Research Article

A Bridge Neural Network-Based Optical-SAR Image Joint Intelligent Interpretation Framework

Meiyu Huang , Yao Xu, Lixin Qian , Weili Shi, Yaqin Zhang, Wei Bao , Nan Wang, Xuejiao Liu, and Xueshuang Xiang

Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China

Correspondence should be addressed to Xueshuang Xiang; xiangxueshuang@qxslab.cn

Received 4 August 2021; Accepted 23 September 2021; Published 12 October 2021

Copyright © 2021 Meiyu Huang et al. Exclusive Licensee Beijing Institute of Technology Press. Distributed under a Creative Commons Attribution License (CC BY 4.0).

The current interpretation technology of remote sensing images is mainly focused on single-modal data, which cannot fully utilize the complementary and correlated information of multimodal data with heterogeneous characteristics, especially for synthetic aperture radar (SAR) data and optical imagery. To solve this problem, we propose a bridge neural network- (BNN-) based optical-SAR image joint intelligent interpretation framework, optimizing the feature correlation between optical and SAR images through optical-SAR matching tasks. It adopts BNN to effectively improve the capability of common feature extraction of optical and SAR images and thus improving the accuracy and application scenarios of specific intelligent interpretation tasks for optical-SAR/SAR/optical images. Specifically, BNN projects optical and SAR images into a common feature space and mines their correlation through pair matching. Further, to deeply exploit the correlation between optical and SAR images and ensure the great representation learning ability of BNN, we build the QXS-SAROPT dataset containing 20,000 pairs of perfectly aligned optical-SAR image patches with diverse scenes of high resolutions. Experimental results on optical-to-SAR crossmodal object detection demonstrate the effectiveness and superiority of our framework. In particular, based on the QXS-SAROPT dataset, our framework can achieve up to 96% high accuracy on four benchmark SAR ship detection datasets.

1. Introduction

With the rapid development of deep learning, remarkable breakthroughs have been made in deep learning-based land use segmentation, scene classification, object detection, and recognition in the field of remote sensing in the past decade [1–4]. This is mainly due to the powerful feature extraction and representation ability of deep neural networks [5–8], which can well map the remote sensing observations into the desired geographical knowledge. However, the current mainstream interpretation technology for remote sensing images is still mainly focused on single-modal data and cannot make full use of the complementary and correlated information of multimodal data from different sensors with heterogeneous characteristics, resulting in insufficient intelligent interpretation capabilities and limited application scenarios. For example, optical imaging is easily restricted by illumination and weather conditions, based on which accurate interpretation cannot be obtained at night or under

complex weather with clouds, fog, and so on. Compared with optical imaging, synthetic aperture radar (SAR) imaging can achieve full-time and all-weather earth observations. However, due to the lack of texture features, it is difficult for SAR images to be interpreted even by well-trained experts. Therefore, gathering sufficient amounts of training SAR data with diverse scenes and accurate labeling is a challenging problem, which heavily affects the deep research and application of SAR images based on intelligent interpretation. To address the above issues, multimodal data fusion [9–12] becomes one of the most promising application directions of deep learning in remote sensing, especially the combined utilization of SAR and optical data because these data modalities are completely different from each other in terms of geometric and radiometric appearance [13–17].

However, the existing optical-SAR fusion techniques mainly concentrate on the matching problem. The proposed optical-SAR image matching methods can be divided into three types: signal-based, hand-crafted feature-based, and

FIGURE 1: The BNN-based optical-SAR image joint intelligent interpretation framework.

deep learning-based approaches. For the signal-based similarity measures, crosscorrelation (CC) [18] and mutual information (MI) [19, 20] have been widely used to the optical-SAR matching tasks. Since MI is an intensity-based statistical measure and has good adaptability to geometric and radiometric changes, it is extensively outperformed CC in optical-SAR image matching. Nevertheless, signal-based approaches that do not contain any local structure information are not robust and accurate enough for matching multisensor images. Feature-based methods commonly utilize invariant key points and feature descriptors. The reason why it comes to a better result may stem from the feature descriptors, which are less sensitive to the geometric and radiometric changes. Many traditional hand-crafted methods have been proposed for optical-SAR image matching, such as SIFT [21], SAR-SIFT [22], and HOPC [23, 24]. However, considering highly divergence between SAR and optical images and the computing ability, the hand-crafted feature-based matching approaches are quite limited to get a further step. Because of the powerful feature extraction and representation learning ability, exploiting convolution neural networks to extract the deep features has achieved a high matching accuracy. The mainstream architecture for optical-SAR image matching is Siamese network [25–28], which is composed of two identical convolutional streams. The dual network is used to extract deep characteristic information of input image pairs; therefore, the deep features in the same space can be measured under the same metric. However, the Siamese network can be only applied to the optical-SAR image matching problem; yet, no subsequent optical-SAR images joint interpretation work.

