Quest Journals
Journal of Software Engineering and Simulation
Volume 11 ~ Issue 2 (2025) pp: 01-08
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Foreign Object Detection Algorithm for Transmission Lines Based on YOLO-World

Shulin Xun[1]
[1]*School of Control and Computer Engineering, North China Electric Power University, Baoding, 071051, China*
Corresponding Author:Shulin Xun

ABSTRACT: *Foreign object detection based on unmanned aerial vehicle (UAV) aerial images is an important part of intelligent inspection of transmission lines. The YOLO series of object detection algorithms have become mainstream methods due to their high accuracy and fast detection capabilities. However, they require predefined target categories, which limits their applicability in open scenarios. YOLO-World enhances its adaptability to open scenarios through visual-language modeling and large-scale dataset pre-training. However, the environment for foreign object detection in transmission lines is complex, and the types of targets are numerous, which are not fully covered in general datasets. To fully leverage the advantages of YOLO-World, it must be fine-tuned to better learn and adapt to the specific foreign object dataset of transmission lines. Experimental results show that the fine-tuned YOLO-World significantly improves detection accuracy and reliability in actual inspections, effectively identifying various unknown foreign objects, thereby ensuring the safe operation of transmission lines.*

KEYWORDS: *Transmission lines; foreign body detection; object detection; YOLO-World*

R*eceived 08 Feb., 2024;* ***R****evised 16 Feb., 2025;* ***A****ccepted 18 Feb., 2025 © The author(s) 2025.*
P*ublished with open access at* ***www.questjournas.org***

I. INTRODUCTION

With the continuous growth of power demand and the gradual expansion of power grids, the safety and stability of transmission lines have become particularly important. However, as transmission lines are constantly exposed to the outdoor environment, they are prone to interference from foreign objects such as bird nests, balloons, kites, and garbage. If these foreign objects are not promptly removed, they may cause power grid failures, power outages, and even serious accidents such as fires. Therefore, the detection and removal of foreign objects on transmission lines have become an important link in ensuring the safe operation of power grids.

The traditional transmission line inspection mainly relies on manual operation, and inspectors need to inspect along the transmission line on foot or using inspection vehicles. Although this method can find part of the hidden dangers, but there are obvious limitations, such as inspection efficiency is low, the operation of the safety hazards are large, the inspector may face the danger of working at height, electromagnetic radiation and so on. In order to overcome the various shortcomings of manual inspection, intelligent inspection technology came into being. Based on drones, robots and advanced computer vision technology, intelligent inspection can obtain real-time images of transmission lines through high-definition cameras or infrared imaging equipment without contacting the transmission lines, and use deep learning algorithms to automatically identify abnormalities in the images. Intelligent inspection not only significantly improves the inspection efficiency, but also realizes high-precision and all-round detection of foreign objects in complex environments, so as to better protect the safe operation of the power grid.

In recent years, the application of deep learning methods for foreign object detection in transmission line images has become a research hotspot and difficulty in the field of smart grid. Literature [1] proposed an improved YOLOv8 model (TFD-YOLOv8), which significantly improves the precision and accuracy of foreign object detection in transmission lines by double-branch downsampling and hybrid augmented attention module. Literature [2] proposed an improved YOLOv3 (YOLOv3-RepVGG) based on RepVGG, which significantly improves the mAP, precision and recall of foreign object detection by enhancing the feature extraction and multi-scale detection capabilities. Literature [3] combines Swin Transformer with YOLOv5 to optimize foreign

object detection in transmission line channels, especially in complex background and small target detection to significantly improve the detection accuracy. Literature [4] improved the feature pyramid pooling and BCE loss function of YOLOv4, which enhanced the target information retention and similar target differentiation ability under the background interference, and improved the detection accuracy. Literature [5] proposed an improved YOLOv7 algorithm for transmission line foreign object detection in response to the problem of slow detection and low accuracy of transmission lines by introducing BiFormer attention mechanism to enhance the ability of small target detection, using the GSConv module to reduce the model complexity in order to improve the detection speed, and using the Shape-IoU loss function to improve the bounding box regression accuracy.

