frontiers | Frontiers in Ecology and Evolution

TYPE Methods
PUBLISHED 22 December 2022
DOI 10.3389/fevo.2022.1083801

OPEN ACCESS

EDITED BY
Yansheng Li,
Wuhan University, China

REVIEWED BY
Chengli Peng,
Wuhan University, China
Qiqi Zhu,
China University of Geosciences
Wuhan, China

*CORRESPONDENCE
Yunsheng Zhang
✉ zhangys@csu.edu.cn

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 29 October 2022
ACCEPTED 06 December 2022
PUBLISHED 22 December 2022

CITATION
Wang X, Zhang Y, Zhang Z, Luo Q and Yang J (2022) GSC-MIM: Global semantic integrated self-distilled complementary masked image model for remote sensing images scene classification.
Front. Ecol. Evol. 10:1083801.
doi: 10.3389/fevo.2022.1083801

GSC-MIM: Global semantic integrated self-distilled complementary masked image model for remote sensing images scene classification

Xuying Wang, Yunsheng Zhang*, Zhaoyang Zhang, Qinyao Luo and Jingfan Yang

School of Geosciences and Info-Physics, Central South University, Changsha, Hunan, China

Masked image modeling (MIM) is a learning method in which the unmasked components of the input are utilized to learn and predict the masked signal, enabling learning from large amounts of unannotated data. However, due to the scale diversity and complexity of features in remote sensing images (RSIs), existing MIMs face two challenges in the RSI scene classification task: (1) If the critical local patches of small-scale objects are randomly masked out, the model will be unable to learn its representation. (2) The reconstruction of MIM relies on the visible local contextual information surrounding the masked regions and overemphasizing this local information will potentially lead the model to disregard the global semantic information of the input RSI. Regarding the above considerations, we proposed a global semantic integrated self-distilled complementary masked image model (GSC-MIM) for RSI scene classification. To prevent information loss, we proposed an information-preserved complementary masking strategy (IPC-Masking), which generates two complementary masked views for the same image to resolve the problem of masking critical areas of small-scale objects. To incorporate global information into the MIM pre-training process, we proposed the global semantic distillation strategy (GSD). Specifically, we introduced an auxiliary network pipeline to extract the global semantic information from the full input RSI and transfer the knowledge to the MIM by self-distillation. The proposed GSC-MIM is validated on three publicly available datasets of AID, NWPU-RESISC45, and UC-Merced Land Use, and the results show that the proposed method's Top-1 accuracy surpasses the baseline approaches in three datasets by up to 4.01, 3.87, and 5.26%, respectively.

KEYWORDS

self-supervised learning (SSL), masked image modeling (MIM), self-distillation, remote sensing images (RSIs), scene classification

1. Introduction

The propose and development of self-supervised learning (SSL) has freed conventional supervised learning from the heavy reliance on large-scale high-quality annotated data, making it a reality to obtain well-performed interpretation models label-free (Wang Y. et al., 2022). Compared to supervised learning (Lu et al., 2022; Zhu et al., 2022a,b) which learn the projection data to label, SSL aims to find correlations between data and drive models to learn features by manually constructing pre-tasks to constitute labels, yielding competitive results in RSI scene classification (Tao et al., 2020; Wang X. et al., 2022), object detection (Ding et al., 2021), and semantic segmentation (Li et al., 2022) tasks.

The dominant self-supervised learning approaches can be divided into two main categories: contrastive and generative (Liu X. et al., 2021). The core of the contrastive learning method (Chen et al., 2020; He et al., 2020) is to pull closer different views of the same image (positive samples) and push apart views of different images (negative samples), after which potential invariant features in the images are prompted to be learned by the model through the determination of the positive and negative samples. Currently, the contrastive self-supervised learning-based approaches obtain promising results in various RSI interpretation tasks (Wang Y. et al., 2022). Some researchers (Li et al., 2021a; Akiva et al., 2022) follow the basic idea of contrastive learning to construct positive and negative samples to accomplish label-free training of RSI interpretation models. In addition, since RSIs contain abundant information, some researchers include geographic features (Ayush et al., 2021; Li et al., 2021b), time series features (Manas et al., 2021), and audio features (Heidler et al., 2021) of RSIs in the contrastive learning process to encourage models to learn the invariance of RS-specific features. However, contrastive SSL has its own inherent limitations for RSI interpretation tasks. Specifically, RSI contains a complex variety of land objects and their corresponding labels are at the scene level in scene classification task. If the RSI containing the same type of object is selected as a negative sample, it will negatively influence the feature learning of such object in the pushing apart process, and vice versa (Zhang et al., 2022). Figure 1 shows a brief introduction of the contrastive learning methods and masked image modeling method. Take the negative pair in Figure 1 as example, if we push this negative pair apart, it will inevitably push the feature of orange building apart, which is clearly not ideal. Therefore, we argue that the conventional contrastive SSL is suboptimal for the RSI interpretation tasks.

The fundamental concept of generative self-supervised approaches is to train models to reconstruct and restore manually corrupted images, which essentially enables the models to learn to present well representations of the original image (Liu X. et al., 2021). Mask image modeling (MIM) as a classical generative self-supervised pre-training paradigm, which leverages a large amount of data to drive the training of vision transformer (ViT) (Dosovitskiy et al., 2020), has achieved competitive results in many image interpretation tasks (Bao et al., 2021; He et al., 2022; Xie et al., 2022). The core idea of MIM is to crop the input image into several local semantically meaningful visual tokens and randomly mask some of them, then train the vision transformer to reconstruct the masked parts based on the adjacent visible parts. Based on the above considerations, we assume that MIM-based generative self-supervised learning is more suitable for training remote sensing image interpretation models as it can flexibly acquire various features embedded inside remote sensing images and obtain well-formed representations of RSI without relying on data augmentation methods or negative samples.

However, when MIM is used for remote sensing image scene classification tasks, it usually encounters the following two issues:

- Since remote sensing images contain complex multi-scale land objects, if the critical local patch token containing a certain type of small-scale object is randomly masked out, the model will not be able to obtain its features, resulting in irreversible information loss.
- The reconstruction mechanism of MIM is essentially a model inference based on local contextual information near the masked region, which causes the model to ignore the global semantic information of the whole input image, which is critical for RSI scene classification.

