Robust Multi-modal Remote Sensing Image Semantic Segmentation Using Tuple Perturbation-based Contrastive Learning

Jinkun Dai[1], Liang Zhou[1,*], Keyi Duan[1], Yangang Zhao[2], Yuanxin Ye[1]

[1] Faculty of Geosiences and Engineering, Southwest Jiaotong University, 611756, Chengdu, China
daijinkun@my.swjtu.edu.cn, (zhouliangmale, seonnyod)@163.com, yeyuanxin@home.swjtu.edu.cn
[2] The Second Topographic Surveying Brigade of Ministry of Natural Resources, 710054, Xian, China - zyg15926@163.com

Keywords: Multi-modal Remote Sensing Image, Contrastive Learning, Tuple Perturbation, Negative samples, Semantic Segmentation.

Abstract

Deep learning models exhibit promising potential in multi-modal remote sensing image semantic segmentation (MRSISS). However, the constrained access to labeled samples for training deep learning networks significantly influences the performance of these models. To address that, self-supervised learning (SSL) methods have garnered significant interest in the remote sensing community. Accordingly, this article proposes a novel multi-modal contrastive learning framework based on tuple perturbation. Firstly, a tuple perturbation-based multi-modal contrastive learning network (TMCNet) is designed to better explore shared and different feature representations across modalities during the pre-training stage and the tuple perturbation module is introduced to improve the network's ability to extract multi-modal features by generating more complex negative samples. In the fine-tuning stage, we develop a simple and effective multi-modal semantic segmentation network (MSSNet), which can reduce noise by using complementary information from various modalities to integrate multi-modal features more effectively, resulting in better semantic segmentation performance. Extensive experiments have been carried out on two published multi-modal image datasets including optical and SAR pairs, and the results show that the proposed framework can obtain superior performance of semantic segmentation than the current state-of-the-art methods in cases of limited labeled samples.

1. Introduction

Remote sensing image semantic segmentation (RSISS) aims to accurately classify each pixel in an image into specific semantic categories, that is, to perform pixel-level image segmentation, so as to more accurately understand and describe different objects and scenes in the image (Zhang et al., 2023). Its main application areas include land cover classification (Hong et al., 2023), cartography (Hong et al., 2019), natural resource management, environmental monitoring, geospatial information analysis and other fields. Therefore, RSISS is a crucial foundational task in the analysis of images.

RSISS mainly includes semantic segmentation of optical, synthetic aperture radar (SAR), infrared, multi-spectral, hyperspectral, or other dataset (Roy et al., 2022). Since it is difficult to break through the limitation of information diversity when only using the single-modal remote sensing image (RSI), the performance of semantic segmentation inevitably encounters bottlenecks in complex scenes. Compared with the semantic segmentation single-modal images, that of multi-modal images can obtain a more comprehensive segmentation map of ground objects by making full use of complementary information between images (Li et al., 2020). Currently, the explosion of remote sensing data makes it possible to obtain multi-modal RSI within the same area, which offers a unique chance to improve semantic segmentation accuracy through leveraging the complementary characteristics of diverse imaging modes (Audebert et al., 2017, Hu et al., 2021, Ye et al., 2024a). However, the MRSISS methods based on supervised deep learning rely heavily on a lot of high-quality labeled samples, which are very hard to get, time-consuming, and labor-intensive (Jaiswal et al., 2021). With the MRSISS technology playing an increasingly crucial role in rapid global development, the issue of insufficient labeled samples is becoming increasingly severe.

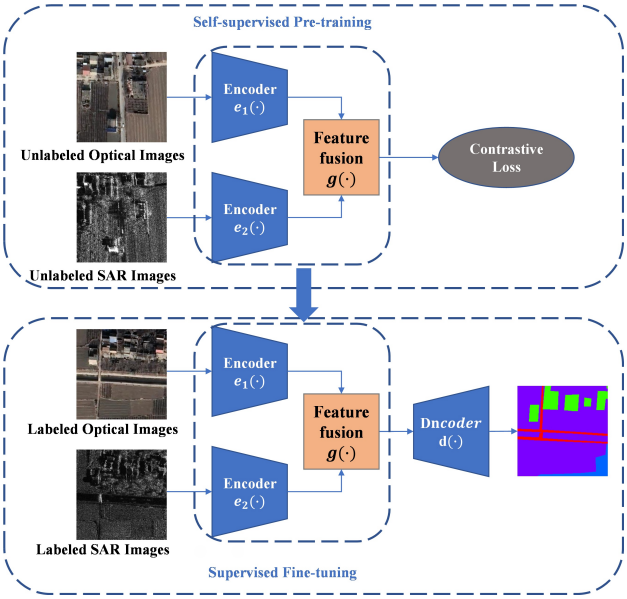

Figure 1. Multi-modal contrastive learning framework.

To solve this problem, self-supervised learning (SSL) (Li et al., 2021, Stojnić and Risojević, 2021) offers an innovative framework. This approach reaches a performance level on these tasks that rivals that of supervised learning, even with a minimal number of labeled samples (Chen and Bruzzone, 2022, Tao et al., 2022). Accordingly, the SSL framework is employed in our MRSISS method (see figure 1). While a vast quantity

of labeled RSI data is challenging to obtain, an abundance of diverse and rich unlabeled RSI data is readily accessible. As a typical self-supervised learning method, contrastive learning has made great achievements in the field of computer vision (CV). A contrastive learning framework of visual representations (SimCLR) (Chen et al., 2020), which uses larger batch-size, more data augmentation, the projection head and other technologies, is present to achieve very good results on ImageNet. The self-distillation with no labels called DINO (Caron et al., 2021), which changes the backbone network from residual network to vision transformer (ViT), and normalizes the output of teacher network, is proposed to effectively improve the stability of model training. Recent researches have integrated contrastive learning into remote sensing, successfully validating its applicability.

