IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 5 July 2024, accepted 29 July 2024, date of publication 2 August 2024, date of current version 2 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3437359

RESEARCH ARTICLE

PW-YOLO-Pose: A Novel Algorithm for Pose Estimation of Power Workers

QIN SU[1,2], JULING ZHANG[3], MINGJU CHEN[1,2], AND HONGMING PENG[1,2]

[1] School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China
[2] Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Yibin 644000, China
[3] Power Internet of Things Key Laboratory of Sichuan Province, Chengdu 610095, China

Corresponding author: Mingju Chen (chenmingju@suse.edu.cn)

This work was supported in part by the Natural Science Foundation of Sichuan under Grant 2023NSFSC1987, in part by the Opening Fund of Power Internet of Things Key Laboratory of Sichuan Province under Grant PIT-F-202303, in part by the Key Science and Technology Program of Zigong Municipality under Grant 2020YGJC25, in part by the Opening Fund of Artificial Intelligence Key Laboratory of Sichuan Province under Grant 2023RYY05, and in part by Sichuan University of Science and Engineering Postgraduate Innovation Fund Project under Grant Y2023289.

ABSTRACT To address the detection challenges of keypoints, such as misdetections and omissions caused by backgrounds, occlusions, small targets, and extreme viewpoints in complex electrical power operation environments for power workers. This study proposes a 2D pose estimation algorithm for power workers based on YOLOv5s6-Pose: PW-YOLO-Pose. In this study, the detection rate of occluded keypoints is improved by embedding the Swin Transformer encoder in the top layer of the backbone network. The proposed BiFPN (a weighted bi-directional feature pyramid network) structure with a small target detection layer improves the detection rate of small target characters and the precision of their keypoints'detection. The keypoint regression precision is improved overall by using CA (coordinate attention) in the model neck and improving the bounding box regression loss function to Wise-IoU. The algorithmic model in this study demonstrates excellent detection and largely meets the real-time requirements on the proposed power worker pose estimation dataset in this study. The $mAP_{0.5}$(The mean average precision when the threshold for object keypoint similarity is set to 0.5.) and $mAP_{0.5:0.95}$ are 93.35% and 64.75% respectively, which are 5.22% and 1.53% higher than the baseline model. The detection time of a single image is 21.3 ms, respectively. It can serve as a valuable theoretical foundation and reference for behavior recognition and state monitoring of power workers in intricate electrical power operation environments.

INDEX TERMS Electrical power operation, pose estimation, YOLO-Pose, detection of keypoints.

I. INTRODUCTION

Energy is an important material basis for economic and social development. Electricity is the core of the system to build modern energy [1]. The normal operation of the power system can't be separated from the hard work of power workers. However, power workers constantly face various dangers in electrical maintenance [2]. Complex electrical operating environments and the improper operations of power workers lead to various safety hazards during electrical power operations [3]. Therefore, developing an intelligent safety warning system for power workers can help improve their safety awareness and reduce the occurrence of accidents during power operations [4]. Skeletal keypoints detection of power workers is a prerequisite for pose recognition and abnormal behavior analysis during operations. It is the core of implementing safety warnings for power workers, providing a theoretical basis and reference for behavior recognition and status monitoring of power workers [5].

Currently, significant advancements have been achieved in deep learning-based algorithms for 2D multi-person human pose estimation both domestically and internationally. In comparison to traditional methods, these deep learning-based approaches offer notable advantages such as improved precision [6], real-time performance, end-to-end learning capabilities, and adaptability to diverse scenarios,

The associate editor coordinating the review of this manuscript and approving it for publication was Zhongyi Guo.

poses, and lighting conditions [7]. Consequently, they have emerged as a prominent research area in contemporary studies.

The design methodologies for deep learning-based 2D multi-person human pose estimation can be broadly classified into two main approaches: top-down and bottom-up methodologies. The top-down human pose estimation algorithm initially identifies all human targets within the image, followed by performing keypoint detection for each individual target. Fang et al. proposed a novel pose estimation framework consisting of three components: symmetric spatial transformation network (SSTN), parameterized pose non-maximal suppression (NMS), and pose-guided proposal generator (PGPG), which efficiently solves the problem of inaccurate human pose estimation with inaccurate human bounding boxes [8]. Li et al. proposed a novel regression paradigm with residual log-likelihood estimation to capture potential output distributions, which enhances human pose regression without any test time overhead [9]. The bottom-up human pose estimation algorithm initially predicts the keypoints of all individuals in the image upon input, followed by a matching algorithm that combines these key points to obtain an individual representation. Subsequently, the joint points are connected to derive the human skeleton map. Cheng et al. proposed HRNet, which is capable of maintaining high-resolution characterization throughout the process as well as connecting multi-resolution sub-networks in parallel for multiple multi-scale fusion, ensuring higher-precision heat maps of keypoints and better positional precision [10]. Cao et al. proposed the OpenPose model, which presents the first bottom-up representation of correlation scores through partial affinity fields, where the partial affinity fields are used to encode the location and orientation of the limb in the image, doing so with high-quality results at a fraction of the computational cost [11].

Currently, there is little research on pose estimation for power operations and complex industrial scenarios. The algorithms mentioned in this study have shown good detection performance on some large public pose estimation datasets. However, the complexity of the electrical power operation environments gives rise to occlusion of keypoints, presence of small targets, and extreme viewpoint poses due to the worker's flexible and changeable limbs as well as camera view during work. The aforementioned situation may lead to the absence of crucial feature information, disrupt the correlation among joints, and result in the omission and misdetection of keypoints. To address the issue of occlusion, Bai et al. proposed CONet [12]. They introduced a new structure called COHead, which uses two branches to separately estimate the poses of the occluder and the occluded. By incorporating an attention mechanism, the network achieves differential learning between the two branches, thereby enhancing feature representation. However, this network structure lacks a multi-scale architecture, making it unsuitable for multi-scale detection in complex scenarios. Gu et al. proposed a new multi-task learning framework for multi-person pose estimation, incorporating body orientation and mutual occlusion information [13]. However, this method uses Mask R-CNN as the baseline model, which inevitably increases inference time compared to single-stage algorithms, making it unsuitable for detection tasks with high real-time requirements. To address the issue of small target human detection, Yuan et al. proposed a clustering-based solution to address the low detection efficiency caused by the sparse and uneven distribution of small targets in human body detection [14]. Roy and Bhaduri enhanced the model's ability to detect small objects and multi-scale targets by introducing additional detection heads and a Swin Transformer encoder [15]. Roy et al. improved the feature extraction capability of the network model by including DenseNet transition blocks before the residual blocks of the original CSPDarknet53 and adding new residual blocks between the backbone and the neck. Subsequently, they used an improved PANet to retain fine-grained local information [16]. Li et al. proposed PF_YOLOv4 [17], which introduces a soft threshold function to handle noise such as light in images, enabling more accurate recognition of small target pedestrians. To address extreme viewpoint poses, Linzhi Huang et al. proposed a new 2D HPE (Human Pose Estimation) dataset based on the WEPDTOF dataset, collected using overhead fisheye cameras indoors. However, due to fisheye lens distortion, most people captured in the dataset are from an overhead view. Jingrui Yu et al. used a game engine to generate a synthetic dataset, but this method often suffers from limited pose variability, typically depicting a narrow range of walking or everyday activities.

The computational complexity of top-down algorithms demonstrates a linear growth pattern in relation to the number of individuals depicted in the image [18]. While, in bottom-up algorithms, the use of heatmaps is common for keypoint detection. Even with sophisticated post-processing techniques, heatmaps can still be challenging to make sufficiently clear. It can be difficult to distinguish between two spatially close and similar keypoints originating from different individuals. YOLO-Pose is a new heatmap-free pose estimation method similar to top-down algorithms. This method does not require the complex post-processing of bottom-up algorithms, nor does it need multiple forward passes like top-down algorithms. The complexity of the YOLO-Pose algorithm is independent of the number of people in a single image, and all keypoints are directly located in a single unidirectional inference. It combines the advantages of both types of algorithms: simple post-processing and constant running time. This study is the first to unify and address the issues of keypoint occlusion, small targets, and extreme viewpoint poses. It proposes a keypoint detection algorithm for workers in power operation environments based on YOLO-Pose (YOLOv5s6-Pose), named PW-YOLO-Pose. This algorithm achieves high precision and good real-time performance. Additionally, a proprietary dataset for pose estimation of power workers in complex electrical scenarios is created. The algorithm proposed in this study, when tested on the

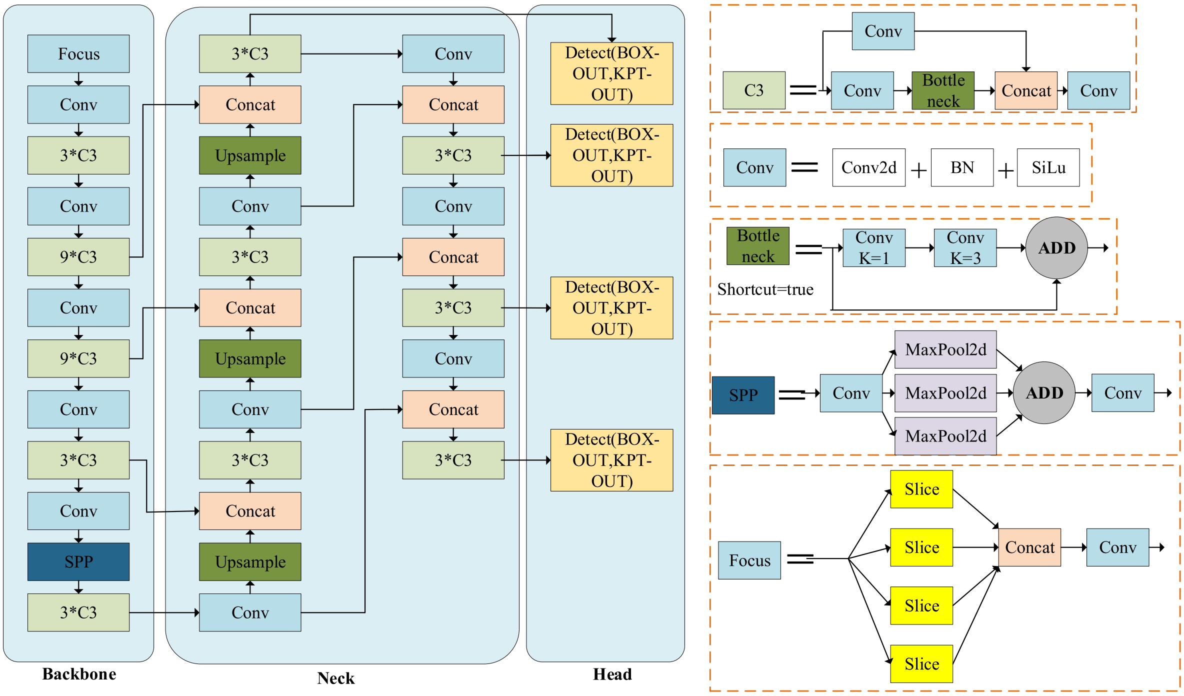


FIGURE 1. Network structure of YOLO-Pose (YOLOv5s6-Pose).

power workers' pose estimation dataset developed in this study, shows an improvement of 10.74% and 7.2% over the current state-of-the-art heatmap-based pose estimation algorithms OpenPose and HRNet, respectively, in terms of $mAP_{0.5}$. In terms of metric $mAP_{0.5:0.95}$(The average detection precision at object keypoint similarity thresholds of 0.5, 0.55, ..., 0.95, respectively), it shows improvements of 6.89% and 2.22%, respectively. Additionally, the number of parameters is reduced by 16M and 15.4M, respectively. The experiments demonstrate that the algorithm proposed in this study has unique advantages in the task of pose estimation for power workers, making it more suitable for real-world applications and improving operational efficiency and safety. Specifically, the contributions of this study are as follows:

- By incorporating the Swin Transformer encoder, this algorithm effectively addresses the challenges of keypionts' misdetection and omissions caused by complex power operation environments with background interference and occlusion.
- This study proposes a BiFPN feature extraction structure with a small target detection layer, enhancing the algorithm's feature fusion and multi-scale detection capabilities. Combined with CA, it improves the precision of keypoints regression for power workers in power operation scenarios, especially in the presence of small targets.
- Modify the bounding box regression loss function to Wise-IoU to make the model focus on anchor boxes of ordinary quality, accelerate the convergence speed of the model, and overall improve the model's regression performance on keypoints.
- A dataset was created for the purpose of power workers pose estimation, and the RePoGen extreme viewpoint poses dataset was introduced as an additional resource to enhance the precision of the algorithm model in predicting keypoints under extreme viewpoint poses. Moreover, this study validates the performance improvement achieved by incorporating the RePoGen extreme viewpoint poses dataset into the dataset this study created.

