(3) The FA-SSD model [32] is proposed to improve the accuracy and recall rate of insulator umbrella disc shedding detection.

2. Materials and Methods

As shown in Figure 3, the overall process of umbrella disc shedding detection included three parts: pre-training and fine-tuning of the defogging model, training with the clear insulator image datasets, and testing with the fogged insulator images.

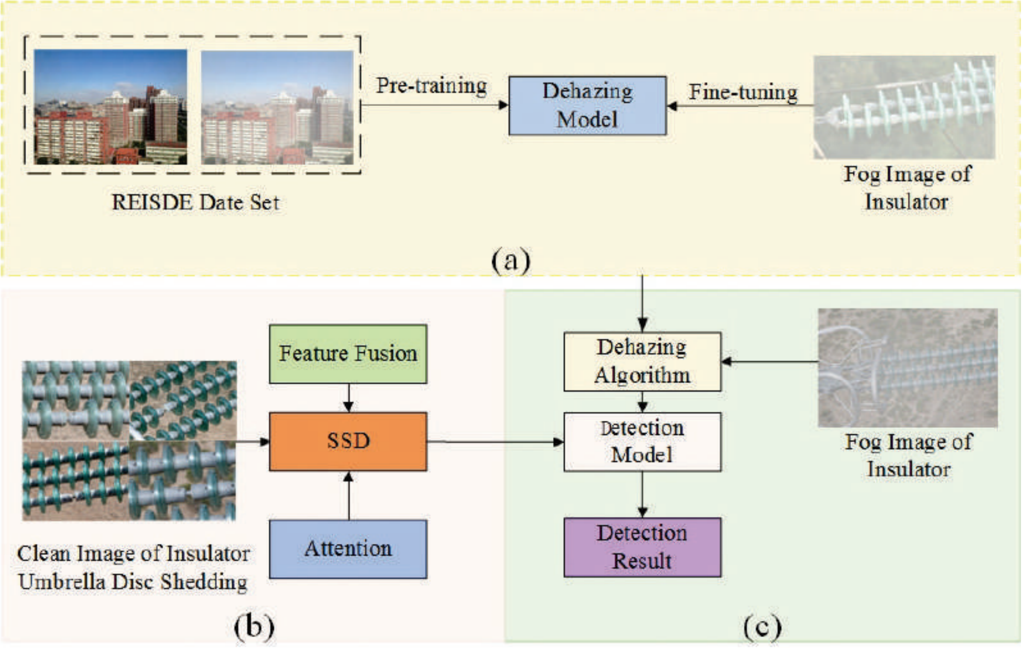


Figure 3. The overall process of umbrella disc shedding detection. (**a**) Dehaze model. (**b**) Training phase. (**c**) Testing phase.

The dehazing model was trained by synthetic foggy images, and the insulators with foggy images were fine-tuned to improve the dehazing effect of the algorithm. A feature fusion module and an attention module were added to the umbrella disc shedding detection model to improve the detection accuracy. In the detection of the insulator umbrella disc shedding, clear images of insulators were used for training, and images of insulators with fog were used for testing.

2.1. Dehazing Model

Inspired by the dehazing algorithm proposed by Chen [33], this paper adopted the method of pre-training and fine-tuning to improve the dehazing effect of the dehazing model. The training of the model was divided into two steps. The first step used a large number of haze-free images and artificially-generated fogged images from the REISDE dataset [34] to train the dehazing model, and the second step used the foggy insulator images to fine-tune the dehazing model to improve the dehazing ability of the dehazing model on fogged insulator images. During fine-tuning, physical priors were guided through the loss function. As shown in Figure 4, the dehazing model had a two-stage framework.

In the pre-training stage, an advanced dehazing model was adopted as the backbone. The pre-training phase used synthetic data for training, resulting in a pre-trained model on the synthetic domain. In the fine-tuning stage, the fog-free image J, transmission map t, and atmospheric light A were obtained through the backbone network. At the same time, three priors, including a dark channel prior, a bright channel prior, and the Contrast Limited Adaptive Histogram Equalization (CLAHE) were introduced, and the model was guided in the form of loss function.

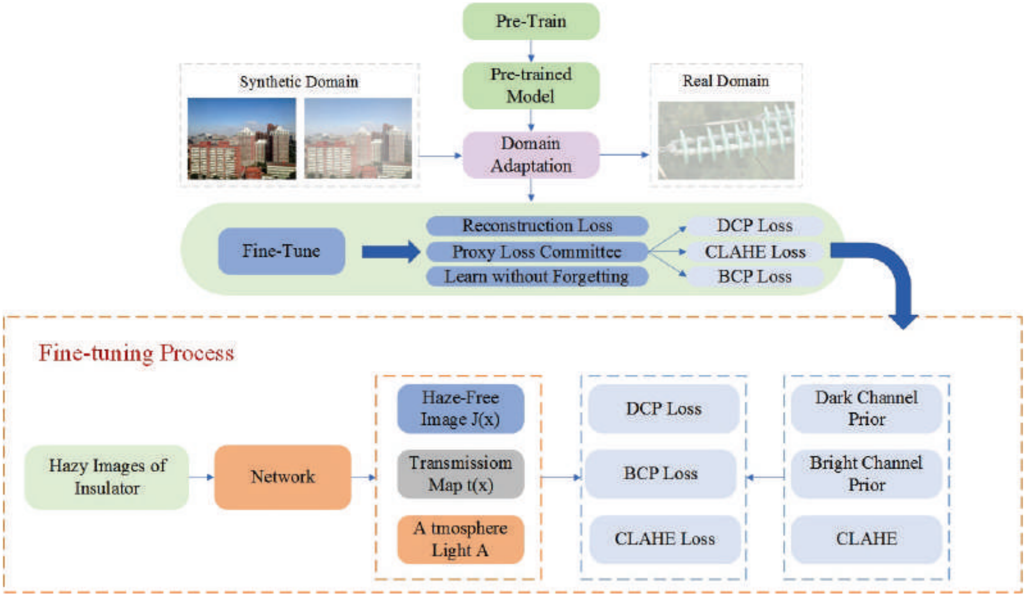


Figure 4. Structure of the dehaze model.

The loss function of the dark channel prior is shown as follows:

$$L_{DCP} = E(t, \widetilde{t}) = t^T L t + \lambda (t - \widetilde{t})^T (t - \widetilde{t}) \quad (1)$$

where t and $\widetilde{t}$ denote the transmission estimates from the DCP and the backbone network, respectively. L is a Laplacian-like matrix.

The loss function of the bright channel prior is shown as follows:

$$L_{BCP} = \left\| t - \widetilde{t} \right\|_1 \quad (2)$$

where t and $\widetilde{t}$ represent the transmission estimates from the BCP and the backbone network, respectively.

The loss function of the CLAHE reconstruction is shown as follows:

$$L_{CLAHE} = \| I - I_{CLAHE} \|_1 \quad (3)$$

where I is the original hazy input, and I_CLAHE is the reconstruction result by J_CLAHE, $\widetilde{t}$, and $\widetilde{A}$.

The role of the three loss functions is different. Dark channel prior greatly advances the model performance on real hazy images, bright channel prior helps make the resulting images brighter and with enhanced contrast, and CLAHE is used to achieve a balance between L_DCP and $LBCP$.

The total loss of the fine-tuning process was obtained by combining the three losses as follows:

$$L_{com} = \lambda_d L_{DCP} + \lambda_b L_{BCP} + \lambda_c L_{CLAHE} \quad (4)$$

where λ_d, λ_b, and λ_c are the tradeoff weights.

2.2. Fa-Ssd Model

Target detection includes target recognition and localization. For CNN, the two are contradictory [35]. Generally speaking, deep feature maps contain more semantic information, which is good for object recognition but not good for object localization; the difference is that the shallow feature map contains more detailed features, which is good for object localization but not good for object recognition. As shown in Figure 5, the SSD model adopts a feature pyramid structure to detect objects of different scales; small objects are detected on the shallow feature maps, and large objects are detected on the deep feature maps.

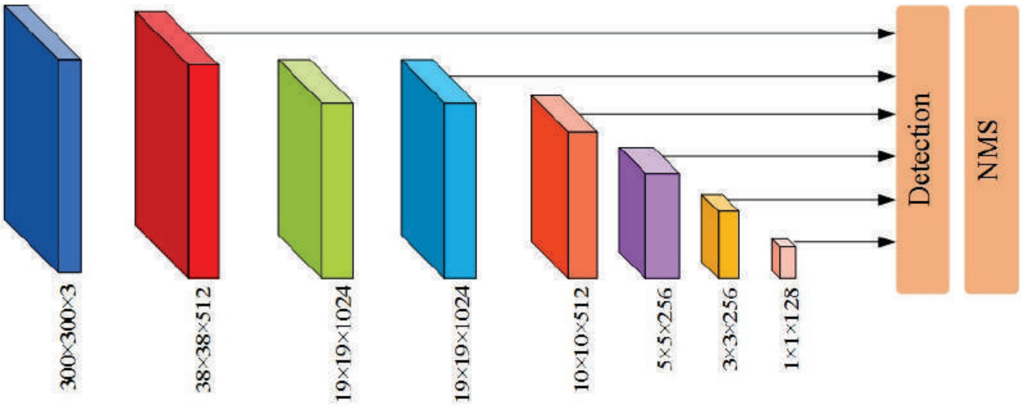


Figure 5. Structure of the SSD model.

However, the problem with this method is that the small target features generated by the shallow layer lack sufficient semantic information, and the detection of small targets still is not effective. In order to improve the detection ability of the SSD model for the insulator umbrella disk shedding, the FA-SSD model is proposed. As shown in Figure 6, the FA-SSD model adds a feature fusion module and an attention module to the SSD model.

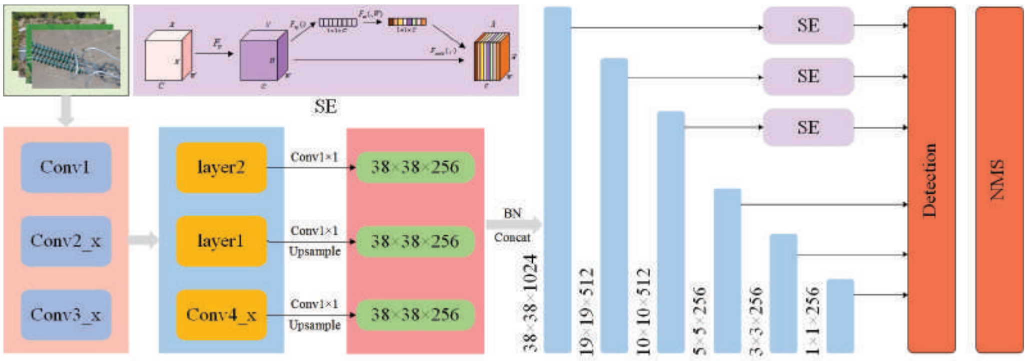


Figure 6. Structure of the FA-SSD model.

First, the insulator images were sent to the ResNet50 [36] feature extraction network to extract the features. Since the shallow feature maps contain richer small target detail information, Conv4_x in ResNet50 and two auxiliary convolutional layers were selected for feature fusion. The feature dimension of Conv4_x was 38 × 38 × 1024, and the feature dimensions of the two auxiliary convolutional layers were 19 × 19 × 512 and 10 × 10 × 512. Then, in order to fuse the features of the three different scales simply and efficiently, the two auxiliary convolutional layers were upsampled using bilinear interpolation to make them the same size as Conv4_x. Finally, the feature map was concatenated and normalized to generate a new feature pyramid structure for the identification and localization of umbrella disc shedding. The parameters of each layer in the structure are shown in Table 1.

On this basis, in order to enhance the network's ability to extract low-level detail features, the SE channel attention module [37] was added to the lowest three layers of the feature pyramid.

Table 1. Input and output dimensions of each layer.

Layer Name	Input	Output
Conv1	300 × 300 × 3	150 × 150 × 64
Conv2_x	150 × 150 × 64	75 × 75 × 256
Conv3_x	75 × 75 × 256	38 × 38 × 512
Conv4_x	38 × 38 × 512	38 × 38 × 1024
layer1	38 × 38 × 1024	19 × 19 × 512
layer2	19 × 19 × 512	10 × 10 × 512

SE Module

The SE learns a set of weight coefficients through a small fully connected network to weigh each channel of the original feature map. In this way, different weights are assigned to each channel to enhance the feature extraction capability of the network. The implementation process of the SE was as follows:

(1) We performed convolution pooling and other operations on the input image to obtain a feature map:

$$u_c = v_c * X = \sum_{s=1}^{c'} v_c^s * x^s \tag{5}$$

where v_c and X represent the convolution kernel and the input image, respectively; v_c^s and x^s represent the convolution kernel and the sth channel of the input image, respectively; and c' represents the number of channels.

(2) We squeezed and compressed the feature map into one-dimensional features:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{6}$$

where H and W represent the width and height of the feature map, respectively.

(3) For excitation, we performed activation operations on multiple channels to extract different features:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{7}$$

(4) We multiplied the obtained weight factor with the corresponding channel feature to obtain a new feature map:

$$\widetilde{x_c} = F_{scale}(u_c, s_c) = s_c \cdot u_c. \tag{8}$$

3. Results

3.1. Experimental Environment

The proposed model used an NVIDIA RTX 2080Ti GPU for training and testing and the Ubuntu 18.04 LTS as the operating system; the training process was accelerated by CUDA 10.1; the computer language was Python 3.6, and the network framework was PyTorch. The batch size was set to 8, the learning rate was 0.003, the preprocessed size of the input image was 300 × 300, and the maximum number of iterations was 7800. The SSD was chosen as the baseline for improvement and comparison purposes.

The datasets used in the dehazing stage included the REISDE dataset and images of fogged insulators. The insulator images used in this paper were the aerial images of transmission line inspection, which were obtained by UAV. The datasets used in the object detection stage consisted of fogged insulator images, as well as fog-free insulator images. Since the insulators were in normal working condition most of the time, the defect images occupied a small proportion of the obtained aerial images. In addition, due to factors such as shooting environment, shooting angle, shooting distance, etc., many images were of poor quality. By cooperating with several power grid companies, we obtained some samples of insulator umbrella disk shedding. Among them, there were 160 images (the number of the insulator umbrella disc shedding was 176) with fog and 480 images (the number of the insulator umbrella disc shedding was 518) without fog. We used the images without fog as the training set and the images with fog as the test set. As shown in Figure 7, the insulator datasets contained glass insulators and ceramic insulators.

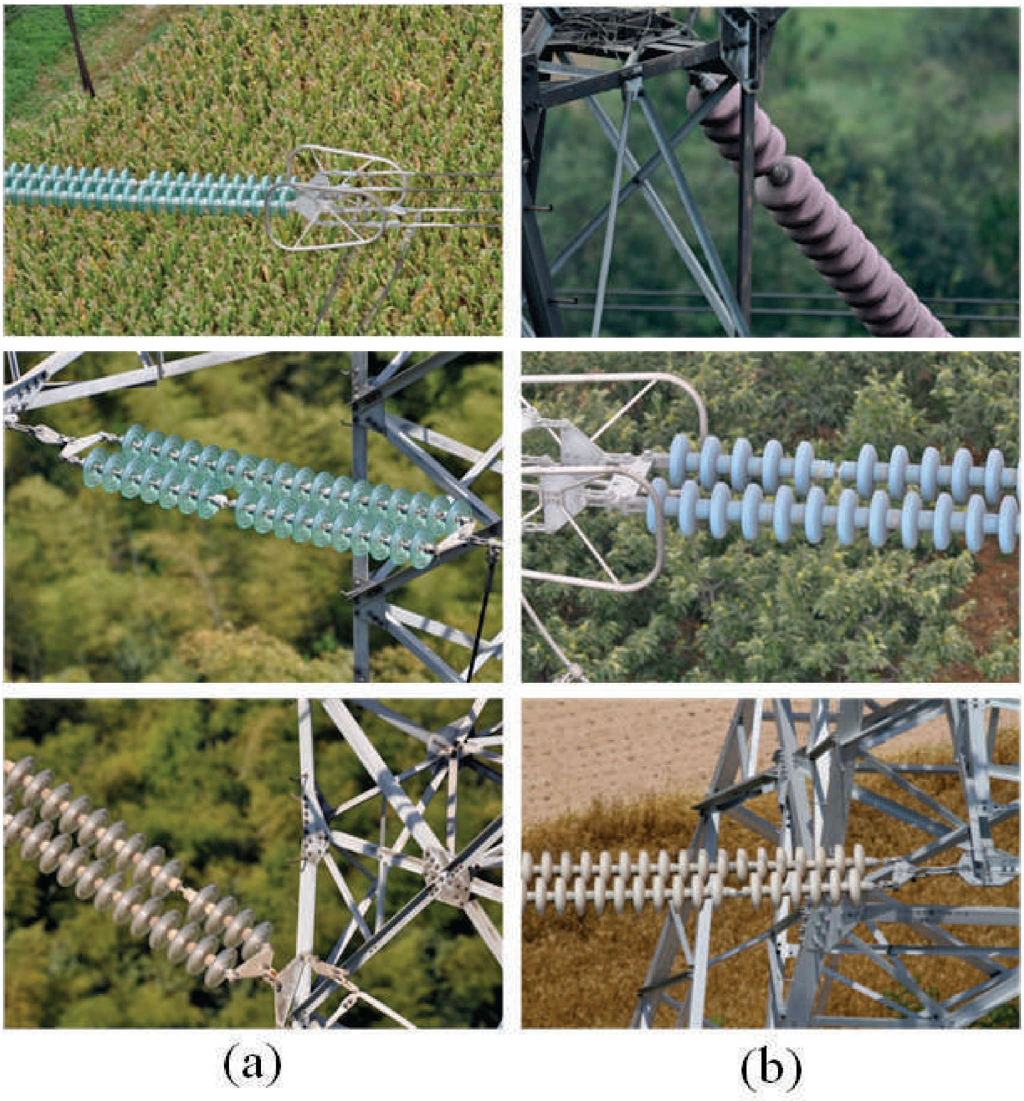


Figure 7. Glass insulators and ceramic insulators. (**a**) Glass insulators. (**b**) Ceramic insulators.

To compare the different models, precision(P), recall(R), and F_1 were used as model evaluation metrics. The higher the value, the better the detection performance of the model.

$$P = \frac{T_P}{T_P + F_P} \tag{9}$$

$$R = \frac{T_P}{T_P + F_N} \tag{10}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{11}$$

where TP and FP denote the number of correctly and incorrectly located defects, respectively. $TP + FP$ is the total number of located defects, and$TP + FN$ is the total number of actual defects. F_1 is the harmonic mean of precision and recall.

3.2. Ablation Experiment of Fa-Ssd Model

In order to verify the effectiveness of the feature fusion module and the attention module, the experiments were conducted on the original SSD model, the SSD model with the feature fusion module, the SSD model with the attention module, and the FA-SSD model. The visualization results of the FA-SSD model and the SSD model are shown in Figure 8.

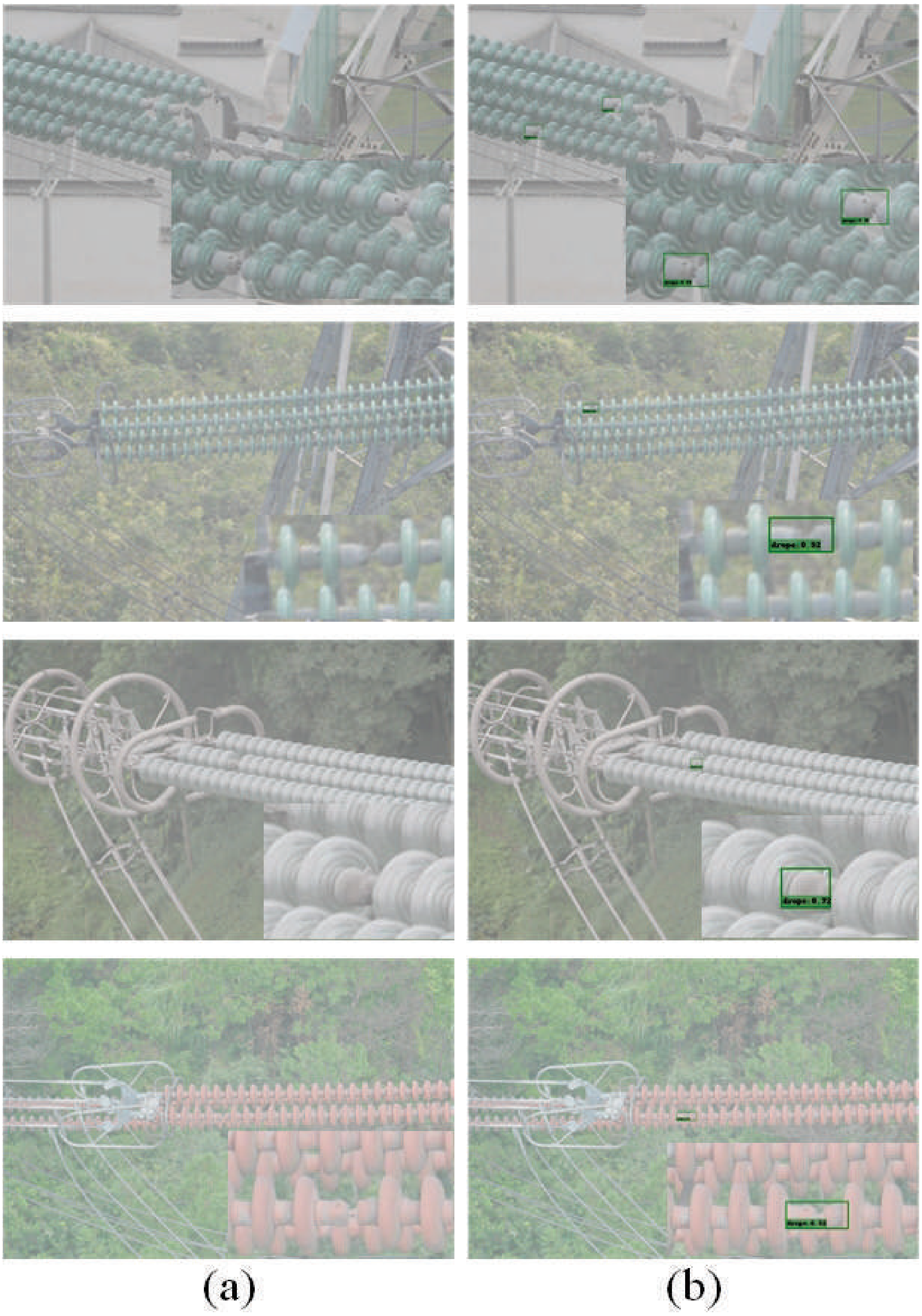


Figure 8. Visualization of SSD and FA-SSD. (**a**) SSD. (**b**) FA-SSD.

In the experiment, the other parameters of the model training were guaranteed to be the same, and the obtained detection results are shown in Table 2.

Table 2. Results of the ablation experiment.

SSD	Feature Fusion	Attention	*P*	*R*	F_1
✓			0.866	0.755	0.806
✓	✓		0.899	0.769	0.828
✓		✓	0.877	0.793	0.832
✓	✓	✓	**0.909**	**0.817**	**0.860**

The detection performance of the FA-SSD was better than the methods that only added the feature fusion module or the attention module. Compared with the original SSD model, the accuracy rate was improved, the recall rate was improved, and the F1 indicator was improved. The experimental results showed that both the feature fusion module and the attention module had a positive effect on the model.

3.3. Compared with Other Methods

In order to further verify the effectiveness of the FA-SSD model in the detection of insulator umbrella disk shedding, under the condition of ensuring the same feature extraction network and hyperparameters, the method in this paper was compared with the commonly used target detection algorithm at this stage. The compared methods included Faster R-CNN [27], YOLOV3 [25], and RetinaNet [38], and the results are shown in Figures 9 and 10 and Table 3.

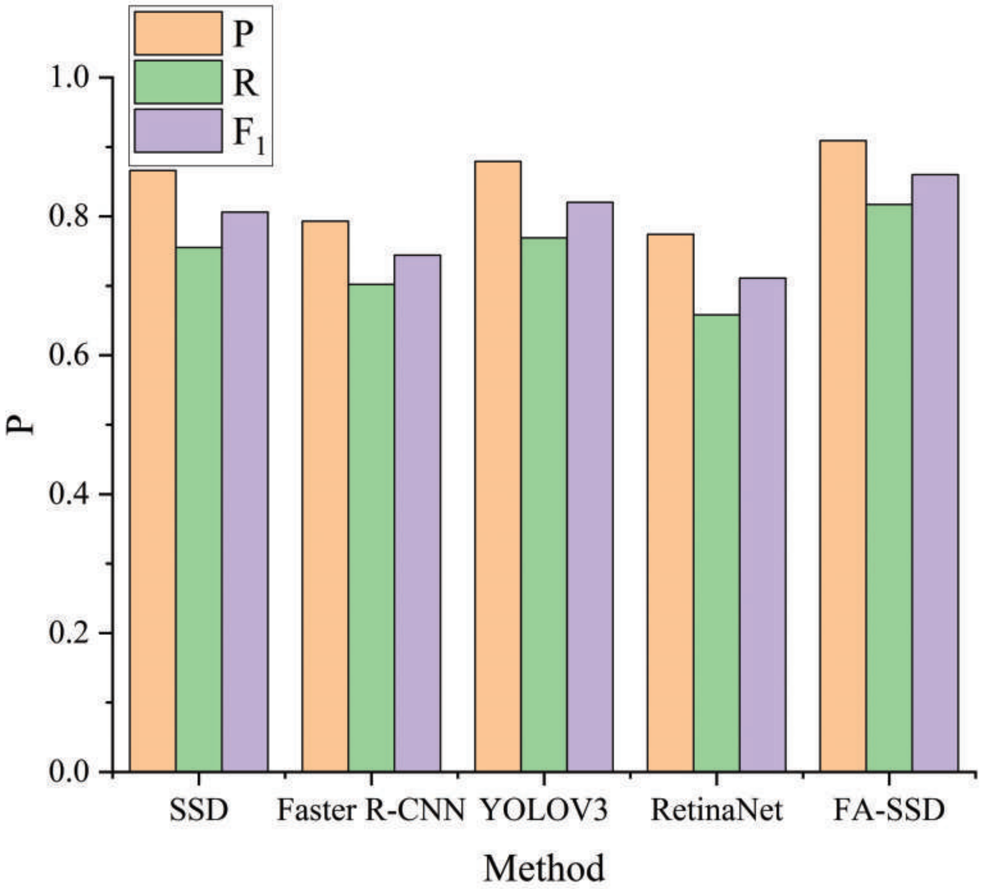


Figure 9. Results of different methods.

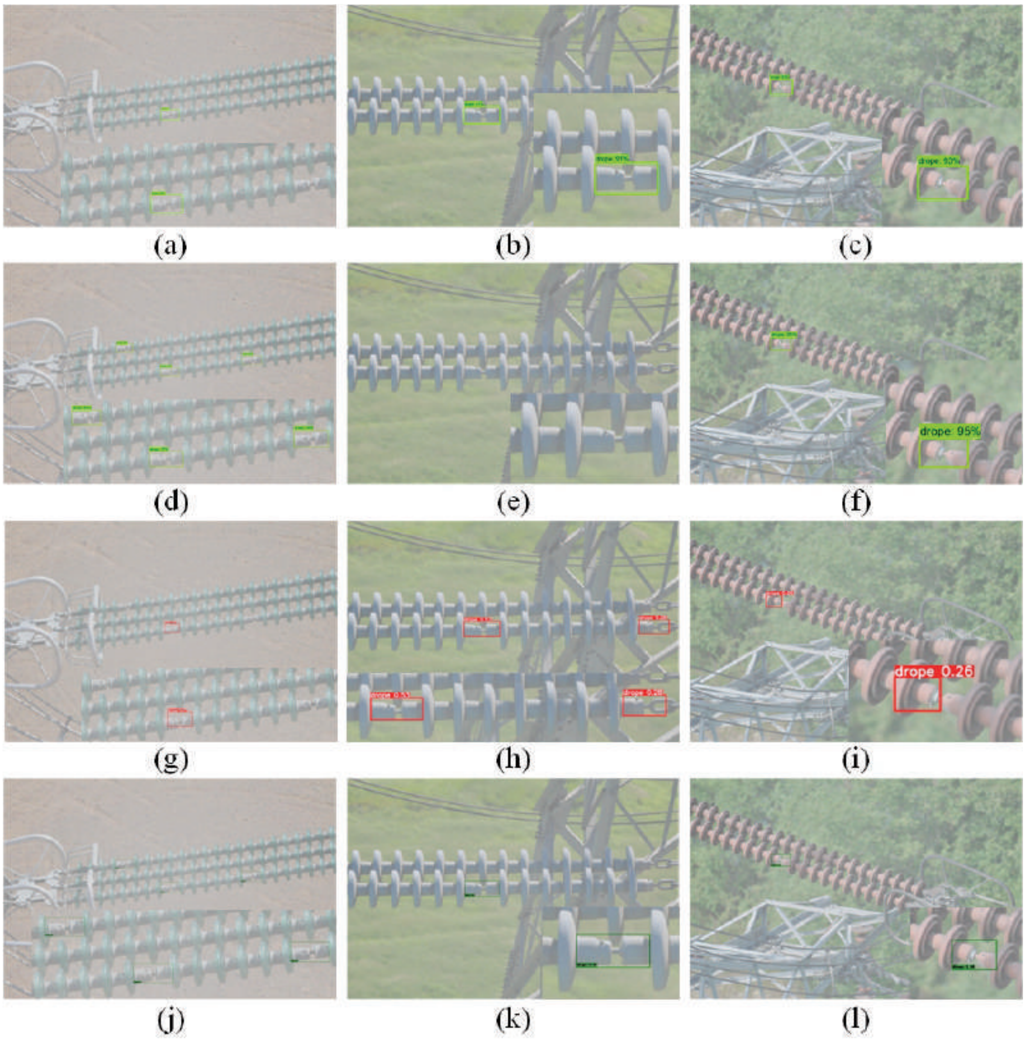


Figure 10. Visualization results of different methods. (**a**–**c**) Faster R-CNN. (**d**–**f**) YOLOV3. (**g**–**i**) RetinaNet. (**j**–**l**) FA-SSD.

Table 3. Results of different methods.

Method	Input Size	*P*	*R*	F_1
SSD [32]	300 × 300	0.866	0.755	0.806
Faster R-CNN [27]	800 × 800	0.793	0.702	0.744
YOLOV3 [25]	300 × 300	0.879	0.769	0.820
RetinaNet [38]	300 × 300	0.774	0.658	0.711
FA-SSD	300 × 300	**0.909**	**0.817**	**0.860**

It can be seen that FA-SSD significantly outperformed SSD and other commonly used object detection algorithms. Compared with other algorithms, the accuracy rate of detecting the umbrella disc shedding was improved on average 8.1%, and the recall rate was improved on average 9.6%. Compared with other target detection algorithms, the FA-SSD algorithm improved the detection accuracy and reduced the missed detection rate.

3.4. Dehazing Algorithm Experiment

As shown in Figure 11, after using the dehazing algorithm to dehaze the hazy images, the pictures became clearer.

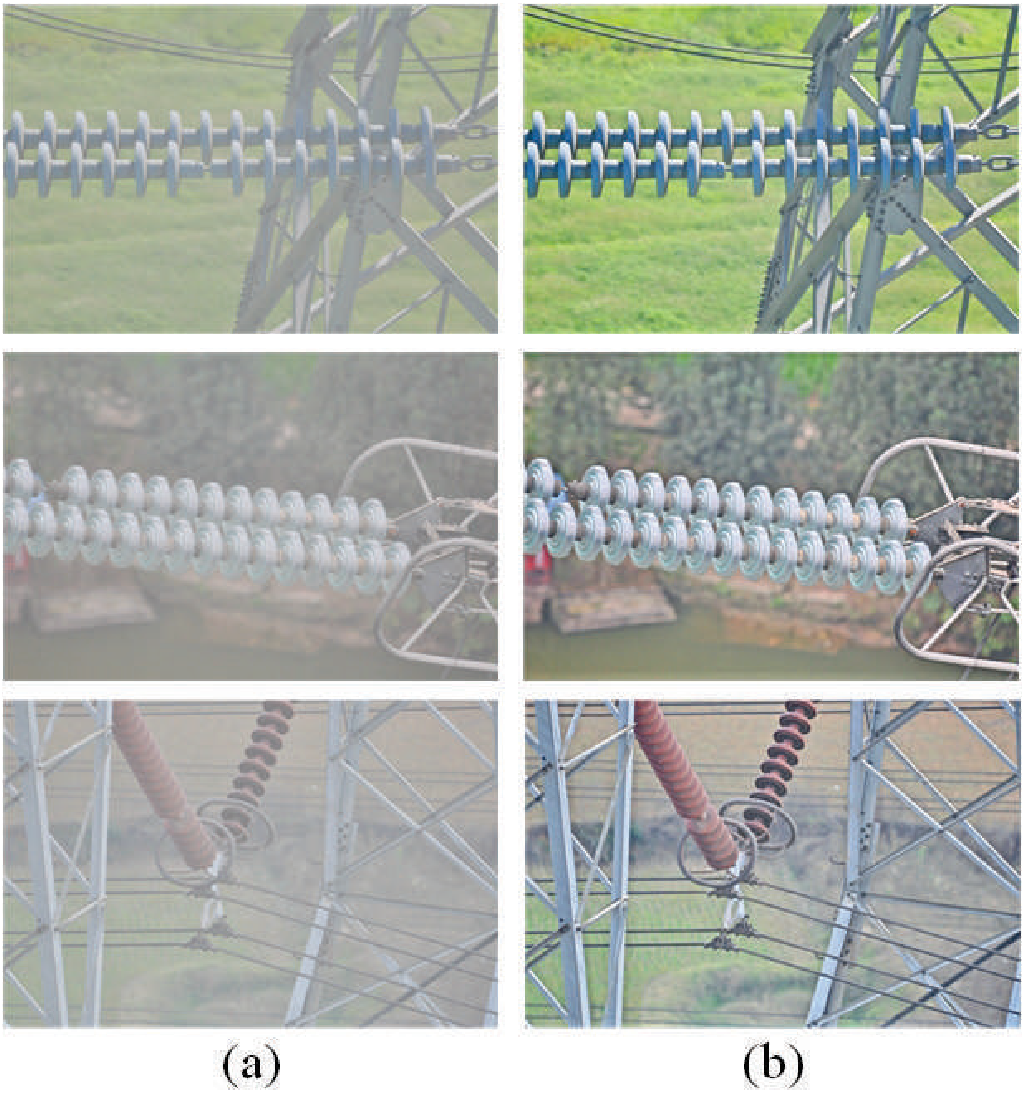


Figure 11. Visualization of Dehazing Algorithms. (**a**) Foggy images. (**b**) Images after dehazing.

In order to verify the effectiveness of the dehazing algorithm proposed in this paper for the detection of insulator umbrella disc shedding in foggy images, the dehazing algorithm proposed in this paper was combined with the target detection algorithm, and the obtained detection results are shown in Figure 12.