In addition, the existing contrastive learning mainly focuses on single-modal RSISS tasks, only concentrating on the learning of single-modal features and ignoring the relationship between multi-modal features (Cha et al., 2022). There is a balance between learned optical feature and SAR feature in MRSISS tasks, and the primary challenge is how to efficiently extract and fuse multi-modal features(Ye et al., 2024b). The contrastive multi-view coding (CMC) (Tian et al., 2020), which extends the task of two views to multiple views, laying the foundation for subsequent multi-modal contrastive learning research is put forward. However, this approach focuses the stronger modalities, neglecting the weaker ones. To solve this problem, the TupleInfoNCE (Liu et al., 2021), which is regarded as a new contrastive learning objective, is present. Inspired by it, we design a multi-modal contrast learning framework based on tuple perturbation, which overcomes the problem of feature interaction between different modalities in the feature extraction process of TupleInfoNCE. During the pre-training stage, we design a tuple perturbation-based multi-modal contrastive learning network (named TMCNet). The pseudo-Siamese feature extraction module is used to extract the features of optical and SAR images respectively to avoid mutual interference of different modal features. Then the tuple perturbation module is employed to better explore the shared and different feature representations between modalities and improve the network's ability to extract multi-modal features by generating more complex negative samples. In addition, unlike with computer vision, which pays more attention to the global features of images, RSISS requires a balance between global features and local features. For this reason, we adopt a loss function combining global contrastive loss and local contrastive loss to produce better feature representations. Moreover, in the fine-tuning stage, we design a simple and effective multi-modal semantic segmentation network (named MSSNet). In such network, a multi-modal feature rectification module (MFRM) is designed to reduce noise by using the complementary information between multi-modalities, and a multi-scale multi-modal feature fusion module (MMFFM) is also designed to achieve an elegant collection of cross-attention and feature augmentation, fusing multi-modal features more closely and improving semantic segmentation more effectively, including low-level feature fusion module (LFFM) and high-level feature fusion module (HFFM).

The primary contributions of this paper are outlined as follows:

1. We apply the self-supervised contrastive learning to the MRSISS and propose a novel multi-modal contrastive learning framework based on tuple perturbation. Validated across diverse datasets, the proposed framework allows the model to autonomously learn from unlabeled images and leverage a minimal set of annotations to direct the tasks involving multi-modal semantic segmentation;
2. We propose a tuple perturbation-based multi-modal contrastive learning network (named TMCNet). The TMCNet is proposed to better enhance the extraction of shared and different feature representations between patterns during the pre-training stage, improving the network's ability to extract multi-modal features by generating more complex negative samples;
3. The MSSNet is proposed, which uses the complementary information between multi-modal to reduce noise and fuse multi-modal features more closely, thereby enhancing semantic segmentation results.

2. Methods

2.1 Overall Network Architecture

Figure 1 shows the multi-modal contrastive learning paradigm. On this basis, we propose a novel multi-modal contrastive learning framework based on tuple perturbation. In detail, this framework includes a pre-trained stage TMCNet and a fine-tuning stage MSSNet, which are used to perform semantic segmentation on multi-modal RSI.

2.2 Tuple Perturbation-based Multi-modal Contrastive Learning Network

Given that contrastive learning improves by making positive samples similar and negative samples dissimilar, its core lies in the generation of positive and negative samples. On this basis, multi-modal contrastive learning considers the differences and complementarities between different modalities and encourages the model to distinguish different samples from different modalities. We design the TMCNet to train the encoder of MSSNet, as shown in figure 2, which consists of the following five key parts:

1) Data augmentation: To incentivize the model to acquire features, we generate four augmented views $\widehat{x}$, $\widetilde{x}$, $\widehat{y}$ and $\widetilde{y}$ from a given set of optical and SAR images x and y by data augmentation t_1 and t_2:

$$\begin{aligned} \widehat{x} = t_1(x), \widetilde{x} = t_2(x) \\ \widehat{y} = t_1(y), \widetilde{y} = t_2(y) \end{aligned} \tag{1}$$

where t_1 = process of random cropping and subsequently resizing to a constant resolution
t_2 = a sequence of enhancements including the steps in t_1, along with random flipping, color distortion, noise introduction, and more

2) Feature extraction: The pseudo-Siamese encoder networks $e_1(\cdot)$ and $e_2(\cdot)$, which are employed to extract global features from the optical and SAR augmented sample instances respectively, enabling the construction of representations for downstream tasks.

$$\begin{aligned} \widehat{f_{x_i}} = \mu\big(e_1(\widehat{x_i})\big), \widetilde{f_{x_i}} = \mu\big(e_1(\widetilde{x_i})\big) \\ \widehat{f_{y_i}} = \mu\big(e_2(\widehat{y_i})\big), \widetilde{f_{y_i}} = \mu\big(e_2(\widetilde{y_i})\big) \end{aligned} \tag{2}$$

where μ = global average pooling
$e_1(\cdot)$, $e_2(\cdot)$ = encoders of the DeepLabV3+

