ITC 2/55
Information Technology and Control
Vol. 55 / No. 2/ 2026
pp. 571-584
DOI 10.5755/j01.itc.55.2.42664

MS-SACNet: Multi-Scale Self-Attention Convolution Network for Vision-Based Anti-Misoperation System

Received 2025/08/29 | Accepted after revision 2025/12/24

HOW TO CITE: Xu, C., Meng, Y., Yu, R., Wang, W., He, X., Luo, Z., Wang, Z. (2026) MS-SACNet: Multi-Scale Self-Attention Convolution Network for Vision-Based Anti-Misoperation System. *Information Technology and Control*, 55(2), 571-584. https://doi.org/10.5755/j01.itc.55.2.42664

MS-SACNet: Multi-Scale Self-Attention Convolution Network for Vision-Based Anti-Misoperation System

Chennan Xu, Yuwei Meng, Rongdong Yu
Zhejiang Energy Digital Technology Co., Ltd; Hangzhou, China

Wei Wang
Zhejiang Sci-Tech University, School of Mechanical Engineering; phone: +370 37 300 35; fax: +370 37 300 352

Xingwei He, Zhou Luo
Zhejiang Energy Digital Technology Co., Ltd; Hangzhou, China

Zhan Wang*
Zhejiang Energy Digital Technology Co., Ltd & Northwestern Polytechnical University, School of Electronics and Information; Hangzhou, China

Corresponding author: zhanwang@zjenergy.com.cn

Keypoint detection and pose estimation are pivotal in various industrial applications such as vision-based anti-misoperation systems. This system typically incorporates computer vision techniques, including keypoint detection models and object detection, where precision and robustness are critical to the effectiveness and reliability of the system. Advanced deep learning models like YOLO, HRNet and EfficientNet are commonly employed in these applications, but challenges remain due to dynamic scale variations and the requirement for high accuracy and reliability. To address this issue, this paper proposed a Multi-Scale Self-Attention Convolution based keypoint detection network(MS-SACNet), which enhances keypoint localization accuracy by effectively extracting fine-grained features from the input feature map, enabling more precise detection across varied scales. Additionally, we incorporate a generative probabilistic Gaussian Mixture Model (GMM) to enhance the model's detection capability by involving feature distributions during the

training stage. The cross-domain data augmentation techniques further improve the model's performance by exposing it to diverse scenarios and noise, thereby improving its robustness in real-world conditions. Our proposed method has been integrated into the deployed system, and its effectiveness has been validated using the both public datasets such as VOC2007 and VOC2012, and our self-made industrial datasets. For example, by integrating MS-SACNet, our framework outperforms YOLOv5 by 5.2 % mAP^{pose}_{50-95} and 2.5 % mAP^{box}_{50-95}, demonstrating superior effectiveness and reliability.

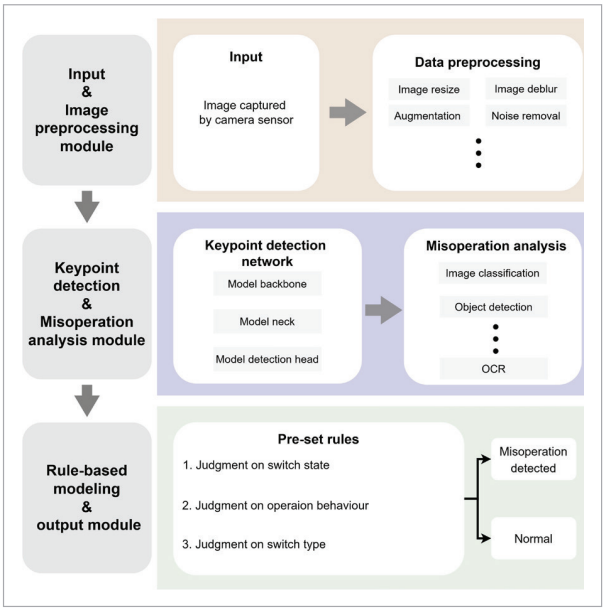
KEYWORDS: Internet of things; One-dimensional dense connected convolutional network; Convolutional neural network; Data balancing processing; Security testing

1. Introduction

In recent years, an increasing number of applications based on computer vision have been applied to traditional enterprises due to the development of artificial intelligence, for instance, the intelligent operation ticket systems in the power plant often integrate computer vision and deep learning techniques to enhance operational efficiency and safety. Ensuring the safe operation of power systems heavily relies on preventing misoperations, the conventional method predominantly involves manual supervision, which brings challenges of low efficiency and risks associated with inconsistent levels of human expertise.

Figure 1

System diagram of the misoperation system adopted with deep learning technologies.



As shown in Figure 1, these systems typically utilize wearable sensors that capture real-time images, serving as inputs for advanced processing. A keypoint detection framework is employed to accurately localize the finger joints of the user in those systems. This process entails analyzing captured images to detect and track finger positions and movements in real-time. Subsequently, various deep learning frameworks, including object detection, image classification, and OCR, are employed based on the localized fingertip positions. The outputs of these frameworks are then fed into a predefined rule-based model to identify instances of misoperation behaviour. By combining these technologies, the intelligent operation ticket system can facilitate safer and more efficient workflows, ensuring that tasks are completed correctly and minimizing the risk of errors. To ensure the overall effectiveness of the system, the performance of the computer vision models plays a crucial role, since keypoint detection and object detection serve as the foundations for accurate localization and target status recognition.

Keypoint detection, a core research domain in computer vision, with significant relevance across various practical applications, such as behavior recognition [17], motion capture [20], human-computer interactions [34] and numerous AI-integrated systems. It follows either top-down [25] or bottom-up approaches [5], and typically focuses on locating keypoints in a target object or person. Significant advancements have been made in this field, particularly with the introduction of deep convolutional neural networks (CNNs) [15] and Transformers [30], and also some notable techniques employed in pose estimation methods. For instance, heat-

map regression [19] techniques predict multiple keypoints for an object with a clear distinction between different keypoints, and the scale-invariant loss metric [21] ensures consistent performance when objects vary in size, etc.

