MFDA: Unified Multi-Task Architecture for Cross-Scene Sea Ice Classification

Yuhan Chen, *Student Member, IEEE*, Qingyun Yan, *Member, IEEE*, Igor Bashmachnikov, Kaijian Huang, Fangru Mu, Zixuan Zhang, Minghuan Xu, and Jiechen Zhao

***Abstract*— Although the extensive research has been conducted on retrieving sea ice variables from synthetic aperture radar (SAR) and multimodal remote sensing data, cross-scene retrieval using regional training models remains a significant challenge. Previous studies have employed multi-task learning (MTL) but have not sufficiently explored the interplay between network architectures and multi-task performance. Moreover, self-supervised learning (SSL) has shown promise in improving tasks with limited training samples, though its potential in sea ice variable retrieval requires further study. To address the challenge of cross-scene retrieval of sea ice variables, we introduce a novel and effective method called multimodal fusion domain adaptation (MFDA), which combines three key strategies: 1) employ SSL methods for multimodal data to pretrain the model, improving its noise sensitivity and promoting a hierarchical understanding of multimodality; 2) propose a unified convolutional and Transformer-based data fusion architecture to enhance the integration of multimodal data and improve semantic understanding; and 3) incorporate a domain adaptation module between the multimodal encoder and the multi-task decoding predictor to facilitate the model's understanding of the semantic gaps between different regional environments. The performance of the proposed MFDA has been extensively evaluated on the Ai4Arctic dataset. The experimental results demonstrate that MFDA achieves superior performance compared with other state-of-the-art (SOTA) sea ice classification approaches for the task of cross-scene sea ice retrieval. The code will be made available at: https://github.com/yuweikong/MFDA.**

***Index Terms*— Arctic sea ice, domain adaptation, masked image modeling (MIM), multi-task learning (MTL), self-supervised learning (SSL), synthetic aperture radar (SAR).**

Received 27 June 2024; revised 24 September 2024; accepted 1 November 2024. Date of publication 7 November 2024; date of current version 19 November 2024. The work of Jiechen Zhao was supported in part by the National Natural Science Foundation of China under Grant 42276251, Grant 42211530033, and Grant 41876212; and in part by the Taishan Scholars Program. The work of Kaijian Huang was supported in part by the Foundation of the Education Department of Guangdong Province, China, under Grant 2021KQNCX091. *(Corresponding author: Jiechen Zhao.)*

Yuhan Chen, Fangru Mu, Zixuan Zhang, Minghuan Xu, and Jiechen Zhao are with Qingdao Innovation and Development Base (Centre), Harbin Engineering University, Qingdao 266000, China (e-mail: zhaojiechen@hrbeu.edu.cn).

Qingyun Yan is with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China.

Igor Bashmachnikov is with the Nansen International Environmental and Remote Sensing Centre, 199034 Saint Petersburg, Russia, and also with the Institute of Earth Science, Saint Petersburg State University, 199034 Saint Petersburg, Russia.

Kaijian Huang is with the School of Electronic Information and Electrical Engineering, Huizhou University, Huizhou 516007, China.

Digital Object Identifier 10.1109/TGRS.2024.3491190

I. Introduction

SEA ice, as a critical component of the cryosphere, exhibits significant variations across diverse spatial and temporal scales [1]. Accurately characterizing the spatial and temporal dynamics of sea ice is of paramount importance for understanding the climate system [2], forecasting climate change [3], planning polar navigation routes [4], and ensuring maritime safety [5]. Remote sensing has emerged as a pivotal approach for rapidly obtaining large-scale, high-quality Earth observation data, thus significantly advancing Earth observation capabilities [6], [7]. Remote sensing data can be leveraged to automatically extract and generate sea ice maps, which serves as an effective means for monitoring sea ice dynamics and provides crucial support for research and applications in related domains [8], [9], [10]. However, traditional expert system-centric methods have faced limitations in meeting the evolving needs of Earth observation data in the era of remote sensing big data, particularly when dealing with complex scenes [11], [12]. Deep learning offers a viable solution to intelligently extract valuable knowledge from diverse Earth observation data, enabling enhanced understanding and monitoring of polar environments [13].

