AP-Pose: Enhancing human pose estimation in complex airport scenes with YOLOv8s-Pose and multi-module optimization

H.Z. Shen [1], X.C. Wang [1], G. Dong [2], Y. Ci [3]
[1] School of Electrical Engineering, Shanghai Dianji University, 201306, Shanghai, China, Shuihua Road 300;
[2] Zhengzhou Coal Industry Group, 100039, Zhengzhou, Henan China;
[3] Beijing Simulation Center, 100039, Beijing, China

Abstract

Recognizing human behavior in airports is crucial for security management and public safety. To address the issue of low accuracy in human behavior recognition caused by significant lighting changes and complex backgrounds in airport settings, we propose an improved human pose estimation method based on YOLOv8s-Pose, called the AP-Pose algorithm. Firstly, the PSA module is added to the end of the backbone network to enhance the network's ability to model global features, capture the global relationships between different regions in the image, and filter out interference from complex backgrounds. This enables accurate human target localization and pose estimation. Second, the WASP module is introduced into the neck network to improve the spatial capture capability of the pose estimation network, thereby enhancing the localization accuracy of human joint points. Finally, the Adapter module is employed to enrich the feature representation for pose estimation by incorporating information from the original detection task layer and fine-tuning the pose estimation task using this information. Experiments were conducted on a self-constructed airport human pose dataset, and the results show that our improved method significantly enhances pose estimation performance, achieving a detection accuracy of 94.9 %, with improvements of 3.3 % in precision (P), 3.3 % in recall (R), and 3.5 % in mAP@0.5 compared to the baseline model. This method enhances pose estimation accuracy and robustness in complex airport scenes while maintaining computational efficiency, achieving a 3.5 % mAP@0.5 improvement over YOLOv8s-Pose with a marginal increase in computational complexity (12.5 GFLOPs vs. 10.8 GFLOPs).

Keywords: human pose estimation, airport scene, ap-pose, adapter module, wasp module, psa module.

Citation: Shen HZ, Wang XC, Dong G, Ci Y. AP-Pose: Enhancing human pose estimation in complex airport scenes with YOLOv8s-Pose and multi-module optimization. Computer Optics 2026; 50(1): 1707. DOI: 10.18287/COJ1707.

Introduction

Advances in AI and computer vision have led to the use of human pose estimation in fields like security, transportation, virtual reality, and human-computer interaction. It's key to effective airport security: detecting and monitoring aberrant behavior, and ensuring personnel behavior. However, varied lighting and complex backgrounds in airports challenge the effectiveness of current human pose estimation algorithms.

YOLOv8s-Pose, a lightweight human pose estimation algorithm based on YOLOv8s, has been widely adopted for real-time human pose estimation tasks due to its efficiency and high accuracy. However, its performance in complex airport scenarios is limited by challenges such as occlusions and background interference, necessitating enhancements that balance accuracy and computational efficiency. Therefore, in this paper, we propose AP-Pose, a human pose estimation algorithm tailored for airport scenarios, which introduces improvements to the YOLOv8s-Pose network.

Therefore, in this paper, we propose AP-Pose, a human pose estimation algorithm tailored for airport scenarios, which introduces the improvements to the YOLOv8s-Pose network. Recalling previous networks, starting with the most traditional standard convolutional, the receptive field of standard convolution is highly localized, as it can only focus on a small region of the input data during each convolution operation. This limitation makes it challenging to capture global context, leading to a narrow range of features being learned. In complex scenarios, localized features are insufficient to capture the full range of human body gestures, which can impair the network's learning ability. The proposed attention mechanism allows for a broader set of features to be learned compared to traditional convolutional neural networks. Therefore, a novel network structure, Transformer [1], is emerged, which incorporates an attention mechanism known as the self-attention mechanism. This mechanism allows each location to attend to all other locations in the image, leading to a more comprehensive understanding of the relationships between different parts of the image. It further strengthens the network's capacity for modeling long-range dependencies. To lower computational complexity, we introduce a Partial Self-Attention (PSA) module, which enhances global representation learning, allowing better spatial relationship extraction in complex airport scenes and improving key point localization precision.

Dilated Convolution [2] introduces spacing in the standard convolution, expanding the receptive field of the convolution kernel and improving the network's ability to capture information. Building upon this, the WASP module enhances the sensory field and spatial capture capability of the network by stacking dilated convolutions at different

scales. This enables the network to not only extract local features of human body key points more accurately but also better understand the spatial relationships between each joint, thereby improving the accuracy of human pose estimation in complex scenes, especially in airport scenarios.

The introduction of the Transformer architecture has significantly advanced the field of Natural Language Processing (NLP). Building on this success, Vision Transformers (ViTs) and Transformer-based models have been integrated into computer vision, advancing tasks like image classification [3, 4, 5], segmentation [6], object detection [7, 8, 9],video understanding [10, 11, 12], 3D object recognition [13], and low-level processing [14, 15, 16]. Thanks to its powerful relational modeling capabilities and large-scale pre-training, Transformer models have greatly enhanced the performance of visual recognition [17, 18]. However, the larger size of pre-trained Transformer models, compared to traditional CNN backbone networks, presents a challenge in efficiently fine-tuning these models for specific downstream tasks. To address this, AdaptFormer proposes a lightweight fine-tuning approach that adapts ViTs from their pre-trained state to a model optimized for particular tasks with fewer parameters. Building on this idea, this paper introduces the Adapter module to improve the alignment of features with keypoint predictions by controlling the intensity of keypoint feature influence, ultimately boosting the accuracy and robustness of human keypoint recognition in airport scenarios. Our experiments (Section 3.4) demonstrate that Adapter-based fine-tuning significantly reduces the number of parameters adjusted, enabling efficient adaptation to airport-specific tasks while preserving pre-trained knowledge.