However, top-down approaches suffer from a significant drawback: the accuracy of keypoint detection is highly dependent on the bounding box precision of target, since the keypoints is regressed from the bounding boxes. Consequently, false positives and false negatives adversely affect subsequent keypoint detection. To overcome this issue, we propose MS-SACNet, and our main contributions are summarized as follows:

1 Proposed MS-SACNet enhanced with several preprocessing techniques to improve feature representation.
2 Developed GMM-based feature augmentation method to generate richer and more informative feature distributions.
3 Constructed a cross-domain training dataset by combining open-source data with our own collected samples, improving model robustness.
4 Verified the performance of MS-SACNet and conducted ablation studies of our proposd methods on the open-source PASCAL VOC dataset to prove the effectiveness.

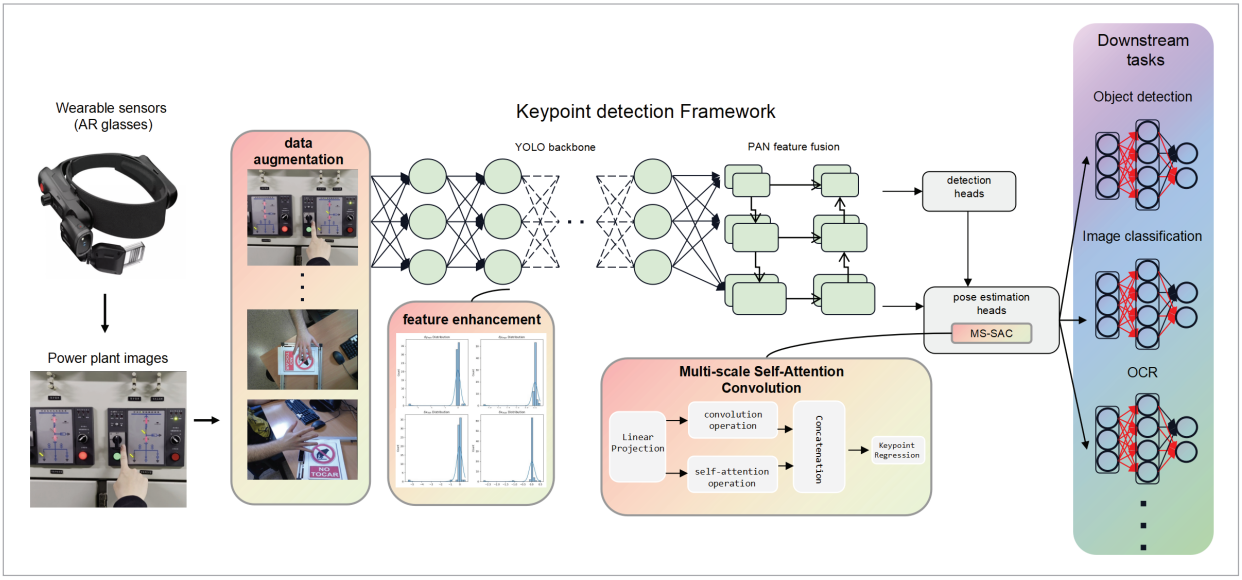
2. Related Works

2.1. Real-time Object Detectors

In recent years, significant efforts have been devoted to developing real-time object detectors, with the YOLO series being one of the most prominent. YOLO networks introduced the classic object detection architecture divided into three parts: backbone, neck, and head. YOLOv3 [28] specifically enhanced feature extraction using Darknet-53, which incorporates more layers and residual connections. YOLOv4 [1] further advanced the architecture by introducing CSPDarknet, SPP, and PAN. YOLOv5 [12] achieved widespread adoption due to its user-friendly design and robust performance, integrating various advancements in model architecture, augmentation techniques, and post-processing methods. The de-

Figure 2

MS-SACNet incorporated with YOLO framework in anti-misoperation system. Wearable AR glasses capture real-time images, which are transmitted to the keypoint detection network to localize the operator's forefinger. The network adopts a YOLO-based backbone and PAN feature-fusion neck, enhanced by cross-domain data and GMM-based feature augmentation. The Multi-Scale Self-Attention Convolution (MS-SAC) is integrated to model both local and global spatial dependencies between keypoints. The detected keypoints support downstream tasks such as object detection, image classification, and OCR for misoperation prevention.



sign of YOLOv6 [16] features an efficient backbone incorporating RepVGG or CSPStackRep blocks, a PAN neck, and an efficient decoupled head with a hybrid-channel strategy. YOLOv7 [31] introduced modifications to the architecture with the Extended Efficient Layer Aggregation Network (E-ELAN). YOLOv8 [13], except for utilizing the novel C3f neural network block, is notable for being an anchor-free model, eliminating predefined anchor boxes to adapt to different object sizes, resulting in higher detection accuracy. The YOLOv9 [32] model includes advancements such as Programmable Gradient Information (PGI) to prevent data loss during gradient updates and the Generalized Efficient Layer Aggregation Network (GELAN) to enhance the architecture.

2.2. Fusion between Convolution and Attention Mechanism

The convolution operation in CNNs uses the kernel to extract the features within an image and has become the most powerful technique in most computer vision tasks [11, 26, 38]. Meanwhile, the attention mechanism has demonstrated outstanding performance in language tasks, such as GPT3 [2], and GPT4 [23]. Given the powerful representation capabilities of the attention mechanism in transformers, extensive research has been conducted to explore its application in vision tasks, and it has led to successful transformer-based variants like the Vision Transformer [37, 6, 18]. The design paradigms of convolution and self-attention are distinct. The convolution operation extracts features over a localized receptive field using predefined kernels, whereas the attention module computes and compares the similarity between pixel pairs and captures more informative features across the entire input feature map. The integration of convolution and self-attention can benefit from both advantages of these two approaches. Some previous work such as SE [10] and CBAM [36] shows that self-attention can be served as an augmentation of the convolution, and ACmix [24] decomposes the operation of convolution and self-attention which provides a more precise integration between them.