Based on the analysis above, we innovatively propose a bridge neural network- (BNN-) based optical-SAR image joint intelligent interpretation framework, which utilizes BNN to enhance the general feature embedding of optical and SAR images to improve the accuracy and application scenarios of specific optical-SAR images joint intelligent interpretation tasks. Completely different from the Siamese network, BNN contains two independent feature extraction networks and projects the optical and SAR images into a subspace to learn the desired common representation where features can be measured with Euclidean distance. The proposed framework is shown in Figure 1. BNN is trained on an optical-SAR image matching dataset to learn the common representation from optical and SAR images so that the BNN model can be transferred to the feature extraction

module for fine-tuning the interpretation model on optical-SAR/SAR/optical image interpretation datasets.

Further, to verify the effectiveness and the superiority of our proposed framework and promote the development of research in optical-SAR image fusion based on deep learning, it is very important to obtain datasets of a large number of perfectly aligned optical-SAR images. Besides, considering the existing optical-SAR image matching dataset either lacks scene diversity due to the huge difficulty in pixel-level matching between optical and SAR images [29], or has a low resolution limited by the remote sensing satellites [14], or covers only a single area [30], which cannot fully exploit the relevance of optical and SAR images, we publish the QXS-SAROPT dataset, which contains 20,000 optical-SAR patch pairs from multiple scenes of a high resolution. Specifically, the SAR images are collected from the Gaofen-3 satellite [31], and the corresponding optical images are from the Google Earth [32]. These images spread across landmasses of San Diego, Shanghai, and Qingdao. The QXS-SAROPT dataset under open access license CCBY is publicly available at <https://github.com/yaoxu008/QXS-SAROPT>.

On this basis, we conduct experiments on the optical-to-SAR crossmodal object detection to demonstrate the effectiveness and superiority of our framework. In particular, based on the QXS-SAROPT dataset, our framework can achieve up to 96% high accuracy on four benchmark SAR ship detection datasets.

The contributions of this paper can be summarized as follows:

- (i) We propose a BNN-based optical-SAR image joint intelligent interpretation framework, which can effectively improve the generic feature extraction capability of optical and SAR images, and thus improving the accuracy and the application scenarios of specific intelligent interpretation tasks for optical-SAR/SAR/optical images
- (ii) We publish the optical-SAR matching dataset: QXS-SAROPT, which contains 20,000 optical-SAR image pairs from multiple scenes of a high resolution of 1 meter to support the joint interpretation of optical and SAR images
- (iii) The BNN-based optical-SAR image joint intelligent interpretation framework is applied to SAR ship detection and achieves high accuracy on four SAR ship detection benchmark datasets

FIGURE 2: Illustration of BNN structure on the SAR-optical matching task. BNN contains two feature extraction networks: optical CNN and SAR CNN, to project the images into a common representation space. The Euclidean distance of the representations of SAR and optical images is regressed to 0 for positive samples and to 1 for negative samples, respectively.

2. Methodology

In this section, the details of the bridge neural network (BNN) and the proposed BNN-based joint interpretation framework are introduced.

2.1. Bridge Neural Network. The bridge neural network (BNN) proposed in [33] is adopted to learn the common representations of optical and SAR images on the optical-SAR image matching tasks, as shown in Figure 2.

Given a SAR-optical image matching dataset $\{X_s, X_o\}$, where $X_s = \{x_s^i\}_{i=1}^N$ is the set of SAR images and $X_o = \{x_o^i\}_{i=1}^N$ is the set of corresponding optical images, here, we consider samples from $S_p = \{x_s^i, x_o^i\}$, whose image pairs from the same region, as positive samples and samples from $S_n = \{x_s^i, x_o^j\} (i \neq j)$ as negative samples. Different from Siamese network, BNN contains two separate feature extraction networks: SAR network $f_s(x_s; \theta_s)$ and optical network $f_o(x_o; \theta_o)$ with parameters (θ_s, θ_o) , respectively, extracting features from SAR and optical images (x_s, x_o) . To decrease the feature dimension, following the feature extraction backbone, we employ a 1×1 convolution layer and a 4×4 average-pooling layer to the feature map. Finally, a linear layer with sigmoid activation function is followed to project the feature map into the n -dimension common feature representations: $z_s = f_s(x_s, \theta_s)$, $z_o = f_o(x_o, \theta_o)$. Then, the BNN outputs the Euclidean distance between z_s and z_o to measure the relevance of the input SAR-optical image pairs, which is described as