Although the above methods are excellent in handling complex backgrounds and small target detection, their applicability in open scenarios is limited by the need to predefine target categories. When faced with undefined new categories, the detection ability of the model may be significantly degraded. In addition, traditional target detection algorithms [6-10] are highly dependent on a large amount of labeled data for training, which makes it difficult for them to flexibly adapt to changing real-world application scenarios. Therefore, this suggests that while improving the detection accuracy, we need to explore detection models with stronger generalization ability and adaptability to better meet the challenges in open scenarios.

II. RELATED WORK

The YOLO-World network model consists of a YOLO detector, a text encoder, and a reparameterizable Visual Verbal Path Aggregation Network (RepVL-PAN). After the model receives the input text, the text encoder transforms it into a text embedding; meanwhile, the image encoder in the YOLO detector extracts multi-scale features from the input image. Subsequently, RepVL-PAN combines the image features with the text embedding through a cross-modal fusion mechanism to enhance the representation of both image and text. Diagram 1 illustrates the network structure of YOLO-World.

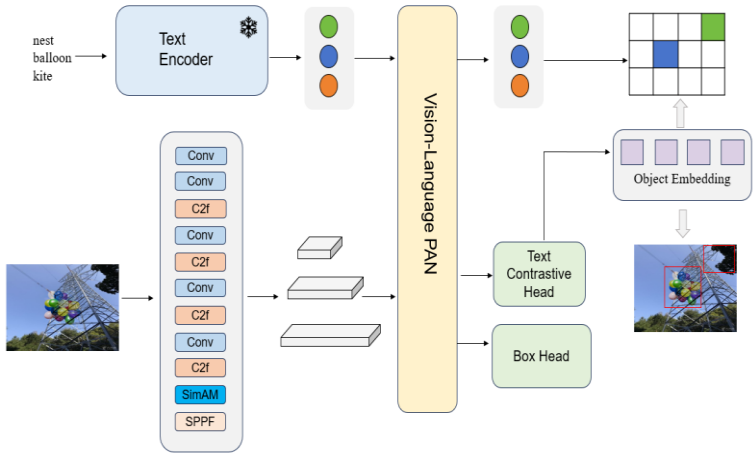

Diagram 1

2.1 YOLO detector

YOLO-World [11] is developed based on YOLOv8 [12] and uses Darknet [13] as the image encoder, which is a highly efficient convolutional neural network architecture first introduced in the YOLO series, and has been optimized many times to provide high computational efficiency and powerful feature extraction capabilities. The network gradually extracts spatial features and contextual information from images through multiple convolutional layers, activation functions and pooling layers, and is particularly good at capturing complex patterns and textures.

In addition, YOLOv8 contains a Path Aggregation Network (PAN) and a Head layer for bounding box regression and object embedding.The PAN enhances the robustness of the model to targets at different scales by fusing features from different layers. Specifically, the PAN extracts high semantic information from deep features while obtaining detail information from shallow features, so that the model can capture the overall contour of large objects as well as detect the details of small objects, thus improving the performance of multi-scale target detection.The Head layer is responsible for generating the final prediction results.

2.2 Text Encoder

When processing the input text, YOLO-World uses a pre-trained Transformer text encoder based on CLIP [14], a multimodal learning model proposed by OpenAI, which learns by large-scale graphic comparisons to align natural language with images, thus creating a stronger visual and linguistic semantic association between vision and language. Unlike traditional text-only language encoders, the CLIP text encoder has the ability to characterize both vision and language through cross-modal training. Given an input text T, the CLIP

text encoder transforms it into a text embedding$w=Textencoder(T)\in R^{C\times D}$, where C is the number of nouns and D is the dimension of the embedding vector. Compared to traditional text-only language encoders, CLIP text encoders not only have excellent text processing capabilities, but also provide stronger visual semantic associations that can effectively link visual objects to textual descriptions. This multimodal feature enables YOLO-World to more accurately understand and detect targets based on natural language descriptions in open scenarios, thus enhancing the generalization ability and applicability of the model.

2.3 RepVL-PAN

The structure of RepVL-PAN proposed by YOLO-World follows the top-down and bottom-up paths in the literature [15] and builds a feature pyramid {P3, P4, P5} through multi-scale image features {C3, C4, C5}. In the top-down path, the model aggregates deep and high semantic information, enabling it to better understand the semantic level of large targets; while in the bottom-up path, the model utilizes shallow features to supplement spatial detail information, thus enhancing the perception of small targets. This bi-directional feature fusion strategy effectively combines the advantages of different scale features, which improves the global semantic understanding of large targets and enhances the detail perception of small targets, ensuring that the model performs well when dealing with targets at different scales, with stronger robustness and adaptability.