II. PRINCIPLE OF YOLO-POSE POSE ESTIMATION ALGORITHM

YOLO-Pose is a novel heat-free pose estimation algorithm based on the YOLO detection framework, which enables end-to-end training [18]. This algorithm integrates object detection and pose estimation tasks into a unified processing pipeline. Such integrated processing allows obtaining both the position and pose information of objects in a single forward propagation process, reducing computational costs and processing time.

The YOLO-Pose algorithm, which is based on the YOLOv5 architectural model, employs CSP-darknet53 [16] as the backbone network and PANet for multi-scale feature fusion. The YOLO-Pose algorithm has 4 detection heads, each containing 2 decoupled heads for predicting bounding boxes and keypoints at different scales. The output includes a human detection box and a skeleton graph connecting 17 keypoints. The network architecture is illustrated in Fig. 1.

YOLO-Pose treats pose estimation as a single-person human detection problem, primarily using the COCO dataset. Within each detected individual, there are 17 key points. And each keypoint information containing $\{x, y, conf\}$, where (x, y) denotes the keypoints' coordinates and $conf$ denotes the keypoints' confidence level. For each anchor of YOLO-Pose, there are 57 elements, 51 of which are predicted by the keypoints header, and 6 are predicted by the boxes' header. The overall prediction definition vector for an anchor with n keypoints is shown in Equation (1):

$$P_v = \{C_{x,} C_y, W, H, box_{conf}, class_{conf}, K_x^1, K_y^1, K_{conf}^1, \ldots \\ \ldots\ldots, K_x^n, K_y^n, K_{conf}^n\} \quad (1)$$

where C_x and C_y respectively represent the horizontal and vertical coordinates of the center point of the anchor boxes; W and H represent the width and height of the anchor boxes respectively; b_{conf} and c_{conf} represent the confidence of the anchor boxes and the confidence of the predicted class respectively.

The YOLO-Pose algorithm utilizes CIoU [19] loss for bounding box supervision. For a ground truth bounding box located at position (i, j) and matched with the k-th anchor boxes of scale s, the loss is defined as:

$$\mathcal{L}_{box}(s, i, j, k) = (1 - CIoU(Box_{gt}^{s,i,j,k}, Box_{pred}^{s,i,j,k})). \quad (2)$$

YOLO-Pose employs Object Keypoint Similarity (OKS) to calculate the regression loss $\mathcal{L}_{kpts}$ for human keypoint coordinates. For each anchor box, the entire pose information is stored. If the ground truth bounding box matches the anchor boxes located at position (i, j) and scale s, the algorithm predicts OKS for each keypoint relative to the anchor boxes center, then sums them to obtain the final OKS loss or keypoints' IOU loss as follows:

$$\mathcal{L}_{kpts}(s, i, j, k) = 1 - \sum_{n=1}^{N_{kpts}} OKS \\ = 1 - \frac{\sum_{n=1}^{N_{kpts}} \exp\left(\frac{d_n^2}{2s^2k_n^2}\right) \delta(v_n > 0)}{\sum_{n=1}^{N_{kpts}} \delta(v_n > 0)} \quad (3)$$

in this expression, d_n represents the Euclidean distance between the predicted position and the ground truth position of the nth keypoint; K_n represents the specific weight of the nth keypoint; s represents the target scale; v_n represents the visibility flag for each keypoint; δ is the impulse function representing that only the $\mathcal{L}_{OKS}$ values of the visible keypoints in the real annotation are computed. The loss function of confidence in keypoints is defined as follows:

$$\mathcal{L}_{kpts\_conf}(s, i, j, k) = \sum_{n=1}^{Nkpts} BCE(\delta(v_n > 0), p_{kpts}^n) \quad (4)$$

In the expression, BCE represents the binary cross-entropy loss function, which determines whether individual keypoints exist through binary classification; p_{kpts}^n represents the predicted confidence of the nth keypoint.

Finally, the total loss function is defined based on all the individual loss functions as follows:

$$\mathcal{L}_{total} = \sum_{s,i,j,k} (\lambda_{cls}\mathcal{L}_{cls} + \lambda_{box}\mathcal{L}_{box} + \lambda_{kpts}\mathcal{L}_{kpts} \\ + \lambda_{kpts\_conf}\mathcal{L}_{kpts\_conf}) \quad (5)$$

where $\mathcal{L}_{cls}$ represents the classification loss, $\lambda_{cls} = 0.5, \lambda_{cls} = 0.5, \lambda_{kpts} = 0.1$ and $\lambda_{kpts\_conf} = 0.5$ are hyper-parameters chosen to balance losses at different scales.

III. PW-YOLO-POSE ALGORITHM MODEL

To address the detection challenges of keypoints, such as misdetections and omissions caused by backgrounds, occlusions, small targets, and extreme viewpoint poses in complex electrical power operation environments for power workers, this study proposes a 2D pose estimation algorithm model specifically designed for detecting keypoints of power workers in complex electrical operation environments based on YOLOv5s6-Pose. We name it PW-YOLO-Pose.

The algorithm proposed in this study mainly made the following improvements on the existing algorithm: 1. Embedding Swin Transformer encoder in the top C3 layer of the backbone network, enhancing the algorithm's capability to detect occluded key points through its unique SW-MSA (Shifted Window Multi-head Self Attention) mechanism. 2. BiFPN with a small target detection layer is proposed to construct a new feature fusion network structure, strengthen the information exchange between shallow features and deep features, enrich the scale and source of feature fusion for power workers' pose estimation, improve the detection ability of keypoints of small-target power workers, and alleviate the problems of misdetections and omissions. 3. The CA is embedded in each C3 layer connected to the head to enhance the sensitivity of the algorithm to the location information of keypoints at different scales and improve the precision of the keypoint regression. 4. Wise-IoU is introduced to redefine the bounding box regression loss function to reduce the harmful gradient generated by low-quality examples, accelerate the convergence speed of the model, and further improve the algorithm's ability to detect the keypoints of the human body as a whole. PW-YOLO-Pose's network architecture is shown in Fig. 2.

A. SWIN TRANSFORMER ENCODER

The Swin Transformer encoder is the core module and basic computational unit of the Swin Transformer [20], whose structure is illustrated in Fig. 3. A single Swin Transformer encoder consists of two Swin Transformer blocks, with one block composed of W-MSA (window-based Multi-head Self-Attention) and MLP(Multilayer Perceptron), and the other composed of SW-MSA (Shifted Window Multi-head Self-Attention) and MLP, where the MLP [21] is a multi-layer perceptron with an embedded GELU activation function. Layer normalization (LN) is applied before each MSA module and MLP, and residual connections are used between each

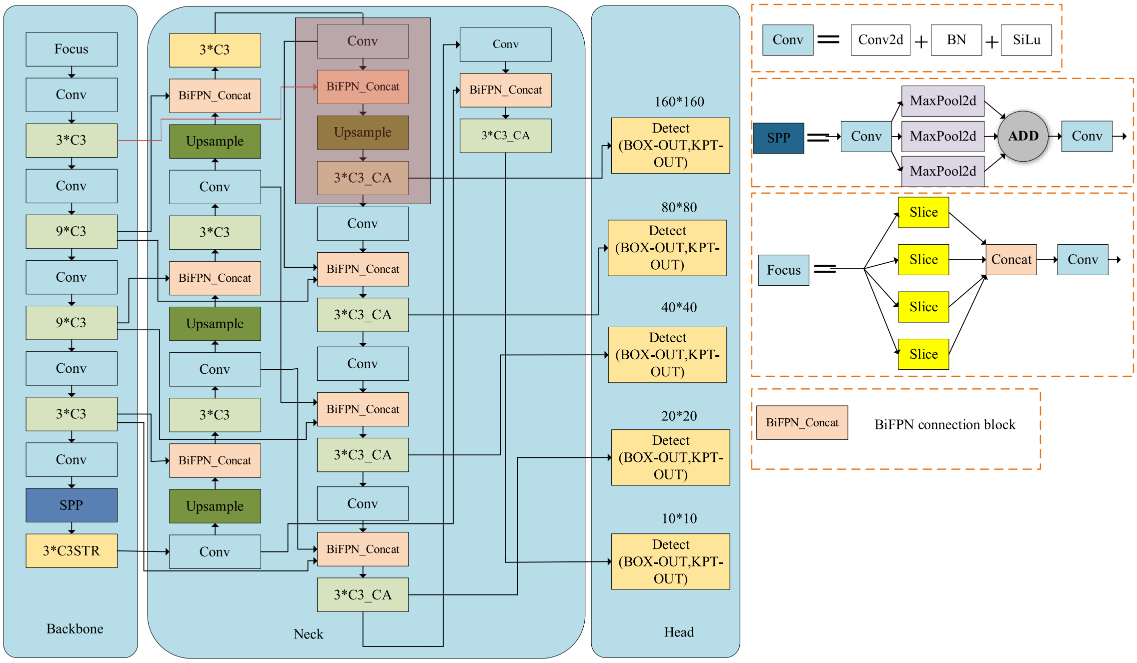


FIGURE 2. Structure of PW-YOLO-Pose network. The PW-YOLO-Pose network structure consists of three parts: the backbone network, the neck, and the head. The red lines represent the new fusion paths. The portion enclosed in the transparent red box represents the newly added small target detection layer in this study. BiFPN_Concat is the module used in this study to construct the BiFPN. The other modules will be introduced in Chapter 3.

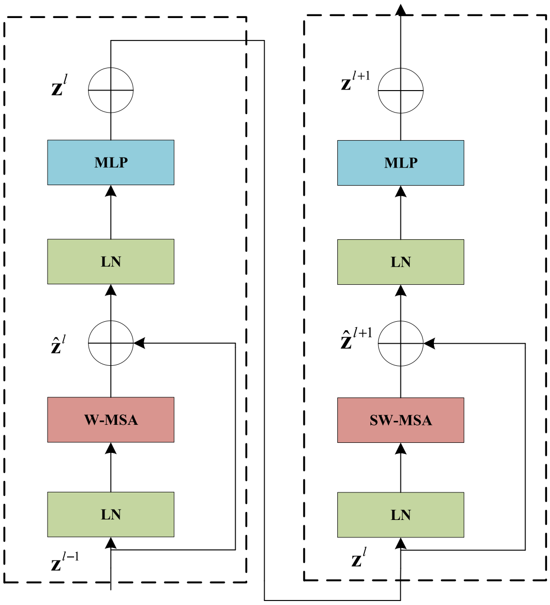


FIGURE 3. Swin transformer encoder structure.

MSA (Multi-head Self-Attention) and MLP. In this study, we solve the global information loss problem caused by constant downsampling in the backbone network by embedding a Swin Transfomer encoder in the top C3 layer of the backbone network, and we name this module C3STR. Its structure is shown in Fig. 4.

Compared to the traditional ViT [22] architecture, the advantages of the Swin Transformer encoder lie in its utilization of both W-MSA, which confines attention computation within each window of the feature map to save computational resources, and SW-MSA, enabling interaction of information between different windows. Additionally, it employs a masking mechanism to isolate invalid information exchange between non-adjacent pixels in the original feature map. Fig. 5 illustrates the principle of SW-MSA, where the self-attention [23] computation is formulated as Equation (6):

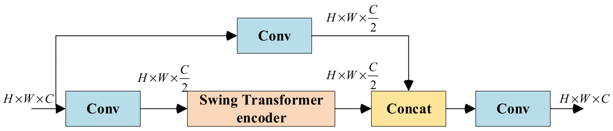

$H\times W\times C$
Conv
$H\times W\times\frac{C}{2}$
Swing Transformer encoder
$H\times W\times\frac{C}{2}$
Conv
$H\times W\times\frac{C}{2}$
Concat
Conv
$H\times W\times C$

FIGURE 4. C3STR structure.

