Journal of Electronic Imaging

JElectronicImaging.org

Multimodal polarization image simulated crater detection

Xin Zhang
Jingjing Zhang
Feng Wang
Xiao Liu
Jun Wu
Teng Li

SPIE. IS&T imaging.org

Xin Zhang, Jingjing Zhang, Feng Wang, Xiao Liu, Jun Wu, Teng Li, "Multimodal polarization image simulated crater detection," *J. Electron. Imaging* **29**(2), 023027 (2020), doi: 10.1117/1.JEI.29.2.023027

Multimodal polarization image simulated crater detection

Xin Zhang,[a] Jingjing Zhang,[a,*] Feng Wang,[b] Xiao Liu,[c] Jun Wu,[d] and Teng Li[a]

[a]Anhui University, Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Electrical Engineering and Automation, Hefei, China
[b]Key Laboratory of Polarized Light Imaging and Detection Technology in Anhui Province, Hefei, Anhui, China
[c]Chinese Academy of Sciences, Anhui Institute of Optics and Fine Mechanics, Key Laboratory of Optical Calibration and Characterization, Hefei, China
[d]Guangzhou and Chinese Academy of Sciences, Institute of Software Application Technology, Nansha, GuangZhou, China

Abstract. Most previous target detection methods are based on the physical properties of visible-light polarization images, depending on different targets and backgrounds. However, this process is not only complicated but also vulnerable to environmental noises. A multimodal fusion detection network based on the multimodal deep neural network architecture is proposed in this research. The multimodal fusion detection network integrates the high-level semantic information of visible-light polarization image in crater detection. The network contains the base network, the fusion network, and the detection network. Each of the base networks outputs a corresponding feature figure of polarization image, fused by the fusion network later to output a final fused feature figure, which is input into the detection network to detect the target in the image. To learn target characteristics effectively and improve the accuracy of target detection, we select the base network by comparing between VGG and ResNet networks and adopt the strategy of model parameter pretraining. The experimental results demonstrate that the simulated crater detection performance of the proposed method is superior to the traditional and single-modal-based methods in that the extracted polarization characteristics are beneficial to target detection.
 [DOI: 10.1117/1.JEI.29.2.023027]

Keywords: target detection; polarization characteristics; multimodal fusion; simulated crater.

Paper 190658 received Jul. 23, 2019; accepted for publication Mar. 24, 2020; published online Apr. 21, 2020.

1 Introduction

In the military field, cameras are usually employed to collect images after a heavy artillery test and the success of this experiment is determined according to the position of the crater in the image. However, because of active or passive interference caused by fog, clouds, and glare, traditional image-based crater-detection methods based on the visible-light band simply cannot meet the basic needs of military research. Detection based on images with polarization[1] is a new approach in which a photoelectric imaging device is used to obtain the target scene radiation, spatial information, spectral information, and polarization information.[2] The evaluation requirements can be initially met using the difference in polarization characteristics between the target and the background to extract the target object. However, this process[3] is complicated, cumbersome, and usually inaccurate.

Recently, researchers have focused on detecting targets in polarization images using physical information such as polarization, texture, and spectral information. We visualize some of the physical features in Fig. 1, where Fig. 1(a) presents a texture image produced using the local binary pattern (LBP) algorithm[4] from a visible-light polarization image, and Fig. 1(b) presents a polarization image of the visible light using the Stokes equation.[5] Previous target detection

