A Feature-Map-Based Method for Explaining the Performance Degradation of Ship Detection Networks

Peng Jia, Xiaowei He, Bo Wang, Jun Li, Qinghong Sheng, and Guo Zhang

***Abstract*—The unknowability of the inner workings limits the magnitude of performance improvement of ship target detection networks in synthetic aperture radar (SAR) images under Gaussian noise. However, none of the existing interpretation methods explain the phenomenon of network changes under noise. The feature map can visually reflect the changes in image delivery in the network, and some metrics can quantitatively characterize the degree of network performance degradation in a noise environment. So, in this article, we propose a comprehensive analysis method that integrates texture and brightness features of the internal feature map of the network to clarify the change process of target features under Gaussian noise. First, we analyzed the degradation of three target detection networks under different levels of Gaussian noise; then, the feature maps of four convolution layers were sampled and visualized for qualitative analysis; finally, the texture and brightness features were extracted for quantitative characterization of the feature amount changes. We experimentally validated the method on publicly available SSDD radar datasets. The networks were extremely sensitive to Gaussian noise, and the mean Average Precision decreased by up to 96.3%. The angular second moment and entropy texture feature values of the feature map could drop and rise 59.10% and 97.81%, respectively, while the brightness value could increase up to 100.92%. This indicates that noise changes the structure of feature maps and reduces the amount of effective information.**

***Index Terms*—Interpretability, ship detection, synthetic aperture radar, visualization.**

I. Introduction

THE all-day, all-weather, and high-resolution characteristics of synthetic aperture radar (SAR) imagery have led to its application in the field of sea surface situational awareness [1], [2], [3], [4]. Therefore, ship target detection on SAR images has become more important in military and civilian fields [5], [6], [7]. However, as SAR technology is increasingly used in military environments, noise techniques for SAR imaging are rapidly evolving. When the test data and the training data meet the condition of independently identically distribution, the convolutional neural networks do show great advantages in feature extraction of complex images [8]. Gaussian noise is simple in principle and can significantly degrade the detection performance of target detection networks, but how it leads to degradation of network performance due to the poor interpretation of the deep learning algorithms themselves remains to be investigated. At present, for the problem of ship target detection on SAR images, many scholars have proposed many innovative algorithms from various perspectives, among which deep learning algorithms are an important breakthrough [9], [10], [11], [12], [13]. Most of these methods are currently dedicated to solving scenario complexity and multiscale problems of ship targets or improving the detection accuracy of nearshore ships [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. If we want to optimize the network to an optimal state in a targeted manner, we must explain the network principle and clarify its internal working state.

Current deep learning explanation methods are divided into two main types: exante design and expost explanation. Exante design methods improve network interpretation by introducing mathematical physical models of known principles or designing a network structure that can be interpreted by itself. Zhang et al. [24] modified the traditional convolutional neural network (CNN) into an interpretable CNN, which allows training without additional annotation of objects for supervision purposes and can explicitly represent the knowledge in the higher hidden layers of the CNN. They also introduced a decision tree structure to clarify, which parts of the training object activate which neurons in the CNN, explaining how the CNN works at the semantic level [25]. Wan et al. proposed a similar idea. They designed a neural-backed decision tree to improve the interpretability of the neural network while ensuring the accuracy of the network as much as possible [26]. In addition, Wang and Yeung [27] combined Bayesian models with deep learning models to improve the interpretability of deep learning. However, the limitations of these approaches are that they somewhat weaken the powerful computational power of neural networks, and it is difficult and tedious to design interpretable networks with satisfactory results by choosing the appropriate mathematical physical model for each task. The existing ship target detection networks have

Manuscript received 5 November 2022; revised 19 January 2023; accepted 22 January 2023. Date of publication 1 February 2023; date of current version 16 February 2023. This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20220888, in part by the National Natural Science Foundation of China under Grant 42271448, and in part by the National Key Laboratory Foundation under Grant 2021-JCJQ-LB-006 and Grant 8676142411442120. *(Corresponding author: Bo Wang.)*

The authors are with the College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: iambeck@163.com; mailhexw@nuaa.edu.cn; wangbo_nuaa@nuaa.edu.cn; jun.li@nuaa.edu.cn; qhsheng@nuaa.edu.cn; guozhang@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3241395

deep hidden layers and more convolution kernels. Thus, it is challenging to tessellate or change a structure into an interpretable model, without any guarantee of accuracy degradation.

The expost interpretation method is based on the trained model and data for analysis. Representation visualization methods use a saliency map as a presentation to clearly show the output results of the layers within the neural network, visualize the attention of the network during sample training, and provide a reference for locating image features [28], [29], [30], [31]. Simonyan et al. [32] proposed two methods for visualizing image classification models and established a link between gradient-based visualization methods for convolutional networks and deconvolution methods. Springberg et al. [33] proposed a "deconvolution approach" to visualize the features learned by the network for the purpose of interpretation. Zhou et al. [34] explained how CNNs work when performing scene classification tasks by showing the object detectors inside the network during training. Furthermore, Bach et al. [35] proposed a method to visualize the contribution of individual pixels to the prediction of a kernel-based classifier in the form of a heat map. Based on the previous attribution methods, Sundararajan et al. [36] proposed the integrated gradient attribution method, which extracts rules from the network by invoking the standard gradient operators of the network several times to facilitate the user to understand the model. In addition, the class activation mapping method is also a common representation visualization method. Zhou et al. [37] explained the important role of the global average pooling layer using the class activation mapping technique (CAM). Selvaraju et al. [38] proposed a gradient-weighted class-activation mapping (Grad-CAM) method to highlight important areas in images, which can be used to explain the principle of image classification. Based on the Grad-CAM method, Chattopadhay et al. [39] proposed the Grad-CAM++ model, which has a better effect in explaining the model prediction process. Moreover, Wang et al. [40] designed Score-CAM, which achieved better visual performance and fairness for interpreting the decision-making process than Grad-CAM and Grad-CAM ++. Fong and Vedaldi [41] attributed network decisions to a feature of the input through meaningful perturbations. Zhang et al. [42] developed a simple and effective method to learn feature maps that reveal the component hierarchy of object components encoded in the convolution layer of a pretrained CNN. However, the abovementioned visualization methods for characterization based on factors such as gradients or perturbations were mainly designed to analyze the regions in the input image that have an impact on the decision and the magnitude of their impact. The process of change of features in feature maps under noise is not addressed, but the methods shed some light on explaining the degradation of target detection network performance under Gaussian noise. We conjectured that we could make a summary of the causes of network degradation under the influence of Gaussian noise based on successive feature maps of the same change process since the visualization method can visualize the phenomenon inside the network.

In addition to representational visualization methods, expost interpretations include sample-based interpretations and natural language interpretations. Sample-based interpretation methods are used to mathematically associate variables with prediction results, such as substitution variables, to explain the effect of information features on the prediction performance of the network. Li et al. [43] constructed an interpretable deep neural network by combining deep learning and sample-based inference with interpretability. Arik and Pfister [44] proposed an approach that combines a coded representation with a small number of samples to achieve high-quality interpretability. Hendricks et al. [45] proposed a model that focuses on visible object recognition properties to explain the fundamentals of network classification decisions from a natural language perspective. These two methods have more demanding conditions and higher complexity than the methods for characterization visualization. They are not applicable to the SAR image ship detection task covered in this article.

Inspired by the representation visualization interpretation method, we designed an interpretation analysis method that integrates the texture and brightness features of feature maps; conducted an experimental validation on the publicly available SSDD radar dataset; conducted a comparative analysis of the feature changes of the target and background in the feature map under different intensity Gaussian noise; corresponded the results to the changes of the network detection performance, and summarized the relevant conclusions.

