remote sensing

MDPI

Communication

Self-Supervised Pre-Training with Bridge Neural Network for SAR-Optical Matching

Lixin Qian [1,2,†], Xiaochun Liu [1], Meiyu Huang [2,†] and Xueshuang Xiang [2,*]

[1] School of Mathematics and Statistics, Wuhan University, Wuchang District, Wuhan 430072, China; qianlixin@whu.edu.cn (L.Q.); xcliu@whu.edu.cn (X.L.)
[2] Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Haidian District, Beijing 100086, China; huangmeiyu@qxslab.cn
* Correspondence: xiangxueshuang@qxslab.cn
† These authors contributed equally to this work.

Abstract: Due to the vast geometric and radiometric differences between SAR and optical images, SAR-optical image matching remains an intractable challenge. Despite the fact that the deep learning-based matching model has achieved great success, SAR feature embedding ability is not fully explored yet because of the lack of well-designed pre-training techniques. In this paper, we propose to employ the self-supervised learning method in the SAR-optical matching framework, in order to serve as a pre-training strategy for improving the representation learning ability of SAR images as well as optical images. We first use a state-of-the-art self-supervised learning method, Momentum Contrast (MoCo), to pre-train an optical feature encoder and an SAR feature encoder separately. Then, the pre-trained encoders are transferred to an advanced common representation learning model, Bridge Neural Network (BNN), to project the SAR and optical images into a more distinguishable common feature representation subspace, which leads to a high multi-modal image matching result. Experimental results on three SAR-optical matching benchmark datasets show that our proposed MoCo pre-training method achieves a high matching accuracy up to 0.873 even for the complex QXS-SAROPT SAR-optical matching dataset. BNN pre-trained with MoCo outperforms BNN with the most commonly used ImageNet pre-training , and achieves at most 4.4% gains in matching accuracy.

Keywords: SAR-optical fusion; image matching; self-supervised learning; representation learning

Citation: Qian, L.; Liu, X.; Huang, M.; Xiang, X. Self-Supervised Pre-Training with Bridge Neural Network for SAR-Optical Matching. *Remote Sens.* **2022**, *14*, 2749. https://doi.org/10.3390/rs14122749

Academic Editors: Yue Wu, Kai Qin, Qiguang Miao and Maoguo Gong

Received: 1 April 2022
Accepted: 1 June 2022
Published: 8 June 2022

1. Introduction

Synthetic Aperture Radar (SAR) and optical imagery are two of the most commonly used modalities in remote sensing since they provide highly complementary content to each other. While optical imagery with good interpretability is easily affected by atmospheric conditions, SAR data can collect information all the time but suffered from serious intrinsic speckle noise. Therefore, fusion information of SAR and optical images can give rise to a better interpretation of the imaged area. For accurate SAR-optical data fusion, identifying corresponding image patches plays a crucial role as a pre-procedure. It remains a widely unsolved challenge to match SAR-optical remote sensing data due to the vast geometric and radiometric differences as shown in Figure 1.

Over the past few decades, the traditional SAR-optical image matching methods can be generally divided into area-based and feature-based approaches. Area-based methods utilize the intensity of pixel values in some regions of the image and the corresponding regional similarity evaluation is calculated, such as cross correlation (CC) [1], structural similarity (SSIM), mutual information (MI) [2,3], and so on. However, owing to the low flexibility and lack of local structure information, the area-based methods fail to avoid information loss in the measure of multimodal image similarity. Therefore, more attempts at SAR and optical image matching have been placed on the feature-based methods. Since feature-based methods rely on the invariant feature points and handcrafted descriptors can

handle the geometric changes well, the feature-based methods generally outperform the area-based methods, for example, scale-invariant feature transform (SIFT) [4], SAR-SIFT [5], HOPC [6,7], etc. However, the handcrafted descriptors based on low-level semantic features are not capable of dealing with highly divergent changes between SAR and optical images. More recently, deep learning-based SAR-optical image matching approaches have achieved great success. Two-tower architecture is the most commonly exploited in multimodal image matching, such as a Siamese or pseudo-Siamese network [8–11], which consists of two convolutional neural networks (CNN) to extract the deep characteristic features—not only with the Siamese network but a novel method called Bridge Neural Network (BNN) [12] to project the multimodal data into a common representation subspace where features can be measured with Euclidean distance.