*Address all correspondence to Jingjing Zhang, E-mail: fannyzjj@ahu.edu.cn

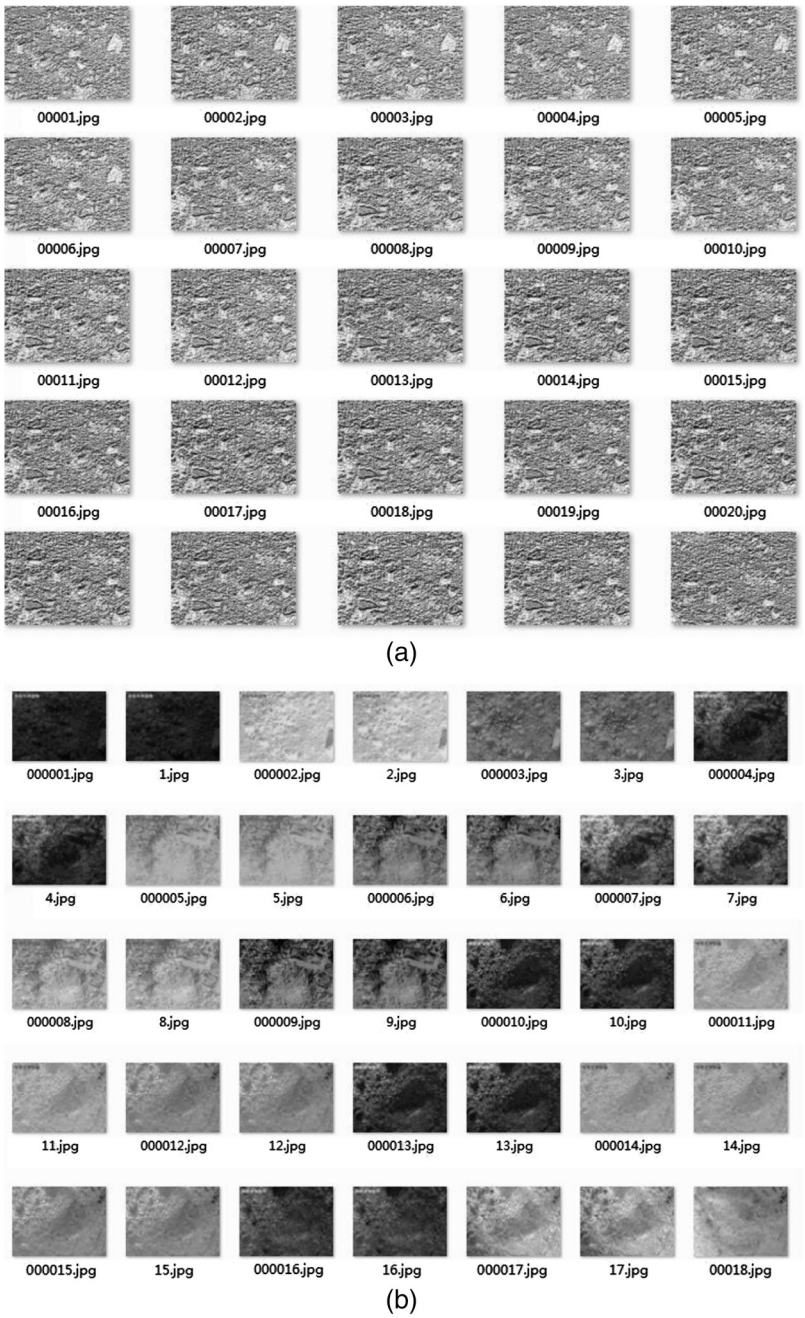

Fig. 1 (a) Example images of the texture diagram of the polarized image of visible light; (b) examples of the visible polarization image representing polarization information through the Stokes equation. (a) Texture image and (b) polarization image.

methods employing visible-light polarization images can be divided into two categories: methods based on prior information[6–10] and methods based on external devices.[11–13] Early methods, such as polarization information fusion enhancement,[6,8] multiband fusion priors,[10] and algorithm prior optimization,[14] mostly use prior parameter estimations of the polarization characteristics. In contrast, external device-based methods obtain these parameters from external conditions, which mainly depend on the visible-light polarization detection system and are based on its mechanical design. These methods are based on the physical information of the image and tend to lose detailed information and focus only on certain feature information. Therefore, the target location cannot be detected accurately in different scenarios. In this study, we obtain these parameters from training data using a deep learning-based approach and an improved multimodal network.

Multimodal deep learning[15] has been used successfully in audio-visual classification as well as in shared-representation learning. Multimodal networks are currently used for target detection in synthetic aperture radar images; for instance, in Ref. 16 in which a multiscale convolutional neural network (CNN) model is used to extract the features learned by multiscale training directly from the image patches to detect built-up areas. Furthermore, Ref. 17 proposed a deep fusion network by adding more base networks and focusing on how they are integrated.

In this study, our goal is to accurately detect a target crater in visible-light polarization images. To achieve this aim, some existing methods[11–13,18] change the equipment and configuration of the camera, for instance, by employing liquid crystal variable retarders.[18] Others[19–24] rely on analyzing the polarization characteristics to highlight the target. The above methods solely focus on learning knowledge representation from a single modality, yet neglect the complementary information from others.[25]

In the proposed method, we obtain and represent the polarization information of the visible-light polarization image using the Stokes equation and the LBP algorithm, respectively. Because the resultant images have a wealth of semantic information, we use a neural network for further processing. That is, we utilize a CNN to learn the target features in the multisource information image, extract semantic information, fuse features, and ultimately detect the target accurately. In contrast to previous methods, we not only use physical information, but also high-level semantic information. In addition, our method can fully automatically detect and mark the location of a target crater in an image.

We present the following contributions in this paper: (1) we collected many real images of simulated craters through a large number of experiments and created a comprehensive dataset consisting of six small datasets based on the physical information of the polarized images. (2) We propose a multimodal fusion detection algorithm that combines the physical information and semantic information of polarized images to detect craters effectively. (3) Our proposed algorithm can quickly and accurately detect craters and mark them automatically. (4) We obtain a lightweight fusion model through experiments comparing different base network frameworks that can be used to efficiently perform target detection tasks.

2 Related Work

Several methods have been proposed to solve the problem of target detection in visible-light polarization images. Some require additional information. For instance, in a polarization imaging detection system, Ref. 18 used a liquid crystal variable retarder as a phase delay device. After the image was acquired by the detection system, targets were detected to obtain their contours and partial details. The system proposed in Ref. 19 used the differences in the polarization characteristics of a target and the background to design a visible-light polarization detection system based on a double line polarizer. Alternatively, Ref. 20 proposed a noncontact road condition detection method that is illuminated by a near-infrared quartz-halogen tungsten lamp. Using a rotating polarizer, images of four polarization direction components were sequentially collected, and then the degree of linear polarization was extracted to detect road conditions (such as icy, wet, and dry surfaces). These methods rely on certain external experimental conditions.

Many methods extract polarization information from visible-light polarization images using the Stokes equation and then fuse certain features of that information to detect targets. In Ref. 21, a calculation based on the Stokes vector and Mueller matrix was proposed that can determine the degree of line polarization in an arbitrary polarization direction. A system for the detection of polarization for low-illumination camouflage targets in multiple directions was implemented. Zhao et al.[22] used a polarization image enhancement method based on Stokes parameters to improve the detection and recognition rate of targets. In Ref. 23, the Stokes vector was used to calculate the polarization angle and degree of each polarization image based on the least-squares method to obtain the polarization degree image. At the same time, the local entropy was extracted from the polarization image and thresholded to obtain a binary image. The polarization degree image and binary image were synthesized to create a composite image that was then decomposed into binary connected domains for target detection.

There are also several methods that use image spatial information and feature fusion. For example, to improve the quality of visible images and the detection rate of artificial targets hidden in natural backgrounds, Ref. 24 proposed a method based on polarization imaging that could highlight artificial targets and provide more details and texture information. The authors of Ref. 26 applied hue, saturation, value-RGB image fusion technology to a polarization correlation-based imaging system and effectively fused multiple polarization images to comprehensively describe target structure and improve target detection and recognition efficiency. In Ref. 27, using multidimensional information from polarization images, a method of suppressing image background based on fused polarization information was proposed for target detection against complex cloud or sea backgrounds.

These previously proposed methods are all based on the physical information of images and prior knowledge. Our proposed approach in this work is fundamentally different in that we employ the physical information of images as a training dataset and then use deep learning to train the model to collect data of the multimodal target information in these images for automatic target detection.

