PowerSAM: Edge-Efficient Segment Anything for Power Systems Through Visual Model Distillation

Nannan Yan[a,b], Yuhao Li[c,b], Yingke Mao[b], Xiao Yu[a], Wenhao Guan[d,*], Jiawei Hou[d,*] and Taiping Zeng[a,*]

[a]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China
[b]State Grid Shanghai Municipal Electric Power Company, Shanghai 201204, China
[c]Bowdoin College
[d]School of Computer Science, Fudan University, Shanghai, China

ARTICLE INFO

Keywords:
Semantic segmentation
Power systems
Visual foundation model
Knowledge distillation

ABSTRACT

The inspection of power system equipment is a critical task for ensuring grid reliability and safety which is labor-intensive, costly, and prone to human error, yet the automation process remains challenging due to complex environmental conditions and the edge device computation burden. In this work, we propose a real-time semantic segmentation framework designed for edge computing, leveraging knowledge distillation from large visual foundation models to compact backbones. Our method integrates a bounding box prompt generator with a segmentation model into a unified architecture, significantly reducing computational complexity while maintaining high segmentation accuracy. A two-stage distillation strategy is employed for further optimization of edge device deployment. Extensive evaluations in the Power System dataset demonstrate that our approach outperforms state-of-the-art methods with high efficiency (20.04 FPS on the NVIDIA Jetson Orin NX) and competitive accuracy (19.456 IoU on power system components segmentation), offering a practical solution for real-time equipment monitoring and inspection in power systems. The code will be available at https://github.com/fudan-birlab/PowerSAM.

1. Introduction

The inspection of substation equipment is essential to ensuring the reliability and safety of electrical power grids. Traditionally, this task has been performed manually by trained personnel, a process that is labor-intensive, costly, and prone to human error. Additionally, manual inspections expose workers to potential safety hazards. As such, there is a critical need for more efficient, reliable, and automated inspection methods. Although automation technologies have made significant progress, they still face challenges due to the computation burden for edge device deployment and the variability of substation environments, which limit their effectiveness in real-world applications.

Traditional perception methods perform well on specific datasets but often struggle with adaptability to diverse scenarios. Power system perception, in particular, must address varying environments and conditions. To improve the adaptability of vision perception models, recent advancements in visual foundation models[1, 9, 6, 14, 18] have led to a paradigm shift in various computer vision tasks, including object detection and image segmentation. These models, with their extensive parameter spaces and sophisticated architectures, are capable of capturing intricate patterns and contextual relationships in data, which results in improved performance across a wide range of tasks. Their ability to generalize to diverse scenarios makes them particularly attractive for challenging environments, such as substation equipment inspection. However, the deployment of such models on edge devices for real-time, on-site inspections remains limited. This is primarily due to the high computational demands and large model sizes, which are incompatible with the resource constraints of edge computing devices.

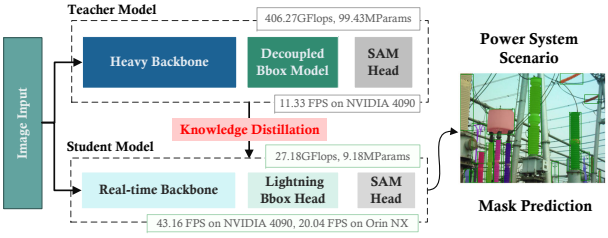

Figure 1: Overview of the PowerSAM Framework for Real-Time Segment Anything in Power System Scenarios.

In this work, we address edge-side power system segmentation by proposing a novel method that combines the strengths of visual foundation models with efficient design strategies tailored for edge devices, achieving environmental adaptability and edge-side practicability. Our approach involves distilling the knowledge from a large, pre-trained visual foundation model (segmentation anything) into a more compact, faster backbone suitable for real-time deployment in power system equipment. We further optimize this process by fine-tuning the mask decoder to ensure that the distilled model retains high segmentation accuracy despite the reduced model size. An overview of our proposed framework is shown in Fig.1. Our method not only preserves the generalization capabilities of large foundation models but also

*Corresponding author
Email addresses: nnyan21@m.fudan.edu.cn (N. Yan); hli3@bowdoin.edu (Y. Li); maoyingke@sh.sgcc.com.cn (Y. Mao); xyu20@fudan.edu.cn (X. Yu); whguan21@fudan.edu.cn (W. Guan); jwhou23@fudan.edu.cn (J. Hou); zengtaiping@fudan.edu.cn (T. Zeng)

reduces the model's footprint, making it suitable for types of equipment with limited computational resources. Through extensive experiments, we demonstrate that our method maintains high segmentation accuracy while achieving real-time performance, making it a viable solution for real-time power system inspections.