To address these issues, this paper adopts the YOLOv8s-Pose network, which is both efficient and highly accurate, as the foundation and improves it. The AP-Pose algorithm is proposed to enhance its robustness and accuracy in airport scenes. The main innovations are as follows:

1. The Adapter module is introduced to integrate raw detection task layer information into the pose estimation task, enriching the feature representation capability of pose estimation and improving its accuracy and robustness through multi-task joint optimization.
2. WASP module optimization: By adding the WASP module to the neck network, the model's spatial capture ability is significantly enhanced, leading to more accurate localization of human joint points in complex background scenarios.
3. The PSA module design, by incorporating the PSA module at the end of the backbone network, strengthens the global modelling capability of deep information, improving adaptability to low-resolution and occluded scenes, and enhances the understanding of the global relationships between different regions in the image.

1. Related work

1.1. Single-person pose estimate

Single-person pose estimation is mainly divided into regression and heatmap-based methods. The regression method directly predicts key point coordinates from input image features. In contrast, heatmap-based methods generate heatmaps for key points, providing richer supervisory information and keypoint confidence and predicting their locations during post-processing.

1.1.1. Regression method

Toshev and Szegedy [19] proposed DeepPose, which uses a cascaded deep neural network to regress keypoint coordinates from images. However, the regression method relies solely on key point information and does not fully leverage contextual features, resulting in insufficient network stability.

Luvizon et al. [20] (2019) further optimized the regression method by introducing the Soft-argmax function to convert heatmaps into coordinates, addressing the differentiability issue when converting the detection network into a regression network. Additionally, Li et al. [21] (2021) designed a Transformer-based cascade network to capture the spatial correlations between joints and appearances through a self-attention mechanism, thereby improving the accuracy of regression. Li et al. [22] (2021) proposed the normalized flow model, named RLE, for modeling the distribution of joint positions and deriving optimization parameters through residual log-likelihood to improve the accuracy of keypoint prediction. However, regression methods still have limited adaptability to occlusion and background interference in complex scenes.

1.1.2. Heatmap-based approach

To enhance supervision beyond key point coordinates and improve CNN training, recent studies use heatmaps as ground truth representations. Since heatmaps are more stable than direct coordinates, most studies now use heatmap-based methods.

Wei et al. [23] (2016) proposed the Convolutional Pose Machine (CPM), which progressively and accurately predicts the positions of key points on the human body through a multi-stage process. In each stage, the convolutional network utilizes the 2D confidence maps generated in the previous stage to make more fine-grained predictions of body part positions. The stacked hourglass network (SHG) proposed by Newell et al. further enhances pose estimation performance through multi-scale feature fusion. Chou et al. [24] (2018) designed a network incorporating adversarial learning using two stacked hourglass networks with generators and discriminators sharing the same structure. The generator is responsible for estimating joint positions and optimizing keypoint positioning effectiveness by distinguishing between

ground truth and predicted heatmaps. While heatmap-based methods are more stable in predicting the location of key points, their performance is degraded in low-resolution or noisy images.

1.2. Multi-person pose estimation

Compared to single-person pose estimation, multi-person pose estimation is more complex, as it requires the simultaneous detection of the number of targets, their positions, and the grouping of their key points. Based on this, the methods are typically classified into top-down and bottom-up approaches.

1.2.1. Top-down approach

Body regions are detected using off-the-shelf body detectors, and pose estimation is then performed individually for each person in each detection frame.Xiao et al. [25] (2018) appended multiple anti-convolution layers to ResNet's final layer, transforming low-resolution features into a high-resolution heatmap. Cai et al. [26] (2020) proposed a multilevel RSN module that learns local features through efficient intra-layer feature fusion. This module combines with a PRM module to balance local and global feature representations.

Early on, the TransPose model was introduced, which generates keypoint heatmaps through the attention layer and learns evidence for fine-grained pose estimation in occluded scenes. Subsequently, Li et al. [27] (2021) proposed the TokenPose model, which is entirely based on Transformer and captures the association between constraint cues and visual appearance through token representations. Ma et al. [28] (2022) introduced PPT for accurate localization of body regions and efficient estimation of multiview postures. Shi et al. [29] (2022) proposed an attentional mechanism for predicting specific body postures. However, while top-down approaches perform better in pose prediction, their efficiency is often affected by an increase in the number of people, and their adaptability to complex scenarios remains limited.

1.2.2. Bottom-up approach

The process of pose estimation entails the detection of all key points within the image, which are subsequently grouped for analysis.Pishchulin et al. [30] (2016) proposed DeepCut, one of the first two-step bottom-up methods, which is a fast body part detector based on R-CNN. The method first detects all candidate body parts and then combines them into a final human pose using integer linear programming (ILP). However, the DeepCut model is computationally expensive. Cao et al. [31] (2017) developed the OpenPose detector, which predicts key point locations and associates key points with the human body through Convolutional Pose Machines (CPMs) combined with heatmaps and Part Affinity Fields (PAFs), providing a substantial improvement in the speed of multi-person pose estimation. Cheng et al. [32] (2020) developed HigherHRNet, an enhanced version of HRNet, which tackles scale variations in bottom-up multi-person pose estimation by introducing an anti-convolution operation to the high-resolution feature maps produced by HRNet. The above methods are computationally complex, insufficiently adaptable to complex scenes, and perform poorly in low-resolution and occlusion conditions while still struggling with scale variations.