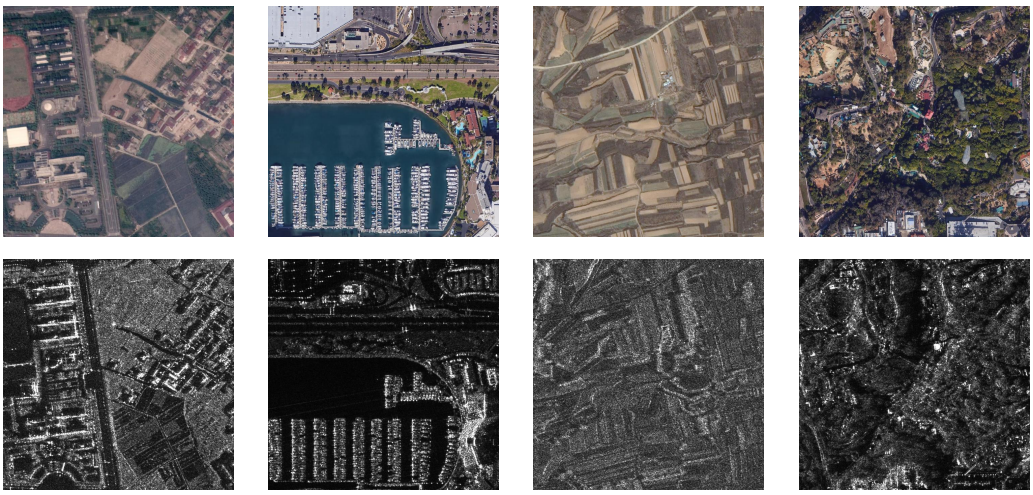


Figure 1. Comparison between SAR (**bottom**) and optical (**top**) imagery of the same scene.

Meanwhile, the ImageNet [13,14] supervised pre-training technique that contains prior knowledge is the most widely adopted in the SAR-related field for the scarcity of labeled SAR images in the past. However, there remain some limitations to using ImageNet pre-trained models for SAR-optical fusion tasks. Images in ImageNet are all optical images, which do not contain SAR information and characteristics; as a result, taking ImageNet supervised pre-training directly can hardly improve the learning ability of SAR images and benefit the SAR-optical fusion task.

As ImageNet pre-training can hardly improve the learning ability for SAR images, the representation learning ability to match models plays a vital role in SAR-optical matching tasks. Recent advances in self-supervised learning for computer vision present competitive results with supervised learning. Without manual annotations, useful representations can be obtained only with the help of some pretext tasks, which is probably achieved by maximizing the mutual information of learned representations. Representations pre-trained by contrastive learning that transferred to downstream tasks: classification, segmentation, and detection tasks lead to a competitive performance with supervised learning [15–21]. A contrastive learning method called Momentum Contrast (MoCo) [19] uses a momentum-update encoder to generate a dynamic dictionary query to save more negative samples with less memory. The MoCo pre-training method has achieved promising results in a variety of downstream tasks.

As discussed above, self-supervised learning can pre-train representations that can be transferred to downstream tasks by fine-tuning. In an attempt to improve multimodal SAR-optical image matching performance, we exploit the self-supervised learning technique to improve the feature learning ability of SAR and optical imagery, respectively. Then, the model is transferred to the SAR-optical matching task. The overall process is illustrated in Figure 2. More specifically, we take MoCo [19] as a pre-training strategy and BNN [12] for matching. By this method, self-supervised pre-training enhances the embedding of SAR

and optical images and benefits SAR-optical matching tasks. Experimental results on three datasets show that self-supervised pre-training leads to a better matching performance. Our main contributions are summarized as follows:

- We propose a framework applying self-supervised learning to SAR-optical image matching, improving the feature learning ability of SAR and optical images.
- We take MoCo and BNN as one of the most representative works in self-supervised learning and multi-modal image matching to make the framework truly implemented.
- For the proposed framework, we conduct lots of experiments to confirm the feasibility and the effectiveness of the self-supervised learning transferred to optical-SAR image matching task, which would encourage further research in this field.