Considering the above limitations, we proposed a global semantic integrated self-distilled complementary masked image model (GSC-MIM) for RSI scene classification, which consist of two strategies: information preserved complementary masking strategy (IPC-Masking) and global semantic distillation strategy (GSD). Existing random masking strategies tend to lose dense and small-scale features in the RSIs. Two versions of the complementary visible-mask patch are generated from the same remote sensing image to maximize the retention of small-scale RS object information when reconstructing them with MIM. The reconstruction of masked patches in MIM relies on local adjacent visible patches and does not emphasize the global scene semantic information of the whole input image. To incorporate global information into the MIM pre-training process, we propose a global semantic distillation strategy (GSD). As an emerging method of SSL, self-distillation (Caron et al., 2021; Cino et al., 2022) learning leverages the network's past self to distillate needed knowledge to the present self to achieve discriminative self-learning. In view of this, we introduce an auxiliary network pipeline to extract global semantic information from the full input image and transfer the global knowledge to the MIM *via* self-distillation.

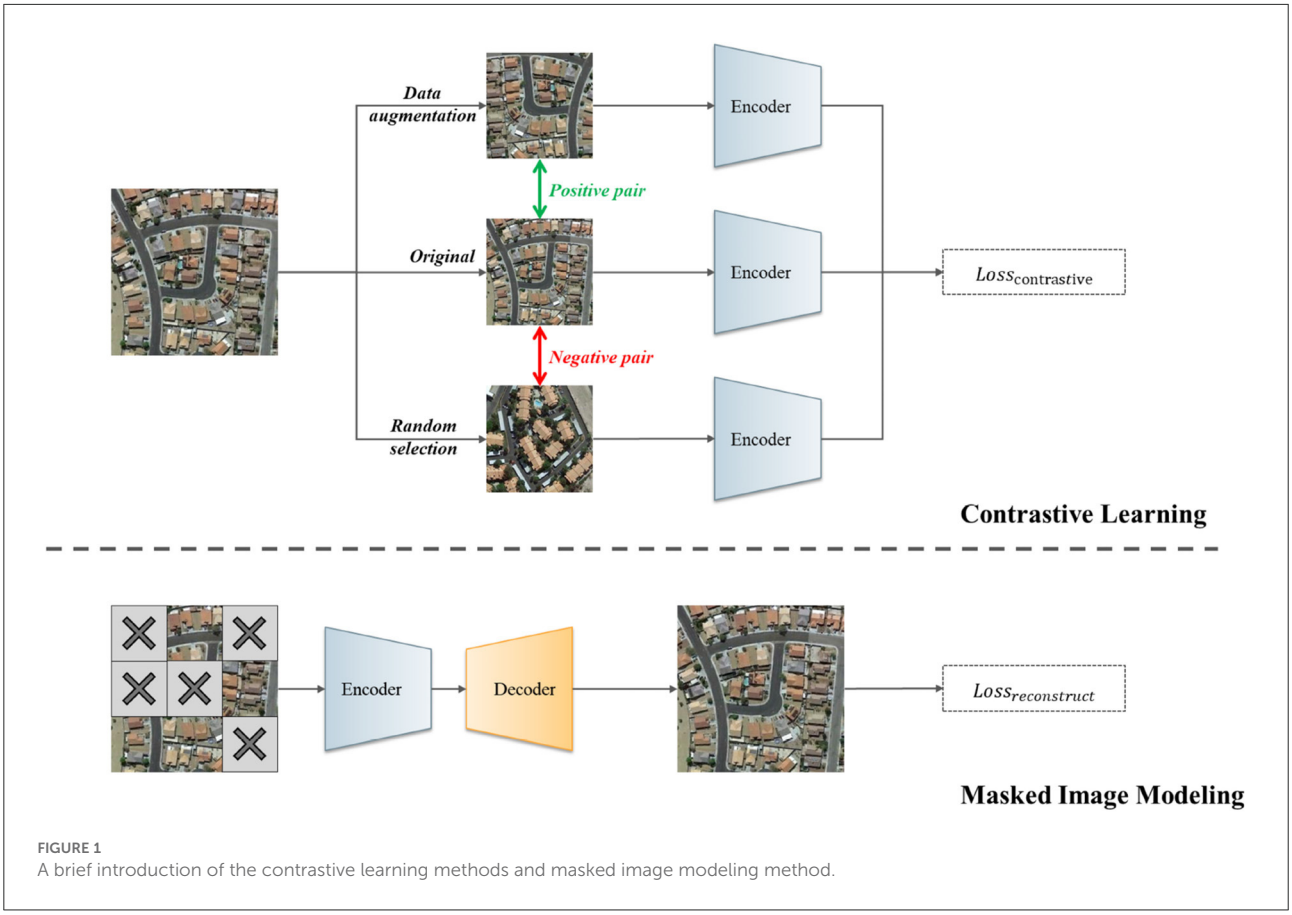

FIGURE 1
A brief introduction of the contrastive learning methods and masked image modeling method.

We use GSC-MIM to obtain the fundamental pre-trained model, and the features obtained by the model can be used for downstream scene classification tasks. We evaluated the performance of the model on three RSI scene classification public datasets, Aerial Image Dataset (AID) (Xia et al., 2017), UC-Merced Land Use (UCM) (Yang and Newsam, 2010), and NWPU-RESISC45 (NWPU45) (Cheng et al., 2017). Experimental results show that our proposed GSC-MIM achieves better results compared to the classical contrastive self-supervised learning, generative self-supervised learning and self-distillation learning approaches. The main contributions of this paper are summarized as follows:

1. We propose an information preserved complementary masking strategy (IPC-Masking), which aims to reduce the loss of small-scale features caused by MIM when used for RSI interpretation tasks. We verified that the information of the original input image could be maximally preserved by the simultaneous reconstruction of its two complementary masked-visible-region views.
2. We highlight the neglect of RSI global semantic information during the MIM training process. We propose a global semantic distillation strategy (GSD) to extract the global semantic information of the input images using additional network pipeline and distill the knowledge into the MIM network to compensate for the lack of global information.
3. Experiments on public datasets show that our proposed GSC-MIM achieves a maximum accuracy improvement of up to 5.26% on the RSI scene classification task under the equivalent conditions. Moreover, the network attention visualization results show that our model captures more highly detailed features of the land objects and locates the global scene information regions of the RSI more precisely.

The rest of this paper is organized as follows: Section 2 introduces the details of our proposed method. Section 3 discusses the experimental and visualization results organized on three public datasets and future work prospects. Section 4 concludes this paper.