$$f(x_s, x_o; \theta_s, \theta_o) = \frac{1}{\sqrt{n}} \|(f_s(x_s; \theta_s) - f_o(x_o; \theta_o))\|_2, \quad (1)$$

where n is the dimension of z_s, z_o . The Euclidean distance indicates whether the input data pairs $\{x_s, x_o\}$ have a potential relation. And the closer distance between them, the more relevant they are. Specifically, the distance between positive samples tends to 0 while the distance between negative samples is close to 1. Therefore, the loss on positive samples and negative samples is set as follows:

$$l_p(S_p; \theta_s, \theta_o) = \frac{1}{|S_p|} \sum_{(x_s, x_o) \in S_p} (f(x_s, x_o; \theta_s, \theta_o) - 0)^2, \quad (2)$$

$$l_n(S_n; \theta_s, \theta_o) = \frac{1}{|S_n|} \sum_{(x_s, x_o) \in S_n} (f(x_s, x_o; \theta_s, \theta_o) - 1)^2. \quad (3)$$

Hence, the problem of learning the common representations of SAR-optical images is taken as a binary classification problem. The overall loss of BNN can be written as

$$l_{BNN}(S_p, S_n; \theta_s, \theta_o) = \frac{l_p(S_p; \theta_s, \theta_o) + \alpha \cdot l_n(S_n; \theta_s, \theta_o)}{1 + \alpha}, \quad (4)$$

where α is a hyperparameter to balance the weights of positive loss and negative loss. Then, the best weights (θ_s^*, θ_o^*) can be obtained via a optimization issue:

$$(\theta_s^*, \theta_o^*) = \arg \min_{\theta_s, \theta_o} l_{BNN}(S_p, S_n; \theta_s, \theta_o). \quad (5)$$

2.2. Optical-SAR Image Joint Intelligent Interpretation. Since BNN projects optical-SAR image patches into a common feature subspace, the model can well mine the correlation between optical and SAR images and thus improving the feature learning ability of optical-SAR images. Based on BNN, we propose the optical-SAR image joint interpretation framework, which can jointly utilize optical and SAR features enhanced by BNN and improve the performance of specific interpretation tasks of optical-SAR/optical images. As depicted in Figure 3, we present two different usage scenarios of the proposed BNN-based optical-SAR image joint intelligent interpretation framework.

As shown in Figure 3(a), for applications of optical and SAR fusion intelligent interpretation, such as object detection, classification, and segmentation, our framework can learn the common representations from SAR and optical images and enhance their feature learning ability for better interpretation performance under all-weather in full-time. As for applications of crossmodal intelligent interpretation, see Figure 3(b), our framework first utilizes BNN to project optical and SAR images into a common feature space and mines their complementary and correlated information through optical-SAR image matching. Then, the feature

FIGURE 3: Two usage scenarios of the proposed BNN-based optical-SAR image joint intelligent interpretation framework. (a) Example illustration of the proposed framework applied to optical and SAR fusion intelligent interpretation. (b) Example illustration of the proposed framework applied to optical to SAR cross-modal intelligent interpretation.

FIGURE 4: Coverage of the SAR images of Gaofen-3 for constructing the QXS-SAROPT dataset. (a) San Diego. (b) Shanghai. (c) Qingdao.

extraction module of BNN for the image modality to be interpreted is used as the pretrained model for feature embedding of the specific crossmodal intelligent interpretation task. In this way, plentiful complementary features transferred from images of the other modality during the learning of common feature space can be used to enhance the feature embeddings to be interpreted. Benefited from the enhanced feature embeddings, our framework can effectively improve the interpretation performance of crossmodal SAR/optical images.