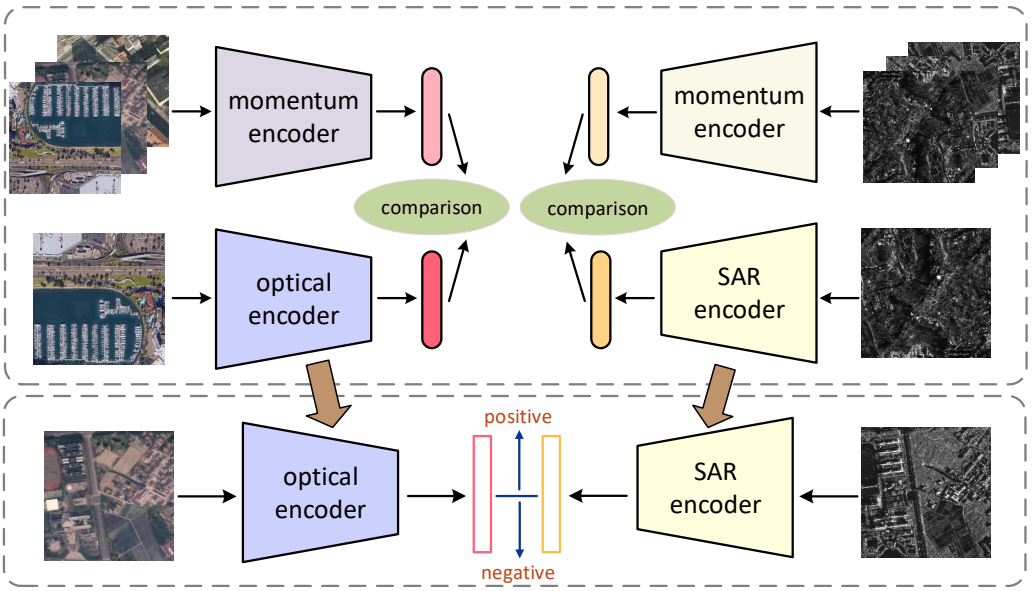


Figure 2. The overall process of models via MoCo pre-training transferred to BNN is described. The first box shows the process of applying MoCo to improve the representation learning ability of SAR and optical images separately. Then, the model is transferred to BNN to deal with the SAR-optical matching task in the second box.

The rest of the paper is organized as follows: Section 2 introduces MoCo, BNN, and the way we combine them in detail. The settings and results of the experiments are shown in Section 3. The discussion and conclusions are drawn in Section 4.

2. Method

The method in this paper consists of two steps: MoCo [19] pre-training and transferring to BNN [12] for matching tasks. In this section, we will introduce the MoCo, BNN, and the way we combine them in detail.

2.1. Momentum Contrast

In [19], they regard contrastive learning as training an encoder for a dictionary look-up task. Momentum Contrast (MoCo) is for building a dynamic large and consistent dictionary on-the-fly. The core of MoCo is maintaining the dictionary as a queue of data samples; therefore, the dictionary size can be decoupled from the mini-batch size for the utilization of the queue to be larger and contain more negative samples.

Given a dataset $X = \{x_i\}_{i=1}^{N}$. , where x_i can be considered as a query sample. Then, we can randomly select the other k samples from the dataset X to form a dictionary $\{d_1, d_2, \ldots, d_k\}$. There is a single sample d_i in the dictionary that matches x_i. Therefore, the sample x_i with the dictionary can be combined into one positive pair $\{x_i, d_i\}$ and k negative pairs $\{x_i, d_j\}(j \neq i)$. To extract the representations, two feature encoders $f(\cdot;\theta_x), f(\cdot;\theta_d)$

are employed with parameters θ_x, θ_d respectively to project the images into the query representations $z_i = f(x_i; \theta_x)$ and the keys of the dictionary $y_i = f(d_i; \theta_d)$. Regarding InfoNCE [16] as the contrastive loss:

$$\mathcal{L}_{contrast} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{s(x, d_+)}{\sum_{x \in X} \sum_{j=1}^{k+1} s(x, d_j)}\right], \tag{1}$$

where $s(\cdot, \cdot)$ is the metric function and d_+ is the unique positive sample. Here, dot product is applied to measure the similarity between latent representations with a temperature hyper-parameter τ:

$$s(x, d) = \exp\left(\frac{f(x; \theta_x) \cdot f(d; \theta_d)}{\|f(x; \theta_x)\| \cdot \|f(d; \theta_d)\|} \cdot \frac{1}{\tau}\right), \tag{2}$$

which achieves high values for positives and low scores for negatives.

Momentum update. The parameters θ_x of query encoder are updated by back-propagation while the parameters θ_d of key encoder are updated by:

$$\theta_d \leftarrow m\theta_d + (1 - m)\theta_x, \tag{3}$$

where $m \in [0, 1)$ is the momentum coefficient. The momentum update makes θ_d evolve more smoothly than θ_x. Consequently, the keys in dictionary are encoded by slightly different encoders.

Compared with memory bank. The memory bank proposed in [15] is composed of the representations of samples. For every mini-batch, the keys are randomly sampled from the memory bank with no back-propagation. It can support a large mini-batch size, but the representations in the memory bank can not be consistent since it is only updated when it has been chosen. In contrast, the momentum update is more memory-efficient and guarantees the consistency of dictionary keys. Comparison is shown in Figure 3.