3 Proposed Method

Most previous deep learning methods depend on single-modal image input and hence require a complicated process to learn effective feature representations and detect the target. Here, we propose a multimodal fusion detection algorithm that is based on the characteristics of visible-light polarization.

Our proposed method can be divided into four main steps. First, the physical information of the visible-light polarization image is obtained by the Stokes equation and the LBP algorithm and used as input to the base networks. Second, target features in the three images are extracted simultaneously through three identical CNNs and the semantic information of the target in the image is learned, thereby reducing the network training time. Third, the output feature images of these three networks are fed into the fusion network. These images are fused and more features are extracted. Finally, the output is fed to the detection network to detect the target in the image. In addition, we used a pretrained model and fine-tuned it to improve the network computational speed and detection accuracy, as explained in detail below. Figure 2 shows the overall multimodal fusion detection network architecture.

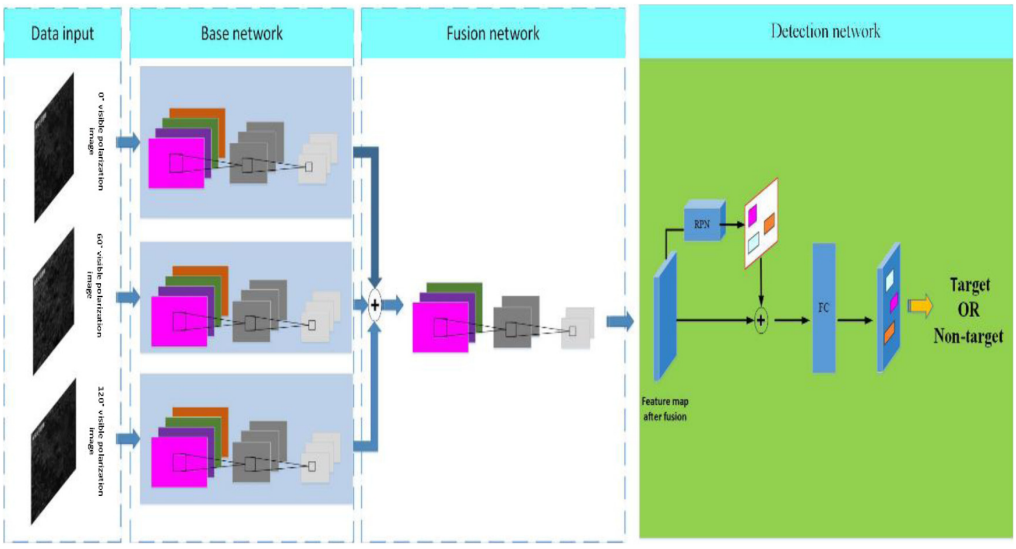

Fig. 2 Multimodal fusion detection network architecture. Different colors represent various layers and the plus sign indicates fusion. The architecture consists of three convolution networks: a base network, fusion network, and detection network. The inputs are three images, and they are converted into features by the base network. The three features are fused by the fusion network to obtain the fused feature, which is then input to the detection network to detect the target.

3.1 *Data Input*

In the network, the input data consist of either real images of simulated craters, which consists of images captured by a visible-light polarization camera, or images converted from the simulated crater dataset. The dataset is introduced in Sec. 4. We labeled all images in the dataset. For each target detection, three images containing different types of target information were input into the base network to extract the features.

3.2 *Base Network*

First, the CNN part of a VGG[28] network was adopted as the base network of our multimodal network for training. We used very small 3×3 receptive fields for convolution with each pixel of the input. However, in the experiment, the training speed was slow because there were too many model parameters. For the final network, we choose a ResNet[29] network with batch normalization layers and activation layers as the base network of the multimodal network. Experiments showed that the output feature of the ResNet-4 network contains the most discriminative information and yields the best detection performance.

Figure 3 shows the network architecture of a residual block, and Fig. 4 presents the network architecture of the base network. Each residual block has a batch normalization layer and an activation layer to avoid the disappearing gradient problem and speed up learning. The architecture of each residual block is the same. Res2, Res3, and Res4 stages are composed of res_units (residual blocks), which extract the initial features of the image and sharpen the edges in the image.

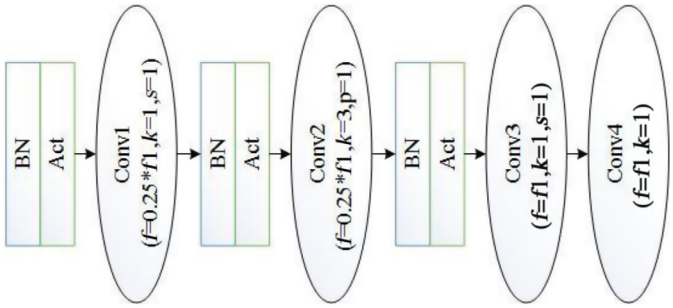

Fig. 3 ResNet50 res_units network architecture. *f*1 represents the input filter numbers, *k* is the kernel, *s* is the stride, and *p* is the padding.

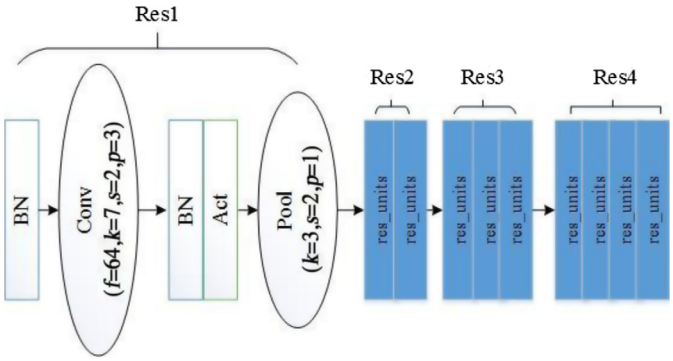

Fig. 4 Base network architecture. *k* is the kernel, *s* is the stride, *p* is the padding, Res2, Res3, and Res4 uses the res_units as the parameters. (a) Simulated crater dataset, (b) IQU dataset, (c) I dataset, (d) Q dataset, (e) U dataset, and (f) P dataset.