Take SAR ship detection as an example. SAR ship detection in complex scenes is a great challenging task. And CNN-based SAR ship detection methods have drawn considerable attention because of the powerful feature embedding ability. Due to the scarce labeled SAR images, the pretraining technique is adopted to support these CNN-based SAR ship detectors. As SAR completely different from optical images, directly leveraging ImageNet [34] pretraining is hardly to obtain a good ship detector. However, our proposed framework can transfer rich texture features from optical images to SAR images to obtain a specific feature extraction model with better SAR feature embedding capabilities. Specifically, our proposed framework resorts to a SAR feature embedding operator from common representation learning based on the optical-SAR image matching task using BNN.

3. QXS-SAROPT Dataset

To fully exploit the relevance of optical and SAR images and verify the effectiveness of our proposed framework, a large perfectly aligned optical-SAR image dataset with diverse scenes of a high resolution is in need. Considering the existing optical-SAR image matching dataset either lack scene diversity or has a low resolution, we have published the QXS-SAROPT dataset, which contains 20,000 pairs of SAR and optical image patches with a high resolution of 1m extracted from multiple Gaofen-3 and Google Earth [32] scenes. As far as we know, QXS-SAROPT is the first dataset to provide high-resolution $1m \times 1m$ coregistered SAR and optical satellite image patches covering over three big port cities in the world: San Diego, Shanghai, and Qingdao. The coverage of these images is shown in Figure 4. Algorithm 1 shows the procedure for the QXS-SAROPT dataset construction. Finally, 20,000 high-quality image patch pairs are preserved in our dataset, some of which are shown in Figure 5 for examples.

4. Experiments

To verify the effectiveness and superiority of our BNN-based optical-SAR image joint intelligent interpretation framework,

1. Select three SAR images acquired by the Gaofen-3 satellite [31], which contain rich land cover types such as inland, offshore, and mountains. The spatial resolution of SAR imagery is $1\text{m} \times 1\text{m}$ for each pixel
2. Download the optical images of the corresponding area from Google Earth [32] with a spatial resolution of $1\text{m} \times 1\text{m}$
3. Cut the whole optical-SAR image pair into several subregion image pairs according to the complexity of land coverage. After that, we can register the subregion image pairs separately instead of directly registering the whole image pair
4. Locate matching points of the subregion optical-SAR image pairs manually, which are selected as the geometrically invariant corner points of buildings, ships, roads, etc
5. Use an existing automatic image registration software to register the subregion optical-SAR image pairs. Optical imagery is registered to the fixed SAR image through the bilinear interpolation method
6. Crop the registered subregion optical-SAR image pairs into small patches of 256×256 pixels with 20% overlap between adjacent patches
7. Double checked all image patches manually to ensure that every image contains meaningful information and texture. Remove indistinguishable or flawed images, such as images with similar scenes, texture-less sea, or visible mosaicking seamlines

ALGORITHM 1: Process of constructing QXS-SAROPT dataset.

FIGURE 5: Some exemplary patch-pairs from the QXS-SAROPT dataset. Top part: Google Earth RGB image patches, bottom part: Gaofen-3 SAR image patches.

the framework is applied to one typical optical-to-SAR cross-modal object detection task.

We conduct optical-to-SAR crossmodal object detection tasks on four benchmark SAR ship detection datasets. We select two representative CNN-based ship detection methods: faster R-CNN [35] and YOLOv3 [36], as the benchmarks in our work. We first utilize BNN to pretrain the feature extraction module for the selected SAR ship detectors, namely, the ResNet50 [8] backbone for faster R-CNN [35] and the Darknet53 [36] backbone for YOLOv3 [36], based on QXS-SAROPT with 14,000 image pairs as the training set and the remaining 6,000 image pairs as the testing set. Then, the pretrained model with better SAR feature embedding capabilities by common representation learning from optical and SAR images is used for fine-tuning the corresponding SAR ship detector.

4.1. Dataset. Four benchmark SAR ship detection datasets are tested: AIR-SARShip-1.0 [37], AIR-SARShip-2.0 [38], HRSID [39], and SSDD [40].

The AIR-SARShip-1.0 dataset consists of 31 high-resolution large-scale 3000×3000 images from the Gaofen-

TABLE 1: Results of optical-SAR image matching using BNN [33] on QXS-SAROPT.