2. Method

In this paper, the YOLOv8s-Pose algorithm is enhanced by incorporating the Adapter module between the detection task layer and the pose estimation layer. Additionally, the PSA module is introduced into the backbone network to improve its feature extraction and global modelling capabilities. The WASP module is also integrated into the neck network to strengthen further the network's ability to capture human pose features. The improved network places a stronger emphasis on the feature representation of human posture, which can be referred to as the human posture estimation network based on the AP-Pose algorithm. Fig. 1 shows its specific structure.

The improved AP-Pose algorithm network consists of three components: Backbone, Neck, and Head. The Backbone serves as the core network, containing modules such as Conv, C2f, SPPF, and PSA, which capture complex feature information through multilayer convolutional and feature extraction structures. Specifically, the SPPF module is responsible for extracting global context features, while the PSA module enhances global relationship modelling through a local self-attention mechanism. The Neck component focuses on feature enhancement and achieves bidirectional multi-scale feature fusion through operations like the WASP module, Upsample, and Concat. The WASP module leverages multiple convolutional layers with varying expansion rates to effectively enhance spatial capture and improve key point localization accuracy. The Head consists of two parts: a target detection head and a human body key point detection head. These two components combine task-level information detection with key point detection via the Adapter module, further enhancing the feature representation capability of pose estimation. The human target position predicted by the target detection head corresponds to the human pose output by the key point detection head. In post-processing, non-maximum suppression (NMS) is used to filter the human detection results, which are then combined with the key point feature map to locate human key points and complete posture estimation.

2.1. PSA module

The Partial Self-Attention Module (PSA) [33] has been developed to address the challenges associated with self-attention mechanisms, which are extensively utilized in various vision tasks due to their advanced global modeling

capabilities. However, these mechanisms' high computational complexity and memory requirements impede efficiency in practical applications.To address this limitation, an efficient Partial Self-Attention Module (PSA) is proposed, which reduces computational costs and resource requirements while maintaining global modeling capabilities for complex and dynamic scenarios in airport scenes. The specific structure of the PSA module is illustrated in Fig. 2.

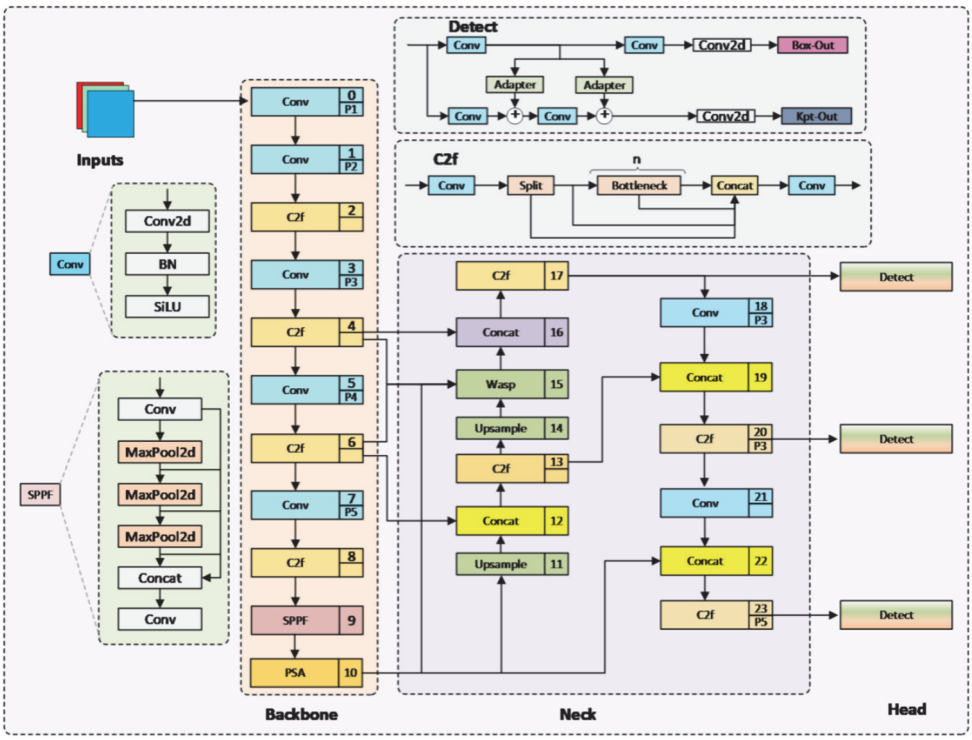

Fig. 1. The structure of the AP-Pose algorithm

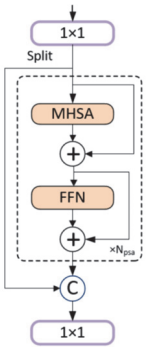

Fig. 2.The structure of the PSA module

The PSA modules reduce computational redundancy through a light-weight design. First, the input features are uniformly split into two parts using a 1×1 convolution, where one part retains the original feature information, and the other is fed into the NPSA block, which consists of the Multihead Self-Attention Module (MHSA) and Feed-forward Network (FFN) for processing to extract long-range dependent features. Subsequently, the two feature parts are concatenated along the channel dimension and fused using another 1×1 convolution to generate the final enhanced feature representation.To improve efficiency, PSA sets query and key dimensions in the MHSA to half of the value dimension, reducing self-attention computational complexity.

