processes

MDPI

Article

An Improved Mask2Former-HRNet Method for Insulator Defect Detection

Yaoran Huo [1,*,†], Lan Xiao [1], Zhenyu Tang [1], Jian Zhou [2], Xu Dai [1], Yuhao Xiao [1] and Xia Fang [3]

[1] Information and Communication Company, State Grid Sichuan Electric Power Company, Chengdu 610041, China
[2] State Grid Sichuan Electric Power Company, Chengdu 610041, China
[3] West China Hospital, Sichuan University, Chengdu 610041, China; fangxia@scu.edu.cn
* Correspondence: hyrchina2012@gmail.com; Tel.: +86-15198280794
† Current address: No.16, Jinhui West Second Street, Wuhou District, Chengdu 610041, China.

Abstract: To solve the problem of scale variation in insulator images captured by drones, caused by the lack of control over angle and distance, which makes it hard to detect subtle defects, this paper proposes an instance segmentation method based on an improved Mask2Former-HRNet model for precise localization and defect detection of transmission line insulators. First, a mask-guided and matching component is added to Mask2Former to reduce the misjudgment rate of insulator defects by including noisy label masks. Second, the HRNet backbone network is used to better capture the spatial and shape information of insulators, as it has a stronger feature transfer ability. Deformable convolutions are introduced to handle deformation issues caused by varying angles in insulator images. Then, an attention mechanism is added to focus on key content, improving the network's attention to crucial information. Finally, experimental results on defect detection of transmission line insulator images captured by drones show that the proposed method increases the detection accuracy by 8.41% and reduces the misjudgment rate by 4.11%. Comparative experiments indicate that the proposed method outperforms existing methods in several evaluation metrics.

Keywords: transmission line inspection; neural network; Mask2former; HRNet; instance segmentation

check for updates

Academic Editor: Yo-Ping Huang

Received: 25 December 2024
Revised: 19 January 2025
Accepted: 23 January 2025
Published: 23 January 2025

Citation: Huo, Y.; Xiao, L.; Tang, Z.; Zhou J.; Xu D.; Xiao Y.; Fang X. An Improved Mask2Former-HRNet Method for Insulator Defect Detection. *Processes* **2025**, *13*, 316. https://doi.org/10.3390/pr13020316

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid growth of drone technology, drone-based imaging has become an important tool for obtaining aerial images. It plays a key role in the inspection of power transmission lines, such as transmission line defect detection [1], insulator defect detection [2], and terminal connection verification in substations [3]. For the power industry, transmission line images are essential for ensuring the stable operation of the power system. Accurate detection and analysis of insulators and their defects are crucial for maintaining the safety and reliability of power systems. However, this task faces several major challenges.

First, the backgrounds of drone-captured images are often complex and diverse, including environments like mountains, farmlands, forests, and deserts. This makes it harder to detect insulators and their defects. Second, drone images may have issues with perspective and distance. Because of the long distance from the insulator, small defects may become blurred or hard to see. Additionally, insulators are often high up on the transmission line, which can cause them to overlap or be blocked by other equipment. This

may lead to partial or full occlusion of insulators in the images, making it impossible to fully see defects. Therefore, detecting insulators and their defects efficiently and accurately is very important, and overcoming these challenges is crucial.

To address these issues, significant research has been carried out worldwide on detecting insulators and their defects. For example, some studies have used special image processing algorithms, such as insulator shape extraction algorithms based on Canny edge detection and generalized Hough transform [4], image segmentation using the Otsu method on the Hue and Saturation components in the HIS color space followed by histogram-based recognition of insulator contours [5], and particle swarm optimization-based ant colony algorithms for detecting insulator contours and counting after median filtering preprocessing [6].

However, traditional image processing methods usually have slow speeds and low accuracy. In contrast, deep learning methods offer faster detection speeds. For example, some studies have used YOLO object detection neural networks [7–10], two-stage object detection models based on region proposal networks and data augmentation [11], and deep convolutional autoencoders [12], combining supervised and unsupervised learning for insulator and defect detection. Some research has also applied Mask R-CNN for wind turbine blade fault diagnosis with good results [13]. However, these methods do not achieve pixel-level localization, which leads to imprecise defect positions and blurry detection areas for insulators and defects in drone-captured images. This limits the effectiveness of insulator and defect detection during transmission line inspections.

To achieve instance segmentation, or pixel-level localization of insulators and their defects, this paper proposes an improved Mask2Former network [14] to address the issue of blurred minor defects in insulators. The backbone network is replaced with an improved HRNet [15] to solve the problem of large-scale variation in insulator images caused by differing perspectives and distances. This method is used for precise localization of the insulators and their defects in transmission lines.

In summary, this paper presents an instance segmentation method based on the improved Mask2Former and HRNet models for detecting and distinguishing insulators and their defects in transmission lines. The proposed method aims to improve detection speed, accuracy, and localization precision. Its application will support the development of adaptive inspection and precise defect localization for transmission line components using drones, enhancing the safety and reliability of power systems.

2. Insulator Defect Detection Based on an Improved Mask2Former-HRNet Method

2.1. Improvement of Mask2Former

Mask2Former is a deep learning method designed for instance segmentation and semantic segmentation tasks. The core idea is to treat image segmentation as a sequence-to-sequence problem, using a transformer model to improve performance and better adapt to practical applications [16–18]. Specifically, Mask2Former uses a framework called Masked Attention Masked Autoencoder (MAMAE) for segmentation.

As shown in Figure 1, in the improved version of Mask2Former, the model still uses a backbone network, pixel decoder, and transformer decoder architecture. The main improvement is in the transformer decoder.

In the improved network, the decoder queries are divided into two parts: the mask-guided and matching components. The mask-guided part takes the label mask image and class embeddings as inputs and directly assigns the predicted result to the corresponding label mask, without requiring a second matching step.