3.3 *Fusion Network*

Our fusion network is similar to a deep fused network[17] architecture and a deep fusion network is multi-input. Network fusion is a process of combining multiple base networks, such as K base networks $\{H^1_{L_1}, \cdots H^K_{L_K}\}$. The conventional fusion, in general, includes two approaches: feature fusion, fusing the feature representations extracted from the networks together, and decision fusion, fusing the scores computed from the networks. Our method focuses on feature fusion and the fusion can be formulated in the function form $H_{(x_0)} = F[H^1_{L_1}(x_0), \ldots, H^K_{L_K}(x_0)]$, where the fusion function $F(.)$, in our method, is the sum of the representations:

$$F[H^1_{L_1}(x_0), \ldots, H^K_{L_K}(x_0)] = \sum_{k=1}^{K} H^K_{L_K}(x_0). \tag{1}$$

3.4 *Detection Network*

In the target detection network, the detection network of faster region-CNN (Faster R-CNN)[30] is used. Faster R-CNN is a two-stage detector mainly consisting of three major components: shared bottom convolutional layers, a region proposal network (RPN), and a region-of-interest (ROI)-based classifier. In our method, we only use the two components of RPN and ROI.

First, feature figures are a shared bottom feature figure. Based on that feature figure, RPN generates candidate object proposals where afterward the ROI-wise classifier predicts the category label from a feature vector obtained using ROI pooling. The training loss is composed of the loss of the RPN and the loss of the ROI classifiers:

$$L_{\text{det}} = L_{\text{rpn}} + L_{\text{rot}}. \tag{2}$$

Both training losses of the RPN and ROI classifiers have two loss terms: one is to classify the accuracy of prediction probability and the other is a regression loss on the box coordinates for better localization.

3.5 *Transfer Learning*

Neural networks are trained with data. They obtain the information from the data and convert this information into the corresponding weights. These weights can be extracted and transferred to other neural networks, which enables us to "transfer" these learned features without having to train another neural network from scratch. Some researchers use a VGG16 or VGG19 pretrained model to perform the initial training on a network. According to the characteristics of the visible-light polarization image, we input visible-light polarization images into a single-model network, extract the features in the network, and train the single model. We use this model as the pretrained model of the multimodal fusion network. Because the pretrained model has good generalization performance, we can use an analogous structure and the weights directly when training using the new dataset to improve the accuracy of target detection.

4 Experiments

In this section, we present the experimental details, experimental results, and discuss the results of other methods. Our experimental datasets are obtained from real images taken by cameras. Of the possible cameras we used the polarization spectrometer and the camera of a spectrometer polarization imaging detection system. The camera operating mode is simultaneous imaging with a single channel; the polarization directions are 0 deg, 60 deg, and 120 deg; and the imaging band was 400 to 1100 nm.

4.1 *Datasets*

We collected visible-light polarization images of craters that we simulated at a test site to verify our methods. The ground types of the test site include soil, sand, grassland, and other types. We created a simulated crater and used the visible-light polarization camera from different angles and different heights to obtain images for the simulated crater dataset, which is also called the uncharacterized dataset. Using the Stokes equation, the images in the simulated crater dataset were processed and used to create a dataset called the incident light intensity and linear polarization information (IQU) dataset, which belongs to the characterized dataset. Using the LBP algorithm, the texture images corresponding to the simulated crater dataset were also obtained. We combined the IQU dataset and texture diagrams with the simulated crater dataset and processed them to get four semicharacterized datasets: the I dataset, Q Dataset, U dataset, and P dataset. The Stokes equation expression is as follows:

$$I = \frac{2}{3}[I(0\ \ \mathrm{deg}) + I(60\ \ \mathrm{deg}) + I(120\ \ \mathrm{deg})], \tag{3}$$