The PSA module is placed at the end of the backbone network to perform global modelling of deep information. Conventional convolutional neural networks (CNNs) struggle to effectively capture the correlation of distant regions in an image due to their limitation to a local receptive field. In contrast, PSA allows each feature location to correlate with all locations in the entire image through the self-attention mechanism, enabling comprehensive global relationship modelling. This design significantly enhances the network's ability to capture long-range dependencies and compensates for the limitations of convolutional operations in global feature modelling. By introducing the PSA module into the deeper layers of the network, the model is better able to understand the relationships between different regions in the image, providing more powerful feature representations for the human pose estimation task. This is especially beneficial in complex airport scenes with occlusion, where the design can accurately capture key human pose features, making the network both efficient and robust.

2.2. WASP module

The WASP Module (Waterfall Atrous Spatial Pooling) [34] processes the multiscale features extracted by the backbone network through a cascaded dilated convolution structure, enabling more efficient multiscale feature representation. Unlike traditional parallel multiscale approaches, the WASP module employs a waterfall hierarchical design. This design not only effectively extends the receptive field but also significantly reduces the number of parameters and computational complexity by stacking dilated convolutions layer by layer.The branches of the WASP module are set to expansion rates of (6, 12, 18) in sequence, with each branch generating a feature map with a different receptive field. These branches capture multi-scale information by stacking layers and eventually fusing their outputs into a rich feature representation.

The sensory field and spatial acquisition capabilities are strengthened by applying dilated convolutions at progressively larger expansion rates, improving the spatial capture ability of the pose estimation network and enabling clearer localization of the body's joint points. In the improved network, the number of fused WASP inputs is reduced to three, and the three output layers from the fused backbone network combine the outputs of the WASPs with the inputs of the target detection head. The specific structure is shown in Fig. 3.

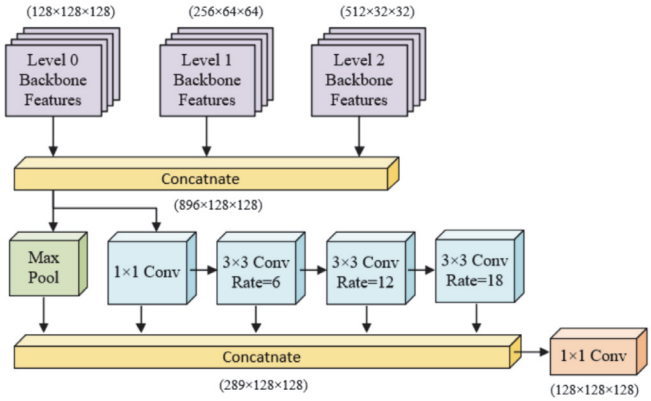

Fig. 3. The structure of the WASP module

The WASP module improves spatial capture in the network by combining larger receptive fields with multi-scale features during extraction, enhancing its understanding and representation of spatial dimensions. The WASP module not only strengthens context modelling in complex scenarios through multi-scale features but also strikes a good balance between computational efficiency and feature representation, providing more accurate and efficient feature support for the improved pose estimation network. With a larger receptive field and multi-scale features, the WASP module is better able to distinguish between interference areas of the human body and the airport background, and can more accurately capture detailed information such as the small joints of the human body.

2.3. Adapter module

In the traditional Transformer model fine-tuning process, a full fine-tuning approach is typically employed, where all the model parameters are adjusted to fit the downstream tasks. While this method offers better performance, it incurs substantial computational overhead. It is prone to causing parameter conflicts between tasks, making it difficult to meet high-efficiency requirements, especially in complex scenarios or multi-task learning.

To address the above issues, AdaptFormer [35] proposes an efficient adaptation method, the Adapter Module (AdaptMLP), which enables efficient fine-tuning of the pre-trained model through a lightweight design while avoiding full tuning of the original parameters. The Adapter module employs a parallel bottleneck structure to separate features into task-independent generic features and task-specific features, which are processed and fused through different branches.The left branch retains the original MLP layer's structure and weights to preserve the generalization capability of the pre-trained model's features. In contrast, the right branch incorporates a new lightweight module composed of a downscaled projection layer ($\mathbf{W}_{\text{down}}$), a ReLU activation function, and an upscaled projection layer ($\mathbf{W}_{\text{up}}$), which is designed to extract and adapt task-relevant feature representations.

The specific formula for the feature transformation in the right branch is as follows:

$$\widetilde{x_l} = \text{ReLU}(\text{LN}(x_l') \cdot \mathbf{W}_{\text{down}}) \cdot \mathbf{W}_{\text{up}}, \tag{1}$$

where x_l' is a specific input feature, $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d \times \hat{d}}$, $\mathbf{W}_{\text{up}} \in \mathbb{R}^{\hat{d} \times d}$, and $\hat{d}$ is the bottleneck dimension, typically satisfying $\hat{d} \ll d$. The outputs of the left and right branches are fused with the scaling s via residual linkage, as shown in the following equation:

$$x_l = \text{MLP}(\text{LN}(x_l')) + s \cdot \widetilde{x_l} + x_l', \tag{2}$$

where s is a trainable scaling parameter used to regulate the impact of task-specific features on the overall network, with this design, the Adapter module achieves precise optimization of task-relevant features while preserving the stability of the pre-trained weights.

We adopt the concept of AdaptFormer and improve it in the head network by using the Adapter module to integrate the original detection task layer information, thereby fine-tuning the pose estimation task and enhancing its feature representation. The specific details are shown in Fig. 4.