2.3. Pose Estimation Framework

The bottom-up paradigm begins by detecting all potential keypoints throughout the entire image, subsequently grouping them into individual sets. Some related works are viewed such as OpenPose [4] and OpenPifpaf [14]. OpenPose introduces Part Affinity Fields (PAFs) to encode association scores between body parts, thereby addressing the matching problem by breaking it down into a series of bipartite matching subproblems. OpenPifpaf, on the other hand, proposes the use of a Part Intensity Field (PIF) for localizing body parts and a Part Association Field (PAF) for their association. Our work follows the top-down paradigm like others [25, 33], which firstly obtains the bounding box for each human body through an object detector and then performs single-person pose estimation within the object bounding boxes. However, this paradigm has several drawbacks that pose challenges for our anti-misoperation system. For instance, significant variations in object size can result in inaccurate keypoint localization, and the performance of keypoint detection is highly dependent on the accuracy of the object detection step. Consequently, integrating top-down methods into real-world applications poses significant challenges.

3. Methodology

3.1. Overview of MS-SACNet Framework

As depicted in Figure 2, AR glasses equipped with camera sensors capture the image in real-time, which transmits the images to the keypoint detection network to identify the forefinger's position. Our keypoint detection network is based on the YOLO architecture, incorporating its traditional backbone structure and PAN neck design. To enhance the model's generalization capability, modifications such as cross-domain data augmentation and GMM-based feature enhancement techniques have been introduced. We incorporate the Multi-scale Self-Attention Convolution architecture, which enables the model to capture both local and global visual representations. We believe that keypoint detection bears a resemblance to sequence data in natural language processing due to the positional relationships between each keypoint. When handling serialized data with transformers, the attention mechanism utilizes Query, Key, and Value matrices. These matrices facilitate the projection of positional and relational information, enabling

the model to capture dependencies and contextual relationships between tokens. It is also essential to understand and encode the spatial positions and relational presentation of each finger joint for accurately capturing the intricate movements and configurations of the fingers. Also, these approaches can be easily transferred to other backbones such as EfficientNet, HRNet and ResNet.

3.2. Cross-domain Data Augmentation and Feature Distribution Modelling Using GMM

While Mosaic data augmentation was initially introduced in YOLOv4 and has been iteratively refined through the latest YOLO version, there are specific reasons why it remains insufficient for our system deployed in the power plant. Firstly, the quantity of raw data collected from the power plant is limited due to confidentiality constraints. Consequently, the insufficient number of data is inadequate for model training. To alleviate this issue, we have customized and sampled the open-source Multi-view Hand Pose (MHP) datasets [8] to augment our dataset.

In our keypoint detection network, the finger keypoints estimator relies on the hand detector following a top-down approach. It is necessary to enhance the capability of the hand detector by incorporating more informative features during the training stage. Since the bounding box distribution on unseen data obtained by an off-the-shelf detector differs from the ground truth, we believe that the difference in this feature distribution can be used as a useful feature representation for the hand detector. In our feature augmentation method, the Real-time Detection Transformer is adopted as the off-the-shelf hand detector, and modelling the bounding box distribution is substituted by modelling the distribution of offsets between the ground truth and the predicted bounding box in both horizontal and vertical directions, which can be calculated by,

$$\delta_{x_{\min}} = \frac{x_{\min}^{\det} - x_{\min}^{gt}}{x_{\max}^{gt} - x_{\min}^{gt}}, \delta_{x_{\max}} = \frac{x_{\max}^{\det} - x_{\max}^{gt}}{x_{\max}^{gt} - x_{\min}^{gt}} \tag{1}$$

$$\delta_{y_{\min}} = \frac{y_{\min}^{\det} - y_{\min}^{gt}}{y_{\max}^{gt} - y_{\min}^{gt}}, \delta_{y_{\max}} = \frac{y_{\max}^{\det} - y_{\max}^{gt}}{y_{\max}^{gt} - y_{\min}^{gt}}. \tag{2}$$

During the training phase, supplementary feature proposals are incorporated by employing a generative, probabilistic Gaussian Mixture Model (GMM) to model the offsets in both the x and y directions. These offsets represent plausible yet unseen variations in bounding-box predictions. These features are subsequently generated through sampling from the GMM. Remember the density function of GMM can be expressed as:

$$p(x) = \sum_{c=1}^{K} \pi_c N(c \mid \mu_c, \Sigma_c), \tag{3}$$

where $N(c|\mu_c, \Sigma_c)$ represents the Gaussian distribution for component c.

After constructing the GMM models by fitting the offset data in both vertical and horizontal directions, additional data points can be generated through ancestral sampling from these models. Figure 3 provides a visualization of the bounding box feature distribution obtained via GMM sampling. The hand detector is trained using a combination of real bounding boxes and GMM-generated augmented proposals, enabling it to better generalize to unseen noise patterns. To prevent overfitting, a hyperparameter is introduced to regulate the proportion of GMM-generated samples within each training batch.

Compared with Mosaic augmentation which primarily increases visual diversity through background and scale changes, our GMM-based method targets structural detection errors that conventional augmentations cannot simulate. By modeling the domain-specific offset distributions produced by off-the-shelf

Figure 3

Distribution of offsets in horizontal and vertical directions sampled by Gaussian Mixture Model, with the origin at (0,0) representing the ground truth.

