Proceedings of 14th IEEE International Conference
on CYBER Technology in Automation, Control, and Intelligent Systems
July 16 - 19, 2024, Copenhagen, Denmark

2024 IEEE 14th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER) | 979-8-3315-0605-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CYBER63482.2024.10749618

A Visual Foundation Model of Image Segmentation for Power Systems

Nannan Yan[1,2], Wenhao Guan[3], Xiao Yu[1], Jiawei Hou[3*], and Taiping Zeng[1*]

Abstract— **The automation of substation equipment inspection is a pivotal development area within the power industry. Traditional substation equipment inspection methods utilizing instance segmentation models trained on specific dataset have shown broad application, however, their generalization performance is limited to specific scenes. To enhance the robustness against intricate environments, we propose a two-stage instance segmentation method based on visual foundation models. In our work, the state-of-the-art object detector YOLOX and the visual foundation model SAM are employed to integrate the high-efficiency 2D detector with the general visual knowledge powered by foundation models trained on large-scale datasets. We utilize YOLOX to generate bounding box prompts which are processed by the pruned and aligned visual foundation model SlimSAM to perform instance segmentation with 68.6% mAP on our validation dataset. The method's effectiveness is validated through extensive comparisons with different model configurations and segmentation prompts, highlighting its robustness and potential for practical application in the domain of substation equipment maintenance and inspection.**

I. INTRODUCTION

The inspection of substation equipment is a critical task that ensures the reliability and safety of the electrical power grid. Traditionally, this process has been conducted manually by trained personnel, which is fraught with challenges such as labor intensity, safety risks, and potential inaccuracies due to human error. Recently, the advent of unmanned inspection methods has presented a significant improvement, offering numerous advantages over traditional approaches, including enhanced safety, reduced labor costs, and the ability to perform inspections in hard-to-reach or hazardous areas.

The effectiveness of unmanned inspection systems heavily relies on the precision of algorithmic detection and segmentation capabilities. While conventional detection algorithms have made strides in addressing these requirements, they often fall short in terms of generalization performance, particularly when dealing with complex scenes and diverse equipment types. The recent rapid development of visual foundation models has introduced a paradigm improvement in the field, promising to overcome the generalization limitations of traditional methods. Visual foundation models, with their vast parameter space and sophisticated architectures, are capable of capturing intricate patterns and relationships within data, leading to improved detection and segmentation performance. They offer the advantage of better generalization to a wide variety of scenarios and can adapt more effectively to the unique challenges posed by substation equipment inspection.

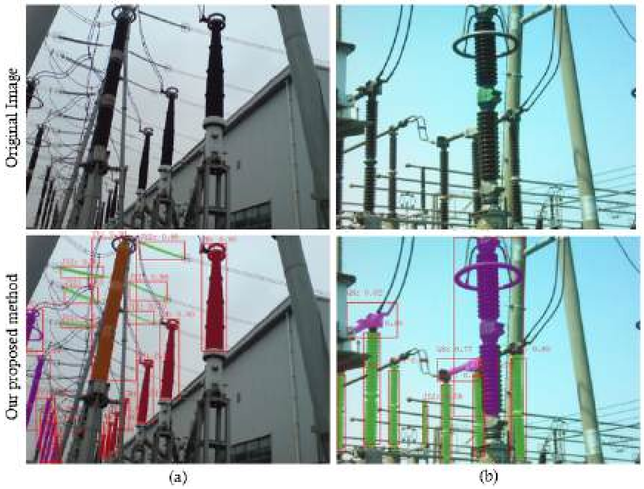

Fig. 1. Examples of the substation equipment instance segmentation inference result. The top images come from our substation equipment dataset, and the bottom images are visualizations of inference using our proposed method.