Remotely sensed data acquired from satellite platforms frequently exhibit temporal discontinuities, variability in spatial resolution, and diminished signal-to-noise ratios, particularly within polar geographic regions [14], [15]. Concurrently, the reference data available for these polar environments are often limited and may originate from a single, localized area, posing significant challenges for the training and deployment of deep learning models in such contexts [16]. Moreover, polar climate systems are highly complex and dynamic, influenced by climate change, anthropogenic activities, and various other factors [17], [18]. Consequently, deep learning models trained solely on constrained datasets may struggle to fully capture the multifaceted complexities inherent to these regions [19], [20].

Advanced deep learning models have been successfully applied to a range of remote sensing and Earth science applications. Recent advancements in the field of sea ice parameter retrieval have garnered significant attention. For example, Ren et al. [21] introduced a dual-attention U-Net for the classification of sea ice and open water. In addition, Song et al. [22] proposed a joint model that integrates spatial and temporal features for sea ice classification within

the Canadian Ice Service ice charts area. Furthermore, De Gelis et al. [23] highlighted the potential of fully convolutional networks (FCNs) for the automatic estimation of sea ice concentration (SIC) using synthetic aperture radar (SAR). In a related study, Huang et al. [24] developed a dual-branch encoder U-Net (DBU-Net) for sea ice-type classification in SAR images from the Beaufort Sea, demonstrating the advantages of end-to-end models in this application. These models often perform exceptionally well in homogeneous regional scenes, where the underlying characteristics, spatial distributions, and typologies are relatively consistent [25]. Nevertheless, the capacity of these sophisticated deep learning methodologies to effectively tackle the challenges inherent in the analysis of diverse spatiotemporal data distributions and the heterogeneity of regional environments is still constrained [26], [27].

Deep learning architectures have demonstrated impressive performance in polar environments characterized by relatively homogeneous regional conditions [7], [28], [29]. However, progress in developing deep learning approaches for broader sea ice classification tasks across diverse polar scenes has been relatively slow [30], [31], [32]. This can be primarily attributed to two key challenges. First, the paucity of large-scale, high-quality, multimodal remote sensing datasets restricts a more comprehensive understanding of cross-domain polar environments. Consequently, the availability of additional high-fidelity datasets would be instrumental in facilitating a more holistic study of sea ice conditions across varied geographical regions [9], [33], [34], [35], [36]. Second, current methodological approaches tend to prioritize optimization for single, localized sea areas rather than enhancing the generalization capabilities of the models, particularly across different latitudinal and longitudinal domains. To better adapt to the unique characteristics of diverse polar regions, it is imperative to place greater emphasis on developing models with robust generalization abilities that can effectively handle a wide range of environmental scenes [9], [33].

The recently released AI4SeaIce public dataset [37] addresses the challenge of limited high-quality multimodal remote sensing data in polar regions, thereby opening new avenues for research in this field. This dataset has facilitated a range of studies, contributing to advancements in SIC [9], [33], [34], [35], [36], the floe size (FLOE) [9], and the stage of development (SOD) [9], [38]. The concept of cross-scene classification using multimodal data has been extensively studied and applied in the field of computer vision [39], [40], [41], [42], [43], [44], providing a promising direction for adapting sea ice monitoring models across diverse polar environments.

Using the AI4SeaIce public dataset, Chen et al. [9] introduced a multi-task learning (MTL) framework and demonstrated significant results, highlighting the potential of multi-task architectures for comprehensive sea ice monitoring in diverse polar environments. Compared with traditional single-task models, MTL may offer enhanced solutions by implementing implicit constraints and shared representations among different but related tasks. While preliminary results have been achieved in sea ice classification tasks, further optimization of the architecture is warranted. Recently, Xiong et al. [45] investigated the effects of multimodal data fusion on model performance when integrating multiple tasks through MTL methods. In addition, the methods of parameter sharing across multiple tasks merit further discussion [46], [47].