The main contributions of this work are as follows:

- We distill knowledge from large-scale visual models into a compact SAM model for real-time edge device performance, ensuring accuracy for power system inspection.
- We design a unified architecture combining a bounding box prompt generator, prompt encoder, and mask decoder, achieving 43.16 FPS for industrial applications.
- We validate the feasibility of deploying our model on NVIDIA Jetson Orin NX, achieving an inference speed of 20.04 FPS, enabling efficient real-time computations in power system environments.

2. Related Work

Visual Foundation Model. Visual foundation models[1, 9, 6, 14, 18] have emerged as a transformative approach in computer vision, offering unparalleled generalization capabilities across diverse tasks. The Segment Anything Model[6] has been a pioneering effort, introducing a prompt-based segmentation framework trained on a vast dataset (SA-1B) of over 1 billion masks. SAM[6] demonstrates remarkable zero-shot generalization to new segmentation tasks, leveraging its combination of a powerful image encoder[15], flexible prompt encoder, and efficient mask decoder[4, 3]. However, the computational demands of SAM[6], with its extensive parameter size, have limited its applicability in real-time and resource-constrained environments. SAM2[9], a successor to SAM[6], extends its capabilities to video segmentation, integrating streaming memory to handle temporal dimensions and refine predictions iteratively. With advancements in segmentation accuracy and efficiency, SAM2 bridges the gap between image and video segmentation, making it suitable for dynamic environments.

Knowledge Distillation. Knowledge distillation has become a critical method for compressing large-scale visual models to enable deployment on resource-constrained edge devices. Techniques like MobileSAM[19] and TinySAM[11] achieve significant reductions in model size by replacing the transformer-based architecture of SAM with lightweight alternatives[12, 16]. Despite their efficiency, these methods often compromise segmentation accuracy due to aggressive model simplifications. SlimSAM[2] introduces an alternate slimming framework and disturbed Taylor pruning, enabling efficient compression with minimal data and achieving near-original performance levels. EdgeSAM[21] further refines this process by incorporating prompt-in-the-loop distillation. These approaches exemplify the trade-offs between computational efficiency and segmentation quality, advancing the feasibility of deploying robust segmentation models on edge devices in real-world scenarios. Together, these efforts underscore the potential to integrate knowledge distillation and visual foundation models for edge-efficient applications in industrial and mobile environments.

3. Proposed Method

In this section, we introduce PowerSAM for real-time segmentation of power systems by distilling knowledge from the Segment Anything Model (SAM), leveraging bounding box prompts. Our approach integrates prompt generation and SAM into a single model, optimizing it for a smaller backbone while retaining accuracy. This enables the model to be deployed in real-time on edge devices like NVIDIA Jetson NX[5], highlighting its efficiency and practicality for on-site power system inspections. Our method pipeline is shown in Fig.2.

3.1. Revisiting Box Prompts Based Segmentation

In our previous work[17], we harnessed the power of both YOLOX[20] and the Segment Anything Model[6] for the segmentation of substation equipment. By feeding images into the high-performing YOLOX object detector[20] and SAM[6]'s image segmentor, we leveraged YOLOX[20] to generate bounding box prompts for substation equipment, which were then used as inputs for SAM[6]'s prompt encoder. This approach allowed SAM[6] to produce high-quality segmentation masks for substation equipment, enabling efficient segmentation of equipment within substation environments.

We have further optimized our method by pruning[2] the models to reduce the data requirements, allowing us to train powerful substation equipment segment anything model with a limited amount of data.