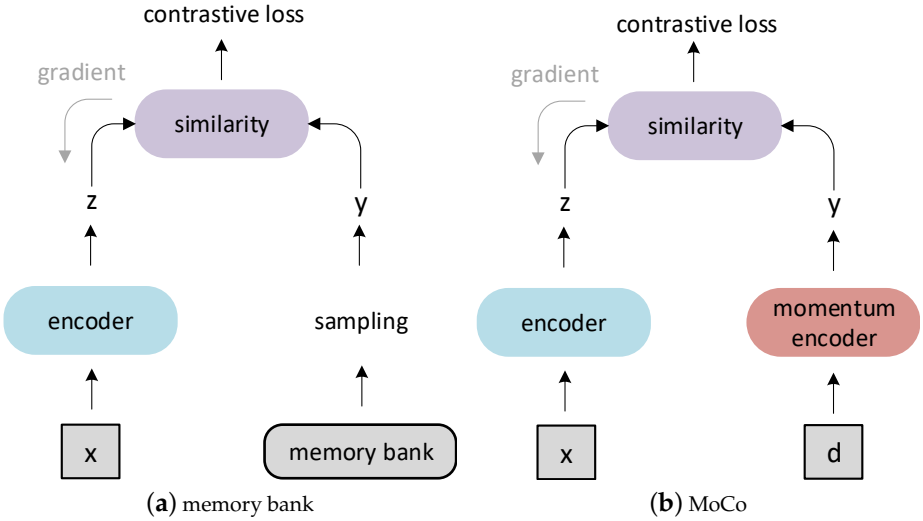


Figure 3. Comparison between memory bank [15] and MoCo [19]. (**a**) Forming a memory bank and sample from it as the key representations; (**b**) MoCo encodes the dictionary on-the-fly by momentum updating an encoder slowly.

Pretext Task. Pretext tasks act as an important strategy to learn representations of data. The pretext tasks in this paper follows [15,19]: random grayscale, random color jittering, and random horizontal flip. In addition, as depicted in [19,22], Batch Normalization (BN) [23] prevents the model from learning good representations. Thus, shuffling BN [19] is employed to solve this problem.

2.2. Bridge Neural Network

Despite the different modalities of SAR and optical images, BNN [12] works like a bridge and is capable of projecting the multimodal images into a common representation subspace. As depicted in Figure 4, given a SAR-optical image dataset $\{X_s, X_o\}$. $X_s = \{x_s^i\}_{i=1}^N, X_o = \{x_o^i\}_{i=1}^N$ are sets of SAR and optical image patches. We construct the positive sample set $S_p = \{x_s^i, x_o^i\}$, where corresponding image pairs are from the same region. Image pairs from $S_n = \{x_s^i, x_o^j\}(i \neq j)$ are from different areas, which we call negative samples. The dual networks respectively extract features from SAR and optical images: $f(\cdot)$ is for features of SAR images while $g(\cdot)$ for optical images. The separate networks reduce the images into the n-dimension latent vectors: $z_s = f(x_s), z_o = g(x_o)$. Then, Euclidean distance is designed to bring the latent representations of positive samples together while pushing negative samples apart in the common representation subspace. The Euclidean distance between z_s and z_o is defined as:

$$h(x_s, x_o) = \frac{1}{\sqrt{n}} \| (f(x_s) - g(x_o)) \|_2, \tag{4}$$

which indicates whether the input data pairs $\{x_s, x_o\}$ have a potential relationship. The distance of positive samples is regressed to 0 while the distance between negative samples converges to 1. Therefore, the regression loss on positive samples and negative samples are as follows:

$$l_p(S_p) = \frac{1}{|S_p|} \sum_{(x_s, x_o) \in S_p} (h(x_s, x_o) - 0)^2, \tag{5}$$

$$l_n(S_n) = \frac{1}{|S_n|} \sum_{(x_s, x_o) \in S_n} (h(x_s, x_o) - 1)^2. \tag{6}$$

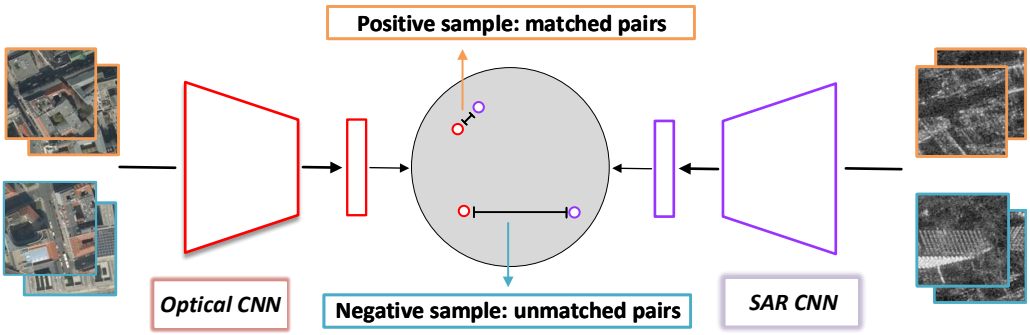


Figure 4. Schematic illustration of the BNN architecture for the SAR-optical image matching task. Optical network and SAR network are used in BNN to project the images from different modalities into a common subspace. The Euclidean distance of the representations of SAR and optical images is pulled close for positive samples and pulled apart for negative samples.