II. Methodology

The interpretation of the degradation of the target detection networks is important for their performance optimization in the environment with Gaussian noise. Current network interpretation methods, both exante design and expost interpretation, have unavoidable problems when migrating to the interpretation of target detection networks in SAR images under Gaussian noise. Most of the exante design methods include inherently interpretable network structures by introducing knowable mathematical physical models. The most comprehensively developed of the expost interpretation methods are the representation visualization methods, but most of them are devoted to analyzing the degree of influence of each region in the image on the decision result while ignoring the overall change pattern. Feature maps, especially those of convolution layers, as the results of feature extraction from convolution kernels, contain low-dimensional features such as brightness and texture, as well as high-dimensional features that are difficult to be distinguished by the human eye. Starting from the feature maps, we analyzed the changes in the low-dimensional features of the same hidden layer feature maps and the changes of the low-dimensional features of different hidden layer feature maps under different intensities of Gaussian noise in target detection networks, first. Then, we corresponded these changes to the performance of target detection networks under different intensities of Gaussian noise. Finally, we summarized the relevant conclusions and partially explained the phenomenon of network performance degradation. To achieve this purpose, this article takes a fourfold approach, with the following aims:

1) characterize the network performance under different intensity Gaussian noise with four accuracy metrics, namely,

precision, recall, F1, and mean Average Precision (mAP), and quantitatively analyze the performance degradation of three networks, namely, Yolov5, Faster R-CNN, and SSD;
2) output the feature maps of four convolution layers of the three networks in four sheets with a layered sampling method;
3) extract the texture features of the images with gray-level co-occurrence matrix (GLCM) to generate two feature volumes, which are combined with the brightness features for quantitative analysis;
4) assess the correspondence between the results of 1), 2) and 3), and summarize the analysis to draw conclusions related to the degradation of network performance under Gaussian noise.

A. Evaluation Indicator

Currently, there are four main metrics to evaluate the accuracy of target detection networks: precision, recall, F1 score, and mAP. In the PASCAL visual object classes (VOC) Challenge [46] and ImageNet Large Scale Visual Recognition Challenge [47], the organizers used the precision/recall curve and mAP as evaluation metrics to measure the effectiveness of the participating teams' target detection. Precision denotes the percentage of all detected objects above the threshold that is correct. Recall denotes the percentage of all positive examples ranked above a given rank. mAP is obtained by combining the average correct rates of all categories in a combined weighted average when the Intersection over Union (IoU) is 0.5. Gidaris and Komodakis [48] proposed the concept of IoU, a metric defined as the overlap rate between the target window generated by the model and the labeled window. The F1 score is a statistical indicator defined as the summed average of precision and recall.

We used precision, recall, F1, and mAP to evaluate neural networks' performance. These four metrics can be defined as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

$$\text{mAP} = \frac{\sum_j^M P_j}{M} \tag{4}$$

where TP, TN, FP, FN, and M represent true positives, true negatives, false positives, false negatives, and type of targets, respectively. The meaning of P_j is determined by

$$P_j^i = \frac{A_j^i}{B_j^i} \tag{5}$$

$$P_j = \frac{\sum_i^N P_j^i}{N} \tag{6}$$

where i, j, A_j^i, B_j^i, and N represent, respectively serial number of the image, serial number of the target category, number of detected targets, and real number of targets and number of samples.

$$\text{IoU} = \frac{\text{area}\,(B_t \cap B_{tt})}{\text{area}\,(B_t \cup B_{tt})}. \tag{7}$$

The definition of IoU can be seen in (7), where B_t and B_{tt} represent predicted and real target boundaries, respectively. And, in this article, IoU was set at 0.5.

B. Feature Map Visualization

In this article, four convolution kernels were selected in four convolution layers as the visualization objects of the feature maps. Since the method in this article started from the variation of the feature maps to explain the network-related principles, we chose four feature layers in three networks with different stages of the feature extraction module. These feature maps are effective for detection and classification. Because the number of convolution kernels in each convolution layer is different, and the maximum can be 1024, stratified sampling was used to extract four convolution kernels in each convolution layer with the total pixel value as the index and visualize the feature map. We first calculated the total pixel values of each convolution kernel feature map according to (8) and arranged them in descending order. Then, we divided the number of convolution kernels in this layer into four equal parts. Finally, the first image of each part was selected and visualized. As images are transmitted in the form of tensors in the network, the process of feature map visualization was to extract the tensor of the target convolution kernel and output it in the form of two-dimensional images

$$\text{Pixel value} = \sum_{i=1}^{N} \sum_{j=1}^{M} \text{value}_i^j \tag{8}$$

here N, M, $(i,\ j)$, and value represent width, length, position of the pixel, and value of the pixel, respectively.

C. Texture and Brightness of Feature Maps

After qualitative analysis, two feature quantities generated by GLCM and brightness value were used for quantitative analysis. Texture features can express information about the spatial arrangement of colors or intensities in selected areas of an image and are important references when analyzing image properties. Texture features consist of spatial relationships between pixels as well as neighboring pixels and are local in nature. Local texture information presents different degrees of repetitiveness. The GLCM is a common method for extracting image texture features, reflecting the pixel correlation by the distance between pixel values as well as the angle. It integrates the information of the image in terms of direction, interval, and magnitude of change and speed, and expresses it through the matrix. Based on GLCM, 14 statistics could be calculated: energy, entropy, contrast, uniformity, correlation, variance, sum average, sum variance, sum entropy, difference variance, difference average, difference entropy correlation information measure, and maximum correlation coefficient. The feature quantities selected in this article were angular second moment (ASM) and entropy

TABLE I
MEAN AND VARIANCE OF LEVEL 0–LEVEL 5 DATASETS WHEN ADDED GAUSSIAN NOISE

level	mean	variance
Level 0	0	0
Level 1	0.1	0.01
Level 2	0.1	0.02
Level 3	0.1	0.03
Level 4	0.2	0.04
Level 5	0.2	0.05

(ENT). ASM can measure the total value of feature map information and ENT can describe the degree of uniformity of the grayscale distribution. They give us some insights in terms of the amount of information. Their equations are shown in (9) and (10) as follows:

$$\mathrm{ASM} = \sum_i \sum_j p(i,j)^2 \tag{9}$$

$$\mathrm{ENT} = -\sum_i \sum_j p(i,j) \log p(i,j) \tag{10}$$

where $p(i,j)$ refers to the normalized GLCM.

When brightness values were calculated, we first counted the pixel values of R, G, and B, and then followed

$$\mathrm{brightness} = \left(0.241 \times r^2 + 0.691 \times g^2 + 0.068 \times b^2\right)^{\frac{1}{2}} \tag{11}$$

where r, g, b represent pixel values for the three channels of the image.

After analyzing the performance degradation of the three networks and the changes in the feature maps under the Gaussian noise by the abovementioned three methods, we synthesized and summarized the results to obtain relevant conclusions.

III. EXPERIMENTAL AND DISCUSSION

A. Datasets

In this article, we used the SSDD dataset of ships in SAR images, constructed in 2017 [44]. It has 1160 images and 2456 ships, with between one and 13 ships per image. The resolution of this dataset is below 3 m. Because the target detection network depends on the dataset, the more images in the dataset, the more accurate the target detection result. Therefore, we expanded the number of noise-free images to 2000, in which the ratio of the training set, validation set, and test set was 81:9:10. Then, based on the noise-free image, we added five different levels of noise using the variance of Gaussian noise as the level classification criterion, which was shown in Table I. In the end, we acquired a total of 12 000 SAR ship images.

B. Network Degradation Characterization

Table II through Table IV represents the results of the Yolov5, Faster R-CNN, and SSD networks trained with a noise-free ship dataset, which we tested with noisy datasets with different levels of Gaussian noise effects after obtaining the three models.

As can be seen from Table II to Table IV, the detection capability of all three target detection networks decreases rapidly as the level of Gaussian noise increases. Among the four evaluation metrics, the recall, F1, and mAP consistently decrease. The precision metrics of Faster R-CNN and SSD fluctuate, differing from the consistent decrease of Yolov5. With noise enhancement, the precision of SSD and Faster R-CNN even increases and reaches 1.0, while the precision of Yolov5 always decreases. The definition of precision [Function (1)], indicates the probability of the actual true target in the sample of detected true targets. As the metric recall, which describes the true target detection rate, always decreases, it is known that this particular phenomenon is caused by a decrease in the number of samples detected as true targets, and it so happens that most of them are true targets. When there is no Gaussian noise, Yolov5 has the best detection, with 95.2% for precision, 91.1% for recall,

TABLE II
CHARACTERIZATION OF YOLOV5 NETWORK DEGRADATION UNDER DIFFERENT LEVELS OF GAUSSIAN NOISE

Noise level	Precision	Recall	F1	mAP
level 0	0.952	0.911	0.93	0.966
level 1	0.114	0.151	0.13	0.068
level 2	0.048	0.219	0.08	0.040
level 3	0.040	0.062	0.05	0.017
level 4	0.012	0.049	0.02	0.006
level 5	0.010	0.016	0.01	0.003

TABLE III
CHARACTERIZATION OF FASTER R-CNN NETWORK DEGRADATION UNDER DIFFERENT LEVELS OF GAUSSIAN NOISE

Noise level	Precision	Recall	F1	mAP
level 0	0.391	0.619	0.48	0.476
level 1	0.760	0.051	0.10	0.087
level 2	0.333	0.003	0.01	0.025
level 3	1.000	0.003	0.01	0.004
level 4	0.000	0.000	0.00	0.000
level 5	0.000	0.000	0.00	0.000