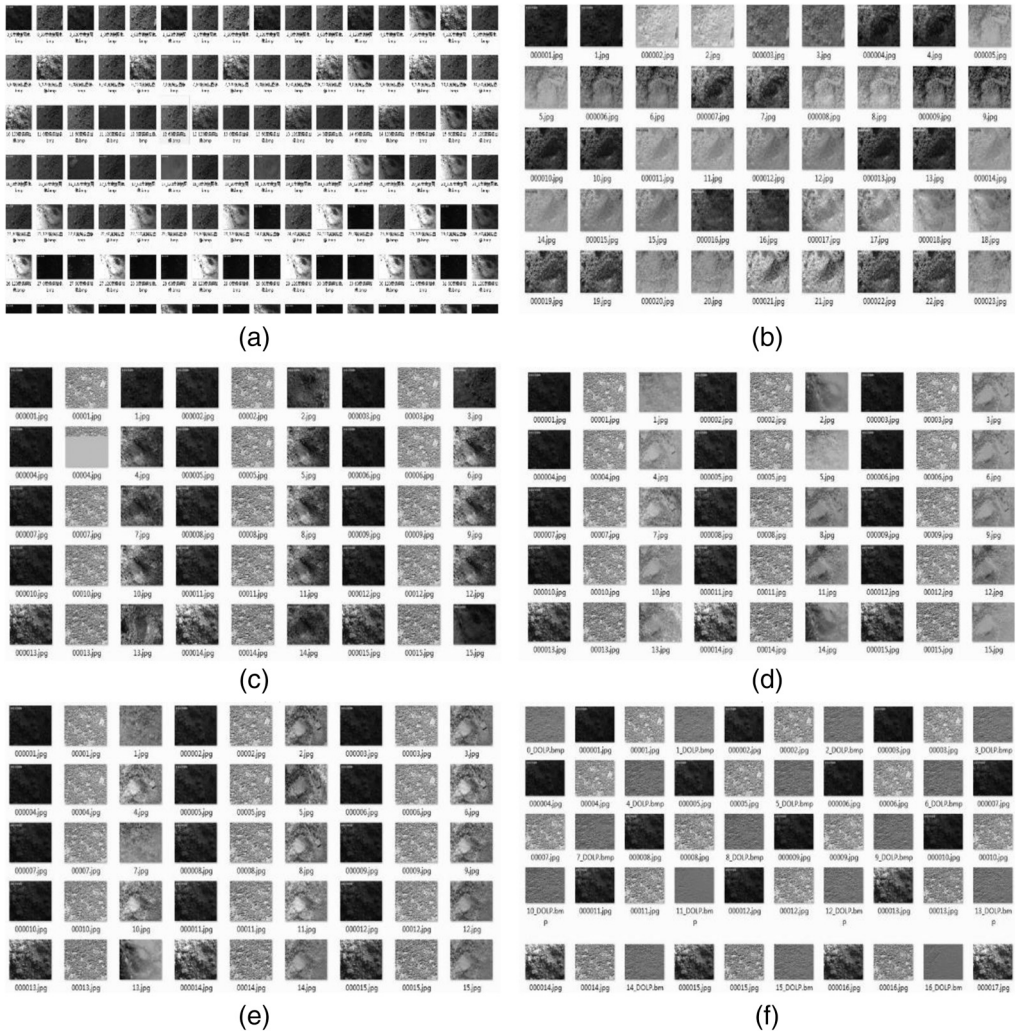

Fig. 5 Some of the images in the six datasets used in experiments are shown: (a) uncharacterized dataset, (b) characterization dataset, (c), (d), (e), and (f) the semicharacterization dataset. (a) Visible light polarization image, (b) polarization image, (c) polarization feature image, (d) mean fusion image, (e) Laplace fusion enhancement, (f) wave fusion enhanced, (g) Brovey fusion image, (h) multifocus fusion enhancement image, and (i) our method.

$$Q = \frac{2}{3}[2I(0\ \text{deg}) - I(60\ \text{deg}) - I(120\ \text{deg})], \tag{4}$$

$$U = \frac{2[I(60\ \text{deg}) - I(120\ \text{deg})]}{\sqrt{3}}, \tag{5}$$

$$P = \frac{\sqrt{Q^2 + U^2}}{I}. \tag{6}$$

In the equation: $I(\theta\ \text{deg})$ represents a polarization image of the polarizing plate at a rotation angle of $\theta(\theta \in [0\ \text{deg}, 360\ \text{deg}])$, I is related to the incident light intensity, Q is related to the linear polarization information in the 0 deg, 60 deg, and 120 deg directions, U is related to the linear polarization information in the 60 deg and 120 deg directions, and P is the degree of polarization.

The simulated crater dataset contains 2403 visible-light polarization images and the IQU dataset contains 2403 characterized images. Moreover, the I dataset contains the simulated crater dataset, corresponding texture images, and 801 I images; the Q, U, and P datasets are similar to the I dataset, except that I images are replaced by the Q, U, and P images, respectively, As is shown in Fig. 5.

4.2 *Training Parameters*

We trained all networks on a NVIDIA Geforce GTX 1080 Ti GPU. The proposed framework was implemented using the MXNet toolbox. The size of the input images was 256×256. The network was trained with a minibatch size of 16, a learning rate of 0.001, 10 training epochs, and a learning rate decay every seven epochs. Further, the optimization method was the Adam optimizer. These parameters were constant in our experiments. We used 80% of the images in the dataset as a training dataset and the remaining 20% of the images as the test dataset.

5 Result

5.1 *Traditional Polarization Detection Experiments*

In Fig. 6, we show the target detection results obtained by different methods, where the final image was obtained using our proposed method. It can be concluded from Fig. 6 that the test results obtained by the previous methods require manual observation, but our method can automatically and accurately detect the target and mark the target area.

5.2 *Single-Modal Experiments*

The results obtained by our method and single-modal methods for the simulated crater dataset are compared. The single-modal networks consist of a faster R-CNN (VGG) network, faster R-CNN (ResNet) network, and the Yolov3[31] network. The resulting models were validated on a test dataset, as shown in Table 1.

According to the precision and mean average precision (mAP) metrics, our model obtains better results than most previous methods. In addition, because two base networks are used in the proposed network, the two models generated are different in size. The ResNet50 network is a better base network than the VGG16 base network. Moreover, its detection precision is higher and model size is smaller, so we obtain a better lightweight network model. In addition, this result demonstrates that using a multimodal network framework and base network to extract features and fuse them to detect targets is effective.

5.3 *ResNet50 Fused Experiments*

In order to get the feature figure with the most target information, we find out which stage of ResNet50 has the best fusion effect based on the receptive field[32] formulation:

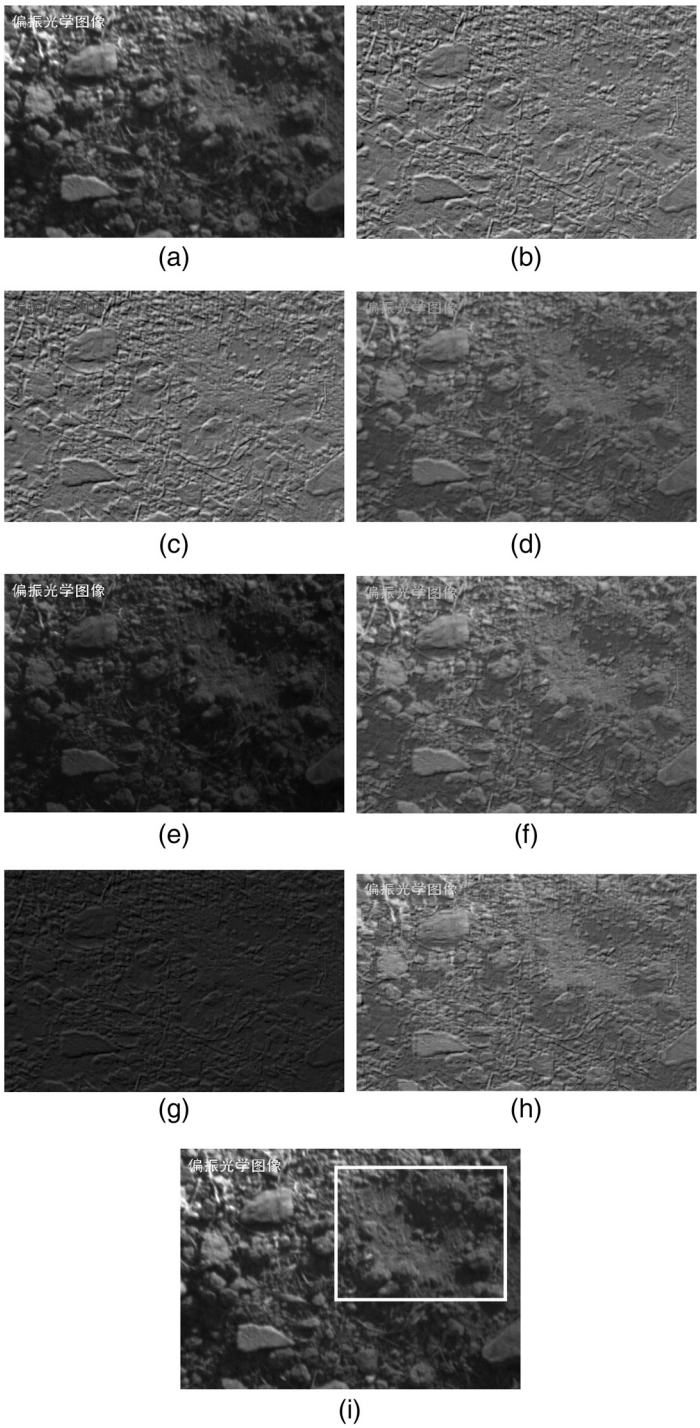

Fig. 6 Above nine pictures in order: the first eight images are obtained by using the previous methods of image fusion to detect the simulated crater, and the last one is generated using our proposed method to detect the simulated crater in the image. (a) Original image, (b) simulated crater real box, (c) faster R-CNN (VGG) test image, (d) faster R-CNN (ResNet) test image, (e) FV1 test image, and (f) FR1 test image.

$$l_k = l_{k-1} + \left[(f_k - 1) * \sum_{i=1}^{k-1} s_i \right], \tag{7}$$

where l_{k-1} is the receptive field of layer $k-1$, f_k is the filter size (height or width, but assuming they are the same here), and s_i is the stride of layer i.

Table 1 Experimental comparison table of the simulated crater dataset. FV represents our method to use VGG16 base networks, FR represents our method to use ResNet50 base networks.

Method	Precision	mAP	Model size
Faster R-CNN (VGG)	66.58%	64.31	546.8 MB
Faster R-CNN (ResNet50)	69.78%	66.06	113.3 MB
Yolov3	68.63%	65.92	246.3 MB
FV	71.35%	67.43	1.5 G
FR	78.50%	69.05	198.3 MB

We hypothesize that if the receptive field of the output in the ResNet_4 network covers the original image, the precision will be maximized. We used the IQU dataset as the input dataset to verify this. The ResNet50 network was split into five parts, namely ResNet_1, ResNet_2, ResNet_3, ResNet_4, and ResNet_5. The features of each stage were outputted and then analyzed. The five results are shown in Table 2. The detection precision increases gradually after each of the outputs of ResNet_1, ResNet_2, ResNet_3, and ResNet_4 is fused. However, when the feature output by ResNet_5 is fused, the detection precision decreases. Therefore, the output of the fourth layer of the ResNet network is the best.

5.4 *Multimodal Experiments*

For each labeled image in each training dataset, three images with different modals were input to the three networks in the base network to obtain three base network features. The outputs were then fed into the fusion network and detection network. Continuous learning in the network generated the final model. We then verified the multimodal-fusion detection model on the verification dataset, and the results show that the precision obtained is better than the precision of the single-modal model.

The model test results for 12 experiments are shown in Table 3. We used precision, mAP, and model size to evaluate our results. Moreover, we set the intersection over the union (IoU) threshold to 0.8 in the experiment. When IoU ≥ 0.8, the simulated crater is correctly detected. As can be observed in Table 1, the test precision of the multimodal fusion detection model is higher than that of the single model. In Table 3, FV represents our method to use VGG16 base networks, CS represents simulated crater dataset, FR represents our method to use ResNet50 base networks. Therefore, FV–CS and FR-CS are experiments using simulated crater datasets, FV-IQU and FR-IQU are experiments using the characterized datasets, and the rest are experiments using semi-characterized datasets. As can be seen from Table 3, the best precision is obtained in the FV-IQU and FR-IQU experiments using the visible-light polarization images characterized by the Stokes equation. These experimental results show that the polarization information is beneficial for

Table 2 Using ResNet50 as the base network, it is judged that the fusion of feature diagrams in a certain part of the base network can make the best detection effect, and five experiments have been carried out and the precision rates have been obtained.

Method	Precision (%)
ResNet_1	79.42
ResNet_2	80.34
ResNet_3	83.75
ResNet_4	88.60
ResNet_5	84.68

Table 3 FV-SC represents the use of a VGG16 base network to experiment with simulated crater dataset. Similarly, FR-SC represents the use of the ResNet50 base network to experiment with simulated crater dataset. Therefore, the name of after '-' represents used the dataset, and before '-' represents the used method.

Data type	Method	Precision (%)	mAP	Model size
Uncharacterized dataset	FV-SC	71.35	67.43	1.5 G
	FR-SC	78.50	69.05	198.3 MB
Characterized dataset	FV-IQU	80.88	70.87	1.5 G
	FR-IQU	88.60	75.63	198.3 MB
Semicharacterized dataset	FV-I	78.93	69.93	1.5 G
	FR-I	85.99	73.02	198.3 MB
	FV-Q	78.01	69.96	1.5 G
	FR-Q	**89.10**	**76.28**	198.3 MB
	FV-U	74.72	68.77	1.5 G
	FR-U	63.85	61.09	198.3 MB
	FV-P	79.68	70.41	1.5 G
	FR-P	86.45	73.56	198.3 MB

target detection. In the multimodal-fusion detection network, the detection results using the ResNet50 base network are better than those using the VGG16 base network. Moreover, we use two different base networks to obtain different model sizes. The size of the network model generated by the VGG-based network was 1.5 G, and the model size obtained by ResNet50-based network was 198.3 MB, which is a factor of 7.7. Hence, it is possible to obtain a lighter multimodal fusion detection model. We, therefore, conclude the following: in multimodal fusion detection networks, polarization information is beneficial for target detection in visible-light polarization images. At the same time, the use of different base networks for target detection leads to different performances.