Hence, we add the loss on positive samples and negative samples up as the BNN loss:

$$l_{BNN}(S_p, S_n) = \frac{l_p(S_p) + \alpha \cdot l_n(S_n)}{1 + \alpha}, \tag{7}$$

where α balances the weights of positive loss and negative loss. Thus, optimizing the loss of BNN can lead to the best networks (f^*, g^*):

$$(f^*, g^*) = \text{argmin}_{f,g}\, l_{BNN}(S_p, S_n) \tag{8}$$

2.3. Transfer MoCo Pre-Trained Model to BNN

As depicted in Figure 2, at the pre-training stage, we first pre-train MoCo to obtain an optical encoder and an SAR encoder to separately learn feature representations and

measure the similarity in the latent space. At the matching stage, the pre-trained encoders transferred to BNN serve as the initialization of the optical network and SAR network for fine-tuning in the SAR-optical matching task. The encoders project the SAR and optical images into a common feature representation subspace, where it is easy to measure similarity, to determine whether they match or not.

3. Experiments

3.1. Dataset

We conduct experiments on three SAR-optical image datasets: SARptical [24], QXS-SAROPT [25], SEN1-2 [26]. **SARptical** is a dataset of over 10,000 pairs of corresponding SAR and optical image pairs, which are from TerraSAR-X and aerial UltraCAM optical images. The images are of 112 × 112 pixels with 100 m × 100 m ground coverage. We use 7577 image pairs for training and 1263 patch pairs for testing. **QXS-SAROPT** contains 20,000 pairs of SAR-optical image patches from GaoFen-3 satellite and Google Earth, covering a variety of land types. The images have a size of 256 × 256 pixels at a pixel spacing of 1 m × 1 m. The dataset is randomly divided into training and testing at a ratio of 7:3. We select the spring subset from four sub-groups in **SEN1-2** dataset, which consists of registered patch-pairs from Sentinel-2 and Sentinel-1—a total of 75,724 patch pairs of size 256 × 256 pixels with spatial distance of 10 m × 10 m, of which 52,799 are for training and 22,925 for testing.

3.2. Implementation

Considering the complexity and difficulty of the matching task, we only take Vgg11 [27] as the feature extraction network for SARptical while Vgg11 [27], ResNet50 [28], and Darknet53 [29] as the backbone for QXS-SAROPT and SEN1-2. More specifically, we encode the SAR and optical images into a 50-dimensional feature representation subspace. It is noted that every mini-batch input contains the same number of positive and negative samples to prevent the model from mode collapse for the unbalanced data distribution. All images are normalized w.r.t. mean and variance in preparation.

MoCo. We take $N - 1$ negative samples (N is the number of training datasets) and set temperature parameter $\tau = 0.07$ and momentum coefficient for updating encoder is set as $m = 0.999$. We use SGD as an optimizer and a mini-batch size of 20. The weight decay is 0.001 and the SGD momentum is 0.9. For SARptical, the learning rate is set as 0.05 for 300 epochs. While training on QXS-SAROPT, the learning rate is 0.001 for the first 250 epochs and 0.0005 for the last 250 epochs. As for SEN1-2, the models are trained with a learning rate of 0.05 for 70 epochs.

BNN. The ratio of positive and negative samples is 1:1 and adjusting factor $\alpha = 1$. For SARptical, the learning rate is set as 0.1 to fine-tune the model for 100 epochs. Meanwhile, BNN with MoCo pre-training on QXS-SAROPT and SEN1-2 are fine-tuned with a learning rate of 0.05 for 10 epochs.

3.3. Results Analysis

We fine-tune BNN to SAR-optical image matching task with MoCo pre-training (MoCo-BNN) on three datasets: SARptical, QXS-SAROPT, and SEN1-2. To demonstrate the superiority of our method, we compare our method with two other initialization methods: no pre-training (NP) and ImageNet pre-training (IP).. Accuracy, precision, and recall score are employed as evaluation metrics. The matching results on three datasets can be seen in Tables 1–3. It is noted that the pair matching results of IP-BNN benchmark can be directly obtained from article [25,30].

Table 1. Results for BNN patch-matching on SARptical with different pre-training methods.

Methods	Accuracy	Precision	Recall
NP-BNN	0.887	0.825	0.993
IP-BNN	**0.913**	0.855	**0.999**
MoCo-BNN(ours)	**0.913**	**0.859**	0.991

Table 2. Results for BNN patch-matching on QXS-SAROPT with different pre-training methods.