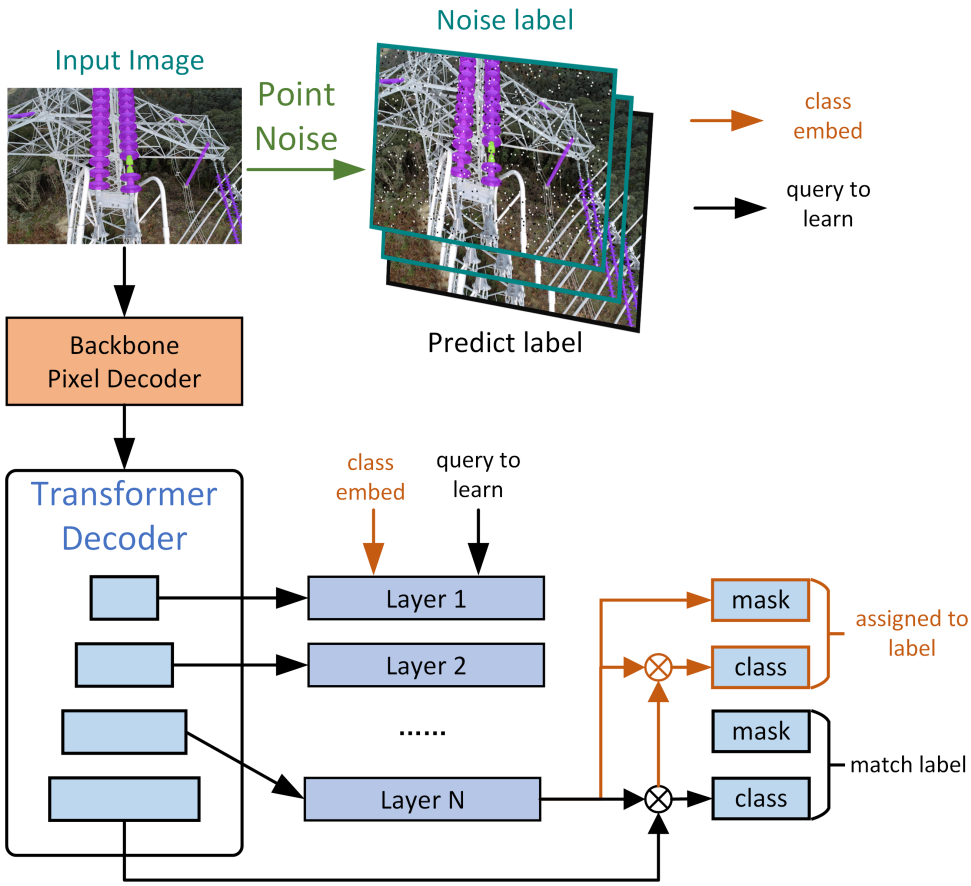


Figure 1. Structure of improved Mask2former.

To reduce the misdetection rate of insulator defects and improve prediction accuracy, additional queries and masks are introduced in the transformer structure of the mask-guided part. In the loss function, the outputs from both the mask-guided and matching parts are assigned to label instances. The predicted results from the mask-guided part are assigned to the corresponding label masks, while the predictions from the matching part proceed with binary graph matching. Following the original network's loss function design, the loss weights for both parts are set equally.

Building on Mask2Former, label masks are added in multiple layers, and different resolutions of feature maps are used in different decoder layers. As a result, the label masks are interpolated to different resolutions when applied to different layers. By adding noisy labels, the mask refinement in each decoder becomes more robust. Since Mask2Former refines masks from previously inaccurate predictions, label masks without noise would reduce the robustness of the extracted features, hindering further improvements in the training process. By inputting noisy masks into the decoder, the model is trained to reconstruct the original masks, slowing the feature extraction rate and increasing the robustness of the final extracted features for insulators and their defects.

2.2. Improvement of HRNet

2.2.1. Backbone Network HRNet

Due to the significant scale variation of the insulators and their defects in the images caused by perspective and distance, this paper replaces the original ResNet backbone [19] of Mask2Former with high-resolution network(HRNet). The HRNet model is shown in Figure 2.

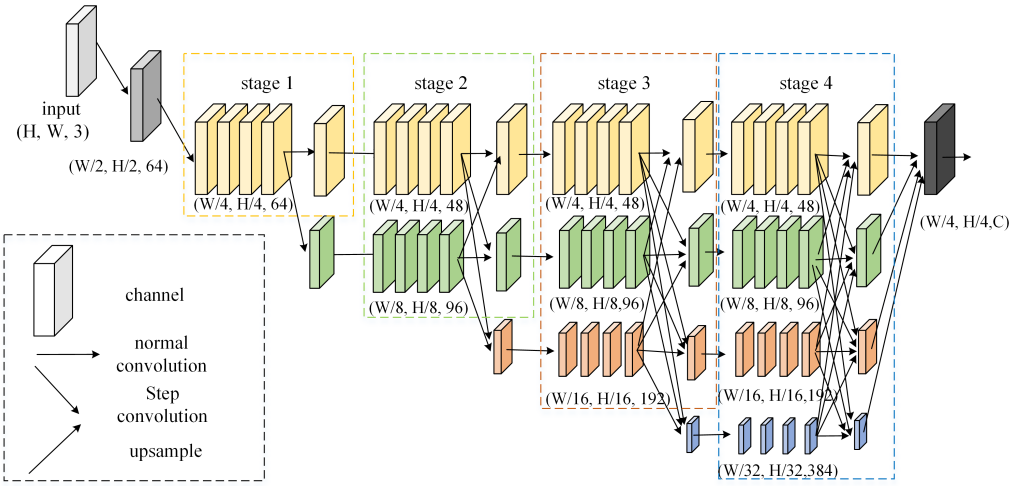


Figure 2. Structure of HRNet.

HRNet is a high-resolution network designed for computer vision tasks such as image classification, object detection, and image segmentation. Its architecture addresses the issue of information loss in traditional networks when processing high-resolution images, while preserving high-resolution and rich semantic information. To achieve this, HRNet uses a multi-branch approach that retains feature maps at multiple resolutions throughout the network. By constructing a multi-branch parallel structure, each branch processes feature maps at different resolutions and exchanges information through high-resolution feature maps. This parallel structure allows HRNet to maintain both high-resolution information and multi-scale feature representations simultaneously.

