Hindawi
Journal of Electrical and Computer Engineering
Volume 2023, Article ID 9772652, 9 pages
https://doi.org/10.1155/2023/9772652

Research Article

FFA-YOLOv7: Improved YOLOv7 Based on Feature Fusion and Attention Mechanism for Wearing Violation Detection in Substation Construction Safety

Rong Chang,[1] **Bingzhen Zhang**,[1] **Qianxin Zhu**,[1] **Shan Zhao**,[2,3] **Kai Yan**,[2,3] **and Yang Yang**[2,3]

[1]*Yuxi Power Supply Bureau, Yunnan Power Grid Corporation, Yuxi 653100, China*
[2]*School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China*
[3]*The Laboratory of Pattern Recognition and Artificial Intelligence, Yunnan Normal University, Kunming 650500, China*

Correspondence should be addressed to Shan Zhao; 220042@ynnu.edu.cn

Received 2 April 2023; Revised 2 May 2023; Accepted 5 June 2023; Published 12 June 2023

Academic Editor: Yang Li

Ensuring compliance with safety regulations regarding wearing is essential for the safety and security of those working on substation construction sites. However, relying on supervisors to monitor workers in real time on the work site or through remote surveillance videos is both unreasonable and inefficient. A deep learning network approach named FFA-YOLOv7 is presented in this study that utilizes an improved version of YOLOv7 to detect violations of worker wearing in real time during power construction site surveillance. In YOLOv7, the feature pyramid network (FPN) of the neck stage is constructed through continuous upsampling and skip connections for feature fusion, after continuous downsampling of the backbone. However, this process can result in the loss of precise shallow position information. To tackle this issue, we have introduced a novel feature fusion pathway to the FPN architecture, enabling each layer not only to fuse feature maps from the same level during the downsampling course but also to fuse feature maps from shallower levels. This approach combines precise positional information from shallow layers with rich semantic information from deep layers. Additionally, we utilized attention after feature fusion in each layer to optimize the feature map fusion effect and achieve better detection accuracy performance. In order to conduct comparative experiments, we trained six variations of the YOLO model as detectors using a dataset gathered from realistic construction sites. The experimental results indicate that our proposed FFA-YOLOv7 attained a detection precision of 95.92% and a recall rate of 97.13%, demonstrating a high level of accuracy and a low rate of missed detections. These outcomes effectively satisfy the requirements for robust and accurate detection of real-world power construction violations.

1. Introduction

The construction of electric power infrastructure is a crucial component for ensuring the smooth transmission and distribution of energy in substations. The safe and efficient construction of these sites is essential for maintaining reliable power grids. Unfortunately, accidents resulting from non-compliant wearing of work clothes are common occurrences in substation construction, jeopardizing the safety of workers and disrupting the normal operation of the site. Ensuring that workers wear the appropriate attire and comply with safety regulations is therefore of utmost importance. For managers overseeing these sites, identifying and addressing violations of safety protocols are crucial for maintaining a safe and productive work environment.

In the past, the assessment of workers' attire and behavior at substation construction sites was primarily carried out through manual inspections performed by security personnel. Nevertheless, this method proved to be both time-consuming and demanding in terms of labor. Moreover, manual inspections may not be able to cover all workers in real time, especially in large-scale construction sites. With the advancement of video surveillance technology, many power construction sites have installed

monitoring systems to transmit video footage to the substation's monitoring and dispatching center through the network. Security officers on duty can monitor workers' activities in real time and identify violations through surveillance video. Nevertheless, the present monitoring methods still depend on manual inspections, neglecting to fully exploit the capabilities offered by intelligent video surveillance technology. This limits the ability to efficiently and accurately identify violations in various construction scenarios. As such, there is a need to utilize intelligent video surveillance technology to develop a more efficient and accurate method for identifying construction violations.

This paper aims to put forward a novel deep learning approach that can detect wearing violations in substation construction sites more efficiently and accurately compared to conventional methods. The proposed network can identify not only straightforward wearing violations but also more intricate ones, by analyzing the distance between objects. Furthermore, the network is trained in an end-to-end manner using a comprehensive dataset that includes authentic images captured from actual power construction sites and synthetic images generated through data augmentation.

The main contributions of this paper are as follows:

(1) In this paper, an enhanced variant of YOLOv7 is proposed, which introduces a new feature fusion pathway within the FPN. The objective is to effectively integrate accurate position information from shallow layers with rich semantic information from deep layers. Additionally, attention mechanisms are incorporated into these fusion layers to enhance the feature representation after fusion.