2. Methodology

RSI has complex geographical features and a multi-scale spatial layout, if the critical patches containing certain types of features are randomly masked during the MIM training process, it will lead to information loss; Moreover, the reconstruction of local patch by MIM may lead the model to emphasize the

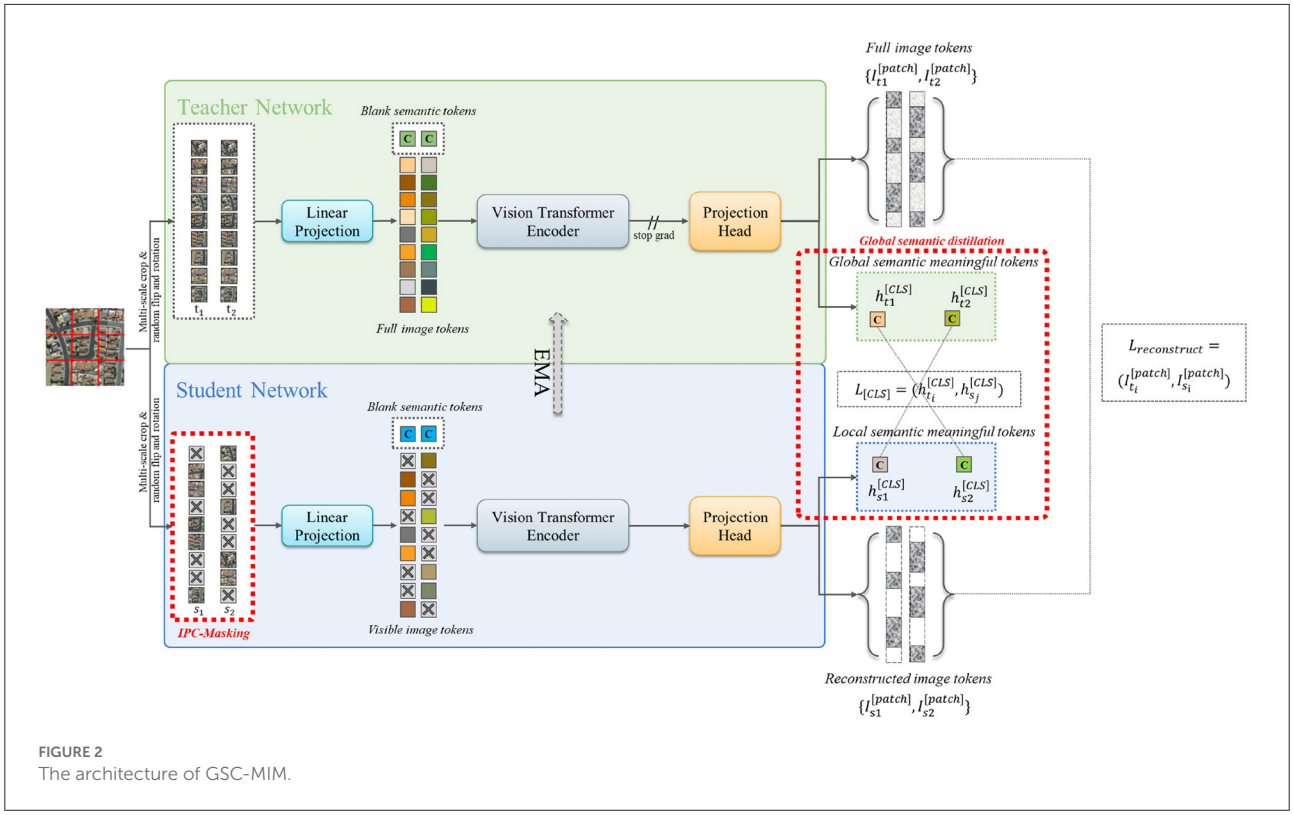

FIGURE 2
The architecture of GSC-MIM.

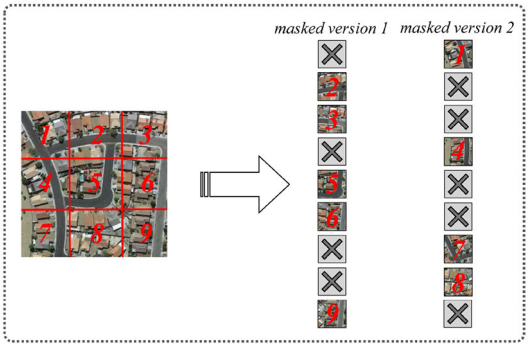

FIGURE 3
The illustration of IPC-masking.

local information and ignore the global information associated with the input image's category. Inspired by the above facts, the GSC-MIM is developed to preserve information of different scales' land objects, and to integrate global meanings into the training process *via* self-distillation. The architecture of GSC-MIM is shown in Figure 2. The architecture consists of two identical structured ViT backbone networks, which we refer to as the teacher and student networks. The network includes two strategies, information-preserved complementary masking strategy (IPC-Masking) and global semantic distillation strategy (GSD), which will be described in the following sections.

2.1. Information-preserved complementary masking strategy (IPC-Masking)

The IPC-Masking is inspired by the conventional MIM's masking strategy. Firstly, for each input image $I_{in} \in \mathcal{R}^{H\times W\times 3}$, we crop it into N patches and feed into the linear projecting layer to obtain patch token sequence $x = \{x_i\}_{i=1}^{N}$. After that, we randomly give each patch token a mask indicator $m \in \{0,1\}^N$, where N is the number of tokens. Specifically, wherever the mask indicator m_i is 1, the corresponding x_i is replaced by a mask $e_{[MASK]}$, which yields vision 1 of the masked image as:

$$\hat{\boldsymbol{x}} \triangleq \left\{\hat{\boldsymbol{x}}_i \mid (1-m_i)\,\boldsymbol{x}_i + m_i \boldsymbol{e}_{[\mathrm{MASK}]}\right\}_{i=1}^{N} \tag{1}$$

Vision 2 of the masked image is complementary to version 1, i.e., the visible-masked tokens of the two images are opposite. The illustration of IPC-Masking is shown in Figure 3. In this work, we perform the IPC-Masking to the input image of the student network. Through the vision transformer backbone and decoder, the reconstructed patch tokens $\{I_{s1}^{[patch]}, I_{s2}^{[patch]}\}$ are generated for comparison.

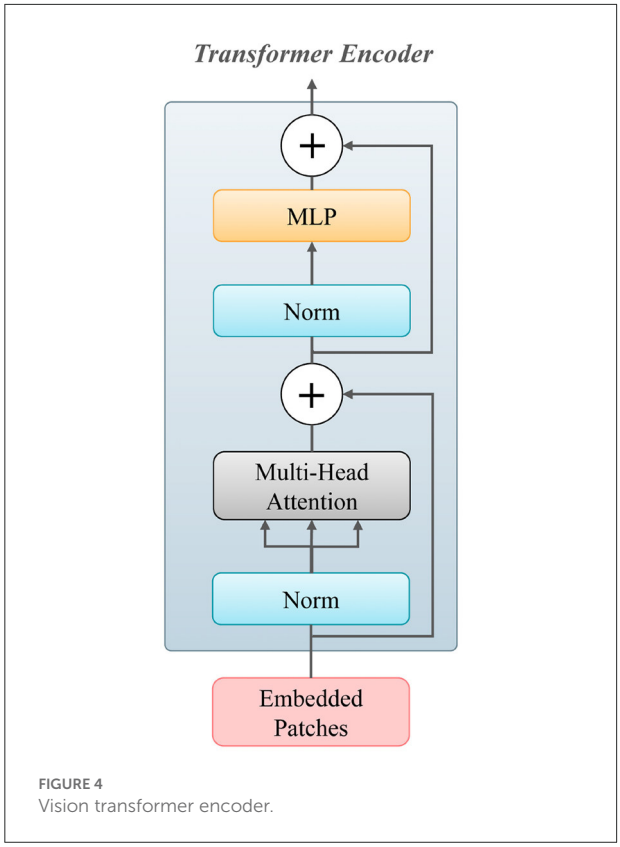

FIGURE 4
Vision transformer encoder.