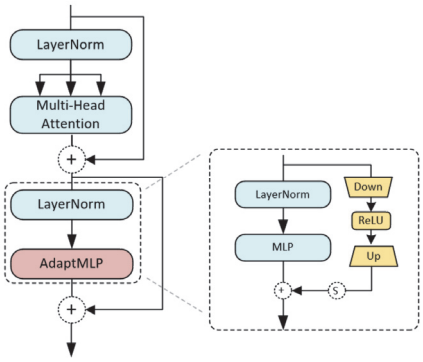

Fig. 4. The structure of the Adapter module

In the AP-Pose algorithm network, the Adapter module is introduced between the detection task layer and the pose estimation layer to enhance the feature transfer and sharing capabilities between tasks. By incorporating contextual information from the detection task into the pose estimation task, the Adapter module effectively enriches the feature representation for keypoint detection. In the airport scenario, where human targets exhibit complex distributions and diverse postures, the Adapter module helps the network better adapt to the complex environment and improve the localization accuracy of key human points through lightweight feature fine-tuning. Moreover, the Adapter module only requires optimizing a small number of parameters compared to a fully fine-tuned approach, significantly reducing the computational overhead for training and inference. The introduction of the Adapter module not only resolves the parameter conflict issue in multi-task learning but also enhances the adaptability and robustness of the AP-Pose algorithm network in complex airport scenarios through efficient feature fusion.

3. Experiments

The study's experimental platform is an Intel(R) Xeon(R) Silver 4214R CPU @ 2.40 GHz with 90 GB of RAM. The GPU is an RTX 3080 Ti with 12 GB of system video memory. It uses Ubuntu 20.04 and Python 3.8, with Pytorch 1.11.0 for machine learning. Both the training and validation sets for the experiments use our self-constructed human pose dataset for airport scenes, with 5672 images in the training set and 1418 images in the validation set. The SGD (Stochastic Gradient Descent) optimizer was used for network optimization during training, with a total of 200 training epochs. The input image size is 640×640, and a cosine annealing learning rate scheduler was employed to adjust the learning rate dynamically. The initial learning rate of 0.01 was gradually decreased at epochs 130 and 150.

To ensure a fair comparison and avoid domain bias from pre-trained models trained on general datasets like COCO, all models in the ablation and comparison experiments were trained from scratch on our airport dataset. The dataset's 5,672 training images, augmented with random scaling and flipping, provided sufficient diversity for convergence, as evidenced by the stable training curves in Fig. 5.

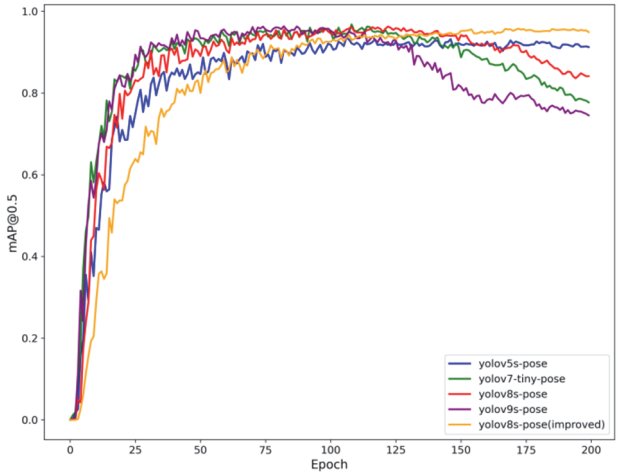

Fig. 5. Comparison curves of mAP@0.5 with different models

Our self-built airport dataset contains 7,090 images covering a variety of scenes in different airport environments, including terminals, check-in counters, and security checkpoints. These images were taken under a variety of lighting conditions, such as daytime, nighttime, and cloudy days, and noise, different lighting changes, and fog effects were added to the background to simulate the complexity of the real environment. Each image may contain crowded scenes with multiple people, luggage occlusions, and dynamic signs, presenting the challenges of high noise interference and changing environments. The dataset was manually annotated in COCO format, with 17 keypoints per person, and split into 5,672 training and 1,418 validation images. Compared to the COCO dataset, which primarily includes general daily activities with an average of 2-3 individuals per image, our dataset emphasizes high-density crowd scenarios (average of 10 individuals per image) and security-relevant poses, presenting unique challenges such as frequent occlusions and background clutter, making it particularly suited for evaluating pose estimation in airport surveillance applications.

Proper preprocessing of the input data is essential before training begins. First, our self-constructed human pose video of an airport scene is cropped into an image dataset for subsequent training. Next, the human pose dataset is labelled, and after labelling, its format is converted into the COCO dataset format for easy training. To enhance dataset diversity and model generalization, data augmentation techniques like random scaling and flipping are applied.In this experiment, the OKS (Object Keypoint Similarity) metric is the evaluation criterion, the OKS is selected: AP50 denotes the prediction accuracy when the OKS threshold is set to 0.5, and mAP denotes the average prediction accuracy when the OKS threshold is varied between 0.5, 0.55, ..., 0.90, and 0.95.

Precision (P) and Recall (R) were computed based on the Object Keypoint Similarity (OKS) metric, with a confidence threshold of 0.5 for keypoint detection. Specifically, a keypoint prediction is considered a true positive if its OKS score with the ground truth exceeds 0.5, following the COCO evaluation protocol. OKS thresholds ranged from 0.5 to 0.95 for mAP calculations, with mAP@0.5 reported as the primary metric for comparison. The OKS score for a person instance is calculated as:

$$\mathrm{OKS} = \frac{1}{K}\sum_{i=1}^{K} \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right), \tag{3}$$

where K is the number of keypoints, d_i is the Euclidean distance between the predicted and ground truth keypoint i, s is the object scale (square root of the bounding box area), and k_i is the per-keypoint constant reflecting annotation uncertainty. This evaluation ensures robust assessment of keypoint predictions across varied scenarios.