3.2. Power Systems Knowledge Distilling

Even with pruning, the depth of the SAM model remains considerable, leading to increased latency (Table.1). In this section, we focus on transferring the semantically rich visual representations of the visual foundation model into a more compact SAM model, specifically designed for real-time inference within power system scenarios.

We initiate the knowledge distillation process by transferring the visual knowledge of the powerful SAM visual transformer backbone[15], which is an expert in power system visual representations, into a lighter, real-time capable ViT backbone. This distilled model maintains the essential characteristics of the original SAM while reducing its depth and parameter quantity, making it suitable for deployment on edge devices with limited computational resources. This distillation process is shown in Fig.3(a).

Subsequently, we fine-tune the mask decoder to bridge any minor inconsistencies between the feature spaces of the distilled backbone and the original ViT backbone. This fine-tuning process is shown in Fig.3(b). This adaptation ensures that the pre-trained decoder can effectively accommodate the

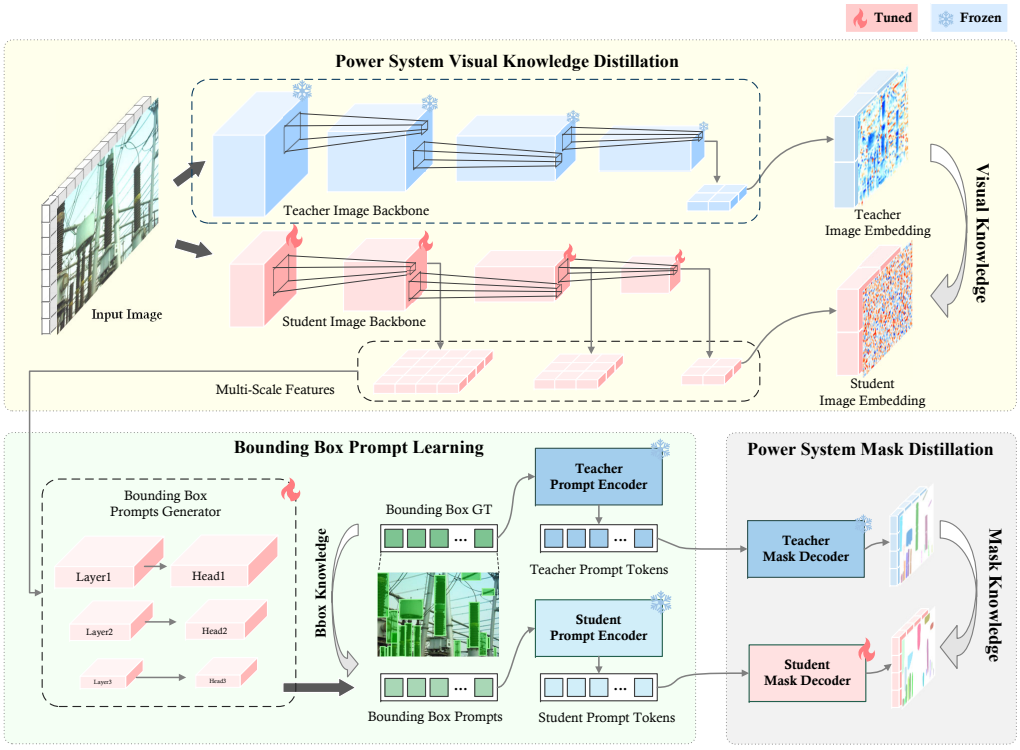

Figure 2: Overview of the real-time segment anything framework PowerSAM for power systems. The top section illustrates a teacher-student knowledge distillation approach, where the teacher model (SAM[6]) distills visual knowledge into a smaller student backbone. The bottom section depicts the integration of a Bounding Box Prompt Generator and a unified prompt encoder to mask decoder architecture, enabling efficient feature utilization and high-performance perception for power system applications.

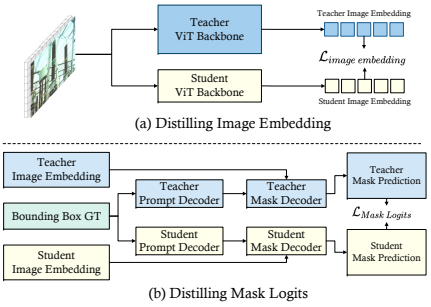

Figure 3: The two distillation stages of the proposed framework for power system segmentation.