TABLE 1 Datasets description.

Datasets	AID	NWPU45	UCM
Classes	30	45	21
Images per class	220–420	700	100
Total images	10,000	31,500	2,100
Spatial resolution (m)	0.5–8	0.2–30	0.3
Image size	600 × 600	256 × 256	256 × 256
Data source	Google Earth	Google Earth	USGS
Published year	2017	2017	2010

In the teacher network, the two different augmented non-masked views of the same image serve as input to obtain their projections $\{I_{t1}^{[patch]}, I_{t2}^{[patch]}\}$.

We then define the training objective of complementary MIM as:

$$\mathcal{L}_{\text{reconstructed}} = -\sum_{i=1}^{N} m_i \cdot I_{t_i}^{[\text{patch}]^{\mathrm{T}}} \log I_{s_i}^{[\text{patch}]} \tag{2}$$

We symmetrize the loss by averaging with the above CE term between two pairs of $I_{t_i}^{[patch]}$ and $I_{s_i}^{[patch]}$.

2.2. Global semantic distillation strategy (GSD)

As a branch of self-supervised learning, self-distillation dexterously transmits knowledge from the model's past itself and is cast as a discriminative objective. The GSD proposed in this paper utilized the framework of self-distillation to generate and integrate global semantic information into the student network.

We followed the self-distillation paradigm proposed in Caron et al. (2021). We adopt the same architecture for student and teacher networks, consisting of backbone f and projection head $h : g = h \circ f$. The parameters of the student network θ_s are updated by back propagation according to the loss function, while the teacher's parameters θ_t are exponentially moving averaged (EMA) of the updated θ_s, which is described as follows:

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s \tag{3}$$

where λ following a cosine schedule from 0.996 to 1 during training.

For vanilla ViT (Dosovitskiy et al., 2020), [*CLS*] token is a learnable embedding to the sequence of embedded patches whose state at the output of the Transformer encode and contains the predictive categorical distributions of the input image x. In GSC-MIM, we utilize [*CLS*] tokens as a proxy for global semantic information and to perform knowledge distillation. For a training set $\mathcal{I}$, an image $x \in \mathcal{I}$ is sampled uniformly and applied two random augmentations, yielding two distorted views t_1 and t_2. We also applied complementary masking to the same image to get two corrupted views s_1 and s_2. After feeding t_1 and t_2, s_1, and s_2 with learnable [*CLS*] tokens into the teacher and student network correspondingly, we get the global semantically meaningful tokens of $h_{t_1}^{[\text{CLS}]}, h_{t_2}^{[\text{CLS}]}$ and $h_{s_1}^{[\text{CLS}]}, h_{s_2}^{[\text{CLS}]}$. Since the global semantics of the input image is modeled by the [CLS] token generated *via* the teacher network with unmasked input. The global context encoding ability of the target student masked image model (MIM) is improved by minimizing the cross-view cross-entropy between the teacher's global semantic meaningful [CLS] token and the student's [CLS] token, which we assumed to be global semantically deficient, formulated as:

$$\mathcal{L}_{[\text{CLS}]} = -h_{t_i}^{[\text{CLS}]^{\mathrm{T}}} \log h_{s_j}^{[\text{CLS}]} \tag{4}$$

2.3. Loss function and architecture

With the information preserved complementary masking strategy and global semantic distillation, we designed the overall training objective as follows:

$$\mathcal{L}_{all} = \lambda_1 \mathcal{L}_{[\text{CLS}]} + \lambda_2 \mathcal{L}_{\text{reconstructed}} \tag{5}$$

TABLE 2 Experimental results comparison with SOTA methods.

Method type	Method	Backbone	Params. (M)	Top-1 Acc (%)		
				Datasets		
				AID	NWPU45	UCM
†	SimCLR	Resnet50	21	61.37	65.16	52.89
	MoCo v3	ViT-small	22	63.67	67.11	50.38
△	MAE	ViT-base	22	65.33	71.56	50.63
	SimMIM	Swin-base	88	**69.49**	**73.01**	54.89
♦	DINO	ViT-small	22	64.26	69.03	53.63
★	GSC-MIM	ViT-small	22	**65.38**	**69.54**	**54.14**
	GSC-MIM	ViT-base	86	68.98	72.47	**55.89**
	GSC-MIM	ViT-large	304	72.03	76.04	None

†Contrastive learning with ResNet/Vision Transformer backbone.
△MIM based methods.
♦Self-distillation based method.
★The proposed GSC-MIM.
The bold values represent the best-performing methods for similar amounts of network parameters.