In this paper, we propose a novel approach that enhances traditional detection models with the powerful capabilities of the visual foundation model. Our method leverages the YOLOX[4] model for generating bounding box prompts, which are utilized by the visual foundation model, specifically the SAM[8], to achieve the instance segmentation results shown in Fig.1. This combination has proven to yield remarkable results in terms of detection and segmentation accuracy, thereby enhancing the overall performance of unmanned inspection systems.

In summary, our contributions to this work are shown as follows:

- We introduce a comprehensive dataset of substation equipment, annotated with polygon masks and class labels, which captures the diversity and complexity of real-world substation environments.
- We present a method pipeline that effectively integrates YOLOX[4] for bounding box generation with the SlimSAM[2] model for instance segmentation, demonstrating a data-efficient approach that requires significantly less training data while maintaining high accuracy.
- We conduct extensive experiments to validate the effectiveness of our proposed method, showcasing its superiority over traditional detection models and highlighting its potential for practical applications in the field of

[*] Co-corresponding authors.
[1] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China
[2] State Grid East China Electric Power Test and Research Institute, Shanghai 200437, China
[3] School of Computer Science, Fudan University, Shanghai, China

979-8-3315-0605-6/24/$31.00 ©2024 IEEE

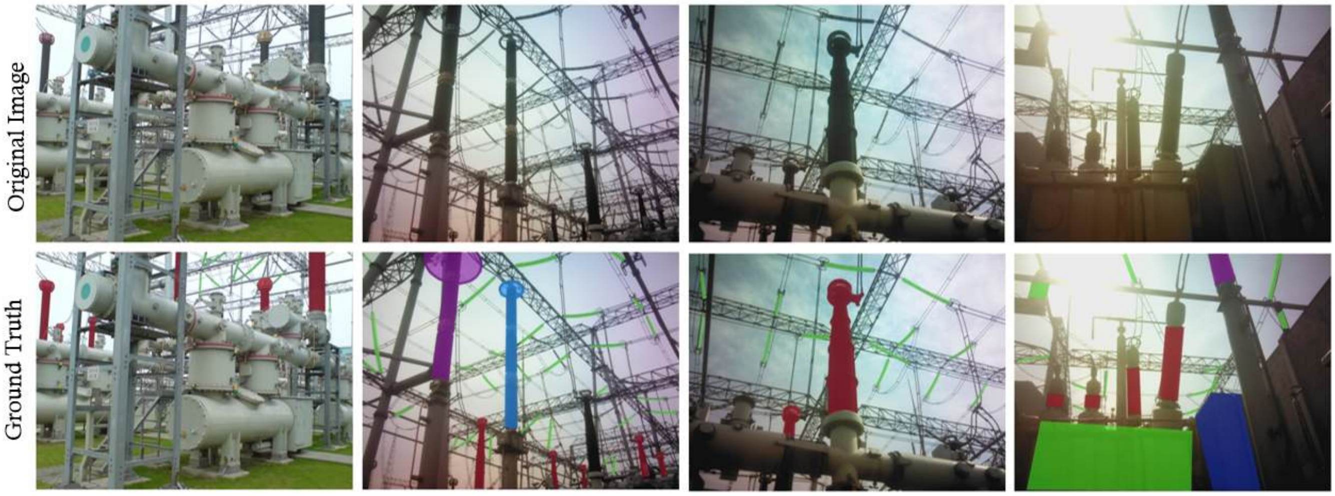

Fig. 2. Examples of the substation equipment dataset. Images were taken from different perspectives and under varying lighting conditions to simulate the real-world application scene.

TABLE I
THE NAME, ABBREV, AND THE AMOUNT IN THE TRAINING SET AND VALIDATION SET OF EACH CCATEGORY

Categories	Abbrev	Train set	Val set
Insulator	JYZ	33683	243
Circuit breaker	QF	1788	8
Current transformer	TA	1961	14
Arrester	FV	2795	30
Disconnector	QS	3460	31
Radiator	SQR	1151	4
Bushing	CM	8457	53
Transformer	T	900	3
Voltage transformer	TV	1895	15
Inductance	L	535	7
Capacitor	C	536	3

substation equipment inspection.