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_k}}+\boldsymbol{B}\right)V \tag{6}$$

In Equation (6), Q, K and V represent the query matrix, key matrix, and value matrix, respectively, obtained by multiplying the input matrix with three trainable parameter matrices; $\sqrt{d_k}$ Represents the square root of the number of channel sequences; $\boldsymbol{QK}^{\mathrm{T}}$ represents the information interaction process between different feature matrices, while $\boldsymbol{QK}^{\mathrm{T}}$ divided by $\sqrt{d_k}$ is to prevent the softmax function's values from becoming too large, which could lead to gradient vanishing in the network; B represents relative positional encoding. Unlike the position encoding in the ViT architecture, Swin Transformer incorporates positional encoding into attention

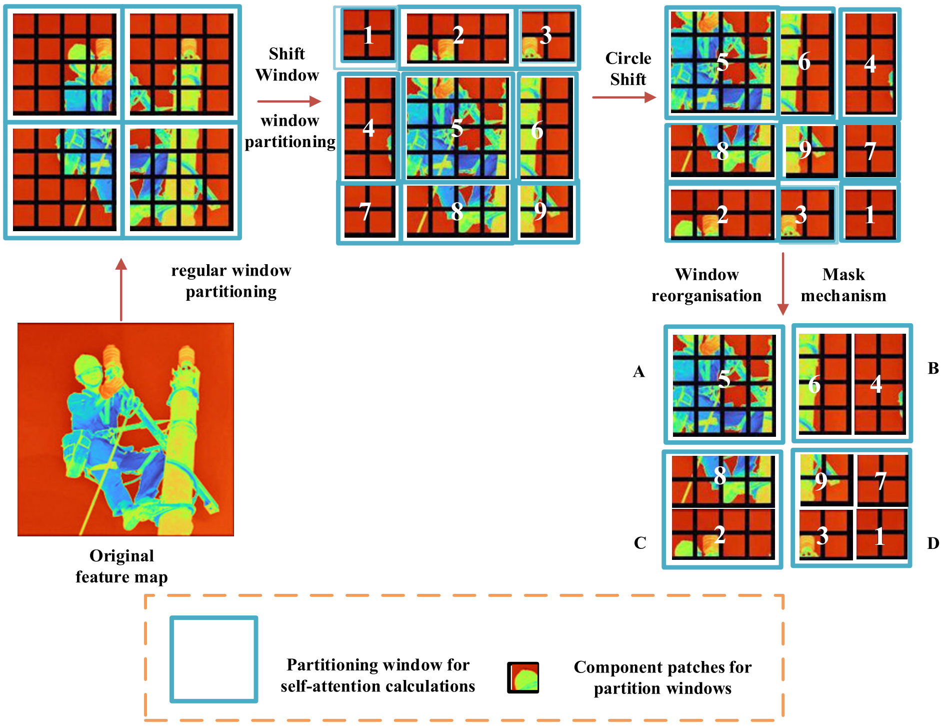

Shift Window
window partitioning
Circle Shift
regular window partitioning
Original feature map
Window reorganisation
Mask mechanism
1
2
3
4
5
6
7
8
9
A
B
C
D
Partitioning window for self-attention calculations
Component patches for partition windows

FIGURE 5. Chematic of SW-MSA self-attention principle. 1-9 represent the 9 Windows that divide the feature map after the shift window self-attention calculation; A, B, C, and D represent four Windows of the same size that have been reassembled after a loop shift and window reorganization operation.

and utilizes relative positional information instead of absolute positional information, enabling it to focus on shifted window information. In summary, embedding the Swin Transformer encoder into the C3 layer, named C3STR, helps the network model in this study to better understand contextual information, effectively capture long-range dependencies, and obtain global information from the feature maps while maintaining relatively low computational costs. Moreover, it can also suppress local interference information to some extent, enhancing the model's sensitivity to occluded keypoint features in complex electrical work environments.

B. BIFPN WITH SMALL TARGET DETECTION LAYER

The YOLO-Pose network model, based on the YOLOv5 architecture, utilizes PANet for feature fusion from the backbone network. The PANet structure is depicted in Fig. 6(b). Deeper feature maps contain richer semantic information suitable for object classification, while shallower feature maps contain better positional information suitable for object localization. PANet builds upon the Feature Pyramid Network (FPN) by establishing a bottom-up pathway [24]. The structure of FPN [25] is shown in Fig. 6(a). This enables the feature maps predicted by PANet to possess both strong semantic and positional information simultaneously. While using PANet relative to FPN has resulted in a significant improvement in detection precision, it also comes with a corresponding increase in computational overhead.

In previous feature fusion methods, the equal treatment of feature information for different scales fails to adequately consider the varying importance of input features, which in reality exhibit unequal contributions to the output features. This suggests that features at specific scales may demonstrate heightened significance and exert a more pronounced influence on the ultimate outcome. In pursuit of this research objective, this study utilizes a novel neck feature fusion network known as BiFPN (a weighted bi-directional feature pyramid network).

This innovative approach incorporates additional weights to each input, facilitating the selective integration of diverse input features [26]. The structure of BiFPN is depicted in Fig. 6(c). The BiFPN model outperforms PANet in several aspects. First, to streamline the network and reduce parameter complexity, nodes with only one input edge and low contribution are eliminated; Second, under the premise that the number of parameters is slightly increased, the original power workers' profile features extracted from the shallow network are weighted and fused with the profile features extracted from the deep network through the jump connection mechanism, which strengthens the information exchange between the shallow features and the deep features, making the network model focus on key information prediction, avoiding the problem of omissions and misdetection caused by the single source of fusion features in the original network, and improving the prediction precision. BiFPN uses a fast normalized weighted feature fusion method, which is computed as

$$O = \sum_{i} \frac{\omega_i}{\varepsilon + \sum_{j} w_j} .I_i \tag{7}$$

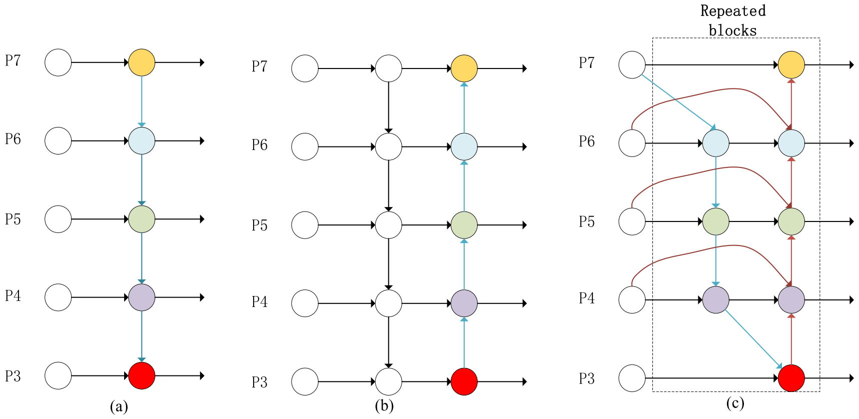


FIGURE 6. Comparison of different feature pyramid network structures.

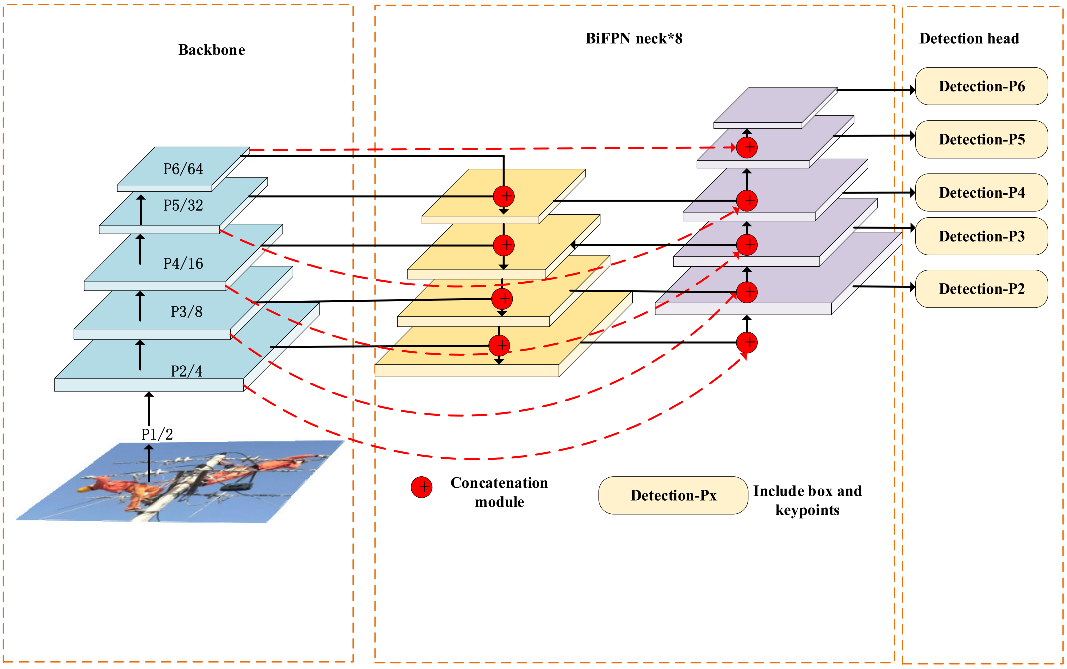


FIGURE 7. Feature Fusion and result prediction network after Introducing BiFPN and Adding Small Target Detection Layer. Detection-P2 is the new detection head for better prediction of bounding boxes and keypoints of small target characters.

where ω_i represents the learnable weights; I_i represents the input features, and $\varepsilon = 0.0001$ is used to avoid instability of the values. The backbone network of YOLO-Pose downsamples the input image to generate feature maps of sizes proportional to 1/2, 1/4, …, 1/64 compared to the original image size. According to principles in computer vision, it is known that deeper feature maps with larger receptive fields are suitable for detecting large objects and contain more global information, while shallower feature maps with smaller receptive fields are suitable for detecting small objects and contain more detailed information [27], [28], [29]. This achieves consistency between the receptive field and target scale. Therefore, feature maps P1, P2, P3, P4, P5, and P6 in Fig. 7 are gradually suitable for detecting targets of increasing sizes. To address the issue of small targets for power workers in electrical operation environments, this study introduces a small target detection layer at the neck, capable of fusing features from the P2 layer of the backbone network, which downsamples the original image twice. Additionally, the detection heads in the network model's head part are increased to five and used for detection, thereby enhancing the algorithm's ability to detect small target individuals and improving the precision of keypoint regression in this situation. The improved network fusion and result prediction structure are illustrated in Fig. 7.

C. C3_CA BLOCK

In this study, the CA (Coordinate Attention) is embedded in each C3 layer connected to the head of the model to improve the sensitivity of the model to the location information of keypoints at different scales, enhance the ability to locate keypoints, and improve the precision of keypoints' regression.

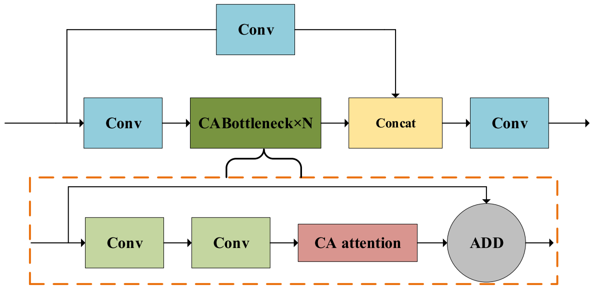


FIGURE 8. C3_CA network structure.

In this study, the C3 module embedded with the attention CA is named C3_CA. The structure of C3_CA is shown in Fig. 8, and the network structure of CA is shown in Fig. 9 below.