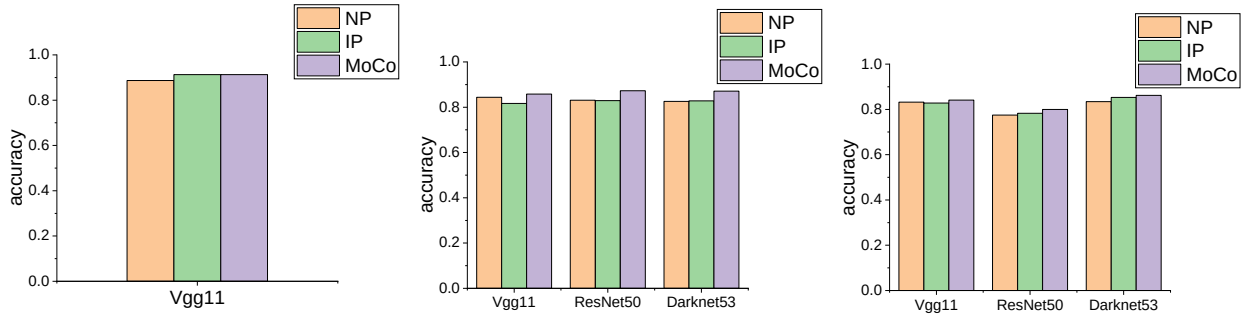
Backbone	Methods	Accuracy	Precision	Recall
Vgg11	NP-BNN	0.844	0.781	0.982
	IP-BNN	0.817	0.744	**0.999**
	MoCo-BNN(ours)	**0.858**	**0.795**	0.990
ResNet50	NP-BNN	0.831	0.750	0.990
	IP-BNN [25,30]	0.829	0.748	0.993
	MoCo-BNN(ours)	**0.873**	**0.808**	**0.995**
Darknet53	NP-BNN	0.826	0.761	0.980
	IP-BNN [25,30]	0.828	0.746	0.995
	MoCo-BNN(ours)	**0.871**	**0.809**	**0.997**

Table 3. Results for BNN patch-matching on SEN1-2 with different pre-training methods.

Backbone	Methods	Accuracy	Precision	Recall
Vgg11	NP-BNN	0.832	**0.800**	0.916
	IP-BNN	0.828	0.760	**0.993**
	MoCo-BNN(ours)	**0.841**	0.787	0.960
ResNet50	NP-BNN	0.775	0.721	0.931
	IP-BNN	0.783	0.722	0.954
	MoCo-BNN(ours)	**0.800**	**0.753**	**0.968**
Darknet53	NP-BNN	0.834	0.775	0.970
	IP-BNN	0.853	0.788	0.993
	MoCo-BNN(ours)	**0.862**	**0.796**	**0.998**

Accuracy performance. Tables 2 and 3 suggest that our MoCo pre-trained models lead to a better matching performance on QXS-SAROPT and SEN1-2.(The bolded number represents the highest score on the backbone.) Especially on QXS-SAROPT, the accuracy of BNN taking MoCo as a pre-training strategy achieves 87.3% and 87.1% and makes a 4.4% improvement, which surpasses the other two methods by large margins. MoCo pre-training also makes an obvious improvement on SEN1-2, indicating that MoCo has a powerful representation learning ability for both SAR and optical images. Furthermore, the results of NP and IP with ResNet50 and Darknet53 as backbone are almost the same, which illustrates that the ImageNet supervised pre-trained models have no capacity for SAR information and characteristics and it hardly makes sense in the SAR-optical image matching problem. It is worth noting that IP-BNN is even worse than NP-BNN with Vgg11 as a backbone on QXS-SAROPT and SEN1-2, which is in line with the claims made in [14], i.e., shallow models can be trained from scratch as long as a proper initialization is used, whereas only when the network is large enough can the ImageNet pre-training model which contains prior knowledge provide a good initialization to fine-tune. Besides, in Table 1, MoCo does not make any progress on SARptical. The reason may stem from the single building scenario in this dataset; the IP method learns better optical features and makes a great performance. When confronted with more complex SAR-optical datasets, optical feature embedding ability is weakened and SAR feature learning ability plays a

major role. We visualize the accuracy results in Figure 5 to better show the comparison of the different pre-training strategies.



(**a**) SARptical (**b**) QXS-SAROPT (**c**) SEN1-2

Figure 5. Visualization of the accuracy results of BNN with NP, IP, and MoCo pre-training on three datasets.

We show matching results of some images pairs in Figure 6. In particular, image pairs in red box are classified correctly by our MoCo-BNN and classified incorrectly by IP-BNN. As shown in the figure, MoCo-BNN not only can distinguish the similar negative image pairs, but also can correctly discriminate among the non-obvious positive sample pairs.