II. DATASET AND PREPROCESSING

The dataset of substation equipment consists of real scene data captured by the State Grid of China, containing 7520 images from different electric power substations in China. The captured scenes encompass various power equipment that require special attention during inspection tasks, such as insulators, circuit breakers, and so on. For detection and instance segmentation tasks, the images were annotated using a set of polygon masks with class labels. In our dataset, 11 different categories of substation equipment are specified. The category name, abbreviation, and the amount of substation equipment in each category in the training set and validation set are shown in Table I.

It can be seen that there is an obvious gap in the amounts of instances in different categories, which is a common and evident challenge in practical detection and segmentation tasks. What's more, to simulate inspection tasks in real-world scenarios, the images encompass substation equipment photographs captured from different perspectives and under varying lighting conditions. A set of samples with images and classified segmentation masks of our dataset are shown in Fig. 2.

For the detection and instance segmentation tasks, label files in the dataset were preprocessed into the COCO format[10] that is widely used. In detail, the segmentation masks in polygon format with semantic labels were first converted into binary masks. Then binary masks were encoded into the COCO RLE format with their provided API, at the same time, the bounding box and mask areas were calculated and saved. This format is used to train both the detection model and the segmentation visual foundation model.

III. METHOD

In this section, we separately introduce the detection model to generate the bounding box prompts for the visual foundation model and the compressed large visual foundation model for segmentation which can be fitted on resource-constrained devices. Our method pipeline is shown in Fig.3.

A. Bounding Box Prompts Generation

In this paper, we use YOLOX[4] as our powerful bounding boxes prompts generator for substation equipment. YOLOX[4] is a single-stage object detector based on YOLOv5[6], which has a light design and achieves high performance in both academic benchmarks and industrial applications. It uses a lightning convolution-based backbone and FPN[9] to extract multi-scale image feature maps with different sizes and channels. Then these multi-scale image feature maps are sent into single-stage detect heads, which have a decoupled design. Each decoupled head has three independent convolutional branches, including Cls. branch, Reg. branch, and IoU. aware branch. The entire network model is shown in Fig.4. Specifically, it uses the anchor-free design, the model directly predicts four values for each object instance: two offsets (horizontal and vertical) from the top-left corner of the grid cell, and the height and width of the predicted bounding box. Similar to anchor-based detectors, a