TABLE IV
CHARACTERIZATION OF SSD NETWORK DEGRADATION UNDER DIFFERENT LEVELS OF GAUSSIAN NOISE

Noise level	precision	recall	F1	mAP
level 0	0.906	0.495	0.64	0.612
level 1	0.933	0.038	0.07	0.098
level 2	0.600	0.008	0.02	0.035
level 3	1.000	0.003	0.01	0.003
level 4	0.000	0.000	0.00	0.000
level 5	0.000	0.000	0.00	0.000

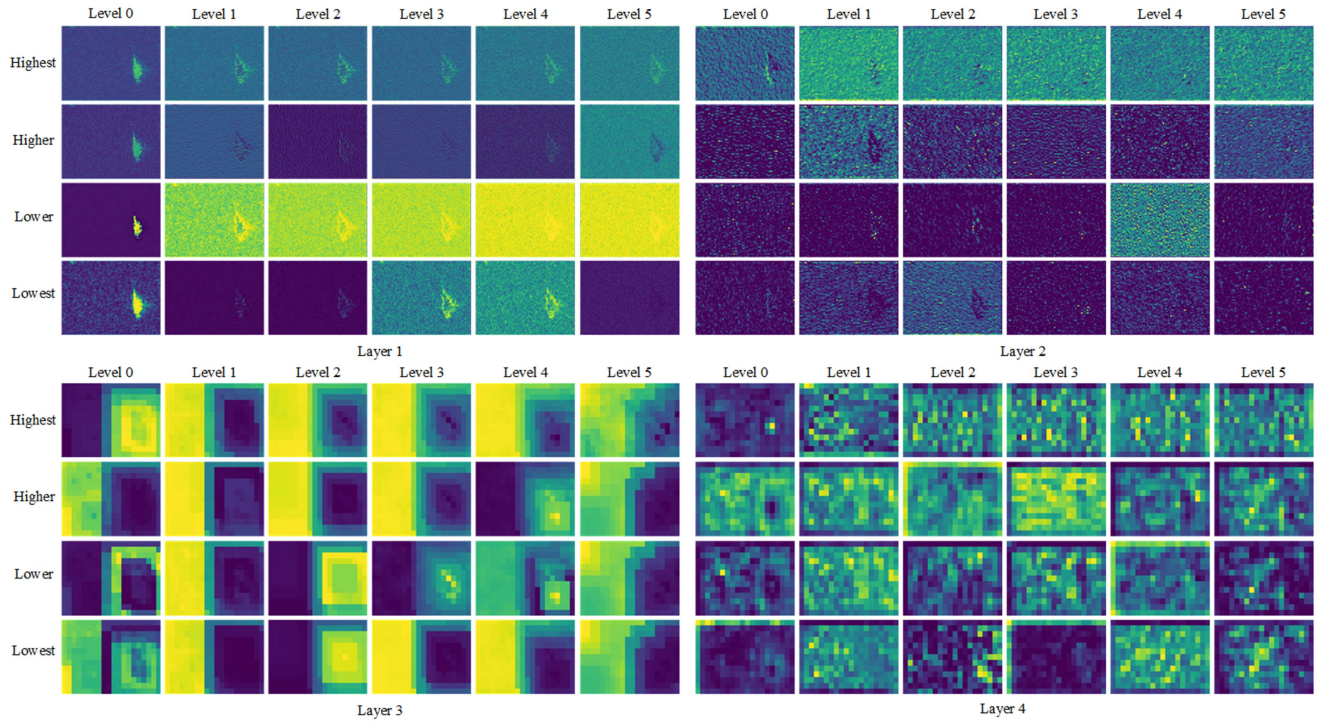

Fig. 1. Feature maps of four convolution layers in Yolov5 for image 1 with different levels of Gaussian noise.

0.93 for F1, and 96.6% for mAP. The precision, recall, F1, and mAP of SSD at this time are 90.6%, 49.5%, 0.64, and 61.2%, respectively, while for Faster R-CNN, the precision, recall, F1, and mAP are 39.1%, 61.9%, 0.48, and 47.6%, respectively. In terms of mAP values, Yolov5 is at least 35% higher. With the enhancement of Gaussian noise, although Yolov5's detection performance decreases the most, it still detects a small number of targets with a value of at least 0.3% for mAP when level 4 and level 5 noise affects it, which is significantly better than the other two networks. This indicates that the impact of Gaussian noise on the target detection network is huge.

C. Feature Map Visualization

Figs. 1–6 represent feature maps of four convolution layers of three target detection networks, named Yolov5, Faster R-CNN, and SSD, respectively.

In layer 1 of Yolov5, we made horizontal and vertical comparisons. The vertical comparison shows that the total pixel value of the image is taken as the standard for stratified sampling, which reflects the brightness features of the whole image rather than the brightness features of the target on the image. Therefore, the edge, brightness, and texture of the target do not decrease as the total pixel value decreases. So, we can find that the lowest and lower phases are visible to the naked eye. The object is brighter than the other two phases, and the edges are clearer, indicating a higher contrast with the background. The effect of feature maps in four different stages is inconsistent, which also indicates that the types and quantities of features extracted by each convolution kernel are not the same, which is also the significance of the weighted results of each convolution kernel. The horizontal comparison shows that with the enhancement of the noise, the target is gradually covered by the noise, especially in the lower stage. Moreover, the brightness of the background increases with the enhancement of the fifth level of noise. In the other stages, the contrast between the background and the target does not decrease linearly with the noise level, but the target is covered, and the background brightness increases. This nonuniform change also reflects the inhomogeneity of the influence of Gaussian noise on each convolution kernel. Therefore, this phenomenon can be understood as follows: The Gaussian noise leads to the reduction of the target features extracted by the neural network, which leads to the rapid decline of the detection effect of the Yolov5 network. It is clear that feature maps gradually become combinations of color blocks as data travels through Yolov5. In layer 4, there is no specific shape of the target in feature maps, because the transfer of the image in the network is a downsampling process, and the further it goes, the bigger the region of the feature map mapped on the original image, and the more high dimensional and abstract the extracted features.

In layer 1 of Faster R-CNN, the vertical comparison shows that with the decrease of the overall pixel value, the target features in the feature maps gradually decrease, and the contrast with the background decreases until the lowest stage when the target is almost invisible. The horizontal comparison shows that similar to Yolov5, with the enhancement of noise, the contrast between the target and background gradually decreases, and the target features disappear obviously, especially in the highest and higher layers. Similarly, as Faster R-CNN deepens,

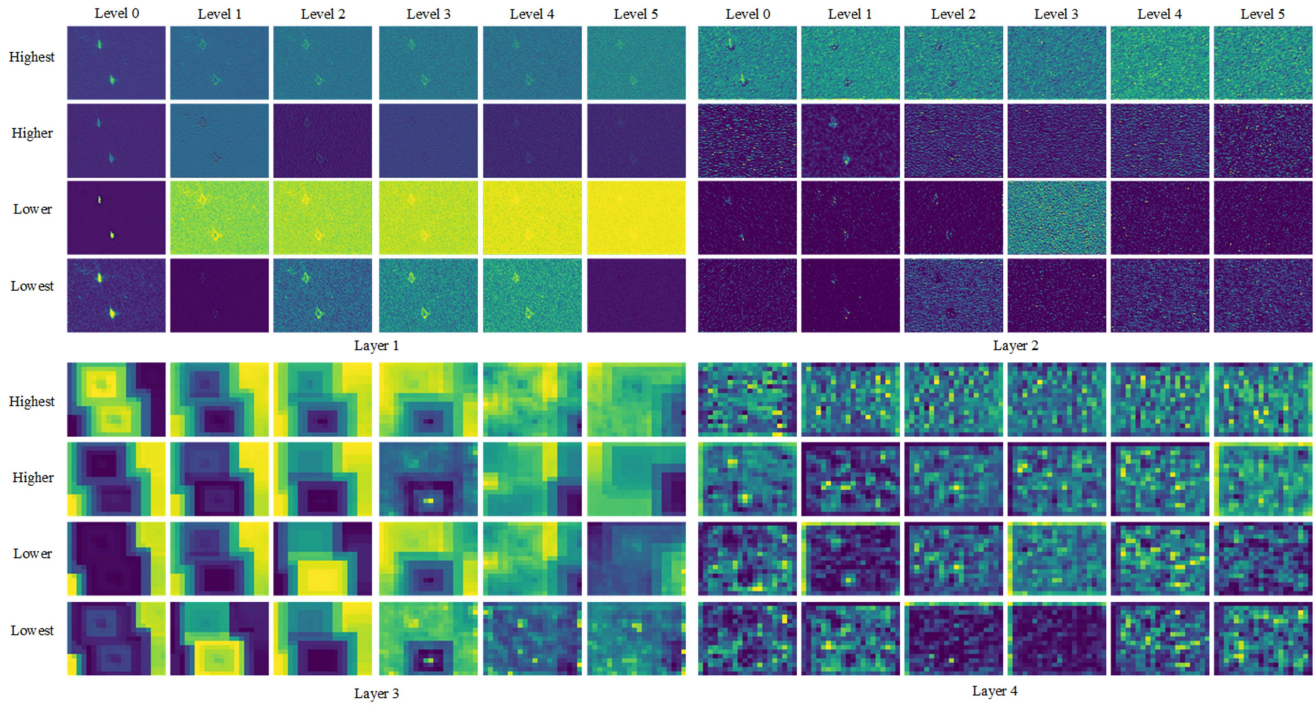

Fig. 2. Feature maps of four convolution layers in Yolov5 for image 2 with different levels of Gaussian noise.