III. METHODOLOGY

In this study, we adopted a fine-tuning method to optimize the foreign body detection algorithm of transmission lines based on YOLO-World. Fine-tuning is a transfer learning technique in which an existing pre-trained model is retrained using domain-specific data so that the model can adapt to a new task. The process of fine-tuning can be understood in terms of the source domain and the target domain. Source domains correspond to general-purpose object detection datasets (such as COCO) that contain large amounts of annotated data, often containing rich general-purpose image content and multiple object classes. On the source domain, YOLO series models have been fully trained and have learned a wide range of visual features such as edges, textures, shapes, etc. The target domain is the foreign object detection data set of our practical application of transmission lines. In the target domain, the type and distribution of data are different from the source domain, and mainly contain specific foreign objects (such as bird's nests, branches, garbage, etc.) in the image of the transmission line. Because the training data for the target domain is usually limited, training a new model directly can result in overfitting or inadequate training. By fine-tuning, we transfer the knowledge learned from the source domain to the target domain. The fine-tuning principle is shown in Diagram 2.

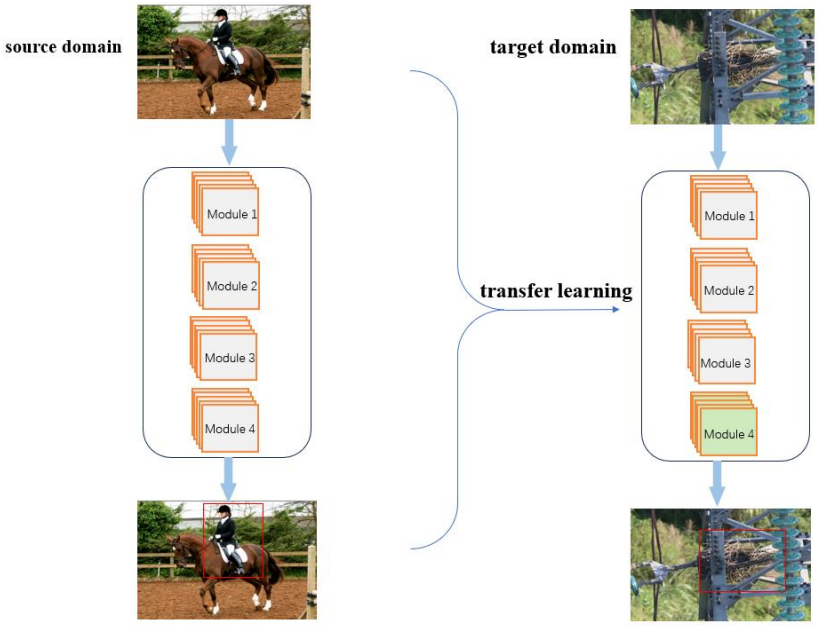

Diagram 2

The detailed steps of the fine-tuning process are as follows: (1) Model initialization: Load the pre-trained YOLO-World model on the source domain dataset and check its weight distribution. (2) Data preparation: the data in the target domain are labeled and pre-processed, and the training and verification data set including the detection of foreign bodies in transmission lines is constructed. Use data enhancement techniques (such as rotation, cropping, brightness adjustment, etc.) to compensate for the lack of data volume.

(3) Specific layer freezing: By analyzing the characteristic differences between the source domain and the target domain, the layers responsible for low-level feature extraction in the model are selectively frozen to ensure that the weights of these layers remain unchanged during training. (4) Fine-tuning training: The unfrozen layer is trained at a low learning rate so that it can adapt to the specific task of the target domain. At the same time, appropriate optimizers and loss functions are used to balance the performance of the model in the target domain. (5) Model validation: Evaluate the performance of the fine-tuning model on the validation data set of the target domain, and use evaluation indicators to measure the detection effect of the model. Our training strategy is to dynamically adjust the proportion of frozen layer and unfrozen layer during training, and gradually thaw more layers, so as to further optimize the overall performance of the model. By refining these intermediate processes, we ensure that each step of the fine-tuning has a clear objective and operational basis, thus optimizing the YOLo-World-based foreign matter detection algorithm for transmission lines.