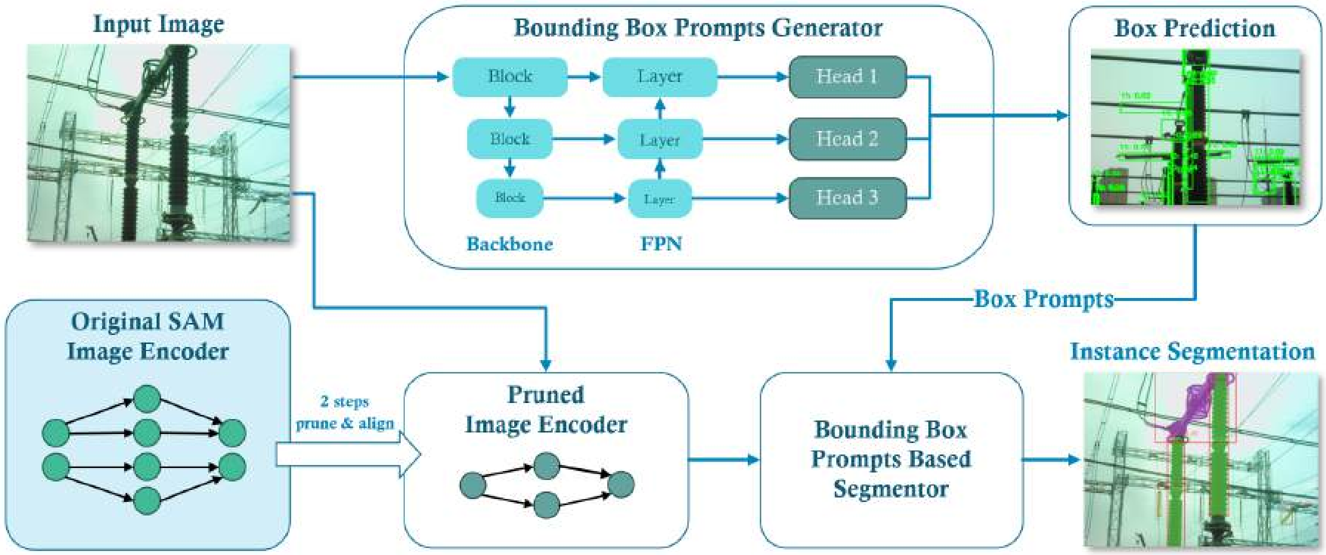

Fig. 3. The pipeline of our proposed method. Substation equipment images would flow into the bounding box prompts generator at the first. After the SAM[8] are pruned and aligned within 2 steps, the same image and box prompts would be processed by bounding box prompts based segmentor to infer the substation equipment instance segmentation results.

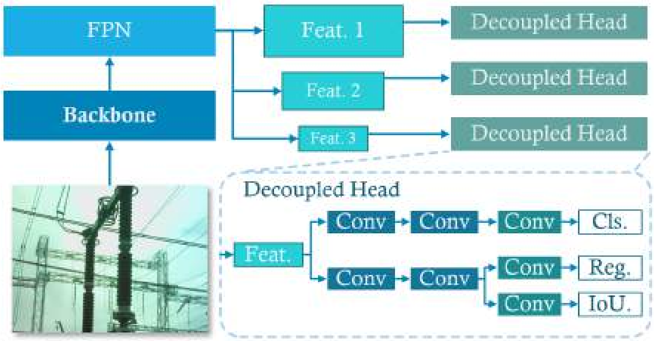

Fig. 4. The structure of YOLOX[4]. Each feat at a different scale connects with a decoupled head predicting bounding boxes.

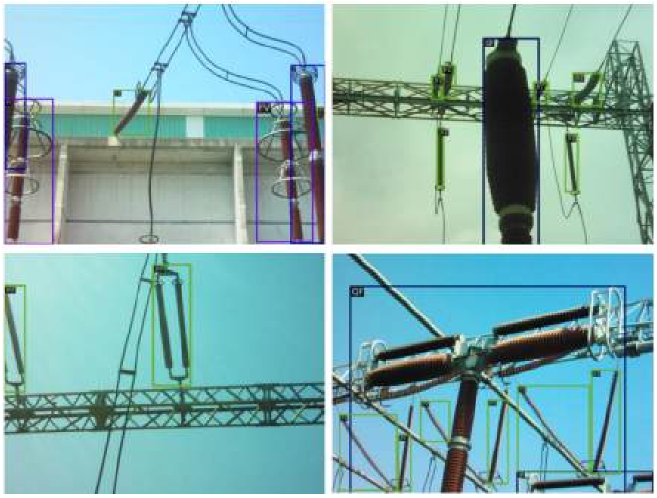

Fig. 5. An example of the accurate substation equipment prompts, inferred by YOLOX[4].

scale range of grid cells is predefined to designate the feature pyramid network (FPN)[9] level for each object. This helps determine the appropriate level of granularity for detecting objects of different sizes. Finally, composing the predicted results of multi-scale level, we could use NMS[12] to figure out final bounding box prompts.