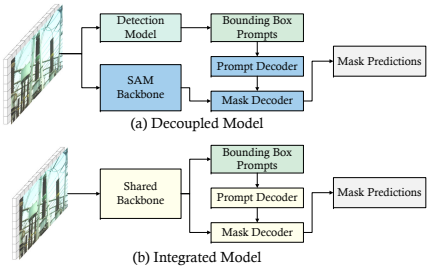

Figure 4: Comparison of decoupled and integrated models for segmentation.

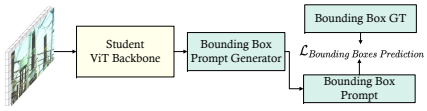

Figure 5: Illustration of the fine-tuning process for the bounding box prompt generator.

subtle differences, thereby enhancing the model's ability to fit the nuances of power system perception scenarios.

Through this process, we have successfully developed a real-time SAM perceiver that is efficient and highly accurate in segmenting substation equipment. This distilled and fine-tuned model compresses the knowledge from a visual foundation model into a smaller model while maintaining the performance necessary for practical applications in power system monitoring and inspection.

3.3. SAM Knowledge Based Prompts Generation

As shown in Fig.4(a), we recognize that YOLOX[20], as a standalone model for generating bounding box prompts, still imposes a significant computational burden on edge devices. We detail the integration of an independent bounding box prompt generator with the SAM[6] into a unified perception model, leveraging the visual semantic knowledge and feature representations of the visual foundation model to its fullest performance, shown in Fig.4(b). As shown in Fig.5, after learning the bounding box prediction by fine-turning the bounding box prompt generator, this integration allows the bounding box prompt generator to achieve perception performance on par with a detector using a carefully

designed backbone while avoiding the redundant time complexity and spatial complexity introduced by the backbone of a standalone bounding box prompt generator.

By consolidating these components into a single model, we have eliminated the need for separate processes and reduced the overall computational overhead. This streamlined approach enables end-to-end real-time segmentation for power systems, ensuring that the model can efficiently perceive and segment substation equipment with minimal latency. The integrated model not only maintains the high accuracy of segmentation but also optimizes the workflow for on-site inspections and monitoring tasks in power system environments. Through this innovation, we have realized a powerful model for practical applications, enhancing the efficiency and reliability of automated substation equipment inspection.

4. Experiments

4.1. Dataset

We utilize a comprehensive dataset of substation equipment that we have curated and annotated[17]. This dataset is designed to capture the diversity and complexity of real-world substation environments, ensuring that our model is tested under conditions that closely mimic actual deployment scenarios. The dataset comprises a total of 7520 images, sourced from various electric power substations across China, and includes a variety of substation equipment that requires meticulous inspection.

Each image in the dataset has been meticulously annotated using polygon masks and class labels, allowing for precise segmentation. The inclusion of 11 different categories of substation equipment ensures that our model must be capable of recognizing and segmenting a broad range of equipment types. To simulate real-world inspection tasks, the images in our dataset are captured from multiple perspectives and under varying lighting conditions. This variability is crucial for training a model that can generalize well across different substation environments and conditions. A subset of these images, along with their corresponding segmentation masks, is depicted in Fig.6.

For the detection and segmentation tasks, the labels in our dataset have been preprocessed into the COCO[7] format, which is widely adopted in the field. This involves converting the polygon format with semantic labels into binary masks, encoding these masks into the COCO[7] RLE format, and calculating and saving the bounding box and mask areas. This standardized format facilitates the training of both the bounding box supervision and the segmentation supervision, ensuring that our pipeline is robust and adaptable to industry standards.

4.2. Implementation Details

4.2.1. Distilled Segment Anything Model

For this work, our student model is the lightweight backbone RepViT[13]. We utilize a checkpoint from the RepViT[13] family as the pre-trained weights for our student model, leveraging the knowledge of the visual foundation model from the RepViT[13] architecture. The images are resized to 1024 pixels.