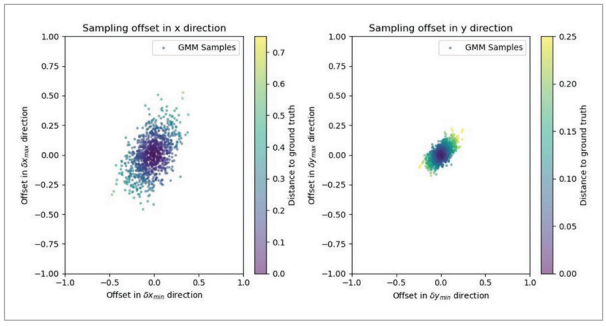
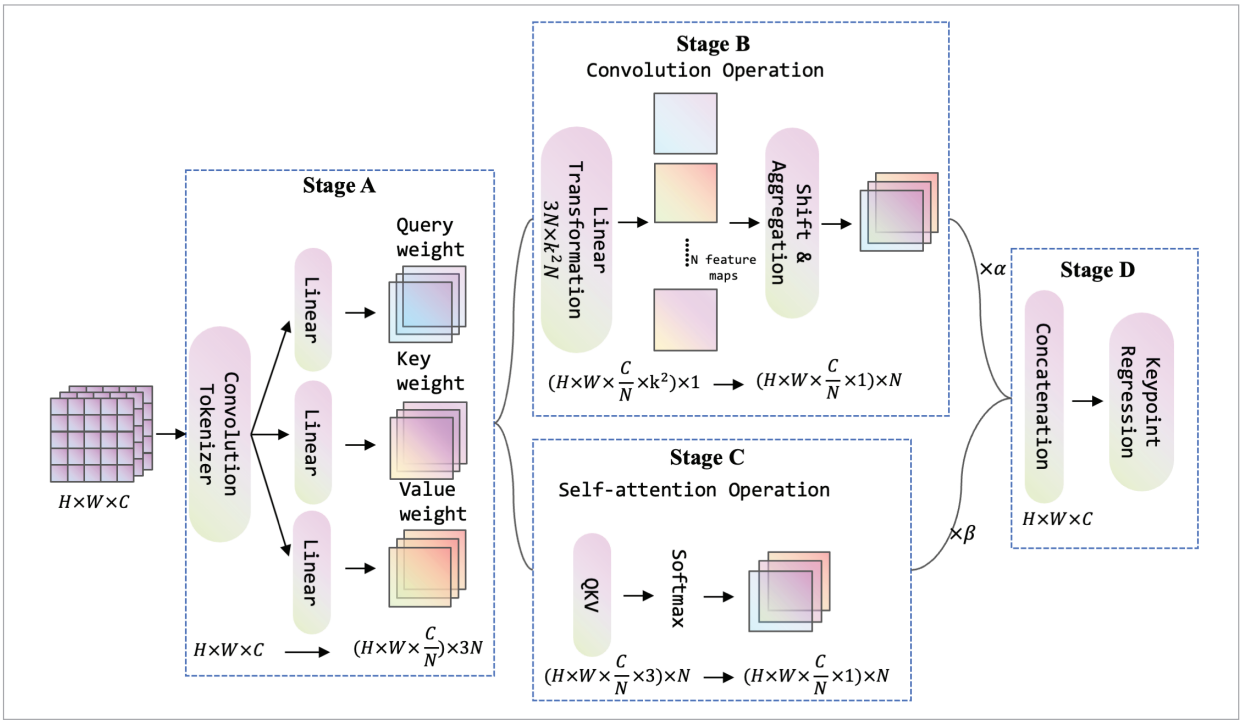


Figure 4

The MS-SAC keypoint detection network operates in multiple stages. In stage A, the input feature map undergoes the process of convolution tokenization to obtain fine-level feature representation, and then the query, key, and value matrices are generated.



detectors, the GMM enables multi-modal probabilistic sampling that reflects realistic noise patterns encountered in industrial environments. This is especially beneficial in top-down pipelines, where hand-detection accuracy directly affects downstream keypoint estimation. Moreover, in confidentiality-restricted settings with limited raw data, the GMM provides an effective way to generate meaningful feature variations without requiring additional images.

3.3. MS-SACNet: Multi-scale Self-attention Convolution

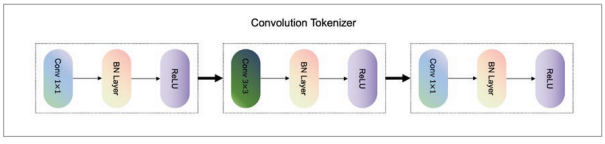
The MS-SACNet is depicted in detail as shown in Figure 4. The convolution tokenizer is utilized in stage A to extract the initial feature map by three basic building blocks, which consist of a non-linear activation layer, batch normalization layer and convolution layer with different kernel sizes to tokenize the input image, as shown in Figure 5. The convolutional layers that employ 1×1 and 3×3 kernels enhance spatial connections by focusing on local regions while also taking into account the broader context of the feature map. After the features are extracted using the convolution tokenizer, the feature map is processed through three linear layers. This step enhances the visual representations and produces the necessary weight matrices for the self-attention calculations in the following stage.

In the stage B and C, the convolution mechanism and self-attention mechanism have been decomposed into separate steps. A standard convolution can be formulated as,

$$g_{ij} = \sum_{p,q} K_{p,q} f_{i+p-\frac{k}{2}, j+q-\frac{k}{2}}, \tag{4}$$

Figure 5

Detailed architecture of the convolution tokenizer.



where $K_{p,q}$ represents the kernel weights at position (p,q), $g_{i,j}$ and $f_{i,j}$ is the pixels at position (i,j) corresponding to the output feature map g and input feature map f, k refers to the kernel size.

After undergoing the linear transformation, we obtained N feature maps and each has the dimension of $(H \times W \times \frac{C}{N} \times k^2) \times 1$. According to the equation (4), we decide to use shift and aggregation operation to decompose the standard convolution procedure, it can be formulated as,

$$Step1: g_{ij} = \sum_{p,q} shift(K_{p,q} f_{i,j}, p - \frac{k}{2}, q - \frac{k}{2}), \quad (5)$$

$$Step2: g_{ij}^{(p,q)} = shift(K_{p,q} f_{i,j}, p - \frac{k}{2}, q - \frac{k}{2}), \quad (6)$$

$$Step3: g_{ij} = \sum_{p,q} g_{ij}^{(p,q)}. \quad (7)$$

Also, for the self-attention operation, the multi-head self-attention with N head can be reformulated as,

$$q_{i,j} = W_q f_{i,j}, k_{i,j} = W_k f_{i,j}, v_{i,j} = W_v f_{i,j}, \quad (8)$$

$$g_{i,j} = \|_{c=1}^{N} \left\{ \sum_{a,b \in N_k(i,j)} SoftMax\left(\frac{(q_{i,j}^c)^T (k_{a,b}^c)^T v_{a,b}^c}{\sqrt{d}} \right) \right\}, \quad (9)$$

where $\|_{c=1}^{N}$ represents the concatenation of N-head self-attention, the $SoftMax\left(\frac{(q_{i,j}^c)^T (k_{a,b}^c)^T v_{a,b}^c}{\sqrt{d}} \right)$ is the equation to calculate the attention.