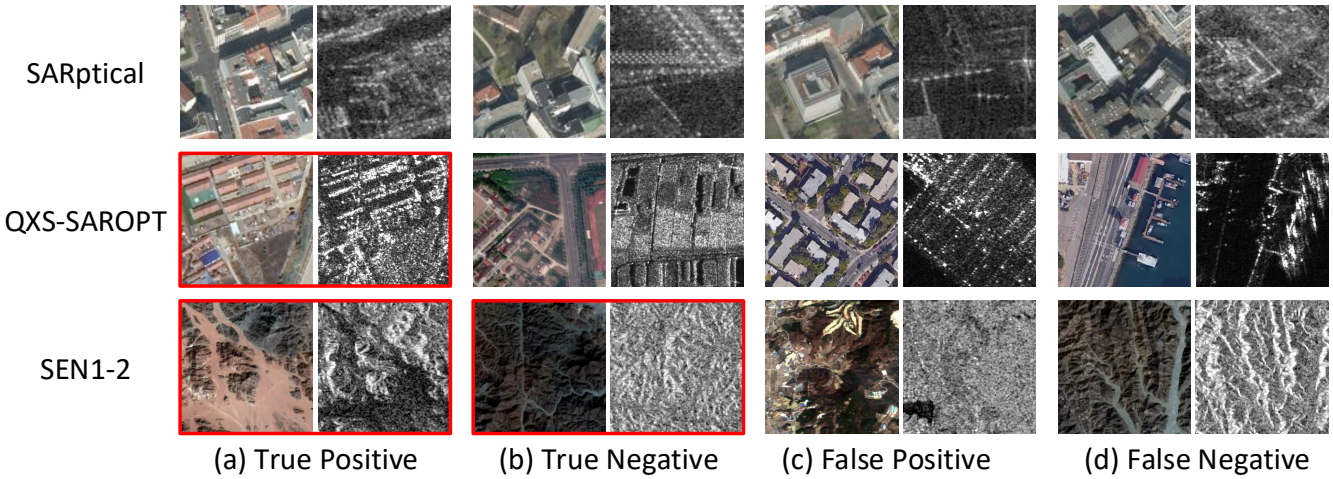
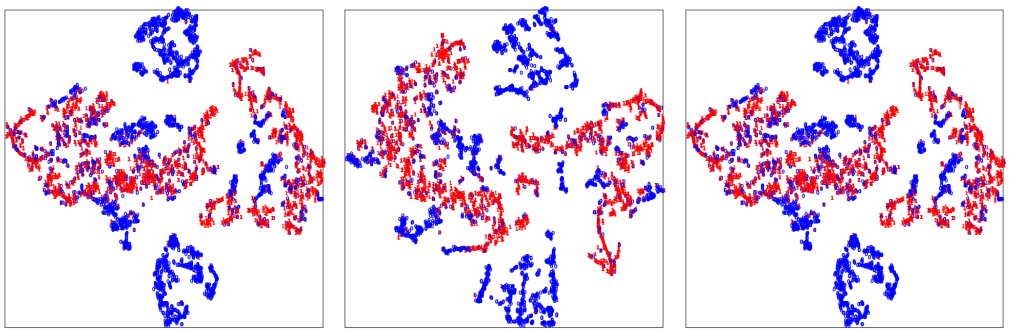


Figure 6. Exhibition of images pairs in different matching results. The red boxes frame the image pairs that MoCo-BNN classifies correctly and IP-BNN classifies incorrectly.

Embedding learning performance. To intuitively compare the feature representations learn by BNN with NP, IP, and MoCo pre-training, we visualize the embedding learning results of the test set of QXS-SAROPT. We use t-distributed Stochastic Neighbor Embedding(t-SNE) to visualize the features extracted by BNN with NP, IP, and MoCo pre-training. We first concatenate the SAR and optical 50-dimensional features to the 100-dimensional features. T-SNE projects the features to two dimensions so that the high-dimensional features are convenient to visualize. As shown in Figure 7, the features of positive samples and negative samples learned by BNN with NP and IP are mixed together while the positive features (class 1) and negative features (class 0) learned by BNN with MoCo pre-training are more gathered in each class and more separate between different classes. Therefore, the MoCo pre-trained BNN can generate a more distinguishable embedding space, leading to a better multi-modal image matching result. The reason may stem from MoCo pre-training leading to a better representation learning ability for SAR and optical images.



(**a**) NP (**b**) IP (**c**) MoCo

Figure 7. Visualization of positive and negative features extracted by BNN with NP, IP, and MoCo pre-training using t-SNE. The red dots represent the positive features, and the blue dots represent the negative features.

4. Discussion and Conclusions

Aiming at improving SAR-optical image matching performance, considering directly fine-tuning on an ImageNet supervised pre-training model as commonly used can hardly benefit for improving the learning ability for SAR images; this paper exploits a self-supervised pre-training to improve the feature learning ability of SAR and optical images respectively. Then, the pre-trained model is transferred to the SAR-optical matching tasks. The experiments demonstrate that self-supervised pre-training leads to a significant improvement.