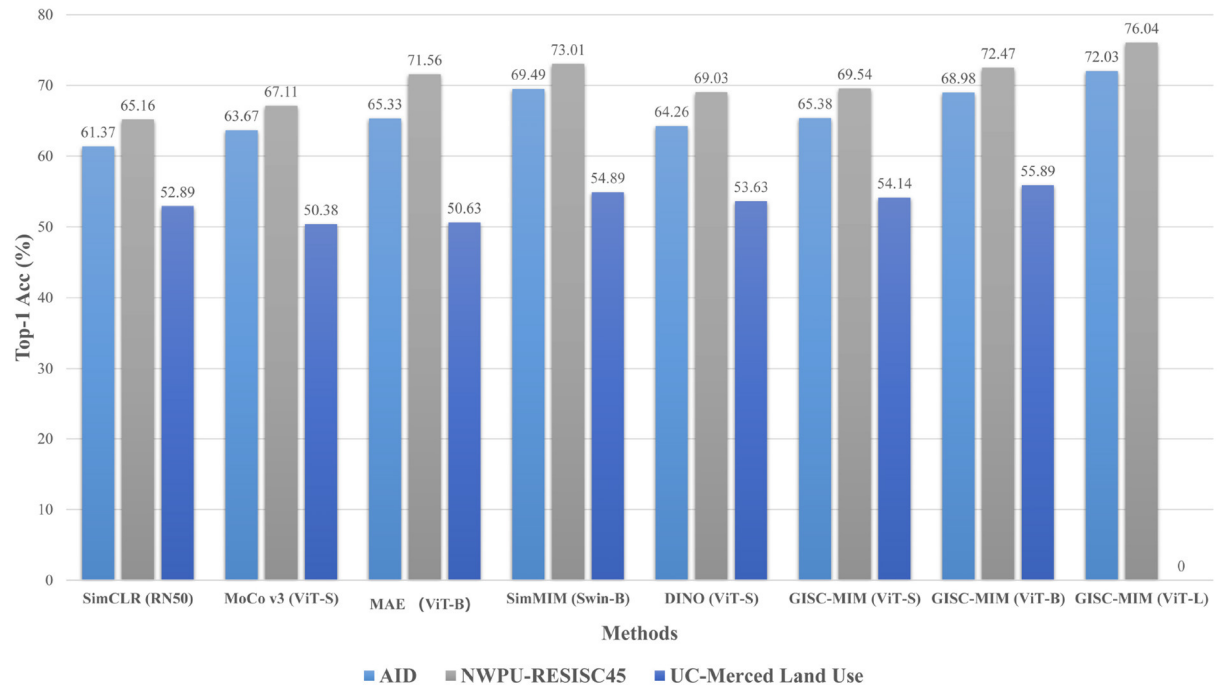

FIGURE 5
Bar chart of experimental results comparison with SOTA methods.

where λ_1, λ_2 represent the loss coefficients that balance the importance of [*CLS*] losses and reconstructed losses.

The backbone f adopted by GSC-MIM is vision transformer (ViT) encoder (Dosovitskiy et al., 2020), shown in Figure 4. We also evaluated the different amounts of backbone parameters of ViT-S/16, ViT-B/16, and ViT-L/16. The projection head h of GSC-MIM is set as 3-layers MLP, which is proven optimal in both Caron et al. (2021) and Zhou et al. (2021). Moreover, to further borrow the capability of semantic abstraction obtained by self-distillation on [*CLS*] tokens, we share the projection head parameters for both patch tokens and [*CLS*] tokens. The scene classification head takes [*CLS*] token as input and is implemented by a MLP projection head with three hidden layers at pre-training time and by a single linear layer at fine-tuning time.

3. Experimentation and results discussion

Three public remote sensing datasets are used for thorough tests to assess the efficacy of the proposed GSC-MIM approach, including parameter analysis, an experimental comparison of several algorithms, and visualization analysis.

TABLE 3 Experimental results comparison with patch size as hyperparameter.

Method	Patch size	Top-1 Acc (%)
GSC-MIM (ViT-Base/p16)	32	66.91
	16	72.47
	8	77.60

3.1. Experimental setup

(1) **Datasets description.** In this paper, the Aerial Image Dataset (AID) (Xia et al., 2017), the NWPU-RESISC45 Dataset (NWPU45) (Cheng et al., 2017), and the UC-Merced Land Use Dataset (UCM) (Yang and Newsam, 2010) are selected to conduct the experiments. With high image size and spatial resolution diversity, these three datasets are challenging for RSI scene classification task. More detailed information about datasets is shown in Table 1.

(2) **Hardware and software environment.** All experiments are carried out on NVIDIA A100 Tensor Core GPU with a software environment of Python 3.7 with PyTorch performing on the Ubuntu system.

(3) **Architecture parameter setup.** For the backbone f, we inherit the ViT models' parameters from Caron et al. (2021) and train the network without loading the pretrained weight. For ViTs, /16 denotes the patch size being 16. We also evaluate patch sizes 8 and 32 in the following experiment. For the projection head h, a $3-layerMLP$ with $l2-normalized$ bottleneck is chosen to generate representations. Since we share h for both patch tokens and [*CLS*] tokens, they all get the output dimension of 8,192.

(4) **Data processing and evaluation metrics.** The proposed GSC-MIM uses multi-scale crop, random flip and random rotation as two different types of data augmentation methodologies to generate the distorted views of t_1, t_2 and s_1, s_2. For the three datasets, 80% of each category is selected as the training set and the remaining 20% served as the testing set. After pretraining, we perform the linear evaluation using 1% of each category in the training set to elaborate on the effectiveness of the proposed GSC-MIM among the compared methods. The linear evaluation metric is the same as the works in Chen et al. (2020); Caron et al. (2021), and Zhou et al. (2021). Furthermore, the Top-1 Accuracy (Top-1 Acc) with visualized histogram is employed to illustrate the classification performance of the proposed method on the three datasets.

(5) **Parameter optimization setup.** We by default pre-train GSC-MIM on the above training datasets and set 16 as default patch size for IPC-Masking. The batch size is set to 64 for ViT-S and ViT-B, and 16 for ViT-L due to the limitation of GPU. For both teacher and student networks, the AdamW (Loshchilov and Hutter, 2017) is employed as the optimizer for better convergence, the corresponding momentum is set to 0.996, and the weight decay is 0.004. For the student network, we employ random MIM, with prediction ratio r uniformly chosen from the range $[0.1, 0.5]$ with a probability of 0.5 and set as 0 with a probability of 0.5. Moreover, we set λ_1 and λ_2 equal to 1 in $\mathcal{L}_{all}$, i.e., we sum $\mathcal{L}_{[\mathrm{CLS}]}$ and $\mathcal{L}_{\mathrm{reconstructed}}$ up without scaling. Finally, we train and evaluate the GSC-MIM for 300 epochs as default. During the first 10 epochs, the learning rate is linearly ramped up to its base value scaled with the entire batch size: $lr = 5e^{-4} \times batch\_size/256$.

3.2. Results and discussion

3.2.1. Performance results

Various state-of-the-art (SOTA) approaches are presented to compare with GSC-MIM on the three datasets to highlight the advantages of the proposed method. The results obtained from the linear evaluation are summarized in Table 2. The training and testing ratios for all listed methods remain the same for a fair comparison. Furthermore, all methods are divided into four categories to show the results effectively, including classical contrastive learning approaches (†) (Chen et al., 2020; He et al., 2020), classical MIM-based methods (△) (He et al., 2022; Xie et al., 2022), self-distillation method (⧫) (Caron et al., 2021), and the proposed GSC-MIM (★). We also provide a bar chart for a more intuitive illustration, as shown in Figure 5. Several analysis can be drawn from these results.

Firstly, the results of GSC-MIM (ViT-S) are significant compared to the conventional contrastive learning approaches with roughly the same amount of parameters such as SimCLR (ResNet50) (Chen et al., 2020) and MoCo v3 (ViT-S) (Chen et al., 2021). These observations suggest that the proposed MIM-based method benefits from the internal signal generated from the image itself compared to the handcraft supervision signal such as data distortion.