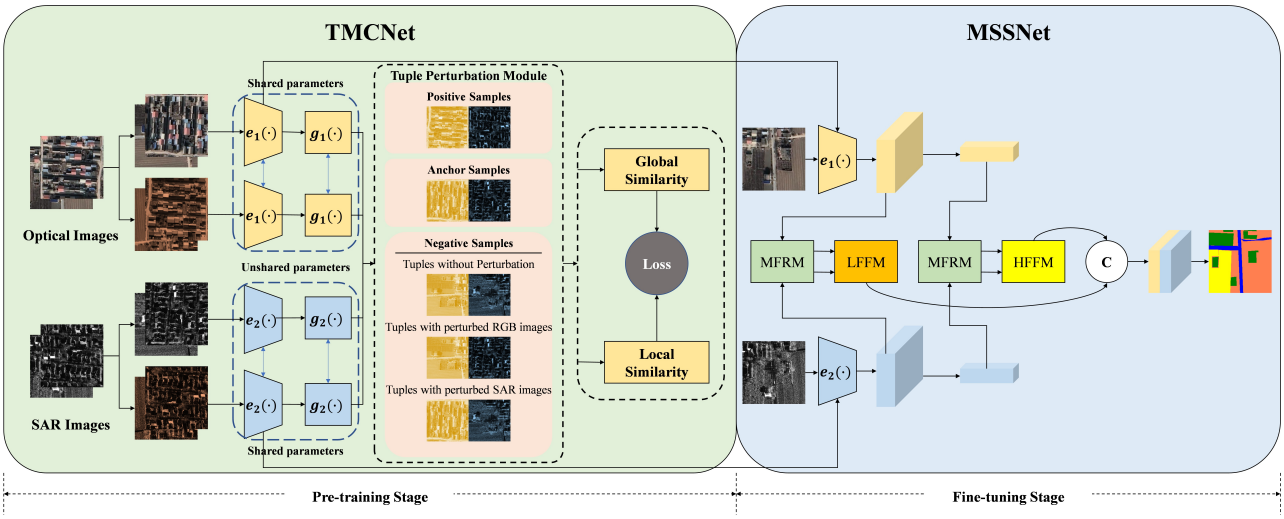

Figure 2. Architecture of tuple perturbation-based contrastive learning framework for MRSISS.

3) Projection head: A nonlinear projection function $g(\cdot)$ further processes the extracted features, mapping them to the layer z, at which the loss is computed.

$$\begin{aligned} \widehat{z_{x_i}} &= g(\widehat{f_{x_i}}) = W^{(2)}R(W^{(1)}\widehat{f_{x_i}}), \widetilde{z_{x_i}} = g(\widetilde{f_{x_i}}) \\ \widehat{z_{y_i}} &= g(\widehat{f_{y_i}}) = W^{(2)}R(W^{(1)}\widehat{f_{y_i}}), \widetilde{z_{y_i}} = g(\widetilde{f_{y_i}}) \end{aligned} \tag{3}$$

where R = rectified linear unit ($ReLU$)

4) Tuple perturbation: Generating challenging negative samples is essential for effective representation learning in contrastive learning, especially in the case of multi-modal contrastive learning for optical and SAR images, in which optical modality tends to dominate the learned representation. In order to alleviate the limitation of ignoring SAR modality and cooperation between modalities in contrastive learning, we propose a tuple perturbation module to generate negative samples, in which different modalities are different for different scenes. This prevents the network from becoming lazy and overly focusing on the optical modality.

Given a set of optical and SAR images, as illustrated in figure 3, we input the optical and SAR images into the pseudo-siamese encoder to extract the features, and then input the features into the tuple perturbation module.

Specifically, the augmentation is applied to N optical and SAR sample pairs from a mini-batch, generating a total of $2N$ pairs. Each augmented pair originating from the same initial pair constitutes positive samples, while the remaining $2(N-1)$ pairs are treated as negatives. Building upon this foundation, the tuple perturbation module generates an additional $4(N-1)$ more challenging negative samples, of which the optical features in $2(N-1)$ sample pairs are consistent with those in the anchor sample, and their corresponding SAR features are perturbed, specifically speaking, SAR features are replaced by SAR features in different regions. Similarly, the other $2(N-1)$ sample pairs keep the SAR features unchanged and replace the optical features. Therefore, when the anchor sample is defined as $\widehat{z_{x_i}} \otimes \widetilde{z_{y_i}}$, the positive sample is $\widetilde{z_{x_i}} \otimes \widetilde{z_{y_i}}$ and the negative samples are $\sum_{j=1,(j\neq i)}^{N}(\widehat{z_{x_j}} \otimes \widetilde{z_{y_j}} + \widetilde{z_{x_j}} \otimes \widetilde{z_{y_j}}) + \sum_{j=1,(j\neq i)}^{N}(\widehat{z_{x_i}} \otimes \widehat{z_{y_j}} + \widetilde{z_{x_i}} \otimes \widetilde{z_{y_j}}) + \sum_{j=1,(j\neq i)}^{N}(\widehat{z_{x_j}} \otimes \widehat{z_{y_i}} + \widetilde{z_{x_j}} \otimes \widetilde{z_{y_i}})$, where $\otimes$ denotes feature fusion.

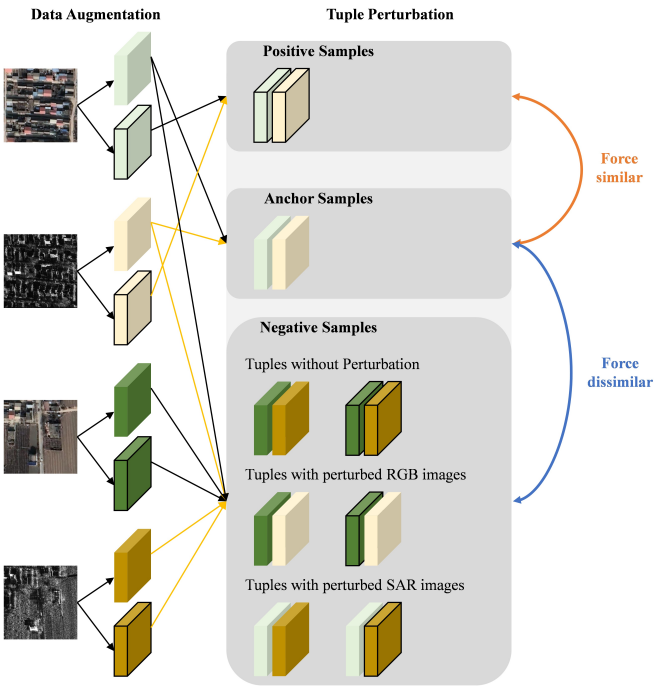

Figure 3. Details of tuple perturbation module.