B. Bounding Box Prompts Based Segmentation

We choose the prompt-based segmentation model SAM[8] for next-step substation equipment detection. For accurate mask generation, we use the substation equipment bounding box coming from the last step, to be substation equipment segmentation prompts. Bounding box prompts provide precise spatial information about the location and extent of objects within an image. By using these detection boxes as prompts, SAM[8] can focus mask prediction efforts specifically within the regions of interest corresponding to detected objects. This enables more accurate and refined mask predictions by limiting the search space to relevant areas, reducing computational overhead, and improving segmentation quality. Utilizing 2D detection boxes as prompts in SAM[8] can streamline the segmentation process by reducing the need for exhaustive pixel-wise computations across the entire image. Focusing computational resources on regions specified by detection boxes can improve efficiency without sacrificing accuracy. The substation equipment bounding box prompts are shown in Fig.5.

As a visual foundation model, SAM[8] was trained by a published dataset. We could not use SAM[8] directly to detect precise substation equipment. As Fig.6 shows, the original SAM[8] could not realize true composition in substation equipment. Because of the lack of substation equipment in the published dataset, we should introduce

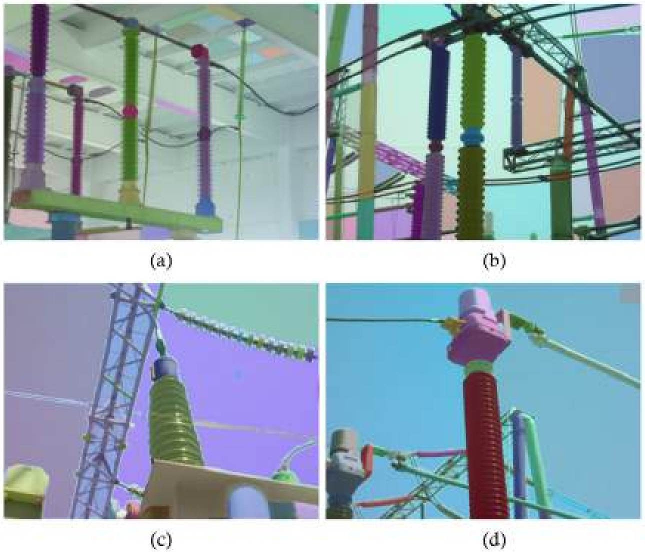

Fig. 6. An example of the Segment Anything model inferencing substation equipment. The original SAM[8] model could not realize the true composition of substation equipment. The insulator, bushing, and current transformer, located in (a) and (b), (c), and (d) respectively, are incorrectly identified into sub-compositions.

substation equipment segmentation knowledge for SAM[8].

In this paper, we employ the SlimSAM[2] method primarily serving as a data-efficient compression method for the Segment Anything Model (SAM[8]). It achieves superior performance with significantly less training data by introducing an alternate slimming framework and disturbed Taylor pruning to effectively compress the model through alternate pruning and distillation of distinct sub-structures while addressing misalignment between pruning objectives and training targets. This approach enhances knowledge inheritance and minimizes divergence from the original model, resulting in substantial performance improvements while reducing training data requirements.

IV. EXPERIMENTS

A. Implementation Details

1) Detection Model: In our experiment, we utilized the mmdetection[1] library for the detection task with the following configurations and settings. We employed the YOLOX[4] model as the chosen detection model. The backbone network was configured as CSPDarknet, while the neck structure consisted of an FPN. The bounding box head was implemented according to YOLOX[4]. Each component of the model employed normalization configurations and activation functions. The loss function consisted of cross-entropy loss, IoU loss, and L1 loss, with different weights assigned to each. Non-maximum suppression (NMS) was applied for post-processing during testing.