As shown in Figure 12, the accuracy and recall of the model proposed in this paper were better than other models. It can be seen that after adding the defogging model, the accuracy and recall rate of the insulator umbrella disc shedding detection of the other models were significantly improved. Among them, the accuracy rate of the model increased by 0.08 on average, and the recall rate increased by 0.06 on average. This is because the clear image obtained by the dehazing algorithm was more conducive to the extraction of the features, thereby improving the detection effect. As shown in Figure 13, after adding the defogging algorithm, the detection effect of the FA-SSD model was significantly improved.

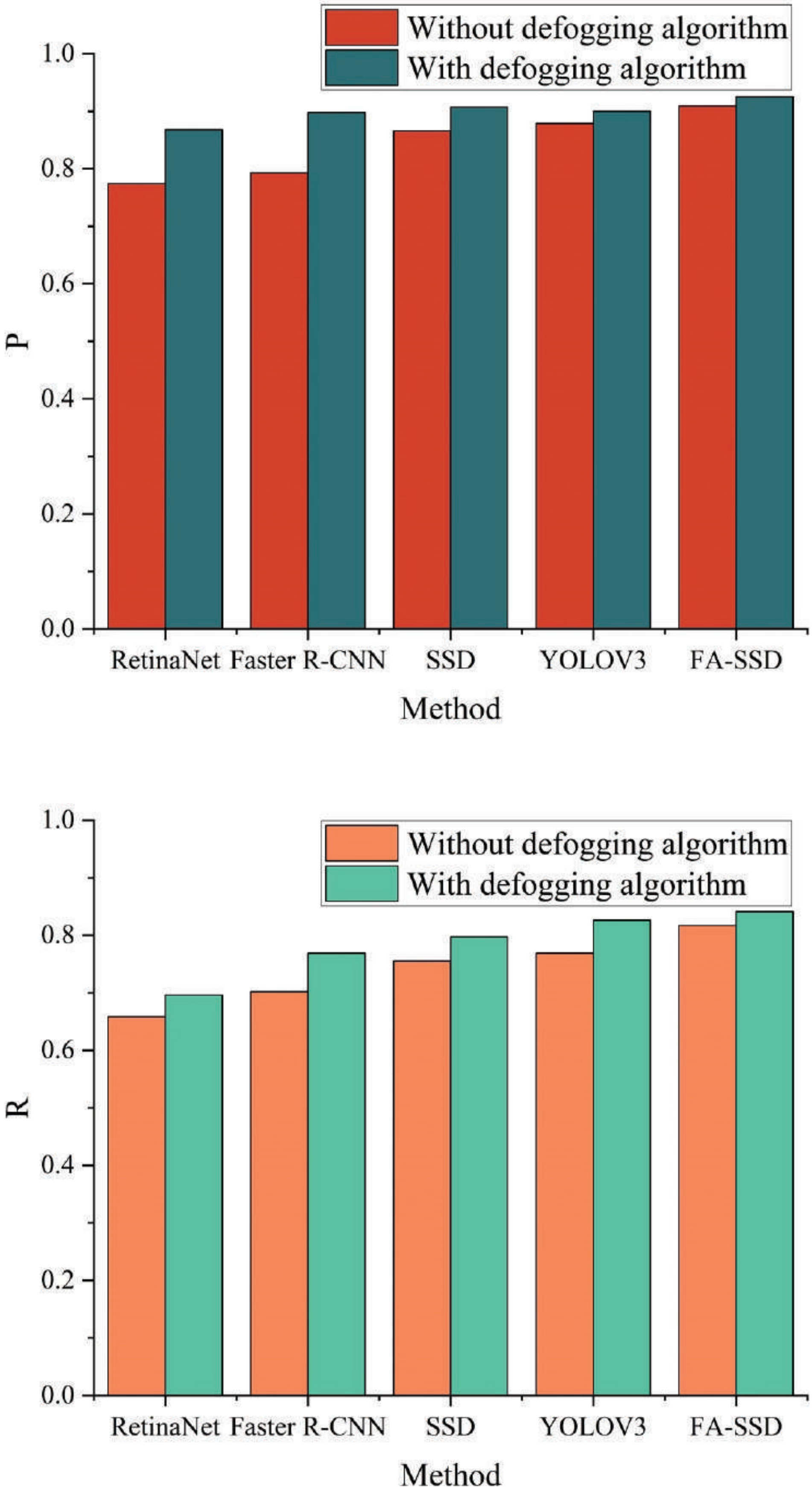


Figure 12. Experimental results before and after adding the defogging algorithm.

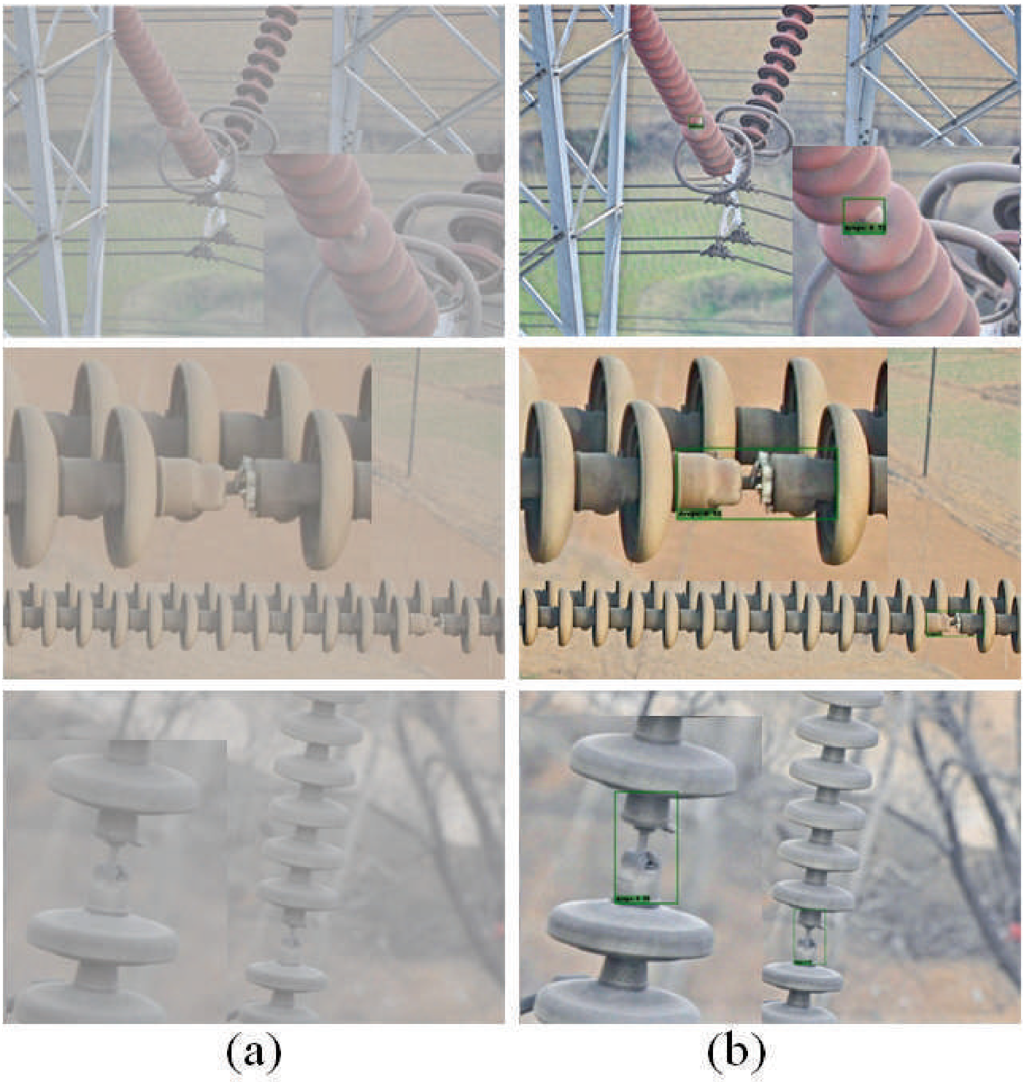


Figure 13. Visualization results of the FA-SSD and FA-SSD with defogging algorithm. (**a**) FA-SSD. (**b**) FA-SSD with defogging algorithm.

4. Discussion

On the basis of realizing the defect detection of insulators with foggy images, combined with the high-speed transmission advantages of 5G technology, real-time detection of insulator defects can be realized, and the necessary processing methods can be taken in time to reduce insulator failures. Compared with the traditional manual inspection, the method in this paper can reduce labor, material resources, and the influence of subjective factors; compared with the currently used UAV inspection, the method in this paper is more in real time. In the context of China's vigorous promotion of a smart grid, this research has important practical significance and good development prospects.

In the future, our research will have the following three aspects. First, we will examine more dehazing algorithms, such as the latest semi-supervised [39] or unsupervised [40] frameworks. Second, we will collect more fogged images of insulators and conduct a joint training strategy to combine image dehazing with defect detection [41]. Third, we will study the defect detection of insulators under a series of complex weather conditions such as sand, rain, and snow and devote ourselves to solving the problem of the defect detection of transmission lines in complex weather, so as to realize all-weather real-time monitoring of transmission lines.

5. Conclusions

Aiming to solve the difficulty of fully extracting effective features from foggy insulator images, as well as the small and difficult to detect proportion of umbrella disk shedding in an image, this paper proposed a detection method for insulator umbrella disk shedding defects that combined a dehazing algorithm and FA-SSD. Through the two-stage algorithm of dehazing and detection, the accurate detection of the insulator umbrella disk shedding in a foggy image was realized. This paper is the first to detect the defects in transmission lines with foggy images, which provides a solution for all-weather monitoring of transmission lines under complex weather conditions.

Author Contributions: Conceptualization, R.X. and X.C.; methodology, K.Y. and X.W.; software, J.W.; validation, R.X., X.C. and J.W.; formal analysis, K.Y.; investigation, X.W.; resources, Y.Z.; data curation, X.W.; writing—original draft preparation, K.Y.; writing—review and editing, R.X.; visualization, J.W.; supervision, X.C.; project administration, Y.Z.; funding acquisition, R.X. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the Science and Technology Project of State Grid Hebei Information and tTelecommunication Branch: Research projects on key technologies and typical application scenarios based on “5G+AR+AI” (No.SGHEXT00SJJS2100036), and this research was funded by the National Natural Science Foundation of China Joint Fund Project Key Support Project: Robot semantic perception and interaction collaboration based on cross-spectral multisensory information fusion in visibility-limited environments (project number U21A20486).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Special thanks are given to the Information and Communication Branch, State Grid Hebei Electric Power and the Department of Automation and North China Electric Power University.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

5G	Fifth Generation Mobile Communication Technology
AI	Artificial Intelligence
HD	High Definition
SSD	Single Shot MultiBox Detector
NMS	Non-maximum suppression
YOLO	You Only Look Once
CNN	Convolutional Neural Network
FA-SSD	SSD Combining Feature Fusion and Attention Mechanism
FPN	Feature Pyramid Network
AR	Augmented Reality
GPU	Graphics Processing Unit
P	Precision
R	Recall
TP	True Positive
FP	False Positive
FN	False Negative
UAV	Unmanned Aerial Vehicle

References

1. Asprou, M.; Kyriakides, E.; Albu, M.M. Uncertainty bounds of transmission line parameters estimated from synchronized measurements. *IEEE Trans. Instrum. Meas.* **2018**, *68*, 2808–2818. [CrossRef]
2. Zhao, Z.; Qi, H.; Qi, Y.; Zhang, K.; Zhai, Y.; Zhao, W. Detection method based on automatic visual shape clustering for pin-missing defect in transmission lines. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 6080–6091. [CrossRef]
3. Park, K.C.; Motai, Y.; Yoon, J.R. Acoustic fault detection technique for high-power insulators. *IEEE Trans. Ind. Electron.* **2017**, *64*, 9699–9708. [CrossRef]
4. Zhai, Y.; Wang, D.; Zhang, M.; Wang, J.; Guo, F. Fault detection of insulator based on saliency and adaptive morphology. *Multimed. Tools Appl.* **2017**, *76*, 12051–12064. [CrossRef]
5. Xia, H.; Yang, B.; Li, Y.; Wang, B. An Improved CenterNet Model for Insulator Defect Detection Using Aerial Imagery. *Sensors* **2022**, *22*, 2850. [CrossRef]
6. Wen, Q.; Luo, Z.; Chen, R.; Yang, Y.; Li, G. Deep learning approaches on defect detection in high resolution aerial images of insulators. *Sensors* **2021**, *21*, 1033. [CrossRef]
7. Deng, C.; Wang, S.; Huang, Z.; Tan, Z.; Liu, J. Unmanned Aerial Vehicles for Power Line Inspection: A Cooperative Way in Platforms and Communications. *J. Commun.* **2014**, *9*, 687–692. [CrossRef]
8. Zhai, Y.; Yang, X.; Wang, Q.; Zhao, Z.; Zhao, W. Hybrid knowledge r-cnn for transmission line multifitting detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–12. [CrossRef]
9. Zhang, C.; Ueng, Y.L.; Studer, C.; Burg, A. Artificial intelligence for 5G and beyond 5G: Implementations, algorithms, and optimizations. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2020**, *10*, 149–163. [CrossRef]
10. Mahmood, A.; Beltramelli, L.; Abedin, S.F.; Zeb, S.; Mowla, N.; Hassan, S.A.; Sisinni, E.; Gidlund, M. Industrial IoT in 5G-and-beyond networks: Vision, architecture, and design trends. *IEEE Trans. Ind. Inform.* **2021**, *18*, 4122–4137. [CrossRef]
11. Liu, X.; Li, Y.; Shuang, F.; Gao, F.; Zhou, X.; Chen, X. Issd: Improved ssd for insulator and spacer online detection based on uav system. *Sensors* **2020**, *20*, 6961. [CrossRef] [PubMed]
12. Stark, J.A. Adaptive image contrast enhancement using generalizations of histogram equalization. *IEEE Trans. Image Process.* **2000**, *9*, 889–896. [CrossRef] [PubMed]
13. Liu, X.; Zhang, H.; Cheung, Y.m.; You, X.; Tang, Y.Y. Efficient single image dehazing and denoising: An efficient multi-scale correlated wavelet approach. *Comput. Vis. Image Underst.* **2017**, *162*, 23–33. [CrossRef]
14. Li, C.; Tang, S.; Kwan, H.K.; Yan, J.; Zhou, T. Color correction based on cfa and enhancement based on retinex with dense pixels for underwater images. *IEEE Access* **2020**, *8*, 155732–155741. [CrossRef]
15. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353.
16. Zhou, F.; Meng, X.; Feng, Y.; Su, Z. SNPD: Semi-Supervised Neural Process Dehazing Network with Asymmetry Pseudo Labels. *Symmetry* **2022**, *14*, 806. [CrossRef]
17. Chen, J.; Yang, G.; Ding, X.; Guo, Z.; Wang, S. Robust detection of dehazed image via dual-stream CNNs with adaptive feature fusion. *Comput. Vis. Image Underst.* **2022**, *217*, 103357. [CrossRef]
18. Zhao, W.; Zhao, Y.; Feng, L.; Tang, J. Attention Enhanced Serial Unet++ Network for Removing Unevenly Distributed Haze. *Electronics* **2021**, *10*, 2868. [CrossRef]
19. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
20. Gao, G.; Cao, J.; Bao, C.; Hao, Q.; Ma, A.; Li, G. A Novel Transformer-Based Attention Network for Image Dehazing. *Sensors* **2022**, *22*, 3428. [CrossRef]
21. Zhang, Z.; Huang, S.; Li, Y.; Li, H.; Hao, H. Image Detection of Insulator Defects Based on Morphological Processing and Deep Learning. *Energies* **2022**, *15*, 2465. [CrossRef]
22. 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]
23. She, L.; Fan, Y.; Wang, J.; Cai, L.; Xue, J.; Xu, M. Insulator Surface Breakage Recognition Based on Multiscale Residual Neural Network. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–9. [CrossRef]
24. Zhang, X.; Zhang, Y.; Liu, J.; Zhang, C.; Xue, X.; Zhang, H.; Zhang, W. InsuDet: A Fault Detection Method for Insulators of Overhead Transmission Lines Using Convolutional Neural Networks. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–12. [CrossRef]
25. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
26. Zhao, W.; Xu, M.; Cheng, X.; Zhao, Z. An Insulator in Transmission Lines Recognition and Fault Detection Model Based on Improved Faster RCNN. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–8. [CrossRef]
27. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. Available online: https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html (accessed on 17 May 2022). [CrossRef]
28. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* **2017**, *36*, 2117–2125.

29. Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; Zhang, L. Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions. *arXiv* **2021**, arXiv:2112.08088.
30. He, Y.; Liu, Z. A Feature Fusion Method to Improve the Driving Obstacle Detection Under Foggy Weather. *IEEE Trans. Transp. Electrif.* **2021**, *7*, 2505–2515. [CrossRef]
31. Hassaballah, M.; Kenk, M.A.; Muhammad, K.; Minaee, S. Vehicle detection and tracking in adverse weather using a deep learning framework. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 4230–4242. [CrossRef]
32. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 August 2016; Springer: Berlin/Heidelberg, Germany, 2016, pp. 21–37.
33. Chen, Z.; Wang, Y.; Yang, Y.; Liu, D. PSD: Principled synthetic-to-real dehazing guided by physical priors. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7180–7189.
34. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.* **2018**, *28*, 492–505. [CrossRef] [PubMed]
35. Li, Z.; Zhou, F. FSSD: Feature fusion single shot multibox detector. *arXiv* **2017**, arXiv:1712.00960.
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
37. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
38. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
39. Li, Y.; Chang, Y.; Gao, Y.; Yu, C.; Yan, L. Physically Disentangled Intra-and Inter-Domain Adaptation for Varicolored Haze Removal. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 5841–5850.
40. Yuntong, Y.; Changfeng, Y.; Yi, C.; Lin, Z.; Xile, Z.; Luxin, Y.; Yonghong, T. Unsupervised Deraining: Where Contrastive Learning Meets Self-similarity. *arXiv* **2022**, arXiv:2203.11509.
41. Li, Y.; Chang, Y.; Yu, C.; Yan, L. Close the Loop: A Unified Bottom-Up and Top-Down Paradigm for Joint Image Deraining and Segmentation. 2022. Available online: https://www.aaai.org/AAAI22Papers/AAAI-678.LiY.pdf (accessed on 17 May 2022).

sensors

MDPI

Article

Two-Level Model for Detecting Substation Defects from Infrared Images

Bing Li, Tian Wang, Zhedong Hu, Chao Yuan * and Yongjie Zhai

Department of Automation, North China Electric Power University, Baoding 071003, China
* Correspondence: chaoyuan@ncepu.edu.cn; Tel.: +86-134-0036-7541

Abstract: Training a deep convolutional neural network (DCNN) to detect defects in substation equipment often requires many defect datasets. However, this dataset is not easily acquired, and the complex background of the infrared images makes defect detection even more difficult. To alleviate this issue, this article presents a two-level defect detection model (TDDM). First, to extract the target equipment in the image, an instance segmentation module is constructed by training from the instance segmentation dataset. Then, the target equipment is segmented by the superpixel segmentation algorithm into superpixels according to obtain more details information. Next, a temperature probability density distribution is constructed with the superpixels, and the defect determination strategy is used to recognize the defect. Finally, experiments verify the effectiveness of the TDDM according to the defect detection dataset.

Keywords: infrared image; substation equipment; defect detection; superpixel segmentation; temperature probability density

Citation: Li, B.; Wang, T.; Hu, Z.; Yuan, C.; Zhai, Y. Two-Level Model for Detecting Substation Defects from Infrared Images. *Sensors* **2022**, *22*, 6861. https://doi.org/10.3390/s22186861

Academic Editor: Jiayi Ma

Received: 18 August 2022
Accepted: 6 September 2022
Published: 10 September 2022

1. Introduction

Substation equipment is an essential part of the power system [1]. Once defects exist in operating equipment, an abnormal temperature usually occurs at the defective parts, triggering thermal failures that can lead to local equipment burnout or even more severe electric power accidents [2]. Therefore, timely and accurate detection of defects in substation equipment is of great significance to the safety and stability of a power system.

Many methods have been studied for defects detection in substation equipment, including dielectric loss measurement [3], UHF (ultra-high frequency) method [4], FDR (frequency domain reflectometry) method [5], and infrared image-based methods [6,7]. The dielectric loss measurement requires off-line preventive testing, which will delay the operation of substation equipment. The complexity of the UHF method makes directly locating defective regions difficult. The FDR method is sensitive only to defects caused by moisture. Early infrared image-based methods for detecting thermal defects in substation equipment require manual intervention, which is time-consuming and costly. However, with the development of smart grids and the successful application of substation inspection robots, a large number of on-site infrared images needed to be inspected urgently. Intelligent defect detection methods have emerged based on computer vision.

Due to the redundant background and the densely packed targets, applying automatic intelligent defect detection methods directly is difficult. Thus, extracting the target equipment in the complex infrared images is required first. Early researchers studied the methods using traditional digital image processing techniques, including threshold-based, region-based, and edge-based methods. Threshold-based methods separate the foreground from the background of an image by selecting a suitable threshold [8], which is simple and efficient but susceptible to noise interference, causing poor robustness. A typical region-based method is the watershed algorithm [9]. It uses the local minima of the image gradient to form a specific region to segment different image parts. However, it is sensitive to the color changes in the object's surface, giving rise to over-segmentation. Edge-based methods extract edge features from the image by edge detection operators such as the Sobel operator [10] and Canny operator [11] to realize the segmentation of an image. Nevertheless, it cannot guarantee the existence of closed, continuous edge regions, and it lacks robustness to noise interference. The recent rapid development of deep learning and imaging technologies has brought innovative ideas for extracted methods from infrared images of substation equipment. Instance segmentation is a classic task in the field of computer vision, which can perform object extraction excellently in images. This task, not only locates and classifies all instances but also segments each instance from the images [12]. Many applications benefit from accurate instance segmentation, including electrical systems [13,14], autonomous driving [15], robotics [16,17], and intelligent transportation systems [18]. Consequently, instance segmentation has become an active research topic in the industry, which benefits its powerful ability of object extraction. Xiong et al. [19] proposed a method based on Mask R-CNN and Bayesian context network to recognize power equipment, which is considered the relationship between objects in a complex background. Ling et al. [20] presented a novel deep learning framework to locate the broken insulators, which is address the problem of low signal-noise-ratio (SNR) setting. To detect the transmission line, a transmission line detection (TLD) algorithm is proposed [21], which is a multitask deep neural network with branched outputs. The deep learning-based methods show excellent performance to extract the target object.

In the stage of defect detection, some promising methods for detecting defects are feature extraction and convolutional neural networks. The key to feature extraction-based approaches is acquiring target ontology features and using classifiers to recognize the extracted features [22,23]. However, the effectiveness of feature extraction and the selection of classifiers are great dependence on personal experience. Convolutional neural networks focus on detecting target defects through an object detection model [24,25]. Li et al. [26] proposed a method of insulator defect location, which is cascades detection and segmentation networks from two levels. In view of the characteristics of insulator defects, Wang et al. [27] presented an improved network to detect the defect of aerial insulator photos. The above method achieved excellent results in defect detection, but requires numerous defective insulator images to train the DCNN. In reality, the infrared images of defective substation equipment are difficult to acquire, and the performance of DCNN is difficult to guarantee. Implementing defect detection of substation equipment in infrared images is still challenging. In an infrared image, the different parts of the target corresponding to different heat generation characteristics. Thus, the temperature feature of the target is used to estimate temperature probability density distribution, which is used to identify defects by the presented strategy. The proposed defect detection part is an unsupervised learning method and is not limited by the dataset. Before that, the superpixel processing is used to provide more details, those details offer more information for defect detection. Meanwhile, it reduces the complexity and time spent on the model.

This study proposes the TDDM for defect detection in electric power substations, which is used in infrared images of substation equipment, e.g., insulator, current transformer, lightning arrester, bushing and voltage transformer. The main contributions of this paper are as follows.

(1) Inspecting the substation equipment from the infrared images with the redundant background and the densely packed targets directly is difficult. The proposed TDDM extracted the target firstly, and then, defect analysis is conducted on a single instance, which is converted to a two-level detection problem.

(2) Superpixel segmentation is conducted on the extracted target equipment to merge adjacent pixels with similar characteristics. The process is used to provide more details and reduce the complexity of the subsequent detection determination.

(3) Based on a Gaussian kernel function, the temperature probability density distribution of the target equipment is constructed, which is used in a defect determination strategy to find the defective areas in infrared images of the target substation equipment.

(4) The experimental results show that the proposed model accurately detects defects in substation equipment in infrared images.

The remainder of this paper is organized as follows. In Section 2, a novel model for detecting these defects in infrared images is provided, including instance segmentation, superpixel segmentation, and defect determination. Section 3 verifies the performance of the proposed model and discusses the influences of superpixel parameters on the results. Section 4 concludes this work.

2. Procedure for the Proposed Model

The model proposed is designed for automatically detecting defects of substation equipment in infrared images.The model transforms defect detection into a two-level detection problem. First, an instance segmentation algorithm directly extracts the target equipment from infrared images with complex backgrounds. After that, a superpixel segmentation algorithm merges similar characteristics and captures the details of the target equipment. Finally, the defect position is determined. Figure 1 is a flowchart of the proposed TDDM procedure.

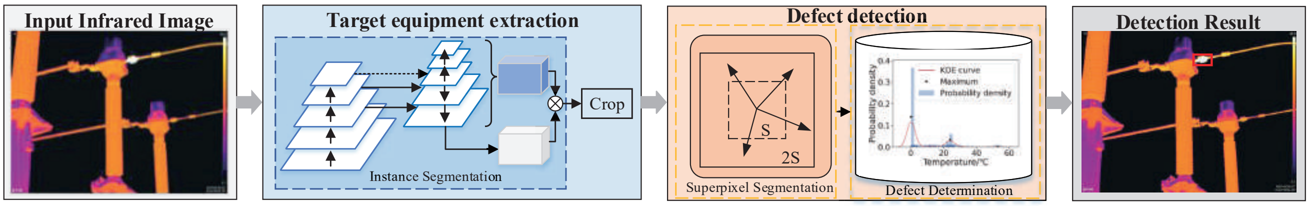


Figure 1. Flowchart of TDDM.

2.1. Instance Segmentation

To detect substation equipment in infrared images, we must first extract the target equipment from the image. Instance segmentation is a basic task of DCNN, which is to extract the target from a complex background and distinguish different instances in the image's foreground [28]. There are three commonly used instance segmentation methods: top-down detection-based methods, bottom-up semantic segmentation-based methods and direct instance segmentation at the pixel level. Top-down detection-based methods perform instance segmentation in a bounding box, such as the Mask R-CNN [29], Mask Scoring R-CNN [30], and YOLACT [31] networks. In bottom-up semantic segmentation-based methods, the pixels are labeled for prediction and clustered [32,33]. The SOLO algorithm [34] performs end-to-end optimization of instance segmentation by mask labeling, which directly segments instances at the pixel level.

This study extracted target equipment images using YOLACT. Its backbone is the feature extraction part used to obtain different resolution feature maps $C_i(i = 2,3,4,5)$ from the input infrared image. The description of specific backbone configuration parameters as shown in Table 1. To obtain the multiscale features, $C_i(i = 2,3,4,5)$ are fused by the horizontal connection with the feature pyramid. Then multiscale features $P_j(j = 3,4,5,6,7)$ are connected to prediction heads for multiscale prediction of objects. There are two branches after the feature pyramid. The one branch predicts the object category, the bounding box, and the mask coefficients; the higher score bounding box is obtained through non-maximum suppression (NMS) [35]. The other branch is a fully convolutional network called protonet, which generates a series of prototype masks based on the feature map P_3. Finally, the prototype masks obtained from protonet are linearly combined with mask coefficients to get m instance $c_m(m \in \{1,2,\cdots,M\}$. We can perform defect analysis on a single instance, removing interference from complex backgrounds.

Table 1. The description of specific backbone configuration parameters.

Layer Name	Structure	Convolution Kernel	Feature Map Size
Input Layer	-	-	640 × 640
Conv1	-	7 × 7 × 64, stride 2	320 × 320
Pool1	Maxpool	3 × 3 × 64, stride 2	160 × 160
Conv2_x	Bottleneck	$\begin{bmatrix} 1 \times 1 \times 64 \\ 3 \times 3 \times 64 \\ 1 \times 1 \times 256 \end{bmatrix} \times 3$	160 × 160
Conv3_x	Bottleneck	$\begin{bmatrix} 1 \times 1 \times 128 \\ 3 \times 3 \times 128 \\ 1 \times 1 \times 512 \end{bmatrix} \times 4$	80 × 80
Conv4_x	Bottleneck	$\begin{bmatrix} 1 \times 1 \times 256 \\ 3 \times 3 \times 256 \\ 1 \times 1 \times 1024 \end{bmatrix} \times 6$	40 × 40
Conv5_x	Bottleneck	$\begin{bmatrix} 1 \times 1 \times 512 \\ 3 \times 3 \times 512 \\ 1 \times 1 \times 2048 \end{bmatrix} \times 3$	20 × 20

2.2. Superpixel Segmentation

In the previous section, the image of each type of target equipment in the infrared image is segmented. In this section, the target equipment is detected individually. To make defect detection easier, we first perform superpixel segmentation. Superpixel segmentation forms superpixels from adjacent pixels in the image of target equipment with similar texture, color, luminance, or other characteristics. Thus, superpixels can be treated as processing units, reducing the complexity and time spent on the subsequent processing of the image [36]. Superpixel segmentation methods are generally classified into graph theory-based methods [37,38] and clustering-based methods [39–41]. Computation of cost functions in graph theory-based methods is complicated. In contrast, clustering-based methods has simple principles and good interpretability. The clustering-based simple linear iterative clustering (SLIC) algorithm obtains uniform compact superpixels, and it has good controllability and low operational complexity than other superpixel algorithms [42].

Inspired by the SLIC algorithm, the proposed model forms adjacent pixels with similar temperature characteristics t and spatial characteristics into superpixels $c_m^n, n \in \{1, 2, \dots, N\}$. Assume that there are I pixels in infrared image c, and the number of superpixels is K. Then the interval between the clustering centers C_k is $S = \sqrt{I/K}$. The pixels $2S$ distance from the clustering center is iteratively clustered based on spatial similarity and temperature similarity, until the maximum number of iterations is reached. The formula for calculating the distance D between pixel i and the cluster center C_k is as follows:

$$D = \sqrt{(\frac{d_t}{m_t})^2 + (\frac{d_{xy}}{m_{xy}})^2}, \tag{1}$$

$$d_t = \sqrt{(t_k - t_i)^2}, \tag{2}$$

$$d_{xy} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2}, \tag{3}$$

where d_t is the temperature distance between pixel i and the cluster center C_k, d_{xy} is the spatial distance between pixel i and the cluster center C_k, m_t and m_{xy} are the maximum temperature distance and spatial distance obtained in the previous iteration, respectively.

Further, the superpixels c_m^n of each instance are obtained, and the corresponding temperature characteristic $T_m^n, n \in \{1, 2, \ldots, N\}$ is calculated by averaging the temperature of pixels in the superpixel. All temperature characteristics of c_m^n lie between the maximum temperature $T_m^{\max}$ and the minimum temperature $T_m^{\min}$, i.e., $T_m^n \in [T_m^{\min}, T_m^{\max}]$.

2.3. Defect Determination

After superpixel segmentation of the target equipment, we inspect the target equipment one by one to determine whether there exist defects. Figure 2 shows the target equipment of the background, normal region, and defective region with different temperature characteristics in the infrared image. The range of temperatures that the defect determination algorithm can identify is even broader than the temperatures range in Figure 2.

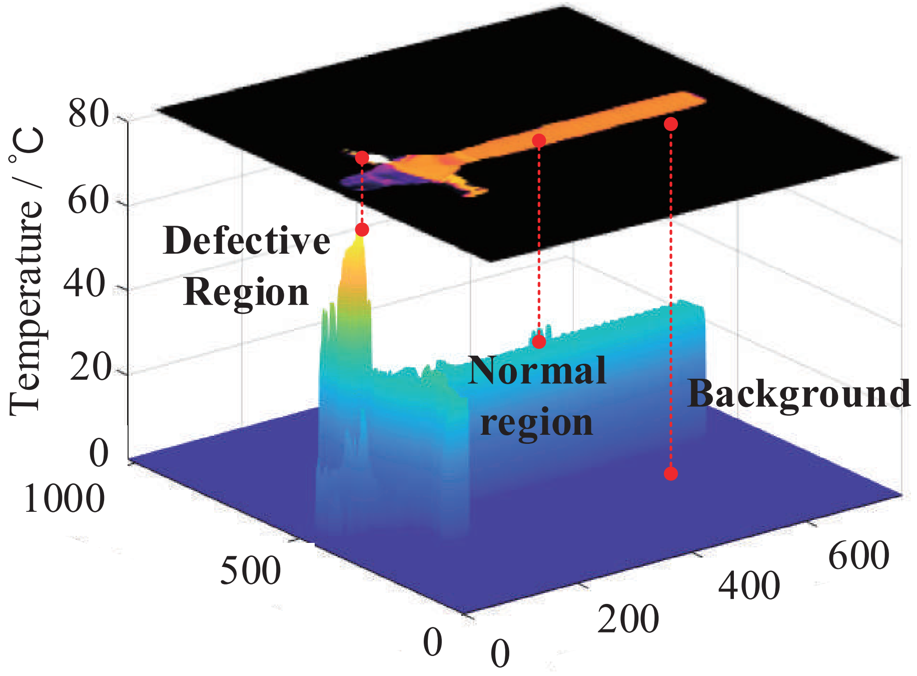


Figure 2. Infrared image of substation equipment and its temperature distribution.

Different temperature characteristics correspond to different temperature probability densities. Thus, we can model the temperature probability density distribution of the instances to determine whether there are defects.

For instance c_m, the temperature probability density T_m^n can be calculated by Equation (4), as shown by the blue histogram in Figure 3.

$$f_m(n) = \frac{T_m^n}{\sum\limits_{i=1}^{N} T_m^i}, n \in [1, N]. \tag{4}$$

However, the temperature probability density data are discretized, which cannot be used directly. Thus, we need to estimate the probability density function to approximate its specific distribution. The common probability density estimation methods include parametric probability density estimation and non-parameter probability density estimation. Kernel density estimation (KDE) [43] is a non-parameter probability density estimation method used to estimate the temperature probability density distribution of the data.