(2) A deep learning approach is proposed for the real-time detection of worker attire violations in surveillance videos obtained from substation construction sites. Additionally, a dataset has been curated using a range of data augmentation techniques. The dataset consists of videos captured in authentic power construction sites and encompasses six commonly encountered targets for the detection of attire violations.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related work. In Section 3, we present a detailed description of the proposed network architecture. To assess the effectiveness of our method, we design experiments in Section 4 and discuss the results in Section 5. Finally, we present our conclusions in Section 6.

2. Related Work

Advancements in video surveillance technology and wireless mobile networks have facilitated real-time monitoring of substations. However, the current video surveillance systems used by electric power enterprises have certain limitations, as examined by Jiangtao et al. [1]. To address these shortcomings, the authors introduced key technologies that can be incorporated into a new substation security video surveillance system. However, it is important to acknowledge that there are still limitations in the current substation video monitoring systems. For instance, Yang et al. [2] highlighted that the video sensor equipment layout on the construction site of substations lacks scientific guidance, which results in incomplete three-dimensional monitoring coverage. They then proposed a video surveillance system that can provide full coverage monitoring for substation construction sites. Moreover, Lu et al. [3] proposed an intelligent monitoring solution for power substations, utilizing big data theory and intelligent analysis algorithms. The objective of this solution is to assist monitoring personnel in comprehending alarm signals and reducing the workload of substation personnel. However, existing video monitoring systems for substations are constrained to basic functionalities such as video capture, storage, and playback. They lack effective video data analysis capabilities. Furthermore, the monitoring of substation workers continues to be conducted manually, without fully capitalizing on the potential of intelligent video surveillance technology.

Deep learning technology has witnessed continuous advancements, particularly in convolutional neural networks (CNNs). Among the notable CNN series, You Only Look Once (YOLO) stands out for its exceptional performance in object detection, offering high accuracy, efficiency, and real-time capabilities. The YOLO series was initially introduced in 2015 with the release of YOLOv1 [4]. This pioneering single-stage detection network addressed the slow reasoning speed issue encountered in two-stage detection networks while maintaining commendable detection accuracy. Subsequent versions, including YOLOv2 [5] and YOLOv3 [6], further improved upon the original model. YOLOv3 introduced the Darknet-53 residual module and the feature pyramid network (FPN) architecture, enabling object prediction at multiple scales and facilitating multiscale fusion. YOLOv4 [7] and YOLOv5 have since incorporated various enhancements based on YOLOv3. The recent YOLOv7 introduced in 2022 [8] introduces the innovative extended ELAN architecture, which enhances the network's self-learning ability without disrupting the original gradient path. Furthermore, YOLOv7 adopts a cascade-based model scaling approach, generating models of different scales to accommodate practical tasks and meet the detection requirements.

Previous research in safety construction monitoring has predominantly concentrated on helmet detection. Several studies have proposed helmet detection methods using YOLOv5, a popular deep learning technology [9, 10]. Additionally, other researchers [11–15] have made advancements in helmet detection by refining networks based on YOLOv5. Furthermore, CNNs have been employed to detect safety vests [16–18], safety belts [19–21], and insulators [22–24] worn by workers in surveillance videos of power substations. While these studies have yielded promising results, they have primarily focused on detecting a single object and are unable to simultaneously detect multiple objects. Consequently, these methods are not well-suited for violation detection tasks in complex power construction sites.

3. Method

3.1. Architecture. Our improvement on the network is based on the YOLOv7 base model. Compared with other models of YOLOv7, the base model of YOLOv7 has fewer parameters and demonstrates superior real-time performance. Ensuring high accuracy is crucial for real-time violation detection applications in substation construction. The improved network structure, FFA-YOLOv7, which is based on YOLOv7, is depicted in Figure 1. Our proposed model incorporates two significant enhancements. Firstly, we introduce a novel feature fusion path within the FPN to effectively merge the rich semantic information from the deep layers with the accurate location information from the shallow layers. This enables us to enhance the representation of features for improved performance. Secondly, we add a new attention module after each feature fusion path to extract inter- and intra-relationships in each fusion source for refinement of the fusion feature. More details about the feature fusion path and attention are available in the next two subsections.

3.2. Feature Fusion Path. In the segmentation process, pixel-level labels often lack global information, making it beneficial to consider larger patches to obtain more comprehensive information. In contrast, object detection tasks relying solely on image-level and bounding box-level labels can obscure crucial information. Edge information in feature maps tends to diminish during continuous downsampling and upsampling. The YOLOv7 model employs extended efficient layer aggregation network (ELAN) to enhance network learning by incorporating a gradient path. However, the model contains numerous convolutional layers, leading to a gradual dilution of location information during continuous extraction of semantic information. In subsequent feature pyramid network (FPN), the ELAN module is repeatedly used to extract features, allowing deeper networks to learn and converge effectively. However, this process significantly weakens the edge position information, which is crucial for accurately generating anchor boxes that fit the target size. Considering larger patches in object detection tasks can also impact the calculation of intersection over union (IoU) between predicted anchor boxes and ground truth boxes, potentially reducing the final detection accuracy, relying solely on image-level and bounding box-level labels may result in the omission of critical information, leading to a decrease in precision during IoU evaluation.