Specifically, input data is preprocessed before being fed into the model for training or inference. We randomly resize images within the specified range (360 to 800 pixels) while ensuring that the resulting size is divisible by 32, which is a requirement for YOLOX[4] to downsample correctly, and the augmentation will be applied every 10 iterations. This introduces variations in scale, which can improve the model's robustness and generalization performance. We employ the CSPDarknet as the backbone network, which could be easily controlled by the deepen and widen factor to determine the depth and width of the network, respectively. We specify the kernel sizes parameter by (5, 9, 13), which is used in the Spatial Pyramid Pooling (SPP)[5] module and helps capture multi-scale information from feature maps. 2nd, 3rd, and 4th feature maps from the backbone will be used as input to subsequent network modules and their output channels are 128, 256, and 512. We employ PAFPN[11] to introduce a more effective feature fusion mechanism, enabling better contextual understanding and object localization across different scales in images, and the output channel is 128. We employ Swish[13] as activate functions in our network. We define the radius 0.25 around the center of the anchor box within which a ground truth box is considered as a positive match.

To improve the model's ability to handle various object appearances, sizes, orientations, and lighting conditions, leading to more robust and accurate predictions, data augmentation techniques were employed to enhance the diversity and robustness of the training data. First, we employed the Mosaic technique, which combines four images into a single mosaic image, providing contextual information and increasing the complexity of the training data. Next, we applied random affine transformations to each mosaic image including random rotations, translations, scaling, and flips. To further enhance the diversity of the training data, we performed MixUp, which blends pairs of mosaic images.

The detection model was trained for 300 epochs with batch size 8 on a single NVIDIA GeForce RTX 4090.

2) Segment Anything Model: In our experiment, committing to better substation equipment segmentation performance, we set the pruning ratio of SlimSAM by 50%. This pruning ratio is particularly suitable for applications like substation equipment segmentation, where high accuracy is required, but data quantity is limited. Specifically, it is based on the SAM-B architecture, which is a balanced version of the original SAM[8] model in terms of the number of parameters and computational efficiency. The model consists of a series of Vision Transformer (ViT)[3] blocks and MLP blocks, which are used for feature extraction from the input images.

The model undergoes an initial pruning phase where parameters are removed based on their importance scores. This is done using a method called disturbed Taylor importance, which aligns the pruning criteria with the optimization objectives of subsequent distillation. The pruning process is focused on the image encoder part of the model, while the original prompt encoder and mask decoder from SAM[8] are retained.

After pruning, the model is refined through a distillation process that aims to recover the performance lost due to pruning. The distillation involves aligning the pruned model

TABLE II
COMPARISON OF DIFFERENT DETECTION MODEL'S AVERAGE PRECISION IN VARIOUS OBJECT SIZE

Object Size	Model	Results of Each Category										
		JYZ	QF	TA	FV	QS	CM	T	TV	L	C	mean
Small	YOLOX-S	0.497	-	0.57	0.377	**0.488**	0.04	-	0.553	0.0	-	0.421
	YOLOX-X	**0.535**	-	**0.642**	**0.453**	0.466	**0.236**	-	**0.751**	-	-	**0.514**
Medium	YOLOX-S	**0.678**	**0.51**	**0.894**	0.698	**0.484**	0.578	0.404	0.49	0.0	0.6	0.534
	YOLOX-X	0.668	0.45	0.835	**0.726**	0.469	**0.639**	**0.601**	**0.555**	0.0	**0.8**	**0.574**
Large	YOLOX-S	**0.815**	0.707	0.725	0.876	0.475	0.764	0.6	0.713	0.838	0.767	0.728
	YOLOX-X	0.762	**0.743**	**0.894**	**0.912**	**0.611**	**0.831**	**0.8**	**0.752**	**0.86**	**0.883**	**0.805**
All	YOLOX-S	**0.631**	0.671	0.699	0.75	**0.43**	0.639	0.432	0.525	**0.716**	0.711	0.620
	YOLOX-X	0.629	**0.721**	**0.786**	**0.782**	0.418	**0.711**	**0.624**	**0.623**	0.712	**0.855**	**0.686**