If there is a sufficiently small temperature region $A = [T_m^{A\min}, T_m^{A\max}]$ with bandwidth h, the probability $P_m(A)$ of T_m^n in A is

$$P_m(A) = \int_A f_m(x)dx \approx f_m(x) \int_{T_m^{A\,\min}}^{T_m^{A\,\max}} dx = f_m(x)h. \tag{5}$$

Suppose the probability of Z out of N data falling into region A is

$$P_m(A) = \frac{Z}{N}. \tag{6}$$

Then the temperature probability density becomes

$$f_m(x) = \frac{Z}{Nh}. \tag{7}$$

The kernel density estimation of Equation 7 using the Gaussian kernel function obtains the temperature probability density function of instance c_m.

$$f_m(x) = \frac{1}{Nh}\sum_{j=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-T_m^j}{h}\right)^2}. \tag{8}$$

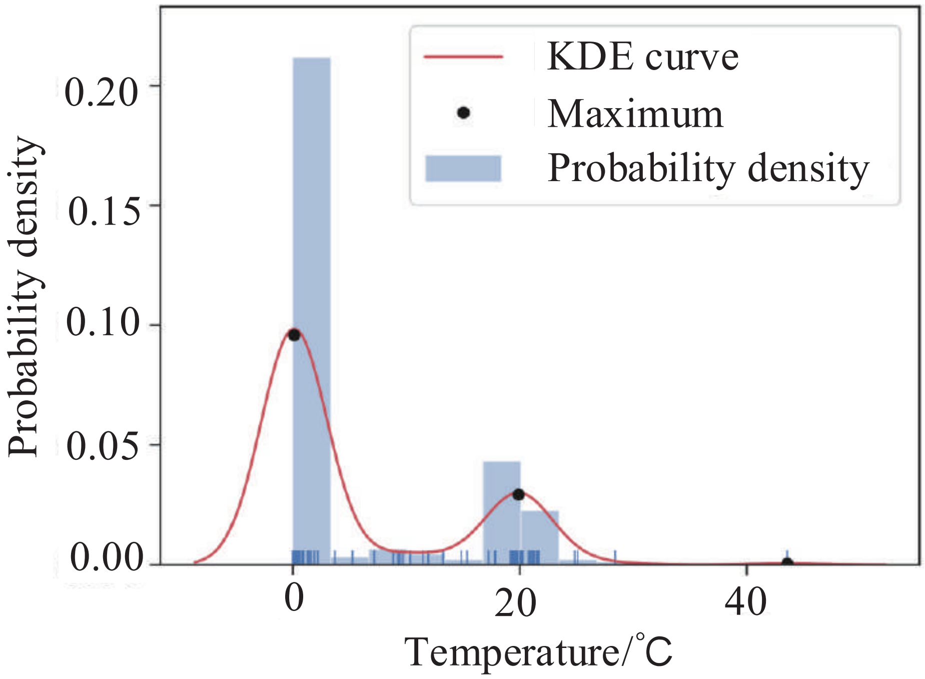


Figure 3. Temperature probability density distribution of c_m.

After that, the temperature probability density distribution function is visualized in Figure 3 by the red curve. The point of local maximum $O_m^q((x_m^q, y_m^q), q = 1, 2, \ldots, Q)$ is obtained, which is denoted by the black dots in Figure 3.

Based on the temperature probability density distribution function, we propose a determination strategy to find defects in infrared images. Due to the different temperature characteristics in the background, normal region, and defective region. Meanwhile, different temperature areas are shown in the temperature probability density distribution. Thus, the presence of O_m^q and $Q \geq 3$ indicate the presence of a defect in the target equipment in this strategy. Then, through the application of the proposed algorithm, x_m^3 is used as the threshold, superpixels c_m^n with temperature characteristics T_m^n higher than x_m^3 are determined to be defective superpixels, automatically. Then, all adjacent defective superpixels are merged to determine the defective regions D_m in instance c_m. Finally, all instances of the infrared image are traversed to obtain all the defective regions automatically. In addition, Algorithm 1 summarizes the whole programming procedure of the proposed TDDM.

Algorithm 1 TDDM

```
Input: Infrared image $c$, Number of superpixels $K$.
Output: All defect regions in the infrared image.
Obtain instance $c_m = \mathrm{Seg}(c), m = 1, 2, \ldots, M$
for $m = 1$to$M$ do
    for $n = 1$to$N$ do
        Compute superpixels $c_m^n$
        Obtain temperature characteristic $T_m^n$
    end for
    Compute temperature probability density distribution $f_m$
    Compute the local maximum $O_m^q(x_m^q, y_m^q)$ of $f_m$, where $q = 1, 2, \ldots, Q$
    if $Q \geq 3$ then
        for $n = 1$to$N$ do
            if $T_m^n > x_m^3$ then
                Determine $c_m^n$ defective
            else
                Determine $c_m^n$ normal
            end if
        end for
        Merge all adjacent defective superpixels to obtain $D_m$
    else
        Output: No defect in the instance.
    end if
end for
```

3. Experiments

3.1. Data Preparation

The experimental infrared images in this article consist of five types of substation equipment, including insulator, current transformer, voltage transformer, bushing, and lightning arrester. The images were captured in a substation by the FLIR T600, where the infrared image resolution is 480 × 360 and the temperature resolution is 0.04 °C. The dataset composition of the substation equipment infrared images in the experiments is illustrated in Figure 4. The instance segmentation dataset is used to train the instance segmentation module, in which the dataset all consists of the normal substation equipment images. The number of each type of equipment is shown in Table 2. In addition, the defect detection dataset is used to evaluate the performance of the TDDM.

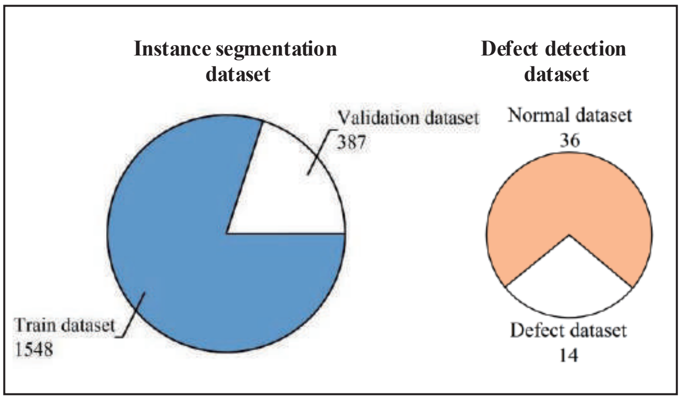


Figure 4. Dataset composition of the substation equipment infrared images.

Table 2. Number of each type of equipment in the instance segmentation dataset.

Equipment	Number
Insulator	919
Current transformer	413
Lightning arrester	289
Bushing	161
Voltage transformer	153

3.2. Instance Segmentation Results and Analysis

The instance segmentation algorithm ran on Ubuntu 18.04LTS with NVIDIA 2080Ti. The training was conducted under the network framework PyTorch through Python3.8, accelerated by CUDA11.2. The current advanced instance segmentation algorithms, including SOLO, Mask R-CNN, MS R-CNN, and YOLACT, were compared experimentally. For training the instance segmentation algorithm, the batch size was set to 2, the SGD optimizer was used, the momentum value was 0.9, the initial learning rate was 0.001, and the number of training iterations was 60 epochs.

To choose the optimal instance segmentation algorithm, a multi-target scene with a complex background was selected for testing. The performance indexes were mAP (mean average precision) and mAR (mean average recall), which are commonly used indexes in the current instance segmentation. SOLO, Mask R-CNN, Mask Scoring R-CNN, and YOLACT were tested on the instance segmentation dataset. The experiment results are shown in Figure 5 and Table 3.

In Table 3, YOLACT had the highest segmentation accuracies compared with the other three algorithms. The values are 67.0% and 74.0% in terms of the mAP and mAR metrics, which were 10.1% and 12.5% higher than the SOLO algorithms. As shown in Figure 5, Figure 5a are the original images and Figure 5f are the ground truth. The four algorithms are intuitively compared in Figure 5b–e, where the white rectangle represents the location of the substation equipment by the model. The pixels of instances belonging to the different categories are marked with different colors. It can be seen from Figure 5 that the YOLACT algorithm accurately located the substation equipment in infrared images and had typically higher quality masks. Thus, this study chose the YOLACT algorithm to segment substation equipment infrared images.

Table 3. Comparison of instance segmentation algorithms.

Method	mAP/%	mAR/%
SOLO	56.9	61.5
Mask R-CNN	63.6	70.4
Mask Scoring R-CNN	65.1	70.9
YOLACT	67.0	74.0

3.3. Compared with Other Superpixel Segmentation Methods

In this section, we compare SLIC [40] to several popular superpixel segmentation algorithms including Felzenszwalb [44], Quickshif [45], and Watershed [46] by the defect detection dataset. The performance of superpixel segmentation is quantitatively evaluated by two metrics, which are boundary recall (BR) and under-segmentation error (UE). BR is the most commonly used metric, which is the percentage of superpixels boundaries coinciding with ground truth boundaries.

$$\mathrm{BR} = \frac{\mathrm{SP}}{\mathrm{GP}}, \tag{9}$$

where SP is the number of segmentation results that meet the condition that should be the ground truth. GP is the total number of the segmentation result. The higher the BR denotes the better performance. UE is the ratio of calculated over-segmented superpixels.

The more approaches zero of the UE, the superpixel approaches the ground truth. UE is defined as follow

$$\mathrm{UE} = -1 + \frac{1}{N}\sum\nolimits_{|u_m \cap u_n| > \omega |u_m|} |u_n|, \tag{10}$$

where u_m and u_n are the pixel sets of the m-th superpixel and ground truth, respectively. ω is set to 0.05 for well established [47]. The lower the UE denotes the better performance.

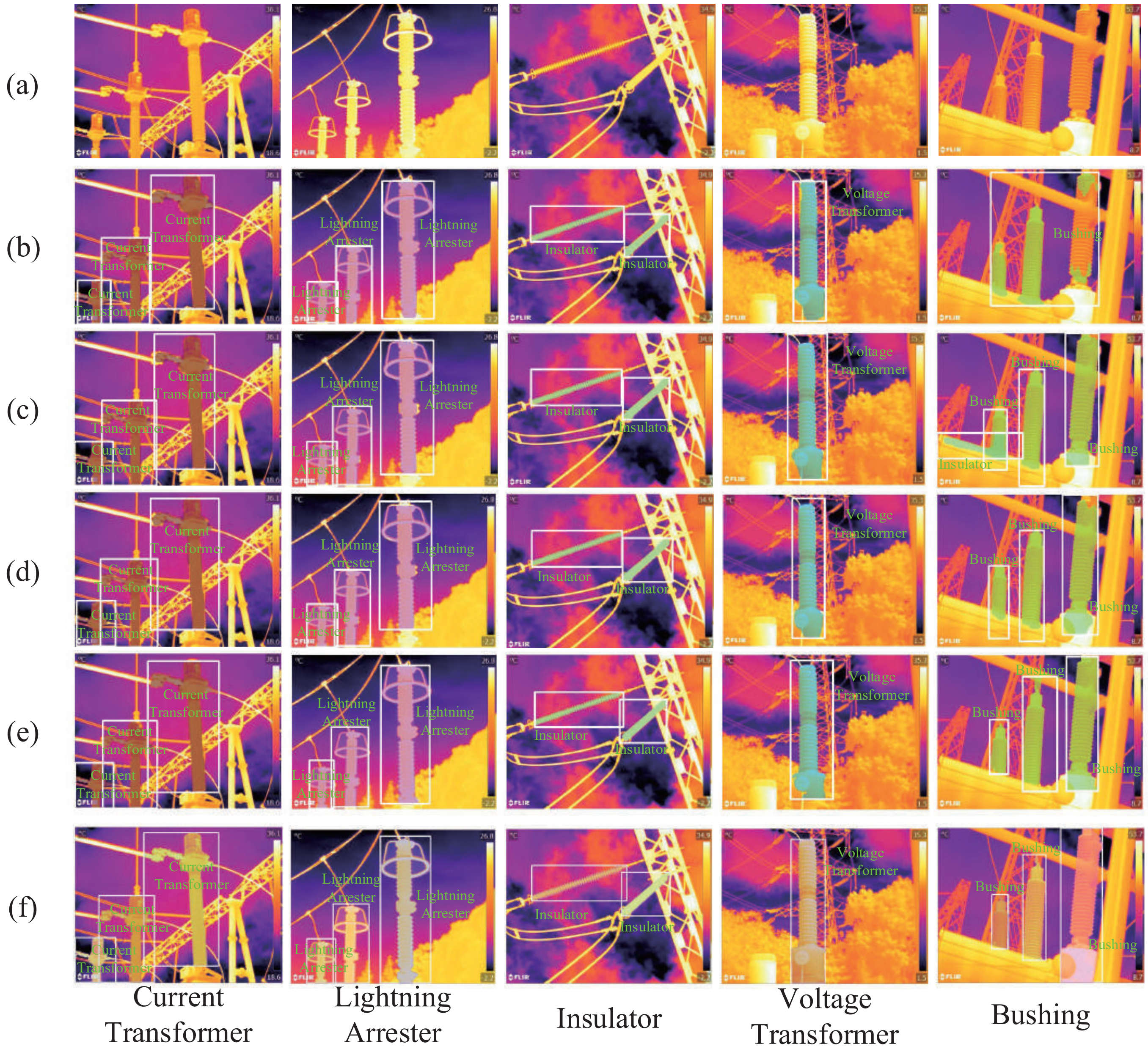


Figure 5. Comparison of segmentation results of different instance segmentation algorithms. (**a**) Original Images. (**b**) SOLO. (**c**) Mask R-CNN. (**d**) Mask Scoring R-CNN. (**e**) YOLACT. (**f**) Ground Truth.

As shown in Figure 6, it illustrates the comparative performance the methods on the defect detection dataset. The numbers of superpixels are set to 250, 500, 750, 1000, 1250, and 1500, respectively. From Figure 6, SLIC, Watershed, and Quickshif all obtain good performance since BR is higher than 0.86. The value of UE in SLIC is the lowest among all methods, this means that better compactness of superpixel segmentation can be achieved.

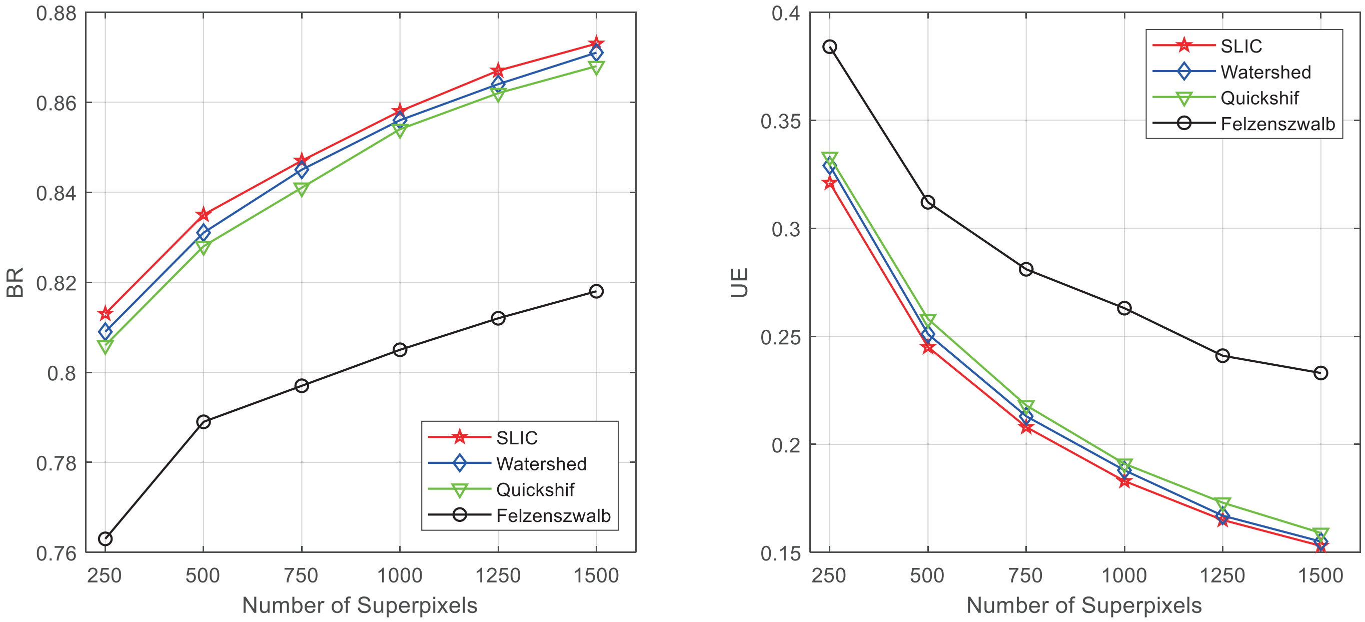


Figure 6. Comparison of superpixel segmentation algorithms performances.

3.4. Defect Detection Results and Analysis

We adopted four widely used metrics for the quantitative evaluations of defect detection performance: precision (P_r), recall (R_e), F_1, and mean running time (mRN). A higher evaluation value indicates better performance, calculated as follows.

$$P_r = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{11}$$

$$R_e = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{12}$$

$$F_1 = \frac{2 * P_r * R_e}{P_r + R_e}, \tag{13}$$

where TP and denote the number of correctly detected defects. TP + FP and TP + FN denote the total number of detected defects and the total number of actual defects, respectively. F_1 is the harmonic mean of P_r and R_e.

We use mean intersection over union (mIoU) to calculate the accuracy of defect region localization. The mIoU is defined as

$$\text{mIoU}(G_T,P_M) = \sum_{m=1}^{M} \frac{\text{Area}(G_T^m \cap P_D^m)}{\text{Area}(G_T^m \cup P_D^m)}, \tag{14}$$

where G_T^m is the ground truth and P_D^m is the predicted region.

To verify the effectiveness of the TDDM, the defect detection datasets are input to the TDDM. To choose the best parameter of the number of superpixels K, we set K from 250 to 3000 with an interval of 250 for the ablation experiments. When $K = 1000$, TDDM has achieved the best defect detection performance. The values of precision, recall, F_1, and mIoU were 0.812, 0.928, 0.866 and 0.831. When $K = 2250$, the model had acceptable precision and recall values performance, but the model running time became longer. Moreover, the running time of TDDM increased with K. Thus, in a word, the selection of an appropriate K is important. Table 4 and Figure 7 show the comparison with a different number of superpixels K to the defect detection dataset.

Table 4. Detection performance for different numbers of *K*.

Number of K	P_r	R_e	F_1	mIoU	mRN
250	0.315	0.428	0.362	0.145	1.47
500	0.190	0.285	0.228	0.098	1.91
750	0.555	0.714	0.624	0.582	2.73
1000	0.812	0.928	0.866	0.831	4.30
1250	0.705	0.857	0.773	0.683	4.36
1500	0.722	0.928	0.812	0.738	5.54
1750	0.764	0.928	0.838	0.796	5.64
2000	0.764	0.928	0.838	0.796	6.82
2250	0.812	0.928	0.866	0.817	6.68
2500	0.764	0.928	0.838	0.796	8.64
2750	0.764	0.928	0.838	0.796	10.42
3000	0.764	0.928	0.838	0.796	11.97

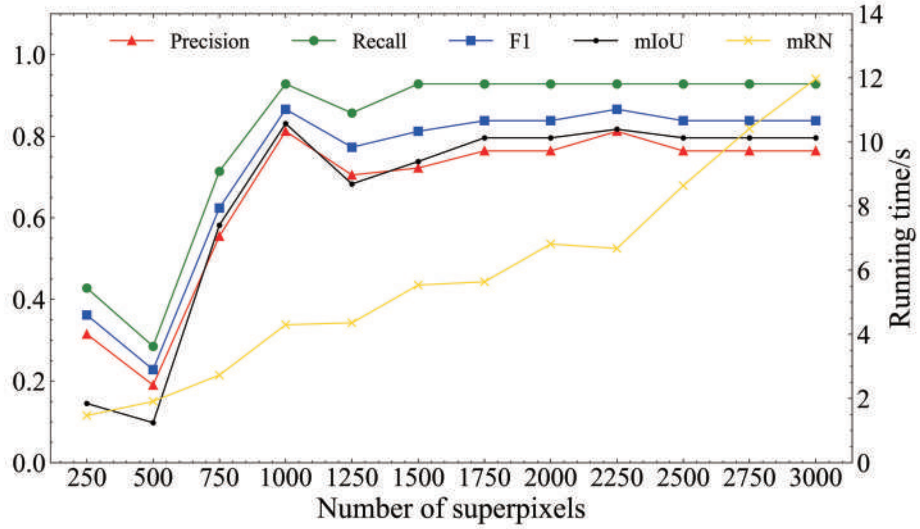


Figure 7. Results of ablation experiments on the number of superpixels.

To evaluate the superiority of the proposed method, some ablation experiments were performed on TDDM. (1) Evaluate the advantage of the superpixel segmentation algorithm (SSA) as a preprocessing for defect detection. (2) Evaluate the advantage of the DCNN + superpixel method for defect detection. Table 5 lists the results of the ablation experiment. As shown in Table 5, the SSA can provide more details and reduce the complexity of the subsequent detection determination. When the objects are extracted firstly by DCNN, the metrics for evaluating accuracy have improved. It indicates that DCNN can overcome the problem of complex background in infrared images. The model achieved superior results when both DCNN and SSA were used. P_r, R_e, F_1, mIoU are reached 0.812, 0.928, 0.866, and 0.831, respectively, which were the highest values.

Table 5. Ablation experiment of TDDM.

DCNN	SSA	P_r	R_e	F_1	mIoU	mRN
✓		0.764	0.928	0.838	0.796	21.34
	✓	0.555	0.714	0.624	0.582	3.62
✓	✓	0.812	0.928	0.866	0.831	4.30

As shown in Figures 8 and 9, the intuitive defect detection process of the TDDM in this paper is on the defect detection dataset.In the intuitive experiment results, the different categories have displayed.

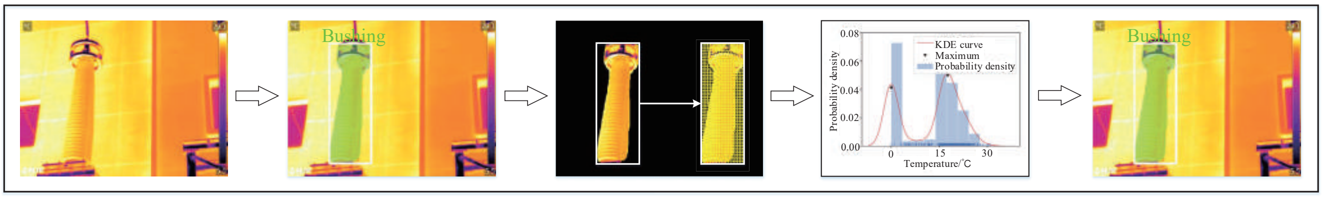
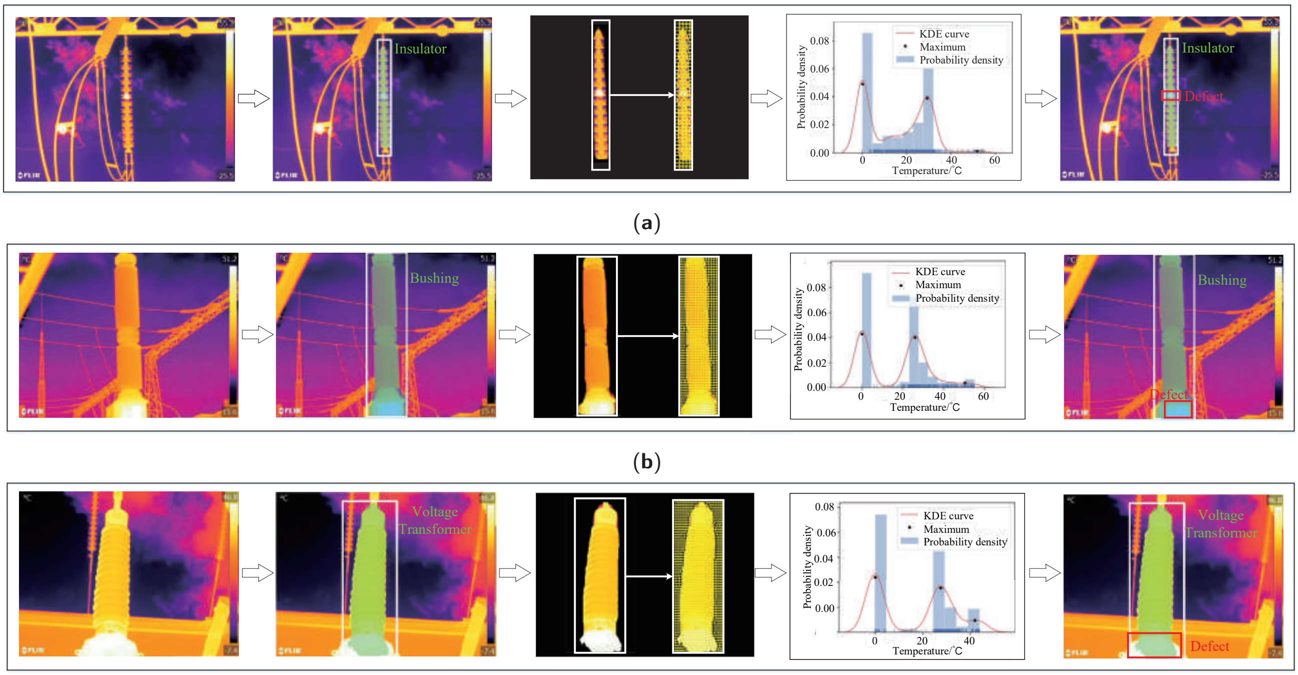


Figure 8. Process of the normal bushing infrared image detection.

Figure 8 shows the process of the normal bushing infrared image detection. In the fourth column, the temperature probability density distribution of the bushing has only two local maxima, which reflects that the substation equipment is no defect. This demonstrates that the TDDM is effectively applied in detecting normal substation equipment.

Figure 9 shows the entire detection flow of the TDDM to the defect-located infrared images. From left to right are the input infrared images, instance segmentation, superpixel segmentation, defect determination, and defect detection results. At the penultimate column, there are three maxima in the temperature probability density distribution of target equipment, representing the equipment exist defect. The target equipment defect detection results are shown in the last column. The white rectangle denotes the target equipment, and the red rectangle represents the location of the defective regions. As can be seen that, TDDM accurately located the defect in substation equipment against a complex background.

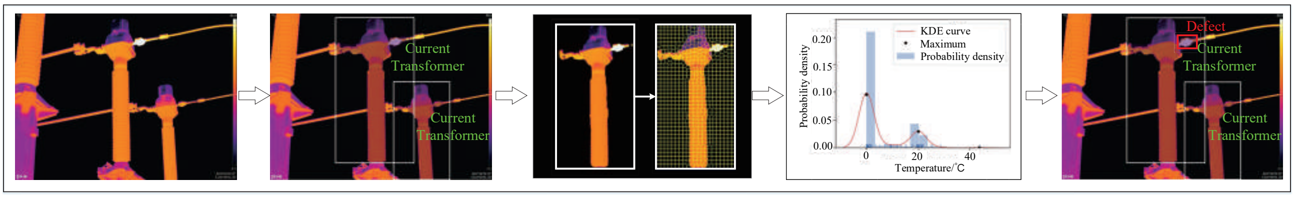


(**a**)

(**b**)

(**c**)

Figure 9. *Cont.*



(**d**)

Figure 9. Process of the defect infrared image detection. (**a**) Insulator. (**b**) Bushing. (**c**) Voltage Transformer. (**d**) Current Transformer.

4. Discussion

In this paper, a two-level model is proposed for the problem of defect detection in substation equipment infrared images. On the basis of extracting substation equipment in the complex background through instance segmentation and superpixel segmentation methods, and realizing defect detection of substation equipment through temperature probability density distribution calculation and adaptive defect detection strategy. Compared with the traditional manual inspection, the proposed method can reduce the resources of labor and material; compared with the end-to-end deep learning method, the presented method in this paper does not require many defect datasets. The operating status of the substation equipment is closely relevant to the stability of the power system, which makes the defects detection of the substation equipment significant.

In the future, our research will not be limited to the substation equipment in this paper and will be applied to other electrical equipment. In fact, according to the characteristic of infrared thermal imaging, the majority of electrical equipment infrared images will show a certain temperature probability density distribution, which is the physical characteristic. The proposed method is based on this characteristic to detect defects precisely. Thus, based on this physical characteristic, we believe the method will be applicable to other cases where may occur defects in electric power, such as medical equipment, airplanes, and industrial equipment.

5. Conclusions

This study proposes a novel defect detection model named TDDM for infrared images of substation equipment. Considering the defective substation equipment infrared images are difficult to acquire, and the data-driven end-to-end model cannot be trained. Thus, the two-level defect detection method is presented. In the proposed TDDM, we take advantage of the fact that the instance segmentation has superior performance to extract the target in the redundant background. Meanwhile, the part of defect detection of TDDM is unsupervised and is not limited by the dataset. Furthermore, we evaluated the proposed model on the defect detection dataset, which accurately detects defects of substation equipment in infrared images. In the future, we would like to combine the RGB information to improve substation inspection tasks. In addition, the technology will be applied to more substation equipment.

Author Contributions: Conceptualization, B.L. and T.W.; methodology, T.W. and Z.H.; software, T.W.; validation, B.L., C.Y. and Y.Z.; formal analysis, Z.H.; investigation, T.W.; resources, Y.Z.; data curation, Y.Z.; writing—original draft preparation, T.W.; writing—review and editing, T.W., Z.H. and C.Y.; visualization, T.W.; supervision, Y.Z.; project administration, B.L.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (NSFC) under grant number U21A20486, 61871182, by the Natural Science Foundation of Hebei Province of China under grant number F2020502009, F2021502008, F2021502013.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Special thanks are given to the Department of Automation and North China Electric Power University.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DCNN	Deep Convolutional Neural Network
TDDM	Two-level Defect Detection Model
UHF	Ultra-high Frequency
FDR	frequency Domain Reflectometry
SNR	Signal-noise-ratio
TLD	Transmission Line Detection
NMS	Non-maximum Suppression
SLIC	Simple Linear Iterative Clustering
KDE	Kernel Density Estimation
mAP	Mean Aaverage precision
mAR	Mean Aaverage Recall
mRN	Mean Running Time
SSA	Superpixel Segmentation Algorithm
TP	True Positive
FP	False Positive
FN	False Negative

References

1. Han, S.; Yang, F.; Jiang, H.; Yang, G.; Zhang, N.; Wang, D. A smart thermography camera and application in the diagnosis of electrical equipment. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–8. [CrossRef]
2. Wang, H.; Zhou, B.; Zhang, X. Research on the remote maintenance system architecture for the rapid development of smart substation in China. *IEEE Trans. Power Deliv.* **2017**, *33*, 1845–1852. [CrossRef]
3. Maina, R.; Tumiatti, V.; Pompili, M.; Bartnikas, R. Dielectric loss characteristics of copper-contaminated transformer oils. *IEEE Trans. Power Deliv.* **2010**, *25*, 1673–1677. [CrossRef]
4. Ozawa, J.; Shindo, K.; Saruta, H.; Yamashita, M.; Takahashi, E. Ultra high frequency electromagnetic wave detector for diagnostic of metal clad switchgear insulation. *IEEE Trans. Power Deliv.* **1994**, *9*, 675–679. [CrossRef]
5. Kwon, G.Y.; Lee, C.K.; Lee, G.S.; Lee, Y.H.; Chang, S.J.; Jung, C.K.; Kang, J.W.; Shin, Y.J. Offline fault localization technique on HVDC submarine cable via time–frequency domain reflectometry. *IEEE Trans. Power Deliv.* **2017**, *32*, 1626–1635. [CrossRef]
6. Zheng, H.; Sun, Y.; Liu, X.; Djike, C.L.T.; Li, J.; Liu, Y.; Ma, J.; Xu, K.; Zhang, C. Infrared image detection of substation insulators using an improved fusion single shot multibox detector. *IEEE Trans. Power Deliv.* **2020**, *36*, 3351–3359. [CrossRef]
7. Wang, B.; Dong, M.; Ren, M.; Wu, Z.; Guo, C.; Zhuang, T.; Pischler, O.; Xie, J. Automatic fault diagnosis of infrared insulator images based on image instance segmentation and temperature analysis. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 5345–5355. [CrossRef]
8. Wang, Y. Improved OTSU and adaptive genetic algorithm for infrared image segmentation. In Proceedings of the 2018 Chinese Control And Decision Conference (CCDC), Shenyang, China, 9–11 June 2018; pp. 5644–5648.
9. Dongmei, W.; Ruyi, W.; Lihua, L. Automatic Detection of Oil Level of Transformer Oil Conservator Based on Infrared Image Segmentation Technology. In Proceedings of the 2011 Fourth International Conference on Intelligent Computation Technology and Automation, Shenzhen, China, 28–29 March 2011; Volume 1, pp. 596–599.
10. Niu, H.; Guo, S.; Xu, T.; Song, T.; Xu, L. Infrared image edge extraction of cable terminal based on improved eight direction Sobel operator. In Proceedings of the 2018 International Conference on Power System Technology (POWERCON), Guangzhou, China, 6–8 November 2018; pp. 3295–3300.
11. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *PAMI-8*, 679–698. [CrossRef]
12. Liu, S.; Qi, X.; Shi, J.; Zhang, H.; Jia, J. Multi-scale patch aggregation (mpa) for simultaneous detection and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3141–3149.
13. Ma, J.; Qian, K.; Zhang, X.; Ma, X. Weakly supervised instance segmentation of electrical equipment based on RGB-T automatic annotation. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9720–9731. [CrossRef]

14. Shu, J.; He, J.; Li, L. MSIS: Multispectral instance segmentation method for power equipment. *Comput. Intell. Neurosci.* **2022**, *2022*, 2864717. [CrossRef]
15. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
16. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 345–360.
17. Hou, H.Q.; Liu, Y.J.; Lan, J.; Liu, L. Adaptive Fuzzy Fixed time Time-varying Formation Control for Heterogeneous Multi-agent Systems with Full State Constraints. *IEEE Trans. Fuzzy Syst.* **2022**, 1–10. [CrossRef]
18. Li, J.; Luo, G.; Cheng, N.; Yuan, Q.; Wu, Z.; Gao, S.; Liu, Z. An end-to-end load balancer based on deep learning for vehicular network traffic control. *IEEE Internet Things J.* **2018**, *6*, 953–966. [CrossRef]
19. Siheng, X.; Yadong, L.; Rui, X.; Ying, D.; Zihan, C.; Yingjie, Y.; Xiuchen, J. Power equipment recognition method based on mask R-CNN and bayesian context network. In Proceedings of the 2020 IEEE Power & Energy Society General Meeting (PESGM), Montreal, QC, Canada, 2–6 August 2020; pp. 1–5.
20. Ling, Z.; Zhang, D.; Qiu, R.C.; Jin, Z.; Zhang, Y.; He, X.; Liu, H. An accurate and real-time method of self-blast glass insulator location based on faster R-CNN and U-net with aerial images. *CSEE J. Power Energy Syst.* **2019**, *5*, 474–482.
21. Li, B.; Chen, C.; Dong, S.; Qiao, J. Transmission line detection in aerial images: An instance segmentation approach based on multitask neural networks. *Signal Process. Image Commun.* **2021**, *96*, 116278. [CrossRef]
22. Dey, D.; Chatterjee, B.; Dalai, S.; Munshi, S.; Chakravorti, S. A deep learning framework using convolution neural network for classification of impulse fault patterns in transformers with increased accuracy. *IEEE Trans. Dielectr. Electr. Insul.* **2017**, *24*, 3894–3897. [CrossRef]
23. Hui, Z.; Fuzhen, H. An intelligent fault diagnosis method for electrical equipment using infrared images. In Proceedings of the 2015 34th Chinese Control Conference (CCC), Hangzhou, China, 28–30 July 2015; pp. 6372–6376.
24. Liu, Y.; Pei, S.; Fu, W.; Zhang, K.; Ji, X.; Yin, Z. The discrimination method as applied to a deteriorated porcelain insulator used in transmission lines on the basis of a convolution neural network. *IEEE Trans. Dielectr. Electr. Insul.* **2017**, *24*, 3559–3566. [CrossRef]
25. Wen, L.; Li, X.; Gao, L.; Zhang, Y. A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Trans. Ind. Electron.* **2017**, *65*, 5990–5998. [CrossRef]
26. Li, X.; Su, H.; Liu, G. Insulator defect recognition based on global detection and local segmentation. *IEEE Access* **2020**, *8*, 59934–59946. [CrossRef]
27. Wang, S.; Liu, Y.; Qing, Y.; Wang, C.; Lan, T.; Yao, R. Detection of insulator defects with improved resnest and region proposal network. *IEEE Access* **2020**, *8*, 184841–184850. [CrossRef]
28. Zhang, H.; Luo, G.; Tian, Y.; Wang, K.; He, H.; Wang, F.Y. A virtual-real interaction approach to object instance segmentation in traffic scenes. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 863–875. [CrossRef]
29. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
30. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask scoring r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6409–6418.
31. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. Yolact: Real-time instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9157–9166.
32. Gao, N.; Shan, Y.; Wang, Y.; Zhao, X.; Yu, Y.; Yang, M.; Huang, K. Ssap: Single-shot instance segmentation with affinity pyramid. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 642–651.
33. Liu, S.; Jia, J.; Fidler, S.; Urtasun, R. Sgn: Sequential grouping networks for instance segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3496–3504.
34. Wang, X.; Kong, T.; Shen, C.; Jiang, Y.; Li, L. Solo: Segmenting objects by locations. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 649–665.
35. Salscheider, N.O. Featurenms: Non-maximum suppression by learning feature embeddings. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 7848–7854.
36. Gaur, U.; Manjunath, B. Superpixel embedding network. *IEEE Trans. Image Process.* **2019**, *29*, 3199–3212. [CrossRef] [PubMed]
37. Liu, M.Y.; Tuzel, O.; Ramalingam, S.; Chellappa, R. Entropy rate superpixel segmentation. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 2097–2104.
38. Ren, X.; Malik, J. Learning a classification model for segmentation. In Proceedings of the Computer Vision, IEEE International Conference on IEEE Computer Society, Nice, France, 13–16 October 2003; Volume 2.
39. Li, Z.; Chen, J. Superpixel segmentation using linear spectral clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1356–1363.
40. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. *Slic Superpixels*; Technical Report; EPFL: Lausanne, Switzerland, 2010.
41. Veksler, O.; Boykov, Y.; Mehrani, P. Superpixels and supervoxels in an energy optimization framework. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 211–224.