Previously, the CBAM [30] attention could also attend to both channel and spatial information. However, CBAM utilizes convolutional operations for spatial attention to leverage positional information, which cannot model long-range dependencies and overlooks the importance of such relationships in visual tasks. In contrast to CBAM, the CA embeds positional information into channel attention, thereby encoding both channel relationships and long-range dependencies. The CA [31] can be specifically divided into coordinate information embedding and coordinate attention generation.

1) COORDINATE INFORMATION EMBEDDING

In order to globally encode spatial information into channel descriptors, it is common to use global pooling. However, this method may lead to the loss of positional information. To address this issue, the CA attention mechanism decomposes global pooling into two separate one-dimensional feature encoding operations. Specifically, for the input feature map, it undergoes global average pooling separately along the width and height directions. In Fig. 9, we can observe that the shape of the input feature map is [C, H, W], where C, H, and W represent the number of channels, height, and width of the feature map, respectively. After performing average pooling along the width direction, the resulting feature map has a shape of [C, H, 1]. At this point, the features are mapped to a higher dimension. The output of the channel c at height h can be represented as:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i). \tag{8}$$

Similarly, for an input feature map with shape [C, H, W], after performing average pooling along the height direction, the resulting feature map has a shape of [C, 1, W]. At this point, the features are mapped to a higher dimension along the width. The output of the channel c at width w can be represented as:

$$z_c^w(w) = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w). \tag{9}$$

These two clever transformations enable the CA to capture long-range dependencies in one spatial direction while preserving precise positional information along another spatial direction. This helps the network in this study to more accurately locate keypoints.

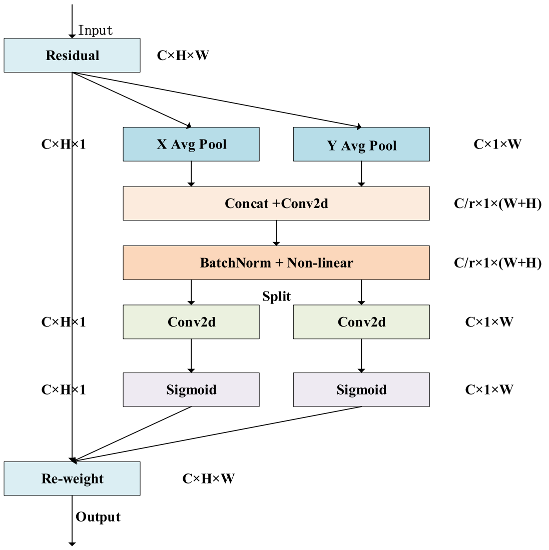


FIGURE 9. CA network structure.

2) COORDINATE ATTENTION GENERATION

After obtaining the feature maps z_c^h and z_c^w in the height and width directions, respectively, generated by Equation (8) and Equation (9). As shown in Fig. 9, first, the Concat operation will be performed on z_c^h and z_c^w. Then, the result after Concatenation will be fed into the 1 × 1 convolution transformation function F1 for dimensionality reduction. Subsequently, normalization and nonlinear activation operations will be carried out to obtain:

$$\mathbf{f} = \delta(F_1([\mathbf{z}^h, \mathbf{z}^w])) \tag{10}$$

The generated feature map $\mathbf{f} \in \mathbb{R}^{\mathrm{C/r\times(H+W)\times 1}}$, where δ represents the non-linear activation function, and r indicates the downsampling factor.

Along the spatial dimension, the feature map **f** is split into $\mathbf{f}^{\mathrm{h}} \in \mathbb{R}^{\mathrm{C/r\times H\times 1}}$ and $\mathbf{f}^{\mathrm{w}} \in \mathbb{R}^{\mathrm{C/r\times 1\times}W}$ using the split operation. Then, each part is upsampled using 1 × 1 convolution and subsequently activated using the sigmoid activation function, resulting in two attention vectors for each direction:

$$\mathbf{g}^h = \sigma(F_h(\mathbf{f}^h)) \tag{11}$$

$$\mathbf{g}^w = \sigma(F_w(\mathbf{f}^w)) \tag{12}$$

The output Y of the generated attention block can be represented as:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{13}$$

D. INTRODUCE WISE-IOU TO IMPROVE CIOU

The well-defined loss function for bounding box regression is crucial for human object detection. Since the YOLO-Pose algorithm resembles a top-down approach, and all keypoints regressions occur within a single forward pass, the precision

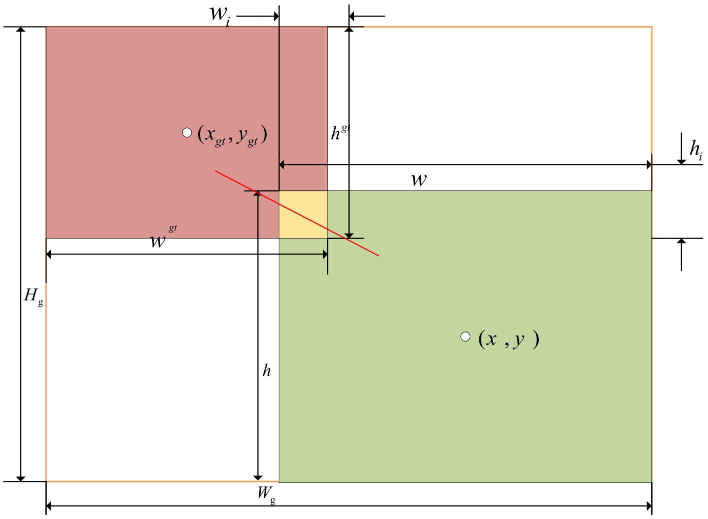


FIGURE 10. Bounding box regression loss function visualization. w_i and h_i represent the width and height of the rectangle intersecting the ground truth bounding boxes and the predicted bounding boxes, respectively.

of human object detection will impact the precision of keypoints' regression.

The YOLO-Pose algorithm employs CIoU as the regression loss for human object bounding boxes. The definition is as follows:

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v \tag{14}$$

CIoU builds upon DIoU [19] by introducing a penalty term to address the issue when the center points of the true bounding box and the predicted bounding box coincide, but their aspect ratios differ while maintaining the same IoU (Intersection over Union) [32]. As shown in Equation (14), where αv represents the newly introduced penalty term. In this equation, v is defined as:

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{15}$$

In Equation (15), $\frac{w^{gt}}{h^{gt}}$ and $\frac{w}{h}$ respectively represent the aspect ratios of the ground truth bounding boxes and the predicted bounding boxes. In Fig. 10: Visualization of bounding box regression loss function, the relationship between the two can be more clearly understood. Here, α is a balancing parameter that assigns priority to IoU, defined as:

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{16}$$

The v in the CIoU formula reflects the difference in aspect ratios rather than the differences in width and height compared to their confidence. Therefore, it can sometimes hinder effective optimization of similarity. Additionally, since the dataset used in this study is manually annotated, it inevitably introduces low-quality examples. CIoU was introduced as a bounding box regression loss function under the assumption that all examples in the training dataset are of high quality. Training with the assumption that all examples in the training dataset are of high quality would emphasize strengthening the fitting capability of the bounding box loss function. Assuming that all training examples in the dataset are of high quality will emphasize strengthening the fitting capability of the bounding box regression loss function. If the fitting capability of the bounding box regression loss function for low-quality samples is blindly strengthened, it will decrease the model's generalization performance. Therefore, this study introduces the bounding box loss function Wise-IoU [33] proposed by Zanjia Tong et al., based on a dynamic non-monotonic focusing mechanism. This allows the algorithm model in this study to focus on anchor boxes of ordinary quality during training, speeding up the model convergence speed, and improving the model's performance in keypoints regression [33].

In response to the unavoidable presence of low-quality examples in the training data, where geometric factors exacerbate the algorithm's penalty on low-quality instances, leading to the inability to optimize high-quality and ordinary samples properly, Wise-IoU v1 with a two-layer distance attention mechanism was first constructed:

$$\begin{aligned} \mathcal{L}_{WIoUv1} &= \mathcal{R}_{WIoU}\mathcal{L}_{IoU} \\ \mathcal{R}_{WIoU} &= \exp(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}) \end{aligned} \tag{17}$$

Next, dynamic non-monotonic FM (focusing mechanism) β is introduced:

$$\beta = \frac{\mathcal{L}^*_{IoU}}{\mathcal{L}_{IoU}} \in [0, +\infty) \tag{18}$$

Finally, by constructing non-monotonic focusing coefficients r and combining them with Equation (17) obtain the Wise-IoUv3 used in this study, which incorporates a dynamic non-monotonic focusing mechanism.

$$\mathcal{L}_{WIoUv3} = r\mathcal{L}_{WIoUv1}, r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{19}$$

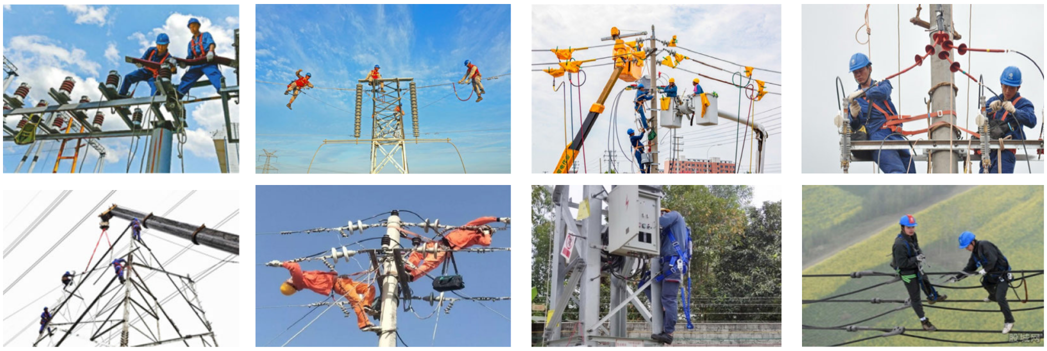


FIGURE 11. Example of the composition of the power workers pose estimation dataset.

In the three equations used to construct Wise-IoUv3 above, $\mathcal{R}_{WIoU} \in [1, e)$ represents the penalty term of Wise-IoU that will significantly strengthen the normal quality $\mathcal{L}_{IoU}$. $\mathcal{L}_{IoU} \in [0, 1]$ represents the IoU loss that will significantly reduce the penalty of Wise-IoU for high quality anchor and focus on the distance between the center point when the anchor overlap well with the predicted bounding boxes. $\overline{\mathcal{L}_{IoU}}$ represents the exponentially moving average with momentum m. FM β represents the outlierness. A smaller outlierness indicates higher quality of anchor boxes, thus assigning them a smaller gradient gain. At the same time, smaller gradient gains are also assigned to predicted boxes with larger outlierness. This approach effectively reduces harmful gradients generated by low-quality training samples, allowing the bounding box regression loss to focus more on anchor boxes of ordinary quality and thereby improving the overall network performance. α and δ are hyperparameters, set to 1.9 and 3 respectively. This study combines the visualization of the bounding boxes regression loss function in Fig. 10. W_g, H_g denote the width and height of the smallest closed box that can wrap both the predicted bounding boxes and the ground truth bounding boxes, respectively. The superscript * in Equation (17) and Equation (18) indicates that these parameters are not involved in back propagation, which eliminates the factors that impede the convergence of the model. x and y represent the horizontal and vertical coordinates of the center point of the predicted bounding box, respectively; x_{gt} and y_{gt} represent the horizontal and vertical coordinates of the center point of the ground truth bounding boxes.