IV. EXPERIMENTAL RESULT

4.1 Experimental Setup

Dataset. The dataset used in this paper is mainly derived from the transmission line foreign object recognition images of the Southern Power Grid, which contains four types of foreign objects: bird's nests, balloons, kites and garbage, totaling 1300 images. According to the ratio of 7:3, the data are divided into training set and test set, which contain 900 and 400 images respectively, and the specific distribution is shown in Figure 1. In the open vocabulary target detection task, the base classes refer to the known categories used in the training phase of the model. The model grasps the features, semantic information and correspondence with the visual data through supervised learning of these base classes; new classes refer to the classes that do not appear in the model training phase but may appear in the inference or testing phase. In the experiments of this paper, bird's nest, kite and garbage are set as the base classes, and balloon as the new class. In order to simulate the open scenario, we process the training set by removing the labeling about the balloon category in it. The hardware and software configurations used for the experiments are shown in Table 1.

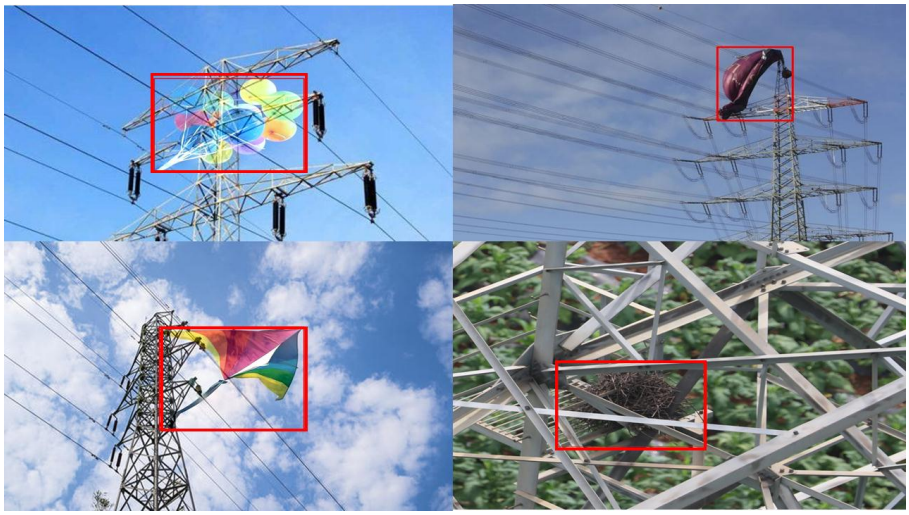

Figure 3: Partial image of the dataset

Table 1: Experimental environment

Hardware Name	version
CPU	Intel(R) Core(TM) i9-9900K CPU @ 3.60GHz
GPU	GeForce RTX 2080
operating system	Ubuntu 18.04
frameworks	Pytorch 1.10.0
Computing Architecture	Cuda 11.3
language	Python3.8

Pre-trained models.YOLO-World uses detection or annotation datasets including Objects365 (V1) [16], GQA [17], and Flickr30k [18]. Among them, GoldG (GQA and Flickr30k) are the images excluded from the COCO dataset. The annotations of the detection datasets used for pre-training all contain bounding boxes and categories or noun phrases. In addition, YOLO-World extends the pre-training data using image-text pairs, i.e., CC3M-Lite

[19], which annotates 246k images by pseudo-labeling methods. The specific pre-training models are shown in Figure 2.

Table 2: Pre-trained models

Model	Pre-train Data	Size
YOLO-Worldv2-S	O365+GoldG	640
YOLO-Worldv2-M	O365+GoldG	640
YOLO-Worldv2-L	O365+GoldG	640
YOLO-Worldv2-XL	O365+GoldG+CC3M-Lite	640

Evaluation indicators.(1) mAP (Mean Average Precision): mAP represents the average precision of all categories, which is the core index to measure the performance of the model in the target detection task. mAP is derived by calculating the average precision (AP) of each category and then averaging the AP values of all categories. It is derived by calculating the average precision (AP) of each category and then averaging the AP values of all categories. mAP is the area of a category under the Precision-Recall curve, which combines the performance of detection precision and recall.

(2) mAP50 is the mAP when the IoU threshold is 0.5. IoU is a measure of the degree of overlap between the predicted bounding box and the real bounding box. mAP50 reflects the performance of the model when the predicted bounding box overlaps with the real bounding box to a certain extent, and it is a more commonly used evaluation criterion.