42. Levinshtein, A.; Stere, A.; Kutulakos, K.N.; Fleet, D.J.; Dickinson, S.J.; Siddiqi, K. Turbopixels: Fast superpixels using geometric flows. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 2290–2297. [CrossRef] [PubMed]
43. Liu, Z.; Shi, R.; Shen, L.; Xue, Y.; Ngan, K.N.; Zhang, Z. Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut. *IEEE Trans. Multimed.* **2012**, *14*, 1275–1289. [CrossRef]
44. Felzenszwalb, P.F.; Huttenlocher, D.P. Efficient graph-based image segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167–181. [CrossRef]
45. Fulkerson, B.; Soatto, S. Really quick shift: Image segmentation on a GPU. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 350–358.
46. Neubert, P.; Protzel, P. Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 996–1001.
47. Ban, Z.; Liu, J.; Cao, L. Superpixel segmentation using Gaussian mixture model. *IEEE Trans. Image Process.* **2018**, *27*, 4105–4117. [CrossRef]

sensors

MDPI

Article

Zero-Shot Image Classification Method Based on Attention Mechanism and Semantic Information Fusion

Yaru Wang [1], Lilong Feng [1], Xiaoke Song [1], Dawei Xu [1,2,*] and Yongjie Zhai [1]

[1] Department of Automation, North China Electric Power University, Baoding 071003, China
[2] State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: xudawei@ncepu.edu.cn; Tel.: +86-176-2781-0027

Abstract: The zero-shot image classification (ZSIC) is designed to solve the classification problem when the sample is very small, or the category is missing. A common method is to use attribute or word vectors as a priori category features (auxiliary information) and complete the domain transfer from training of seen classes to recognition of unseen classes by building a mapping between image features and a priori category features. However, feature extraction of the whole image lacks discrimination, and the amount of information of single attribute features or word vector features of categories is insufficient, which makes the matching degree between image features and prior class features not high and affects the accuracy of the ZSIC model. To this end, a spatial attention mechanism is designed, and an image feature extraction module based on this attention mechanism is constructed to screen critical features with discrimination. A semantic information fusion method based on matrix decomposition is proposed, which first decomposes the attribute features and then fuses them with the extracted word vector features of a dataset to achieve information expansion. Through the above two improvement measures, the classification accuracy of the ZSIC model for unseen images is improved. The experimental results on public datasets verify the effect and superiority of the proposed methods.

Keywords: image classification; attention mechanism; matrix decomposition; attributes; word vectors

Citation: Wang, Y.; Feng, L.; Song, X.; Xu, D.; Zhai, Y. Zero-Shot Image Classification Method Based on Attention Mechanism and Semantic Information Fusion. *Sensors* **2023**, *23*, 2311. https://doi.org/10.3390/s23042311

Academic Editor: Paweł Pławiak

Received: 4 January 2023
Revised: 15 February 2023
Accepted: 16 February 2023
Published: 19 February 2023

1. Introduction

In recent years, deep learning algorithms have made rapid progress in the image recognition field, but they require significant human and material resources to obtain a sufficient quantity of manually annotated data [1]. In many practical applications, a large quantity of labeled data is difficult to obtain, and the variety of objects is increasing, which requires the computer training process to constantly add new samples and new object types [2,3]. The problem of how to use computers and existing knowledge to classify and identify samples with insufficient or even completely missing label data has become a pressing problem. For this reason, ZSIC [4] was created. It is a technique that trains a learning model to predict and recognize data without class labels (unseen classes) based on some sample data with class labels (seen classes), supplemented by relevant common-sense information or a priori knowledge (auxiliary information) [5,6].

To achieve ZSIC, a popular strategy is to learn the mapping or embedding between the semantic space of classes and the visual space of images based on seen classes and the semantic description of each category. Semantic descriptions of categories usually include attributes [7], word vectors [8], gaze [9], and sentences [10]. At present, the embedded-based methods [11–15] are used to learn visual-to-semantic, semantic-to-visual, or latent intermedium space, so that visual and semantic embedding can be compared in shared space. Then, the unseen classes are classified by nearest neighbor search.

Most of the existing embedding methods, either based on end-to-end convolution neural networks or deep features, emphasize learning the embedding between global visual features and semantic vectors, which leads to two problems [16]. First, there are only slight differences between some features of seen and unseen classes. For some datasets, the inter-class difference is even smaller than the intra-class. Therefore, global image features cannot effectively represent fine-grained information, which is difficult to distinguish in semantic space. Second, compared to visual information, semantic information is not rich enough. The attribute features of categories are usually based on manual annotation, rely on professional knowledge, and are limited by the dimension of visual cognition. The dimension of attribute features is usually not high, and as intermediate auxiliary information, the amount of information is insufficient [17]. The word vectors are mostly obtained through models such as word2vec [18], GloVe [19], or fastText [20]. Relatively speaking, the word vectors may contain more noise and are difficult to combine with human prior knowledge; thus, their interpretability and discriminability are poor. Therefore, the imbalanced supervision from the semantic and visual space can make the learned mapping easily overfitting to seen classes. Inspired by the attention mechanism in the field of natural language processing, a few methods [16,21–23] introduce attention thinking into ZSIC. These methods learn regional embedding of different attributes or similarity measures based on attribute prototypes and learn to distinguish partial features, but they ignore the global features and the information imbalance of semantic and visual space.

Based on the above observation, this paper proposes an improved ZSIC model. The main contributions are as follows:

(1) A feature attention mechanism is designed, and an image feature extraction module based on the attention mechanism is built. The features in different regions of the image are assigned attention weights to distinguish the key and non-key local features, and then the local features are fused with the global features.

(2) A semantic information fusion module based on matrix decomposition is built. The matrix decomposition method is used to transform the binary features of attributes into continuous features and transform their dimensions to be the same as word vectors. In addition, attribute features are fused with word vector features to obtain more accurate and richer fused semantic features as a priori category features.

(3) The improved ZSIC model promotes the alignment of semantic information and visual features. Experiments on the public dataset show that the improved ZSIC model improves image classification accuracy.

2. Related Work

2.1. ZSIC Methods

Recent ZSIC methods focus on learning better visual–semantic embeddings. The core idea is to learn a mapping between the visual and attribute/semantic domains and transfer semantic knowledge from seen to unseen classes according to the similarity measure. Some methods [11,12,24,25] follow the visual-to-semantic mapping direction and align visual features and semantic information in semantic space. However, when high-dimensional visual features are mapped to a low-dimensional semantic space, the shrink of feature space would aggravate the hubness problem [26,27] that in some instances in the high-dimensional space becomes the nearest neighbors of a large number of instances. To tackle these problems, some methods [13,14,28–30] map semantic embedding to visual space and treat the projected results as class prototypes. Shigeto et al. [31] experimentally proved that the semantic-to-visual embedding is able to generate more compact and separative visual feature distribution with the one-to-many correspondence manner, thereby mitigating the hubness issue. Ji et al. [32] also follow the inverse mapping direction from semantic space to visual space and proposed a semantic-guided class imbalance learning model which alleviates the class-imbalance issue in ZSIC. In addition, for the class-imbalance issue, the generative models have been introduced to learn semantic-to-visual mapping to generate visual features of unseen classes [33–37] for data augmentation. Currently, the generative ZSIC is usually based on variational autoencoders (VAEs) [37], generative adversarial nets (GANs) [33], and generative flows [34]. However, the performance of this type of method greatly depends on the quality of generated visual features or images, which is difficult to guarantee, and the mode is prone to mode collapse. Furthermore, to alleviate the hubness issue, common space learning is also employed to learn a common representation space for interaction between visual and semantic domains [15,38,39]. However, these embedded-based models only use the global feature representation, ignoring the fine-grained details in the image, and the training results are not satisfied for the poorly identified features.

2.2. Attention Mechanism

The concept of attention was first introduced into natural language processing tasks. In particular, because soft attention is differentiable and can learn parameters by back-propagation of the model, it has been widely used and developed in computer vision tasks. Zhu et al. [40] applied an attention mechanism in the facial expression recognition task and proposed a cascade attention-based recognition network by a hybrid of the spatial attention mechanism and pyramid feature to improve the accuracy of facial expression recognition under uneven illumination or partial occlusion. Sun et al. and Liu et al. applied an attention mechanism in the semantic segmentation task of remote sensing images. They proposed a multi-attention-based UNet [41] and an attention-based residual encoder [42], respectively. Through channel attention and spatial attention, the capability of fine-grained features was improved. The above attention mechanism includes (i) feature aggregation and (ii) a combination of channel attention (global attention) and spatial attention (local attention), which are common branches of the attention mechanism. In addition, Obeso et al. [43] proved that the global and local attention mechanism in deep neural networks works well with the human visual attention mechanism. Inspired by the above works, several researchers incorporated an attention mechanism into models for ZSIC. For example, Yang et al. [16] proposed a semantic-aligned reinforced attention model to discover invariable features related to class-level semantic attributes from variable intra-class vision information, and thereby to avoid misalignment between visual information and semantic representations. Xu et al. [21] jointly learned discriminative global and local features using only class-level attributes to improve the attribute localization ability of image representation. Chen et al. [22] proposed an attribute-guided transformer network to enhance discriminative attribute localization by reducing the relative geometry relationships among the grid features. Yang et al. [23] proposed to learn prototypes via placeholders and proposed semantic-oriented fine-tuning for preliminary visual–semantic alignment. These methods locate salient regions according to semantic attributes and ignore meaningless information to promote the alignment between a visual space and a semantic space. Compared with these methods, we also consider the combination of local features and global features, as well as the imbalance of information in semantic and visual space.

3. Materials and Methods

The basic embedding-based ZSIC model framework is shown in Figure 1.

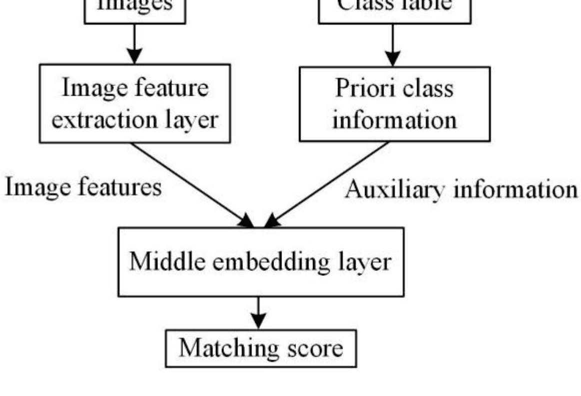


Figure 1. Basic embedding-based ZSIC model framework.

The image feature extraction layer uses a deep CNN to extract image features and input them to a middle embedding layer. A priori class information (auxiliary information) is usually attribute features or word vector features. In the middle embedding layer, the correlation between image features and a priori class information is calculated. Let the total number of seen classes be n and a priori class feature vector of the i-th seen class be β_i, whose dimension is m. In the training stage of the model, the images $\boldsymbol{x}_i$ belonging to the i-th seen class are input into the image feature extraction layer to extract m-dimensional image feature vectors $\boldsymbol{\alpha}_{x_i}$; $\boldsymbol{\alpha}_{x_i}$ and β_i are input into the middle embedding layer, and a relationship similarity $(\boldsymbol{\alpha}_{x_i}, \beta_i)$ between $\boldsymbol{\alpha}_{x_i}$ and β_i is established to obtain the matching score. Cosine distance is used to calculate the matching score. Compared with the European distance, cosine distance is more consistent with the distance calculation form of the high-dimensional vector, and its formula is

$$\text{score} = \text{similarity}(\boldsymbol{\alpha}_{x_i}, \beta_i) = \frac{\sum_{k=1}^{m} a_k b_k}{\sqrt{\sum_{k=1}^{m} a_k{}^2}\sqrt{\sum_{k=1}^{m} b_k{}^2}} \tag{1}$$

where $\boldsymbol{\alpha}_{x_i} = [a_1, a_2, \ldots, a_m]$ and $\beta_i = [b_1, b_2, \ldots, b_m]$.

In order to match the image feature vectors and the prior class feature vectors belonging to the same class as closely as possible, that is, to maximize the matching score, the loss function is used as follows:

$$\text{loss} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{\alpha}_{x_i} \cdot \beta_i}{\|\ \boldsymbol{\alpha}_{x_i}\| \cdot \|\beta_i\|} \tag{2}$$

In the testing stage of the model, the image feature vectors of unseen classes are extracted through the feature extraction layer and then matched with the prior class feature vectors corresponding to each class in the middle embedding layer. When the matching score is the highest, the corresponding class is the prediction class of the input image.

Using the above model framework, the improved embedding-based ZSIC model is shown in Figure 2. Details are as follows.

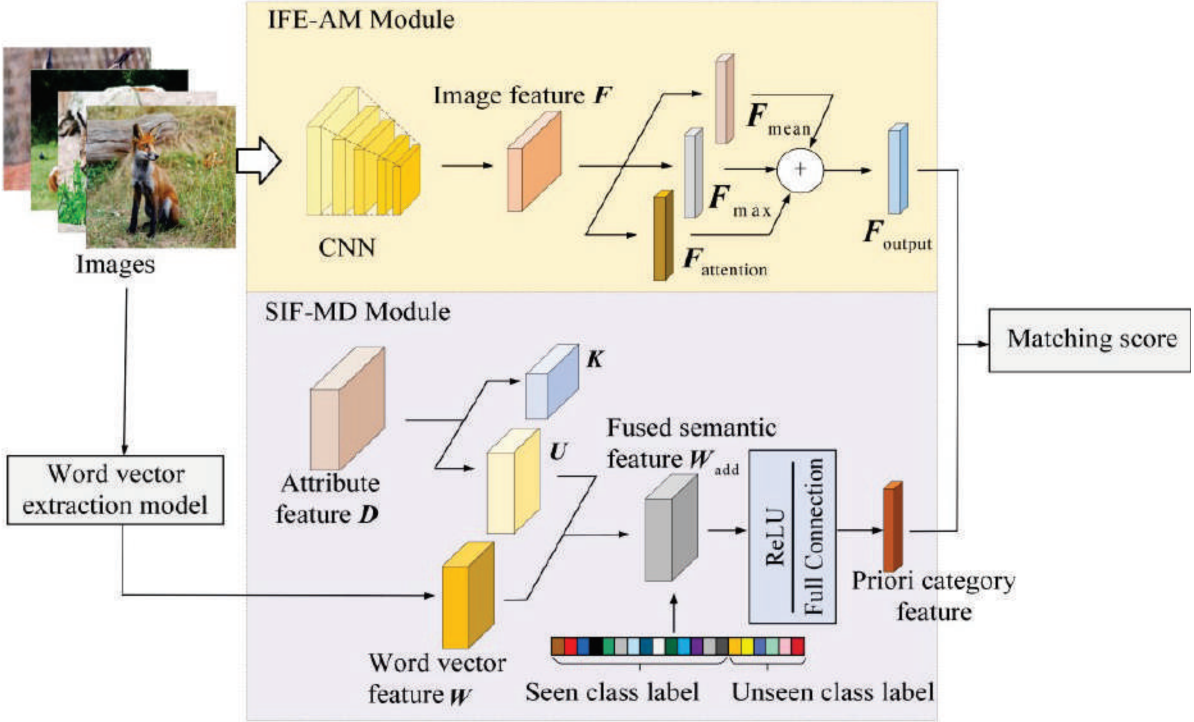


Figure 2. Improved ZSIC model.

3.1. IFE-AM Module

In ZSIC tasks, image features need to be matched with a priori class features, while image features extracted by CNN correspond to a whole image, so they lack discrimination. Therefore, an image feature extraction module based on an attention mechanism (IFE-AM) is constructed (as shown in Figure 2) to focus high-level image features on the key regions of the input image, in order to reduce the deviation from the priori class features and improve the degree of matching. The typical convolutional neural networks VGG-19 and ResNet-34 are taken as examples to illustrate the attention mechanism designed in this paper.

The flowchart of the spatial attention mechanism that weights the feature vector of each position is shown in Figure 3.

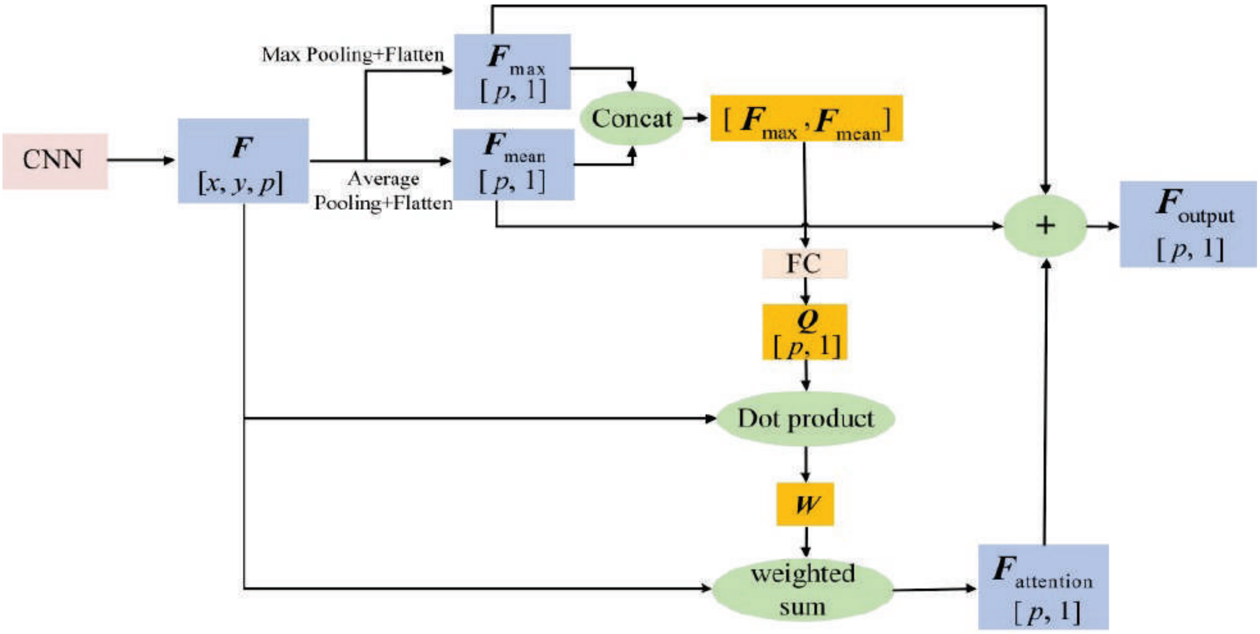


Figure 3. Flowchart of the attention mechanism.

Let the output features of the last layer of the CNN be $\boldsymbol{F}$, with dimension $[x, y, p]$, which contains p channels. For $\boldsymbol{F}$, set window $[x, y]$, and use max pooling and average pooling to obtain two p-dimensional feature vectors $\boldsymbol{F}_{\max}$ and $\boldsymbol{F}_{\text{mean}}$, respectively, and then concatenate them to obtain $[\boldsymbol{F}_{\max}, \boldsymbol{F}_{\text{mean}}]$. Then, $[\boldsymbol{F}_{\max}, \boldsymbol{F}_{\text{mean}}]$ is connected to the fully connected (FC) layer, the hidden layer unit is set as p, and a p-dimensional query vector $\boldsymbol{Q}$ is output for feature selection of the attention mechanism. The feature map of the i-th channel in $\boldsymbol{F}$ is recorded as $\boldsymbol{f}_i$, $i = 1, 2, \ldots, p$, and its size is $x \times y$; the feature vector of the j-th position in $\boldsymbol{F}$ is recorded as $\boldsymbol{l}_j$, $j = 1, 2, \ldots, x \times y$, and its size is $p \times 1$. Calculate the dot product of $\boldsymbol{Q}$ and $\boldsymbol{l}_j$ to obtain the feature weight w_j of the j-th position, and then use the softmax function for normalization to obtain the feature weight matrix $\boldsymbol{W}$. The formula is as follows:

$$\boldsymbol{W} = \text{softmax}\left(w_j\right) = \text{softmax}\left(\text{dot}(\boldsymbol{Q}^T, \boldsymbol{l}_j)\right) \tag{3}$$

The feature values at different positions in $\boldsymbol{f}_i$ are weighted and summed according to the weight matrix $\boldsymbol{W}$, and $\boldsymbol{F}_{\text{attention}}$ is output.

Finally, based on the idea of residual connection, the feature vectors $\boldsymbol{F}_{\max}$, $\boldsymbol{F}_{\text{mean}}$, and $\boldsymbol{F}_{\text{attention}}$ are summed to obtain the final output eigenvector $\boldsymbol{F}_{\text{output}}$.

3.2. SIF-MD Module

ZSIC methods rely on prior class information to complete the transfer from seen classes to unseen classes, so accurate and informative class description information is the key. Currently, the commonly used a priori class description information includes attribute features and word vector features. In order to make the two types of a priori class description information complementary and improve the amount of information, a semantic information fusion module based on matrix decomposition (SIF-MD) is constructed, as shown in Figure 2.

Usually, the dimensions of manually set attribute information is small, and the attribute features are all binary features of 0 or 1, which are relatively sparse and independent; the dimensions of word vectors are relatively large, which are characterized by continuity between [–1, 1]. To carry out information fusion, the matrix decomposition method is used to transform the binary features of attributes into continuous features and transform their dimensions to be the same as word vectors. The architecture diagram of the matrix decomposition of attributes is shown in Figure 4.

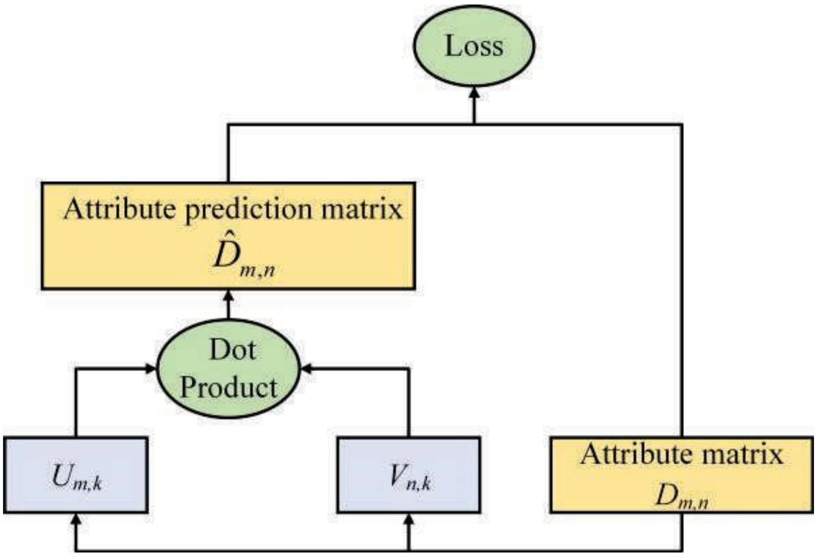


Figure 4. Architecture diagram of the matrix decomposition of attributes.

First, use attribute matrix $\boldsymbol{D}$ $(M \times N)$ to represent n-dimensional attribute vectors of m classes, which is decomposed into $\boldsymbol{U}$ $(M \times K)$ and $\boldsymbol{V}$ $(N \times K)$ with the equation

$$\boldsymbol{D} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} \tag{4}$$

where k is the dimension of the matrix decomposition. Make $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ as close as possible to $\boldsymbol{D}$, that is, fitting attribute feature $\boldsymbol{D}$ through matrix $\boldsymbol{U}$ and matrix $\boldsymbol{V}$. The loss function is the mean squared error MSE (mean squared error) method:

$$\text{loss} = \sum_{i=1}^{M}\sum_{j=1}^{N}\left(D_{i,j} - \hat{D}_{i,j}\right)^2 \tag{5}$$

$$\hat{D}_{i,j} = \boldsymbol{U}_i\boldsymbol{V}_j^{\mathrm{T}} \tag{6}$$

where $\boldsymbol{U}_i$ denotes the vector in the i-th row of matrix $\boldsymbol{U}$, $i = 1, 2, \ldots, M$, and $\boldsymbol{V}_j$ denotes the vector in the j-th row of matrix $\boldsymbol{V}$, $j = 1, 2, \ldots, N$.

To prevent overfitting, the L2 canonical term is added to Formula (5):

$$\text{loss} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(D_{i,j} - \hat{D}_{i,j}\right)^2 + \lambda\left(\|\boldsymbol{U}_i\|_{\mathbf{1}} + \|\boldsymbol{V}_j\|_{\mathbf{1}}\right) \tag{7}$$

Each row in $\boldsymbol{U}$ is a k-dimension vector, which matches the dimension of the word vector of the corresponding class. The matrix $\boldsymbol{U}$ and the word vector matrix $\boldsymbol{W}(m \times k)$ are summed in certain weight proportions as fused semantic features $\boldsymbol{W}_{\text{add}}$, which are given by

$$\boldsymbol{W}_{\text{add}} = \alpha \boldsymbol{W} + (1 - \alpha)\boldsymbol{U} \tag{8}$$

where α is a parameter with a range of [0, 1]; $\boldsymbol{W}_{\text{add}}$ is a fused semantic feature, retaining the content of attribute features and word vector features.

4. Experiment Results

The experiment is based on the 4× 1080Ti GPU server of Ubuntu16.04, the Python 3.6 virtual environment is built through Anaconda, and deep learning frameworks of TensorFlow1.2.0 and Keras2.0.6 are installed.

The top-1 accuracy and top-3 accuracy were used to evaluate the classification results of the zero-shot classification model on the test set. The training set and test set were randomly selected four times to obtain four groups of experimental results, and the average classification accuracy was recorded.

4.1. Dataset

The experiment was conducted based on the Animals with Attributes 2 (AwA2) [27] dataset. AwA2 is a public dataset for attribute-based classification and zero-shot learning, and it is publicly available at http://cvml.ist.ac.at/AwA2, accessed on 9 June 2017. The dataset contains 37,322 images and 50 animal classes, and each class has an 85-dimensional attribute vector. It is a coarse-grained dataset that is medium-scale in terms of the number of images and small-scale in terms of the number of classes. In experiments, we followed the standard zero-shot split proposed in reference [9], that is, 40 classes for training and 10 classes for testing. The training set and test set do not intersect. Among the training set, 13 classes were randomly selected for validation to perform a hyperparameter search.

4.2. Ablation Experiment of IFE-AM Model

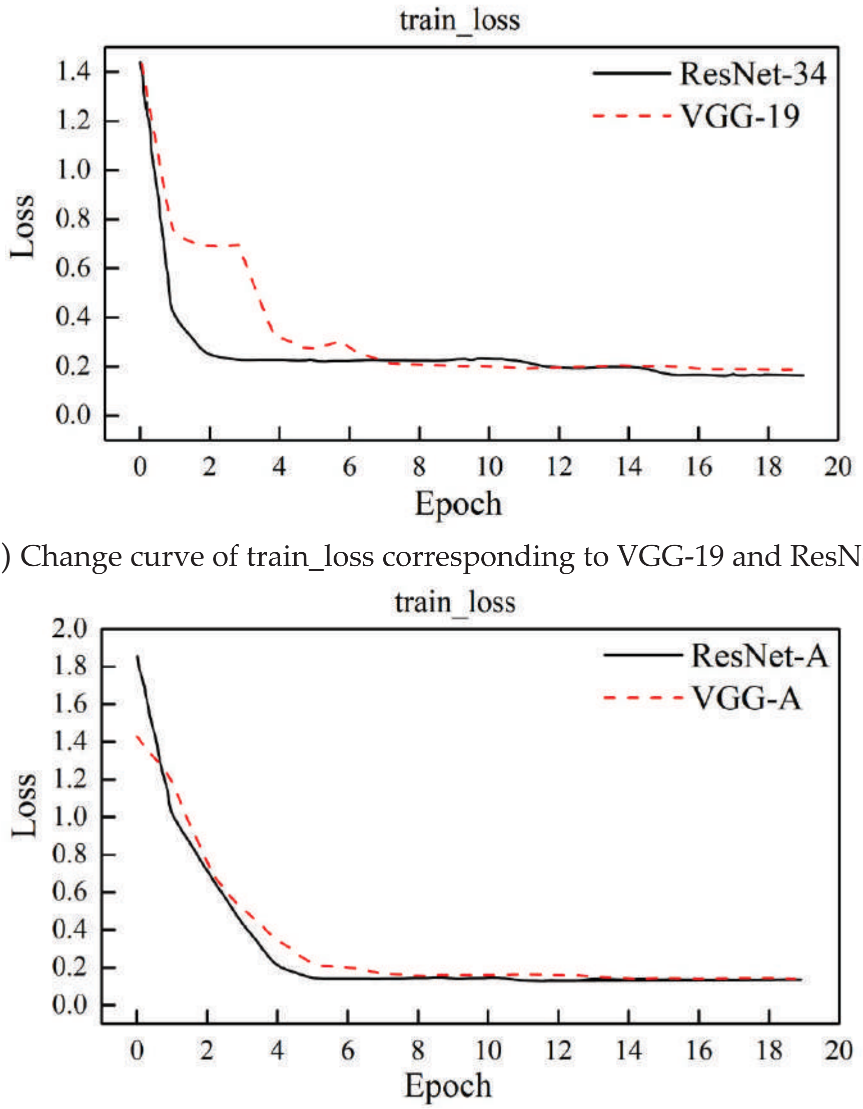
According to the model structure shown in Figure 2, the experiments were conducted with the representative VGG-19 and ResNet-34 as the backbone networks, which are called VGG-A and ResNet-A, respectively. The image features were extracted by the pre-improved and improved networks, and the attribute features of the dataset were used to conduct experiments.

4.2.1. Training Loss and Classification Accuracy

When the model is trained, the training loss is calculated according to Formula (2). Figure 5 shows the change curves of the training loss (train_loss) corresponding to different feature extraction networks.

Table 1 shows the epochs required for training and train_loss values corresponding to different feature extraction networks, as well as the classification accuracy (top-1 and top-3) of the test set.

Figure 5 and Table 1 show that the train_loss of the ResNet-34 model decreases faster than the VGG-19 model. The final train_loss of the VGG-19 and ResNet-34 models tends to be stable, but the train_loss of the ResNet-34 model is lower. From the decreasing trend in train_loss, the train_loss of the VGG-19 model fluctuates greatly, and the decreasing process of train_loss of the ResNet-34 model is more stable. The ResNet-A model is also superior to the VGG-A model in decreasing speed and the stability of train_loss. This shows that the ResNet-34 model with residual connections can realize matching between image features and prior class features faster, better, and more stably. In addition, for both the VGG-A model and ResNet-A model, although their train_loss overall declines slightly slower, their required training epoch and loss value after stabilization are significantly lower than those of the original VGG-19 and ResNet-34 networks. This shows that the IFE-AM module proposed in this paper, as a feature-weighted focusing strategy, improves the model's ability to capture image features in space, thus realizing further fitting of deep features; additionally, the attention mechanism is based on the method of weighted information fusion, which makes the acquisition and update of information more stable, thus achieving a faster and more stable fitting effect.



(**a**) Change curve of train_loss corresponding to VGG-19 and ResNet-34

(**b**) Change curve of train_loss corresponding to VGG-A and ResNet-A

Figure 5. Change curves of train_loss.

Table 1. Test results.

Feature Extraction Network	IFE-AM	Epochs	Train_Loss	Top-1 (%)	Top-3 (%)
VGG-19		17	0.174	40.1	53.1
ResNet-34		16	0.155	41.7	56.1
VGG-A	√	13	0.147	43.2	60.9
ResNet-A	√	5	0.139	43.3	63.9

For the image classification results of the test set, the top-1 and top-3 of the ResNet-34 model are all larger than those of the VGG-19 model, which shows that its residual structure has a good effect on the fitting of deep image features. The top-1 and top-3 of the ResNet-A model are higher than those of the VGG-19 and ResNet-34 models without the attention mechanism, which shows that the attention mechanism can focus the features of spatial attention and effectively improve the generation of image features and the matching effect with prior class features. The accuracies of VGG-A and ResNet-A are similar, but the top-3 of ResNet-A is significantly improved, which shows that the ResNet-A model can obtain more accurate image features in high-dimensional space, making the distance between classes farther, the distance within classes closer, and the matching effect with semantic features better.