Eventually, the feature map is obtained after concatenation. The two hyper-parameters α and β are two learnable parameters during training. The feature map resulting from concatenation in stage D integrates informative features from both the convolution and self-attention operations. Keypoints can be predicted using a convolution layer, and then decoded into image coordinates to determine the locations of the desired finger joint keypoints.

4. Experiments

To evaluate the effectiveness of our proposed framework, we performed experiments on both the public PASCAL VOC dataset [7] and a self-built industrial dataset. The PASCAL VOC dataset is a widely used benchmark in computer vision, designed to support research on object detection, classification, seg-

Table 1
Comparison of different networks with our keypoint detection framework using self-made industrial dataset.

Model	mAP_{50}^{box} (%)	mAP_{50-95}^{box} (%)	mAP_{50}^{pose} (%)	mAP_{50-95}^{pose} (%)	Params	FLOPs
YOLOv5	93.5	88.1	89.4	85.0	2.5M	7.5G
YOLOv5†	96.0(+2.5)	90.6(+2.5)	95.3(+5.9)	90.2(+5.2)	2.6M	7.6G
YOLOv8	93.7	88.5	91.4	88.8	3.1M	8.3G
YOLOv8†	97.3(+3.6)	90.8(+2.3)	95.7(+4.3)	91.7(+2.9)	3.2M	8.4G
YOLOv9	94.5	89.1	92.5	89.6	2.0M	7.8G
YOLOv9†	96.8(+2.3)	91.5(+2.4)	96.0(+3.5)	92.8(+3.2)	2.1M	8.0G
$EfficientNet_{b3}$	94.9	90.7	92.8	89.6	12.2M	22.1G
$EfficientNet_{b3}^{\dagger}$	95.3(+0.4)	92.1(+1.4)	93.4(+0.6)	90.2(+0.6)	12.3M	22.2G
$HRNet_{w18}$	94.2	90.4	93.5	90.3	13.8M	64.3G
$HRNet_{w18}^{\dagger}$	95.0(+0.8)	91.9(+1.5)	94.3(+0.8)	90.4(+0.1)	13.9M	64.4G
$ResNet_{w18}$	94.3	90.2	92.0	89.1	13.4M	35.4G
$ResNet_{w18}^{\dagger}$	94.6(+0.3)	90.7(+0.5)	93.4(+1.4)	90.0(+0.9)	13.5M	35.4G

† represents the adoption of our proposed method

mentation, and action recognition. Among its most popular releases, VOC2007 and VOC2012 contain annotated images covering 20 object categories, including animals, vehicles, household items, and people. It has 16551 train images and 4952 validation images in total.

For our self-made dataset, we utilized a combination of publicly available MHP dataset [8] and a custom-built fingertip dataset for our research. The MHP portion comprises 1582 unique images, each featuring a distinct hand instance, selected after eliminating redundant hand poses from the original dataset. The original MHP datasets contain 21 keypoints per hand instance, capturing detailed finger anatomy. In contrast, our customized datasets are simplified, retaining only the 4 keypoints specific to the forefinger as required by the Anti-misoperation system of power plant operation tickets. Our self-made fingertip dataset includes 509 images containing 533 hand instances, which are captured by the optical camera with 1920×1080 pixels. The datasets used for this research have been initially divided into a training set and a validation set with a ratio of 4:1, with 1685 images for training and 406 images for validation. An example of the collected image from the power plant and the labeled image is shown in Figure 6.

4.1. Implementation Detail

The experiments were implemented using PyTorch, with the neural network trained for 500 epochs with early stopping, and a batch size of 16. Input images were resized to 480×270 pixels. The AdamW optimizer was employed, with an initial learning rate of 0.01, adjusted using the cosine scheduling scheme throughout training. All experiments were performed on one NVIDIA 3090 GPU.

Figure 6

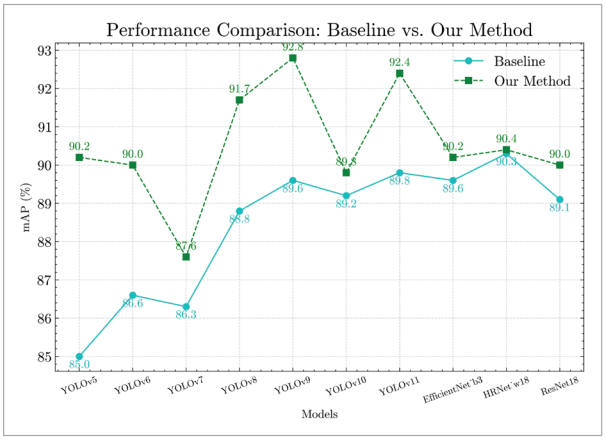
An example of the our self-made industrial datasets.

4.2. Evaluation Metrics

The evaluation metrics for the keypoint detection experiments consist of several major components. The performance of both bounding box and keypoint predictions is assessed using Mean Average Precision(mAP). For the hand bounding box detection, mAP provides a comprehensive representation of the accuracy and error rate of the object detection model in both recognizing and localizing hands and is widely used to evaluate the detection accuracy of advanced object detection models such as the YOLO series. For keypoint detection, mAP is computed by averaging across various Object Keypoint Similarity (OKS) thresholds.

Figure 7

Comparison of the baseline models and the adoption of the MS-SACNet.



The mAP metric is calculated by incorporating Precision, Recall, and Intersection over Union (IoU) for object detection models. IoU quantifies the overlap between the predicted bounding box and the ground truth bounding box. Precision is defined as the ratio of correctly predicted samples to the total number of predicted samples, indicating the likelihood of correct predictions by the model. The recall represents the ratio of correctly predicted samples to all relevant ground truth samples, reflecting the model's ability to accurately identify all relevant instances. For the model size and computational cost evaluation in this experiment, the Params and FLOPs are used. The Params refers to the total number of parameters of the model, which reflects the size of the model. The FLOPs represents the number of float-

ing-point operations, which reflects the amount of computation in the model and can be used to measure the complexity of the model.