Secondly, we compared the classical MIM-based methods such as MAE (ViT-B) (He et al., 2022) and SimMIM (Swin-B) (Xie et al., 2022). The proposed GSC-MIM (ViT-B) outperformed MAE by 0.91–5.26%. The improvement for Top-1 Acc indicates that the global semantic integrated strategy can help the model better acquire the input image's global long-dependence features and achieve higher classification results. Concerning no significant increments found between GSC-MIM

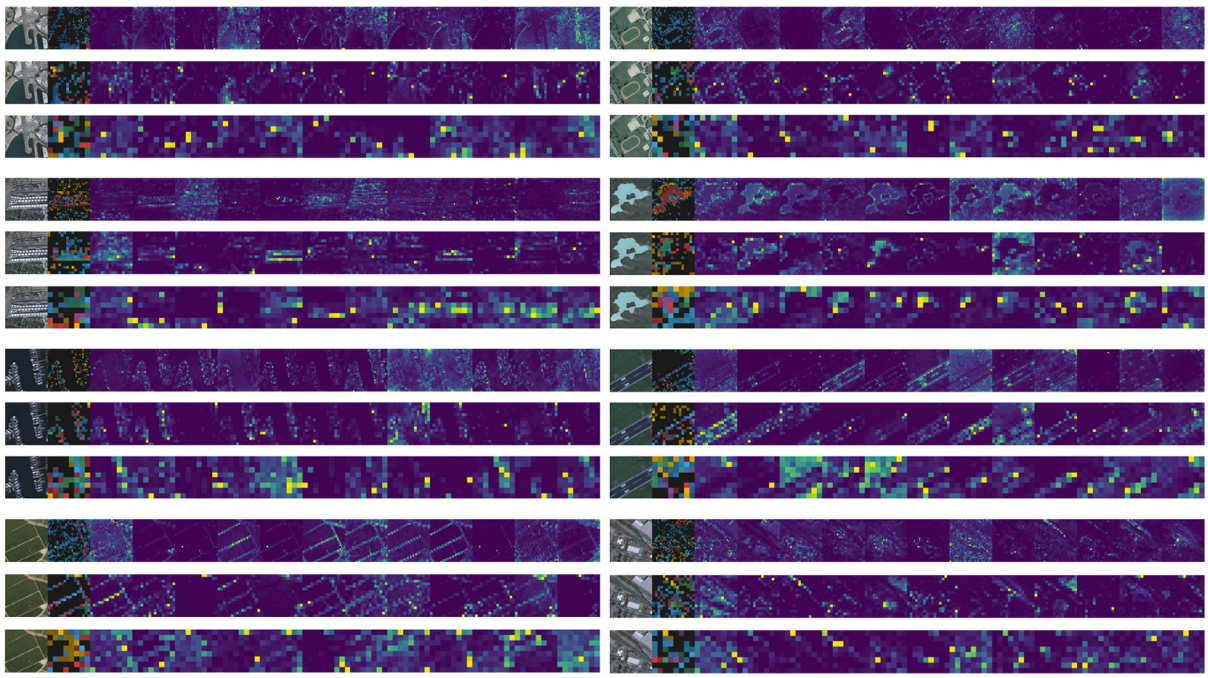

FIGURE 6
Attention map visualization diagrams with patch size of 8 **(top)**, 16 **(middle)**, and 32 **(bottom)**.

and SimMIM, we suppose it is due to the advancement of Swin Transformer (Liu Z. et al., 2021) as a backbone network.

Thirdly, another observation from the results is the accuracy increment compared to DINO(ViT-B) (Caron et al., 2021), which is attributed to the information preserved complementary masking strategy of GSC-MIM that prevents the network from losing detailed information. We will provide visualization results in subsection 3.2.3. to further support our suggestion.

It also should be mentioned that under the scenario of limited training samples of UCM dataset, it is easy to overfit for the proposed GSC-MIM based on large amounts of parameters backbone such as ViT-L (Params. 304 M), and leads to unauthentic results.

3.2.2. Patch size as hyperparameter

Since small-scale features usually occupy relatively small areas in the RSI, the patch size theoretically affects the ability of the model to capture the small-scale features and the detailed information in the RSI. To further evaluate the impact of different patch sizes on the network effects, we conduct patch-size-specific experiments on the NWPU45 dataset. All experiments follow the default hyperparameters excluding patch size. The results are shown in Table 3. It is a rather evident tendency that classification accuracy is increasing with the decreasing of patch size. It is also worth noting that the training time of the network also increases substantially. This result is probably caused by two reasons:

- A small patch size will force the model to reconstruct more detailed information and thus learn fine-grained features of RSI.
- Smaller patch size is more suitable for small-scale objects. When very small-scale objects, such as cars, are in remote sensing images, only a smaller patch size can retain adequate information.

We also provide attention map visualization diagrams to support our statements, shown as Figure 6.

3.2.3. Visualization analysis

This section uses two visualization methods to interpret the network effects. We also perform the same visualization experiment for DINO (ViT-B) (Caron et al., 2021) for a comparison. All the images are randomly selected from NWPU45 Dataset. As mentioned before, the mask of critical patches will cause information loss, so we proposed an information-preserved complementary masking strategy to mitigate this negative impact. To illustrate the effectiveness, we visualize the attention map of the last layer of each attention head in ViT. The results are shown in Figure 7. Note that the colorful images are the visualization of [*CLS*] tokens. It can be