To address the issue of location information loss resulting from the extensive use of ELAN, we introduced additional feature fusion paths in the FPN of YOLOv7 to achieve better fusion of semantic and location information. Figure 2 illustrates the newly proposed feature fusion path. In the backbone network, the feature maps at each scale level are first extracted using the ELAN block and then downsampled using the MPConv block. We incorporated feature fusion paths in the subsequent FPN process, allowing each layer of the FPN to receive not only feature maps of the same size from the previous FPN layer and downsampling process but also feature maps from shallower layers that have not undergone the ELAN block of that layer. This feature fusion path enhances the aggregation of the initial feature pyramid and provides the necessary details for coordinate regression, thereby improving the accuracy of the one-stage object detector.

3.3. Attention Module. With the exception of the top layer, the feature maps in the FPN are obtained through a fusion process involving the previous layer and the two adjacent layers in the downsampling process. However, these three sources exhibit distinct representations of semantic levels and spatial locations due to their generation via different skip connections and upsampling pathways. The feature maps from the shallow pathways contain precise location information and fine-grained features, whereas those from the deep pathways exhibit richer semantic information and coarse-grained features. Consequently, a selective mechanism is required to filter out and retain effective feature information representations when fusing the three feature maps.

To improve the selection of feature information among each set of feature maps, a new select mechanism called the Channel Refinement Attention Module (CRAM) is proposed in this paper. CRAM is built upon the Channel Attention Module (CAM) [25]. As depicted in Figure 3, the three source feature maps are initially concatenated, and then a CAM is employed to extract the inter-group channel relationship. Subsequently, another CAM is applied to capture the intra-group relationships after summing the three feature maps. The final refined output is obtained by sequentially multiplying the concatenated feature map with the two CAMs. In summary, the CRAM of feature map $F \in R^{H\times W\times C}$ can be defined as follows:

$$\begin{aligned}
\mathrm{CAM}(F) &= \sigma(\mathrm{MLP}(\mathrm{AP}(F)) + \mathrm{MLP}(\mathrm{MP}(F))),\\
M_{\mathrm{intra}} &= \mathrm{CAM}\left(x^{l+1}_{\mathrm{FPN}} + x^{l}_{\mathrm{Down}} + x^{l-1}_{\mathrm{Down}}\right),\\
M_{\mathrm{inter}} &= \mathrm{CAM}\left(x^{l+1}_{\mathrm{FPN}}, x^{l}_{\mathrm{Down}}, x^{l-1}_{\mathrm{Down}}\right),\\
\mathrm{CRAM} &= M_{\mathrm{inter}} \otimes \left[\left(M_{\mathrm{intra}} \otimes x^{l+1}_{\mathrm{FPN}}\right) + \left(M_{\mathrm{intra}} \otimes x^{l}_{\mathrm{Down}}\right)\right.\\
&\quad \left. + \left(M_{\mathrm{intra}} \otimes x^{l-1}_{\mathrm{Down}}\right)\right],
\end{aligned} \tag{1}$$

where σ is the sigmoid function and MLP represents the multilayer perception layer. AP and MP denote average pooling and max pooling operations, respectively. $[\cdot]$ denotes the concatenation operation, and $\otimes$ denotes the element-wise multiplication between feature maps and attention maps. l denotes the feature level, and the larger the l value, the deeper the layer.

To complement the refined feature maps generated by the CRAM, the Spatial Attention Module (SAM) [25] is introduced. The SAM is incorporated to specifically