(3) mAP75 is the mAP at an IoU threshold of 0.75. This metric is more stringent because it only considers those cases where the predicted bounding box has a high overlap with the true bounding box. mAP75 is usually lower than mAP50 because it is evaluated by a more stringent criterion.

(4) mAPs usually refers to the mAP on small objects. detection of small objects is a challenge when evaluating target detection models because they occupy fewer pixels in the image and are relatively difficult to recognize. mAPs specifically measures the performance of the model in detecting small objects.

(5) mAPm is the mAP on medium sized objects. this metric focuses on the performance of the model in detecting medium sized objects.

(6) mAPl is the mAP on large objects. this metric measures the model's performance in detecting large-sized objects. Typically, large objects are relatively easy to detect.

4.2 Detection Effectiveness

We initialize YOLO-World with the weights of YOLO-Worldv2-s and fine-tune it. The total number of training times is 500, the initial learning rate is set to 0.00002, the input picture size is 640x640, and all categories are covered during evaluation. Figure 2 shows the changing trend of loss value during training. When the 480th epoch was reached, the loss curve began to change slowly and gradually became stable.

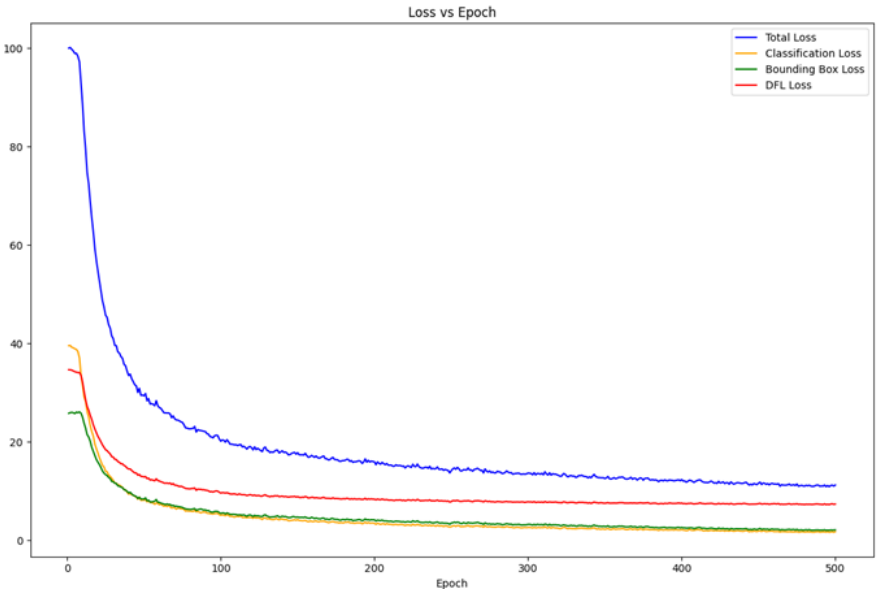

Figure 2: Loss curve change

We performed fine-tuning on different pre-trained models respectively, and the experimental results after fine-tuning are shown in Table 3.

Table 3: Experimental results for different pre-trained models

Model	mAP	mAP50	mAP75	mAPs	mAPm	mAPl
YOLO-Worldv2-s	1.37	2.48	1.34	0.27	7.4	1.40
YOLO-Worldv2-m	1.70	3.60	1.30	0.50	4.60	1.70
YOLO-Worldv2-L	0.50	2.00	3.00	0.00	1.00	0.70
YOLO-Worldv2-XL	7.70	15.2	7.50	1.60	7.30	7.90

The YOLO-Worldv2-s model performs relatively weakly, with a mAP of only 1.37, indicating that the model has a low overall accuracy. It performs better in medium target detection, but is less effective in small and large target detection. mAP50 is 2.48, indicating that the detection accuracy at low IoU thresholds is also not very high. Overall, YOLO-Worldv2-s is suitable for simple or low-complexity tasks, and is less suitable for large and small targets.YOLO-Worldv2-m, like YOLO-Worldv2-s, performs relatively well in medium target detection, but falls short in small and large targets, and is suitable for lightweight scenarios.YOLO-Worldv2-L model performs the best, but is less effective in small and large targets. Worldv2-L model performs the worst, while the YOLO-Worldv2-XL model is far ahead in overall performance, especially in detecting medium and large targets. It is suitable for target detection tasks in complex scenes.