For the distillation process, we set the image embedding to 256. We also fuse features into one feature, and apply pixel-wise feature distillation with a loss weight of 1.0. The training is scheduled for 100 epochs with a warm-up period of 5 epochs. The base learning rate is set to 1.0×10^{-1}, and the warm-up learning rate is 5.0×10^{-7}. The learning rate scheduler follows a cosine annealing pattern, with a decay rate of 0.1 applied every 30 epochs. The weight decay is configured to 0.01, and the gradient norm clipping is set to 5.0. For the distilling process of the mask decoder, we load the pre-trained segment anything model and continue training for an additional 50 epochs with a reduced base learning rate of 3.2×10^{-3}. We apply a combination of distillation losses on the mask decoder, including Binary Cross Entropy loss with a weight of 5.0 and Dice loss with a weight of 5.0. We allow up to 16 bounding box prompts per image. We set the multi-mask output to 4, allowing the model to generate multiple masks for each prompt.

4.2.2. Bounding Box Prompt Generator Fine-Tuning

We employ the YOLOX-S[20] model as our bounding box prompt generator with a shared RepViT[13] backbone, which is also used in the Distilled SAM. The shared RepViT[13] backbone is frozen during the bounding box prompt generator fine-tuning process to maintain the learned features. This setup allows the detection head to adapt to the specific characteristics of our substation equipment dataset while leveraging the knowledge distilled from the shared backbone. The fine-tuning process is shown in Fig.5. The data pre-process within the Bounding Box Prompt Generator is configured to normalize the input images and the pad size is set to be divisible by 32. The data augmentations employ random scaling and affine transformations, enhancing the model's robustness to different image sizes and orientations.

The model's input resolution is adjusted to 1024×1024. The model is trained for 300 epochs. The optimizer uses SGD with a base learning rate of 0.01, momentum of 0.9, and weight decay of 5.0×10^{-4}. The learning rate schedule includes a quadratic warm-up for the first 5 epochs, followed by a cosine annealing schedule until 15 epochs before the end, where a constant learning rate is maintained. During fine-tuning, we focus on optimizing the bounding box prompt generator to generate accurate bounding box prompts for our substation equipment instances.

4.3. Evaluation Comparisons

4.3.1. Performance of Power System Segmentation

We evaluate our proposed method against state-of-the-art models on the Power System dataset. Our method leverages the output of the YOLOX[20] Head to generate bounding boxes, which serve as prompts for our SAM[6] head. We employ mIoU (mean Intersection over Union) as the metric for evaluating mask accuracy.

Using a two-stage distillation process, we transferred visual knowledge from a large-scale SAM[6] model into the