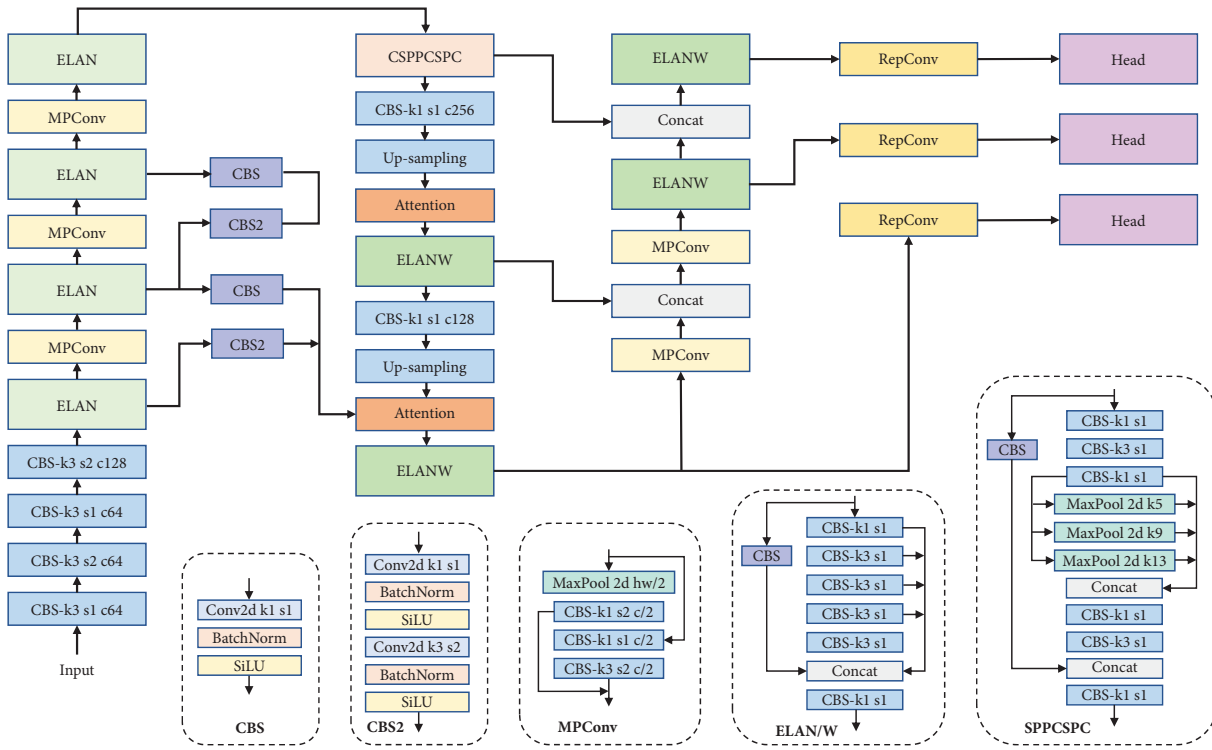

Figure 1: The architecture of the proposed FFA-YOLOv7.

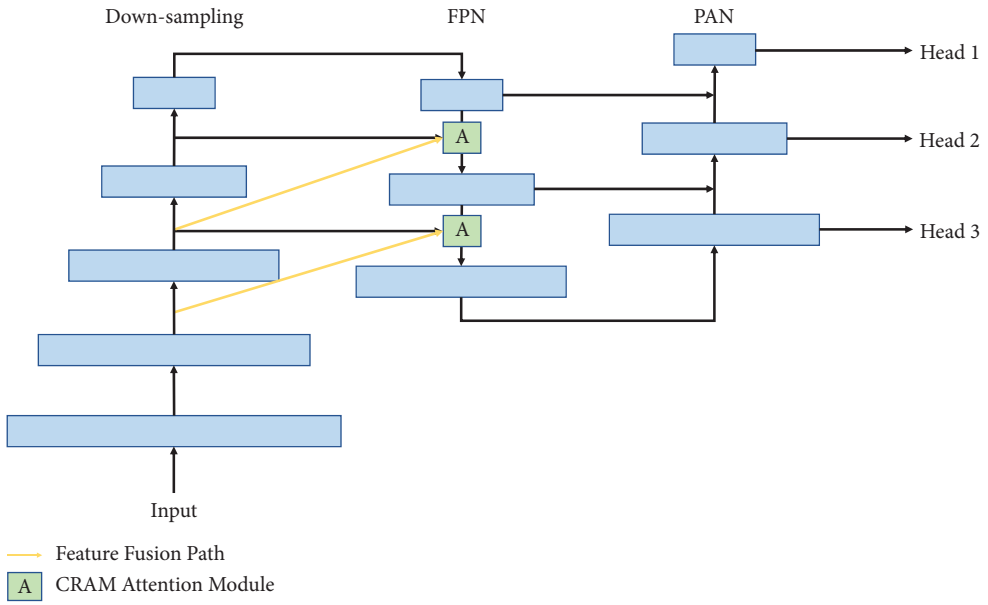

Figure 2: The structure of the feature fusion path.

emphasize the more accurate location of semantic and spatial information on the feature map, leading to an improved accuracy of feature representation. By integrating both CRAM and SAM, a feature map is obtained that encompasses both channel refinement and spatial refinement features. This feature map serves as an effective input for subsequent detection and recognition tasks.