5) Contrastive loss: Contrastive loss encourages similarity between positive sample pairs and dissimilarity among negative sample pairs. From the initial N samples in a mini-batch, augmentation results in $4N-2$ samples. A positive pair arises from augmenting the same initial sample, while $4(N-1)$ samples compose the negative set. The contrastive loss $\mathcal{L}_C$ is thus defined as:

$$\mathcal{L}_C = \frac{1}{2N}\sum_{i=1}^{N}(\ell(\widehat{m_i} + \widetilde{m_i}) + \ell(\widetilde{m_i} + \widehat{m_i})) \tag{4}$$

With:

$$\ell(\widehat{m_i} + \widetilde{m_i}) = -\log \frac{exp(sim(\widehat{z_{x_i}} \otimes \widehat{z_{y_i}}, \widetilde{z_{x_i}} \otimes \widetilde{z_{y_i}})/\tau)}{\sum_{x\in\Lambda_-} exp(sim(\widehat{z_{x_i}} \otimes \widehat{z_{y_i}}, xy)/\tau)}$$

where xy = negative samples

In this work, the final loss is defined by combining the global and local contrastive losses:

$$\mathcal{L} = \gamma \cdot \mathcal{L}_G + (1 - \gamma) \cdot \mathcal{L}_L \tag{5}$$

where γ = a constant of 0.5
$\mathcal{L}_G$ = global contrastive loss
$\mathcal{L}_L$ = local contrastive loss

2.3 Multi-modal Semantic Segmentation Network

After the pre-training stage, the feature extractor produces better representation ability. To further improve the segmentation performance of the feature extractor, the MSSNet is proposed. The network takes independent optical and SAR images, along with their pixel-to-pixel corresponding labels, as the input. To extract more optical and SAR features from RSI, different levels of feature representations can be obtained using the different residual blocks of the pre-trained feature extractor. In general, feature extractor consists of some stages, each containing a sequence of residual blocks. In this scenario, low-level and high-level feature representations can be extracted from different residual blocks of the pre-trained feature extractor, allowing for the acquisition of feature information at different abstraction levels. Recognizing the features of different modalities have their own unique noise information, we design the MFRM to rectify optical and SAR features to minimize the impact of noise on segmentation results. In the feature fusion stage, we separately design the LFFM and HFFM for the feature fusion of optical and SAR images after feature rectification, so as to enhance information interaction and improve segmentation performance. Finally, the deconvolution operation is used to sample the fused features to the original input size and merge them to produce the segmentation map.

3. Experiments and Result

3.1 Data Description

We assess the proposed TMCNet alongside other self-supervised approaches, as well as the proposed MSSNet alongside other multi-modal semantic segmentation approaches on two multi-modal datasets for MRSISS.

Dataset	Pre-training	Train	Test
YESeg-OPT-SAR	2231	800	1431
WHU-OPT-SAR	3528	800	2728

Table 1. Sample size of two typical multi-modal datasets (256×256 pixels).

3.1.1 YESeg-OPT-SAR dataset: The dataset is a sufficiently labeled RSI for land use segmentation, which consists of 2231 high-resolution optical images and SAR images. Both the spatial size of optical and SAR images is 256×256 at a 0.5-meter resolution. The images in the dataset are completely labeled with seven pixel-level categories, including background, bare land, vegetation, trees, houses, water, roads, and others. YESeg-OPT-SAR dataset is presented on https://github.com/yeyuanxin110/YESeg-OPT-SAR.

3.1.2 WHU-OPT-SAR dataset: This dataset is a large-scale optical and SAR joint land use segmentation dataset, which is built by using the optical images of GaoFen-1 satellite and the SAR image of GaoFen-3 satellite in the same area. The dataset consists of 100 images, spanning an area totaling 51448.56 square kilometers, at a 5-meter resolution. Enhanced with pixel-level annotations, these images are ideal for deep learning-driven land use segmentation tasks. Seven labels categorize the pixels into distinct classes such as farmland, urban, rural, aquatic, forested areas, roads, and miscellaneous features. The WHU-OPT-SAR dataset, alongside its corresponding annotations, is accessible at https://github.com/AmberHen/WHU-OPT-SAR-dataset.

3.1.3 Evaluation Metrics: To evaluate the proposed framework, we apply OA and $Kappa$ coefficients to determine the accuracy across the test datasets for downstream tasks. These metrics are described as:

$$\begin{aligned} OA &= \frac{TP}{N} \\ Kappa &= \frac{OA - p_e}{1 - p_e} \\ p_e &= \frac{a_1 \times b_1 + ... + a_c \times b_c}{N \times N} \end{aligned} \tag{6}$$

where TP = total number of correctly predicted pixels
N = total number of pixels
p_e = pixel-wise accuracy
a_c = number of actual pixels of class c
b_c = number of predicted pixels of class c

3.2 Evaluation Metrics and Implementation Details

3.2.1 Implementation Details: All experiments were conducted on an NVIDIA GeForce RTX 3090 GPU and an Intel Core i7-12700KF CPU. The source code is available at https://github.com/yeyuanxin110/TMCNet-MSSNet. To assess the performance of our proposed multi-modal contrastive learning framework based on tuple perturbation, we compare TMCNet with other typical contrastive learning models (i.e. Jigsaw, Inpainting, momentum contrast v2 (MoCo v2) (Ren et al., 2022), SimCLR and GLCNet (Li et al., 2022)) in the self-supervised pre-training stage and compare MSSNet with other classic multi-modal semantic segmentation networks (i.e. MCANet (MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification, 2022), dual-stream high resolution network (DDHRNet) (Ren et al., 2022) and cross-modal gated fusion network (CMGFNet) (Hosseinpour et al., 2022)) in the fine-tuning stage. For the two datasets, 800 pairs of the data are randomly chosen for training, while the remaining data serves as test samples in our experiments. Table 1 displays the specifics regarding the training and test samples.