5.5 *Qualitative Results on a Visible-Light Polarization Image*

Our trained multimodal fusion detection model was tested on the images of the simulated crater dataset. Figure 7 shows the detection results of a visible-light polarization image. The red boxes represent the true position and size of the crater, and the blue boxes indicate the coordinates predicted by various models. We selected a visible-light polarization image taken on sand for detection. There is a clear difference between the target and background in images taken of craters on soil and grass, but the sand is light in color and reflective, so a target in sand is difficult to accurately detect. However, the use of this challenging image better demonstrates that our method can accurately detect targets in images.

Figure 7(a) shows the original image of the simulated crater, where the crater is located in the top right of the image. Figure 7(b) shows the ground truth. Figures 7(c) and 7(d) both show the prediction results of a single-modal method. In these images, the red and blue boxes do not overlap much, indicating that the prediction results are poor. Figures 7(e) and 7(f) are the results of our method. Here, the overlapping parts of the two boxes are large. These figures show that the multimodal fusion detection model using the ResNet-based network better detects the simulated crater. These results demonstrate that our proposed multimodal fusion detection networks can take full advantage of multiple types information provided by visible-light polarization images. Of these proposed models, the lightweight multimodal fusion detection model yields the best detection performance on the simulated crater dataset.

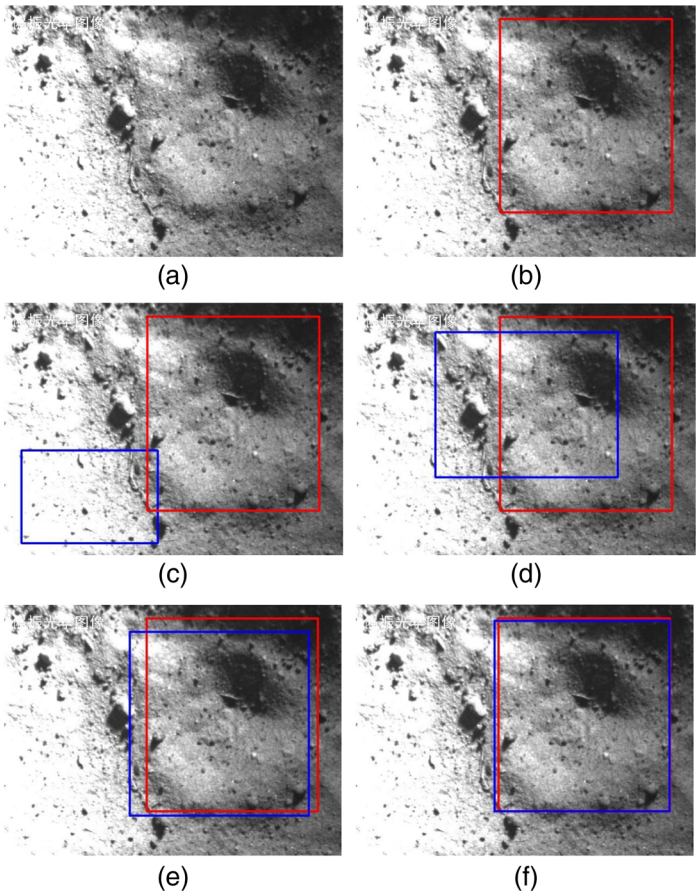

Fig. 7 The test results of different models on the same image.

6 Conclusion

In this paper, a multimodal-fusion detection algorithm based on a multimodal network architecture was proposed. Its aim is to accurately detect targets in visible-light polarization images. ResNet50 was selected as the base network to extract multiscale features of the target in the input image. Then the target features were fused using the fusion network, which obtains a target multifeature output. Finally, the detection network is trained and the target is detected. The experimental results show that precision of target detection can be improved by adding polarization features, and the multimodal-fusion detection network can effectively detect targets in visible-light polarization images. Of the evaluated approaches, the lightweight multimodal fusion detection model has a good detection performance on simulated craters in visible-light polarization images.

Acknowledgments

This work was supported by Anhui Provincial Natural Science Foundation (Grant No. 1808085MF209); supported by Open Research Foundation of Key Laboratory of Polarization Imaging Detection Technology Anhui Province (Grant No. 2019KJS030009); and supported by Open Research Foundation of Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education. This work was also supported by Guangzhou Science and Technology Plan Project Funding, China (Grant No. 201907010020).

References

1. J. Ward et al., "An acousto-optic based hyperspectral imaging camera for security and defence applications," *Proc. SPIE* **7835**, 78350U (2010).