4. Experiment

4.1. Dataset Construction. To comprehensively validate the effectiveness and practicality of the data, the experimental dataset used in this study comprises six distinct categories in the power grid context: (1) Helmet-wearing, (2) Helmet-not-wearing, (3) Seatbelt-wearing, (4) Seatbelt-not-wearing, (5) Ladder, and (6) Insulator. Among these, the first five

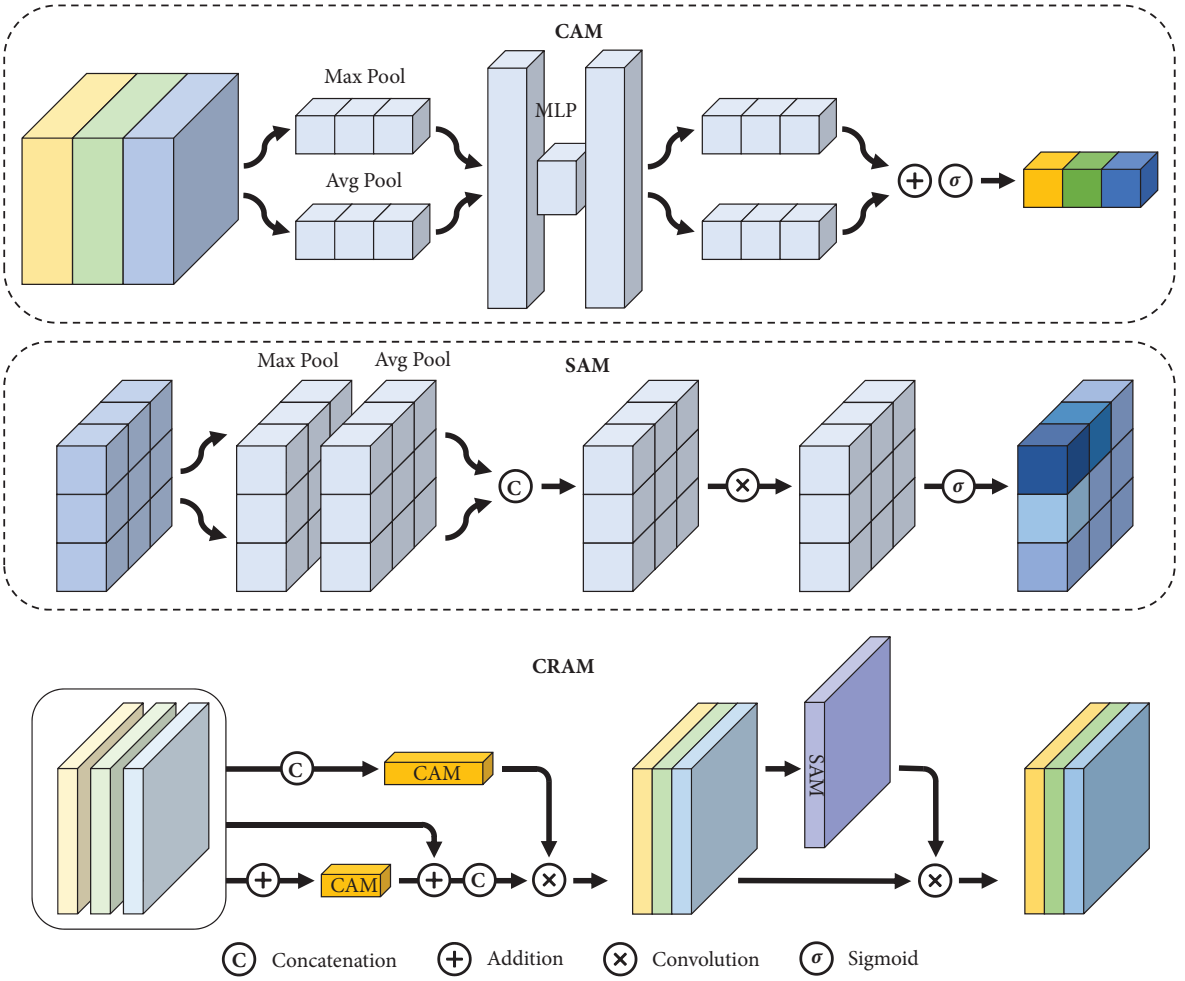

Figure 3: The structure of the CRAM.

categories include original images sourced from the internet or captured on-site during power operation activities. The images in the last category are exclusively obtained from real construction sites of substations. Figure 4 showcases a selection of image samples extracted from the dataset.