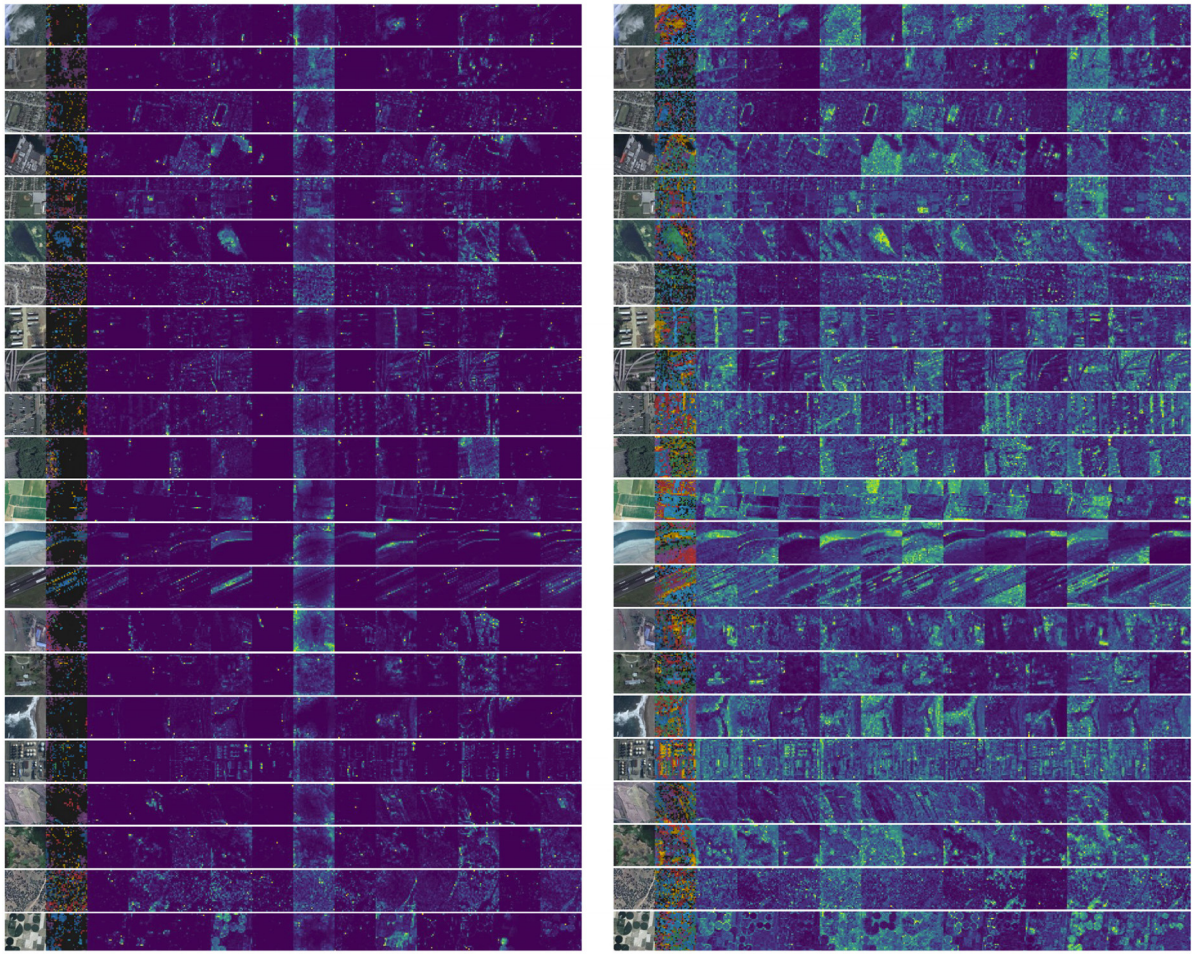

FIGURE 7
Attention map visualization for GSC-MIM **(left)** and DINO **(right)**.

seen that GSC-MIM shows visually more vital ability to detect multi-scale land objects and can separate different land objects or different parts of one land object apart, while DINO tends to capture the ambiguous semantic boundaries of land objects.

Since scene labels are highly related to key areas and objects for RSI classification task, and accurately locating the attention regions is conducive to improving classification results, the proposed GSC-MIM integrates the input image's global semantic information into the network. To better reflect the key attention regions, we generated the energy map of the selected samples. As shown in Figure 8, it is evident that the GSC-MIM-generated distribution of attention regions is accurate than DINO model.

3.2.4. Discussion and future work

In this paper, we propose GSC-MIM, aiming to address two main problems encountered when applying MIM-based generative self-supervised learning to RSI scene classification tasks: the small-scale information absence problem and the global semantic information neglect problem, for which we propose the information-preserved complementary masking strategy (IPC-Masking) and semantic distillation strategy (GSD). Experimental results show that our model can obtain up to 5.26% accuracy improvement on the RSI scene classification task compared to the benchmark approaches.

We find that by introducing IPC-Masking, the network's ability to capture small-scale and fine-grained features is improved, and this ability is further enhanced as the patch size is reduced. Intuitively, the effectiveness of the network shows an inverse relationship with the patch size. However, when the patch size is excessively small (taking a single pixel as an example), the contained semantic information is lost; when the patch size is excessively large (assuming a whole input image as an example), the MIM mechanism loses its ability to work, so how to find the appropriate patch size for different RSI datasets is a problem worth further investigation.

Moreover, our experiments demonstrate that the model has accurate responses for the scene semantically meaningful

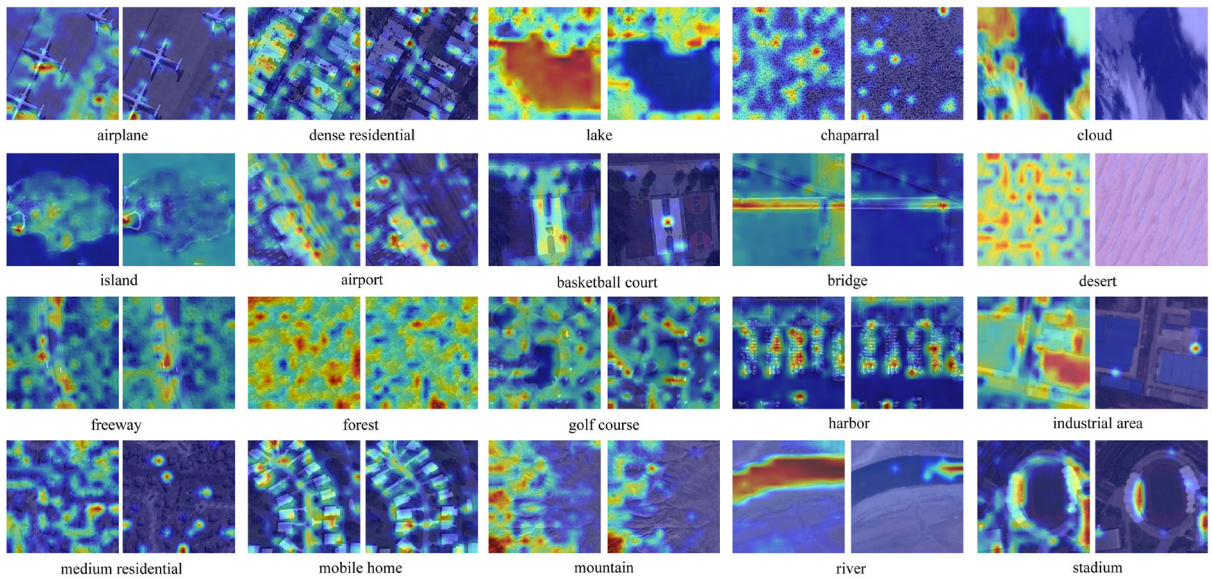

FIGURE 8
Energy map visualization for GSC-MIM **(left)** and DINO **(right)**.

regions when global information is introduced as a supervised signal. This proves the significance of global knowledge for using MIM in RSI scene classification tasks. Due to the specificity of the RSI imaging view, the critical information that determines the semantics of RSI scenes is often mixed with the background. When the global scene information of the field is not well-defined, it might not bring significant gain. Further clustering of the obtained global semantic futures to get more accurate scene information is a possible beneficial direction for future work.