We explored the effect of freezing the first n layers of the pretrained model on the experimental results. Freezing the first few layers of the pre-training model was able to obtain better results even with a smaller amount of data. This is because the first few layers have already learned some generalized features that can be directly used in new tasks. By fine-tuning only the later layers, the model can learn task-specific features even on less data. Table 4 shows the experimental results of freezing the first n layers.

Table 4: Experimental results for N layers before freezing

Pre-freeze N-layer	mAP	mAP50	mAP75	mAPs	mAPm	mAPl
N=1	0.40	0.90	0.40	18.3	0.10	0.50
N=2	0.30	0.70	0.20	0.00	0.60	0.30
N=3	0.40	1.30	0.20	1.70	1.60	0.40
N=4	0.30	1.30	0.00	0. 00	0. 20	0. 40

From the experimental results, it can be seen that the choice of the number of frozen layers needs to be weighed against the needs of the task. Freezing fewer layers (e.g., 1 layer) favors small target detection, while freezing a moderate number of layers (e.g., 3 layers) performs better for medium target detection. If too many layers are frozen, the model loses its adaptability to a specific task and performs poorly especially in high IoU or small target detection. Therefore, a reasonable choice of the number of frozen layers can effectively improve the model performance when the amount of data is small.

4.3 Fine-tuning the CLIP text encoder

This experiment also compares the results of experiments with and without fine-tuning the text encoder, and for fine-tuning we fine-tuned the CLIP text encoder with a learning factor of 0.01. The experimental results are shown in Table 5.

Table 5: Experimental results of fine-tuning the CLIP text encoder

	mAP	mAP50	mAP75	mAPs	mAPm	mAPl
freeze	13.7	24.8	13.4	2.70	7.40	14.8
fine-tuning	14.2	25.5	13.7	4.40	8.70	14.5

The frozen CLIP text encoder model performs relatively consistently, especially well when dealing with large targets. However, the performance on small and medium targets is more average, probably because the model, without fine-tuning, does not capture fine-grained features in a given task well. The fine-tuned text encoder shows a small improvement in all metrics, especially in its ability to detect small and medium targets. This suggests that with fine-tuning in the presence of small amounts of data, the model can be better adapted to the specific task requirements, thus improving the accuracy of detection. Overall, the fine-tuned text encoder is able to obtain better results than the frozen text encoder, especially in the detection of small and medium targets, and is suitable for task scenarios with limited amounts of data.

aper fine-tuned YOLO-World by building a transmission line foreign body detection dataset, aiming to solve the problem of predefined categories that occur in traditional transmission line foreign body detection algorithms. In future engineering applications, more abundant images of suspended foreign bodies in

transmission lines can be collected according to actual needs to further enhance the performance of the model. By continuously expanding and optimizing the data set, the method in this paper will be able to identify more types of foreign bodies, thus further improving the operation and maintenance efficiency of transmission lines. The promotion and application of this technology will provide a strong guarantee for the safe and stable operation of the power system, and promote the intelligent development of the power industry.

4.4 Other datasets

In addition to fine-tuning on the training dataset without balloons, we also constructed three different base class images as training datasets to fine-tune the model: dataset 1 (balloons, bird's nests, and garbage), dataset 2 (balloons, garbage, and kites), and dataset 3 (balloons, garbage, and kites), respectively.

Table 6: Results of fine-tuning experiments with different training datasets

Training datasets	mAP	mAP50	mAP75	mAPs	mAPm	mAPl
dataset 1	13.4	24.7	12.0	1.0	9.7	14.1
dataset 2	0.50	1.40	0.30	0.0	1.00	0.60
dataset 3	1.10	1.90	0.90	40.8	0.8	1.10

Without training certain objects (e.g., kites, garbage, and bird nests), the model showed some generalization ability for the detection of these objects in the validation set. According to the results in Table 6, the model performed relatively well on the mAPs on dataset 3, suggesting that it may rely on similar small object features for detection. On the detection of dataset 1, the model is still able to have some generalization ability on larger-sized objects despite not having seen a kite in training, showing some image feature migration ability.