Data augmentation is commonly used to generate additional samples for detection objects in training data that are insufficient. Its principle involves creating a new dataset by applying various data augmentation methods to the existing dataset. In this study, five key data augmentation techniques were employed, including (1) object segmentation and background fusion, (2) partial erasing, (3) affine transformation, (4) brightness transformation, and (5) clarity transformation. These operations simulate common imaging condition variations in real-world surveillance, such as changes in viewpoint, distance, background, clarity, and illumination.

We generated a dataset consisting of 14,960 images (including 30% data augmentation images) by applying image augmentation techniques to each class of images. An 8 : 2 ratio was employed to split the dataset into a training dataset and a validation dataset. In order to maintain class balance, the number of images in each class was adjusted accordingly. Table 1 presents the parameters utilized for data augmentation. Furthermore, a separate testing dataset, consisting of 500 images captured from realistic power construction surveillance scenarios, was assembled. This testing dataset encompasses all six classes. A comprehensive breakdown of the dataset composition can be found in Table 2.

4.2. Evaluation Criteria. In this experiment, we employed four commonly used evaluation metrics to assess the performance of the detection model: precision (P), recall (R), $F1$ score ($F1$), and mean average precision (mAP). P, R, $F1$, and mAP can be formulated as follows:

$$\begin{aligned} P &= \frac{T_P}{T_P + F_P}, \\ R &= \frac{T_P}{T_P + F_N}, \\ F1 &= \frac{2 * P * R}{P + R}, \\ AP_i &= \int_0^1 P(R)dR, \\ mAP &= \frac{1}{N}\sum_i^N AP_i, \end{aligned} \tag{2}$$

Figure 4: Dataset of six classes of detection objects.

Table 1: Data augmentation parameter settings.

Parameter name	Value
Rotation range	±15°
Translation range	±0.2
Scale range	0.8–1.2
Pixel intensity range	0.5–1.5 per channel
Contrast range	0.5–2 per channel
Horizontal flip probability	0.5
Gaussian distribution	$\mu = 0,\ \sigma = 0.05 \times 255$
Gaussian blur probability	0.5
Gaussian kernel size	3 pixels

Table 2: Composition of the dataset.

Class	Value	Validation	Total
Ladder	3384	846	4230
Insulator	480	120	600
Helmet (with and without)	3680	920	4600
Safety belt (with and without)	4424	1106	5530
Total	11968	2992	14960

where F_P (false positive) denotes the number of objects that were incorrectly detected, F_N (false negative) is used to represent the number of objects that were missed in the detection, and T_P (true positive) represents the number of correctly detected objects. $T_P + F_P$ represent the total number of detected objects, and $T_P + F_N$ represent the total number of actual objects in the dataset. The $F1$ score is the harmonic mean of precision (P) and recall (R), providing a balanced measure of the model's performance. AP (average precision) is computed by calculating the area under the precision-recall curve (PR curve), which describes the trade-off between precision and recall at different thresholds. mAP is the average of AP values across different classes. N represents the number of classes in the test samples. In object detection, higher precision values indicate fewer false detections, while higher recall values indicate fewer missed detections. Therefore, achieving high precision and recall is crucial for accurate and comprehensive object detection.

4.3. Experimental Configuration. This experiment was conducted using the PyTorch deep learning framework on an NVIDIA RTX-3090 GPU. The experimental setup and hyperparameters are as follows. We utilized the Adam optimizer [26] to train our model with an initial learning rate of $1e-3$. The momentum was set to 0.937, and the weight decay was set to $5e-4$. The weights of the convolutional layers were initialized using the Kaiming normalization method. The remaining hyperparameters followed the default values specified in the YOLOv7-s official code. The model was trained for 400 epochs with a batch size of 16.

5. Results

Our proposed FFA-YOLOv7 aims to achieve high accuracy and efficiency in detecting violations in practical power construction sites. To objectively evaluate the performance of our proposed model, this study conducted a comparative analysis with five state-of-the-art object detection models. These models include YOLOv5-s, YOLOv5-m, YOLOv5-l, YOLOv7, and YOLOv7-x. We utilized pretrained weights from the YOLO framework and trained the models on our own constructed dataset. The dataset consists of 14,960 images with 30% data augmentation. By comparing the performance of different models on the same dataset, we are able to provide an objective assessment of the performance of our proposed model. We also utilized a separate testing dataset consisting of 500 images captured from realistic power construction surveillance scenarios. By comparing the results obtained from these different models, we can evaluate the effectiveness and performance of our proposed approach in detecting objects accurately and robustly in the specific context of power construction surveillance.

The detection results after 400 epochs of training are shown in Table 3. YOLOv5-s, with its simple model structure, achieves the best speed performance but the worst precision performance. YOLOv5-m and YOLOv5-l, with more complicated model structures, exhibit better precision but slower speed. The YOLOv7-based models, on the other hand, generally perform better than the YOLOv5 models. In

Table 3: Comparison of performance among state-of-the-art object detection models.