3.1. Ablation studies

To evaluate the effectiveness of each module, a series of ablation studies were conducted using our self-constructed airport field human pose dataset. These studies aimed to verify the feature extraction capability of each module in the model and assess their impact on the accuracy of human key point prediction. No pre-trained models were loaded during the training process. Each experiment tests the addition of a single module (Adapter, WASP, or PSA) to the YOLOv8s-Pose baseline, with the final AP-Pose model incorporating all three modules. The results of the experiments are shown in Table 1.

Tab. 1. Ablation study on individual modules of AP-Pose

Method	P	R	mAP@0.5
YOLOv8s-Pose (Baseline)	91.6%	91.6%	91.7%
YOLOv8s-Pose + Adapter	93.4%	93.4%	93.5%
YOLOv8s-Pose + WASP	92.2%	92.2%	91.7%
YOLOv8s-Pose + PSA	94.2%	94.3%	94.6%
AP-Pose (All Modules)	94.9%	94.9%	95.2%

As the table shows, the Adapter, WASP, and PSA modules improve model performance. Adding the Adapter module alone results in a 1.8 % increase in Precision (P), a 1.8 % increase in Recall (R), and a 1.8 % improvement in mAP@0.5 compared to the baseline, demonstrating its effectiveness in enhancing feature representation through multi-task optimization. These results substantiate the efficacy of the implemented improvement strategy. The fine-tuning approach, which incorporates target-detected features into the human keypoint prediction path, enhances feature representation, improving the model's ability to predict key points. When adding the WASP module, the improved model shows a 0.6 % increase in P and a 0.6 % increase in R, validating the effectiveness of this enhancement. The multi-scale fusion mechanism enhances the receptive field of feature maps, extracts contextual information, and improves detection accuracy.

Furthermore, the PSA module has been shown to enhance performance metrics such as P, R, and mAP@0.5 by 2.6 %, 2.7 %, and 2.9 %, respectively, when compared to the baseline model, demonstrating the effectiveness of this strategy. The incorporation of the self-attention mechanism improves global feature modelling and enhances the capture of key point-related regional features. Finally, the simultaneous addition of the Adapter, WASP, and PSA modules leads to a 3.3 % increase in P, a 3.3 % increase in R, and a 3.5 % increase in mAP@0.5, significantly enhancing human pose estimation performance.

3.2. Comparison experiment

To assess the effectiveness of the AP-Pose algorithm, we conducted comparative experiments with various human pose estimation algorithms, using our self-constructed airport scene human pose dataset for training and validation. None of the methods employed pre-trained models, and the experiments were performed on the validation set.

Tab. 2. Performance comparison of different models

Method	P	R	mAP@0.5
YOLOv8s-pose	91.6%	91.6%	91.7%
YOLOv5s-pose	91.4%	91.4%	91.9%
YOLOv7-tiny	93.4%	93.4%	93.1%
YOLOv9s-pose	91.9%	91.9%	90.6%
AP-Pose	94.9%	94.9%	95.2%

Through experiments, we found that the improved model outperforms other human pose estimation algorithms in terms of Precision (P), Recall (R), and mAP@0.5. Specifically, the Precision of the predicted skeletal points increased by 3.5 %, 1.5 %, and 3 % compared to the YOLOv5s-pose, YOLOv7-tiny, and YOLOv9s-pose algorithms, respectively. Additionally, mAP@0.5 improved by 3.3 %, 2.1 %, and 4.6 %, respectively. These results further validate that our improved algorithm, AP-Pose, enhances human feature extraction and contributes to the human keypoint regression task. The specific performance of the comparison algorithms is shown in Tab. 2, and the comparison curves are presented in Fig. 5. Fig. 6 shows the detection results of the AP-Pose human pose estimation algorithm.

Fig. 6. AP-pose detection results in challenging airport scenes (first row from left to right 1-3, second row from left to right 4-6): (1) hunched in front of cluttered landing gear, (2) hidden under propeller, (3) cluttered background with signage, (4) low contrast background, (5) low light background, (6) high contrast background

3.2.1. Generalization on COCO keypoints dataset

To demonstrate the generalizability of AP-Pose beyond airport-specific scenarios, we evaluated its performance on the COCO Keypoints validation set (5,000 images), which includes diverse daily activities with an average of 2-3 individuals per image. Tab. 3 compares AP-Pose with baseline models, where all models are pre-trained with the backbone on the ImageNet dataset and the whole model on the COCO dataset to ensure a fair comparison. AP-Pose achieves an mAP@0.5 of 69.2 %, outperforming YOLOv8s-Pose by 2.4 %, indicating its robustness across varied scenarios. These results complement our airport dataset experiments, confirming AP-Pose's adaptability to both general and specialized domains.