IV. ANALYSIS OF EXPERIMENTS AND RESULTS

A. COLLECTION AND PROCESSING OF EXPERIMENTAL DATASETS

Part of the dataset used in this study is captured from real scenes inside substations, while the remaining portion is collected from publicly available images on the internet. These images cover various scenarios such as substation operations, high-altitude operations on power towers, operations on high-voltage power lines, emergency operations, etc. As shown in Fig. 11, some examples are listed in this study.As evident from the dataset examples shown above, the working environment in power operations is highly complex, prone to occlusion, small targets, and uncommon extreme viewpoints. However, most official COCO datasets for pose estimation consist primarily of images with normal viewpoints, fewer small targets. This is a significant reason why this dataset was created for this study. To enhance the robustness and generalization capability of the model, this study utilized common data augmentation techniques such as random cropping, random flipping, HSV color space transformation, Mosaic, and Mixup, provided by the algorithm. Additionally, to address extreme viewpoint angle poses, this study introduced a new synthetic data generation method called RePoGen (Rare Poses Generator), proposed by Miroslav Purkrabek et al. The RePoGen dataset, generated using this method, serves as a supplement to the dataset used in this study.

The authors mentioned in the original paper that incorporating this dataset as a supplement to the COCO-keypoint dataset has improved the performance of some algorithm models. Through subsequent analysis of experimental results, it was found that using this computationally synthesized virtual dataset with extreme viewpoint pose data has improved the detection precision of the models in this study. Examples of the RePoGen [34] dataset are shown in Fig. 12. This study dataset is divided into training and test sets in an 8:2 ratio. The composition of this study's dataset is shown in Table 1. The images in this study mostly have occlusions, so they have not been classified.

In this study, the dataset was annotated using the Microsoft official COCO Annotator for key points and bounding boxes. The annotation format follows the COCO format, including annotations for 17 keypoints and 1 bounding box per person.

The label numbers corresponding to the keypoints are shown in Fig. 13, where 0 represents the nose, 1 represents the left eye; 2 represents the right eye; 3 represents the left ear; 4 represents the right ear, 5 represents the left shoulder, 6 represents the right shoulder; 7 represents the left elbow; 8 represents the right elbow; 9 represents the left wrist; 10 represents the right wrist; 11 represents the left hip; 12 represents the right hip; 13 represents the left knee; 14 represents the right knee; 15 represents the left ankle; 16 represents the right ankle. During the labeling process, coco-annotator provides

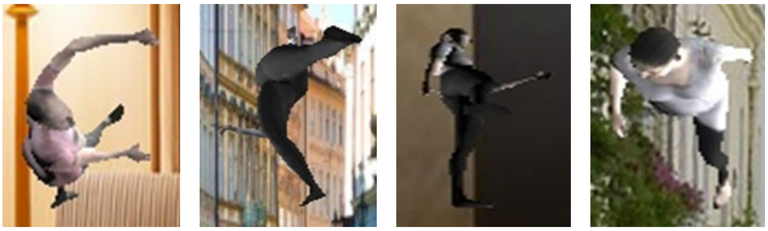
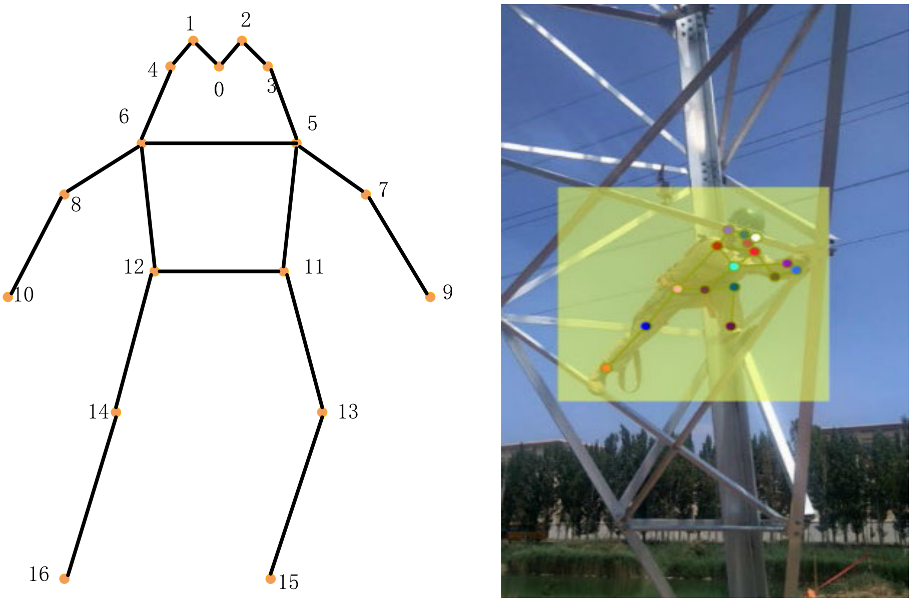


FIGURE 12. Example of RePoGen synthesized extreme viewpoint poses dataset.

TABLE 1. Composition of datasets.

Type of date	Number of data
Small target	1360
Extreme viewpoint	1366
RePoGen	772
Total	3498



(a)

(b)

FIGURE 13. COCO human body keypoints skeletal map and labeling example: (a) skeletal map. (b) examples of labeling.

three keypoint visual states i.e. LABELED VISIBLE, which is the presence of keypoints as stated in this study. It is represented in the algorithm as a value of 2. The second state is LABELED NOT_VISIBLE, which is described in this study as the keypoint is occluded and is represented by a value of 1 in the algorithm. The third state is NOT LABELED, which means that the keypoint is not visible outside the field of view, it is not labeled in this study and is represented by a value of 0 in the algorithm. Fig. 13 shows a human skeleton map in COCO format consisting of 17 keypoints with an example of the labeling in this study.

B. SETTING OF EXPERIMENT ENVIRONMENT AND PARAMETERS

The PW-YOLO-Pose: a 2D pose estimation method for power workers in the power operation environment proposed in this study was tested on Windows 10 with a 64-bit operating system, programming software using PyCharm Community Edition, deep learning framework PyTorch, and GPU acceleration using CUDA and cuDNN. The comprehensive parameters of the test platform are shown in Table 2.

During training, this study set the number of epochs to 300, batch size to 16, initial weight decay coefficient to 0.0005,

TABLE 2. Configuration of the experiments environment.

Name	Configuration
CPU	Intel(R)Xeon(R)CPUE-52695v4@2.10GHz，
GPU	NVIDIA TITAN Xp(4)
Memory	256G
PyTorch	Torch1.12.1+cu113.
CUDA	11.4
cuDNN	8.3

initial learning rate to 0.001, and final learning rate to 0.02. The optimizer used was stochastic gradient descent (SGD) with a momentum of 0.937. This study employed a warm-up strategy during training to stabilize the model. The warm-up initial bias learning rate was set to 0.01, and the number of epochs for warm-up was set to 3. During training, input images were uniformly resized to 640 × 640 dimensions. For transfer learning, this study utilized weights pretrained on the COCO 2017 dataset.

C. EVALUATION METRICS

To evaluate the similarity between the ground truth keypoints and the predicted keypoints, this study adopts the official evaluation standard OKS (Object Keypoint Similarity) from the COCO human keypoints detection dataset:

$$L_{oks} = \frac{\sum_i \left[\exp\left(-\frac{d_i^2}{2s^2 k_i^2} \right) \delta(v_i > 0) \right]}{\sum_i \delta(v_i > 0)} \tag{20}$$

In Equation (20), i represents the keypoints number; d_i represents the Euclidean distance between the predicted location and the true location of the ith keypoint; K_i represents the keypoints' specific weight; s represents the target scale; δ is the impulse function, which indicates that only the value of L_{oks} is calculated for the visible keypointss in the true annotation; and v_i denotes the visibility flag for the ith keypoints. The three visual states represented by v_i are illustrated in this study in this chapter A

This study uses Precision and Recall as evaluation metrics for human target detection. They are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

where TP represents the number of true positive samples correctly classified; FP represents the number of false positive samples incorrectly classified; FN represents the number of false negative samples incorrectly classified. In this study, $mAP_{0.5}$ and $mAP_{0.5:0.95}$ are used as evaluation metrics for keypoint detection. Here, $mAP_{0.5}$ represents the average detection precision of keypoints when the threshold of L_{oks} is 0.5, and $mAP_{0.5:0.95}$ represents the average detection precision of keypoints when the threshold of L_{oks} is 0.5,0.55,...0.90,0.95, respectively. Model size is evaluated using the parameter count, and the inference speed of the model on a single image is utilized to evaluate the model's inference speed.

D. ANALYSIS OF RESULTS

1) ABLATION EXPERIMENT

To assess the impact of each improvement module and the RePoGen dataset on the overall performance of the model, this study conducted ablation experiments. The design of the ablation experiments is presented in Table 3, and the experimental results are shown in Table 4. From the results in Table 4, it can be observed that the PW-YOLO-Pose models in this study exhibited improvements in various indicators compared to the baseline model (YOLOv5s6-Pose) on the electrical power operation pose estimation datasets created in this study. In terms of human target detection, the precision of the algorithm in this study is 92.61%, up 1.06% from the baseline, and the recall rate is 89.18%, which is a large improvement, up 5.2% from the baseline model. At the same time, it also indicates that through the improvement methods proposed in this paper, the algorithm has indeed improved its recognition rate for power operation personnel targets in complex power operation environments. For the algorithm proposed in this study, due to its specific keypoint regression method, the recognition rate of human targets will directly affect the probability of keypoint omissions. In terms of keypoints detection, $mAP_{0.5}$ and $mAP_{0.5:0.95}$ are 93.35% and 64.75%, respectively, which are 5.22% and 1.53% higher than the baseline model. The detection time of a single image is 21.3 ms, which basically meets the requirements of real-time detection.

From the results of improvement 2 in Table 4, it is evident that introducing the Swin Transformer encoder in this study led to a slight increase in parameters compared to the baseline model, but the average detection precision of keypoints improved significantly. The reason behind this improvement lies in the unique characteristics of the Swin Transformer encoder, such as the shift window multi-head self-attention computation and relative position encoding, which help the algorithm model capture long-distance dependencies better, enhance the model's global modeling capability, and

TABLE 3. Design of the ablation experiment. STDL stands for small target detection layer. √ means that the improved methods is used, - means that it is not used.

Method	C3STR	C3_CA	BiFPN	STDL	Wise-IoU	RePoGen
1	-	-	-	-	-	-
2	✓	-	-	-	-	-
3	✓	✓	-	-	-	-
4	✓	✓	✓	-	-	-
5	✓	✓	✓	✓	-	-
6	✓	✓	✓	✓	✓	-
7	✓	✓	✓	✓	✓	✓

TABLE 4. Experimental results of ablation.

Method	Precision/%	Recall/%	$mAP_{0.5}$ / %	$mAP_{0.5:0.95}$ / %	Params/M	Time/ms	GFLOPs
1	91.55	83.98	88.13	63.22	12.5	16.9	17.3
2	91.31	86.56	91.31	64.53	12.7	19.6	27.3
3	92.61	84.21	91.64	64.60	12.7	20.1	27.3
4	89.85	87.08	92.01	64.64	12.9	20.7	27.6
5	91.94	88.53	92.59	64.70	13.1	21.3	31.1
6	92.82	87.81	92.90	64.72	13.1	21.3	31.1
7	92.61	89.18	93.35	64.75	13.1	21.3	31.1

improve the detection ability of occluded keypoints. Comparing the results of improvement 3 with improvement 2, it can be observed that the CA attention mechanism enhances the sensitivity of the algorithm model to keypoint positional information, thereby improving the precision of keypoint regression. Comparison between improvement 4 and improvement 3 demonstrates that BiFPN can better integrate multi-scale information. Moreover, comparing improvement 5 with improvement 4 indicates that the proposed small object detection layer in this study enhances the detection ability of small target power workers and their keypoints by the algorithm model. Contrasting improvement 6 with improvement 5, and referring to Fig. 14, it becomes evident that Wise-IoU can accelerate the model convergence speed and improve the precision of keypoint regression. Finally, by comparing improvement 7 with improvement 6, it can be inferred that the algorithm model of this study has enhanced its ability to detect extreme viewpoint human targets and regress their keypoints by learning the RePoGen extreme viewpoint poses dataset, thereby improving the overall precision of keypoints regression. In summary, combined with Table 4, it can be concluded that although the algorithm model of this study has slightly increased in parameters and detection time compared to the original algorithm model, it has achieved a significant improvement in detection precision.