4.2.2. Feature Segmentation

According to the model shown in Figure 4, for VGG-A and ResNet-A, the image feature $\boldsymbol{F}_{\text{output}} = \boldsymbol{F}_{\max} + \boldsymbol{F}_{\text{mean}} + \boldsymbol{F}_{\text{attention}}$ is split, and $\boldsymbol{F}_{\max}$, $\boldsymbol{F}_{\text{mean}}$ and $\boldsymbol{F}_{\text{attention}}$ are, respectively, output to the next layer for comparison with $\boldsymbol{F}_{\text{output}}$. The accuracy of the final image classification is shown in Tables 2 and 3.

Table 2. Comparison of different image features in the VGG-A model.

Image Features	Attention	Feature Fusion	Top-1 (%)	Top-3 (%)
$F_{\max}$			39.9	45.0
F_{mean}			40.3	51.1
$F_{\text{attention}}$	√		40.9	51.9
F_{output}	√	√	42.3	60.9

Table 3. Comparison of different image features in the ResNet-A model.

Image Features	Attention	Feature Fusion	Top-1 (%)	Top-3 (%)
$F_{\max}$			39.1	41.1
F_{mean}			41.7	56.1
$F_{\text{attention}}$	√		42.9	61.1
F_{output}	√	√	43.3	63.9

As shown in Tables 2 and 3, the image classification results of the improved ResNet-A model based on the attention mechanism are better than those of the VGG-A model. Whether it is the VGG-A or ResNet-A model, the image classification accuracy corresponding to different image features satisfies $\boldsymbol{F}_{\text{output}} > \boldsymbol{F}_{\text{attention}} > \boldsymbol{F}_{\text{mean}} > \boldsymbol{F}_{\max}$, which verifies the effect of image feature extraction based on the spatial attention mechanism. Inspired by the idea of residual connection, the three features are superposed to obtain $\boldsymbol{F}_{\text{output}}$, which fuses the information of different features and finally obtains the optimal image classification result.

4.3. Ablation Experiment of SIF-MD Module

Since the above experiments verified that ResNet-A and $\boldsymbol{F}_{\text{output}}$ are better, the following further experiments are conducted on these bases. Three models of word2vec, GloVe, and fastText were used to extract the word vector features of each class in the dataset, with a dimension of 256. The attribute features of the dataset were decomposed according to Formulas (4)–(7), and the loss threshold value was set as 0.1. Then, the decomposed attributes were weighted and fused with word vector features extracted by word2vec, GloVe, and fastText, respectively, according to Formula (8). The fusion parameter α was set as [0, 1] and the step size as 0.1.

The image classification experiment of the test set was repeated five times, and the average value of the top-1 was taken. The experimental results corresponding to different word vectors and different fusion parameters α are shown in Table 4. Figure 6 more intuitively shows the changing trend of top-1 accuracy with α when different word vectors are used as auxiliary information.

Table 4. Image classification top-1 accuracy of the test set.

Word Vector	***α***										
	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
word2vec	43.1	43.1	43.1	43.3	43.7	43.8	43.8	44.0	44.3	44.5	44.2
GloVe	43.1	44.3	44.6	44.6	44.6	44.7	45.0	45.1	45.8	45.3	44.7
fastText	43.1	43.0	43.3	43.6	43.2	42.8	42.5	42.5	42.2	42.2	42.1

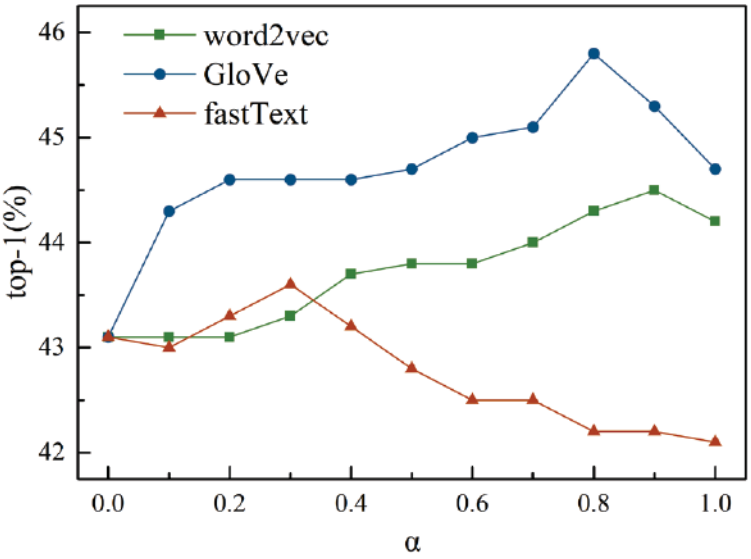


Figure 6. Changing trend of top-1 accuracy of image classification.

As shown in Figure 6, the top-1 accuracy of the word vector extracted by GloVe as prior class features is significantly higher than that extracted by word2vec or fastText. As shown in Table 4, when *α* = 0, that is, only the attribute features are used as the prior class feature, the top-1 accuracy of image classification is 43.1%. When *α* = 1, that is, only word vectors are used as prior class features, the top-1 accuracies corresponding to word2vec and GloVe are 44.2% and 44.7%, respectively, which are better than the results when only attribute features are used, while the top-1 accuracy corresponding to fastText is lower than the results when only attribute features are used. For the word vectors extracted by word2vec, GloVe, and fastText, the fusions with attribute feature all have positive effects. For the word2vec word vector, when the fusion weight *α* = 0.8 and 0.9, the top-1 accuracy is 1.2% and 1.4% higher than that of the attribute vector only and 0.1% and 0.3% higher than that of the word vector only, respectively. For the fastText word vector, when the fusion weight *α* = 0.2, 0.3, and 0.4, the top-1 accuracy is 0.2%, 0.5%, and 0.1% higher than that of the attribute vector only and 1.2%, 1.5%, and 1.1% higher than that of the word vector only, respectively. For the GloVe word vector, when the fusion weight *α* = 0.6, 0.7, 0.8, and 0.9, the top-1 accuracy is 1.9%, 2.0%, 2.7%, and 2.2% higher than that of the attribute vector only and 0.3%, 0.4%, 1.1%, and 0.6% higher than that of the word vector only, respectively. The results show that it is meaningful to fuse attribute features and the word vector features.

5. Discussions

To verify the effectiveness of the method proposed, the method is compared with the baseline model and existing classical models. The baseline model only uses the deep learning network ResNet-34 or VGG-19 to extract image features and uses attributes or word vectors as auxiliary information. The results of the comparative experiment are shown in Table 5 and Figure 7. In the table, “ResNet-34 + attribute” refers to the model that uses ResNet-34 to extract image features and uses attributes as auxiliary information. The image classification results were evaluated with top-1 accuracy. The experimental results of IAP, CONSE, and CMT adopt the results given in references [27,31]. The dataset and the splits of the training set and test set in the experiments of all methods are the same as that of our method, and no methods were pre-trained by large datasets (such as ImageNet).

Table 5. Image classification results of different methods.

	Method	**Top-1 (%)**
1	ResNet-34 + attribute	41.7
2	ResNet-34 + word2vec	42.3
3	ResNet-34 + GloVe	42.7
4	ResNet-34 + fastText	40.6
5	VGG-19 + attribute	40.1
6	VGG-19 + word2vec	40.4
7	VGG-19 + GloVe	41.2
8	VGG-19 + fastText	39.9
9	IAP	35.9
10	CONSE	44.5
11	CMT	37.9
12	ours	45.8

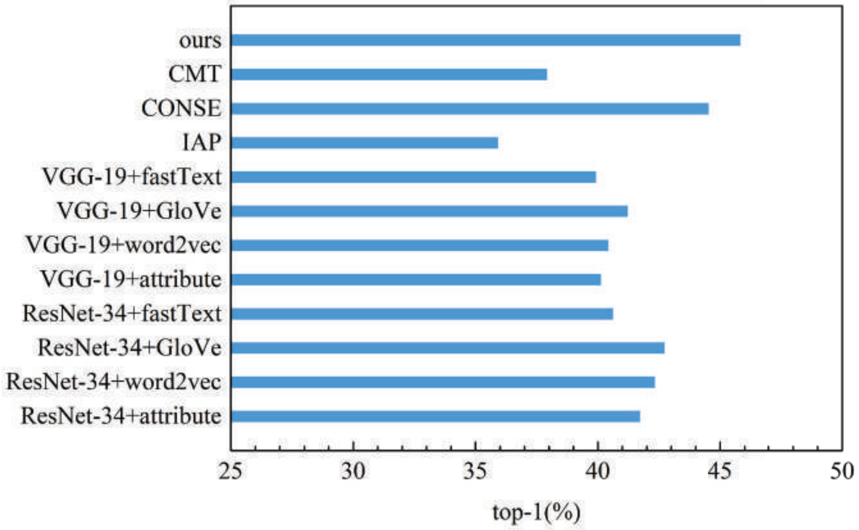


Figure 7. Top-1 accuracy comparison of different methods.

As shown in Table 5 and Figure 7, for the baseline model, the top-1 accuracy of the model using ResNet-34 to extract image features is higher than that of the model using the VGG-19 network; the top-1 accuracy of the model using word vectors extracted by word2vec or GloVe as auxiliary information is higher than that of the model using attributes; and the top-1 accuracy of the “ResNet-34 + GloVe” method is the highest, with a value of 42.7%. The top-1 accuracy of our method is 3.1% higher than that of the “ResNet-34 + GloVe” method. For existing classical methods, IAP detects unseen classes based on attribute transfer between classes, the attribute features are limited by the dimension of visual cognition, and the amount of information is insufficient. CONSE uses CNN to extract image features without distinguishing the importance of different regional features, and only uses word vectors extracted by word2vec as auxiliary information. CMT uses Sparse Coding to extract image features and uses a neural network architecture to learn the word vectors of categories. Although more semantic word representations are learned by using local and global contexts, the discrimination of word vectors is poor, and the imbalanced supervision between semantic features and visual features is still large. Our method assigns attention weights to different regions of the image through the SIF-MD module and strengthens the key features highly related to semantic information. In addition, it alleviates the imbalanced supervision issue between semantic features and visual features through IFE-AM module. These improvements promote the alignment of visual features and semantic information and make the matching degree of the two higher, which is very important for ZSIC. Thus, the top-1 accuracy of our method is 9.9% higher than IAP, 1.3% higher than CONSE, and 7.9% higher than CMT. The above experimental results prove the effectiveness of our method.

6. Conclusions

To improve the accuracy of the ZSIC model based on embedded space, the IFE-AM model and SIF-MD module are constructed in this paper. After the existing CNN is used to extract the image feature map, the max pooling, average pooling, and spatial attention methods are used to obtain three feature vectors, and then they are fused as the final image features. The attribute matrix of the dataset is decomposed to match its dimensions with the extracted word vector, and then the attribute and word vector are weighted and fused as auxiliary information of the improved ZSIC model.

Experiments were conducted on a public dataset. First, the ablation experiment of the IFE-AM model was carried out. The experimental results show that the top-1 and top-3 accuracies corresponding to ResNet-A are 1.6% and 7.8% higher than those of ResNet-34, respectively; the top-1 and top-3 accuracies corresponding to VGG-A are 3.1% and 7.8% higher than those of VGG-19, respectively. Then, the ablation experiment of the SIF-MD module was carried out. The experimental results show that the top-1 accuracies of using fused semantic information as auxiliary information are significantly higher than that of using attribute or word vector alone. Third, comparative experiments were carried out, and the results show that the accuracy of the proposed method is significantly higher than the baseline method and several existing classical methods.

For different types of semantic information, the fusion parameter is not fixed and needs to be determined by experiments. How to derive the value of the fusion parameter in theory is our future work. A small- to medium-sized dataset is considered in our work, and larger data scenarios will be explored in the future.

Author Contributions: Conceptualization, Y.W. and D.X.; methodology, Y.W. and L.F.; software, L.F.; validation, L.F. and X.S.; data curation, X.S.; writing—original draft preparation, L.F.; writing—review and editing, X.S. and Y.Z.; supervision, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the following projects: Natural Science Foundation Youth Science Fund Project of Hebei Province (No. F2021502008): "Research on Zero-shot Fault Detection Method for Transmission Lines with Domain Knowledge"; the General Project of the Fundamental Research Funds for the Central Universities (No. 2021MS081): "Research on Transmission Lines Fault Detection Method with Multimodal Information"; and Open Research Fund of The State Key Laboratory for Management and Control of Complex Systems (No. 20220102): "Research on Interactive Control Method of Snake-Like Manipulator".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Special thanks are given to North China Electric Power University.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study, in the collection, analyses or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ZSIC	Zero-shot image classification
CNNs	Convolutional neural networks
IFE-AM	Image feature extraction module based on an attention mechanism
SIF-MD	Semantic information fusion module based on matrix decomposition
AwA2	Animals with Attributes 2
FC	Fully connect

References

1. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef]
2. Sun, X.; Gu, J.; Sun, H. Research progress of zero-shot learning. *Appl. Intell.* **2021**, *51*, 3600–3614. [CrossRef]
3. Li, L.W.; Liu, L.; Du, X.H.; Wang, X.; Zhang, Z.; Zhang, J.; Liu, J. CGUN-2A: Deep Graph Convolutional Network via Contrastive Learning for Large-Scale Zero-Shot Image Classification. *Sensors* **2022**, *22*, 9980. [CrossRef]
4. Palatucci, M.; Pomerleau, D.; Hinton, G.E. Zero-shot learning with semantic output codes. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 1410–1418.
5. Li, Z.; Chen, Q.; Liu, Q. Augmented semantic feature based generative network for generalized zero-shot learning. *Neural Netw.* **2021**, *143*, 1–11. [CrossRef]
6. Ohashi, H.; Al-Naser, M.; Ahmed, S.; Nakamura, K.; Sato, T.; Dengel, A. Attributes' Importance for Zero-Shot Pose-Classification Based on Wearable Sensors. *Sensors* **2018**, *18*, 2485. [CrossRef]
7. Wu, L.; Wang, Y.; Li, X.; Gao, J. Deep attention-based spatially recursive networks for fine-grained visual recognition. *IEEE Trans. Cybern.* **2018**, *49*, 1791–1802. [CrossRef]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances In Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
9. Lampert, C.; Nickisch, H.; Harmeling, S. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 453–465. [CrossRef]
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
11. Xu, W.J.; Xian, Y.Q.; Wang, J.N.; Schiele, B.; Akata, Z. Attribute prototype network for zero-shot learning. *Neural Inf. Process. Syst.* **2020**, *33*, 21969–21980.
12. Xie, G.S.; Liu, L.; Jin, X.B.; Zhu, F.; Zhang, Z.; Qin, J.; Yao, Y.Z.; Shao, L. Attentive region embedding network for zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 9376–9385.
13. Li, K.; Min, M.R.; Fu, Y. Rethinking zero-shot learning: A conditional visual classification perspective. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3583–3592.
14. Zhang, L.; Xiang, T.; Gong, S. Learning a deep embedding model for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Vattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2021–2030.
15. Chen, S.M.; Xie, G.S.; Liu, Y.Y.; Peng, Q.M.; Sun, B.G.; Li, H.; You, X.G.; Ling, S. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Neural Inf. Process. Syst.* **2021**, *34*, 16622–16634.
16. Zhu, Y.Z.; Tang, Z.; Peng, X.; Elgammal, A. Semantic-guided multi-attention localization for zero-shot learning. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
17. Jayaraman, D.; Kristen, G. Zero-shot recognition with unreliable attributes. In Proceedings of the International Conference on Neural Information Processing Systems, Montreal, QC, USA, 8–13 December 2014; pp. 3464–3472.
18. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3111–3119.
19. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
20. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of tricks for efficient text classification. *arXiv* **2016**, arXiv:1607.01759.
21. Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; Akata, Z. Attribute prototype net-work for zeroshot learning. *arXiv* **2020**, arXiv:2008.08290.
22. Chen, S.; Hong, Z.; Liu, Y.; Xie, G.S.; Sun, B.; Li, H.; Peng, Q.; Lu, K.; You, X. Transzero: Attribute-guided transformer for zero-shot learning. *arXiv* **2021**, arXiv:2112.01683. [CrossRef]
23. Yang, Z.; Liu, Y.; Xu, W.; Huang, C.; Zhou, L.; Tong, C. Learning prototype via placeholder for zero-shot recognition. *arXiv* **2022**, arXiv:2207.14581.
24. Chen, L.; Zhang, H.-W.; Xiao, J.; Liu, W.; Chang, S. Zero-shot visual recognition using semantics preserving adversarial embedding networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1043–1052.

25. Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1425–1438. [CrossRef]
26. Liu, Y.; Zhou, L.; Bai, X.; Gu, L.; Harada, T.; Zhou, J. Information bottleneck constrained latent bidirectional embedding for zero-shot learning. *arXiv* **2020**, arXiv:2009.07451.
27. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-Shot Learning-A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 9. [CrossRef]
28. Zhao, B.; Wu, B.; Wu, T.; Wang, Y. Zero-shot learning posed as a missing data problem. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2616–2622.
29. Wang, D.; Li, Y.; Lin, Y.; Zhuang, Y. Relational knowledge transfer for zero-shot learning. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2145–2151.
30. Changpinyo, S.; Chao, W.L.; Gong, B.; Sha, F. Synthesized classifiers for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5327–5336.
31. Shigeto, Y.; Suzuki, I.; Hara, K.; Shimbo, M.; Matsumoto, Y. Ridge Regression, Hubness, and Zero-shot Learning. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, 7–11 September 2015; pp. 135–151.
32. Ji, Z.; Yu, X.; Yu, Y.; Pang, Y.; Zhang, Z. Semantic-guided class-imbalance learning model for zero-shot image classification. *IEEE Trans. Cybern.* **2021**, *52*, 6543–6554. [CrossRef]
33. Chen, S.-M.; Wang, W.J.; Xia, B.H.; Peng, Q.M.; You, X.G.; Zheng, F.; Shao, L. Free: Feature re-finement for generalized zero-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 122–131.
34. Li, J.; Jing, M.M.; Lu, K.; Ding, Z.; Zhu, L.; Huang, Z. Leveraging the invariant side of generative zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 7402–7411.
35. Keshari, R.; Singh, R.; Vatsa, M. Generalized zero-shot learning via over-complete distribution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13300–13308.
36. Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; Akata, Z. Generalized zero- and few-shot learning via aligned variational autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 8247–8255.
37. Shen, Y.; Qin, J.; Huang, L.; Liu, L.; Zhu, F.; Shao, L. Invertible zero-shot recognition flows. In Proceedings of the European Conference on Computer Vision, 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 614–631.
38. Yao-Hung, H.T.; Huang, L.-K.; Salakhutdinov, R. Learning robust visual-semantic embeddings. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3591–3600.
39. Yu, Y.; Ji, Z.; Li, X.; Guo, J.; Zhang, Z.; Ling, H.; Wu, F. Transductive zero-shot learning with a self-training dictionary approach. *IEEE Trans. Cybern.* **2018**, *48*, 2908–2919. [CrossRef]
40. Zhu, X.L.; He, Z.L.; Zhao, L.; Dai, Z.C.; Yang, Q.L. A Cascade Attention Based Facial Expression Recognition Network by Fusing Multi-Scale Spatio-Temporal Features. *Sensors* **2022**, *22*, 1350. [CrossRef]
41. Sun, Y.; Bi, F.; Gao, Y.E.; Chen, L.; Feng, S.T. A Multi-Attention UNet for Semantic Segmentation in Remote Sensing Images. *Symmetry* **2022**, *14*, 906. [CrossRef]
42. Liu, R.; Tao, F.; Liu, X.; Na, J.; Leng, H.; Wu, J.; Zhou, T. RAANet: A Residual ASPP with Attention Framework for Semantic Segmentation of High-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3109. [CrossRef]
43. Obeso, A.M.; Benois-Pineau, J.; Vazquez, M.S.G.; Acosta, A.Á.R. Visual vs internal attention mechanisms in deep neural networks for image classification and object detection. *Pattern Recognit.* **2022**, *123*, 108411. [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.

sensors

MDPI

Article

A New Bolt Defect Identification Method Incorporating Attention Mechanism and Wide Residual Networks

Liangshuai Liu *, Jianli Zhao, Ze Chen, Baijie Zhao and Yanpeng Ji

Electric Power Research Institute, State Grid Hebei Electric Power Co., Ltd., Shijiazhuang 050013, China
* Correspondence: liuliangshuai214@163.com; Tel.: +86-186-3390-0355

Abstract: Bolts are important components on transmission lines, and the timely detection and exclusion of their abnormal conditions are imperative to ensure the stable operation of transmission lines. To accurately identify bolt defects, we propose a bolt defect identification method incorporating an attention mechanism and wide residual networks. Firstly, the spatial dimension of the feature map is compressed by the spatial compression network to obtain the global features of the channel dimension and enhance the attention of the network to the vital information in a weighted way. After that, the enhanced feature map is decomposed into two one-dimensional feature vectors by embedding a cooperative attention mechanism to establish long-term dependencies in one spatial direction and preserve precise location information in the other direction. During this process, the prior knowledge of the bolts is utilized to help the network extract critical feature information more accurately, thus improving the accuracy of recognition. The test results show that the bolt recognition accuracy of this method is improved to 94.57% compared with that before embedding the attention mechanism, which verifies the validity of the proposed method.

Keywords: deep learning; bolt defect recognition; wide residuals; double attention

Citation: Liu, L.; Zhao, J.; Chen, Z.; Zhao, B.; Ji, Y. A New Bolt Defect Identification Method Incorporating Attention Mechanism and Wide Residual Networks. *Sensors* **2022**, *22*, 7416. https://doi.org/10.3390/s22197416

Academic Editor: Sylvain Girard

Received: 29 August 2022
Accepted: 27 September 2022
Published: 29 September 2022

1. Introduction

Bolts are the most numerous and widely distributed fasteners in transmission lines. As they play an important role in maintaining the stable operation of the lines, it is necessary to inspect the abnormal state of the bolts promptly so as to guarantee the safe and steady operation of the lines [1,2]. At present, the use of unmanned aerial vehicles (UAV) equipped with high-resolution cameras for transmission line inspection is not only safer and more efficient [3], but also can integrate deep learning-based image processing technology, which remarkably improves the quality and speed of inspection work. It is of great significance to study the bolted defect image recognition method based on deep learning.

Since the LeNet model was proposed, convolutional neural network models have shown considerable potential in image recognition tasks and have continued to develop. AlexNet [4] further increased the network depth and won the ImageNet challenge in 2012, and then ZFNet [5] and Google Inception Network (GoogLeNet) [6] were proposed one after another. Visual Geometry Group Network (VGGNet) [7] uses 16 convolutional layers and fully connects layers to improve the image recognition accuracy. However, the deepening of the network is not infinite. With the deepening of the number of network layers, problems caused by the deep network such as gradient disappearance and gradient explosion also emerge. The residual network (ResNet) proposed in [8] employs a jump connection method which effectively reduces the parameter number of the network, improves the training speed of the network, and ensures high accuracy. It is an effective solution to the problem that deep neural networks are difficult to train. Based on this, wide residual networks (WRNs) [9] further improve the model performance and increase the recognition accuracy by adding the number and width of convolutional layers to the residual blocks.

Currently, deep learning has been comprehensively used in bolt detection [10], defect classification [11], etc. In [12], the authors used multi-scale features extracted by cascade regions with a convolutional neural network (Cascade R-CNN) to build a path aggregation feature pyramid, which completes bolt defect identification. In [13], the authors enhanced the model complexity and improved the image recognition accuracy through the combined utilization of multiple algorithms. In [14], the authors used wide residuals as the backbone network and selected the optimal structure to achieve effective recognition of bolt defects by adjusting the network-widening dimension. In [15], a bolt defect data augmentation method was proposed based on random pasting, and it effectively expanded the number of bolt defect samples and improved the accuracy of defect recognition. However, due to the small size of the bolt itself, the bolt image features of the aerial transmission line are difficult to extract, and the bolt defect recognition effect is not satisfactory. The above method did not take into account the features of the bolt itself when improving the model.

The attention mechanism can help the network improve the feature extraction ability of the image [16,17]. It is a bionic of human vision that enables the acquisition of detailed information and the suppression of irrelevant information by allocating more attention to the target area. In the domain of deep learning, the attention mechanism uses the feature map to learn a new weight distribution, which is imposed on the original feature map. This weighting not only preserves the original information of the image extracted by the original network, but also enhances focus on the target region, effectively improving the performance of the model. The attention mechanism is not a complete network structure, but a plug-and-play lightweight module. When this module is embedded in the network, it can reasonably allocate computational resources and significantly increase the neural network performance at the cost of a finite increase in the number of parameters. Thus, it has received much attention in detection, segmentation, and recognition tasks because of its practicality and robustness [18–20]. Currently, it can be classified into three categories: spatial domain, channel domain, and hybrid domain. The squeeze and excitation attention network (SENet) [21] and efficient channel attention networks (ECA-Net) [22] are both of single-way attention frames that help the network detect or identify targets better by aggregating information in the spatial domain or channel domain and adaptively learning new weights. These networks are more concise than those with multi-way attention. The selective kernel network (SK-Net) [23] decomposes the feature map into feature vectors by decomposition, aggregation, and matching. In this way, the network is able to extract more detailed feature information. The convolutional block attention module (CBAM) [24] aggregates spatial and channel information to guide the model to focus on the key target regions in the image, while channel attention (CA) improves the ability to capture targets by aggregating one-dimensional channel and spatial information to relate the location relationships between targets in the feature graph. In [25], the authors proposed a dynamic supervised knowledge distillation method for bolt defect recognition and classification by applying knowledge distillation techniques to the bolt defect classification task and combining spatial channel attention. This method effectively improves the accuracy of bolt defect classification. In [26], the authors used an attention mechanism to locate the possible regions of the bolt in the image and then combined it with a deconvolutional network to build a model to achieve accurate detection of the bolt. This is an attention-based mechanism for transmission tower bolt detection. In [27], the authors embedded a dual-attention mechanism in faster regions with a convolutional neural network (Faster R-CNN) to analyze and enhance visual features at different scales and different locations, which effectively improved the bolt detection accuracy.

Although these methods improve the recognition or detection accuracy of bolts to some extent, they are all based on improving the feature expression capability of bolts without improving the model by combining bolt features. In order to identify bolt defects more accurately, by combining the attention mechanisms, we introduce bolt knowledge into the model and study the bolt defect recognition method incorporating dual attention in this paper. WRN is used as the backbone network, and the attention-wide residual network is designed by embedding squeeze and excitation networks [21] and coordinate attention [28] to enhance the network's perception of features in the spatial dimension and channel dimension. The network was designed to enhance its ability to perceive features in the spatial dimension and channel dimension, extracting richer feature information. It is combined with the prior knowledge of bolts to achieve high-accuracy recognition of bolt defects.

2. Materials and Methods

In this work, WRN is used as the backbone network, and the number of channels is 16 × k, 32 × k, and 64 × k, a total of three levels. Among them, three wide residual blocks are in the first level, four wide residual blocks are in the second level, and six wide residual blocks are in the third level. The width factor k is taken as 2. The attention-wide residual network is designed by fusing the attention mechanism in the WRN, so as to enhance the extraction ability for bolt features and improve the accuracy of defect recognition. The overall structure is shown in Figure 1. Firstly, SENet attention is added to each level in the WRN to enhance the network's ability to capture bolt defect features and output higher-quality feature maps. Secondly, CA attention based on structural prior knowledge is imported in combination with the spatial location relationship of pins and nuts on bolts, which enables the network to better utilize the feature location relationship and thus improve the accuracy of bolt defect recognition.

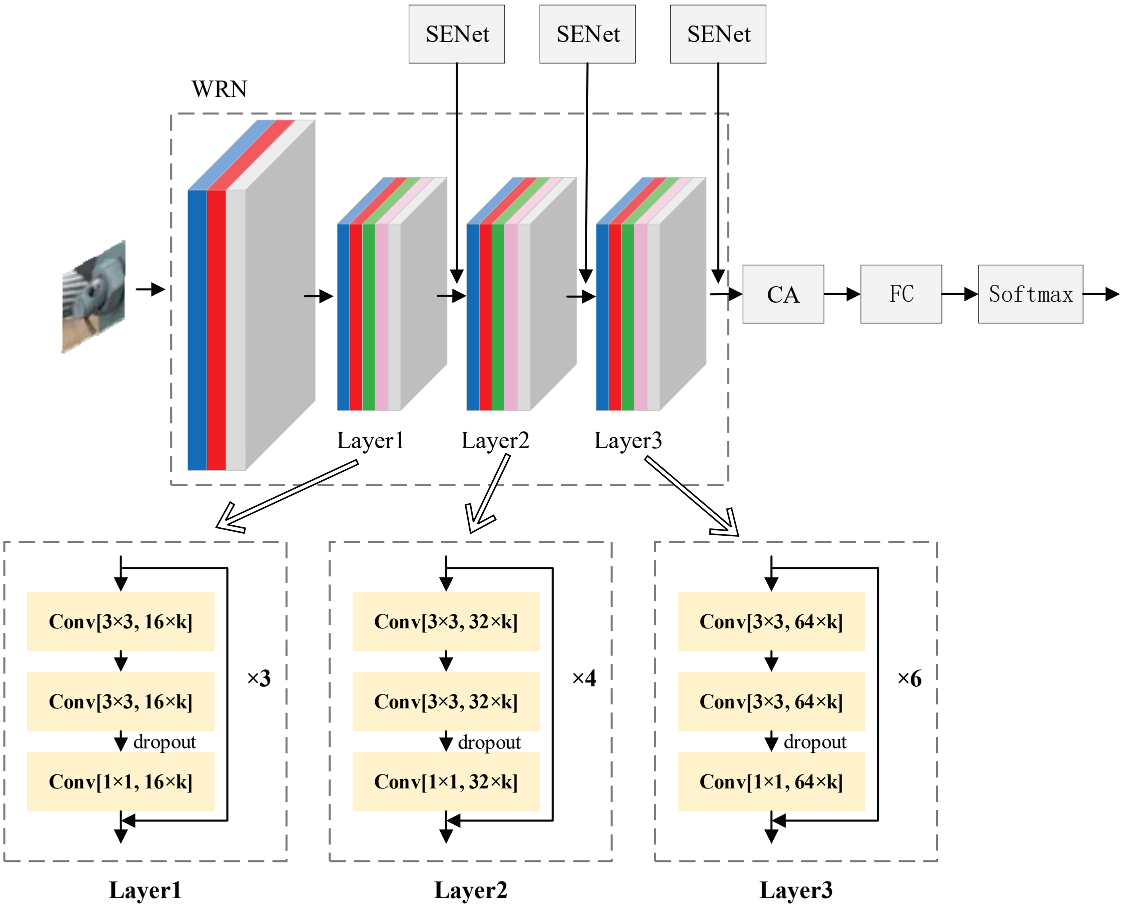


Figure 1. Attention to wide residual network structure.

2.1. WRN Framework for Fusing Channel Attention

A residual network consists of a residual block. It is a constant mapping of shallow features to deeper features using a jump connection so that the residual block can learn more feature information based on the input features and effectively solve the degradation problem caused by deeper networks. However, as the number of network layers increases, the residual block itself cannot be better expressed. A new type of residual approach, WRN, which widens the number of convolutional kernels in the original residual block, was proposed. It effectively improves the utilization of the residual block, reduces the model parameters, speeds up the computation, and makes it possible to obtain a better training result without a deeper network layer. In addition, WRN adds a dropout between the convolutional layers in the residual block to form a wide ResNet block, which has the effect of improving the performance of the network. The relationship between the ResNet block and the wide ResNet block is shown in Figure 2, where 3 × 3 indicates the size of the convolution kernel, N is the number of channels, and k indicates the width factor.

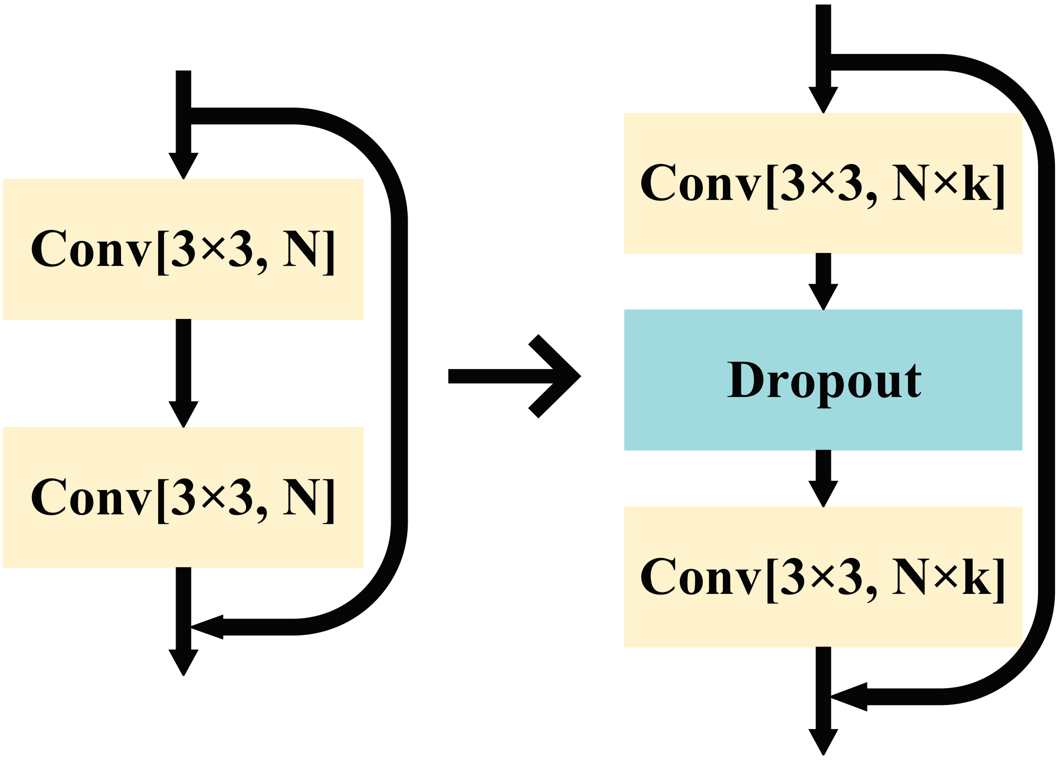


Figure 2. Schematic diagram of the relationship between ResNet block (**left**) and wide-ResNet block (**right**).

SENet attention can aggregate the information from the input features at the spatial level and adaptively acquire new weight relationships through learning. These weight relationships represent the importance of different regions in the feature map, making the network focus on key regions in the feature map as a whole. It helps the information transfer in the network and continuously updates parameters in the direction that is beneficial to the recognition task.