4. Conclusion

This paper proposed a global semantic integrated self-distilled complementary masked image model, called GSC-MIM, for RSI scene classification. In GSC-MIM, the information preserved complementary masking strategy is proposed to prevent the information loss of local patches. The global semantic distillation strategy is employed to integrate the input image's global semantic information into the network and achieve label-free training. Experiments on the AID, NWPU-RESISC45, and UCM datasets indicate that the proposed GSC-MIM can better catch features of multi-scale land objects and achieves competitive classification accuracies.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

XW: conceptualization, methodology, software, and writing—original draft preparation. YZ: supervision. ZZ: software and validation. QL: data curation. JY: writing—reviewing and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (Grants 41871364 and 42171376) and supported by the High Performance Computing Platform of Central South University.

Acknowledgments

We thank reviewers for valuable comments on the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Akiva, P., Purri, M., and Leotta, M. (2022). "Self-supervised material and texture representation learning for remote sensing tasks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (New Orleans, LA), 8203–8215. doi: 10.1109/CVPR52688.2022.00803

Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., et al. (2021). "Geography-aware self-supervised learning," in *Proceedings of the IEEE International Conference on Computer Vision* (Montreal, QC), 10181–10190. doi: 10.1109/ICCV48922.2021.01002

Bao, H., Dong, L., and Wei, F. (2021). BEiT: BERT pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE International Conference on Computer Vision* (Montreal, QC), 9630–9640. doi: 10.1109/ICCV48922.2021.00951

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). "A simple framework for contrastive learning of visual representations," in *Proceedings of the IEEE International Conference on Machine Learning*, Vol. 119, 1597–1607.

Chen, X., Xie, S., and He, K. (2021). "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE International Conference on Computer Vision* (Montreal, QC), 9640–9649. doi: 10.1109/ICCV48922.2021.00950

Cheng, G., Han, J., and Lu, X. (2017). Remote sensing image scene classification: benchmark and state of the art. *Proc. IEEE* 105, 1865–1883. doi: 10.1109/JPROC.2017.2675998

Cino, L., Mazzeo, P. L., and Distante, C. (2022). "Comparison of different supervised and self-supervised learning techniques in skin disease classification," in *IEEE International Conference on Image Information Processing* (Lecce: Springer), 77–88. doi: 10.1007/978-3-031-06427-2_7

Ding, J., Xie, E., Xu, H., Jiang, C., Li, Z., Luo, P., and Xia, G.-S. (2021). Unsupervised pretraining for object detection by patch reidentification. *arXiv preprint arXiv:2103.04814*. doi: 10.1109/TPAMI.2022.3164911

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (New Orleans, LA), 16000–16009. doi: 10.1109/CVPR52688.2022.01553

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 9729–9738. doi: 10.1109/CVPR42600.2020.00975

Heidler, K., Mou, L., Hu, D., Jin, P., Li, G., Gan, C., et al. (2021). Self-supervised audiovisual representation learning for remote sensing data. *arXiv preprint arXiv:2108.00688*.

Li, H., Li, Y., Zhang, G., Liu, R., Huang, H., Zhu, Q., et al. (2022). Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–14. doi: 10.1109/TGRS.2022.3147513

Li, W., Chen, H., and Shi, Z. (2021a). Semantic segmentation of remote sensing images with self-supervised multitask representation learning. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* 14, 6438–6450. doi: 10.1109/JSTARS.2021.3090418

Li, W., Chen, K., Chen, H., and Shi, Z. (2021b). Geographical knowledge-driven representation learning for remote sensing images. *IEEE Trans. Geosci. Remote Sensors* 60, 1–16. doi: 10.1109/TGRS.2021.3115569

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* 1–20. doi: 10.1109/TKDE.2021.3090866

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of International Conference on Computer Vision* (Montreal, QC), 10012–10022. doi: 10.1109/ICCV48922.2021.00986

Loshchilov, I., and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, W., Tao, C., Li, H., Qi, J., and Li, Y. (2022). A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data. *Remote Sens. Environ.* 270, 112830. doi: 10.1016/j.rse.2021.112830

Manas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., and Rodriguez, P. (2021). "Seasonal contrast: unsupervised pre-training from uncurated remote sensing data," in *Proceedings of International Conference on Computer Vision* (Montreal, QC), 9414–9423. doi: 10.1109/ICCV48922.2021.00928

Tao, C., Qi, J., Lu, W., Wang, H., and Li, H. (2020). Remote sensing image scene classification with self-supervised paradigm under limited labeled samples. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2020.3038420

Wang, X., Zhu, J., Yan, Z., Zhang, Z., Zhang, Y., Chen, Y., et al. (2022). LaST: label-free self-distillation contrastive learning with transformer architecture for remote sensing image scene classification. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2022.3185088

Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., and Zhu, X. X. (2022). Self-supervised learning in remote sensing: a review. *arXiv preprint arXiv:2206.13188*. doi: 10.1109/MGRS.2022.3198244

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., et al. (2017). AID: a benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* 55, 3965–3981. doi: 10.1109/TGRS.2017.2685945

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., et al. (2022). "SimMIM: a simple framework for masked image modeling," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (New Orleans, LA), 9653–9663. doi: 10.1109/CVPR52688.2022.00943

Yang, Y., and Newsam, S. (2010). "Bag-of-visual-words and spatial extensions for land-use classification," in *ACM SIGSPATIAL GIS* (New York, NY), 270–279. doi: 10.1145/1869790.1869829

Zhang, Z., Wang, X., Mei, X., Tao, C., and Li, H. (2022). FALSE: false negative samples aware contrastive learning for semantic segmentation of high-resolution remote sensing image. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. doi: 10.1109/LGRS.2022.3222836

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., et al. (2021). iBOT: image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.

Zhu, Q., Lei, Y., Sun, X., Guan, Q., Zhong, Y., Zhang, L., et al. (2022a). Knowledge-guided land pattern depiction for urban land use mapping: a case study of Chinese cities. *Remote Sens. Environ.* 272, 112916. doi: 10.1016/j.rse.2022.112916

Zhu, Q., Sun, Y., Guan, Q., Wang, L., and Lin, W. (2022b). A weakly pseudo-supervised decorrelated subdomain adaptation framework for cross-domain land-use classification. *IEEE Trans. Geosci. Remote Sens.* 60, 1–13. doi: 10.1109/TGRS.2022.3170335