Furthermore, we exploit a self-supervised pre-training paradigm to improve the feature learning ability of multi-modal images. However, this paper only did experiments on SAR-optical images and only for matching tasks. It is believed that our method not only can be adaptive to different kinds of remote sensing images, such as multi-spectral images, hyperspectral images and so on, but also can be transferred to different tasks, such as objective detection.

However, the experiments were only conducted on one self-supervised method MoCo, and one matching network BNN is a major deficiency. Furthermore, it has been verified in [15–21] that a large mini-batch size is necessary for self-supervised learning to learn a good representation. Nonetheless, it still works well when we train the MoCo with mini-batch size 20, which is much smaller than the commonly used mini-batch size in self-supervised learning. In the future, more experiments on a Siamese network and other self-supervised learning methods will be carried out to confirm the effectiveness of self-supervised pre-training in SAR-optical image matching.

Author Contributions: L.Q. wrote the manuscript and performed all the experiments; M.H. performed the experiment analysis and revised the manuscript. L.Q. and M.H. contributed equally to this work. M.H., X.L. and X.X. supervised the study. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by Beijing Nova Program of Science and Technology (Grant Number: Z191100001119129).

Data Availability Statement: Publicly available datasets were analyzed in this study. SARptical [24] can be downloaded at http://www.sipeo.bgu.tum.de/downloads/SARptical_data.zip. QXS-SAROPT [25] can be found at https://github.com/yaoxu008/QXS-SAROPT. SEN1-2 [26] is available at https://mediatum.ub.tum.de/1436631.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Burger, W.; Burge, M.J. *Principles of Digital Image Processing: Core Algorithms*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2010.
2. Walters-Williams, J.; Li, Y. Estimation of mutual information: A survey. In Proceedings of the International Conference on Rough Sets and Knowledge Technology, Gold Coast, Australia, 14–16 July 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 389–396.
3. Suri, S.; Reinartz, P. Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas. *IEEE Trans. Geosci. Remote Sens.* **2009**, *48*, 939–949. [CrossRef]
4. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
5. Dellinger, F.; Delon, J.; Gousseau, Y.; Michel, J.; Tupin, F. SAR-SIFT: A SIFT-like algorithm for SAR images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 453–466. [CrossRef]
6. Ye, Y.; Shen, L. Hopc: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 9. [CrossRef]
7. Ye, Y.; Shan, J.; Bruzzone, L.; Shen, L. Robust registration of multimodal remote sensing images based on structural similarity. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2941–2958. [CrossRef]
8. Zagoruyko, S.; Komodakis, N. Learning to compare image patches via convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4353–4361.
9. Merkle, N.; Luo, W.; Auer, S.; Müller, R.; Urtasun, R. Exploiting deep matching and SAR data for the geo-localization accuracy improvement of optical satellite images. *Remote Sens.* **2017**, *9*, 586. [CrossRef]
10. Mou, L.; Schmitt, M.; Wang, Y.; Zhu, X.X. A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes. In Proceedings of the 2017 Joint Urban Remote Sensing Event (JURSE), Dubai, United Arab Emirates, 6–8 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
11. Hughes, L.H.; Schmitt, M.; Mou, L.; Wang, Y.; Zhu, X.X. Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 784–788. [CrossRef]
12. Xu, Y.; Xiang, X.; Huang, M. Task-Driven Common Representation Learning via Bridge Neural Network. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5573–5580.
13. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
14. He, K.; Girshick, R.; Dollár, P. Rethinking imagenet pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4918–4927.
15. Wu, Z.; Xiong, Y.; Yu, S.X.; Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3733–3742.
16. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
17. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multiview coding. *arXiv* **2019**, arXiv:1906.05849.
18. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Vienna, Austria, 12–18 July 2020; pp. 1597–1607.
19. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
20. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv* **2020**, arXiv:2006.09882.
21. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *arXiv* **2021**, arXiv:2103.03230.
22. Henaff, O. Data-efficient image recognition with contrastive predictive coding. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 12–18 July 2020; pp. 4182–4192.
23. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 448–456.
24. Wang, Y.; Zhu, X.X. The sarptical dataset for joint analysis of sar and optical image in dense urban area. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 6840–6843.
25. Huang, M.; Xu, Y.; Qian, L.; Shi, W.; Zhang, Y.; Bao, W.; Wang, N.; Liu, X.; Xiang, X. The QXS-SAROPT Dataset for Deep Learning in SAR-Optical Data Fusion. *arXiv* **2021**, arXiv:2103.08259.
26. Schmitt, M.; Hughes, L.H.; Zhu, X.X. The SEN1-2 dataset for deep learning in SAR-optical data fusion. *arXiv* **2018**, arXiv:1807.01569.
27. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
29. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
30. Bao, W.; Huang, M.; Zhang, Y.; Xu, Y.; Liu, X.; Xiang, X. Boosting ship detection in SAR images with complementary pretraining techniques. *arXiv* **2021**, arXiv:2103.08251.