To highlight the adaptability of our algorithm to complex power operation environments, three specific scenarios with significant occlusion, extreme viewpoint poses, and small target individuals were selected for visual comparison. In Fig. 15, it can be observed that the original algorithm model (YOLOv5s6-Pose) performs poorly in situations with significant occlusions because occlusions severely disrupt the correlation between joints, resulting in low precision in some keypoints regressions. In contrast, our algorithm embeds the Swin Transformer encoder in the top C3 layer of the

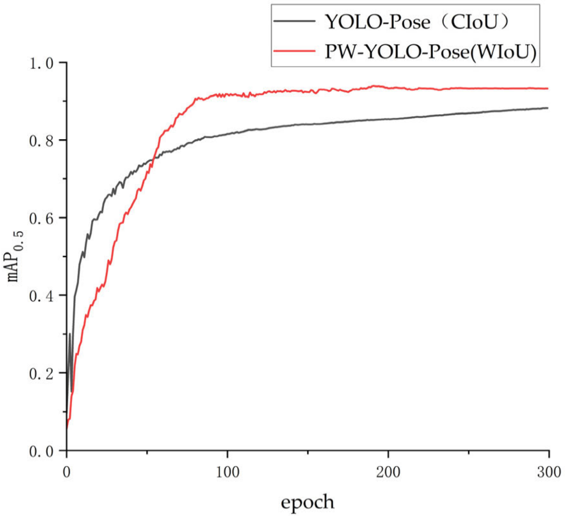


FIGURE 14. The comparison of $mAP_{0.5}$ ($mAP_{0.5}$ represents the average detection precision of keypoints when the threshold of L_{oks} is 0.5).

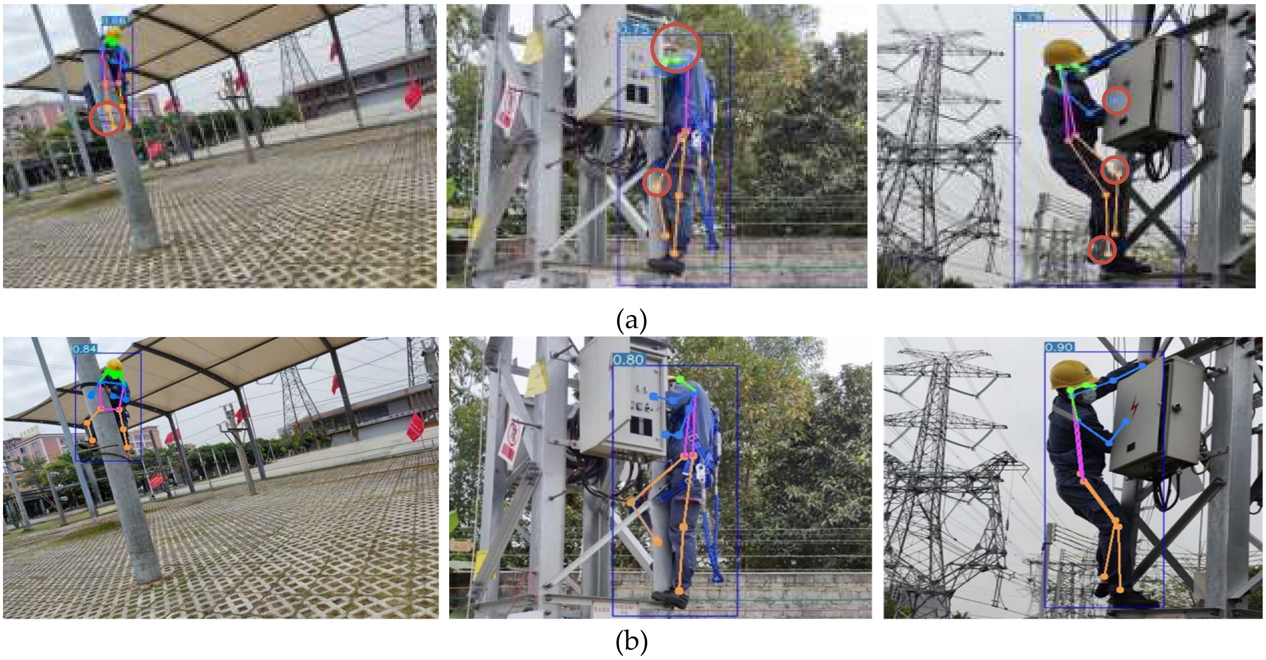


FIGURE 15. Comparison of the effect of pose estimation in the case of occlusion. (a) Original algorithm. (b) Algorithm of this study. The red circles labeled in the figure represent the points that the original algorithm incorrectly detected in this case.

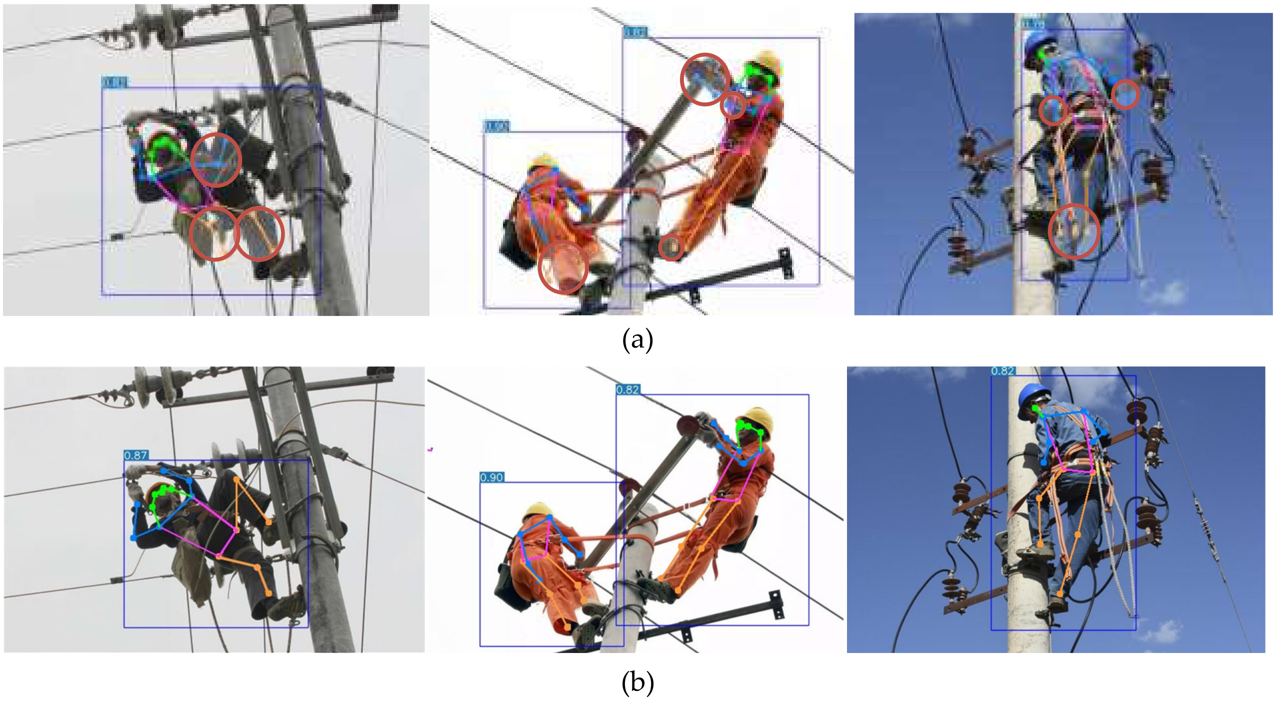
backbone network, enhancing information exchange between adjacent pixels through the SW-MSA, thereby improving the model's perception of spatial positional information. Therefore, compared to the original algorithm, our algorithm can still accurately regress keypoints. In Fig. 16, severe errors in keypoint regression occur in some extreme viewpoint poses situations with the original algorithm. Our algorithm embeds the CA in each C3 layer connected to the detection head and uses the RePoGen virtual extreme viewpoint poses dataset as a supplement to our self-made dataset, thereby improving the precision of keypoints' regression under extreme viewpoint poses conditions. In the case of small targets shown in Fig. 17, the original algorithm's precision of keypoint regression is insufficient, and there are even cases where background information is detected as human targets. In contrast, our algorithm can still maintain good detection performance in this special scenario. The main reason is that we replaced the PANET used for feature extraction in the original algorithm with the more powerful feature extraction capability of BiFPN. Additionally, we introduced a small target detection layer, further enhancing the feature extraction capability of our network in this special scenario.

2) COMPARISON EXPERIMENTS OF DIFFERENT POSE ESTIMATION ALGORITHMS

In order to further verify the improvement effect of the algorithms in this study, two current mainstream algorithmic models based on generating heatmaps and one algorithmic model based on regression coordinates with a similar structure to the algorithmic model in this study are selected for comparison. They are OpenPose, HRNet-W32, and YOLOXs-Pose, respectively. The experimental results are shown in Table 5. Compared to other algorithmic models, the model in this study has the lowest number of parameters but the highest detection precision for keypoints under the power workers pose estimation dataset produced in this study. In Comparision to OpenPose, the algorithmic model of this study reduces the number of parameters by 16M, $mAP_{0.5}$ and $mAP_{0.5:0.95}$ improves by 10.74% and 6.89% respectively. Compared with HRNet-W32, the model parameters are reduced by 15.4M, $mAP_{0.5}$ and $mAP_{0.5:0.95}$ improved by 7.2% and 2.22%, respectively. Compared with YOLOXs-Pose, the number of model parameters is reduced by 21.3 M, $mAP_{0.5}$ and $mAP_{0.5:0.95}$ improved by 4.12% and 1.27%, respectively.

Fig. 18 illustrates the pose estimation performance of the three compared algorithms in occluded scenarios. In cases where occlusion occurs in the forward view, the two heatmap-based pose estimation methods exhibit slightly better keypoint detection performance than YOLOx-Pose. However, when the human body is in a side-view angle causing occlusion, YOLOx-Pose demonstrates better detection performance.

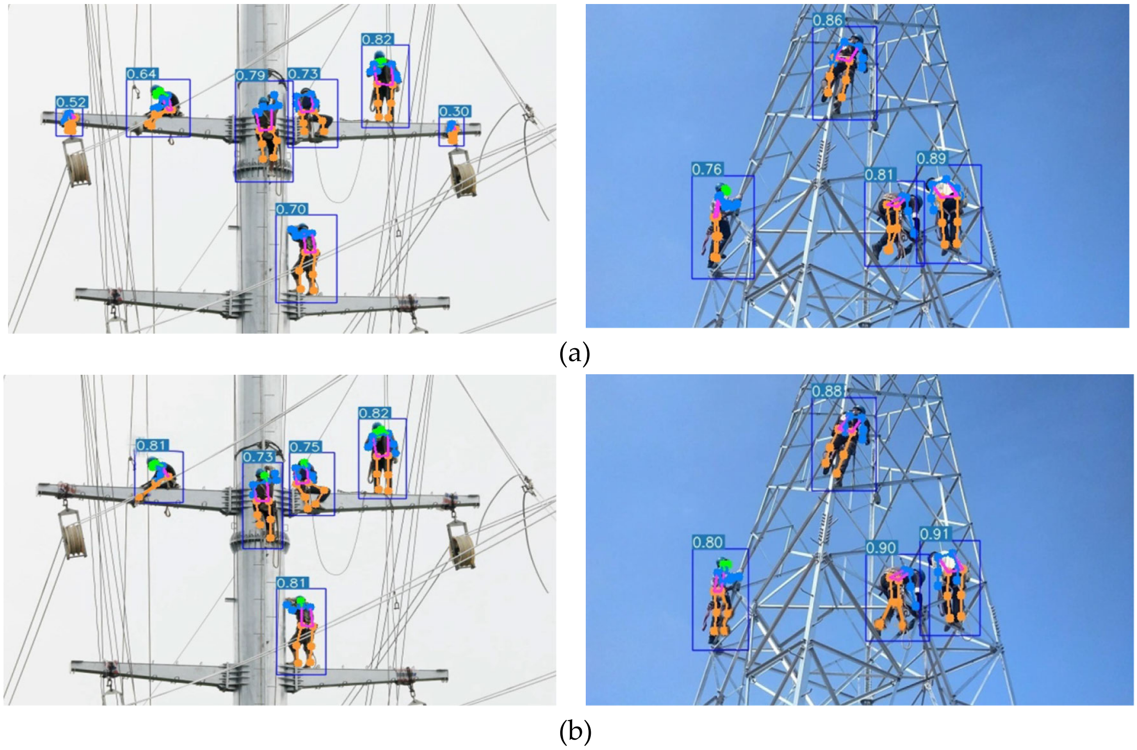
Fig. 19 shows the pose estimation performance of the three compared algorithms in extreme viewpoint poses situations. It is noticeable that the heatmap-based pose estimation algorithms perform poorly in some extreme viewpoint poses scenarios. This is because even with complex post-processing to optimize the heatmap, the heatmap remains unclear in such extreme cases, and is significantly influenced by background factors in the complex power operation environment. Compared to the two heatmap-based pose estimation algorithms, YOLOXs-Pose algorithm exhibits better detection performance in these situations. However, as shown in Fig. 19 (c),



(a)

(b)

FIGURE 16. Comparison of the effect of extreme viewpoint poses estimation. (a) Original algorithm. (b) Algorithm of this study. The red circles labeled in the figure represent the points that the original algorithm incorrectly detected in this case.