Model	P (%)	R (%)	mAP (%)	$F1$ (%)	Speed (ms)
YOLOv5-s	94.62	95.32	96.45	94.93	8.6
YOLOv5-m	95.14	95.33	96.67	95.22	12.0
YOLOv5-l	95.18	96.02	96.71	95.53	14.0
YOLOv7	95.76	95.92	97.64	95.84	8.8
YOLOv7-x	95.55	96.45	97.82	96.00	11.5
FFA-YOLOv7	95.92	97.13	98.16	96.50	9.6

Table 4: Performance of FA-YOLOv7 on the power construction dataset.

Class	P (%)	R (%)	mAP (%)	$F1$ (%)
No safety belt	88.62	96.51	96.23	92.40
Safety belt	94.23	95.88	95.01	95.04
No helmet	95.58	96.72	96.32	96.15
Helmet	96.82	97.75	97.53	97.28
Insulator	85.83	82.45	88.57	84.10
Ladder	89.03	89.29	90.35	89.16
All	91.69	93.10	94.00	92.36

Figure 5: The detection performance of the FFA-YOLOv7 model. The operators wearing safety helmets and safety belts are highlighted with green bounding boxes. Those who are not wearing safety helmets, not wearing safety belts, or violating safety regulations are marked with red bounding boxes. The insulators and ladders are, respectively, indicated by orange and blue bounding boxes.

comparison, our proposed model outperforms all of the aforementioned models as for P, R, mAP, and $F1$.

In order to validate the effectiveness of the proposed FFA-YOLOv7 model in this paper, we conducted a comprehensive evaluation on a dataset collected from real-world power construction sites. The dataset encompasses a wide range of detection objects from all six classes. The evaluation aimed to measure the performance of our model in accurately detecting these objects. Table 4 presents the test results of the proposed method, providing a detailed list of metrics for each class and an overall evaluation of the model's performance. Additionally, the corresponding detection results are showcased in Figure 5, providing a visual demonstration of the accuracy and effectiveness of our proposed method. The outcomes presented in Table 4 and Figure 5 highlight the excellent detection accuracy achieved by our FFA-YOLOv7 model. It successfully detects objects from all six classes across diverse practical power operation sites, exhibiting both class-specific and overall high-performance capabilities. It is evident from the results that the method proposed in our work demonstrates effectiveness in target detection within the field of power construction monitoring.

Furthermore, we conducted an evaluation of the detection speed and concurrency of our proposed system using the testing dataset. The results demonstrate that our network achieves a detection speed of 9.6 ms per image, indicating its high efficiency in practical applications. Additionally, the system exhibits remarkable concurrency, allowing it to simultaneously record and detect up to 30 different video streams at a real-time frame rate of 30 FPS.

6. Conclusions

In this study, we propose FFA-YOLOv7, an improved version of YOLOv7, for detecting worker-wearing violations in substation construction. The process of downsampling and upsampling usually leads to location information loss, and the edge positioning accuracy and detection performance will be further affected. To address the issue, a new feature fusion path is presented to synthesize rich semantic information and precise location information from deep layers and shallow layers, respectively. Additionally, attention modules are incorporated to refine the fused features. Furthermore, we establish a dataset to compensate for the limited training samples, enabling better detection performance in realistic power construction scenarios. Compared to other YOLO-based detection methods, our proposed FFA-YOLOv7 achieves the highest detection accuracy (96.5%) without compromising detection speed. Experimental results on a dataset collected from realistic power construction sites demonstrate that FFA-YOLOv7 exhibits superior accuracy and robustness in detecting violations in practical power construction scenarios.

Data Availability

The data used to support the findings of this study are available on request from the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The science and technology program "Research on remote safety control technology of power field operation based on infrared and visible multi-source image fusion" funded by the China Southern Power Grid provided funding for this effort. This work was also partially supported by the Yunnan Province Ten Thousand Talents Program.

References

[1] H. Jiangtao, "Discussion on the construction of substation security video surveillance system," *IOP Conference Series: Materials Science and Engineering*, vol. 563, p. 32004, 2019.

[2] J. Yang, Y. Wang, X. He et al., "Optimized configuration of video surveillance layout for substation construction site for full coverage surveillance," in *Proceedings of the 2022 7th Asia Conference on Power and Electrical Engineering (ACPEE)*, pp. 1932–1939, IEEE, Hangzhou, China, April 2022.