Compared to the original Mask2Former using ResNet, HRNet has advantages in multi-scale representation capabilities. While ResNet typically processes features at a single resolution, HRNet's multi-branch parallel structure can retain low-, medium-, and high-resolution features simultaneously, improving the ability to detect objects of varying sizes. Furthermore, HRNet maintains higher spatiotemporal consistency. During the construction of the feature pyramid, dense connectivity ensures spatiotemporal consistency, allowing each location in the feature map to benefit from features at all resolutions, rather than relying on fixed-resolution information. This spatiotemporal consistency enhances the performance of Mask2Former in object segmentation tasks.

Additionally, HRNet has stronger feature transfer capabilities. By preserving high-resolution features, HRNet captures detailed information from the image, which is crucial for detecting subtle structures and fine details, particularly for small targets like insulator defects. In contrast, ResNet may experience information loss in lower-level features. Therefore, HRNet's superior feature transfer ability allows it to better capture the spatial and shape information of objects.

2.2.2. Deformable Convolutional Networks (DCNs)

In conventional convolution, a fixed and regular sampling grid is used. For example, when a 5×5 convolution kernel slides over an image, each element of the kernel corresponds to a fixed location in the input image, and the convolution result is computed based on the values at these locations. However, this approach has limitations in efficiency and accuracy when handling images with complex spatial variations. For instance, in the insulator images discussed in this paper, shape deformations are caused by varying angles.

To address this issue, a DCN [20] was proposed. A DCN does not use fixed and regular sampling grids; instead, it allows for "shifting" or "deforming" the sampling locations during the convolution process. This enables the network to adapt to more complex spatial

transformations, improving performance and accuracy. The network structure of a DCN is shown in Figure 3.

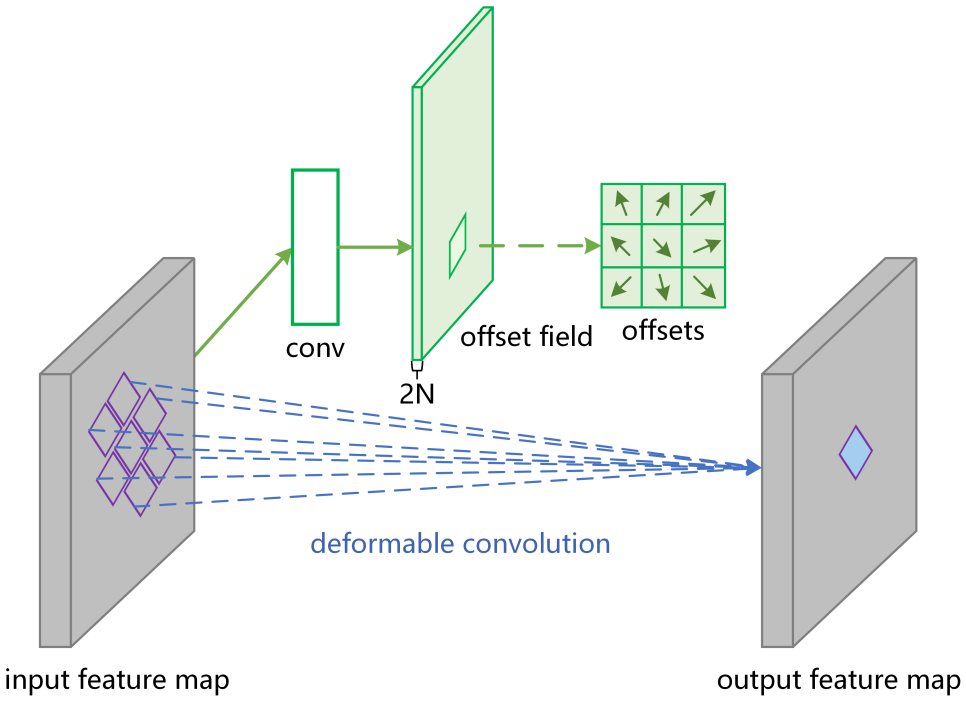


Figure 3. Structure of a DCN.

In the residual structure of HRNet (similar to the residual structure of ResNet), this work introduces a DCN before each residual addition. The specific structure is shown in Figure 4, where "C" represents the number of channels in the feature layer. By introducing a DCN, we can better handle complex spatial variations, such as shape deformations, thereby improving the model's performance and accuracy.

2.2.3. Attention Mechanism

The essence of the attention mechanism originates from studies of attention in the human brain [21]. It allows the system to automatically and rapidly focus on relevant targets based on various features, while ignoring irrelevant factors. This mechanism enables neural networks to pay attention to key elements in an input image while suppressing irrelevant ones.

Zhu Xizhou, Cheng Dazhi, and others [22] explored the impact of deformable convolutions and four different factors: query (query); key (key); content (value); and query, content, and relative position. They carried out this investigation using only key content, and the use of relative position affecting target segmentation performance in Transformer attention modules [23]. The experimental results showed that, compared to a standalone Transformer attention module, deformable convolution combined with the key-content-based Transformer attention mechanism achieved the best trade-off between accuracy and efficiency.

Based on these findings, deformable convolution and the Transformer attention module considering only the key content (key attention) are integrated into the residual block (basic block) of the HRNet network as attention blocks. The residual block structure of this attention block is shown in Figure 5.

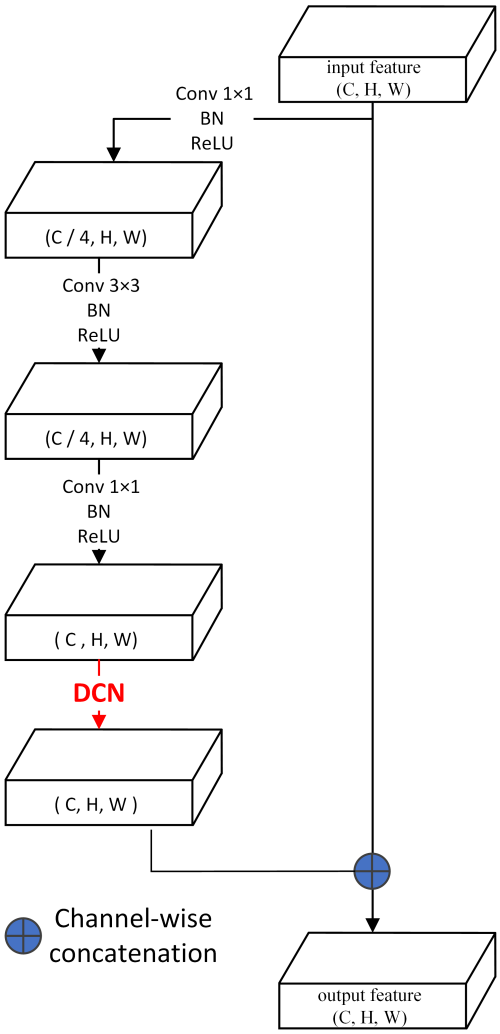


Figure 4. Residual structure of fused DCN.