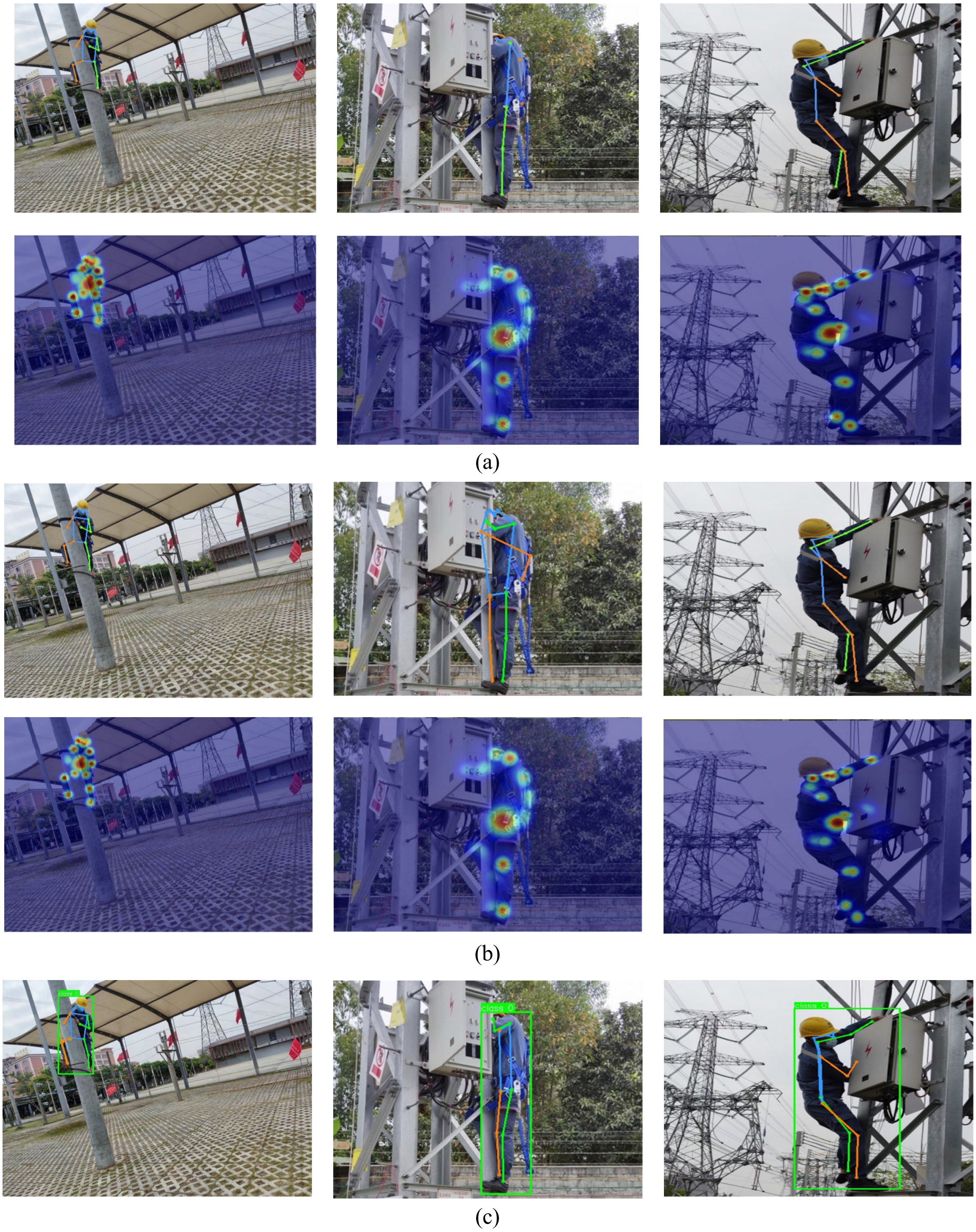
(a)

(b)

FIGURE 17. Comparison of the effect of small target pose estimation. (a) Original algorithm. (b) Algorithm of this study.

this algorithm's detection is also susceptible to background factors in some cases, leading to misdetection in human targets and keypoints.

Fig. 20 presents the pose estimation performance of the three compared algorithms in scenarios with small targets. OpenPose performs the worst, with severe omissions of


(a)

(b)

(c)

FIGURE 18. Comparison of the detection effect of different pose estimation algorithms in the case of occlusion. (a) Detection effects of the OpenPose algorithm and its generated heatmaps. (b) Detection effects of the HRNet-W32 algorithm and its generated heatmaps. (c) Detection effects of YOLOXs-Pose algorithm.

keypoints and human target. This is attributed to OpenPose using the VGG19 first 10 layers as the backbone network for feature extraction, followed by two stages. In this process, feature information flows in one direction, and there is not enough communication between high-layers, mid-layers, and low-layers feature information, resulting in severe loss of low-level information. However, in the image pyramid, the feature maps in the lower layers contain more detailed information, which is crucial for predicting keypoints of small targets. HRNet-W32 maintains high resolution throughout

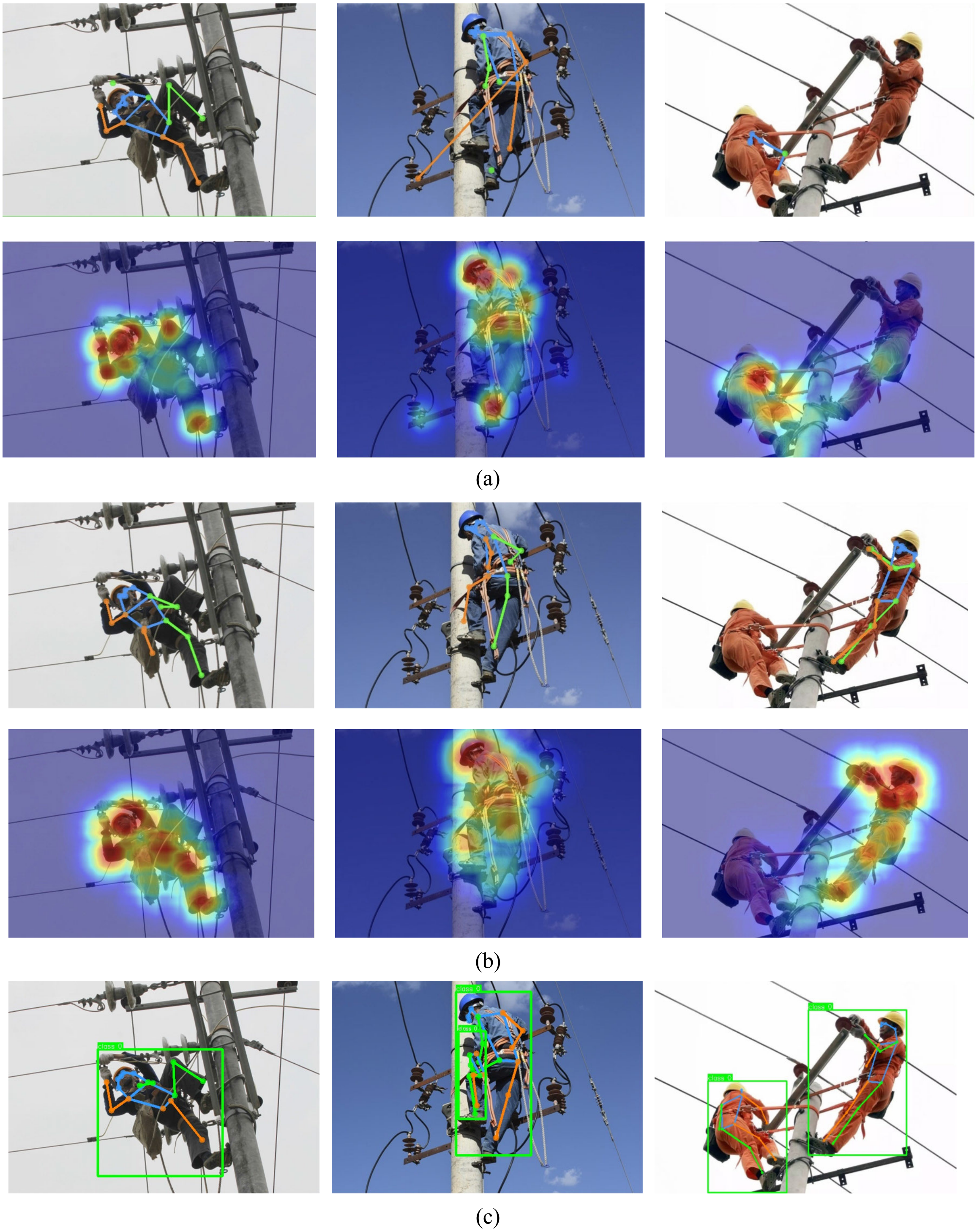
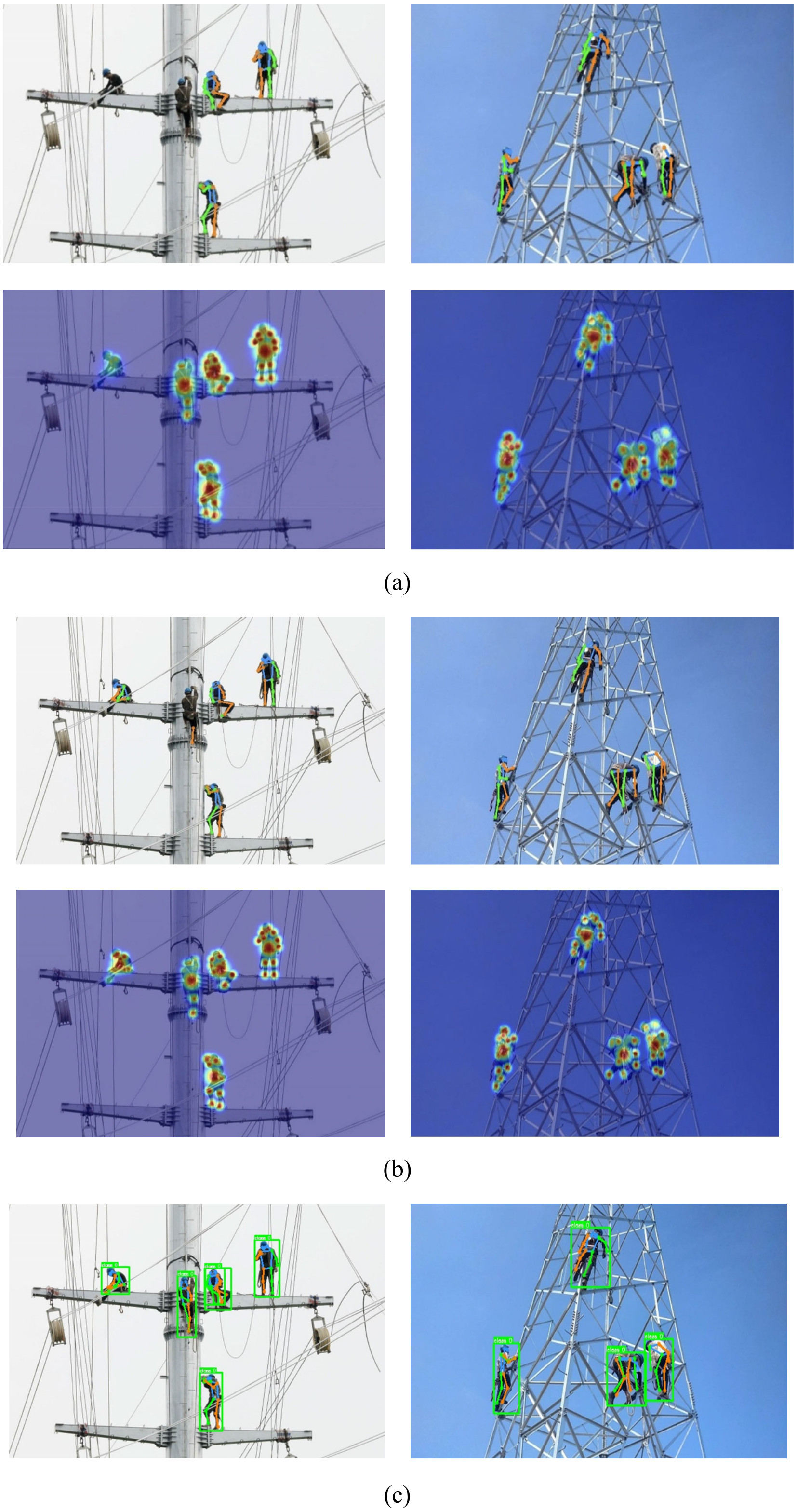