Backbone	Accuracy	Precision	Recall
ResNet50 [8]	0.829	0.748	0.993
Darknet53 [36]	0.828	0.746	0.995

3 satellite. 21 images are randomly selected as training and validation data, and the remaining 10 images are used for testing. The AIR-SARShip-2.0 dataset includes 300 images of size 1000×1000 with resolutions ranging from 1 m to 5 m from the Gaofen-3 satellite. 210 images are randomly selected as training and validation data, and the remaining 90 images are used for testing. Images in AIR-SARShip-1.0 and AIR-SARShip-2.0 datasets are cropped into 512×512 pixels with a 0.5 overlap. The HRSID dataset contains 5604 SAR images of size 800×800 and is divided into training and testing set at a ratio of 13 : 7. The resolutions of images in the SSDD dataset range from 1 m to 15 m, and 1160 images are divided into 928 images for training and 232 images for testing.

For faster R-CNN [35], we directly input each image of AIR-SARShip-1.0, AIR-SARShip-2.0, and HRSID into the

FIGURE 6: Image pairs with different matching results by BNN [33] for a test subset of QXS-SAROPT. The image pairs correctly predicted as matched and unmatched pairs are considered as the true positive and true negative, respectively. False positive and false negative are, respectively, unmatched and matched images pairs that are wrongly predicted.

network and resize images in SSD to 1000×600 pixels. As for YOLOv3 [36], all the images are resized to 416×416 pixels. Because of the necessity of the multiscale training strategy for YOLOv3, no data augmentation except for scaling is applied.

4.2. Parameter Settings

4.2.1. Optical-SAR Image Matching. The BNN model with ResNet50 [8] and Darknet53 [36] as backbone is both trained in SGD for 200 epochs with a batch size of 20. The initial learning rate is set as 0.01 and then divided by a factor of 2 at the 30th and 100th epochs. The SAR and optical images are encoded into a 50-dimensional feature representation subspace. The ratio of positive and negative samples is set as 1 : 1 and the adjusting factor $\alpha = 1$.

4.2.2. SAR Ship Detectors. For the faster R-CNN [35] benchmark, all models are trained with SGD for 14 epochs with 0.0001 weight decay and 0.9 momentum, and the batch size is set to 8. The initial learning rate is 0.02 and is then divided by 10 at the 8th and 12th epochs. For the YOLOv3 [36] benchmark, all models are trained with SGD for 240 epochs with 12 images per minibatch. The initial learning rate is set as 0.001 and is then divided by 10 at the 160th and 200th epochs. The IoU threshold is set as 0.5 when training and testing for rigorous filtering of the bounding boxes with low precision. Warm-up [8] is used at the first 500 iterations during the training stage to avoid gradient explosion. The same settings are applied for all experiments for a fair comparison.

4.3. Results

4.3.1. Optical-SAR Image Matching. Table 1 shows the results of optical-SAR image matching using BNN [33] on the QXS-SAROPT dataset, which suggests that BNN achieves an outstanding performance on both ResNet50 [8] and Darknet53 [36] backbone. Specifically, the matching accuracy based on ResNet50 and Darknet53 reach up to 82.9% and 82.8%, respectively, demonstrating that BNN can learn useful common representations and well predict the relationship between SAR and optical images. Image pairs with different matching results by BNN are shown in Figure 6.

TABLE 2: Results of optical-SAR image matching using BNN [33] on different sizes of training sets on ResNet50.

Training set size	Accuracy	Precision	Recall
4000 (25%)	0.731	0.656	0.972
8000 (50%)	0.793	0.714	0.977
16000 (100%)	0.829	0.748	0.993

To explore the relationship between the training set size and matching results, we randomly select 4000 and 8000 optical-SAR image pairs as the training sets to train BNN on ResNet50. Table 2 shows the results of three sizes of training sets, which indicates that BNN can learn a good common representation even with a small number of training image pairs, and more training data can lead to better matching results. Besides, we show the accuracy, precision, and recall curves of BNN with 8000 image pairs as the training set in Figure 7, which show the convergence process of BNN.

4.3.2. SAR Ship Detection. Table 3 shows the average precision (AP) of SAR ship detection results on four ship detection datasets using ImageNet pretraining-based SAR ship detector (ImageNet-SSD) and our BNN-based SAR ship detector (BNN-based-SSD) pretrained on QXS-SAROPT. As shown in Table 3, compared with ImageNet-SSD, the AP of detection results is generally improved by BNN-based-SSD. Especially on SAR ship detection dataset AIR-SARShip-1.0 [37], 1.32% and 1.24% performance improvement can be achieved using two-stage detection benchmark: faster R-CNN [35] and one-stage detection benchmark: YOLOv3 [36], respectively. Average precision of ImageNet-SSD and BNN-based-SSD during the training on the test set of HRSID and AIR-SARShip-2.0 with two detectors are displayed in Figure 8. Taking YOLOv3 in AIR-SARShip-2.0 dataset as an example, BNN-based-SSD achieves higher average precision than ImageNet-SSD during the whole training process, indicating the significant improvement of our BNN-based-SSD. Similar phenomena are also presented on other datasets for both benchmarks, demonstrating the superiority of our BNN-based optical-SAR images joint intelligent interpretation framework. All these improved performances prove that our framework can well enhance