Implementation Details on TMCNet: In order to control the variables and ensure the reliability of the experiment, we apply the same multi-modal semantic segmentation network in the downstream task, and its feature extractor is ResNet50. Therefore, Jigsaw, Inpainting, SimCLR, MoCo v2, GLCNet and TMCNet are only designed to train the ResNet50 feature extractor in the contrastive learning pre-training stage. We use the Adam optimizer to optimize the proposed model. During the pre-training stage, the batch size is set to 32, the initial learning rate is set to 0.01 with the cosine decay schedule, and 100 epochs

are executed. Moreover, for the proposed TMCNet, we fuse optical features and SAR features in an appropriate ratio. The model achieving the minimum loss during the pre-training stage is preserved for subsequent tasks.

Implementation Details on MSSNet: In the pre-training stage, we use TMCNet to pre-train the feature extractor, while in the downstream multi-modal semantic segmentation task, we use a variety of multi-modal semantic segmentation networks for comparative experiments. For multi-modal semantic segmentation training, a sparse dataset comprising 800 pairs of self-supervised data is employed. The Adam optimizer is utilized for refining the suggested model. In the fine-tuning phase, the batch size is set to 32 while the initial learning rate is set to 0.01 with the cosine decay schedule, and 100 epochs are executed.

3.3 Accuracy Analysis on TMCNet

This section employs the test datasets from section 3.1 to contrast TMCNet with alternative approaches. We perform both qualitative and quantitative evaluations on the two datasets for the mentioned methods.

3.3.1 Qualitative Accuracy Analysis: In this section, we qualitatively evaluate the performance of the proposed TMCNet on MRSISS tasks on multiple datasets with limited labels, and compare it with other contrastive learning methods. figures 7 and 8 respectively show examples of visualization results of self-supervised tasks on both kinds of dataset. The other self-supervised semantic segmentation methods produce rich segmentation maps. However, it is challenging to balance the optical and SAR features where the optical features are interfered by SAR features, resulting in a decrease in the accuracy of segmentation maps. Although CMC considers the problem of multi-modal features, it ignores that the data of different modalities have different requirements for encoder weights, resulting in the accuracy improvement not obvious. Fully considering the differences between optical and SAR images, our TMCNet uses the pseudo-siamese network, which uses unshared parameters, to extract the optical features and SAR features, and uses tuple perturbation module to improve the performance of the feature extraction network, thus fully utilizing the information in optical and SAR images to improve the segmentation accuracy. The ground objects fail to be well identified due to the SAR images seriously affected by noise. However, the proposed TMCNet is capable of eliminating the influence of noise covering from the visual perspective. In general, our proposed model demonstrates improved capability in distinguishing land cover categories and producing more realistic segmentation maps.

3.3.2 Quantitative Accuracy Analysis: Table 2 reports the quantitative OA and kappa using the proposed method as well as the other self-supervised methods for the optical and SAR data, respectively, for the YESeg-OPT-SAR dataset and WHU-OPT-SAR dataset. The YESeg-OPT-SAR dataset has a high resolution of 0.5 meters, while the WHU-OPT-SAR dataset is of a low resolution of 5 meters. The evaluation data shows that the proposed TMCNet outperforms the rest by obtaining the highest OA, kappa and outperforms others in most class-wise segmentation accuracies. Each method will be further discussed below.

Dataset	YESeg-OPT-SAR		WHU-OPT-SAR	
Method	OA	Kappa	OA	Kappa
Jiasaw	63.58	27.67	74.49	58.94
Inpainting	52.3	22.45	74.95	59.63
MoCo v2	56.33	20.49	77.47	64
SimCLR	59.93	37.38	67.41	50.97
GLCNet	62.58	32.63	73.28	56.32
CMC	57.7	26.93	76.46	63.8
TMCNet (Proposed)	**79.82**	**65.73**	**82.79**	**73.37**

Table 2. OA and Kappa on the YESeg-OPT-SAR and WHU-OPT-SAR datasets (in %) on TMCNet and other self-supervised networks.

Jigsaw: The employed self-supervised technique forms its task through puzzle-solving exercises. An image is segmented into several pieces which are then randomized. Afterward, they are fed into a CNN to discern the contextual links between these segments. As can be seen from table 2, compared with the other contrastive learning methods, Jigsaw has a good overall effect in processing remote sensing images, and its OA is high, and Kappa is average. From figure 4 and 5, we can see that its ability to distinguish complex ground objects has no obvious advantage over other methods;

Inpainting: Inpainting stands as a common technique for generating self-supervised signals, centering around the concept of image reconstruction. As can be seen from table 2, compared with the other contrastive learning methods, the accuracy of Inpainting on the YESeg-OPT-SAR dataset is poor, and it is impossible to identify the ground objects with inconspicuous features. In contrast, the accuracy of Inpainting on the WHU-OPT-SAR dataset is similar to the other contrastive learning methods. And it also can be seen from figure 4 and 5 that it is easy to mistake the ground objects and cannot accurately identify the water body, because it is easily affected by SAR image noise;