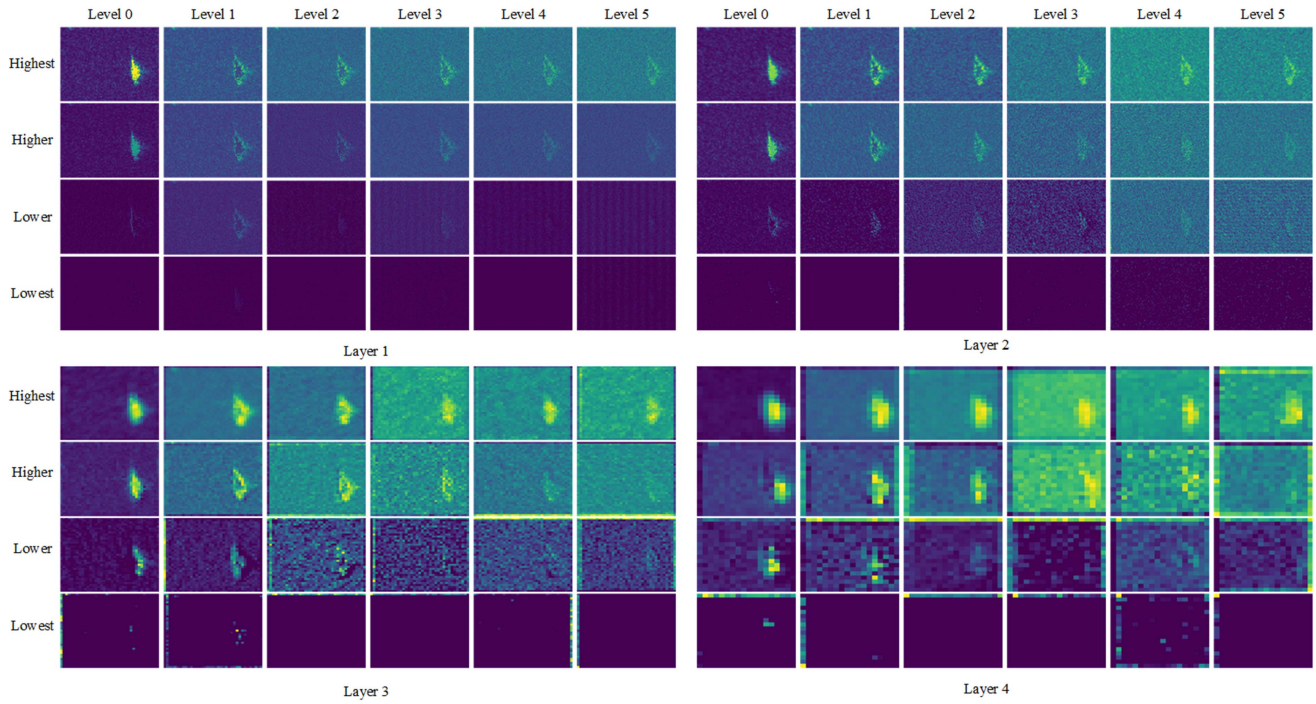

Fig. 3. Feature maps of four convolution layers in Faster R-CNN for image 1 with different levels of Gaussian noise.

feature maps gradually become abstract and targets pattern fades away. However, in layer 4 of Faster R-CNN, different from the abstraction of the last convolution layer of Yolov5, the target and background can still be clearly distinguished in the feature map. The vertical and horizontal changes are consistent with layer 1. With the enhancement of noise, the contrast between the target and background decreases, and the background brightness increases significantly, resulting in the disappearance of target features. Therefore, it can be concluded that because the Gaussian noise covers the important features of the target, the features

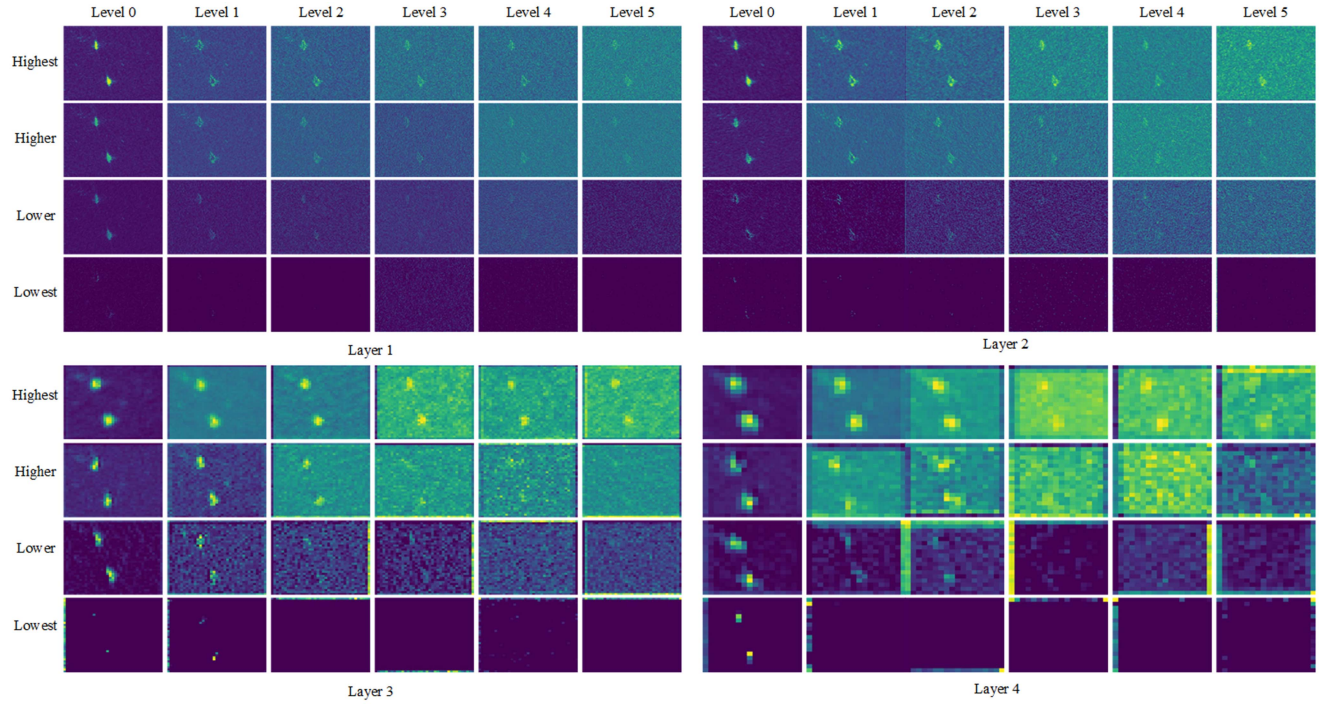

Fig. 4. Feature maps of four convolution layers in Faster R-CNN for image 2 with different levels of Gaussian noise.