[3] P. Lu, F. Zhang, and S. Fan, "Research on substation intelligent monitoring scheme under big data environment," in *Proceedings of the 2018 China International Conference on Electricity Distribution (CICED)*, pp. 1530–1536, IEEE, Tianjin, China, September 2018.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, pp. 779–788, Washington, DC, USA, June 2016.

[5] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, Honolulu, Hawaii, USA, May 2017.

[6] A. Farhadi and J. Redmon, "Yolov3: an incremental improvement," *Computer Vision and Pattern Recognition*, Vol. 1804, Springer, Berlin, Germany, 2018.

[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," p. 10934, 2020, https://arxiv.org/abs/2004.10934.

[8] C. Y. Wang, A. Bochkovskiy, and L. H. Y. M. Yolov7, "Trainable bag-of-freebies sets new state-of-the-art for real-time object 285 detectors," 2022, https://arxiv.org/abs/2207.02696.

[9] A. Hayat and F. Morgado-Dias, "Deep learning-based automatic safety helmet detection system for construction safety," *Applied Sciences*, vol. 12, no. 16, p. 8268, 2022.

[10] D. Fu, L. Gao, T. Hu, S. Wang, and W. Liu, "Research on safety helmet detection algorithm of power workers based on improved yolov5," *Journal of Physics: Conference Series*, vol. 2171, p. 12006, 2022.

[11] J. Zhang, P. Qu, C. Sun, and M. Luo, "Safety helmet wearing detection algorithm based on improved yolov5," *Journal of Computer Applications*, vol. 42, no. 4, p. 1292, 2022.

[12] L. Wang, Y. Cao, S. Wang et al., "Investigation into recognition algorithm of helmet violation based on yolov5-cbam-dcn," *IEEE Access*, vol. 28, pp. 56–89, 2022.

[13] N. Ni and C. Hu, "Automatic detection of safety helmet based on improved yolo deep model," in *Advanced Intelligent Technologies for Industry*, Springer, Berlin, Germany, 2022.

[14] Z. Xu, Y. Zhang, J. Cheng, and G. Ge, "Safety helmet wearing detection based on yolov5 of attention mechanism," *Journal of Physics: Conference Series*, vol. 2213, p. 12038, 2022.

[15] C. Sun, S. Zhang, P. Qu et al., "Mca-yolov5-light: a faster, stronger and lighter algorithm for helmet-wearing detection," *Applied Sciences*, vol. 12, no. 19, p. 9697, 2022.

[16] Y. Tamaazousti, D. Egorov, A. Benzine, U. Asif, M. Amri, and A. Benaichouche, *Smarthse: Comprehensive Ai-Based Safety Monitoring Solution for Work-Sites*, ADIPEC, Abu Dhabi, United Arab Emirates, 2022.

[17] Z. Li, W. Xie, L. Zhang et al., "Toward efficient safety helmet detection based on yolov5 with hierarchical positive sample selection and box density filtering," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[18] R. Tyagi and K. Thomas, "Multiple safety equipment's detection at active construction sites using effective deep learning techniques," in *Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1271–1276, IEEE, Tirunelveli, India, June 2022.

[19] C. Fang, H. Xiang, C. Leng, J. Chen, and Q. Yu, "Research on real-time detection of safety harness wearing of workshop personnel based on yolov5 and openpose," *Sustainability*, vol. 14, no. 10, p. 5872, 2022.

[20] W. Zhu, Y. Shu, and S. Liu, "Power grid field violation recognition algorithm based on enhanced yolov5," *Journal of Physics: Conference Series*, vol. 2209, p. 12033, 2022.

[21] B. Zhao, H. Lan, Z. Niu, H. Zhu, T. Qian, and W. Tang, "Detection and location of safety protective wear in power substation operation using wear-enhanced yolov3 algorithm," *IEEE Access*, vol. 9, pp. 125540–125549, 2021.

[22] H. Xia, B. Yang, Y. Li, and B. Wang, "An improved centernet model for insulator defect detection using aerial imagery," *Sensors*, vol. 22, no. 8, p. 2850, 2022.

[23] Z. Zhang, S. Huang, Y. Li, H. Li, and H. Hao, "Image detection of insulator defects based on morphological processing and deep learning," *Energies*, vol. 15, no. 7, p. 2465, 2022.

[24] Q. Li, F. Zhao, Z. Xu, J. Wang, K. Liu, and L. Qin, "Insulator and damage detection and location based on yolov5," in *Proceedings of the 2022 International Conference on Power Energy Systems and Applications (ICoPESA)*, pp. 17–24, IEEE, Guangzhou, China, July 2022.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, Tel Aviv, Israel, December 2018.

[26] D. P. Kingma, J. Ba, and Adam, "A method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.