MoCo v2: MoCo v2 builds upon instance-based contrastive learning, striving for a broader scope of negative samples that surpasses typical batch sizes. To achieve this, it introduces a dynamic queue for storing negative sample features and employs a momentum update mechanism for the encoder, mitigating representation consistency issues caused by rapid encoder adjustments. From table 2, we can see that MoCo v2 cannot make full use of optical and SAR features, and its performance on both datasets is greatly affected by the quality of image labels. From figure 4 and 5, we can see that there are a lot of errors in the segmentation map;

SimCLR: The SimCLR method also employs instance-wise contrastive learning, which promotes the similarity of positive samples enhanced from a single sample and the dissimilarity of negative samples enhanced from various samples within a mini-batch. It can be seen from table 2 that, compared with MoCo v2, the performance of SimCLR is improved on some datasets, but the optical and SAR features are not balanced, and its segmentation accuracy is poor. From figure 4 and 5, we can see that it is difficult to identify complex ground objects due to the unfavorable effects of SAR features;

GLCNet: The GLCNet method, like SimCLR, is based on the idea of instance-wise contrastive learning. Compared with SimCLR, it emphasizes the equilibrium between global and local feature learning in RSISS tasks for remote sensing images. As can be seen from table 2, the performance of GLCNet has some limitations on the WHU-OPT-SAR dataset, but it has been improved on the YESeg-OPT-SAR dataset. From figure 4 and 5,

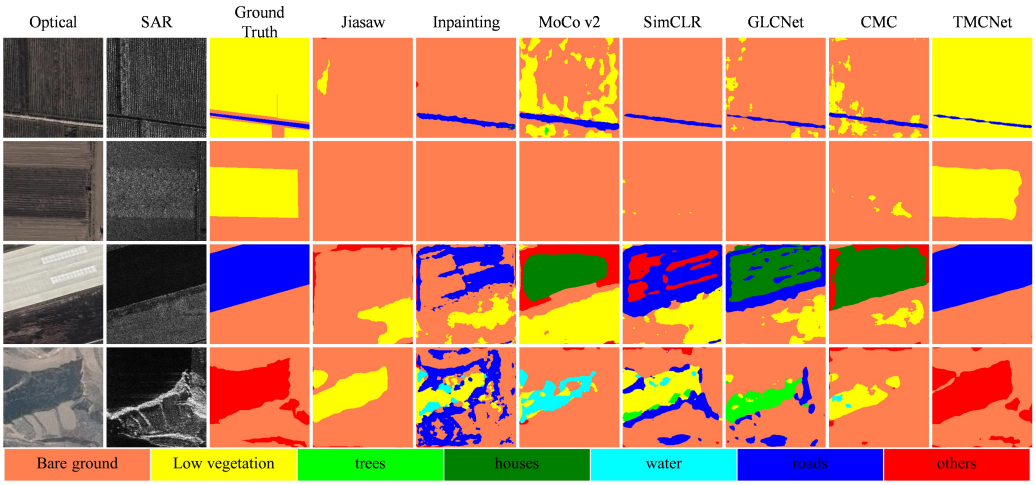

Figure 4. Examples of visualization results on the YESeg-OPT-SAR dataset on TMCNet and other self-supervised networks.

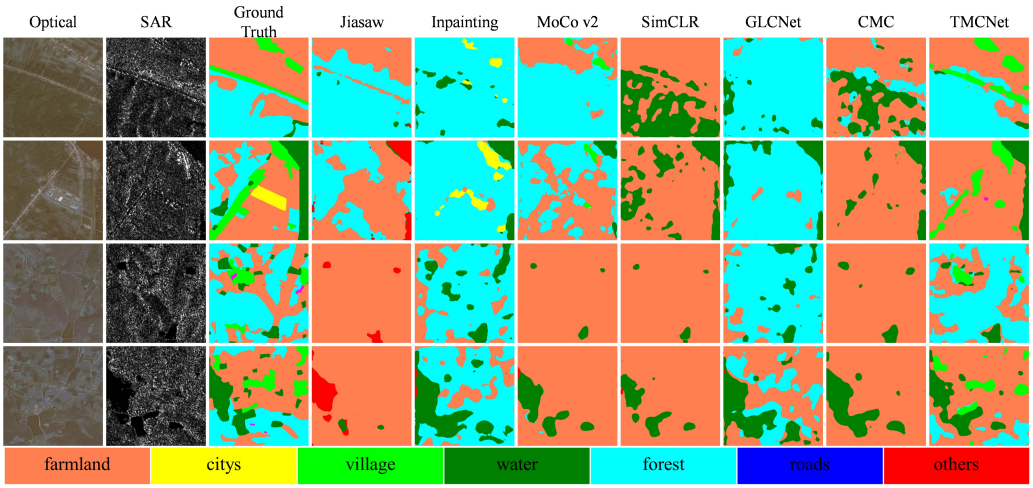

Figure 5. Examples of visualization results on the WHU-OPT-SAR dataset on TMCNet and other self-supervised networks.

it can be seen that the performance of GLCNet in roads identification is significantly reduced, and there are obvious errors in the segmentation of other categories;

CMC: CMC is a deep learning method, and its goal is to train the shared information characteristics of multiple sensing perspectives by maximizing the mutual information of different perspectives in the same scene. Replacing different perspectives with different modalities can be used in MRSISS tasks, and the more modalities, the better the effect. As can be seen from table 2, the performance of CMC in processing multi-modal datasets has been significantly improved, especially on the WHU-OPT-SAR dataset. From figure 4 and 5, we can see that the segmentation of ground objects can be roughly judged, but the integrity of ground objects cannot be guaranteed;

TMCNet: The information provided in table 2 illustrates that TMCNet has achieved the best accuracy, and its performance has been significantly improved on the YESeg-OPT-SAR dataset and the WHU-OPT-SAR dataset. For example, on the WHU-OPT-SAR dataset, its comprehensive accuracy far exceeds all existing contrastive learning methods, although it is still difficult to identify village and other categories. For the YESeg-OPT-SAR dataset, TMCNet has more obvious advantages, and achieves the best results in all categories. It can be clearly observed from figure 7 that TMCNet achieves the best performance, which can accurately identify low vegetation and other ground objects. From figure 8, although TMCNet misclassifies ground objects, it is still superior to other methods. The above experimental results prove that TMCNet can be applied to various types of MRSISS tasks, and adapted to datasets with different resolutions. In comparison to other self-supervised methods, TMCNet obtains the most comprehensive and accurate segmentation results overall.