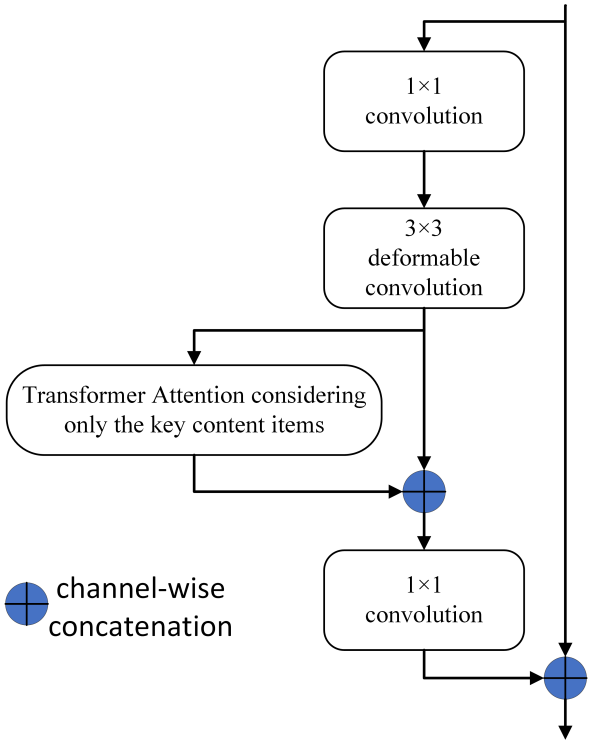


Figure 5. Diagram of the structure with residual blocks incorporating an attention mechanism.

3. Experiments and Analysis

3.1. Embedded Platform Setup

The system is built on the NVIDIA Jetson TX2 with Seeed Studio platform (NVIDIA Corp., Santa Clara, CA, USA), featuring an NVIDIA-ARM architecture with a 1024-core NVIDIA Ampere architecture GPU, 32 Tensor cores, and 4 MB of L3 cache. This setup ensures exceptional performance for handling complex computational tasks, including the execution of deep learning models. The platform benefits from the acceleration provided by cuDNN and CUDA, enabling high-speed computations.

The experimental platform includes a custom-built quadcopter drone, equipped with the Pixhawk V4 flight controller to manage the drone's attitude and flight trajectory. The drone is fitted with a HDR camera capable of capturing high-definition video at 30 frames per second, ensuring the image quality meets the requirements for defect detection. The onboard computational unit, based on the NVIDIA Jetson TX2, is responsible for real-time processing and analysis, including the detection and analysis of insulator defects during flight. Figure 6 shows the appearance and configuration of the drone and computational platform. To evaluate the real-time performance of our method, we conducted inference time tests on the NVIDIA Jetson platform with Seeed Studio platform and RTX3090. The results are shown in Table 1.

Table 1. Ablation experiment comparison results of HRNet, DCN and key attention.

Platform	Inference Time/ms	Video Memory/%	FPS
NVIDIA Jetson	96.1	1286	10.4
NVIDIA RTX 3090	24.4	5134	41.0

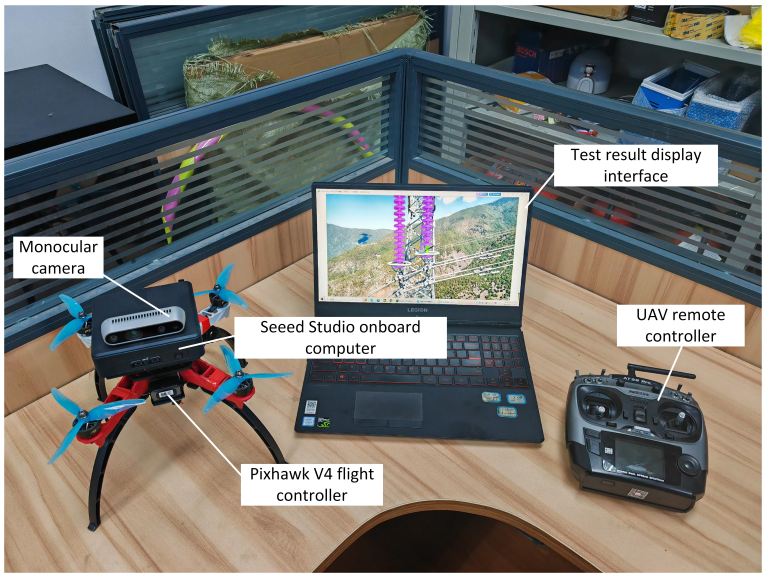


Figure 6. Unmanned aerial vehicle experimental platform.

3.2. Experimental Data Preparation

The data for this experiment were collected from aerial videos taken by an onboard HDR camera during real drone inspections. The videos have a resolution of 3648 × 5472 and a frame rate of 30 fps. We selected 1000 frames from these videos, which were taken from different angles and backgrounds. These frames were divided into training, validation, and testing sets in a 7:2:1 ratio. The original images were manually labeled using LabelImg software v.1.8.3 to create accurate labels. This dataset can be used for tasks like instance

segmentation, object detection, and semantic segmentation. It contains 2737 instances. Sample data are shown in Figure 7. The dataset of different conditions is shown in Figure 8.



Figure 7. Typical samples of annotated dataset.



Figure 8. Dataset of different conditions. (**a**) Cloudy. (**b**) Sunset. (**c**) Sunny.