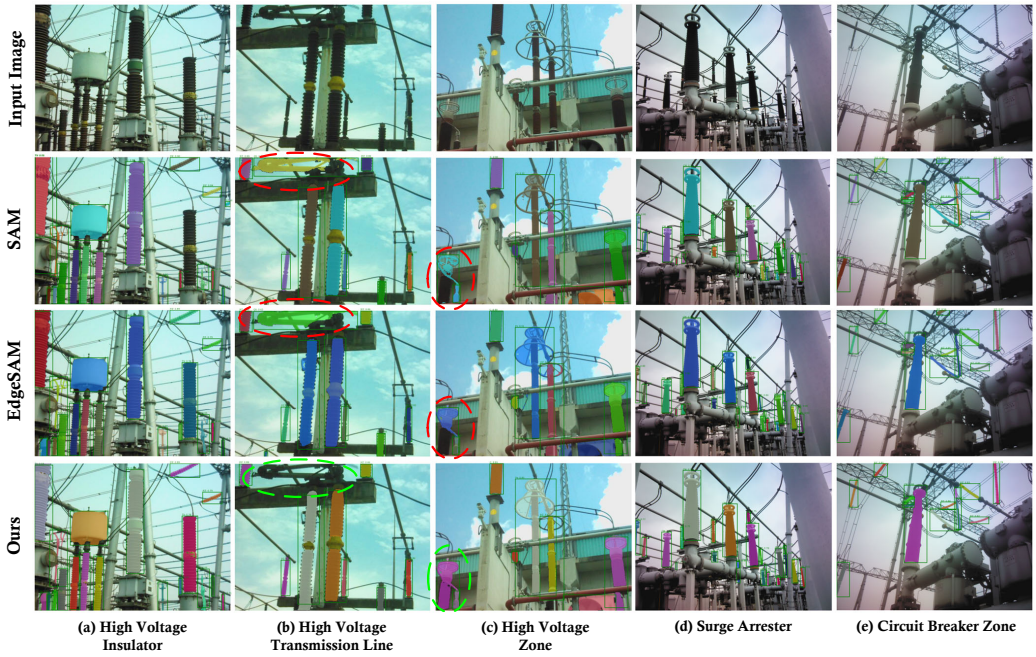

Figure 6: Some examples of segmentation results are shown with the results of SAMs in the second and third lines and our proposed method in the fourth line. The original images are shown on the top line. Several examples show that our proposed method can better segment the substation equipment in complex scenes.

Table 1
Comparison of mIoU (%) and FPS(N_{frames}/s) on the power system semantic segmentation for different models with box and point prompts. In this table, we present the ranking of key performance metrics: the best-performing metric is highlighted in **bold**, the second-best is **bold and underlined**, and the third-best is *italicized and underlined*. PowerSAM distills the power system visual knowledge of visual foundation model features to smaller backbones and combines the bounding box prompt generator and SAM model into an integrated model, achieving a significant enhancement in computational efficiency while keeping the segmentation precision.

Model	Backbone	Box Prompt			Point Prompt			FPS↑
		Box	+ 1 Pt.	+ 2 Pt.	Point	+ 1 Pt.	+ 2 Pt.	
SAM-B[6]	ViT-Base[15]	**19.578**	*40.678*	**49.533**	58.006	64.188	66.512	11.33
SAM-H[6]	ViT-Huge[15]	20.520	35.983	44.020	**59.668**	**66.249**	**69.059**	3.79
SAM2-L[9]	Hiera-Large[10]	**34.341**	**40.870**	44.431	**63.696**	**67.605**	**71.031**	8.87
Pruned SAM[2, 17]	Pruned ViT-Base[15, 2]	19.061	40.232	47.345	57.379	64.098	66.365	13.83
EdgeSAM[21]	RepViT-M1[13]	19.368	35.745	46.468	56.143	63.308	*68.256*	*30.11*
PowerSAM-B	RepViT-M1[13]	18.945	**41.012**	**50.318**	57.959	64.258	67.626	**41.54**
PowerSAM-S	RepViT-M0[13]	*19.456*	40.323	*48.702*	*58.292*	*64.616*	68.096	**43.16**

efficient RepViT-M0.6[13] backbone. In the first stage, we distilled image embeddings, and in the second stage, the backbone and prompt encoder were frozen while distilling SAM's lightweight mask decoder. The YOLOX[20] head was fine-tuned using bounding box annotations from the dataset. This approach ensured efficient knowledge transfer while maintaining accuracy. As shown in Table.1, our method achieves high computational efficiency and segmentation precision, comparable to larger models like SAM with ViT-Base[15], ViT-Huge[15], and SAM2[9] with Hiera-Large[10] backbones and multi-scale design.

As shown in Table.1, our proposed method achieves an optimal balance between accuracy and efficiency, making it a robust solution for real-time segmentation in power system environments. By distilling the capabilities of large visual foundation models into a compact ViT[13] backbone, we enable high-performance segmentation suitable for deployment on edge devices, addressing the practical demands of power system inspection tasks.

4.3.2. Comparison of Decoupled and Integrated Model

To further validate the efficiency of our proposed approach, we compared the Decoupled Model and Integrated Model configurations in terms of computational complexities, including GFLOPs and MParams, and inference speed (ms). The Decoupled Model separates the SAM[6] and

Table 2
Comparison of GFLOPs, MParams, and Latency (ms) between knowledge-separated Decoupled (D) Models and knowledge-shared Integrated (I) Models with various backbones. † indicates that the Prompt Encoder and Mask Decoder are designed based on SAM2[9] architecture and initialized with its pre-trained weights. Decoupled models use separate modules for SAM[6] and bounding box prompt generation[20], while knowledge-shared Integrated models consolidate them into a single architecture for efficiency.