FIGURE 19. Comparison of the detection effect of different pose estimation algorithms in extreme viewpoint poses. (a) Detection effects of the OpenPose algorithm and its generated heatmaps. (b) Detection effects of the HRNet-W32 algorithm and its generated heatmaps. (c) Detection effects of YOLOXs-Pose algorithm.

the feature extraction process by parallelizing multiple resolution branches and continuously exchanging information between different branches, thus having strong semantic information and precise positional information simultaneously. Therefore, in this scenario, HRNet-W32's detection performance is superior to OpenPose. Compared to OpenPose, YOLOXs-Pose, which has a similar structure to our research algorithm, exhibits better performance in small target pose estimation. This is because YOLOXs-Pose uses PANET, allowing more information fusion between different



(a)

(b)

(c)

FIGURE 20. Comparison of the detection effect of different pose estimation algorithms in small target human state. (a) Detection effects of the OpenPose algorithm and its generated heatmaps. (b) Detection effects of the HRNet-W32 algorithm and its generated heatmaps. (c) Detection effects of YOLOXs-Pose algorithm.

feature layers, and is more adaptable to keypoint prediction at different scales. However, there is still a gap in precision between YOLOx-Pose and our research algorithm. The reason lies in our research network structure, which not only

TABLE 5. Comparison of detection performance of different algorithms.

Models	$mAP_{0.5}$ / %	$mAP_{0.5:0.95}$ / %	Params/M	GFLOPs
Baseline	88.13	63.22	12.5	17.3
OpenPose	82.61	57.86	29.1	263.1
HRNet-W32	86.15	62.53	28.5	7.1
YOLOXs-Pose	89.23	63.48	34.4	30.8
PW-YOLO-Pose	93.35	64.75	13.1	31.1

utilizes the stronger feature extraction capability of BiFPN but also adds a small target detection layer to improve the regression precision of keypoints of small target individuals. Overall, our research algorithm demonstrates better detection of keypoint and pose estimation for power workers in power operation environments.

V. CONCLUSION

For complex electrical power operation environments, this study proposes a pose estimation algorithm for electrical workers in electrical power operation environments based on YOLO-Pose, termed PW-YOLO-Pose. It is primarily aimed at addressing challenges of keypoints' omission and misdetection for power workers in complex environments of electrical power operation due to background complexity, occlusion, small targets, and extreme viewpoint. The effectiveness of the algorithm is validated through training and testing on a dataset specifically created for power workers pose estimation in this study.

To enhance the detection rate of occluded keypoints, a Swin Transformer encoder is embedded in the top C3 layer of the backbone network. Additionally, a BiFPN with an added small target detection layer is utilized to improve the detection rate of small target individuals and the regression precision of their keypoints. The incorporation of a CA into the C3 layer connected to the model head enhances the sensitivity of the algorithm to positional information, thereby improving overall keypoints' regression precision. The introduction of the Wise-IoU bounding box loss function prevents harmful gradients caused by low-quality anchor boxes annotated due to human factors during training, accelerating model convergence speed and enhancing overall detection performance.

In this study, the RePoGen dataset, generated using a novel synthetic data generation method proposed by Miroslav Purkrabek et al., is used as a supplement to the dataset created for this study to improve the model's detection performance under extreme viewpoint poses. Experimental results show that compared to the original algorithm model, the $mAP_{0.5}$ (The average detection precision based on object keypoint similarity threshold of 0.5.) and the $mAP_{0.5:0.95}$ (The average detection precision at object keypoint similarity thresholds of 0.5, 0.55, …, 0.95, respectively) have improved by 5.22% and 1.53%, respectively. Compared to OpenPose and HRNet, which use heatmaps for pose estimation, our algorithm shows improvements of 10.74% and 7.2% in $mAP_{0.5}$, respectively. Moreover, in terms of model parameters, our algorithm is more lightweight than these two algorithms. In conclusion, the proposed algorithm model outperforms current mainstream human keypoint detection network models in complex environments of electrical power operation, particularly in extreme viewpoint poses and small target states, offering theoretical support for monitoring the status and identifying behaviors of electrical workers. Of course, our proposed algorithm is not only suitable for environments of electrical power operation but also applicable in other settings, such as construction site scenarios.

However, our proposed algorithm still has certain limitations. In cases with extensive occlusion, the regression of key points remains suboptimal. Nonetheless, this is a crucial issue that needs to be addressed in complex scenarios. To address this issue, we plan to implement the following improvements: (1) Data augmentation, specifically by increasing the proportion of heavily occluded samples; (2) Using multi-task learning methods, such as combining pose estimation with instance segmentation tasks to share feature representations, allowing the model to better understand occlusions and improve its robustness. Additionally, the complexity of our algorithm model has significantly increased compared to the baseline model, which is detrimental to real-time detection. Moving forward, we will continue to pursue a smaller number of model parameters and reduced algorithm complexity. Finally, we aim to deploy the model on edge AI devices such as the Jetson NANO, utilizing TensorRT for model inference acceleration and DeepStream for video acceleration to maintain good real-time detection performance. We also plan to conduct application tests in actual electrical work scenarios.

REFERENCES

[1] R. Hafezi and M. Alipour, "Energy security and sustainable development," in *Affordable and Clean Energy: Encyclopedia of the UN Sustainable Development Goals*. Cham, Switzerland: Springer, 2020.

[2] M. Chen, Z. Lan, Z. Duan, S. Yi, and Q. Su, "HDS-YOLOv5: An improved safety harness hook detection algorithm based on YOLOv5s," *Math. Biosciences Eng.*, vol. 20, no. 8, pp. 15476–15495, 2023.

[3] M. Chen, T. Liu, J. Zhang, X. Xiong, and F. Liu, "Digital twin 3D system for power maintenance vehicles based on UWB and deep learning," *Electronics*, vol. 12, no. 14, p. 3151, Jul. 2023.

[4] Z. Chang, Y. Deng, J. Wu, X. Xiong, M. Chen, H. Wang, and X. Xie, "Safety risk assessment of electric power operation site based on variable precision rough set," *J. Circuits, Syst. Comput.*, vol. 31, no. 14, Sep. 2022, Art. no. 2250254.

[5] Y. Xi, Z. Zhang, and W. Wang, "Low-light image enhancement method for electric power operation sites considering strong light suppression," *Appl. Sci.*, vol. 13, no. 17, p. 9645, Aug. 2023.

[6] B. Jiang, S. Chen, B. Wang, and B. Luo, "MGLNN: Semi-supervised learning via multiple graph cooperative learning neural networks," *Neural Netw.*, vol. 153, pp. 204–214, Sep. 2022.

[7] M. Ben Gamra and M. A. Akhloufi, "A review of deep learning techniques for 2D and 3D human pose estimation," *Image Vis. Comput.*, vol. 114, Oct. 2021, Art. no. 104282.

[8] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2353–2362.

[9] J. Li, "Human pose regression with residual log-likelihood estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11025–11034.

[10] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "HigherhrNet: Scale-aware representation learning for bottom-up human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2020, pp. 5386–5395.

[11] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[12] X. Bai, X. Wei, Z. Wang, and M. Zhang, "CONet: Crowd and occlusion-aware network for occluded human pose estimation," *Neural Netw.*, vol. 172, Apr. 2024, Art. no. 106109.

[13] Y. Gu, H. Zhang, and S. Kamijo, "Multi-person pose estimation using an orientation and occlusion aware deep learning network," *Sensors*, vol. 20, no. 6, p. 1593, Mar. 2020.

[14] G. Yuan, T. Ye, H. Fu, L. Wang, and Z. Wang, "Clustering based detection of small target pedestrians for smart cities," *Sustain. Energy Technol. Assessments*, vol. 52, Aug. 2022, Art. no. 102300.

[15] A. M. Roy and J. Bhaduri, "DenseSPH-YOLOv5: An automated damage detection model based on DenseNet and Swin-transformer prediction head-enabled YOLOv5 with attention mechanism," *Adv. Eng. Informat.*, vol. 56, Apr. 2023, Art. no. 102007.

[16] A. M. Roy, R. Bose, and J. Bhaduri, "A fast accurate fine-grain object detection model based on YOLOv4 deep neural network," *Neural Comput. Appl.*, vol. 34, no. 5, pp. 3895–3921, Mar. 2022.

[17] K. Li, Y. Zhuang, J. Lai, and Y. Zeng, "PFYOLOv4: An improved small object pedestrian detection algorithm," *IEEE Access*, vol. 11, pp. 17197–17206, 2023.

[18] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "YOLO-pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2636–2645.

[19] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12993–13000.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[21] H. Taud and J.-F. Mas, "Multilayer perceptron (MLP)," in *Geomatic Approaches for Modeling Land Change Scenarios*, 2018, pp. 451–455.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[23] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.

[24] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PaNet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9197–9206.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[26] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[27] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[28] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, "Scale-aware trident networks for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6054–6063.

[29] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 385–400.

[30] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[31] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2021, pp. 13713–13722.

[32] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 516–520.

[33] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IoU: Bounding box regression loss with dynamic focusing mechanism," 2023, *arXiv:2301.10051*.

[34] M. Purkrabek and J. Matas, "Improving 2D human pose estimation in rare camera views with synthetic data," 2023, *arXiv:2307.06737*.

QIN SU received the Diploma degree in engineering from Sichuan University of Science and Engineering, in 2018. He is currently pursuing the master's degree in electronic information. He received a Senior Software Engineer Certificate. His research interests include pose estimation and object detection.

JULING ZHANG is currently a Senior Engineer with the Power Internet of Things Key Laboratory of Sichuan Province, specializing in electric power data processing and skilled in all kinds of electric power the IoT related technologies, including but not limited to data acquisition and processing, sensor networks, and the IoT security.

MINGJU CHEN received the Ph.D. degree in image processing from the Southwest University of Science and Technology. He is currently an Associate Professor with Sichuan University of Science and Engineering. His research interests include image processing and intelligent information processing.

HONGMING PENG received the Diploma degree in engineering from Sichuan University of Science and Engineering, in 2018. He is currently pursuing the master's degree in electronic information. He received a Senior Software Engineer Certificate. His research interests include image segmentation and target detection.