Tab. 3. Performance comparison on COCO keypoints dataset

Method	P	R	mAP@0.5
YOLOv8s-Pose	65.3%	66.1%	66.8%
YOLOv5s-Pose	64.3%	64.8%	65.5%
YOLOv7-Tiny	66.3%	66.4%	67.4%
YOLOv9s-Pose	65.2%	65.6%	65.9%
AP-Pose	67.9%	68.7%	69.2%

3.3. Computational complexity analysis

To address the computational efficiency of AP-Pose, we evaluated its complexity in terms of floating-point operations (FLOPs), inference time, and parameter count, compared to the baseline YOLOv8s-Pose. The experiments were conducted on an RTX 3080 Ti GPU with CUDA 12.4, cuDNN 8.0, and PyTorch 2.1 as the deep learning framework. The input images had a resolution of 640×640. As shown in Table 4, AP-Pose requires 12.5 GFLOPs, an average inference time of 8.2 ms per image, and 11.8M parameters, compared to 10.8 GFLOPs, 7.5 ms, and 11.2M parameters for YOLOv8s-Pose. The marginal increase in complexity is primarily due to the addition of the PSA and Adapter modules, which introduce lightweight self-attention and feature fusion mechanisms. Despite this, AP-Pose's 3.5 % mAP@0.5 improvement justifies the trade-off, making it suitable for real-time airport surveillance applications.

Tab. 4. Computational complexity comparison

Method	FLOPs	Inference Time (ms)	Parameters (M)	mAP@0.5
YOLOv8s-Pose	10.8	7.5	11.2	91.7%
AP-Pose	12.5	8.2	11.8	95.2%

3.4. Fine-Tuning experiments

To validate the efficiency of the Adapter module for fine-tuning, as emphasized in the introduction, we conducted experiments using a YOLOv8s-Pose model pre-trained on the COCO Keypoints dataset. The pre-trained model was fine-tuned on our airport dataset with two configurations: (1) full fine-tuning, adjusting all parameters, and (2) Adapter-based fine-tuning, optimizing only the Adapter module's parameters (approximately 10 % of the total). As shown in Tab. 5, Adapter-based fine-tuning achieves a comparable mAP@0.5 of 94.7 % with significantly lower computational overhead (0.12M parameters adjusted vs. 11.2M for full fine-tuning), confirming the Adapter module's effectiveness for task-specific optimization in airport scenarios.

Tab. 5. Fine-Tuning performance with COCO pre-trained model

Method	P	R	mAP@0.5	Parameters(M)
YOLOv8s-Pose (Scratch)	91.6%	91.6%	91.7%	11.2
Full Fine-Tuning	94.8%	94.8%	95.0%	11.2
Adapter-Based Fine-Tuning	94.6%	94.6%	94.7%	0.12
AP-Pose (Scratch)	94.9%	94.9%	95.2%	11.8

Conclusion

Human pose estimation algorithms face significant challenges in terms of accuracy and robustness when applied to airport scenes with complex backgrounds. The YOLOv8s-Pose algorithm, while being a lightweight and efficient method for human pose estimation, excels in detection speed and accuracy but still suffers from issues such as leakage and misdetection in airport environments. It also faces challenges such as strong background interference, occlusion of human targets, and variations in pose changes. To address these problems, this paper improves the YOLOv8s-Pose algorithm and proposes an enhanced method specifically designed for human pose estimation in airport scenes, referred to as the AP-Pose algorithm.

Specifically, this paper first efficiently transfers information from the detection task layer to the pose estimation task by introducing the Adapter module. This approach exploits detection features to enhance pose estimation, optimize regression accuracy of human key points, and address the issue of inadequate feature expression caused by the complex airport background. Second, the WASP module is introduced to enhance the network's spatial capture capability by utilizing a multi-scale dilated convolution design. This enables the network to more accurately capture the multi-level and multi-scale spatial features of human key points in airport scenarios, improving both keypoint localization accuracy and the detection capability of human targets. Finally, the PSA module enhances the network's ability to model long-range dependencies between targets and backgrounds in airport scenarios by modelling global relationships through the partial self-attention mechanism, thereby significantly improving pose estimation performance in airport environments.

Experiments on a self-constructed human pose dataset for airport scenes demonstrate that the improved algorithm significantly enhances human pose estimation in airport environments. The AP-Pose model demonstrates 94.9% detection accuracy, 3.3% Precision(P) enhancement, 3.3% Recall(R) enhancement, and 3.5% mAP@0.5 improvement compared to the baseline model. The experimental results confirm the effectiveness and robustness of AP-Pose in complex airport scenarios, where it not only better adapts to these environments but also significantly reduces both leakage and false detection rates.

References

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30. DOI: 10.5040/9781350101272.00000005 .

[2] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.