FIGURE 7: Training process of optical-SAR image matching using BNN on 8000 image pairs in QXS-SAROPT dataset. (a) Accuracy. (b) Precision. (c) Recall.

TABLE 3: Average precision of SAR ship detection results on four ship detection datasets using ImageNet-SSD pretrained on ImageNet and BNN-based-SSD pretrained on QXS-SAROPT.

(a)				
	AIR-SARShip-1.0 [37]		HRSID [39]	
	Faster R-CNN [35]	YOLOv3 [36]	Faster R-CNN [35]	YOLOv3 [36]
ImageNet-SSD	0.8720	0.8712	0.8878	0.8364
BNN-based-SSD	0.8852	0.8836	0.8932	0.8454
(b)				
	AIR-SARShip-2.0 [38]		SSDD [40]	
	Faster R-CNN [35]	YOLOv3 [36]	Faster R-CNN [35]	YOLOv3 [36]
ImageNet-SSD	0.8487	0.8300	0.9624	0.9447
BNN-based-SSD	0.8582	0.8470	0.9679	0.9468

the feature extraction capability of SAR ship detectors by common representation learning utilizing BNN and thus boosting ship detection in SAR images even with no additional annotation information of ships. To qualitatively compare these two methods, we visualize some detection results in Figure 9, which shows that our BNN-based-SSD clearly outperforms ImageNet-SSD and significantly reduces the missed detections and false alarms.

5. Conclusion and Future Work

In this paper, we propose a bridge neural network- (BNN-) based optical-SAR image joint intelligent interpretation framework, which can effectively improve the generic feature extraction capability of optical and SAR images by mining their feature correlation through matching tasks with BNN, and then improve the accuracy and application scenarios of specific optical-SAR image joint intelligent interpretation tasks. In order to fully exploit the correlation between optical and SAR images and ensure the great representation learning ability of BNN, we publish the QXS-SAROPT dataset containing 20,000 optical-SAR patch pairs from multiple scenes of a high resolution of 1 meter. Experimental results on the optical-to-SAR crossmodal object detection task demonstrate the effectiveness and superiority

of our framework. It is noted that based on the QXS-SAROPT dataset, our framework can achieve up to 96% high accuracy in SAR ship detection.

This research is in its early stage. In the future, we will consider exploring the performance of the proposed framework on optical-SAR fusion intelligent interpretation tasks, such as classification of land use and land cover and building segmentation. To support the research in intelligent interpretation fusing optical-SAR data, we will add label annotations and positions for scenes/objects of interest to every patch pair of the QXS-SAROPT dataset. In addition, to further explore the potential value of the QXS-SAROPT dataset, we are going to release an improved version of the dataset in the future, which will cover more land areas with versatile scenes and different sized patch pairs suitable for various optical-SAR data fusion tasks.

At a more macroscopic level, there are plentiful aspects that deserve deeper investigation. Currently, our approach to interpreting multimodal remote sensing images is verified by experiments on the ground. However, the onboard processing of remote sensing images will be a trend in the future.

Unfortunately, running deep learning models tends to be a high-power consumption process, let alone the tight constraints of onboard memory and computing resources. In

FIGURE 8: Average precision of ImageNet-SSD and BNN-based-SSD during the training on test set of HRSID and AIR-SARShip-2.0 with faster R-CNN and YOLOv3.

FIGURE 9: Visualization of comparison results between ImageNet-SSD and BNN-based SSD. The red and green rectangles figure out the ground truth and correctly predicted boxes. The orange and yellow circles denote the false alarms and missing detection.

this case, deep learning model compression is an effective and necessary technique to achieve onboard processing in our future work. The purpose of model compression is to achieve a model with fewer parameters, calculation amount, and less RAM to run without significantly diminished accuracy. Popular model compression methods include pruning [41], quantization [42], low-rank approximation and sparsity [43], and knowledge distillation [44, 45].