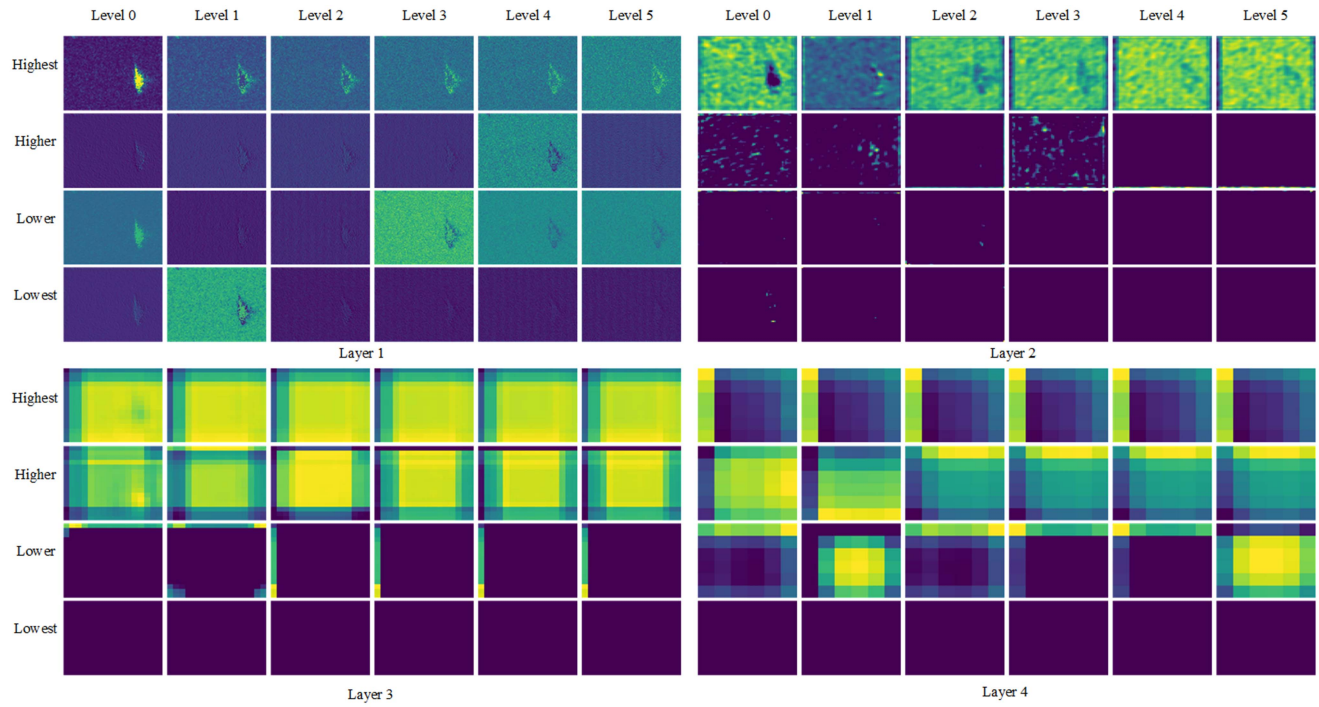

Fig. 5. Feature maps of four convolution layers in SSD for image 1 with different levels of Gaussian noise.

extracted by the network are reduced, which reduces the target detection accuracy of Faster R-CNN.

In layer 1 of SSD, the vertical comparison shows that similar to Yolov5, the brightness of the target does not decrease with the decrease in pixel value, because the total pixel value of the image reflects the global feature of the image. The horizontal comparison shows that similar to the other two networks, with the enhancement of noise, the contrast between the target and the background gradually decreases, and the target is gradually covered by the background. Therefore, it can be concluded

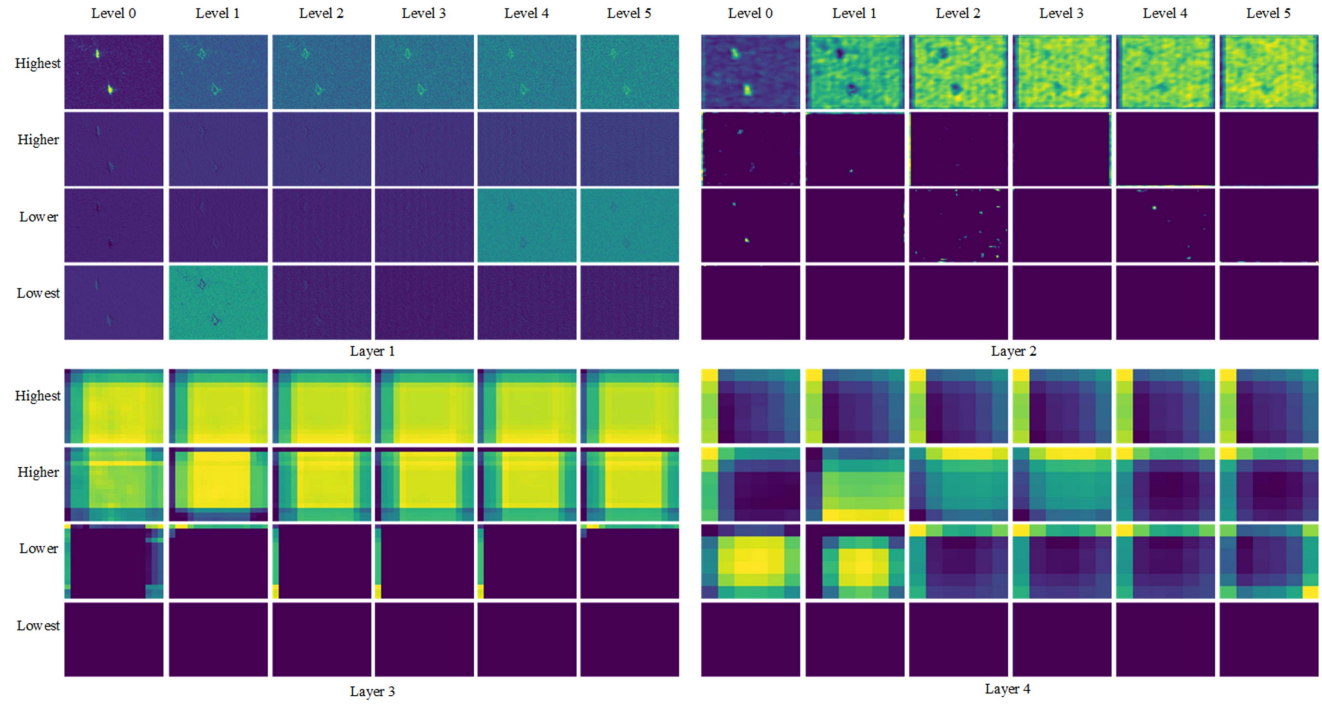

Fig. 6. Feature maps of four convolution layers in SSD for image 2 with different levels of Gaussian noise.

TABLE V
GRAY-LEVEL COINCIDENCE MATRIX OF IMAGE 1 IN YOLOV5

	Layer1		Layer2		Layer3		Layer4	
	ASM	ENT	ASM	ENT	ASM	ENT	ASM	ENT
Level 0	2.3380	3.5518	1.5439	5.1608	0.5677	8.8670	0.7687	7.4158
Level 1	1.9768	3.9818	1.2565	6.0356	0.7557	7.4113	0.5834	8.6902
Level 2	1.7457	4.5131	1.3053	5.8337	0.7218	8.1438	0.5930	8.4653
Level 3	1.2370	5.6721	1.6759	4.9411	0.8521	7.5503	0.6170	8.5489
Level 4	1.2000	5.7709	1.0319	6.5022	0.7197	8.3633	0.6010	8.5356
Level 5	1.4593	4.7608	1.6059	5.0640	0.4388	9.4381	0.6010	8.3428

that with the enhancement of Gaussian noise, the target features extracted by the network are reduced, which degrades the detection performance of the SSD network. In layer 4, the features extracted from the network are abstract, and it can be seen that with the enhancement of the noise, the feature map hardly changes, indicating that the compression noise affects the shallow convolution layer more.

By comparing layer 1 of the three networks under the condition of no noise, it can be found that the convolution kernels of Yolov5 can extract more target information, and the edge, brightness, and texture features are more obvious, which can partly explain why the detection effect of Yolov5 network is significantly better than that of the other two networks under the condition of no noise. With the enhancement of Gaussian noise, the target features that can be extracted by the three networks decreases, so the detection effect decreases rapidly.

D. Feature Map Texture and Brightness

Tables V–X show the changes in the two feature quantities, ASM and ENT, of the feature maps of the four convolution layers of Yolov5, Faster R-CNN, and SSD target detection networks calculated from GLCM under the influence of Gaussian noise with different intensities. Here, the values of ASM and ENT were the sum of the values of four stages, named highest, higher, lower, and lowest.