Backbone	Structure	GFLOPs↓	MParams↓	Latency (ms)↓
ViT-B[15]	D	406.27	99.43	88.22
ViT-H[15]	D	2770.90	644.58	263.39
Hiera-L †[10]	D	830.36	225.84	112.61
Pruned ViT-B[2, 17]	D	132.36	35.33	72.32
RepViT-M1[13]	D	57.93	18.52	33.21
RepViT-M0[13]	D	48.60	15.94	30.37
RepViT-M1[13]	I	36.51	11.76	24.06
RepViT-M0[13]	I	**27.18**	**9.18**	**23.16**

the Bounding Box Prompt Generator into two independent components, while the Integrated Model combines them, significantly reducing computational overhead without compromising prediction accuracy.

In the Decoupled Model, SAM uses the RepViT[13] backbone for image embedding and mask prediction, with YOLOX-S[20] for bounding box generation, operating at a resolution of 640×480. In contrast, the Integrated Model combines the Mask Decoder and Bounding Box Head into a unified backbone with a higher resolution of 1024×1024, eliminating redundancies and improving efficiency. As shown in Table.2, the Integrated Model significantly reduces computational complexity. For instance, with the RepViT-M0[13] backbone, GFLOPs dropped from 48.60 to 27.18 and MParams from 15.94 to 9.18, while latency improved from 30.37 ms to 23.16 ms, enabling real-time edge device applications.

As shown in Table.2, by integrating the bounding box generator and mask decoder, the results illustrate that our Integrated Model is both faster and more computationally efficient than the Decoupled Model. This efficiency makes it ideal for real-time edge device computations in power system scenarios, significantly reducing latency and hardware resource requirements while maintaining robust segmentation performance.

4.3.3. Performance of Bounding Box Prompt Generator

The performance of the Bounding Box Prompt Generator is a critical component of our method, enabling efficient and accurate bounding box generation tailored for power system scenarios. This module directly accepts multi-scale image features from the ViT[15] backbone, with varying resolutions and feature dimensions, to accommodate the diverse sizes and shapes of substation equipment.

Our Bounding Box Prompt Generator follows the configuration of the YOLOX[20] box head while integrating PAFPN[8] for feature fusion. The PAFPN[8] processes the multi-scale outputs of the ViT[15] backbone, enhancing compatibility with the unique demands of power system applications. For the RepViT-M0[13] backbone, PAFPN[8] takes in feature maps with dimensions $(80, 160, 320)$, representing three different resolutions. These multi-scale features ensure the model effectively captures both large receptive fields and fine-grained details in the input images, then PAFPN[8] outputs fused features with a unified dimension of 128, which are directly fed into the multi-scale bounding box head. This unified representation improves the generator's ability to detect and localize equipment of varying sizes, from small components to large transformers.

Table 3
Comparison of bounding box performance on the Power System dataset between YOLOX-S[20] and our Bounding Box Prompt Generator (BBPG).

Metric	YOLOX-S[20]	BBPG (Ours)
AP (Overall) ↑	0.620	0.501
AP@0.5 ↑	0.858	0.716
AP@0.75 ↑	0.721	0.565
AR (Overall) ↑	0.695	0.533
AR (Small) ↑	0.421	0.345
AR (Medium) ↑	0.534	0.348
AR (Large) ↑	0.728	0.662
GFLOPs ↓	34.28	24.20
MParams ↓	8.94	5.12

As shown in Table.3, the evaluation on the Power System dataset demonstrated that the generator achieves 24.20 GFLOPs and 5.12 MParams, significantly lower than YOLOX-S[20]'s 34.28 GFLOPs and 8.94 MParams. Despite the reduction, its AR score of 0.662 for larger objects is competitive with YOLOX-S[20], demonstrating its ability to balance performance and efficiency by leveraging visual foundation model knowledge effectively.