The images in our dataset were captured by an onboard HDR camera, which was positioned far from the transmission line corridor. The camera took wide-angle photos of the cables and towers, so the background takes up most of the image. The background covers much more space than the target objects, and many insulator strings are hidden by the towers and cables. Because of this, the dataset is challenging to work with.

3.3. Evaluation Metrics

The experimental images in this study use three segmentation evaluation metrics to assess the model's detection performance, accuracy (ACC), intersection over union (IoU), and false-positive rate (FPR), to visually demonstrate the prediction accuracy of the proposed network model. Their calculation formulas are as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$IoU = \frac{Areaof Overlap}{Areaof Union} \tag{3}$$

where true positive (TP) is the number of samples correctly predicted as positive by the model. True negative (TN) is the number of samples correctly predicted as negative. False positive (FP) is the number of samples incorrectly predicted as positive. False negative (FN) is the number of samples incorrectly predicted as negative. Area of overlap is the number of pixels that overlap between the predicted region and the ground truth region. Area of union is the number of pixels covered by both the predicted region and the combined ground truth region.

4. Results Comparison and Analysis

4.1. Ablation Experiment

To verify the improvement in performance of the Mask2former network model, ablation experiments were conducted using insulator string samples with different backgrounds as research objects. The experiments involved sequentially adding the Mask2former improvement strategies, HRNet, DCN modules, and attention mechanisms to the Mask2former network. The specific experimental data are shown in Table 2. As indicated in Table 2, modifying the backbone structure of the network, using DCN, and adding the attention mechanism to the residual block all improved the network's performance, both in terms of individual target accuracy and overall effectiveness, although the degree of improvement varied. The experiments were conducted on an NVIDIA RTX 3090, with the batch size set for 300 epochs, and the total training time was approximately 40 h. Furthermore, we completed an ablation experiment for HRNet, DCN and key attention. These results are shown in Table 3. The heatmaps of the results before and after key attention are shown in Figure 9.

As shown in Table 2, the adoption of the Mask2former improvement strategy resulted in an average increase of 5.1% in accuracy and IoU, with a 2.5% reduction in the FPR. Improving the backbone structure with HRNet led to an average increase of 4.7% in accuracy and IoU. Incorporating the DCN resulted in an average increase of 2.6% in accuracy and IoU. Adding the attention mechanism after the original Mask2former model led to an average increase of 4.7% in accuracy and IoU. With the combined improvement strategies, the proposed algorithm achieved an average increase of 7.72% in accuracy and IoU, with a 4.1% reduction in the false-positive rate.

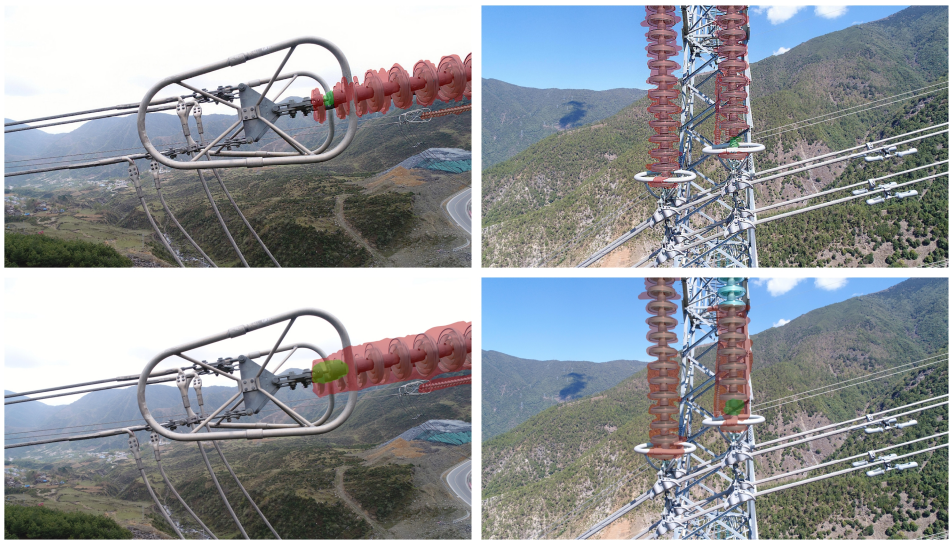


Figure 9. Heatmaps of the results before and after key attention. Bottom: no key attention. Top: key attention.

Table 2. Comparison results of ablation experiments using different improvement measures.

Model	Mask2former	Improvement	HRNet	DCN	Key Attention	ACC/%	IoU/%	FPR/%
1	✓					75.56	76.36	11.23
2	✓	✓				80.66	82.23	8.72
3	✓		✓			79.65	81.35	9.43
4	✓			✓		78.11	81.93	9.57
5	✓				✓	80.21	82.22	8.99
6	✓	✓	✓			81.78	82.89	8.24
7	✓	✓		✓		80.28	81.98	8.67
8	✓	✓			✓	80.99	82.28	8.79
9	✓		✓	✓		81.02	82.87	8.82
10	✓		✓		✓	81.07	82.89	8.62
11	✓			✓	✓	81.63	82.80	8.21
12	✓	✓	✓	✓		82.03	83.25	7.72
13	✓	✓	✓		✓	82.22	83.37	7.61
14	✓	✓		✓	✓	82.73	83.79	7.69
15	✓		✓	✓	✓	82.78	83.56	7.89
16 (proposed algorithm)	✓	✓	✓	✓	✓	83.28	84.77	7.12

Table 3. Ablation experiment comparison results of HRNet, DCN and key attention.

Method	ACC/%	IoU/%	FPR/%	FPS
ResNet	72.13	73.21	17.20	40.2
HRNet	76.18	75.63	14.52	45.3
No Key Attention	72.13	73.21	17.20	40.2
Key Attention	73.12	72.51	15.63	38.1
Normal Convolution	72.13	73.21	17.20	40.2
DCN	73.76	75.28	16.32	42.6

Figures 10 and 11 show the comparison of the network performance before and after the improvement. Before the improvement, instance segmentation was performed using the original Mask2former network. The insulator defect areas were inaccurately located, with ACC and IoU values of only 75.56% and 76.36%, respectively. Additionally, there were instances of missed detections, and the false-positive rate (FPR) reached 11.23%.