Table V shows that with the increase of the noise intensity, the value of ASM decreases in layer 1 of Yolov5, and the value of the entropy increases in this layer. Values of ASM and ENT change in opposite directions. The former drops up to 48.67%, while the latter rises to 62.48%. This shows that with the enhancement of Gaussian noise, the values of each pixel of feature maps in layer 1 of Yolov5 become gradually close, and the randomness enhances. It means the feature structure of feature maps is changed. Since Gaussian noise affects the whole image, convolution kernels learned by the image with noise have a lot more Gaussian noise features than noise-free images. ASM and ENT change differently at layer 2 and layer 3 than they do at layer 1. The values of ASM even increase. We speculate that this is a method for Yolov5 to resist Gaussian noise, and also proof that different convolution kernels extract different features.

TABLE VI
GRAY-LEVEL COINCIDENCE MATRIX OF IMAGE 2 IN YOLOV5

	Layer1		Layer2		Layer3		Layer4	
	ASM	ENT	ASM	ENT	ASM	ENT	ASM	ENT
Level 0	2.3082	3.0166	1.9281	4.3663	0.7036	8.3421	0.6119	8.3710
Level 1	2.2069	3.5015	2.3252	3.4237	0.4540	9.4432	0.5978	8.3083
Level 2	1.3259	5.2496	1.4289	5.4615	0.4509	9.2779	0.6874	7.8703
Level 3	1.2506	5.5702	1.3179	5.8363	0.7143	8.1350	0.7667	7.6382
Level 4	1.2085	5.6790	1.1914	6.0186	0.7082	8.1038	0.5780	8.5743
Level 5	1.7214	4.5254	1.3522	5.6514	0.7917	7.8178	0.6062	8.3685

TABLE VII
GRAY-LEVEL COINCIDENCE MATRIX OF IMAGE 1 IN FASTER R-CNN

	Layer1		Layer2		Layer3		Layer4	
	ASM	ENT	ASM	ENT	ASM	ENT	ASM	ENT
Level 0	2.2157	3.4594	2.1214	3.6231	2.3298	3.5239	1.8243	4.7733
Level 1	1.6448	4.6558	2.4285	3.0722	2.1852	4.0031	1.9518	5.0344
Level 2	1.7335	4.8264	2.0673	4.0068	1.9203	4.6101	1.9513	4.7499
Level 3	1.3935	5.4350	1.8261	4.5648	1.8330	4.8310	1.7315	5.5550
Level 4	1.7242	4.8701	1.4000	5.2069	2.0068	4.2422	1.6333	5.5661
Level 5	1.0033	6.4665	1.7566	4.6433	2.0168	4.4846	1.7100	5.4843

TABLE VIII
GRAY-LEVEL COINCIDENCE MATRIX OF IMAGE 2 IN FASTER R-CNN

	Layer1		Layer2		Layer3		Layer4	
	ASM	ENT	ASM	ENT	ASM	ENT	ASM	ENT
Level 0	2.1365	3.3927	2.2539	3.3878	2.4684	3.2313	2.1851	4.1358
Level 1	1.8830	4.2196	2.4287	2.9749	1.9744	4.2826	1.9490	4.7172
Level 2	1.8172	4.5430	1.9818	4.0457	2.1342	4.2060	1.7879	5.3606
Level 3	0.8801	6.8669	1.7963	4.4847	1.7061	5.0718	1.9321	5.0288
Level 4	1.4181	5.4560	1.5325	4.9105	1.4566	3.4390	1.5648	5.7375
Level 5	1.6239	5.1210	1.8858	4.3520	1.9595	4.4310	1.5260	5.5941

TABLE IX
GRAY-LEVEL COINCIDENCE MATRIX OF IMAGE 1 IN SSD

	Layer1		Layer2		Layer3		Layer4	
	ASM	ENT	ASM	ENT	ASM	ENT	ASM	ENT
Level 0	2.1504	3.5980	2.7961	2.9978	2.2358	4.3916	1.4252	6.3689
Level 1	1.1298	6.0098	3.2765	1.6746	2.2148	4.4794	1.4439	6.2851
Level 2	0.8792	7.1173	3.0964	2.1761	2.3203	4.2761	1.5088	6.0537
Level 3	0.8963	7.0290	2.7988	2.9328	2.3005	4.1697	1.8418	5.1911
Level 4	1.1628	5.9499	3.1324	2.1446	2.2948	4.1747	1.8383	5.2024
Level 5	1.0513	6.5746	3.1112	2.2258	2.2434	4.2599	1.4479	6.4986

TABLE X
GRAY-LEVEL COINCIDENCE MATRIX OF IMAGE 2 IN SSD

	Layer1		Layer2		Layer3		Layer4	
	ASM	ENT	ASM	ENT	ASM	ENT	ASM	ENT
Level 0	1.5425	4.7223	3.2990	1.4670	1.9590	5.0526	1.4895	6.2204
Level 1	1.2506	5.5021	3.1055	2.1942	2.2700	4.3978	1.4581	6.1640
Level 2	0.9128	6.9687	3.0469	2.4296	2.2838	4.2243	1.5027	6.0258
Level 3	0.8890	7.1618	3.1105	2.2154	2.2792	4.2296	1.4999	6.0354
Level 4	1.1298	6.2392	3.1222	2.1918	2.2359	4.2987	1.4610	6.0779
Level 5	1.0919	6.3415	3.1043	2.2274	2.2122	4.3664	1.4289	6.3686

However, in layer 4 of Yolov5, the changes of ASM and ENT are similar to those in layer 1, but the magnitude is significantly reduced. Compared with the case without noise, the former drops up to 24.10%, while the latter rises to 17.18%. This indicates that the similarity between every pixel is gradually increasing. The descending amplitude and direction of each convolution kernel are different, which indicates that the influence degree of noise on each convolution kernel is different. Table VI shows changes in ASM and ENT of another image with different levels of Gaussian noise in Yolov5. On the whole, feature map changes of the four convolution layers are similar, but the change results in layer 2 are different from those in the first image. This indicates that the performance of different images in the same network is not the same, although the noise with the same parameters is added, on the whole, it conforms to the same change rule.

Compared with Yolov5, in Table VII, layer 1 of Faster R-CNN has enhanced disorder of ASM and ENT change. If we add the values for the highest, higher, lower, and lowest phases as a whole, the value of ASM tends to decrease with increasing noise intensity, especially for the fifth noise level, up to 54.72%. On the contrary, the value of ENT has an upward trend, up to 86.93%. This indicates that Gaussian noise destroys the texture structure of the whole image and makes every pixel homogeneous. The nonlinear change relationship between noise level and the values of ASM and ENT reflects the nonlinear operation mechanism of Faster R-CNN. ASM and ENT change at Layer 2 and Layer 3 similar to their change at Layer 1. The change of ASM in layer 4 of Faster R-CNN is different from that in other layers in that exhibits a rise followed by a fall, but the magnitude is small and can be approximated as constant. The value of ENT still shows a small rise and can also be approximated as constant. This approximate invariance indicates that the effect of noise on this convolution layer is not significant. Table VIII shows changes in ASM and ENT of another image with different levels of Gaussian noise in Faster R-CNN. On the whole, feature map changes of the four convolution layers are similar. However, the values of ASM of feature maps with noisy images are always lower than the value of feature maps with noise-free images, and the values of ENT change the opposite.

As Table IX shows, in layer 1 of SSD, the value of ASM decreases by a maximum of 59.11%, and the value of ENT increases by a maximum of 97.81%. So, a conclusion similar to the abovementioned two networks can be drawn: The enhancement of noise leads to a change in the feature structure of feature maps, and the pixel values on the feature maps are gradually close to each other, resulting in the masking of target features and the degradation of SSD network detection performance. In another three layers, there is a small change between feature maps of images with different levels of Gaussian noise. It can also be seen from Figs. 5 and 6 that the feature maps of layer 2, layer 3, and layer 4 do not change significantly with the increase in noise intensity.

Based on the analysis results of the three networks, it can be concluded that with the introduction of Gaussian noise, feature structure of feature maps extracted by the network convolution kernel changes, the similarity of each pixel in the feature maps increases, the texture structure decreases, and the high-dimensional feature structure also changes. These reasons make the decision basis of the network decrease, and it is difficult to make a correct judgment, thus causing a decrease in network detection accuracy. Among these convolution layers analyzed, reduction of ASM can reach up to 59.11%, and the minimum approximation is 0. Increase of ENT value can reach up to 97.81%, and the minimum approximation is also 0.

Tables XI–XIII show the brightness feature extracted from the feature maps of the four convolution layers of the Yolov5, Faster R-CNN, and SSD networks.