After fusing SENet attention in the WRN, the network first compresses the spatial dimension of the feature map of the input SENet through global average pooling, aggregating spatial information to perceive richer global features of the image and enhancing the network expression capability. The SENet attention structure diagram is shown in Figure 3. The global average pooling operation generates a feature map of C × 1 × 1 (where C represents the number of channels) to obtain the global information of channels. Then, the correlation between channels is captured by the two fully connected layers with the activation function of ReLu, and the normalized channel weights are then generated by the sigmoid activation function. At this point, the channel weights of dimension C × 1 × 1 can be multiplied with the input features of dimension C × H × W (where H represents the feature map of height, W represents the feature map of width) as a new parameter, i.e., the aligned channel dimension C. For each H × W matrix, a channel coefficient c is multiplied to obtain the output features C × H × W after SENet attention optimization, which enhances the key region features and suppresses irrelevant features to improve the performance of the network.

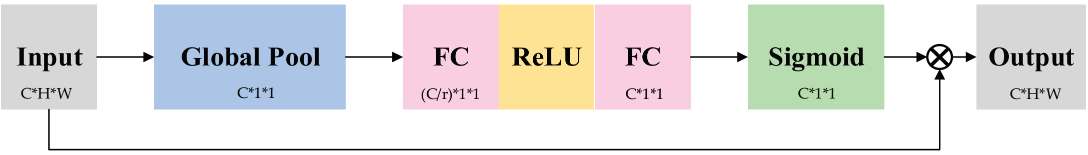


Figure 3. SENet attention structure diagram.

The attention weights are multiplied by the input features to obtain the output features F, as follows:

$$F = \delta(MLP(Pool(F_0))) \times F_0 \tag{1}$$

where F_0 denotes the input features, δ and MLP denote the sigmoid activation function and neural network operation, respectively, and $Pool$ represents the pooling operation.

2.2. CA Attention with Integrated Knowledge

The WRN incorporating SENet attention is enhanced to extract bolt features. However, according to the prior knowledge of the bolt, pins distribute at the head of the bolt while nuts usually locate at the root of the bolt, and these positional relationships are fixed. In order to further improve the bolt defect recognition accuracy using the bolt position information, we add CA attention to the output section of the WRN to enhance the positional relationships of the target. The CA attention structure is shown in Figure 4. First, CA attention decomposes the input features into a horizontal perceptual feature vector of dimension C × H × 1 and a vertical perceptual feature vector of dimension C × 1 × W by global averaging pooling in both directions. The one-dimensional feature vectors in the horizontal and vertical directions are as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} F_c(h, i) \tag{2}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} F_c(j, w) \tag{3}$$

where H and W represent the height and width, respectively, h, w, i, and j represent the location coordinates in the feature map, c represents the number of channels, ${z_c}^h$ represents the one-dimensional feature vector in the horizontal direction, ${z_c}^w$ represents the one-dimensional feature vector in the vertical direction, and F_c represents the input feature map.

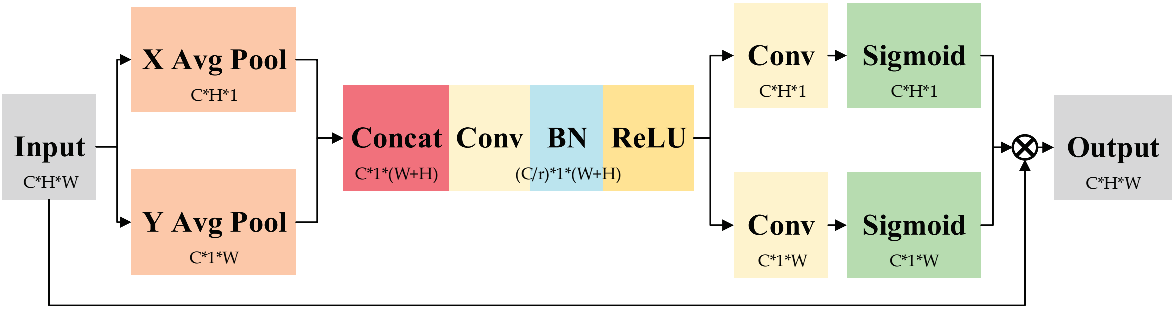


Figure 4. CA attention structure diagram.

In this process, the attention mechanism establishes long-term dependencies in one spatial direction and preserves precise location information in the other, helping the network locate key feature regions more accurately. It also gives the network a better global sensory view of the field as well as rich feature information. Next, the perceptual feature vectors in both directions are aggregated, and the feature mapping is obtained by dimensionality reduction through 1 × 1 convolution. Unique feature mappings are generated using two one-dimensional features.

$$f = MLP([z^h, z^w]) \tag{4}$$

where $[z^h, z^w]$ represents the stitching operation of two one-dimensional features, and f is the feature mapping of spatial information in the encoding process of horizontal and vertical directions. Finally, the feature mapping is decomposed and normalized by the Sigmoid function to obtain the attention weights in the two directions, and the attention weights in the two directions are multiplied with the input features of dimensionality C × H × W to obtain the output features of dimensionality C × H × W. The two directional weights and output features are as follows:

$$g^h = \delta(T(f^h)) \tag{5}$$

$$g^w = \delta(T(f^w)) \tag{6}$$

$$F(i,j) = F_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{7}$$

where T represents the convolution operation and $F(i, j)$ is the output feature. After the feature map is processed by CA attention, it is easier for the network to capture the key feature information in the map using location information, and the relationship between channels is more obvious.

3. Test Results and Analysis

3.1. Test Data and Settings

Dataset Construction: We constructed a transmission line bolt defect recognition dataset by cropping and optimizing transmission line aerial images based on the Overhead Transmission Line Defect Classification Rules (for Trial Implementation). Tests were conducted to verify the effectiveness of this method. The dataset was divided into three categories, namely normal bolts, missing pin bolts, and missing nut bolts. There are a total of 6327 images, of which 2990 were normal bolts, 2802 were missing pin bolts, and 535 were missing nut bolts, and the training set and test set were divided in a ratio of 4:1. The samples of each category are shown in Figure 5.

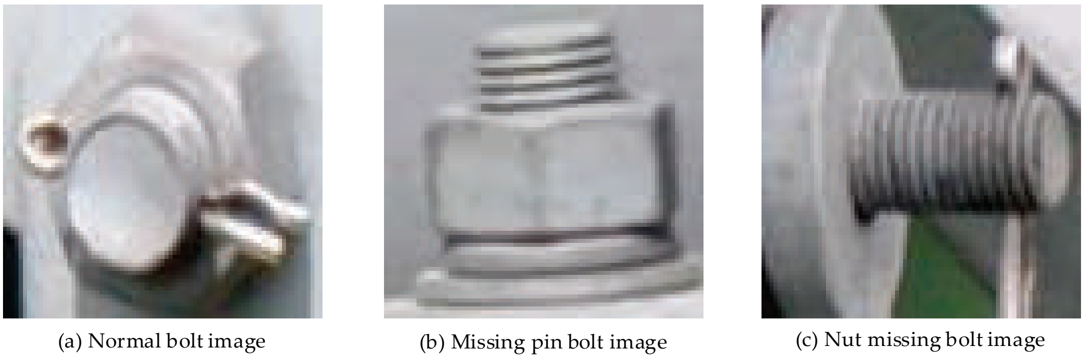

(a) Normal bolt image (b) Missing pin bolt image (c) Nut missing bolt image

Figure 5. Three categories of bolt image samples.

Test Settings: The test hardware environment was Linux Ubuntu 16.04, and the GPU used is an NVIDIA GeForce 1080Ti with 11 GB of RAM. The test parameters were a batch size of 64, an epoch count of 200, and a learning rate of 0.1. We used the model to perform a recognition validation on the test set after the model completes an epoch training, obtain and save the accuracy and loss function values of the model on the test set, and take the highest recognition accuracy on the test set as the model evaluation metric after the model completes training. The accuracy rate was chosen as the evaluation index, and the formula is shown in Equation (8), where *TP* is the number of correctly predicted positive samples, *TN* is the number of correctly predicted negative samples, *FN* is the number of incorrectly predicted negative samples, and *FP* is the number of incorrectly predicted positive samples.

$$\text{Accuracy} = \frac{TP}{TP + TN + FP + FN} \tag{8}$$

3.2. Ablation Tests and Analysis

In order to verify the effectiveness of this method in the actual bolt defect recognition task, we compared the accuracy of the test set under different methods by ablation experiments separately, as shown in Table 1. As can be seen, the recognition accuracy of the base model WRN was 93.31%, an improvement of 0.58% after adding SENet attention. This is because the SENet attention mechanism acquired richer bolt features by compressing spatial information, which enhanced the expressiveness of the network. With the addition of CA attention to the model, the attention mechanism builds long-term dependencies in space and the network is more likely to use the location relationships to capture key feature information, resulting in a 0.72% increment in recognition accuracy. The recognition accuracy of the model was improved by 1.26% after embedding both SENet attention and CA attention. The mutual association between the attentions further improved the network's performance and it has accomplished a more accurate bolt defect recognition task.

Table 1. Ablation test results.

Method	Accuracy (%)
WRN	93.31
WRN + SENet	93.89
WRN + CA	94.03
Ours	94.57

Figure 6 shows the variation curve of the recognition accuracy of the model on the test set as the number of training rounds increases. As can be seen, between epochs of 1 and 60, the accuracy of the model has the fastest rising trend, but the fluctuation is large, and the model has not learned efficient defect recognition ability. Between 60 and 120 epochs, the model's learning task is initially completed, but the accuracy curve is still fluctuating. As the model was trained iteratively, the fluctuation of the accuracy curve gradually decreased after 120 epochs, and finally stabilized after 160 epochs.

Figure 7 shows the loss descent curves of different networks on the training set during the training process. As can be seen, the loss function convergence curves of the model training process under different approaches are compared. The first convergence was between epochs 1 and 60, during which the WRN model had the highest initial value, the WRN plus SENet had the slowest convergence, and the WRN plus CA attention had the fastest convergence. The second convergence was between epochs 60 and 120, and the third was between epochs 120 and 160. In these two convergence domains, the convergence rates and convergence trends of the four models were more or less the same, and the loss function convergence curves of each model showed slight fluctuations. The convergence trend of WRN is the weakest. WRN plus SENet and WRN plus CA attention are similar, and the convergence trend of our proposed method is the best.

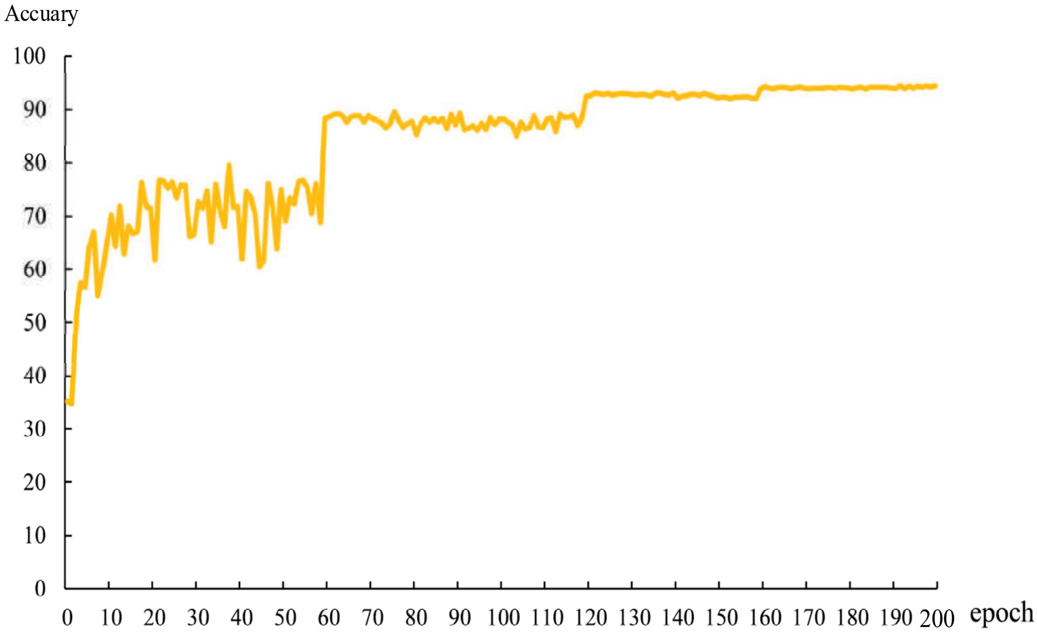


Figure 6. Accuracy curve on test set.

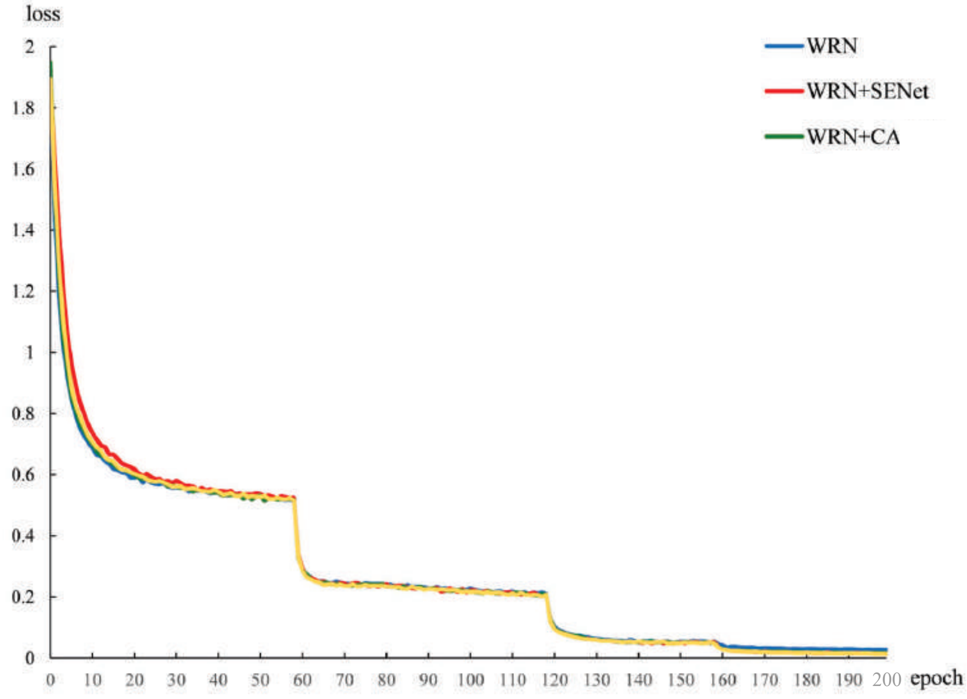
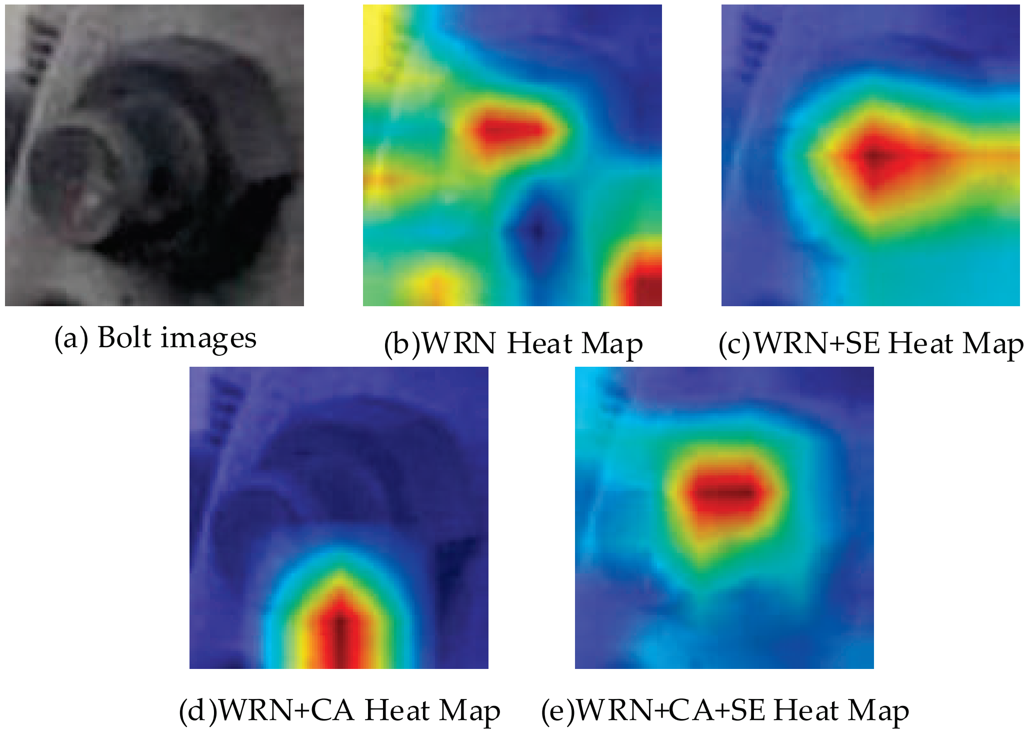


Figure 7. Convergence curve of the model training loss function.

In order to demonstrate the improvement in model performance by attention more intuitively, we used the gradient-weighted class activation mapping (Grad-CAM) [29] algorithm to visualize the feature maps before and after the model improvement, as shown in Figure 8. In this test, a bolt image with missing pins was used as the reference. It can be seen from the figure that the attention area of the features extracted by WRN only is relatively scattered, which is not conducive to the recognition of the bolt by the model. Our method incorporates both SENet attention and CA attention, and the extracted feature map is more significant and discriminative compared with the previous ones. Our method effectively removes redundant information and allows the model to better distinguish bolt categories.



(a) Bolt images

(b)WRN Heat Map

(c)WRN+SE Heat Map

(d)WRN+CA Heat Map

(e)WRN+CA+SE Heat Map

Figure 8. Visualization of the bolt feature map.

3.3. Comparative Tests and Analysis

In these tests, we compared the recognition accuracy of different recognition models for bolt defects in the test set, as shown in Table 2. WRN has the highest accuracy of 93.31%, 3.94% higher than VGG16, and 0.86% and 0.64% higher than ResNet50 and ResNet101, respectively. It fully demonstrates the feasibility and superiority of the backbone network selected in this paper, and paves the way for the next model improvement.

Table 2. Ablation test results.

Recognition Model	Accuracy of Bolt Defect Recognition %
VGG16	89.37
ResNet50	92.45
ResNet101	92.67
WRN	93.31

Meanwhile, we compared the recognition accuracy of each bolt before and after the improvement in the test set, as shown in Figure 9. As can be seen from the figure, after the improvement, the recognition accuracy was increased by 0.77% for normal bolts, 1.24% for missing pin bolts, and 1.76% for missing nut bolts. The accuracy improvement for normal bolts is less, while the accuracy improvement for bolts with missing pins and bolts with missing nuts is more significant with the help of the attention mechanism. This shows that the joint attention-wide residual method proposed in this paper is effective for bolt defect recognition. Embedding SENet attention into each layer to improve the ability of model feature extraction and combining CA attention to focus more accurately on the area of pin or nut in the figure helps the model to better discriminate the bolt category and improve the recognition accuracy.

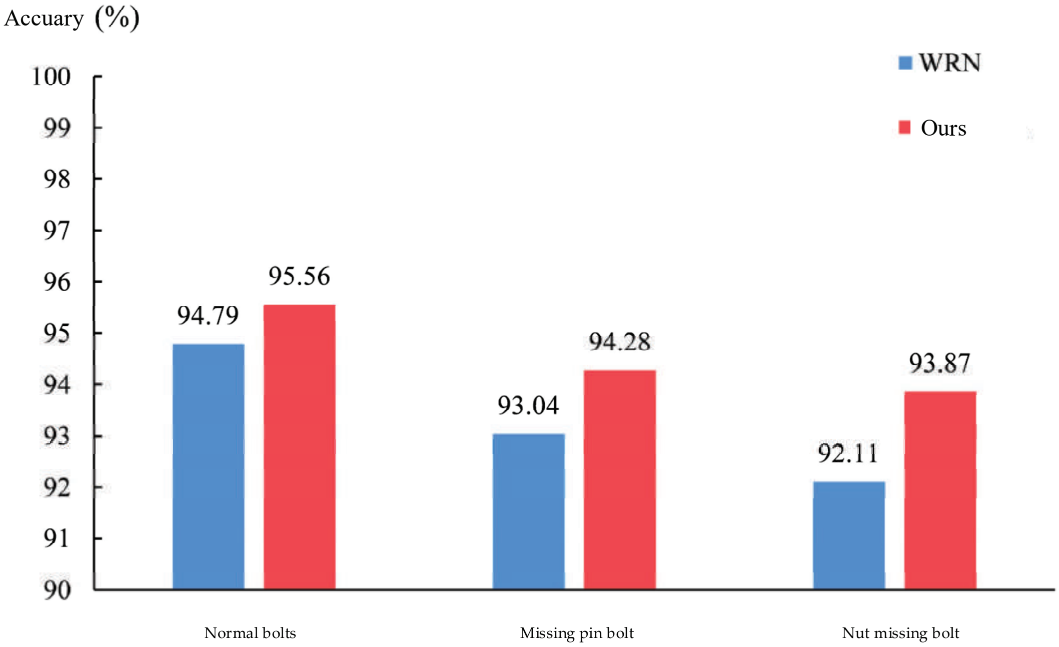


Figure 9. Comparison of classification accuracy before and after model improvement.

4. Conclusions

In order to identify bolt defects more accurately, by taking WRN as the backbone network, we address the problem of difficult extraction of bolt features and the fixed position relationship of pins and nuts on top of the bolts. A new bolt defect identification method incorporating an attention mechanism and wide residual networks is proposed, embedding SENet and CA attention and fusing bolt knowledge. The proposed method can locate the key feature areas with better precision through collaborative space and channel information so as to help the model to improve the recognition accuracy. The proposed method has been validated on a homemade transmission line bolt defect recognition dataset. The test results show that the accuracy of this method was improved by 1.26% compared with that before improvement, which lays a foundation for the transmission line bolt defect detection task.

Author Contributions: Conceptualization, L.L. and J.Z.; methodology, L.L. and J.Z.; software, L.L.; validation, Z.C.; formal analysis, L.L. and B.Z.; investigation, Y.J.; resources, L.L.; data curation, L.L.; writing—original draft preparation, J.Z.; writing—review and editing, L.L.; visualization, Z.C.; supervision, Y.J.; project administration, L.L.; funding acquisition, L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the State Grid Hebei Electric Power Provincial Company Science and Technology Project Fund Grant Project: Research on visual defect detection technology of power transmission and transformation equipment based on deep knowledge reasoning (project number: kj2021-016).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Special thanks are given to Power grid Technology center, Electric Power Research Institute, State Grid Hebei Electric Power Co., Ltd., Shijiazhuang 050013, China.

Conflicts of Interest: All authors have received research grants from Electric Power Research Institute, State Grid Hebei Electric Power Co., Ltd. None of the authors have received a speaker honorarium from the company or own stock in the company. None of the authors have been involved as consultants or expert witnesses for the company. The content of the manuscript has not been applied for patents; none of the authors are the inventor of a patent related to the manuscript content.

Abbreviations

The following abbreviations are used in this manuscript.

UAV	Unmanned Aerial Vehicle
GoogLeNet	Google Inception Network
VGGNet	Visual Geometry Group Network
ResNet	Residual Network
WRN	Wide Residual Networks
Cascade R-CNN	Cascade Regions with Convolutional Neural Network
SENet	Squeeze and Excitation Attention Network
ECA-Net	Efficient Channel Attention Networks
SK-Net	Selective Kernel Network
CBAM	Convolutional Block Attention Module
CA	Channel Attention
Faster R-CNN	Faster Regions with Convolutional Neural Network
Grad-CAM	Gradient-Weighted Class Activation Mapping

References

1. Zhao, Z.; Qi, H.; Qi, Y.; Zhang, K.; Zhai, Y.; Zhao, W. Detection Method Based on Automatic Visual Shape Clustering for Pin-Missing Defect in Transmission Lines. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 6080–6091. [CrossRef]
2. Han, Y.; Han, J.; Ni, Z.; Wang, W.; Jiang, H. Instance Segmentation of Transmission Line Images Based on an Improved D-SOLO Network. In Proceedings of the 2021 IEEE 3rd International Conference on Power Data Science, Harbin, China, 26 December 2021; pp. 40–46.
3. He, T.; Zeng, Y.; Hu, Z. Research of Multi-Rotor UAVs Detailed Autonomous Inspection Technology of Transmission Lines Based on Route Planning. *IEEE Access* **2019**, *7*, 114955–114965. [CrossRef]
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; MITP: Boston, MA, USA, 2012; pp. 1097–1105.
5. Zeiler, D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerlan, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 818–833.
6. Szegedy, C.; Liu, W.; Jia, Y. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–9.
7. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
8. Xie, S.; Girshick, R.; Dollá, P. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1492–1500.
9. Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.
10. Wang, C.; Wang, N.; Ho, S.-C.; Chen, X.; Song, G. Design of a New Vision-Based Method for the Bolts Looseness Detection in Flange Connections. *IEEE Trans. Ind. Electron.* **2020**, *67*, 1366–1375. [CrossRef]
11. Xiao, L.; Wu, B.; Hu, Y. Missing Small Fastener Detection Using Deep Learning. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–9. [CrossRef]
12. Wang, H.; Zhai, X.; Chen, Y. Two-stage pin defect detection model based on improved Cascade R-CNN. *Sci. Technol. Eng.* **2021**, *21*, 6373–6379.
13. Zhao, Y.Q.; Rao, Y.; Dong, S.P. Survey on deep learning object detection. *J. Image Graph.* **2020**, *25*, 629–654.
14. Qi, Y.; Jin, C.; Zhao, Z. Optimal Knowledge Transfer Wide Residual Network Transmission Line Bolt Defect Image Classification. *Chin. J. Image Graph.* **2021**, *26*, 2571–2581.
15. Zhao, W.; Jia, M.; Zhang, H.; Xu, M. Small Target Paste Randomly Data Augmentation Method Based on a Pin-losing Bolt Data Set. In Proceedings of the 2021 IEEE 3rd International Conference on Power Data Science, Harbin, China, 26 December 2021; pp. 81–84.
16. Brauwers, G.; Frasincar, F. A General Survey on Attention Mechanisms in Deep Learning. *IEEE Trans. Knowl. Data Eng.* **2021**. [CrossRef]

17. Sun, J.; Jiang, J.; Liu, Y. An Introductory Survey on Attention Mechanisms in Computer Vision Problems. In Proceedings of the 2020 6th International Conference on Big Data and Information Analytics (BigDIA), Shenzhen, China, 4–6 December 2020; pp. 295–300.
18. Li, Y.-L.; Wang, S. HAR-Net: Joint Learning of Hybrid Attention for Single-Stage Object Detection. *IEEE Trans. Image Process.* **2020**, *29*, 3092–3103. [CrossRef] [PubMed]
19. Guo, Z.; Huang, Y.; Wei, H.; Zhang, C.; Zhao, B.; Shao, Z. DALaneNet: A Dual Attention Instance Segmentation Network for Real-Time Lane Detection. *IEEE Sens. J.* **2021**, *21*, 21730–21739. [CrossRef]
20. Lian, S.; Jiang, W.; Hu, H. Attention-Aligned Network for Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 3140–3153. [CrossRef]
21. Hu, J.; Shen, L.; Albanie, S. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
22. Wang, Q.; Wu, B.; Zhu, P. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 11531–11539.
23. Li, X.; Wang, W.; Hu, X. Selective kernel networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 510–519.
24. Woo, S.; Park, J.; Lee, J.Y. CBAM: Convolutional block attention module. In Proceedings of the Computer Vision-ECCV 2018, Munich, Germany, 8–14 September 2018; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–19.
25. Zhao, Z.; Jin, C.; Qi, Y. Image Classification of Bolt Defects in Transmission Lines Based on Dynamic Supervised Knowledge Distillation. *High Volt. Technol.* **2021**, *47*, 406–414.
26. Weitao, L.; Huimin, G.; Qian, Z.; Gang, W.; Jian, T.; Meishuang, D. Research on Intelligent Cognition Method of Missing status of Pins Based on attention mechanism. In Proceedings of the 2021 IEEE 4th Advanced Information Management, Communicates Electronic and Automation Control Conference, Chongqing, China, 18–20 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1923–1927.
27. Qi, Y.; Wu, X.; Zhao, Z.; Shi, B.; Nie, L. Faster R-CNN Aerial Photographic Transmission Line Bolt Defect Detection Embedded with Dual Attention Mechanism. *Chin. J. Image Graph.* **2021**, *26*, 2594–2604.
28. Hou, Q.B.; Zhou, D.Q.; Feng, J.S. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 13708–13717.
29. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

sensors

MDPI

Article

Multi-Patch Hierarchical Transmission Channel Image Dehazing Network Based on Dual Attention Level Feature Fusion

Wenjiao Zai † and Lisha Yan *,†

College of Engineering, Sichuan Normal University, Chengdu 610101, China; zaiwenjiao@sicnu.edu.cn
* Correspondence: yanlisha@stu.sicnu.edu.cn; Tel.: +86-187-8384-2873
† These authors contributed equally to this work.

Abstract: Unmanned Aerial Vehicle (UAV) inspection of transmission channels in mountainous areas is susceptible to non-homogeneous fog, such as up-slope fog and advection fog, which causes crucial portions of transmission lines or towers to become fuzzy or even wholly concealed. This paper presents a Dual Attention Level Feature Fusion Multi-Patch Hierarchical Network (DAMPHN) for single image defogging to address the bad quality of cross-level feature fusion in Fast Deep Multi-Patch Hierarchical Networks (FDMPHN). Compared with FDMPHN before improvement, the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) of DAMPHN are increased by 0.3 dB and 0.011 on average, and the Average Processing Time (APT) of a single picture is shortened by 11%. Additionally, compared with the other three excellent defogging methods, the PSNR and SSIM values DAMPHN are increased by 1.75 dB and 0.022 on average. Then, to mimic non-homogeneous fog, we combine the single picture depth information with 3D Berlin noise to create the UAV-HAZE dataset, which is used in the field of UAV power assessment. The experiment demonstrates that DAMPHN offers excellent defogging results and is competitive in no-reference and full-reference assessment indices.

Keywords: transmission channels; non-homogeneous fog; dual attention; DAMPHN; image defogging

Citation: Zai, W.; Yan, L. Multi-Patch Hierarchical Transmission Channel Image Dehazing Network Based on Dual Attention Level Feature Fusion. *Sensors* **2023**, *23*, 7026. https://doi.org/10.3390/s23167026

Academic Editors: Ke Zhang and Yincheng Qi

Received: 17 July 2023
Revised: 4 August 2023
Accepted: 4 August 2023
Published: 8 August 2023

1. Introduction

UAVs have been increasingly employed in power inspection to find safety problems effectively [1]. However, in hilly regions, advection fog, uphill fog, and valley fog are frequently encountered [2,3], causing critical portions of transmission lines or towers to become fuzzy or even wholly concealed and decreasing fault detection accuracy. Image-defogging technology can be used to address the appeal issues. However, the non-homogenous fog is challenging for the current homogenous fog removal method. Additionally, the initial non-homogeneous defogging method FDMPHM exploits residual connections between several levels and ignores the issues with channel redundancy and unequal pixel distribution in cross-level fusion. Based on this, we suggest the Dual Attention Level Feature Fusion Multi-Patch Hierarchical Network (DAMPHN), which aims to enhance the cross-level fusion method of FDMPHN and produce superior defogging effects. Haze non-uniformity is not considered in power inspection image defogging studies due to a lack of non-homogeneous haze datasets. Therefore, to create a dataset that may represent non-homogeneous haze in mountainous places (UAV-HAZE), this paper ingeniously combines image depth measurements with 3D Berlin noise. The suggested DAMPHN performs better in color preservation and haze removal than the other four advanced approaches and can complete the picture preprocessing of transmission channels, according to numerous experiments on three open datasets and UAV-HAZE.

1.1. Related Work

Model-based parameter estimation and model-free picture enhancement methods are currently the main single-image fog removal categories. Additionally, future images for machine vision services will be of higher quality because of advancements in CCD imager technology [4]. Some researchers have used the image defog technique to preprocess photos based on high-quality photographs for transmission channels.

1.1.1. Model-Based Parameter Estimation Method

By predicting the transmission matrix $t(x)$ and global atmospheric light A from the haze graph $J(x,\lambda)$, these approaches, based on the atmospheric scattering model [5], provide images $I(x,\lambda)$ that are devoid of haze. In Equation (1), the atmospheric scattering model is displayed.

$$I(x,\lambda) = t(x)J(x,\lambda) + A(1 - t(x)) \tag{1}$$

$$t(x) = e^{-\beta(\lambda)d(x)} \tag{2}$$

where $d(x)$ denotes the depth of the scene and $\beta(\lambda)$ the scattering coefficient. Both the early dark channel prior (DCP) [6] and the color decay prior (CAP) [7] were put out and offered concepts for further study. Convolutional neural networks (CNN) were later developed, and Cai et al. [8] used CNNs with various kernel parameters for the first time to extract the distinctive information of dark channel, color attenuation, maximum contrast, and hue disparity to solve the parameters. Li et al. [9] equalized $t(x)$ and A as a parameter based on Formula (1) and applied CNN and residual connection to get this parameter. Zhang et al. [10] used the Dense-Net and U-net networks, respectively. A Densely Connected Pyramid Dehazing Network (DCPDN) was subsequently proposed based on the joint discriminator of adversarial networks and the optimization parameter estimate of the edge retention loss function. To achieve adaptive fusion, Li et al. [11] employed a multi-stage deep convolutional network to estimate $t(x)$ and A and added a memory network and a two-level attention mechanism to determine the weight of findings at each stage. To filter haze residuals step by step and achieve dehazing, Li et al. [12] modified Formula (1) to be task-oriented and assembled recurrent neural networks based on encoder-decoder and space. Bai et al. [13], who combined $t(x)$ and A into a single parameter and calculated it using the depth pre-defamer. The progressive feature fusion module and the picture recovery module were created to improve parameter estimation.

1.1.2. Model-Free Image Enhancement Method

This technique uses a coding-decoding structure to directly learn the link between the haze/clear image mapping and integrates attention mechanisms, feature fusion, and other techniques to enhance the dehazing performance. Das et al. [14] introduced the Fast Deep Multi-Patch Hierarchical Network (FDMPHN) and Fast Multi-Scale Hierarchical Network (FDMSHN) by improving the loss function, which was inspired by literature [15]. According to Wang et al. [16], a heterogeneous twin network was suggested, U-Net was used to extract haze features, and a detail enhancer network was set up to improve image details. Liu et al. [17] proposed an attention-based multi-scale defogging network (GridDehazeNet), which introduced a channel attention mechanism to improve feature fusion ability among multiple scales. A feature fusion attention network with a channel and pixel focus that prioritizes high-frequency and dense hazy areas was proposed by Qin et al. [18]. To improve the ability to extract edge texture features, Wang et al. [19] created the edge branch module based on the multi-level attention dehazing module and the feature fusion module based on Laplace gradient prior knowledge. Using extended convolution in the multi-scale part, channel attention mechanism in the cross-level fusion part, and frequency domain loss in the loss function part, Yang et al.'s [20] combination of FDMPHN and FDMSHN methods to obtain dense feature maps produced good results. A transfer attention technique was created by Wang et al. [21] to deal with non-uniform noise in images. To focus on the non-uniform hazy region and address the issues of artifacts and excessive smoothing, Zhao et al. [22] developed a dynamic attention module based on the dual attention mechanism. Guo et al. [23] suggested a self-paced half-course learningdriven attention image-generating technique based on the dual attention mechanism to enhance the ability to clear regions with considerable brightness disparities of fog.