Improving the backbone network yields better results for insulators, as insulators face significant issues with image scale variation. HRNet outperforms ResNet in multi-scale tasks.



Figure 10. Results before network improvement.



Figure 11. Results after network improvement.

Using a DCN significantly improves defect localization for insulators. This is because the scale of insulator defects remains relatively consistent with the insulator itself, making them less sensitive to scale variations. Additionally, defects are usually small and occupy a relatively small portion of the image. A DCN can capture more features along the axis of the insulator, significantly improving the model's performance.

It is important to note that since insulator defects are fine-grained targets, using instance segmentation for detection can lead to discrepancies between computational performance and the perceived or actual results. Even if the human eye deems pixel recognition to be correct, the intersection of predicted pixels and true pixels may not be substantial. Furthermore, the dataset is manually annotated, and slight errors in contour annotation can cause the computational results to be lower. Experiments on contour annotation errors and detection results, as shown in Table 4, confirm that the relatively low computational accuracy does not affect the practical use of the model.

Table 4. Influence of contour labeling errors on inspection results.

Contour Labeling Error/%	ACC/%	IoU/%	FPR/%
5	81.58	82.35	7.01
10	80.36	81.56	7.47
15	79.68	80.35	7.78
20	79.02	79.66	8.86

4.2. Comparison with Existing Methods

To verify the superiority of the proposed method, we compared it with existing segmentation or detection frameworks, such as SOLOv2 [24], Yolact++ [25], Mask Scoring RCNN [26], SparseInst [27], PatchDCT [28], and the proposed method.

The results of the comparison experiments are shown in Table 5 and Figure 12. Overall, the proposed method maintains a leading position in terms of key metrics such as accuracy (ACC), intersection over union (IoU), and false-positive rate (FPR). Specifically, in terms of detection accuracy for insulator defects in power transmission lines, the proposed method outperforms existing instance segmentation algorithms. Specifically, the proposed method achieves 11.15% higher ACC and 8.35% higher IoU than SparseInst, with a 9.71% reduction in FPR. It also outperforms PatchDCT by 5.89% in ACC, 8.49% in IoU, and achieves a 7.91% reduction in FPR.

The loss and accuracy curves during the training process of the proposed network are shown in Figure 13. Furthermore, the visual results on the test set are presented in Figure 14. From the figures, it is evident that despite the insulator defects being somewhat blurry and partially occluded, the proposed network is able to accurately segment the insulator and its defects.

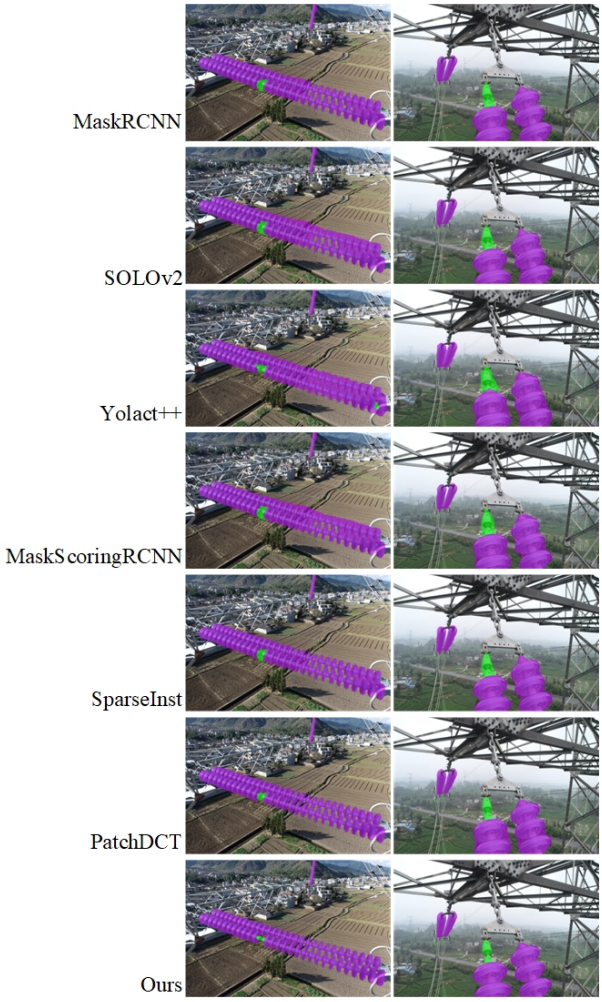


Figure 12. Comparison of visual testing results of different models.

Table 5. Experimental results of different methods.

Method	ACC/%	IoU/%	FPR/%
Mask R-CNN	71.28	72.15	17.43
SOLOV2	70.84	71.39	18.52
Yolact++	46.89	46.33	31.43
Mask Scoring RCNN	69.65	63.24	19.98
SparseInst	72.13	76.42	16.83
PatchDCT	77.39	76.28	15.03
Ours (All Conditions)	83.28	84.77	7.12
Ours (Sunny)	83.52	85.63	6.33
Ours (Cloudy)	82.18	83.47	7.32
Ours (Sunset)	81.12	82.53	8.03

The comparison experiments show that the proposed improvement strategy enhances the performance of the original Mask2former. This allows it to accurately detect and segment insulators and their defects in high-resolution aerial images. The improvement enables more precise analysis of defect issues in power transmission lines during UAV inspections. As a result, it helps with flight path planning and fault diagnosis, contributing to the safety and reliability of the power system.

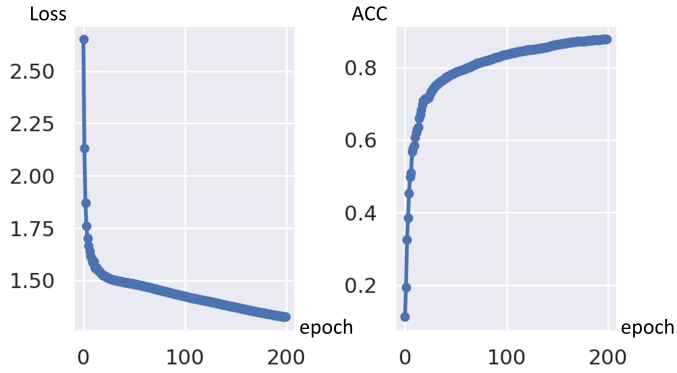


Figure 13. Loss and ACC variation curves for the training process of the model in this paper.