The brightness feature is a global feature that integrates the target and background. At layer 1 of Yolov5, we took the brightness values for the highest, higher, lower, and lowest phases as one. As we can see, the overall brightness value increases as the noise increases, and the range is up to 100.92%. In the image, the size of the target is limited, and under normal conditions, the target is the brightest part of the image. Therefore, the enhancement of the feature map brightness caused by noise is more likely to enhance the background brightness. With the enhancement of background brightness, the target is gradually covered, and the features that can be extracted by the network are reduced, resulting in the degradation of network detection performance. It is worth noting that the influence of Gaussian noise on the image is global. The noise feature extracted from the feature maps will also cover the features of the target, making the target difficult to be found. Using the same analysis method as layer 1, we can find that the growth of brightness value is not linearly related to the noise level, and the growth range is significantly reduced or even negative in layers 2, layer 3, and layer 4. So, what causes this nonlinear change? We speculate that the nonlinear mechanism of convolution calculation leads to the different features extracted from each convolution kernel. Since Gaussian noise affects the overall image structure, some regional features extracted from a single convolution kernel will not necessarily linearly weaken due to noise enhancement. Since the object only occupies a small area on the image, the brightness value depends on the number of background features and noise features extracted by the convolution kernel.

In layer 1 of Faster R-CNN, the overall brightness value increases with the enhancement of noise. The maximum increase in total brightness is 68.14%. Similarly, we can speculate that the increase in brightness values is mainly on the background, resulting in a reduction in the contrast between the target and the background. In layer 2, layer 3, and layer 4 of Faster R-CNN, the overall brightness value increases with the enhancement of noise. The largest increases in total brightness in these three layers are 93.92%, 90.28%, and 83.12%. Thus, the same conclusion can be reached for layer 1.

In layer 1 of SSD, similar to layer 1 of Faster R-CNN, the overall brightness value increases with the enhancement of noise, and the largest is 49.14%. Therefore, the two conclusions are the same. At layer 2, layer 3, and layer 4 of SSD, the value of brightness can go up, down, or stay the same, which indicates that the noise has little influence on the convolution layer. And this also indicates that the brightness features of the deeper convolution layer of SSD have little influence on the detection results.

TABLE XI
BRIGHTNESS FEATURES OF YOLOV5

	Image 1				Image 2			
	Layer1	Layer2	Layer3	Layer4	Layer1	Layer2	Layer3	Layer4
Level 0	256.3595	229.2766	429.4733	324.4119	234.5883	255.4965	415.2857	381.4352
Level 1	416.4599	330.0895	497.6813	457.8250	426.3282	256.7401	466.7088	338.6969
Level 2	386.3499	304.9133	497.9661	386.7297	449.8585	279.1277	536.1572	339.5994
Level 3	487.5063	270.5443	510.0862	424.7098	502.9846	319.0463	556.8190	371.1953
Level 4	515.0893	328.9031	511.7484	432.0801	511.2712	288.6722	544.2760	394.6568
Level 5	501.5886	287.2410	510.5833	347.1983	444.9502	274.0201	496.3829	417.1504

TABLE XII
BRIGHTNESS FEATURES OF FASTER R-CNN

	Image 1				Image 2			
	Layer1	Layer2	Layer3	Layer4	Layer1	Layer2	Layer3	Layer4
Level 0	192.2567	190.4695	196.8017	218.5835	190.8355	189.4837	195.3539	196.1460
Level 1	244.0413	255.6896	295.8874	290.9484	240.5667	262.5391	286.4642	322.7007
Level 2	270.3794	280.0724	346.0648	312.2919	275.2326	292.5413	345.3556	363.7831
Level 3	287.4899	297.4726	366.0805	400.2828	292.2068	316.3728	381.3717	409.5102
Level 4	320.5030	371.0490	353.3930	379.2174	321.0181	363.4356	380.1445	431.4154
Level 5	323.2630	360.5389	374.4770	363.0939	312.2860	377.9556	390.4987	341.8902

TABLE XIII
BRIGHTNESS FEATURES OF SSD

	Image 1				Image 2			
	Layer1	Layer2	Layer3	Layer4	Layer1	Layer2	Layer3	Layer4
Level 0	271.0358	278.5668	415.2008	374.5528	224.1914	189.2853	426.9612	364.47373
Level 1	345.9573	213.8095	416.3599	394.0424	339.2672	264.6036	425.7086	395.82023
Level 2	266.1413	272.3065	427.1205	332.9979	269.8917	289.2497	424.3757	337.29209
Level 3	370.6177	278.8051	422.6213	311.6213	272.8679	294.4803	423.3058	337.00792
Level 4	404.2354	291.3499	422.0786	311.4777	350.8423	290.4634	422.7371	286.74107
Level 5	362.3768	292.9403	422.4325	417.6848	286.7411	297.8722	421.9447	302.60068

According to the abovementioned analysis, in the convolution layer analyzed in this article, with the increase of the Gaussian noise intensity, the maximum increase of the brightness feature can reach 100.92%, and the lowest increase can be approximated to 0. Based on the brightness feature changes of the four convolution layer feature maps of the three target detection networks, it can be concluded that the Gaussian noise leads to the enhancement of the background brightness and the weakening of the contrast between the target and the background to cover the features of the target and lead to the degradation of the detection effect of the network. At the same time, the image brightness will be increased generally after the feature of Gaussian noise is extracted from the convolution kernels. As the depth of the convolution layer deepens, the influence of brightness on the decision results decreases, so the brightness decreases sometimes. All of these phenomena prove the nonlinearity of the convolution kernel and the fact that each convolution kernel has its preference.

Based on the qualitative and quantitative analysis at the feature map level above, it is clear that with the enhancement of Gaussian noise, the feature structure extracted from feature maps changes. Therefore, we can conclude that Gaussian noise causes a decrease in the effective features extracted by the target detection network by masking features such as texture and brightness, which directly leads to a decrease in the prediction results. And, we can learn that the features extracted from each convolution kernel are different, and they do not change linearly with the enhancement of Gaussian noise when processing regional features. At the same time, with the deepening of the convolution layer, the influence of texture and brightness characteristics on the results is no longer obvious. To improve the detection effect, we should start from the two perspectives of suppressing image noise and improving the feature extraction ability of the network. On the one hand, the data set processing module and feature extraction module should be optimized. On the other hand, the convolution kernels need to be designed to extract effective features in a noisy environment to reduce the impact of noise.

IV. CONCLUSION

In this article, we proposed a method to explain the performance degradation of target detection networks by integrating quantitative and qualitative analysis at the feature map level. First, we used four accuracy metrics, namely, precision, recall, F1, and mAP, to characterize the performance of target detection networks under different intensity Gaussian noise. The degradation of precision, recall, F1, and mAP for Yolov5 was 94.2%, 89.5%, 0.92, and 96.3%, respectively; the degradation of precision, recall, F1, and mAP for Faster R-CNN was 100%, 61.9%, 0.48, and 47.6%, respectively; and the degradation of precision, recall, F1, and mAP for SSD was 100%, 49.5%, 0.64, and 61.2%, respectively. The results indicate that the target detection network is extremely sensitive to Gaussian noise in SAR images. Second, we used a hierarchical sampling method to extract the

feature maps of four convolution layers of each network and extract their texture and brightness features. In the quantitative analysis stage, we found that as the noise intensity increased, the value of ASM tended to decrease, with a maximum of 59.10% and a minimum of approximately 0; the value of ENT tended to increase, with a maximum of 97.81% and a minimum of approximately 0; and the value of brightness tended to increase, with a maximum of 100.92%. We can conclude that the Gaussian noise degrades the network performance by masking the features of targets and changing the feature structure of feature maps. The features extracted from each convolution kernel are different, and they do not change linearly with the enhancement of Gaussian noise when processing regional features. At the same time, with the deepening of the convolution layer, the influence of texture and brightness characteristics on the results is no longer obvious. Subsequently, the network can be optimized from two perspectives: suppressing the dataset noise and improving the feature extraction capability of the network.

The proposed method in this article only provides a preliminary analysis and explanation of the causes of network performance degradation under Gaussian noise from the feature map perspective. In the future, the causes of network degradation can be explained from more perspectives, and the structure can be targeted and optimized to achieve better ship detection results.

REFERENCES

[1] X. Wang and C. Chen, "Ship detection for complex background SAR images based on a multiscale variance weighted image entropy method," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 184–187, Feb. 2017, doi: 10.1109/LGRS.2016.2633548.

[2] J. Zhao, Z. Zhang, W. Yu, and T.-K. Truong, "A cascade coupled convolutional neural network guided visual attention method for ship detection from SAR images," *IEEE Access*, vol. 6, pp. 50693–50708, 2018, doi: 10.1109/ACCESS.2018.2869289.

