint search for network architecture, pruning and quantization policy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2078–2087.
30. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. Scaled-yolov4: Scaling cross stage partial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13029–13038.
31. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
32. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Cham, Switzerland, 2014; pp. 740–755.
33. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
34. Yu, F. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
35. Hoefler, T.; Alistarh, D.; Ben-Nun, T.; Dryden, N.; Peste, A. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *J. Mach. Learn. Res.* **2021**, *22*, 10882–11005.
36. Liu, Z.; Sun, M.; Zhou, T.; Huang, G.; Darrell, T. Rethinking the value of network pruning. *arXiv* **2018**, arXiv:1810.05270.
37. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11963–11975.