Furthermore, the formation of SAR images from echoes is the first inevitable step of SAR data processing nowadays, based on the algorithms such as back projection [46], compressed sensing [47], or signal processing. Therefore, the SAR application pipeline consists of multiple operations and varieties of complex calculations. In our future work, we will attempt to develop a deep learning framework that performs an integrating SAR processing workflow end to end, from the reflected echoes to the interpretation results. This will help to reduce the complexity of the onboard processor and further improve the processing efficiency.

Data Availability

The QXS-SAROPT dataset released by this work is publicly available at <https://github.com/yaoxu008/QXS-SAROPT> under open access license CCBY. Four SAR ship detection datasets used in this paper are publicly available. AIR-SARShip-1.0 and AIR-SARShip-2.0 can be accessed at <http://radars.ie.ac.cn/web/data/getData?dataType=SARDataset>. HRSID is available at <https://github.com/chaozhong2010/HRSID>. SSDD can be downloaded at <https://github.com/CAESAR-Radi/SAR-Ship-Dataset>.

Conflicts of Interest

All authors declare no possible conflicts of interests.

Authors' Contributions

X. Xiang and M. Huang conceived the idea of this study and supervised the study. M. Huang, X. Yao, L. Qian, and W. Bao conducted the experiments. M. Huang and L. Qian performed data analysis. M. Huang, X. Yao, and L. Qian contributed to the writing of the manuscript. L. Qian, W. Shi, Y. Zhang, N. Wang, and X. Liu participated in the construction of the QXS-SAROPT dataset. Meiyu Huang, Yao Xu, and Lixin Qian contributed equally to this work.

Acknowledgments

This is supported by the Beijing Nova Program of Science and Technology under Grant Z191100001119129 and the National Natural Science Foundation of China 61702520.

References

- [1] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: a technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.
- [2] X. X. Zhu, D. Tuia, L. Mou et al., "Deep learning in remote sensing: a comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.
- [3] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community," *Journal of Applied Remote Sensing*, vol. 11, no. 4, 2017.
- [4] G. Tsagkatakis, A. Aidini, K. Fotiadou, M. Giannopoulos, A. Pentari, and P. Tsakalides, "Survey of deep-learning approaches for remote sensing observation enhancement," *Sensors*, vol. 19, no. 18, article 3929, 2019.
- [5] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep Convolutional Neural Networks for LvcSr," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8614–8618, Vancouver, BC, Canada, 2013.
- [6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, San Diego, CA, USA, 2015.
- [7] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.
- [9] M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: an ever-growing relationship," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 4, pp. 6–23, 2016.
- [10] Z. Zhang, G. Vosselman, M. Gerke, D. Tuia, and M. Y. Yang, "Change detection between multimodal remote sensing data using siamese cnn," 2018, <https://arxiv.org/abs/1807.09562>.
- [11] P. Feng, Y. Lin, J. Guan et al., "Embranchment cnn based local climate zone classification using sar and multispectral remote sensing data," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6344–6347, Yokohama, Japan, 2019.
- [12] Z. Zhang, G. Vosselman, M. Gerke, C. Persello, D. Tuia, and M. Y. Yang, "Detecting building changes between airborne laser scanning and photogrammetric data," *Remote sensing*, vol. 11, no. 20, article 2417, 2019.
- [13] M. Schmitt, F. Tupin, and X. X. Zhu, "Fusion of sar and optical remote sensing data—challenges and recent trends," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 5458–5461, Fort Worth, TX, USA, 2017.
- [14] M. Schmitt, L. Hughes, and X. Zhu, "The sen1-2 dataset for deep learning in sar-optical data FUSION," *Remote Sensing & Spatial Information Sciences*, vol. IV-1, no. 1, pp. 141–146, 2018.
- [15] Q. Feng, J. Yang, D. Zhu et al., "Integrating multitemporal sentinel-1/2 data for coastal land cover classification using a multibranch convolutional neural network: a case of the yellow river delta," *Remote Sensing*, vol. 11, no. 9, article 1006, 2019.
- [16] S. C. Kulkarni and P. P. Rege, "Pixel level fusion techniques for sar and optical images: a review," *Information Fusion*, vol. 59, pp. 13–29, 2020.
- [17] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Multimodal bilinear fusion network with Second-Order attention-based channel selection for land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1011–1026, 2020.