3.4 Accuracy Analysis on MSSNet

In this subsection, we utilize the test datasets from Section 3.1 to compare MSSNet with other methods. Both qualitative and quantitative analyses are conducted separately for these methods on the two datasets.

3.4.1 Qualitative Accuracy Analysis: We qualitatively evaluate the performance of the proposed MSSNet on MRSISS tasks on multiple datasets with limited labels, comparing it to other multi-modal semantic segmentation networks. Figure 6 and 7 respectively show examples of visualized results of multi-modal semantic segmentation tasks on the YESeg-OPT-SAR dataset and the WHU-OPT-SAR dataset. Although

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-77-2024 | © Author(s) 2024. CC BY 4.0 License.

Dataset	YESeg-OPT-SAR		WHU-OPT-SAR	
Method	OA	Kappa	OA	Kappa
CMGFNet	78.72	62.88	78.15	65.27
DDHRNet	77.84	61.01	74.76	59.94
MCANet	77.18	60.87	75.46	60.37
MSSNet (Proposed)	**79.82**	**65.73**	**82.79**	**73.37**

Table 3. OA and Kappa on the YESeg-OPT-SAR and WHU-OPT-SAR datasets (in %) on MSSNet and other multi-modal semantic segmentation networks.

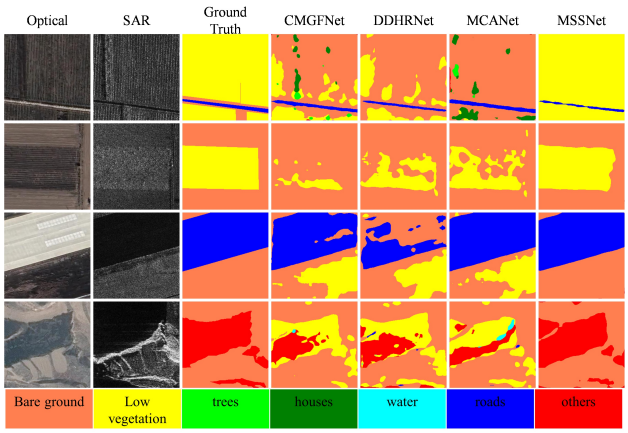

Figure 6. Examples of visualization results on the YESeg-OPT-SAR dataset on TMCNet and other self-supervised networks.

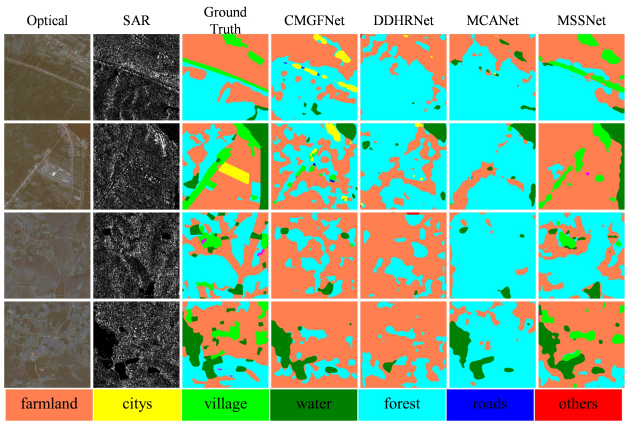

Figure 7. Examples of visualization results on the WHU-OPT-SAR dataset on TMCNet and other self-supervised networks.

other multi-modal semantic segmentation methods can generate a significant number of meaningful segmentation maps, the segmentation accuracy is not high enough because of the failure to deeply explore and integrate the features of optical and SAR images. Taking full account of the shared and different features of optical and SAR images, on the basis of multi-modal feature correction, MSSNet fuses low-level and high-level features respectively to improve the segmentation accuracy.

3.4.2 Quantitative Accuracy Analysis: Table 3 reports the quantitative OA and kappa using the proposed model as well as the other multi-modal semantic segmentation methods for the optical and SAR data, respectively, for the YESeg-OPT-SAR dataset and WHU-OPT-SAR dataset. The evaluation data shows that the proposed MSSNet not only achieves the highest OA, kappa but also outperforms other multi-modal semantic segmentation methods in most class-wise segmentation accuracies. Each method will be further discussed below.

CMGFNet: CMGFNet employs dual encoders to capture multi-tiered features from diverse modalities, providing a pair of fusion techniques to amalgamate traits from two sources. It introduces the Gated Fusion Module (GFM) to effectively merge cross-modal features. Moreover, it adopts a top-down approach to blend deep, high-level attributes with more superficial low-level features for multi-tiered fusion. Table 3 shows that the performance of CMGFNet is acceptable. From figure 6 and 7, we can see that it has strong limitations;

DDHRNet: DDHRNet intricately merges features from SAR and optical sources within each of its branches, leveraging the disparate data to enhance the segmentation quality, unaffected by cloud coverage. Table 3 demonstrates the limitations of DDHRNet in multi-modal semantic segmentation. Figure 6 and 7 reveal that DDHRNet does not improve the shortcomings of CMGFNet;

MCANet: MCANet, inspired by the renowned Deeplabv3+ framework, is designed to classify land use utilizing both optical and SAR imagery. This network processes these images and their pixel-level annotations independently. From table 3, it can be seen that the performance of MCANet has not been improved compared with DDHRNet and CMGFNet. From figure 6 and 7, we can see that it also fails to improve the shortcomings of other schemes;

MSSNet: As can be seen from table 3, MSSNet obtains the highest accuracy on the YESeg-OPT-SAR dataset and the WHU-OPT-SAR dataset. From figure 6, we can see that the segmentation visualized result is excellent, especially the low vegetation will not be wrongly classified. From figure 7, we can see that it can accurately identify the ground object on the WHU-OPT-SAR dataset, but it is difficult for other multi-modal semantic segmentation networks to identify the ground object.