1.1.3. Transmission Channel Image Dehazing Method

Recently, researchers have used it in power inspection after taking inspiration from the appeal algorithm. Liu et al. [24] created their own UAV picture collection for transmission line inspection and used the DCPDN approach to achieve dehazing. To address the drawbacks of the DCP method, Zhang et al. [25] divided the sky region by fusing the Canny operator and gradient energy function to obtain a more accurate atmospheric light value, and Zhai et al. [26] optimized the quadtree segmentation method. Both techniques were then applied to the image dehazing of transmission line monitoring systems. To remove haze from photographs of an insulator umbrella disk in transmission lines, Xin et al. [27] coupled a limited-contrast adaptive histogram equalization method with the dark channel, bright channel, and these methods. Gao et al.'s [28] use of DCP to remove haze from fixed-point monitoring photographs of a tower or pole was likewise based on this technique. Yan et al. [29] created their dataset for UAV power inspection and used FDMPHN to achieve dehazing.

1.2. Motivation and Contribution

The model-based parameter estimate methods produces improved outcomes in the area of picture fog removal. However, the overall image that DCP restored is dark, and color distortion can easily happen in areas of bright light. The reduction impact is weak when the depth of field shift in the image is not visible or when there is haze, as CAP is dependent on the color saturation of the image. To maximize the fog removal effect, later researchers used CNN to estimate the parameters $t(x)$ and A. However, both the parameter estimation methods based on CNN [8,10,11] and the parameter estimation method after the improved atmospheric scattering model [9,12,13] are subject to artifacts, color distortion, and haze residues because of the shortcomings of the atmospheric scattering model. Although the model-free image enhancement methods are not limited by the model, it depends on the ability of the network to extract and fuse the haze features. Only residual connections are used in the multi-patch network FDMPHN for cross-level feature fusion, disregarding channel differences and pixel distribution non-uniformity. Therefore, when the non-uniform characteristics of haze or the fog area are strong, it is easy for haze residue and detail blur to appear. Later researchers enhanced the network's capacity for feature extraction by improving the attention mechanism [17–23], but it was also challenging to address the issue of non-uniform fog.

In the area of fog removal in power inspection images, Refs. [24–28] all use a uniform haze dataset created based on an atmospheric scattering model as the foundation for their analyses, neglecting the non-uniform characteristics of haze distribution in natural settings. As a result, it is only appropriate for processing images with uniform haze distribution. It performs poorly when dealing with powerful light sources and non-uniform haze, and the image quality after recovery is also subpar. Furthermore, power inspection picture fog removal is still in the uniform haze removal stage, and it is challenging to make progress due to the relative paucity of non-uniform haze datasets [30]. Therefore, this paper suggests a Dual Attention Level Feature Fusion Multi-Patch Hierarchical Network (DAMPHN) to enhance the defogging effect of UAV inspection photos of transmission lines in mountainous terrain. This work's key contributions can be summed up as follows:

1. It is suggested to use a Dual Attention Level Feature Fusion Multi-Patch Hierarchical Network (DAMPHN) that combines an encoder-decoder module with a Dual Attention Level Feature Fusion (DA) module. The experimental results show that the network has low color distortion and a good defogging effect.
2. DA module is proposed. DA makes use of channel attention, pixel attention, and residual connection to enhance the multi-patch layered network's cross-level feature function strategy. The DA module has strong feature fusion capabilities, as demonstrated by numerous ablation tests.

3. By calculating picture depth information and inserting 3D Berlin noise of various frequencies, 2225 pairs of non-homogeneous haze/clear images datasets are constructed based on the actual situation. The dataset can, as closely as possible, mimic the characteristics of haze dispersal in mountainous regions. Later, it is employed to support DAMPHN training and testing, which can enhance the ability of UAV inspection photos of transmission lines in mountainous locations to remove fog.

Figure 1 illustrates the specifics of our implementation strategy for DAMPHN-based image preprocessing of mountain areas' transmission channel images. Based on this, Section 2 details the DAMPHN network structure. It also includes the encoder-decoder and DA module's unique construction and the loss function needed for network training. The datasets required for the ablation and application experiments and the creation of the training parameters are described in Section 3. The usefulness of the suggested DA and DAMPHN is first demonstrated in Section 4 through several ablation experiments, after which many algorithms are trained and tested using real haze photos of mountain power transmission routes and UAV-HAZE datasets. Section 5 discusses and analyzes the experimental results. In Section 6, several conclusions are made.

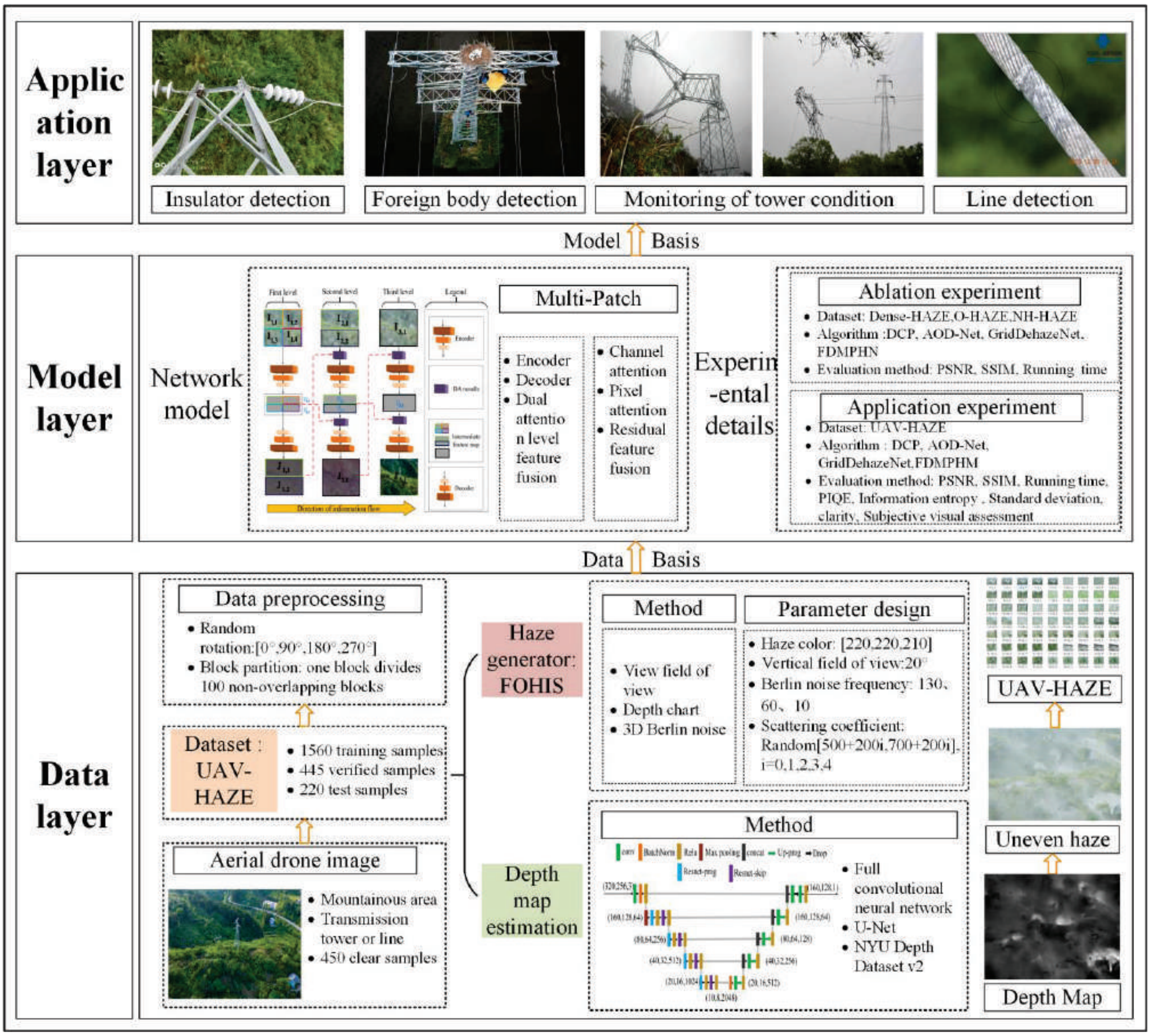

Application layer
Insulator detection
Foreign body detection
Monitoring of tower condition
Line detection
Model Basis
Model layer
Network model
Multi-Patch
• Encoder
• Decoder
• Dual attention level feature fusion
• Channel attention
• Pixel attention
• Residual feature fusion
Experimental details
Ablation experiment
• Dataset: Dense-HAZE,O-HAZE,NH-HAZE
• Algorithm :DCP, AOD-Net, GridDehazeNet, FDMPHN
• Evaluation method: PSNR, SSIM, Running time
Application experiment
• Dataset: UAV-HAZE
• Algorithm : DCP, AOD-Net, GridDehazeNet,FDMPHM
• Evaluation method: PSNR, SSIM, Running time, PIQE, Information entropy , Standard deviation, clarity, Subjective visual assessment
Data Basis
Data layer
Data preprocessing
• Random rotation:[0°,90°,180°,270°]
• Block partition: one block divides 100 non-overlapping blocks
Dataset : UAV-HAZE
• 1560 training samples
• 445 verified samples
• 220 test samples
Aerial drone image
• Mountainous area
• Transmission tower or line
• 450 clear samples
Haze generator: FOHIS
Depth map estimation
Method
• View field of view
• Depth chart
• 3D Berlin noise
Parameter design
• Haze color: [220,220,210]
• Vertical field of view:20°
• Berlin noise frequency: 130、60、10
• Scattering coefficient: Random[500+200i,700+200i], i=0,1,2,3,4
Method
• Full convolutional neural network
• U-Net
• NYU Depth Dataset v2
UAV-HAZE
Uneven haze
Depth Map

Figure 1. Implementation scheme of image preprocessing of mountain transmission channel based on DAMPHN.

2. Materials and Methods

In this study, the encoder-decoder and DA module-based DAMPHN are suggested. This section’s first paragraph introduces DAMPHN’s architecture and design principles, as well as those of its submodules. The training and optimization of the DAMPHN loss function are covered in the second section.

2.1. DAMPHN

DAMPHN network is a multi-level structure, and each level comprises corresponding encoders and decoders. The potential of hierarchical feature fusion is further enhanced by a Dual Attention Level Feature Fusion module (DA). Figure 2 displays the structure in its entirety. Figure 2 depicts DAMPHN with i hierarchical structure, where each level processes 4, 2, and 1 picture blocks, respectively, and when $i = 1, 2, 3$. The j block of level i is represented as $I_{i,j}$ if the input image is I. The first layer then divides I into 4 blocks, identified as $I_{1,1}$, $I_{1,2}$, $I_{1,3}$, and $I_{1,4}$, both vertically and horizontally. I is divided vertically into two blocks, designated as $I_{2,1}$ and $I_{2,2}$, by the second stratum. I is directly inputted into the third layer, which is represented as $I_{3,1}$.

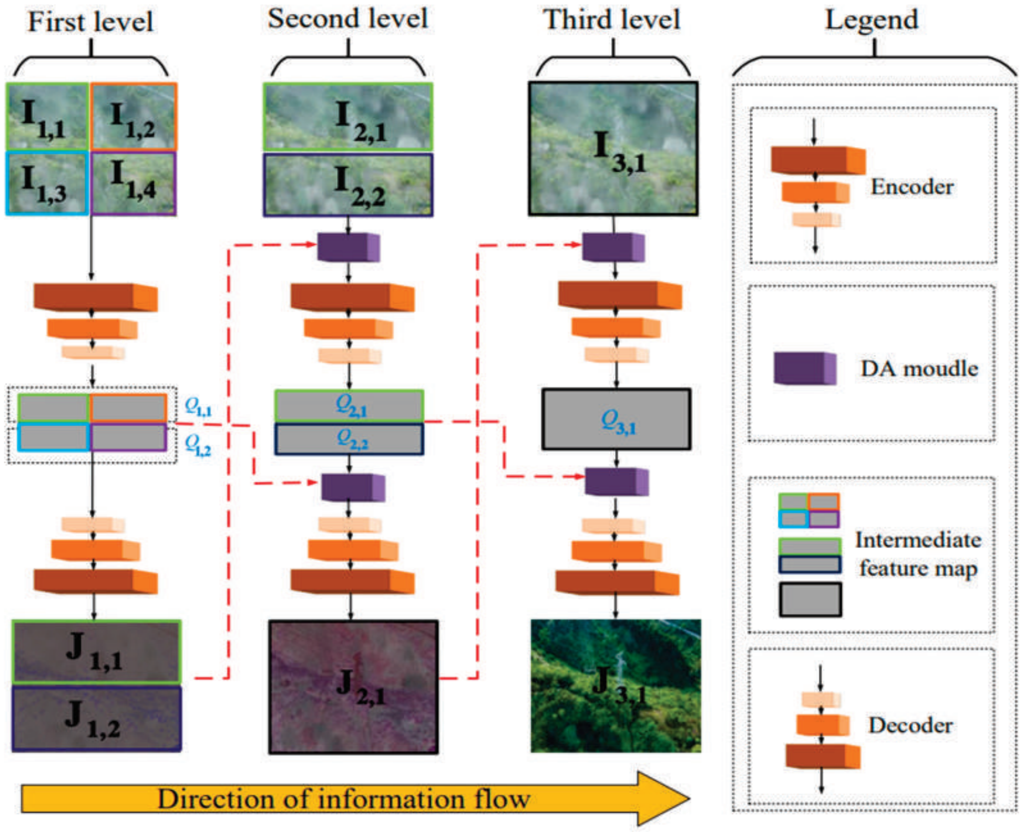


Figure 2. DAMPHN network structure.

The pair of encoder decoders that make up each level are denoted as Enc_i and Dec_i, respectively. The encoding feature $Q_{i,j}$ can be retrieved after the input picture $I_{i,j}$ has sequentially been through the encoder and DA module. In particular, see Equation (3).

$$Q_{i,j} = \begin{cases} Cat\left[Enc_i\left(I_{i,2j-1}\right), Enc_i\left(I_{i,2j}\right)\right], i = 1, j\epsilon 1, 2 \\ Enc_i\left(DA\left(I_{i,j}, J_{i-1,j}\right)\right), i = 2, j\epsilon 1, 2 \\ Enc_i\left(DA\left(I_{i,j}, J_{i-1,j}\right)\right), i = 3, j = 1 \end{cases} \quad (3)$$

The local feature output $J_{i,j}$ of all levels can be acquired after the DA module and decoder. $J_{3,1}$ represents the final dehazing image after DAMPHN feature extraction from the local to the overall concept. The specifics are presented in Equation (4):

$$J_{i,j} = \begin{cases} Dec_i\left(Q_{i,j}\right), i = 1, j\epsilon 1, 2 \\ Dec_i\left(DA\left(Cat\left[Q_{i,j}, Q_{i,2j}\right], Cat\left[Q_{i-1,j}, Q_{i-1,2j}\right]\right)\right), i = 2, j = 1 \\ Dec_i\left(DA\left(Q_{i,j}, Cat\left[Q_{i-1,j}, Q_{i-1,2j}\right]\right)\right), i = 3, j = 1 \end{cases} \quad (4)$$

2.1.1. Encoder-Decoder

The encoder is used to extract the feature data from the image, while the decoder reconstructs the image using the feature data. Three convolution layers and three residual modules (Resblock × 3) make up the encoder in this study. The decoder has a similar design to the encoder, with three residual modules, two transposed convolution layers, and one convolution layer. In order to generate a haze-free image and restore the image scale, decoder transposition convolution is utilized. Figure 3 depicts its network structure.

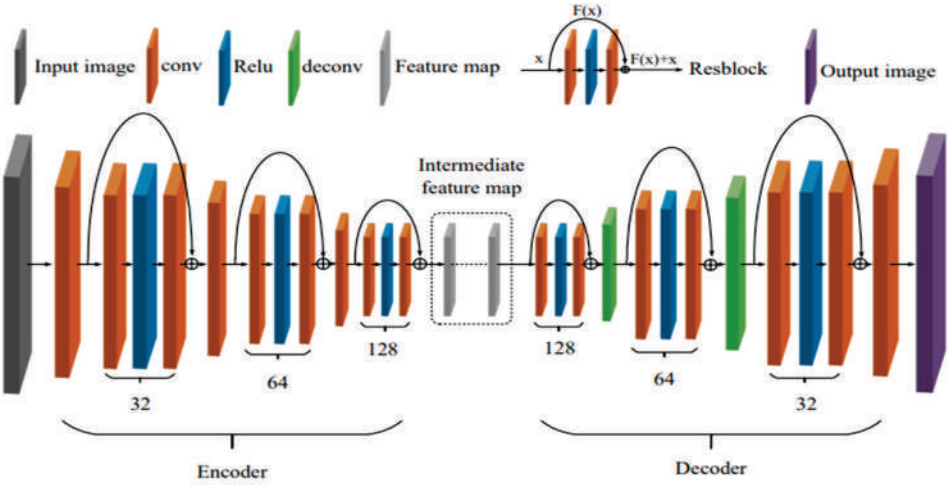


Figure 3. Encoder-decoder module structure.

2.1.2. DA Module

After going through the encoder-decoder during the hierarchical fusion process, the local feature $J_{i,j}$ is produced from the foggy picture I input at the first and second levels. The convolution transformation of $Q_{i,j}$ yields each channel of $J_{i,j}$. As a result, the residual connection in the original FDMPHN network is employed directly in cross-level fusion, and the uneven and redundant channel direction in the fusion feature process is not considered. Additionally, the residual splicing method does not consider the uneven distribution of picture pixels, and the encode-decoder in the original FDMPHN network relies on pixel domain mapping to understand the intricate relationship between the hazy image and the clear image. This led to the development of the DA module provided in this paper, as seen in Figure 4.

The channel domain feature response is first collected by adding the channel attention layer, and subpar or duplicated features are suppressed. Second, by including a pixel attention layer to concentrate on regions of the image with uneven pixel distribution, we may enhance the fusion process' attention to dense haze or high-frequency regions. After stitching, input the channel attention layer (Ca_layer) and pixel attention layer (Pa_layer), assuming that the feature picture of the current level is $F_C \epsilon R^{H\times W\times C}$ and the feature picture of the previous level is $F_U \epsilon R^{H\times W\times C}$. $F_{CA} \epsilon R^{H\times W\times C}$ and $F_{PA} \epsilon R^{H\times W\times C}$ are obtained. Finally, this paper obtains the output F of the final DA module using the convolution joint processing channel and the outcomes of pixel attention processing to make up for the information lost in the extraction process of dual attention layers.

$$F_{CA} = Ca\_layer(Cat[F_C, F_U]) \quad (5)$$

$$F_{PA} = Pa\_layer(F_{CA}) \tag{6}$$

$$F = Cat[conv(F_{PA}), F_{PA}, F_{CA}, Cat[F_C, F_U]] \tag{7}$$

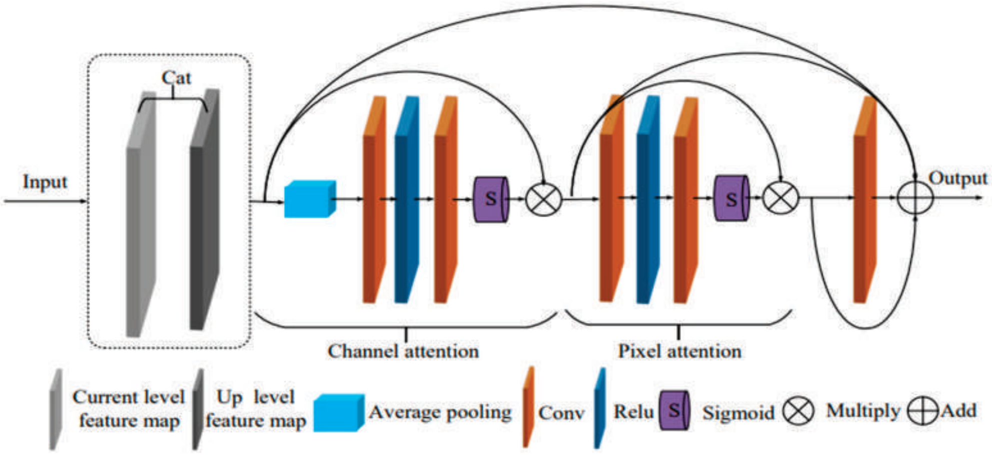


Figure 4. DA module structure.

2.2. Loss of DAMPHN

The total loss function L of DAMPHN is shown in Equation (8), where, respectively, L_r, L_p, and L_{tv} stand for reconstruction loss, perception loss, and total variational loss.

$$L = \alpha_r L_r + \alpha_p L_p + \alpha_{tv} L_{tv} \tag{8}$$

- Reconstruction loss L_r;

Determine the difference between the clear pictures J pixel and the N DAMPHN defogging images J_n. MAE and MSE are combined linearly. L_r can be written as:

$$L_r = \alpha_{r1} \frac{1}{N} \sum_{i=1}^{N} \| J_n - J \| + \alpha_{r2} \frac{1}{N} \sum_{i=1}^{N} \| J_n - J \|^2 \tag{9}$$

- Perception loss L_p;

The VGG16 network was used to calculate features using the pre-trained model. The network's convolution layers (Conv1-2, Conv2-2, and Conv3-2) were utilized to calculate differences, designated as $\varphi(\cdot)$, and extract features. L_p is written as:

$$L_p = \frac{1}{C_K W_K H_K} \sum_{K=1}^{3} \| \varphi_K(J_n) - \varphi_K(J) \| \tag{10}$$

- Total variation loss L_{tv}.

L_{tv} is calculated by computing the gradient amplitude of the dehazing image to reduce noise and keep the image smooth. $\nabla_x(\cdot)$ and $\nabla_y(\cdot)$ in Equation (11), respectively, are used to obtain the gradient matrix of the picture in the horizontal and vertical directions.

$$L_{tv} = \| \nabla_x(J_n) \|_2 + \| \nabla_y(J) \|_2 \tag{11}$$

3. Experiment Setup

3.1. Dataset

3.1.1. Ablation Experimental Dataset

The datasets for the ablation experiment were chosen from three standard datasets from the IEEE CVRP NTIRE Seminar: Dense-HAZE [31], O-HAZE [32], and NH-HAZE [33].

Dense-HAZE includes 55 identical pairs of dense haze/clear images. From the sample, 1–45 pairings were chosen for training, 46–50 pairs for verification, and 51–55 pairs for testing in this study. O-HAZE includes 45 sets of outdoor, non-homogeneous haze/clear images. From that set, 1–35 pairs were chosen for training, 36–40 pairs for verification, and 41–45 pairings for testing in this study. Fifty-five non-homogeneous haze/clear image pairs are included in NH-HAZE. In this study, 1–45 were selected for training, 46–50 for verification, and 51–55 for testing.

3.1.2. Self-Built Transmission Channel Inspection Dataset (UAV-HAZE)

In haze image imaging, because it is often manifested as loss of image visibility, the atmospheric extinction coefficient σ can solve the $\beta(\lambda)$ in Equation (12).

$$\beta(\lambda) = \frac{3.912}{\sigma} \tag{12}$$

Additionally, visibility varies depending on height. Therefore, the depth value of the scene and the vertical field of view of the camera are used to estimate the elevation values of the pixels and their distribution characteristics are calculated to replicate the distribution and color characteristics of genuine haze. To imitate the color features of haze, Formula (1) includes the haze color value I_{al} as follows:

$$I(x,\lambda) = t(x)J(x,\lambda) + A(1 - t(x)) \times I_{al} \tag{13}$$

Taking into account the mountain haze's irregularly distributed properties. Non-uniform haze is created using 3D Berlin noise, and a haze generator called FOHIS [34] is suggested. They are used to mimic non-uniform haze by making three Berlin noises of varying amplitudes and frequencies, which are then merged with Equation (13) and multiplied by $\beta(\lambda)$.

$$P\_noise = \frac{1}{3}\sum_{i=1}^{3}\frac{P\_noise_i}{2^i - 1} \tag{14}$$

In light of FOHIS, this work estimated the picture depth value in order to synthesize the mountain transmission into the UAV-HAZE dataset [35]. In the synthesis process, the I_{al} of the three-color channels of the image RGB is set to [220,220,210], respectively, to simulate the color characteristics of the blue-white mountain fog. Then, to imitate the distribution features of mountain haze, the vertical field of view of the camera is adjusted to 20°. This is combined with the depth value of picture pixels, and the pixel elevation value is calculated. The non-uniform properties of mountain haze were then simulated by creating 3D Berlin noise with three distinct frequency values (f = 130, 60, 10). Finally, the data [700–900], [900–1100], [1100–1300] and [1300,1500] were chosen as the extinction coefficients in Equation (12) using 450 mountain transmission channel photos obtained by UAV inspection as the original dataset. A total of 2225 non-uniform simulated haze/clear images of various concentrations make up UAV-HAZE, which is divided into training sets, verification sets, and test sets in a ratio of 7:2:1. There are 1560 pairs in the training set, 445 pairs in the verification set, and 220 teams in the test set.

3.2. Implementation Details

NVIDIA GeForce RTX3090 (24 GB) was the platform used for the experiment. Data preprocessing involves cropping each training image into 100 non-overlapping image blocks with a size of 120 × 160 pixels and unifying the image resolution of the training set across Dense-HAZE, O-HAZE, NH-HAZY, and UAV-HAZE to 1200 × 1600. The image blocks were simultaneously rotated at random angles of 0, 90, 180, and 270 degrees. The Adam optimizer is initially employed in DAMPHN network training with exponential decay rates γ_1 = 0.9, γ_2 = 0.999, starting learning rates 1×10^{-4}, and batch sizes 100. We also adjusted the learning rate using an equally spaced strategy with step size = 10 and gamma = 0.1. Then, the hyperparameters of the loss function are set to α_r = 1, $\alpha_p = 6 \times 10^{-3}$,

α_{tv} = 2 × 10^{-8}, α_{r1} = 0.6, α_{r2} = 0.4. Finally, when the verification set loss function is stable, the training is stopped and the best model is obtained.

4. Experiment Results

4.1. Ablation Experiment

Two phases of the ablation experiment were conducted. The first and second sections, respectively, confirm the reliability of the DA module and the DAMPHN network.

4.1.1. DA Module

Due to the low cross-level fusion quality of the original multi-patch algorithm FDMPHN, the DA module is proposed in this study. In order to reduce the complexity of the algorithm, the encoder-decoder structure of FDMPHN is diminished. The three sets of experiments listed below are explicitly included in this section:

(I) The network encoder-decoder has six residual modules (Resblock × 6) using only FDMPHN.

(II) The approach suggested in this work builds on (I) by adding a DA module (FDMPHN + DA). A DA module plus six residual modules (Resblock × 6) make up the network encoder-decoder.

(III) To optimize (II) and DAMPHN, the solution presented in this research uses just three residual modules (Resblock × 3).

- Quantitative evaluation

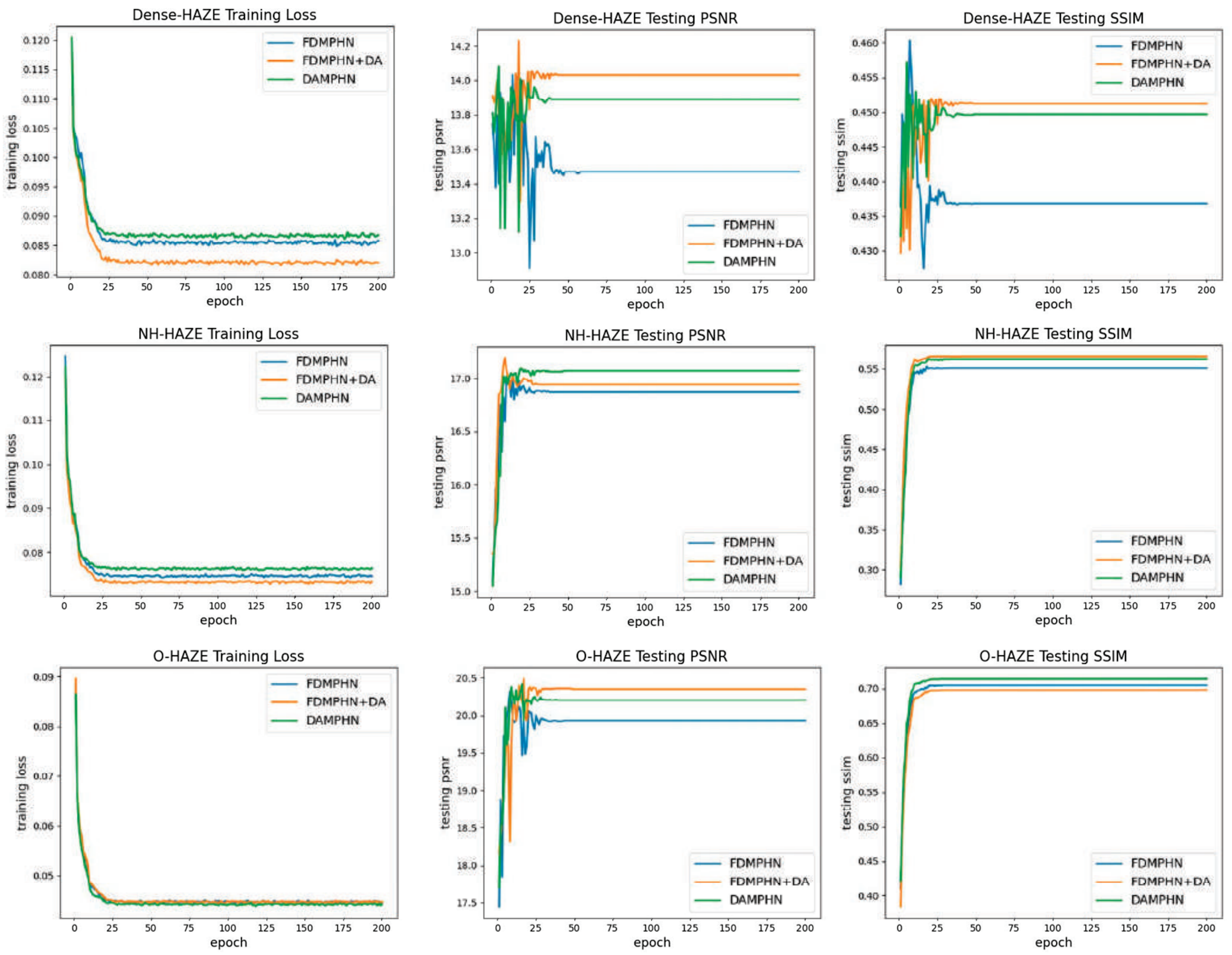
PSNR [36], SSIM [37], and APT were chosen for quantitative evaluation in this section of the experiment. The visual noise and distortion decrease as the PSNR value rises. The recovery of structural properties such as image brightness and contrast is measured by SSIM. The dehazing is better the higher the value. Table 1 displays the precise outcomes of the three groups of studies. In Table 1, when (I) and (II) are compared, the addition of the DA module raised PSNR and SSIM in the three datasets by an average of 0.35 dB and 0.0073, whereas APT rose by 19% (0.007 s). Comparing (I) and (III), the average PSNR and SSIM in the three datasets are raised by 0.30 dB and 0.011, respectively, and APT is shortened by 11% (0.003 s), respectively, after the encode-decoder structure is streamlined.

Table 1. Results of DA module ablation experiments.

Method		Dense-HAZE			O-HAZE			NH-HAZE		
		PSNR	**SSIM**	**APT**	**PSNR**	**SSIM**	**APT**	**PSNR**	**SSIM**	**APT**
(I)	FDMPHN	13.47	0.4369	0.031	19.93	0.7045	0.030	16.87	0.5512	0.030
(II)	FDMPHN + DA	14.03	0.4512	0.036	20.35	0.6976	0.035	16.94	0.5656	0.037
(III)	DAMPHN	13.89	0.4497	0.027	20.20	0.7138	0.027	17.07	0.5621	0.027

- Convergence analysis

This section assessed the convergence using the dynamic curves for training loss, PSNR, and SSIM. On Dense-HAZE, O-HAZE, and NH-HAZE, Figure 5 displays the training losses, PSNR, and SSIM for the FDMPHN, FDMPHN+DA, and DAMPHN approaches, respectively. Figure 5 shows the training and testing of the three approaches on three separate datasets, with the training losses, PSNR, and SSIM information displayed in the rows and columns, respectively. Figure 5a illustrates how the training loss for the aforementioned approaches steadily lowers as the number of iterations increases and gradually stabilizes at 35–40 rounds. In Figure 5b,c, all three approaches converge after 200 rounds, and the DA module performs better regardless of how complicated or straightforward the encoder-decoder structure is.



(**a**) (**b**) (**c**)

Figure 5. Training loss curve and test PSNR and SSIM curve. (**a**) Training loss. (**b**) Testing PSNR. (**c**) Testing SSIM.

4.1.2. DAMPHN Network

To more accurately evaluate DAMPHN, we further conducted quantitative, qualitative, and convergence evaluation on three datasets, Dense-HAZE, O-HAZE, and NH-HAZE, with DCP [6], AOD-Net [9], FDMPHN [14], and GridDehazeNet [17], respectively.

- Quantitative evaluation

PSNR, SSIM, and APT are also used to gauge how well various techniques remove haze. The outcomes of the quantitative comparison are displayed in Table 2. In Table 2, the blue values represent the optimal values, and the underlined values represent the sub-optimal values. In the three datasets, the PSNR and SSIM values of DAMPHN are 3.72 dB and 0.0666 higher than those of DCP on average, and ART is 94% shorter. The defog quality of AOD-Net in the Dense-HAZE dataset is comparable to that of DAMPHN. However, on the non-uniform haze datasets O-HAZE and NH-HAZE, the PSNR and SSIM values of DAMPHN are increased by 1.72 dB and 0.0446 compared with the average value of AOD-Net. The effect of GridDehazeNet on the fog removal in the three datasets has its own advantages compared with the method in this paper. Specifically, DAMPHN is, on average, 0.38 dB higher than GridDehazeNet's PSNR value, but the SSIM value is lower than GridDehazeNet's 0.025. Finally, compared with FDMPHN in the three datasets, the

PSNR and SSIM values of DAMPHN are increased by 0.30 dB and 0.011 on average, and ART is shortened by 11%.

Table 2. Results of DAMPHN Network quantitative comparison.