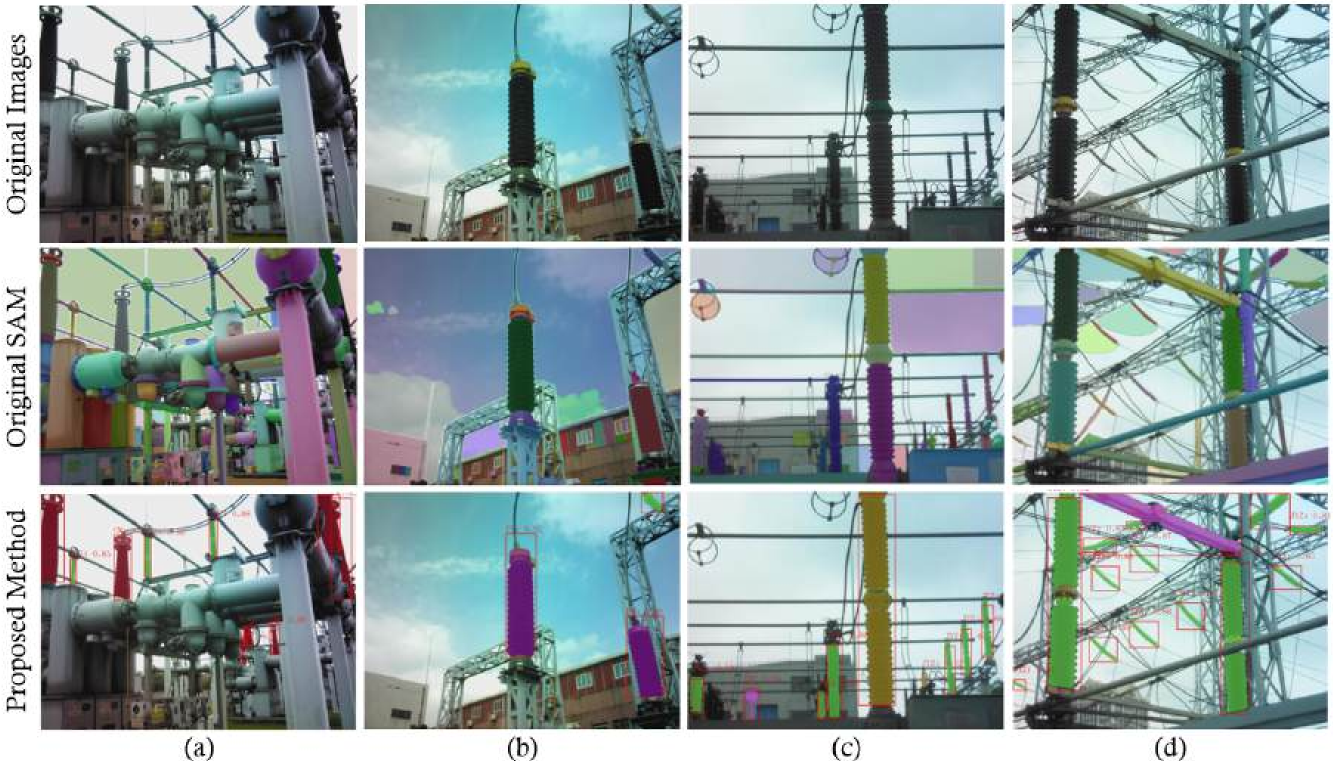

Fig. 7. Some examples of the instance segmentation results of the origin SAM[8] and our proposed method, are shown on the second line and third line respectively. The original images are shown on the top line. Several examples show that our proposed method can better segment the substation equipment in complex scenes.

with the original model at both the embedding and bottleneck levels. We train SlimSAM using the SAM-B model, which uses the ADAM[7] optimization algorithm with a batch size of 4. The total training duration is 20 epochs, with the learning rate initialized at $1e^{-4}$ and reduced by half if validation performance does not improve for 4 consecutive epochs. The model is trained on only less than 0.1% (7467 images) of the substation equipment dataset, demonstrating its data efficiency. The entire compression and training process of SlimSAM is completed on a single Nvidia RTX 4090 GPU.