- [18] W. Burger and M. J. Burge, *Principles of Digital Image Processing: Core Algorithms*, 2010, Springer Science & Business Media.
- [19] J. Walters-Williams and Y. Li, "Estimation of mutual information: a survey," in *Rough Sets and Knowledge Technology. RSKT 2009, Lecture Notes in Computer Science*, P. Wen, Y. Li, L. Polkowski, Y. Yao, S. Tsumoto, and G. Wang, Eds., pp. 389–396, Springer, Berlin, Heidelberg, 2009.
- [20] S. Suri and P. Reinartz, "Mutual-information-based registration of terrasars-x and ikonos imagery in urban areas," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 2, pp. 939–949, 2010.
- [21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [22] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "Sar-sift: a sift-like algorithm for sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 453–466, 2015.
- [23] Y. Ye and L. Shen, "Hopc: a novel similarity metric based on geometric structural properties for multi-modal remote sensing image MATCHING," *Remote Sensing and Spatial Information Sciences*, vol. III-1, pp. 9–16, 2016.
- [24] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2941–2958, 2017.
- [25] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4353–4361, Boston, MA, USA, 2015.
- [26] N. Merkle, W. Luo, S. Auer, R. Müller, and R. Urtasun, "Exploiting deep matching and sar data for the geolocalization accuracy improvement of optical satellite images," *Remote Sensing*, vol. 9, no. 6, article 586, 2017.
- [27] L. Mou, M. Schmitt, Y. Wang, and X. X. Zhu, "A cnn for the identification of corresponding patches in sar and optical imagery of urban scenes," in *2017 Joint Urban Remote Sensing Event (JURSE)*, pp. 1–4, Dubai, United Arab Emirates, 2017.
- [28] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in sar and optical images with a pseudo-siamese cnn," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 784–788, 2018.
- [29] Y. Wang and X. X. Zhu, "The sarptical dataset for joint analysis of sar and optical image in dense urban area," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6840–6843, Valencia, Spain, 2018.
- [30] J. Shermeyer, D. Hogan, J. Brown et al., "Spacenet 6: Multi-sensor all weather mapping dataset," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 196–197, Seattle, WA, USA, 2020.
- [31] Q. Zhang, "System design and key technologies of the gf-3 satellite," *Acta Geodaetica et Cartographica Sinica*, vol. 46, no. 3, pp. 269–277, 2017.
- [32] <https://earth.google.com/>.
- [33] Y. Xu, X. Xiang, and M. Huang, "Task-driven common representation learning via bridge neural network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5573–5580, 2019.
- [34] O. Russakovsky, J. Deng, H. Su et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.
- [36] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, <https://arxiv.org/abs/1804.02767>.
- [37] S. Xian, W. Zhirui, S. Yuanrui, D. Wenhui, Z. Yue, and F. Kun, "Air-sarship-1.0: High resolution sar ship detection dataset," *Journal of Radars*, vol. 8, no. 6, pp. 852–862, 2019.
- [38] "2020 gaofen challenge on automated high-resolution earth observation image interpretation," 2020, <http://en.sw.chreos.org>.
- [39] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.
- [40] J. Li, C. Qu, and J. Shao, "Ship Detection in Sar Images Based on an Improved faster r-Cnn," in *2017 SAR in Big Data Era: Models, Methods and Applications (BIGSAR DATA)*, pp. 1–6, Beijing, China, 2017.
- [41] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural network," *Advances in Neural Information Processing Systems*, MIT Press, 2015.
- [42] S. Han, H. Mao, and W. J. Dally, "Deep compression: compressing deep neural networks with pruning, trained quantization and human coding," in *Proceedings of International Conference on Learning Representations*, San Juan, Puerto Rico, 2016.
- [43] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Speeding up Convolutional Neural Networks with Low Rank Expansions," in *Proceedings of the British Machine Vision Conference*, University of Nottingham, UK, 2014.
- [44] C. Bucilua, R. Caruana, and A. Niculescumizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 535–541, 2006.
- [45] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Advances in Neural Information Processing Systems*, MIT Press, 2014.
- [46] L. A. Gorham and L. J. Moore, "Sar image formation toolbox for matlab, in Algorithms for Synthetic Aperture Radar Imagery XVII," *International Society for Optics and Photonics*, vol. 7699, pp. 769–906, 2010.
- [47] R. Baraniuk and P. Steeghs, "Compressive radar imaging," in *2007 IEEE Radar Conference*, pp. 128–133, Waltham, MA, USA, 2007.