Method	Dense-HAZE			O-HAZE			NH-HAZE		
	PSNR	SSIM	APT	PSNR	SSIM	APT	PSNR	SSIM	APT
DCP [6]	11.60	0.3854	0.406	15.66	0.6753	0.440	13.28	0.4650	0.416
AOD-Net [9]	13.85	0.4714	0.023	18.19	0.6950	0.010	15.64	0.4918	0.009
GridDehazeNet [17]	13.50	0.4721	0.026	19.82	0.7108	0.026	16.70	0.6101	0.026
FDMPHN [14]	13.47	0.4369	0.031	19.93	0.7045	0.030	16.87	0.5512	0.030
DAMPHN (ours)	13.89	0.4497	0.027	20.20	0.7138	0.027	17.07	0.5621	0.027

- Qualitative assessment

The experiment's visual comparison component is the main focus here. Among the images, the haze distribution in the first and second rows is more uniform, and the haze distribution in the third and fourth rows is uneven. The DCP results in Figure 6 reveal color distortion and a significant degree of residual haze. The image's color changes to dark yellow after AOD-Net fog removal, and a significant quantity of haze residue remains in the non-uniform haze area. GridDehazeNet has a good fog effect when the haze distribution is relatively uniform, but the image's color after fog removal is darker than that of the clear picture. In addition, in the case of non-uniform haze, GridDehazeNet also shows many haze residues. The image's overall color after fog removal by FDMPHN is closer to the clear image when the haze distribution is more uniform. Still, the color distortion appears on the ground of the first line of the picture. Regarding non-uniform haze, FDMPHN has a good de-fogging effect, but its de-noising solid ability also causes image smoothing, resulting in blurred details. DAMPHN is visually similar to FDMPHN. However, in the enlarged area of the fourth row of the image, the DAMPHN haze residue is less.

- Convergence analysis

In this experiment section, the convergence is assessed using the change curves of PSNR and SSIM with the number of training rounds. Figure 7 shows the results of each round of PSNR and SSIM tests for four de-fogging techniques on three datasets. DCP has the fastest convergence rate. AOD-Net uses a relatively lightweight CNN structure in the parameter estimation process, which has poor stability and the slowest convergence rate. When the PSNR value of the current verification set is assumed to be greater than the previous results during GridDehazeNet training, the round model is optimal. Under dynamic control, its convergence rate ranks fourth. The FDMPHN and DAMPHN set the hyperparameters before training, and the validation set is used to optimize the hyperparameter settings. Therefore, both FDMPHN and DAMPHN converge faster. Specifically, in Figure 7a, DAMPHN converges faster than FDMPHN. In Figure 7b,c, FDMPHN and DAMPHN converge at similar speeds. Therefore, DAMPHN in this paper is in second place in terms of convergence speed.

4.2. Transmission Channel Image

4.2.1. Synthetic Dataset UAV-HAZE

DAMPHN can be utilized to clear haze from Sichuan's mountainous areas' transmission channel scenery. This section is based on the dataset created in Section 3.1.2, UAV-HAZE. With this collection of data, DCP [6], AOD-Net [9], FDMPHN [14], GridDehazeNet [17], and DAMPHN, the approach in this article, are each examined in turn. This section evaluates both the algorithm's quantitative and qualitative performance.

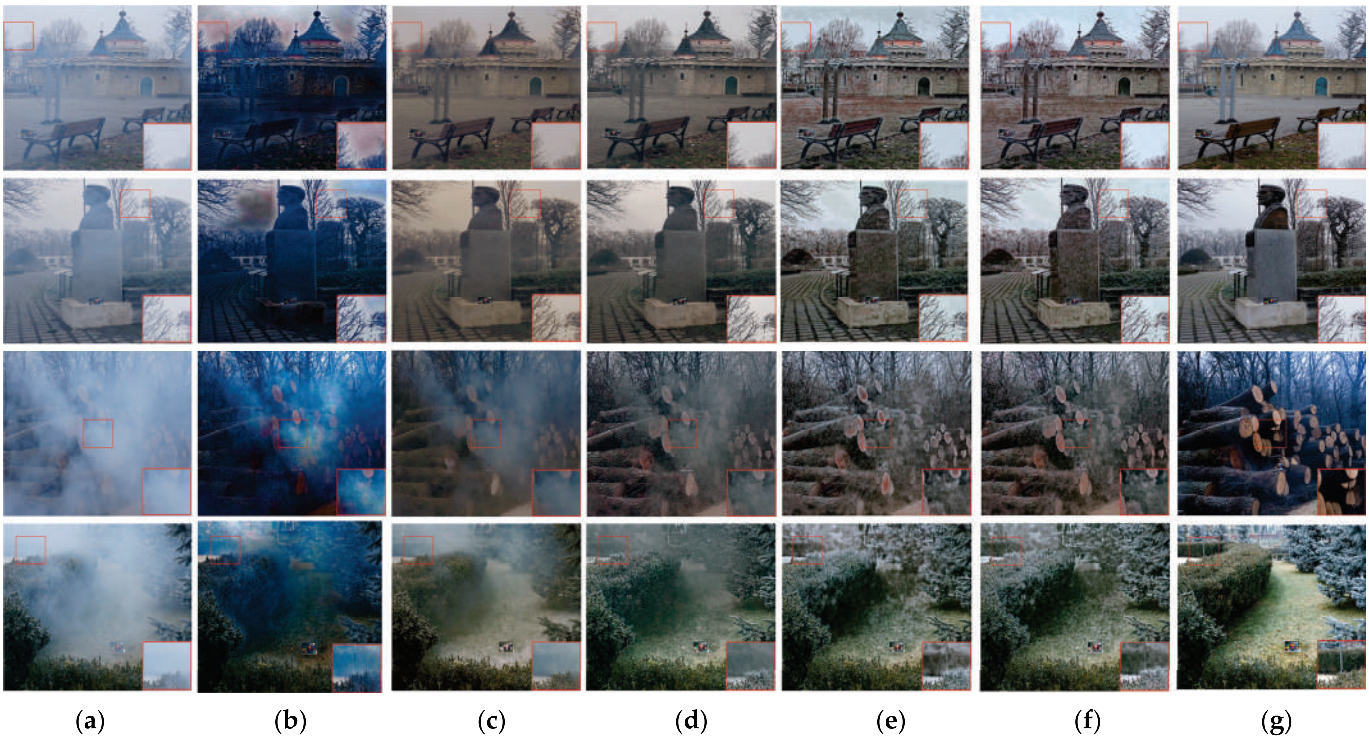


Figure 6. NH-HAZE and O-HAZE dehazing results. (**a**) Hazy. (**b**) DCP. (**c**) AOD-Net. (**d**) GridDehazeNet. (**e**) FDMPHN. (**f**) DAMPHN. (**g**) Ground truth.

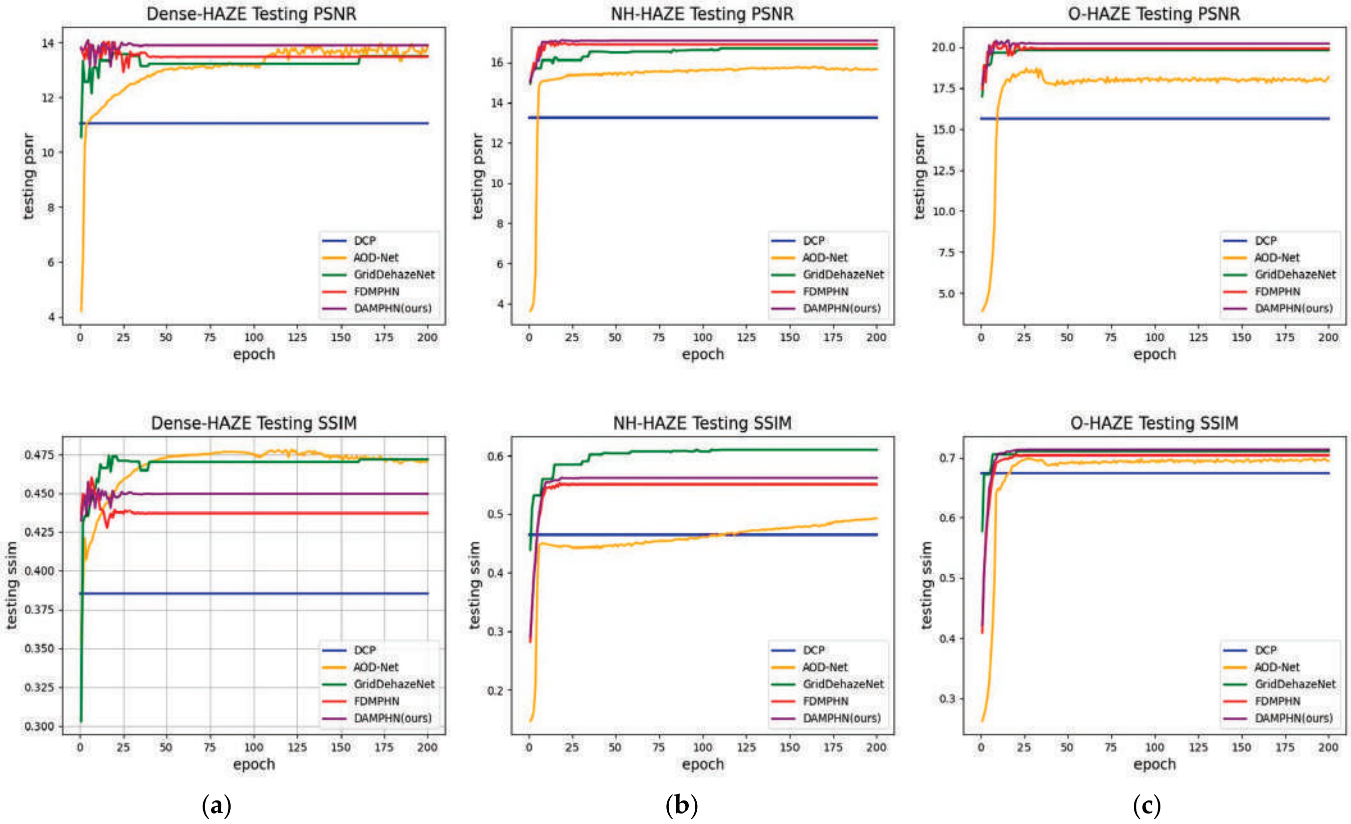


Figure 7. PSNR and SSIM test curves. (**a**) Dense-HAZE. (**b**) NH-HAZE. (**c**) O-HAZE.

- Quantitative evaluation

PSNR, SSIM, and APT were chosen as evaluation indicators. Table 3 presents the experimental outcomes. In Table 3, the blue font is the optimal value, and the underlined value is the sub-optimal value. The PSNR of DAMPHN is optimal, SSIM and ART are sub-optimal. In this study, DAMPHN's PSNR and SSIM values are 7.26 dB and 0.0588 greater than DCP's, respectively. APT barely makes up 4% of DCP techniques. PSNR and SSIM are 9.32 dB and 0.2057 greater in DAMPHN than in AOD-Net, although APT is 14 times higher. The PSNR value of DAMPHN is 0.26 dB higher, and the SSIM value is 0.0007 dB lower than GridDehazeNet. DAMPHN's SSIM value is the same as FDMPHN's, but its PSNR is 0.04 dB higher, and its APT is 94% shorter.

Table 3. Quantitative comparison results on UAV-HAZE.

Method	PSNR	SSIM	APT
DCP [6]	19.97	0.8851	0.352
AOD-Net [9]	17.92	0.7382	0.001
GridDehazeNet [17]	26.98	0.9476	0.015
FDMPHN [14]	<u>27.20</u>	<u>0.9439</u>	0.234
DAMPHN (ours)	27.24	<u>0.9439</u>	<u>0.014</u>

- Qualitative assessment

Figure 8 displays the outcomes of the qualitative comparison between DAMPHN and the techniques mentioned above. DCP has a positive impact in the mist area, according to the analysis of Figure 8. The color of the third row seems distorted when the haze density is excellent, or the randomness of its distribution features is substantial. When dealing with non-uniform haze, AOD-Net's primary result is that a significant amount of haze is left in the processed image, the details are blurred, and there is evident color distortion. The fog removal quality of GridDehazeNet is superior to that of the first two techniques. However, some fog was still present close to the first row's wires and the fourth row's poles and towers. In this study, the FDMPHN and DAMPHN techniques can recover the picture tower's detailed information with excellent clarity and superb color fidelity. FDMPHN does, however, have a trace amount of haze residue in the first row's wire area.

4.2.2. Real Image

The actual utility of DAMPHN was confirmed by the refit project from Gangu to Erlang Mountain in Shuzhou and the real hazy photographs of the Sichuan-Tibet network project. The approach was evaluated using both quantitative and qualitative methodologies.

- Quantitative evaluation

Five non-reference image quality evaluation indexes, including information entropy, standard deviation, clarity, perception-based image quality evaluation method (PIQE) [38], and APT, were chosen for quantitative evaluation because there were insufficient clear reference examples. The more relevant information an image carries, the higher its information entropy. The image's standard deviation is used to assess its contrast; the lower the standard deviation, the more stable the image is. The greater the value, the higher the sharpness, which is defined as the variance of calculating the absolute value of Laplace. Block effects, blur, and noise distortion are calculated using PIQE, and a lower value corresponds to less distortion. In Table 4, the experimental findings are displayed.

In Table 4, the underlined value and the blue text represent the ideal and sub-optimal values, respectively. This approach performs the best regarding clarity and PIQE, comes in second for ART, and comes in third for information entropy and standard deviation. This approach has reduced standard deviation and higher assessment indices compared to DCP. The proposed method has a clear benefit over AOD-Net regarding image quality, but it takes four times as long to operate. DAMPHN has higher evaluation indexes than

GridDehazeNet, except for lower information entropy. DAMPHN is superior to FDMPHN in various assessment indices compared to FDMPHN before improvement, except for the picture information entropy, which is less than 0.17.

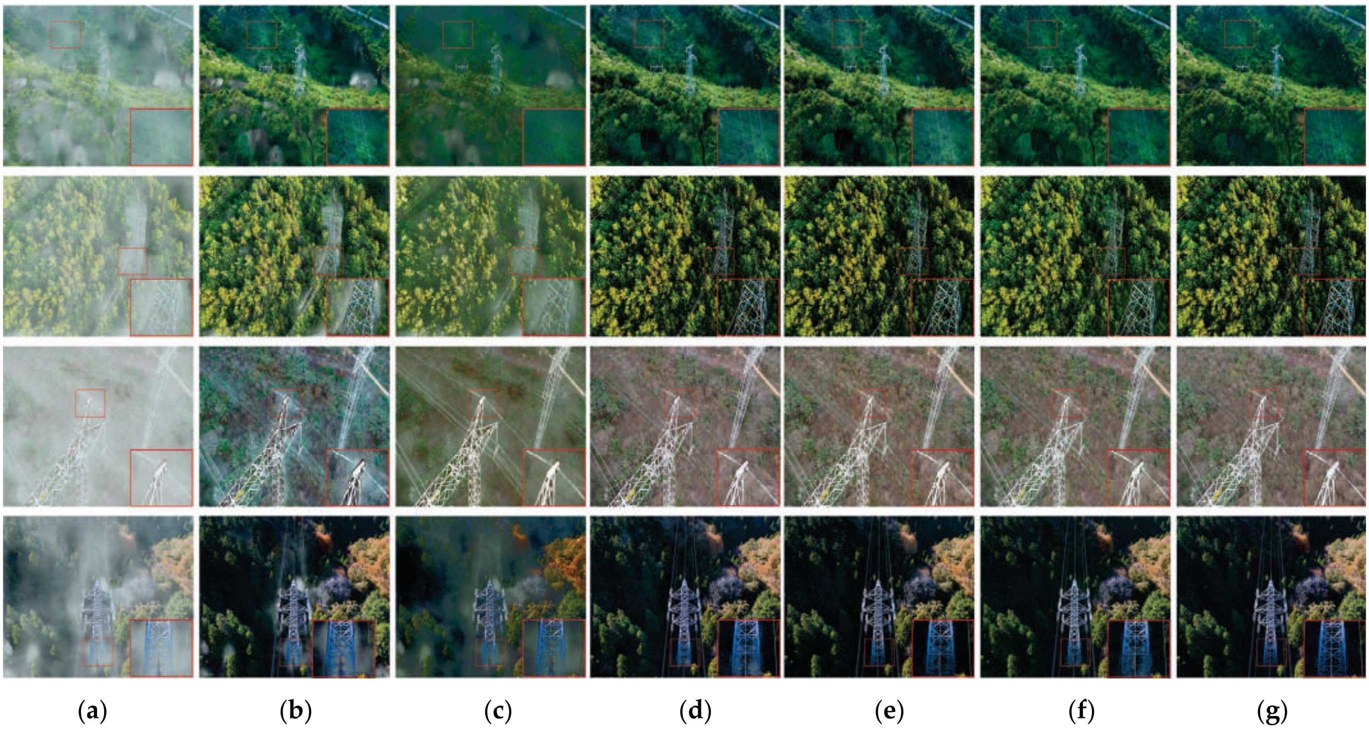


Figure 8. Results of UAV-HAZE dehazing. (**a**) Hazy. (**b**) DCP. (**c**) AOD-Net. (**d**) GridDehazeNet. (**e**) FDMPHN. (**f**) DAMPHN. (**g**) Ground truth.

Table 4. Results of quantitative evaluation of real images.

Method	Information Entropy	Standard Deviation	Clarity	PIQE	APT
DCP [6]	17.78	32.19	459.86	27.51	0.342
AOD-Net [9]	16.10	45.93	452.79	28.87	0.005
GridDehazeNet [17]	18.28	<u>41.61</u>	<u>470.18</u>	24.90	0.021
FDMPHN [14]	<u>18.10</u>	42.38	465.21	<u>24.48</u>	0.270
DAMPHN (ours)	17.93	41.92	536.11	23.98	<u>0.020</u>

- Qualitative assessment

Figure 9 displays two transmission channel views of the retrofitting project from Gangu to Erlang Mountain in Shuzhou and the haze reduction effect of four groups of the Sichuan-Tibet interconnection project. Uphill fog, uphill fog, advection fog, and radiation fog are all depicted in lines 1 through 4. Intuitive examination reveals that the color of DCP is severely altered and turns blue-purple in the sky area. AOD-Net effectively removes haze. However, it has glaring issues with blurred details and intensified hue. Although GridDehazeNet effectively removes fog, there is still some fog in the third-row valley and second-row tower areas. The image is also slightly lavender once the fog has been eliminated, for instance, the first row's valley fog area and the fourth row's pole tower area. In places with high haze density, such as the tower area in the second row and the valley area in the third row, FDMPHN has a competitive dehazing impact but leaves haze residue behind. This technique also results in color distortion, as seen in how the first row of trees on an ascent turned yellow. After adding a DA module, DAMPHN may now pay closer attention to areas with dense fog and a non-uniform haze. As a result, the method suggested in this paper removes fog more thoroughly than GridDehazeNet and FDMPHN in the first-row and third-row valley areas. Additionally, there is no purple or yellowing in terms of color preservation.

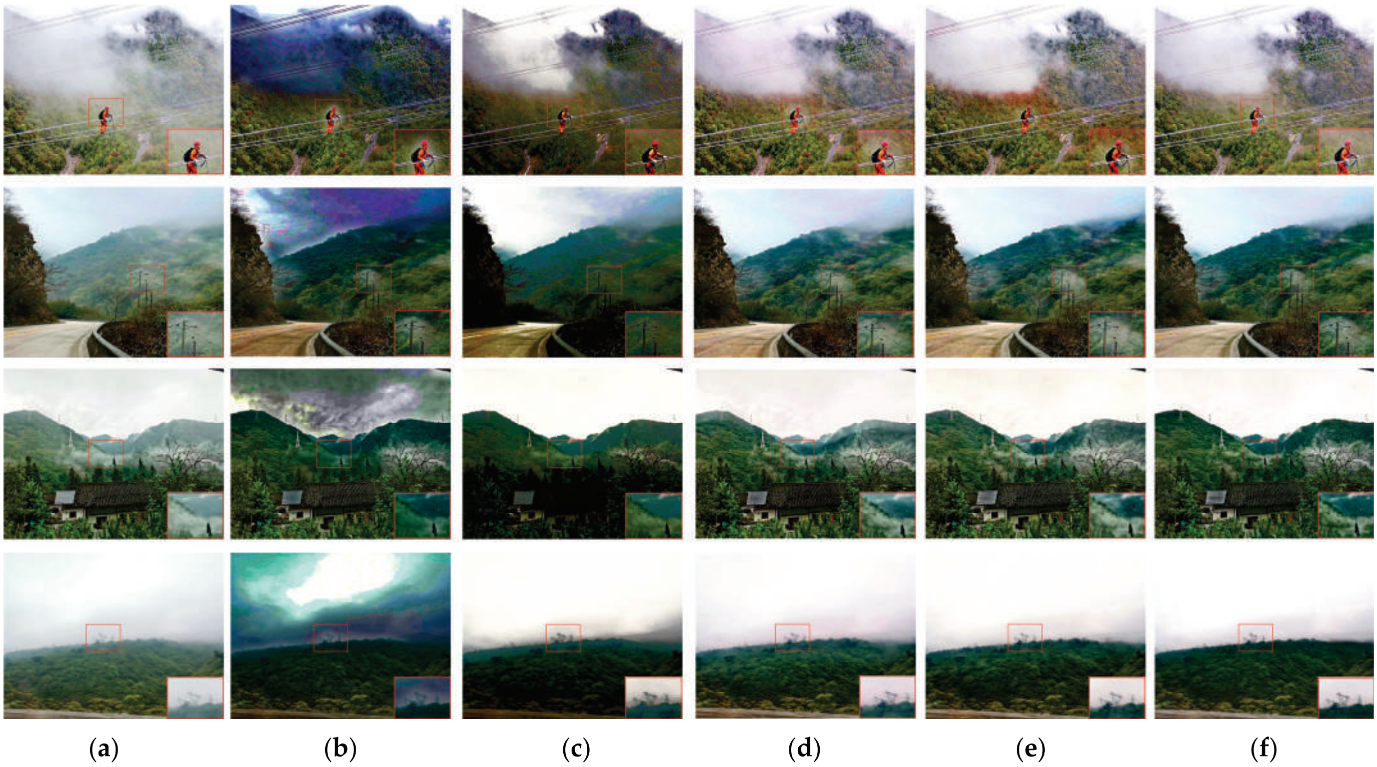


Figure 9. Dehazing result of real transmission channel image. (**a**) Hazy. (**b**) DCP. (**c**) AOD-Net. (**d**) GridDehazeNet. (**e**) FDMPHN. (**f**) DAMPHN.

5. Discussion

In this paper, the issue of transmission line haze that is unevenly dispersed in mountainous places was studied. A DAMPHN is introduced, an innovative non-uniform haze-defogging network model put forth in this research to facilitate picture preprocessing for UAV transmission channel inspection in mountainous terrain. Similarly, the DAMPHN network model is universal. DAMPHN can be used for preprocessing other images in fog environments, such as unmanned visual perception, surveillance video (road traffic, transmission lines), and tachographs. DCP, AOD-Net, GridDenzeNet, and FDMPHN were utilized in numerous tests using open datasets (Dense-HAZE, O-HAZE, and NH-HAZE) and self-built datasets (UAV-HAZE) to demonstrate the efficacy of DAMPHN.

Notably, because the assumption of uniform distribution of air concentration in the atmospheric scattering model limits both DCP and AOD-Net, the error of estimating parameters is significant in dense fog and non-homogeneous haze. DAMPHN is a multi-level end-to-end fog removal network that seeks to remove fog by discovering the relationship between the haze and clear image mapping. DAMPHN does not, therefore, need to estimate the parameters; instead, it relies on the dataset's basis, and the higher the base, the higher the quality of fog removal. GridDehazeNet solves the problem of feature fusion between different scales in multi-scale networks by introducing channel attention. DAMPHN solves the problem of feature fusion between different levels in multi-patch networks by introducing channel and pixel attention mechanisms. GridDehazeNet has vital artifact removal, so the SSIM value is stronger than DAMPHN. DAMPHN pays attention to the problem of uneven pixel distribution, pays attention to the removal of non-uniform fog, and has a strong denoising ability and high PSNR value. FDMPHN is identical to a multi-patch defogging network, but the residual connections in hierarchical fusion restrict how well it can fuse features. The pixel attention layer of the DAMPHN's DA module is designed to pay attention to areas with unequal haze distribution. In contrast, the channel attention layer is designed to appropriately evaluate the channel domain properties. DAMPHN has a better defogging impact as a result than FDMPHN.

Additionally, the frequently used image segmentation algorithms U-Net and GridNet have produced effective outcomes in image segmentation and picture defogging via innovation. DCPDN solves parameter A using the U-Net network. GridDehazeNet proposes a multi-scale attention network based on GridNet. They both have superior defogging effects. With dual U-Net, Amyar et al. [39] created a multi-task and multi-scale network structure that was effectively used for lung tumor segmentation, classification, and prediction. However, DAMPHN accomplishes picture fog removal from the local to the global by helping the feature extraction of the bigger patch image from the top layer with the detailed feature of the lower layer. From the overall to the local picture segmentation, image fog removal, and other tasks, U-Net will employ the more comprehensive information collected from the bottom layer to aid in the development of smaller receptive field information. Consequently, the two networks' designs have produced successful outcomes in their respective domains.

In conclusion, the DAMPHN approach offers an excellent defogging effect, less color distortion, and quick processing speed. In a location with a lot of fog, it is impossible to eliminate it entirely, and the details are hazy. DAMPHN can improve the defog effect by enhancing the encoder-decoder structure, feature extraction, and reconstruction skills, all of which were influenced by U-Net in the field of image segmentation, or by combining with the conventional image edge previous knowledge to increase the texture information and boost the fog removal effect.

6. Conclusions

This paper proposes that DAMPHN can achieve a good defog effect and restore the color and brightness of the image. The network encoder-decoder module and DA module are composed. The former can learn the mapping relationship between haze and clear pictures and has a strong feature extraction ability. The latter enhances the feature fusion ability by empowering the combination of channel attention and pixel attention. However, in excessive haze density, it cannot be entirely removed, and the details are hazy. Future work will improve the haze removal effect by enhancing texture information through edge prior and enhancing the encoder-decoder structure. Additionally, using 3D Berlin noise and image depth information to simulate haze's non-uniform distribution characteristics is not only just restricted to UAV mountain transmission channel inspection; it can also be applied to a broader range of situations to enhance generalization performance.

Author Contributions: Conceptualization, methodology, and writing—review and editing, W.Z.; software, validation, data curation, and writing—original draft preparation, L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the following project: Innovation and entrepreneurship training program for college students of China: "Vision Engine—Leader of Clear Processing of unmanned image".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle
FDMPHN	Fast Deep Multi-Patch Hierarchical Network
DAMPHN	Dual Attention Level Feature Fusion Multi-Patch Hierarchical Network
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
APT	Average Processing Time
DCP	Dark Channel Prior
CAP	Color Decay Prior
CNN	Convolutional neural networks
AOD-Net	All-in-One Dehazing Network
DCPDN	Densely Connected Pyramid Dehazing Network
FDMSHN	Fast Deep Multi-Scale Hierarchical Network
DA	Dual Attention Level Feature Fusion
Ca_layer	Channel attention layer
Pa_layer	Pixel attention layer
PIQE	Perception-based Image Quality Evaluation

References

1. Li, X.; Li, Z.; Wang, H.; Li, W. Unmanned aerial vehicle for transmission line inspection: Status, standardization, and perspectives. *Front. Energy Res.* **2021**, *9*, 713634. [CrossRef]
2. Zhang, T.; Tang, Q.; Li, B.; Zhu, X. Genesis and dissipation mechanisms of radiation-advection fogs in Chengdu based on multiple detection data. *Meteorol. Sci. Technol.* **2019**, *47*, 70–78.
3. Zhao, L.; Zuo, X.; Zhang, S.; Lu, Y. On restoration of mountain haze image based on non-local prior algorithm. *Electron. Opt. Control* **2022**, *29*, 55–58.
4. Imran, A.; Zhu, Q.; Sulaman, M.; Bukhtiar, A.; Xu, M. Electric-Dipole Gated Two Terminal Phototransistor for Charge-Coupled Device. *Adv. Opt. Mater.* **2023**, 2300910. [CrossRef]
5. Swinehart, D.-F. The beer-lambert law. *J. Chem. Educ.* **1962**, *39*, 333–335. [CrossRef]
6. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353.
7. Zhu, Q.; Mai, J.; Shao, L. A fast single image haze removal algorithm using color attenuation prior. *IEEE Trans. Image Process.* **2015**, *24*, 3522–3533.
8. Cai, B.; Xu, X.; Jia, K.; Qing, C.; Tao, D. DehazeNet: An end-to-end system for single image haze removal. *IEEE Trans. Image Process.* **2016**, *25*, 5187–5198. [CrossRef]
9. Li, B.; Peng, X.; Wang, Z.; Xu, J.; Feng, D. AOD-Net: All-in-one dehazing network. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4770–4778.
10. Zhang, H.; Patel, V.-M. Densely connected pyramid dehazing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3194–3203.
11. Li, Y.; Miao, Q.; Quyang, W.; Ma, Z.; Fang, H.; Dong, C.; Quan, Y. LAP-Net: Level-aware progressive network for image dehazing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3276–3285.
12. Li, R.; Pan, J.; He, M.; Li, Z.; Tang, J. Task-oriented network for image dehazing. *IEEE Trans. Image Process.* **2020**, *29*, 6523–6534. [CrossRef]
13. Bai, H.; Pan, J.; Xiang, X.; Tang, J. Self-guided image dehazing using progressive feature fusion. *IEEE Trans. Image Process.* **2022**, *31*, 1217–1229. [CrossRef]
14. Das, S.-D.; Dutta, S. Fast deep multi-patch hierarchical network for nonhomogeneous image dehazing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 482–483.
15. Zhang, H.; Dai, Y.; Li, H.; Koniusz, P. Deep stacked hierarchical multi-patch network for image deblurring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 5978–5986.
16. Wang, K.; Yang, Y.; Li, B.; Cui, L. Uneven image dehazing by heterogeneous twin network. *IEEE Access* **2020**, *8*, 118485–118496. [CrossRef]
17. Liu, X.; Ma, Y.; Shi, Z.; Chen, J. Griddehazenet: Attention-based multi-scale network for image dehazing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7314–7323.
18. Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; Jia, H. FFA-Net: Feature fusion attention network for single image dehazing. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11908–11915.
19. Wang, C.; Shen, H.-Z.; Fan, F.; Shao, M.-W.; Yang, C.-S.; Luo, J.-C.; Deng, L.-J. EAA-Net: A novel edge assisted attention network for single image dehazing. *Knowl.-Based Syst.* **2021**, *228*, 107279. [CrossRef]

20. Yang, K.; Zhang, J.; Fang, Z. Multi-patch and multi-scale hierarchical aggregation network for fast nonhomogeneous image dehazing. *Comput. Sci.* **2021**, *48*, 250–257.
21. Wang, K.; Duan, Y.; Yang, Y.; Fei, S. Uneven hazy image dehazing based on transmitted attention mechanism. *Pattern Recognit. Artif. Intell.* **2022**, *35*, 575–588.
22. Zhao, D.; Mo, B.; Zhu, X.; Zhao, J.; Zhang, H.; Tao, Y.; Zhao, C. Dynamic Multi-Attention Dehazing Network with Adaptive Feature Fusion. *Electronics* **2023**, *12*, 529. [CrossRef]
23. Guo, Y.; Gao, Y.; Liu, W.; Lu, Y.; Qu, J.; He, S.; Ren, W. SCANet: Self-Paced Semi-Curricular Attention Network for Non-Homogeneous Image Dehazing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1884–1893.
24. Liu, J.; Jia, R.; Li, W.; Ma, F.; Wang, X. Image dehazing method of transmission line for unmanned aerial vehicle inspection based on densely connection pyramid network. *Wirel. Commun. Mob. Comput.* **2020**, *2020*, 1–9.
25. Zhang, M.; Song, Z.; Yang, J.; Gao, M.; Hu, Y.; Yuan, C.; Jiang, Z.; Cheng, W. Study on the enhancement method of online monitoring image of dense fog environment with power lines in smart city. *Front. Neurorobotics* **2022**, *16*, 299. [CrossRef]
26. Zhai, Y.; Jiang, L.; Long, Y.; Zhao, Z. Dark channel prior dehazing method for transmission channel image based on sky region segmentation. *J. North China Electr. Power Univ.* **2021**, *48*, 89–97.
27. Xin, R.; Chen, X.; Wu, J.; Yang, K.; Wang, X.; Zhai, Y. Insulator Umbrella Disc Shedding Detection in Foggy Weather. *Sensors* **2023**, *22*, 4871. [CrossRef]
28. Gao, Y.; Yang, J.; Zhang, K.; Peng, H.; Wang, Y.; Xia, N.; Yao, G. A New Method of Conductor Galloping Monitoring Using the Target Detection of Infrared Source. *Electronics* **2022**, *11*, 1207. [CrossRef]
29. Yan, L.; Zai, W.; Wang, J.; Yang, D. Image Defogging Method for Transmission Channel Inspection by UAV Based on Deep Multi-patch Layered Network. In Proceedings of the Panda Forum on Power and Energy (PandaFPE), Chengdu, China, 27–30 April 2023; pp. 855–860.
30. Wang, K.; Yang, Y.; Fei, S. Review of hazy image sharpening methods. *CAAI Trans. Telligent Syst.* **2023**, *18*, 217–230.
31. Ancuti, C.-O.; Ancuti, C.; Sbert, D.; Timofte, R. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1014–1018.
32. Ancuti, C.-O.; Ancuti, C.; Timofte, R.; De, C. O-haze: A dehazing benchmark with real hazy and haze-free outdoor images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 754–762.
33. Ancuti, C.-O.; Ancuti, C.; Timofte, R. NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 444–485.
34. Zhang, N.; Zhang, L.; Cheng, Z. Towards simulating foggy and hazy images and evaluating their authenticity. In Proceedings of the Neural Information Processing: 24th International Conference, Guangzhou, China, 14–18 November 2017; pp. 405–415.
35. Harsányi, K.; Kiss, K.; Majdik, A.; Sziranyi, T. A hybrid CNN approach for single image depth estimation: A case study. In Proceedings of the International Conference on Multimedia and Network Information System, Hong Kong, China, 1 June 2019; pp. 372–381.
36. Wang, Z.; Bovik, A.-C. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.* **2009**, *26*, 98–117. [CrossRef]
37. Wang, Z.; Bovik, A.-C.; Sheikh, H.-R.; Simoncelli, E.-P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612.
38. Venkatanath, N.; Praneeth, D.; Bh, M.-C.; Channappayya, S.; Medasani, S. Blind image quality evaluation using perception based features. In Proceedings of the 2015 Twenty First National Conference on Communications (NCC), Munbai, India, 27 February–1 March 2015; pp. 1–6.
39. Amyar, A.; Modzelewski, R.; Vera, P.; Morard, V.; Ruan, S. Multi-task multi-scale learning for outcome prediction in 3D PET images. *Comput. Biol. Med.* **2022**, *151*, 106208. [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.

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
www.mdpi.com

Sensors Editorial Office
E-mail: sensors@mdpi.com
www.mdpi.com/journal/sensors

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.

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-0015-5