Although the aforementioned solution has yielded some positive results, it remains challenging to apply in cross-scene tasks, where reference data are limited. The rise of masked image modeling (MIM) technology has led to a widespread adoption of self-supervised learning (SSL) methods in the remote sensing field [48], [49], [50], [51]. These methods involve establishing auxiliary tasks and learning features from a substantial amount of unlabeled data to create an initial feature extraction model. Subsequently, this initial model is fine-tuned for specific downstream tasks. MIM has been extensively applied in remote sensing tasks, demonstrating effective performance even when reference samples are limited. For instance, Cao et al. [52] introduced an MIM approach that incorporates contrast loss, utilizing pretraining methods to capture the underlying representations of images, thereby showcasing the robust feature extraction capabilities of the MIM pretraining model. In addition, Lin et al. [53] developed a spatial–spectral masked autoencoder (SS-MAE) tailored for multimodal data, which employs a lightweight convolutional model during the training phase to enhance local feature modeling. The SS-MAE framework underscores the potential of MIM in multimodal tasks and emphasizes the importance of integrating both local and global features.

However, some limitations persist when directly applying MIM [54] to multimodal remote sensing images for semantic segmentation. First, predictions based on raw pixel values contain substantial redundant spatial domain information and typically exhibit low-order statistical features, which may hinder the model's ability to capture high-level semantic information [55]. Second, the inherent differences between observed data and natural RGB images make it challenging for the original single-layer decoder architecture to perform complex modeling of multimodal image features, thereby impeding the convergence of the model [48]. Finally, directly using a single decoder to model deep features may cause dense and small objects to be lost in multimodal images, resulting in incomplete semantic information and challenges in image reconstruction. Recently, Chen and Yan [55] confirmed that using the low-frequency information of remote sensing images as the target for training MIM models can accelerate convergence and reduce model noise. Moreover, Wang et al. [56] revealed the effectiveness of MIM using multilevel constraints, which can accelerate the convergence of the model and promote its multiscale semantic understanding of the input data. These methods have been proven to be straightforward and convenient for multimodal tasks.

In this article, we investigate strategies to enhance model performance for sea ice parameter retrieval tasks under cross-scene conditions. First, we utilize the AI4SeaIce public dataset to create a cross-scene task dataset for sea ice classification, aiming to improve the applicability of Arctic sea ice classification across different scenes. Second, we propose a multiscale low-pass filter mask image model (MLFMIM)