[3] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[4] Liang Y, Ge C, Tong Z, et al. Not all patches are what you need: Expediting vision transformers via token reorganizations. arXiv preprint arXiv:2202.07800, 2022.
[5] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows.Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.DOI:10.1109/iccv48922.2021.00986.
[6] Xie E, Wang W, Yu Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems, 2021, 34: 12077-12090.
[7] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers.European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.
[8] Chi C, Wei F, Hu H. Relationnet++: Bridging visual representations for object detection via transformer decoder. Advances in Neural Information Processing Systems, 2020, 33: 13564-13574.
[9] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020.
[10] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding?ICML. 2021, 2(3): 4.
[11] Arnab A, Dehghani M, Heigold G, et al. Vivit: A video vision transformer.Proceedings of the IEEE/CVF international conference on computer vision. 2021: 6836-6846. DOI: 10.1109/iccv48922.2021.00676.
[12] Fan H, Xiong B, Mangalam K, et al. Multiscale vision transformers.Proceedings of the IEEE/CVF international conference on computer vision. 2021: 6824-6835. DOI: 10.1109/iccv48922.2021.00675.
[13] Chen S, Yu T, Li P. Mvt: Multi-view vision transformer for 3d object recognition. arXiv preprint arXiv:2110.13083, 2021.
[14] Chen H, Wang Y, Guo T, et al. Pre-trained image processing transformer.Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 12299-12310. DOI: 10.1109/cvpr46437.2021.01212.
[15] Liang J, Cao J, Sun G, et al. Swinir: Image restoration using swin transformer.Proceedings of the IEEE/CVF international conference on computer vision. 2021: 1833-1844. DOI: 10.1109/iccvw54120.2021.00210.
[16] Wang Z, Cun X, Bao J, et al. Uformer: A general u-shaped transformer for image restoration.Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 17683-17693. DOI:10.1109/cvpr52688.2022.01716.
[17] Chen X, Xie S, He K. An empirical study of training self-supervised vision transformers.Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9640-9649. DOI: 10.1109/iccv48922.2021.00950.
[18] Pan T, Song Y, Yang T, et al. Videomoco: Contrastive video representation learning with temporally adversarial examples.Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 11205-11214. DOI: 10.1109/cvpr46437.2021.01105.
[19] Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks.Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1653-1660. DOI: 10.1109/cvpr.2014.214.
[20] Luvizon D C, Tabia H, Picard D. Human pose regression by combining indirect part detection and contextual information. Computers Graphics, 2019, 85: 15-22. DOI: 10.1016/j.cag.2019.09.002.
[21] Li K, Wang S, Zhang X, et al. Pose recognition with cascade transformers.Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 1944-1953. DOI: 10.1109/cvpr46437.2021.00198.
[22] Li J, Bian S, Zeng A, et al. Human pose regression with residual log-likelihood estimation.Proceedings of the IEEE/CVF international conference on computer vision. 2021: 11025-11034. DOI: 10.1109/iccv48922.2021.01084.
[23] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines.Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 4724-4732. DOI: 10.1109/cvpr.2016.511.
[24] Chou C J, Chien J T, Chen H T. Self adversarial training for human pose estimation.2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018: 17-30. DOI: 10.23919/apsipa.2018.8659538.
[25] Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking.Proceedings of the European conference on computer vision (ECCV). 2018: 466-481.
[26] Cai Y, Wang Z, Luo Z, et al. Learning delicate local representations for multi-person pose estimation.European conference on computer vision. Cham: Springer International Publishing, 2020: 455-472.
[27] Li Y, Zhang S, Wang Z, et al. Tokenpose: Learning keypoint tokens for human pose estimation.Proceedings of the IEEE/CVF International conference on computer vision. 2021: 11313-11322. DOI: 10.1109/iccv48922.2021.01112.
[28] Ma H, Wang Z, Chen Y, et al. Ppt: token-pruned pose transformer for monocular and multi-view human pose estimation.European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 424-442.
[29] Shi D, Wei X, Li L, et al. End-to-end multi-person pose estimation with transformers.Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11069-11078. DOI: 10.1109/cvpr52688.2022.01079.
[30] Pishchulin L, Insafutdinov E, Tang S, et al. Deepcut: Joint subset partition and labeling for multi person pose estimation.Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4929-4937. DOI: 10.1109/cvpr.2016.533.
[31] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields.Proceedings of the IEEE conference on computer vision and pattern recognition.2017: 7291-7299. DOI: 10.1109/cvpr.2017.143.
[32] Cheng B, Xiao B, Wang J, et al. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation.Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5386-5395. DOI: 10.1109/cvpr42600.2020.00543.
[33] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems, 2025, 37: 107984-108011.
[34] Artacho B, Savakis A. Bapose: Bottom-up pose estimation with disentangled waterfall representations.Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 528-537. DOI: 10.1109/wacvw58289.2023.00059.

[35] Chen S, Ge C, Tong Z, et al. Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems, 2022, 35: 16664-16678.

Authors' information

Hanzhi Shen received the bachelor's degree in electrical engineering and automation from Dalian Jiaotong University, Dalian, Liaoning, China. She is currently studying for a master's degree in Electrical Engineering at Shanghai Dianji University. Her research interests are computer vision and human pose estimation.
E-mail: *17866543979@163.com*

Xichao Wang is an Associate Professor at Shanghai Dianji University.He graduated from Nanjing University of Aeronautics and Astronautics, majoring in Navigation Guidance and Control, with a PhD degree in Engineering in 2014. Prior to that, he pursued his master's degree at Xi'an University of Science and Technology and graduated in 2009.From 2016 to 2017, he worked at Shanghai ZhiShi Intelligent Technology Co. Ltd, as a machine vision researcher and robot hand-eye collaboration project leader. Since August 2014, he has been teaching at Shanghai Dianji University. For more than a decade, his research focuses on key technologies in the fields of machine vision, artificial intelligence, and flight control. As the main person in charge, he has completed a number of national natural funds, etc.
E-mail: *wangxc@sdju.edu.cn*

Gang Dong is a researcher of Zheng Coal Group. His research interests include artificial intelligence and machine learning. His work focuses on developing AI-driven technologies for industrial automation and intelligent monitoring. He is involved in projects related to defect detection, safety personnel behavior monitoring, and predictive maintenance.
E-mail: *dg5594@aliyun.com*

Yan Ci is a researcher at Beijing Simulation Center. His research focuses on the areas of artificial intelligence, computer vision and deep learning. His work focuses on developing advanced AI models and algorithms for real-world applications. He is involved in projects related to target detection, pose estimation and scene understanding.
E-mail: *ciyan1995@qq.com*.

Received March 18, 2025. The final version – September 23, 2025.