B. Evaluation Comparisons

1) Different detection model parameters: Firstly, we use different CSPDarknet[14] parameters to train YOLOX[4], including YOLOX-s and YOLOX-x. For YOLOX-s, we control the deepen factor of the CSPDarknet parameter by 0.33, which can affect the capacity and representational power of the backbone network, and it would be multiplied by 3 then rounded to determine the number of the Bottleneck block within each CSPLayer. We control the width of the CSPDarknet parameter by 0.5, which can influence the number of channels in the network's convolutional layers, potentially affecting its ability to capture and represent features at different levels of abstraction, and it would be multiplied by 64 to set the output dimension of the first convolutional layer in the CSPDarknet.

For YOLOX-x, the network's deepen factor is increased to 1.33 compared to the usual model. This implies that the number of layers in the network is multiplied. A higher deepen factor typically allows the network to learn more complex representations, which can be beneficial for capturing finer details in the substation equipment data. The network's widen factor is increased to 1.25, which refers to the number of channels or filters in the convolutional layers, is increased. A wider network can process more information in parallel, which can improve the model's ability to extract a richer set of features from the input data. This can lead to better performance, especially when dealing with high-resolution images or complex scenes of substation equipment. The in channels of PAFPN are increased by 320, 640, and 1280, and out channels are increased by 320. The CSP blocks are controlled by 4, which are designed to enhance feature propagation. The experimental results are shown in TABLE II.

2) Different segmentation prompts: In this work, we experimented with two prompt-based substation equipment instance segmentation approaches, including our proposed method and the grid points-based method. Our proposed method begins with the YOLOX[4] that identifies objects in the image and generates a bounding box for each detected object. The predicted bounding boxes serve as prompts. Subsequently, the SAM[8] uses these bounding boxes as input to focus on predicting the precise segmentation mask within the box.

The grid points-based method divides the image into a regular grid and predicts the instance segmentation mask of the object at each grid point. The SAM[8] utilizes the grid points to produce segmentation mask prompts and then infer the precise mask prediction. Specifically, the grid points-based method also uses the SAM[8] model, while sampling numbers of points along each side of the image as segmentation mask prompts and then predicting final substation equipment instance segmentation. For sampling numbers of points, several parameters constructing a grid of points to sample across the image are defined. We set the sampling points number on one side of the image by 32, which means we sample 1024 points within an image. The grid points are normalized to the image dimensions and are used as prompts for the SAM[8] model to predict instance masks. For each grid point, the SAM[8] model uses it to predict an instance mask. Mask predictions are filtered based on the predicted IoU and stability score to ensure that only high-quality masks are considered. We filter the final instance segmentation mask by the predicted mask IoU threshold of 0.88 and stability score threshold of 0.9. Because of evenly distributed grid point prompts throughout the image, the grid points-based method can capture the object in the whole picture. The instance segmentation results of these methods are shown in Fig.7.

V. CONCLUSIONS

In conclusion, this paper has presented a novel approach for the automation of substation equipment inspection, leveraging the integration of YOLOX[4] and SlimSAM models to achieve high-precision instance segmentation. The proposed method addresses the challenges posed by the complex distribution of substation equipment across various scenarios and limited data quantity conditions, which traditional detection and segmentation algorithms struggle to overcome. The proposed approach not only enhances the reliability and safety of substation equipment inspections but also paves the way for more efficient and accurate unmanned substation equipment inspection systems. Future work could explore further optimizations and integrations with other powerful vision foundation models to push the boundaries of automated substation equipment inspection technologies even further.

REFERENCES

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[2] Zigeng Chen, Gongfan Fang, Xinyin Ma, and Xinchao Wang. 0.1% data makes segment anything slim. *arXiv preprint arXiv:2312.05284*, 2023.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[6] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*, 2022.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[11] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[12] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006.

[13] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[14] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020.