4.3. Experimental Results on YOLO Series and other CNN Networks

We conducted comparative experiments across different networks and datasets. We select the smallest sized configuration for all networks in the experiment. The experimental results are presented in Table 1, demonstrate that our proposed MS-SACNet and feature enhancement techniques outperform the traditional design of neural networks.

Using YOLOv9 for implementation, the integration of the MS-SACNet leads to even more substantial gains, enhancing performance 3.5 % mAP^{pose}_{50} and 3.2 % $mAP^{pose}_{50\text{-}95}$ compared with the baseline. The original YOLO pose detection head can be easily replaced by our MS-SACNet to enhance the model capacity. Unlike traditional YOLO networks, where the keypoint regressor in the model head relies on convolutional layers and struggles to extract fine-grained details, our MS-SACNet captures richer and more detailed image features, leading to improved baseline performance.

Following our experiments with YOLO models, primarily based on CSP-Darknet53 backbones with slight modifications across various versions, we aim to further investigate the performance of our methods using alternative backbone architectures. We replaced CSP-Darknet53 with models such as EfficientNet, HRNet, and ResNet. These models have distinct characteristics, for instance, we selected the $EfficientNet_{b3}$, which offers an efficient balance between accuracy and computational cost, $ResNet_{18}$ as a classic model in computer vision tasks, and $HRNet_{w18}$, which preserves high-resolution representations throughout the entire network. The experiment results presented in Table 1 allow us to assess the impact of different backbone characteristics on overall model performance and keypoint detection accuracy.

As shown in Figure 7, the integration of the MS-SACNet has a substantial impact on the models, significantly enhancing their performance. From the experiment results, our methods are capable of achieving higher mAP for bounding box prediction and keypoint detection, which clearly shows the effectiveness of our proposed methods.

4.4. Experimental Results Using PASCAL VOC Dataset

From the experimental results in Table 2, it can be observed that the integration of MS-SACNet consistently enhances the performance of the YOLO detection framework on the PASCAL VOC benchmark. Across all three YOLO variants, the adoption of MS-SACNet yields higher precision, recall, and mAP values compared with their original baselines. For instance, in YOLOv5, the precision improves from 81.3% to 84.8%, recall from 68.4% to 70.6%, and mAP from 76.8% to 80.9%. The improvements are consistent across different YOLO architectures, confirming the generalization ability of MS-SACNet and its adaptability to various backbone designs.

Table 2

Comparison of different YOLO networks. with MS-SACNet on PASCAL VOC dataset.

Model	P (%)	R (%)	mAP^{box}_{50} (%)	$mAP^{box}_{50\text{-}95}$ (%)
YOLOv5	81.3	68.4	76.8	54.9
YOLOv5†	84.8	70.6	80.9	57.1
YOLOv8	81.6	69.8	78.4	57.3
YOLOv8†	83.3	71.0	80.7	59.4
YOLOv9	81.2	68.5	76.7	54.9
YOLOv9†	84.6	70.3	79.8	58.5

† represents the adoption of MS-SACNet.

4.5. Ablation Study of the MS-SACNet

Moreover, the ablation experiments presented in Table 3 provide a comparison between the MS-SACNet and other attention mechanisms, such as the NL, SE, CBAM, GCNet and ECANet. The NL block models long-range dependencies in feature maps by considering the relationships between all pixel positions. The SE module is first introduced in Squeeze-and-Excitation Networks, which utilize global average pooling, MLP and ReLU to adaptively recalibrate feature maps by modelling the interdependencies between channels. The CBAM is an extension of SE, applying attention to both the spatial and channel dimensions. The GCNet introduces a global context block for global context modelling. The ECANet incorporates a local cross-channel interaction strategy to boost model performance.

Figure 8
Detection results of our proposed methods on YOLOv9 and its baseline, compared with the ground truth image from the dataset.

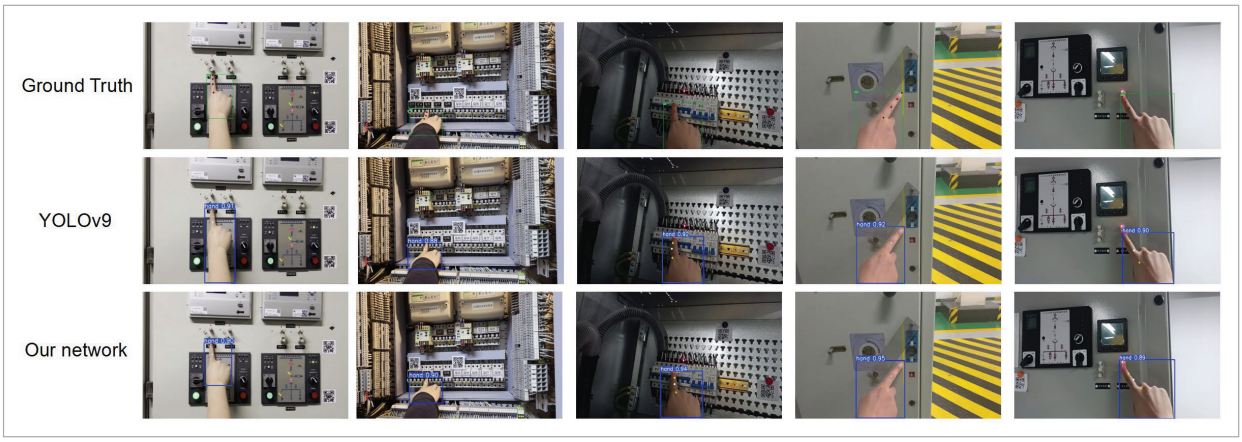


Table 3
Comparison of MS-SACNet with different attention mechanism using YOLOv9 as baseline model.