The above experimental results demonstrate that MSSNet can be applied to various types of MRSISS tasks, and can adapt to the datasets with different resolutions. Comparing other multi-modal semantic segmentation methods, MSSNet provides the most comprehensive and accurate segmentation results overall.

4. Conclusion

In this work, we integrate contrastive learning with the task of MRSISS, learning from an abundance of unlabeled images to extract both shared and different optical features and SAR features with the intent to lessen the dependence on labeled samples. Moreover, since existing multi-modal contrastive learning methods predominantly concentrate on the stronger modality at the expense of the weaker one, these approaches could prove less optimal for MRSISS tasks that depend on multi-modal features. To address this issue, we propose a new multi-modal contrastive learning framework of remote sensing image semantic segmentation, which consists of a tuple perturbation-based multi-modal contrastive learning network named TMCNet and a multi-modal semantic segmentation network named MSSNet. The tuple perturbation module within TMCNet makes the features of different modalities

obtained by feature extractor fused in a more compact way, which tends to lead to better semantic segmentation performance. The proposed MSSNet uses the complementary information between images to reduce noise and fuse multi-modal features more closely. The experimental results show that TMCNet and MSSNet have better performance in MRSISS tasks than other state of art methods. Additionally, the increase in self-supervised pre-training samples correlates with performance enhancements. Given that a large amount of remote sensing data is readily obtainable, our method harbors the potential for substantial practical utility.

In this article, we primarily focus on optical and SAR images in our multi-modal semantic segmentation experiment. In future work, we will investigate additional modal images, including infrared, multi-spectral, and hyperspectral, to improve the versatility of the model.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (No.42271446) and key topics of scientific and technological research plan of China Railway Design Corporation (2023A0240104).

References

Audebert, N., Saux, B. L., Lefèvre, S., 2017. Beyond RGB: Very High Resolution Urban Remote Sensing With Multimodal Deep Networks. *ArXiv*, abs/1711.08681.

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9630–9640.

Cha, K., Seo, J., Choi, Y., 2022. Contrastive Multiview Coding With Electro-Optics for SAR Semantic Segmentation. *IEEE Geoscience and Remote Sensing Letters*, 19, 1-5.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations.

Chen, Y., Bruzzone, L., 2022. Self-Supervised SAR-Optical Data Fusion of Sentinel-1/-2 Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-11.

Hong, D., Yokoya, N., Chanussot, J., Zhu, X. X., 2019. An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing. *IEEE Transactions on Image Processing*, 28(4), 1923-1938.

Hong, D., Zhang, B., Li, H., Li, Y., Yao, J., Li, C., Werner, M., Chanussot, J., Zipf, A., Zhu, X. X., 2023. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sensing of Environment*, 299, 113856.

Hosseinpour, H., Samadzadegan, F., Javan, F. D., 2022. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184, 96-115.

Hu, J., Zhao, G., You, S., Kuo, C., 2021. Evaluation of multimodal semantic segmentation using rgb-d data. 47.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., Makedon, F., 2021. A survey on contrastive self-supervised learning.

Li, H., Li, Y., Zhang, G., Liu, R., Huang, H., Zhu, Q., Tao, C., 2022. Global and Local Contrastive Self-Supervised Learning for Semantic Segmentation of HR Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-14.

Li, J., Zhou, P., Xiong, C., Hoi, S. C. H., 2021. Prototypical contrastive learning of unsupervised representations.

Li, X., Lei, L., Sun, Y., Li, M., Kuang, G., 2020. 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*, 13, 1011-1026.

Liu, Y., Fan, Q., Zhang, S., Dong, H., Funkhouser, T., Yi, L., 2021. Contrastive multimodal fusion with tupleinfonce. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 734–743.

MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification, 2022. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106, 102638.

Ren, B., Ma, S., Hou, B., Hong, D., Chanussot, J., Wang, J., Jiao, L., 2022. A dual-stream high resolution network: Deep fusion of GF-2 and GF-3 data for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102896.

Roy, S. K., Deria, A., Hong, D., Rasti, B., Plaza, A. J., Chanussot, J., 2022. Multimodal Fusion Transformer for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-20.

Stojnić, V., Risojević, V., 2021. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1182–1191.

Tao, C., Qi, J., Lu, W., Wang, H., Li, H., 2022. Remote Sensing Image Scene Classification With Self-Supervised Paradigm Under Limited Labeled Samples. *IEEE Geoscience and Remote Sensing Letters*, 19, 1-5.

Tian, Y., Krishnan, D., Isola, P., 2020. Contrastive multiview coding. A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (eds), *Computer Vision – ECCV 2020*, Springer International Publishing, Cham, 776–794.

Ye, Y., Yang, C., Gong, G., Yang, P., Quan, D., Li, J., 2024a. Robust optical and SAR image matching using attention-enhanced structural features. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–12.

Ye, Y., Zhang, J., Zhou, L., Li, J., Ren, X., Fan, J., 2024b. Optical and SAR Image Fusion Based on Complementary Feature Decomposition and Visual Saliency Features. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-15.

Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R., 2023. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24(12), 14679-14694.