2. H. Dongmei, L. Quan, and N. Guocheng, "Low contrast target detection based on visible light polarization imaging system," *Laser Optoelectron. Prog.* **54**(6), 061101 (2017).
3. Q. Li et al., "Ultraviolet-visible polarimetric imaging and image fusion technology with high resolution and large field-of-view," *Acta Opt. Sin.* **39**(6), 0611001 (2019).
4. H. Liu, "Improved LBP used for texture feature extraction," *Comput. Eng. Appl.* **50**(6), 182–185 (2014).
5. Y. Yulong et al., "Phase delay error analysis of wave plate of division-of-amplitude full Stokes simultaneous polarization imaging system," *Acta Phys. Sin.* **68**(2), 024203 (2019).
6. Z. Wu, S. Zhou, and X. He, "Experimental study on underwater objects detection based on polarization imaging," *Laser Optoelectron. Prog.* **55**(8), 081101 (2018).
7. Z. T. Chen, X. B. Sun, and Y. L. Qiao, "Cloud detection over ocean from PARASOL/POLDER3 satellite data," *J. Remote Sens.* **22**(6), 996–1004 (2018).
8. Y. Liang, W. Yi, and H. Huang, "Detection of target on ocean background based on polarization imagery fusion," *J. Atmos. Environ. Opt.* **11**(1), 60–67 (2016).
9. X. Li and Q. Huang, "Target detection for infrared polarization image in background of desert," in *IEEE 9th Int. Conf. Commun. Software and Networks*, Vol. 9, pp. 779–782, 792 (2016).
10. Z. Gang et al., "Multi-band polarimetric image fusion based on IHS and wavelet transform," *Comput. Meas. Control* **13**(9), 992–994 (2005).
11. Y. Xun et al., "Recognition of camouflage targets by polarization spectral imaging system," *J. Appl.Opt.* **37**(4), 537–541 (2016).
12. X. Mengxi et al., "Research on three-channel synchronous polarization imaging and target detection method," *J. Instrum.* **34**(11), 2408–2417 (2013).
13. X. Wang et al., "Hyperspectral polarization characteristics of typical camouflage target under desert background," *Laser Optoelectron. Prog.* **55**(5), 051101 (2018).
14. Z. Yan et al., "Infrared surface target enhancement based on virtual variational polarization," *Syst. Eng. Electron.* **37**(5), 992–997 (2015).
15. J. Ngiam et al., "Multimodal deep learning," in *Int. Conf. Mach. Learn.*, pp. 689–696 (2011).
16. J. Li, R. Zhang, and Y. Li, "Multiscale convolutional neural network for the detection of built-up areas in high-resolution SAR images," in *IEEE Int. Geosci. Remote Sens. Symp.*, IEEE (2016).
17. J. Wang et al., "Deeply-fused nets," arXiv:1605.07716 (2016).
18. X. Zongjie, "The research on target detection technology based on polarization imaging," Dissertation
19. W. Chen et al., "Polarization detection of marine targets covered in glint," *Infrared Laser Eng.* **46**(S1), 63–68 (2017).
20. Y. Huizhen et al., "Noncontact road condition detection method based on degree of linear polarization," *Instrum. Technol. Sens.* **08**, 89–91, 105 (2017).
21. H. Yanhua et al., "Influence of multi-polarization direction on DOLP extraction of camouflage target under low illumination," *Laser Infrared* **48**(06), 744–749 (2018).
22. Z. Rong, "Research on image enhancement technology for visible light polarization imaging," Dissertation (2016).
23. M. Jie and Y. Zihan, "Water surface target detection method based on polarization characters," CN 106682631 A (2016) (in Chinese).
24. Z. Rong et al., "Visible light image enhancement based on polarization imaging," *Laser Technol.* **40**(02), 227–231 (2016).
25. F. Nian et al., "Multi-modal knowledge representation learning via webly-supervised relationships mining," in *Proc. 25th ACM Int. Conf. Multimedia*, pp. 411–419 (2017).
26. Z. Jiamin et al., "Application of image fusion in polarization correlated imaging," *Infrared Laser Eng.* **47**(12), 1226002 (2018).
27. Z. Xiaojie, L. Sha, and L. Xiang, "Evaluation of background suppression method based on polarization information fusion," *Shanghai Aerosp.* **36**(01), 23–28 (2019).
28. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 702, arXiv:1409.1556 [cs] (2014).
29. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, pp. 770–778 (2016).

30. S. Ren et al., "Faster R-CNN: towards real-time object detection with region proposal networks," in *Neural Inf. Process. Syst.*, pp. 91–99 (2015).
31. J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," in *Comput. Vision Pattern Recognit.* (2018).
32. N. Ketkar, "Convolutional neural networks," in *Deep Learning with Python*, Apress, Berkeley, California (2017).

Xin Zhang received her bachelor's degree in automation from Suzhou University, Suzhou, China, in 2018. She is pursuing her master's degree of pattern recognition at Anhui University, Hefei, China. She is currently developing techniques to solve target detection problems. Her research interests include deep learning, support vector machine, and algorithm.

Jingjing Zhang received her BS degree from Hefei University of Technology, China, in 1997, MS and PhD degrees from Hefei Institute of Physical Science, Chinese Academy of Sciences, China, in 2000 and 2009. She is an associate professor with the School of Electrical Engineering and Automation, Anhui University, Hefei, China. She is mainly engaged in the research on image processing, pattern recognition, and remote sensing information processing.

Feng Wang received his PhD graduated from the Anhui Institute of Optics and Fine Mechanics CAS, in 2007. He is engaged in directional polarization imaging detection and remote sensing information processing. He is currently the director of Key Laboratory and a doctoral tutor.

Xiao Liu received his PhD in optical from University of Chinese Academy of Sciences, Beijing, China, in 2013. Currently, he is an assistant research fellow of Hefei Institute of Physical Science, Chinese Academy of Sciences, Hefei, China. At the same time, he is an assistant research fellow of Key Laboratory of Optical Calibration and Characterization, Chinese Academy of Sciences, Hefei, China. His research interest includes polarization remote sensing and photonics image detection.

Jun Wu graduated from the University of Tokyo in Japan, major in computer science. He used to work as a researcher in INRIA (Institut national de recherche en informatique et en automatique) in France. He worked as director of IVA (Intelligent Video Analysis) Laboratory in Institute of Software Application Technology, Guangzhou and Chinese Academy of Science. He was awarded special allowances experts of the State Council in 2018.

Teng Li received BS degree from University of Science and Technology of China (USTC) in 2001, MS degree from Institute of Automation, Chinese Academy of Sciences (CASIA) in 2004, and PhD from Korea Advanced Institute of Science and Technology (KAIST) in 2010. He received IEEE T-CSVT best paper award in 2014, and ChinaMM excellent paper award in 2007, and ICIMCS best paper award in 2009. He is currently a professor with Anhui University.