Figure 3: Transmission Line Balloon Recognition Results

For the foreign object images obtained from power inspection, this paper shows the detection results on the balloon category with the training set categories of bird's nest, kite, and garbage. Its specific recognition results are shown in Figure 3.

V. CONCLUSION

This paper fine-tuned YOLO-World by building a transmission line foreign body detection dataset, aiming to solve the problem of predefined categories that occur in traditional transmission line foreign body detection algorithms. In future engineering applications, more abundant images of suspended foreign bodies in transmission lines can be collected according to actual needs to further enhance the performance of the model. By continuously expanding and optimizing the data set, the method in this paper will be able to identify more types of foreign bodies, thus further improving the operation and maintenance efficiency of transmission lines. The promotion and application of this technology will provide a strong guarantee for the safe and stable operation of the power system, and promote the intelligent development of the power industry.

REFERENCES

[1]. Xue, A., Jiang, E., Zhang, W., Lin, S., and Mi, Y., Detection of foreign bodies in transmission line channels based on the fusion of Swin Transformer and YOLOv5, Journal of Shanghai Jiaotong University, 2023. pp. 1–22.

[2]. Zhang, Q., Zhu, T., Xiao, S., Yang, Z., Zeng, H., Zhang, C., and Li, G., YANG Zhong, ZENG Huarong, ZHANG Chi, LI Guotao, Foreign object detection of high voltage transmission line based on improved YOLOv4 algorithm, Applied Science and Technology, 2023. 50(04): pp. 59–65.

[3]. Yu, Y., Qiu, Z., Zhou, Y., Zhu, X., and Wang, Q., Foreign Body Detection for Transmission Lines Based onConvolutional Neural Network and ECOC-SVM, Smart Power, 2022. 50(03): pp. 87-92+107.

[4]. Sun, Y., and Li, J., YOLOv7-tiny Transmission Line Foreign Object DetectionAlgorithm Based on Channel Pruning, Computer Engineering and Applications, 2024. 60(14): pp. 319–328.

[5]. Xiong, H., and Xiong, W., Foreign object detection method for transmission linesbased on improved YOLOv3, Journal of Shanghai Dianji University, vol. 26, no. 06, pp. 350-355+366, 2023.

[6]. Wang, Y., Feng, L., Song, X., Qu, Z., Yang, K., Wang, Q., and Zhai, Y., ZHAl Yongjie, TFD-YOLOv8: a transmission line foreign object detection method, Journal of Graphics, 45(05): pp. 901–912, 2024.

[7]. Zhang, H., Zhou, H., Li, S., and Li, P., Improved YOLOv3 foreign body detection method in transmission line, Laser Journal, 2022. 43(05): pp. 82–87.

[8]. Xu, Y., Tang, H., and Xiao, S., XlAo Shenping, Improved YOLOv5s-based Detection of Foreign Objects in Transmission Lines, Electric Engineering, 2023. 21: pp. 54-57+62.

[9]. Tang, X., Shen, W., Zhu, M., and Bao, W., The foreign object detecting algorithm for transmission lines based on theimproved YOLOv4', Journal of Anhui University(Natural Science Edition), 2021, 45(05): pp. 58–63.

[10]. Zhou, Y., and Liao, B., Foreign object detection in transmission lines based on improved YOLOv7 algorithm, Journal of North China Electric Power University(Natural Science Edition), 2024. pp. 1–9.

[11]. Cheng T, Song L, Ge Y, et al. Yolo-world: Real-time open-vocabulary object detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. pp. 16901-16911.

[12]. Varghese, R. and M. Sambath. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. in 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). 2024. IEEE: pp. 1-6.

[13]. Redmon, J., S. Divvala, R. Girshick, et al. You only look once: Unified, real-time object detection. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. pp. 779-788.

[14]. Radford, A., J.W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. in International Conference on Machine Learning. PMLR, 2021. pp. 8748-8763.

[15]. Liu, S., L. Qi, H. Qin, et al. Path aggregation network for instance segmentation. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. pp. 8759-8768.

[16]. Shao, S., Z. Li, T. Zhang, et al. Objects365: A large-scale, high-quality dataset for object detection. in Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019. pp. 8430-8439.

[17]. Hudson, D.A. and C.D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 6700-6709.

[18]. Plummer, B.A., L. Wang, C.M. Cervantes, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. in Proceedings of the IEEE International Conference on Computer Vision. 2015. pp. 2641-2649.