Figure 14. The visual results of the proposed network on the test set.

5. Discussion

5.1. Advantages and Limitations of the Proposed Method

This paper presents an improved insulator defect detection method based on Mask2Former-HRNet, which incorporates HRNet, DCN, and attention mechanisms to significantly enhance the model's detection accuracy and anti-interference capabilities. Experimental results show that the proposed method outperforms existing methods across various evaluation metrics, particularly excelling in handling complex backgrounds and small targets.

However, there are some limitations to the proposed method. Firstly, the model's computational cost is relatively high, especially on embedded platforms, where the inference speed is slower. Future research could focus on optimizing the model structure to reduce computational requirements, enabling faster real-time processing. Secondly, the model's performance in extreme weather conditions (such as heavy rain or fog) still requires improvement. Future work can consider incorporating more data augmentation techniques and multi-modal fusion methods to further enhance the model's robustness. We conducted a detailed analysis of the misclassified samples in the test set. Figure 15 presents several typical misclassification cases. The analysis results show that misclassifications mainly occur under the following conditions:

- Complex background interference: When there are objects with similar colors or textures near the insulator, the model tends to misclassify them as defects. For example, in Figure 15a, the tower near the insulator is mistakenly identified as the insulator itself.
- Occlusion issues: When the insulator is partially occluded by other equipment, the model struggles to accurately detect the complete defect area. In Figure 15b, the lower

half of the insulator is blocked by the tower, resulting in the model failing to identify the crack.

- Low-light conditions: In low-light environments, image quality deteriorates, making it difficult for the model to clearly distinguish defects. In Figure 15c, the image taken at dusk shows blurred surface details of the insulator, and the model misclassifies it as normal.



Figure 15. Detection results for scenarios with higher FPR. (**a**) Similar colors and textures; (**b**) Occlusion; (**c**) Low-light condition.

5.2. Future Research Directions

Future research can be extended in the following areas:

- Multi-view Fusion: Combine images from different angles to reduce the impact of occlusion and improve detection accuracy.
- Lightweight Models: Design more lightweight network structures to reduce computational costs, making the model suitable for resource-constrained embedded platforms.
- Multi-modal Perception: Introduce multi-modal sensors, such as infrared and LiDAR, to enrich perception information and enhance detection capabilities.
- Online Learning: Develop online learning algorithms to allow the model to continuously update and optimize based on new data obtained during actual inspection processes, maintaining optimal performance.

5.3. Challenges in Practical Applications

In practical applications, the proposed method faces several challenges. For instance, during drone inspections, complex electromagnetic interference may be encountered, affecting communication and positioning accuracy. Additionally, long-duration flight missions impose higher demands on battery life. Future research could explore how to optimize drone flight path planning to reduce unnecessary energy consumption and extend flight durations.

6. Conclusions

This study addresses the limitations of existing insulator localization methods in terms of accuracy and defect recognition by proposing a UAV-based insulator defect detection platform and developing an improved Mask2Former-HRNet method for detecting and localizing defects in transmission line insulators.

Firstly, by introducing mask-guided and matching components into Mask2Former, the false-positive rate was significantly reduced, and prediction accuracy was enhanced. In the ablation experiments, implementing these improvement strategies resulted in an average increase of 5.1% in accuracy (ACC) and intersection over union (IoU), while the false-positive rate (FPR) decreased by 2.5%. Secondly, replacing ResNet with HRNet as the backbone network enhanced multi-scale feature representation capabilities and spatiotemporal consistency, making the model more robust when processing high-resolution images. A further incorporation of deformable convolutional networks (DCNs) effectively

addressed complex spatial transformations caused by varying perspectives, improving the model's performance in handling shape deformations. Specifically, adding DCNs led to an average increase of 2.6% in ACC and IoU, and a reduction of 0.42% in FPR. Lastly, integrating the key attention mechanism boosted the network's ability to focus on critical regions, further enhancing overall detection performance. The combined improvements culminated in the proposed algorithm achieving an ACC of 83.28%, an IoU of 84.77%, and an FPR of 7.12%.

In comparison with existing methods, the improved Mask2Former-HRNet method demonstrated superior performance across all key metrics. Specifically, compared to traditional Mask R-CNN, the proposed method achieved a 12% increase in ACC, a 12.62% increase in IoU, and a 10.31% reduction in FPR. When compared to the latest PatchDCT method, it showed a 6.89% improvement in ACC, an 8.49% enhancement in IoU, and a 7.91% reduction in FPR. These results confirm that the proposed method offers significant advantages in insulator defect detection tasks, providing higher detection accuracy and lower false-positive rates.

In conclusion, the improved Mask2Former-HRNet method proposed in this study, through multifaceted optimizations, significantly enhances the defect detection performance of transmission line insulators. Experimental results validate the method's effectiveness and superiority in practical applications, highlighting its potential for ensuring the safety and reliability of power systems. Future research will focus on further optimizing the model structure and expanding its application to larger-scale real-world scenarios, thereby advancing the intelligent development of UAV-based power inspections.

Author Contributions: Conceptualization, Y.H.; methodology, Y.H., L.X., Z.T. and J.Z.; investigation, X.D. and Y.X.; supervision X.F.; data curation Y.H.; writing—original draft, Y.H.; writing—review and editing, X.F.; project administration, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by State Grid Sichuan Electric Power Company Technology Project(521947230002). This funding is supported by Information and Communication Company, State Grid Sichuan Electric Power Company and Sichuan University.

Data Availability Statement: Data are not available due to commercial restrictions.

Conflicts of Interest: Authors Yaoran Huo, Lan Xiao, Zhenyu Tang, Xu Dai and Yuhao Xiao were employed by Information and Communication Company, State Grid Sichuan Electric Power Company and author Jian Zhou was employed by State Grid Sichuan Electric Power Company. The remaining author declares that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflicts of interest.