4.4. Segmentation On Edge Computation

We transfer our consolidated model to edge devices, which integrates bounding box prompt generation and SAM-based semantic segmentation. We use the Jetson Orin NX[5] 16GB for its robust AI performance, which includes an Ampere GPU, Arm Cortex-A78AE 64bit CPU, and 16 GB of

LPDDR5 memory. This hardware provides up to 100 Sparse INT8 TOPs and 50 Dense INT8 TOPs, making it an ideal choice for running complex neural networks like our visual foundation model in real-time on edge devices. And the TensorRT SDK plays a pivotal role in our deployment pipeline by optimizing our model for the Jetson Orin NX platform. It accelerates deep learning inference through advanced optimizations. We utilized TensorRT's API to integrate our model, enabling it to run efficiently on the Jetson Orin NX's GPU architecture. Our model achieves 49.88 ms per frame on Jetson Orin NX. This enables the model to monitor and inspect power systems in real-time, significantly improving the efficiency and safety of substation operations.

5. Conclusions

This work presents an efficient real-time segmentation framework for power systems by distilling knowledge from large-scale visual foundation models into a lightweight architecture tailored for edge computation. By integrating a bounding box prompt generator with segmentation tasks, we achieved significant reductions in computational complexity while maintaining competitive accuracy. Extensive experiments demonstrated our method's effectiveness, achieving over 20.04 FPS on edge devices like NVIDIA Jetson Orin NX[5]. This approach provides a robust and practical solution for real-time monitoring and inspection in resource-constrained environments, setting a benchmark for deploying advanced vision models in industrial applications.

Acknowledgements

This work was supported by the research program of the State Grid Shanghai Municipal Electric Power Company (Grant no.52095023000D) and the Fudan startup funding (T. Zeng).

References

[1] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660.

[2] Chen, Z., Fang, G., Ma, X., Wang, X., 2023. 0.1% data makes segment anything slim. arXiv preprint arXiv:2312.05284 .

[3] Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1290–1299.

[4] Cheng, B., Schwing, A., Kirillov, A., 2021. Per-pixel classification is not all you need for semantic segmentation. Advances in neural information processing systems 34, 17864–17875.

[5] Karumbunathan, L.S., 2022. Nvidia jetson agx orin series, a giant leap forward for robotics and edge ai applications, technical brief. Online at https://www. nvidia. com/content/dam/en-zz/Solutions/gtcf21/jetson-orin/nvidia-jetson-agx-orin-technical-brief. pdf .

[6] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026.

[7] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer. pp. 740–755.

[8] Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path aggregation network for instance segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8759–8768.

[9] Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al., 2024. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 .

[10] Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al., 2023. Hiera: A hierarchical vision transformer without the bells-and-whistles, in: International Conference on Machine Learning, PMLR. pp. 29441–29454.

[11] Shu, H., Li, W., Tang, Y., Zhang, Y., Chen, Y., Li, H., Wang, Y., Chen, X., 2023. Tinysam: Pushing the envelope for efficient segment anything model. arXiv preprint arXiv:2312.13789 .

[12] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention, in: International conference on machine learning, PMLR. pp. 10347–10357.

[13] Wang, A., Chen, H., Lin, Z., Han, J., Ding, G., 2024. Repvit: Revisiting mobile cnn from vit perspective, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15909–15920.

[14] Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T., 2023. Seggpt: Towards segmenting everything in context, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1130–1140.

[15] Wang, Y., Huang, R., Song, S., Huang, Z., Huang, G., 2021. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. Advances in neural information processing systems 34, 11960–11973.

[16] Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L., 2022. Tinyvit: Fast pretraining distillation for small vision transformers, in: European conference on computer vision, Springer. pp. 68–85.

[17] Yan, N., Guan, W., Yu, X., Hou, J., Zeng, T., 2024. A visual foundation model of image segmentation for power systems, in: 2024 IEEE 14th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), IEEE. pp. 569–574.

[18] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything: Unleashing the power of large-scale unlabeled data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10371–10381.

[19] Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S., 2023. Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289 .

[20] Zheng, G., Songtao, L., Feng, W., Zeming, L., Jian, S., et al., 2021. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 .

[21] Zhou, C., Li, X., Loy, C.C., Dai, B., 2023. Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. arXiv preprint arXiv:2312.06660 .