Model	mAP^{pose}_{10}(%)	mAP^{pose}_{50}(%)	$mAP^{pose}_{50\text{-}95}$(%)
Baseline	94.3	92.5	89.6
NL	94.7(+0.4)	93.0(+0.5)	89.9(+0.3)
SE	96.0(+1.7)	93.2(+0.7)	90.2(+0.6)
CBAM	97.3(+3.0)	94.5(+2.0)	90.9(+1.3)
GCNet	97.9(+3.6)	94.1(+1.6)	91.0(+1.4)
ECANet	97.4(+3.1)	95.3(+2.8)	92.2(+2.6)
MS-SACNet	98.8(+4.5)	96.0(+3.5)	92.8(+3.2)

Compared with NL, SE and CBAM, the integration of the MS-SACNet demonstrates superior performance over the YOLOv9 baseline with 3.5 % mAP^{pose}_{50} and 3.2 % $mAP^{pose}_{50\text{-}95}$ due to its self-attention mechanism that effectively captures the spatial positions and relational representations of each finger joint, making it particularly well-suited for our applications.

4.6. Ablation Study of the GMM-based feature Augmentation Method

To further validate the effectiveness of the proposed GMM-based feature augmentation, we conducted a comprehensive ablation study using three representative YOLO detectors. Table 4 summarizes the performance under three augmentation settings: (1) no augmentation, (2) Mosaic-only augmentation, and (3) the proposed GMM-based augmentation. Across all models, incorporating Mosaic provides noticeable gains by enriching background and scale diversity; however, the improvement remains limited when the detector encounters unseen localization noise.

contrast, the GMM-based augmentation consistently delivers the highest performance, with accuracy gains of 2.5%–4.0% over Mosaic-only training and 5%–7% over the baseline without augmentation. This confirms that explicitly modelling the statistical distribution of detector-induced bounding-box offsets introduces realistic structural variations that appearance-based augmentations cannot provide.

Table 4
Comparison of different YOLO networks with or without Mosaic and GMM-based feature augmentation method.

Model	Use Mosasic	Use GMM-based method	mAP^{box}_{50} (%)	$mAP^{box}_{50\text{-}95}$ (%)
YOLO v5	×	×	90.6	83.3
	√	×	93.5	88.1
	√	√	96.0	90.6
YOLO v8	×	×	89.4	82.6
	√	×	93.7	88.5
	√	√	97.3	90.8
YOLO v9	×	×	90.3	82.4
	√	×	94.5	89.1
	√	√	96.8	91.5

5. Discussion and Limitations

The visualized experimental results in Figure 8 demonstrate that the keypoints predicted by our proposed network exhibit greater precision compared to the original YOLOv9 model, validating the superior performance of our detection system for fingertip detection in industrial environments. However, there remains substantial potential for improvement in our keypoint detection network. One key issue is enhancing the model's robustness in real-world scenarios, where variations in hand poses, background clutter, and lighting conditions can impact detection accuracy. Additionally, image blur during transmission, especially in real-time detection applications, poses another significant challenge. This blur can degrade the quality of the input data, leading to reduced model performance and inaccurate keypoint detection. Addressing these factors is critical for ensuring the system's reliability and accuracy in dynamic, real-world environments, particularly in time-sensitive tasks.

The current dataset primarily focuses on power plant environments, with raw images collected via AR glasses, which limits the model's adaptability to different applications and settings. To improve the framework's performance, it will be necessary to develop or access larger datasets that include a wider range of scenarios and environmental conditions. Expanding the dataset in this way would allow for better generalization of the model across diverse industrial applications, further strengthening its performance and reliability. In addition, the important challenge involves computational efficiency and real-time deployment. While the proposed method performs well, realizing real-time deployment on resource-constrained industrial edge devices presents significant challenges in terms of computation and power efficiency. High-resolution inputs and dense convolutional operations can increase latency, potentially hindering real-time performance in safety-critical scenarios. Future work will therefore explore lightweight model compression techniques such as pruning, quantization, and knowledge distillation, as well as hardware-friendly neural architectures optimized for edge computation. Additionally, integrating the system into industrial control pipelines will require attention to communication delays, reliability, and fault tolerance.

6. Conclusion

This paper presents a novel keypoint detection framework designed for use in vision-based intelligent anti-misoperation ticket systems within power plants. We employed the MS-SAC module as the keypoint detection head in the framework. Additionally, we integrated feature enhancement techniques based on the Gaussian Mixture Model, which significantly improves the robustness and accuracy of keypoint detection. These enhancements contribute to a more reliable system, especially in complex industrial environments.

Our main contribution can be summarized as follows: We collected and annotated a dataset of hand images

from diverse industrial settings and integrated it with publicly available Multi-Hand Pose (MHP) datasets after filtering out redundant hand poses, to create a more diverse and representative dataset for training. Our proposed detection framework enhances the conventional YOLO neural network by incorporating the MS-SAC module, which integrates a convolution tokenizer and self-attention mechanism to capture detailed and pixel-aware features for precise keypoint detection. Furthermore, we introduce new data and feature augmentation strategies to improve model robustness, enabling it to generalize better across different environments and conditions. Extensive experimental results validate the effectiveness and efficiency of our proposed keypoint detection framework, demonstrating superior performance in detecting finger keypoints in industrial settings.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, author-ship, and/or pub- lication of this article.

Data Sharing Agreement

The PASCAL VOC dataset is available online at http://host.robots.ox.ac.uk/pascal/VOC/. The self-made datasets used during the current study are available from the corresponding author on reasonable request.