References

1. Yubo, W.; Junli, S.; Ye, Z.; Jianyong, L.; Lei, Y.; Botao, L. Real-Time Transmission Conductor Defect Detection Method Based On Improved Yolov7. *South. Grid Technol.* **2023**, *17*, 127–134. [CrossRef].
2. Junpeng, M.; Xingtao, S.; Shuo, L.; Fuzhai, L.; Zhifeng, T.; Xianqiang, Y. Feasibility Study of Defect Detection Tech-nology for Tub Insulators Based On Ultrasonic Guided Waves. *High Volt. Technol.* **2019**, *45*, 3941–3948. [CrossRef].
3. Jiafu, C.; Zhenhua, L.; Huiqin, L.; Ye, T.; Difei, S.; Shidan, L.; Lei, C. Substation Virtual Terminal Connection Veri-fication Technology Based On Deep Learning and Image Recognition. *Guangdong Electr. Power* **2024**, *37*, 73–79. [CrossRef].
4. Kai, Y.; Ying, X. Study On Tie Method of Aerial Insulator Image Detection and Recognition. *Power Syst. Big Data* **2015**, *18*, 51–53.
5. Shaoping, Z.; Zhong, Y.; Xiaoning, H.; Huaiqun, W.; Yuanzheng, G. Defects Detection and Positioning for Glass Insulator from Aerial Images. *J. Terahertz Sci. Electron. Inf. Technol.* **2013**, *11*, 609–613.
6. Ting, F.; Ming, H.J. Detection and Localization of Insulator Defects in Aerial Images. *Comput. Sci.* **2016**, *43*, 222–225.
7. Souza, B.J.; Stefenon, S.F.; Singh, G.; Freire, R.Z. Hybrid-YOLO for classification of insulators defects in transmission lines based on UAV. *Int. J. Electr. Power Energy Syst.* **2023**, *148*, 108982. [CrossRef]

8. Ninghui, H.; Shijie, W.; Junfu, L.; Hao, Z.; Liangfang, W.; Xiu, Z. Research On Infrared Image Missing Insulator Detection Method Based On Deep Learning. *Power Syst. Prot. Control* **2021**, *49*, 132–140. [CrossRef].
9. Chao, H.; Xiaogang, G.; Lingqin, H.; Shengyang, L. Improved Yolo V4 Model for Insulator Defect Detection Using Aerial Imagery. *Electron. Meas. Technol.* **2023**, *46*, 175–181. [CrossRef].
10. Shuang, R.; Jizai, S.; Kai, Y.; Jiming, Q.; Xiangyu, W.; Yonggen, C. Improved Insulator Image Detection Model for Yolov4. *Guangdong Electr. Power* **2023**, *36*, 94–101. [CrossRef].
11. Tao, X.; Zhang, D.; Wang, Z.; Liu, X.; Zhang, H.; Xu, D. Detection of power line insulator defects using aerial images analyzed with convolutional neural networks. *IEEE Trans. Syst. Man Cybern. Syst.* **2018**, *50*, 1486–1498. [CrossRef]
12. Deng, F.; Luo, W.; Wei, B.; Zuo, Y.; Zeng, H.; He, Y. A novel insulator defect detection scheme based on Deep Convolutional Auto-Encoder for small negative samples. *High Volt.* **2022**, *7*, 925–935. [CrossRef]
13. Zhang, C.; Wen, C. Fault Detection of Wind Turbine Blade Based on Improved Mask R-Cnn. *Renew. Energy Resour.* **2020**, *38*, 1181–1186. [CrossRef].
14. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299. [CrossRef].
15. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [CrossRef] [PubMed]
16. Sheng, J.C.; Liao, Y.S.; Huang, C.R. Apply Masked-attention Mask Transformer to Instance Segmentation in Pathology Images. In Proceedings of the 2023 Sixth International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 30 June–3 July 2023; pp. 342–345. [CrossRef].
17. Guo, S.; Yang, Q.; Xiang, S.; Wang, S.; Wang, X. Mask2Former with Improved Query for Semantic Segmentation in Remote-Sensing Images. *Mathematics* **2024**, *12*, 765. [CrossRef]
18. Yuan, Y.; Hou, S.; Wu, X.; Wang, Y.; Sun, Y.; Yang, Z.; Yin, S.; Zhang, F. Application of deep-learning to the automatic segmentation and classification of lateral lymph nodes on ultrasound images of papillary thyroid carcinoma. *Asian J. Surg.* **2024**, *47*, 3892–3898. [CrossRef]
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef].
20. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets v2: More Deformable, Better Results. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA, USA, 15–20 June 2019; pp. 9300–9308. [CrossRef].
21. Mashour, G.A.; Roelfsema, P.; Changeux, J.P.; Dehaene, S. Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron* **2020**, *105*, 776–798. [CrossRef] [PubMed]
22. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An Empirical Study of Spatial Attention Mechanisms in Deep Networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6687–6696. [CrossRef].
23. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. In Proceedings of the 57th ANNUAL Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy, 28 July–2 August 2019; Korhonen, A.; Traum, D.; Marquez, L., Eds.; pp. 2978–2988.
24. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. SOLOv2: Dynamic and Fast Instance Segmentation. In Proceedings of the Advances in Neural Information Processing Systems 33, Neurips 2020, Virtual, 6–12 December 2020; Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; Lin, H., Eds.; Volume 33.
25. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. YOLACT plus plus Better Real-Time Instance Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1108–1121. [CrossRef] [PubMed]
26. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask Scoring R-CNN. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA, USA, 15–20 June 2019; pp. 6402–6411. [CrossRef].
27. Cheng, T.; Wang, X.; Chen, S.; Zhang, W.; Zhang, Q.; Huang, C.; Zhang, Z.; Liu, W. Sparse Instance Activation for Real-Time Instance Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, LA, USA, 18–24 June 2022; pp. 4423–4432. [CrossRef].
28. Shen, X.; Yang, J.; Wei, C.; Deng, B.; Huang, J.; Hua, X.; Cheng, X.; Liang, K. DCT-Mask: Discrete Cosine Transform Mask Representation for Instance Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021, Nashville, TN, USA, 20–25 June 2021; pp. 8716–8725. [CrossRef].

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.