[3] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 3–22, Nov. 2018, doi: 10.1016/j.isprsjprs.2018.04.003.

[4] M. Huang et al., "A bridge neural network-based Optical-SAR image joint intelligent interpretation framework," *Space Sci. Technol.*, vol. 2021, pp. 1–10, Oct. 2021, doi: 10.34133/2021/9841456.

[5] R. L. Paes, J. A. Lorenzzetti, and D. F. M. Gherardi, "Ship detection using TerraSAR-X images in the campos basin (Brazil)," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 3, pp. 545–548, Jul. 2010, doi: 10.1109/LGRS.2010.2041322.

[6] C. Chen, C. He, C. Hu, H. Pei, and L. Jiao, "A deep neural network based on an attention mechanism for SAR ship detection in multiscale and complex scenarios," *IEEE Access*, vol. 7, pp. 104848–104863, 2019, doi: 10.1109/ACCESS.2019.2930939.

[7] Z. Deng, H. Sun, S. Zhou, and J. Zhao, "Learning deep ship detector in SAR images from scratch," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4021–4039, Jun. 2019, doi: 10.1109/TGRS.2018.2889353.

[8] A. Chen, Y. Xie, Y. Wang, and L. Li, "Knowledge graph-based image recognition transfer learning method for on-orbit service manipulation," *Space Sci. Technol.*, vol. 2021, pp. 1–10, Aug. 2021, doi: 10.34133/2021/9807452.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[10] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[11] X. Xu, X. Zhang, and T. Zhang, "Lite-YOLOv5: A lightweight deep learning detector for on-board ship detection in large-scene sentinel-1 SAR images," *Remote Sens.*, vol. 14, no. 4, Jan. 2022, Art. no. 4, doi: 10.3390/rs14041018.

[12] X. Wu, Z. Zhang, W. Zhang, Y. Yi, C. Zhang, and Q. Xu, "A convolutional neural network based on grouping structure for scene classification," *Remote Sens.*, vol. 13, no. 13, Jan. 2021, Art. no. 13, doi: 10.3390/rs13132457.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[15] T. Tian, Z. Pan, X. Tan, and Z. Chu, "Arbitrary-oriented inshore ship detection based on multi-scale feature fusion and contextual pooling on rotation region proposals," *Remote Sens.*, vol. 12, no. 2, Jan. 2020, Art. no. 2, doi: 10.3390/rs12020339.

[16] T. Zhang, X. Zhang, and X. Ke, "Quad-FPN: A novel quad feature pyramid network for SAR ship detection," *Remote Sens.*, vol. 13, no. 14, Jan. 2021, Art. no. 14, doi: 10.3390/rs13142771.

[17] J. Jiao et al., "A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection," *IEEE Access*, vol. 6, pp. 20881–20892, 2018, doi: 10.1109/ACCESS.2018.2825376.

[18] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, May 28 2020, doi: 10.1109/JSTARS.2020.2997081.

[19] Z. Sun, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "BiFA-YOLO: A novel YOLO-Based method for arbitrary-oriented ship detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 21, Jan. 2021, Art. no. 21, doi: 10.3390/rs13214209.

[20] Z. Sun et al., "An anchor-free detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7799–7816, Jul. 26, 2021, doi: 10.1109/JSTARS.2021.3099483.

[21] Z. Lin, K. Ji, X. Leng, and G. Kuang, "Squeeze and excitation rank faster R-CNN for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 751–755, May 2019, doi: 10.1109/LGRS.2018.2882551.

[22] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, Aug. 2017, Art. no. 8, doi: 10.3390/rs9080860.

[23] Z. Cui, X. Wang, N. Liu, Z. Cao, and J. Yang, "Ship detection in large-scale SAR images via spatial shuffle-group enhance attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 379–391, Jan. 2021, doi: 10.1109/TGRS.2020.2997200.

[24] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proc. 2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, Art. no. 88278836, doi: 10.1109/CVPR.2018.00920.

[25] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting CNNs via decision trees," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6254–6263, doi: 10.1109/CVPR.2019.00642.

[26] D. Ho et al., "NBDT: Neural-backed decision trees," EECS Department, University of California, Berkeley, 2020. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-65.html

[27] H. Wang and D.-Y. Yeung, "Towards Bayesian deep learning: A framework and some existing methods," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3395–3408, Dec. 2016, doi: 10.1109/TKDE.2016.2606428.

[28] P.-J. Kindermans et al., "The (Un)reliability of saliency methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller Eds. Cham, Switzerland: Springer, 2019, pp. 267–280, doi: 10.1007/978-3-030-28954-6_14.

[29] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *ITU J.: ICT Discoveries*, vol. 1, no. 1, p. 3948, 2017.

[30] F. Grün, C. Rupprecht, N. Navab, and F. Tombari, "A taxonomy and library for visualizing learned features in convolutional neural networks," Jun. 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1606.07757

[31] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, Aug. 2019, Art. no. 8, doi: 10.3390/electronics8080832.

[32] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Workshop Int. Conf. Learn. Representations*, 2014, Art. no. 1232.

[33] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. 3rd Int. Conf. Learn. Representations, Workshop Track*, San Diego, CA, USA, May 7-9, 2015. [Online]. Available: http://arxiv.org/abs/1412.6806

[34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," in *Proc. 2015 Int. Conf. Learn. Representations*, May 7-9, 2015. [Online]. Available: https://dspace.mit.edu/handle/1721.1/96942

[35] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLos One*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140, doi: 10.1371/journal.pone.0130140.

[36] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.

[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929, doi: 10.1109/CVPR.2016.319.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[39] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[40] H. Wang et al., "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 111–119, doi: 10.1109/CVPRW50498.2020.00020.

[41] R. Fong and A. Vedaldi, "Explanations for attributing deep neural network predictions," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller Eds. Cham, Switzerland: Springer, 2019, pp. 149–167, doi: 10.1007/978-3-030-28954-6_8.

[42] Q. Zhang, X. Wang, R. Cao, Y. N. Wu, F. Shi, and S.-C. Zhu, "Extraction of an explanatory graph to interpret a CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3863–3877, Nov. 2021, doi: 10.1109/TPAMI.2020.2992207.

[43] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Proc. 32nd AAAI Conf. Artif. Intell. 38th Innovative Appl. Artif. Intell. Conf. 8th AAAI Symp. Educational Adv. Artif. Intell.*, New Orleans, Louisiana, USA, Feb. 2018, Art. no. 35303537.

[44] S. O. Arik and T. Pfister, "ProtoAttend: Attention-based prototypical learning," *J. Mach. Learn. Res.*, vol. 21, no. 1, Jun. 2022.

[45] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," *Comput. Vis. ECCV 2016*, Cham, pp. 3–19, 2016, doi: 10.1007/978-3-319-46493-0_1.

[46] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010, doi: 10.1007/s11263-009-0275-4.

[47] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[48] S. Gidaris and N. Komodakis, "LocNet: Improving localization accuracy for object detection" in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 789–798, doi: 10.1109/CVPR.2016.92.

Peng Jia received the B.S. degree in spatial information engineering and the M.S. degree in cartography and geographic information systems from Information Engineering University, Henan, China, in 2002 and 2005, respectively.

His current research interests include photogrammetry and remote sensing.

Xiaowei He received the B.S. degree in information engineering in 2021 from Nanjing University of Aeronautics and Astronautics, Nanjing, China, where she is currently working toward the M.S. degree in optical engineering.

Her research interests include remote sensing classification and radar target detection.

Bo Wang received the B.S. degree in remote sensing science and technology, the M.S. degree in geomatics engineering, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2010, 2012, and 2015, respectively.

He is currently an Assistant Professor with the College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests include spatial information extraction and remote sensing image processing.

Jun Li received the B.S. degree in remote sensing science and technology, the M.S. degree in geomatics engineering, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2015, 2018, and 2021, respectively.

He is currently an Associate Research Fellow with the College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His research interests include remote sensing image processing and deep learning.

Qinghong Sheng received the B.S. degree in photogrammetry and remote sensing, the M.S. degree in cartography and geography information system, and the Ph.D. degree in photogrammetry and remote sensing techniques from Wuhan University, Wuhan, China, in 2000, 2004, and 2008, respectively.

She is currently a Professor with the College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her current research interests include spatial information extraction and SAR target detection.

Guo Zhang received the B.S. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2000 and 2005, respectively.

He is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing and a part-time Professor with the College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interest includes quantitative remote sensing in space geometry.