References

1. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y. M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv Preprint arXiv:2004.10934, 2020. DOI: 10.48550/arXiv.2004.10934.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. Language Models are Few-Shot Learners. arXiv Preprint arXiv:2005.14165, 2020. DOI: 10.48550/arXiv.2005.14165.
3. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. arXiv Preprint arXiv:1904.11492, 2019. https://doi.org/10.1109/ICCVW.2019.00246
4. Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 1302-1310. https://doi.org/10.1109/CVPR.2017.143
5. Chen, X., Yuille, A. Parsing Occluded People by Flexible Compositions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, 3945-3954. https://doi.org/10.1109/CVPR.2015.7299020
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. arXiv Preprint arXiv:2010.11929, 2021. DOI: 10.48550/arXiv.2010.11929.
7. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A. The PASCAL Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 2010, 88(2), 303-338. https://doi.org/10.1007/s11263-009-0275-4
8. Gomez-Donoso, F., Orts-Escolano, S., Cazorla, M. Large-Scale Multiview 3D Hand Pose Dataset. Image and Vision Computing, 2019, 81, 25-33. https://doi.org/10.1016/j.imavis.2018.12.001
9. He, K., Zhang, X., Ren, S., Sun, J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90
10. Hu, J., Shen, L., Sun, G. Squeeze-and-Excitation Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745
11. Huang, K., Yang, J., Wang, J., He, S., Wang, Z., He, H., Zhang, Q., Lu, G. Granular3D: Delving into Multi-Granularity 3D Scene Graph Prediction. Pattern Recognition, 2024, 153, 110562. https://doi.org/10.1016/j.patcog.2024.110562
12. Jocher, G. YOLOv5 by Ultralytics [Computer Software]. GitHub Repository, 2020. Available at: https://github.com/ultralytics/yolov5

13. Jocher, G., Qiu, J., Chaurasia, A. Ultralytics YOLO [Computer Software]. GitHub Repository, 2023. Available at: https://github.com/ultralytics/ultralytics

14. Kreiss, S., Bertoni, L., Alahi, A. PifPaf: Composite Fields for Human Pose Estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 11969-11978. https://doi.org/10.1109/CVPR.2019.01225

15. Krizhevsky, A., Sutskever, I., Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 2012, 25, 1097-1105.

16. Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., Li, Y., Zhang, B., Liang, Y., Zhou, L., Xu, X., Chu, X., Wei, X., Wei, X. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. arXiv Preprint arXiv:2209.02976, 2022. DOI: 10.48550/arXiv.2209.02976.

17. Li, S., Yu, P., Xu, Y., Zhang, J. A Review of Research on Human Behavior Recognition Methods Based on Deep Learning. Proceedings of the International Conference on Robotics and Computer Vision (ICRCV), 2022, 108-112. https://doi.org/10.1109/ICRCV55858.2022.9953244

18. Liu, F., Zheng, Q., Tian, X., Shu, F., Jiang, W., Wang, M., Elhanashi, A., Saponara, S. Rethinking the Multi-Scale Feature Hierarchy in Object Detection Transformer (DETR). Applied Soft Computing, 2025, 175, 113081. https://doi.org/10.1016/j.asoc.2025.113081

19. Luo, Z., Wang, Z., Huang, Y., Wang, L., Tan, T., Zhou, E. Rethinking the Heatmap Regression for Bottom-Up Human Pose Estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 13264-13273. https://doi.org/10.1109/CVPR46437.2021.01306

20. Ma, X., Lv, N. Human Motion Capture Data Segmentation Based on ST-GCN. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, 8260-8264. https://doi.org/10.1109/ICASSP48485.2024.10446553

21. Maji, D., Nagori, S., Mathew, M., Poddar, D. YOLO-Pose: Enhancing YOLO for Multi-Person Pose Estimation Using Object Keypoint Similarity Loss. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, 2636-2645. https://doi.org/10.1109/CVPRW56347.2022.00297

22. Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T. S., Shi, H. Image Super-Resolution with Cross-Scale Non-Local Attention and Exhaustive Self-Exemplars Mining. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 5690-5699. https://doi.org/10.1109/CVPR42600.2020.00573

23. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al. GPT-4 Technical Report. arXiv Preprint arXiv:2303.08774, 2024. DOI: 10.48550/arXiv.2303.08774.

24. Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., Huang, G. On the Integration of Self-Attention and Convolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 805-815. https://doi.org/10.1109/CVPR52688.2022.00089

25. Pishchulin, L., Jain, A., Andriluka, M., Thormählen, T., Schiele, B. Articulated People Detection and Pose Estimation: Reshaping the Future. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, 3178-3185. https://doi.org/10.1109/CVPR.2012.6248052

26. Raman, R., Jiang, W., Bakshi, S. PODE: Panoptic-Aware Object-Wise Depth Estimation for Occlusion Prediction. IEEE Transactions on Consumer Electronics, 2025. https://doi.org/10.1109/TCE.2025.3610083

27. Redmon, J., Farhadi, A. YOLO9000: Better, Faster, Stronger. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 7263-7271. https://doi.org/10.1109/CVPR.2017.690

28. Redmon, J., Farhadi, A. YOLOv3: An Incremental Improvement. arXiv Preprint arXiv:1804.02767, 2018. DOI: 10.48550/arXiv.1804.02767.

29. Tan, M., Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the International Conference on Machine Learning (ICML), 2019, 6105-6114.

30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017, 30, 5998-6008.

31. Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y. M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, 7464-7475. https://doi.org/10.1109/CVPR52729.2023.00721

32. Wang, C.-Y., Yeh, I.-H., Liao, H.-Y. M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. arXiv Preprint arXiv:2402.13616, 2024. https://doi.org/10.1007/978-3-031-72751-1_1

33. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B. Deep High-Resolution Representation Learning for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(10), 3349-3364. https://doi.org/10.1109/TPAMI.2020.2983686

34. Wang, L., Fang, Y. Research on Application of Perceptive Human-Computer Interaction Based on Computer Multimedia. Proceedings of the International Conference on Intelligent Design (ICID), 2022, 281-284. https://doi.org/10.1109/ICID57362.2022.9969748

35. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 11531-11539. https://doi.org/10.1109/CVPR42600.2020.01155

36. Woo, S., Park, J., Lee, J.-Y., Kweon, I. S. CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV), 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

37. Yang, B., Wang, X., Xing, Y., Cheng, C., Jiang, W., Feng, Q. Modality Fusion Vision Transformer for Hyperspectral and LiDAR Data Collaborative Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17, 17052-17065. https://doi.org/10.1109/JSTARS.2024.3415729

38. Yu, H., Wang, J., Wang, Z., Yang, J., Huang, K., Lu, G., Deng, F., Zhou, Y. A Lightweight Network Based on Local-Global Feature Fusion for Real-Time Industrial Invisible Gas Detection with Infrared Thermography. Applied Soft Computing, 2024, 152, 111138. https://doi.org/10.1016/j.asoc.2023.111138