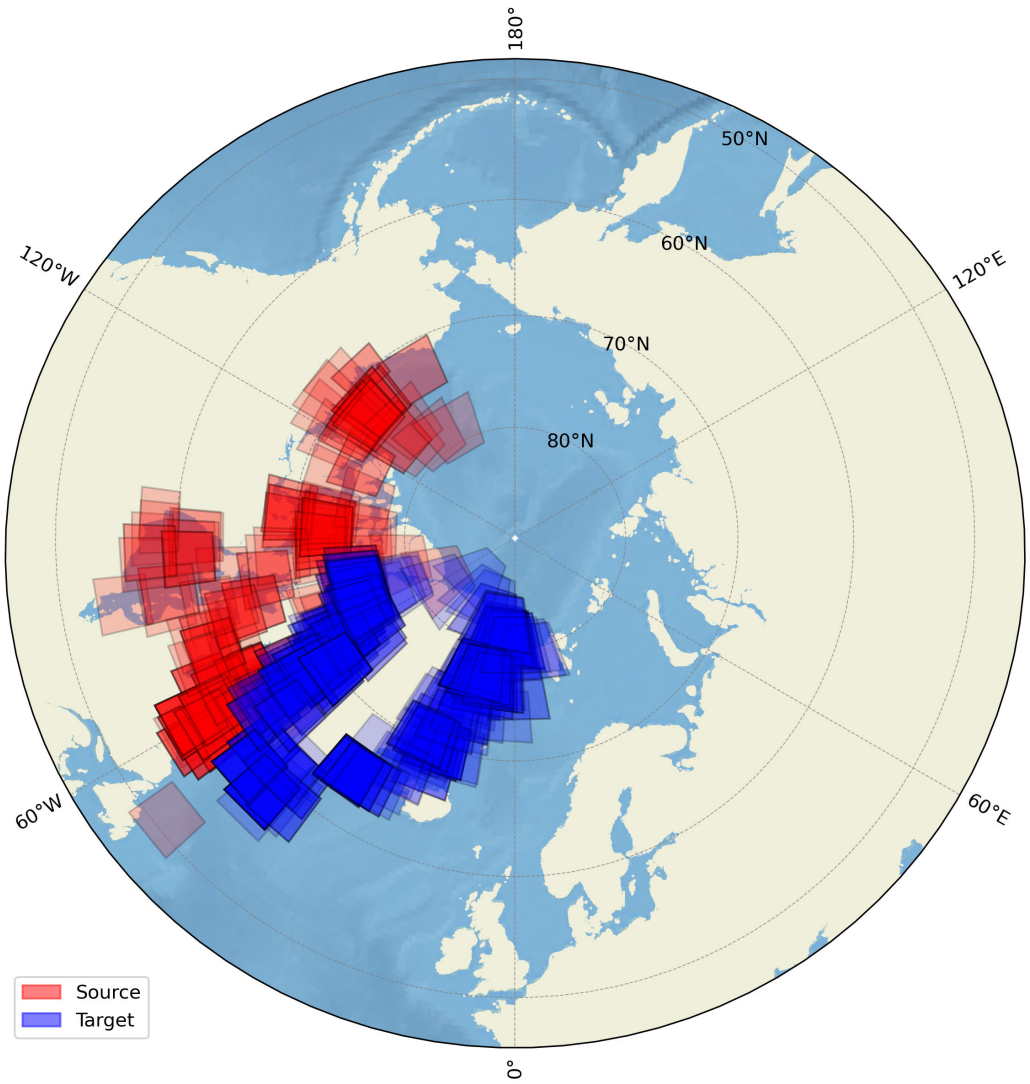

Fig. 1. Dataset region visualization, source, and target in the Arctic region.

for pretraining multimodal data fusion, which compels the encoder to extract robust features from multimodal data across various scenes. Finally, we design a unified architecture termed multimodal fusion domain adaptation (MFDA) to enhance the effectiveness of MTL in sea ice parameter retrieval and facilitate the domain adaptation process. This architecture integrates convolutional and Transformer components to create a multimodal encoder, utilizing the aforementioned MLFMIM for weight initialization. In the decoder phase, we implement the domain adaptive conversion on the generated hierarchical features to address discrepancies between the source and target domain features.

In summary, the main contributions of our work are as follows.

1) We propose a novel MIM architecture, termed MLFMIM, designed for self-supervised pretraining of multimodal data. Unlike previous methods, MLFMIM utilizes low-pass filtered signals as training targets to reduce the influence of redundant spatial information and noise on model performance. In addition, we introduce a local multiscale reconstruction method that provides explicit guidance for multiple deep layers of the model, thereby effectively capturing multiscale features.
2) Based on the characteristics of multimodal data, we propose the MFDA for sea ice parameter retrieval using limited regionally labeled samples. This architecture accounts for the correlation among data at different resolutions and constructs a relatively independent dual-branch model for feature fusion. It integrates convolutional and Transformer components within distinct branches, allowing the model to effectively focus on both local and global features.
3) We incorporated a multi-task domain adaptation module between the encoder and decoder of the MFDA to address discrepancies between the source domain and the target domain.

II. Datasets

This study leverages the Ai4Arctic remote sensing dataset [37] as the primary research material. The spatial distribution of the dataset is shown in Fig. 1. The dataset encompasses two key geospatial domains: the center of the Canadian

TABLE I
DATA VARIABLES AND RESOLUTIONS

Variable Description	Variable Abbreviation	Original Resolution	Resolution Training
Sentinel-1 EW Data	HH	40 × 40 m	0.4 × 0.4 km
	HV	40 × 40 m	0.4 × 0.4 km
	Lat	-	1.6 × 1.6 km
	Lon	-	1.6 × 1.6 km
AMSR2 Data	18.7GHz H	14 × 22 km	1.6 × 1.6 km
	18.7GHz V	14 × 22 km	1.6 × 1.6 km
	36.5GHz H	7 × 12 km	1.6 × 1.6 km
	36.5GHz V	7 × 12 km	1.6 × 1.6 km
ERA5 Data	10m windspeed	0.25 × 0.25°	1.6 × 1.6 km
	2m air temperature	0.25 × 0.25°	1.6 × 1.6 km
	total column water vapor	0.25 × 0.25°	1.6 × 1.6 km
	total column cloud liquid water	0.25 × 0.25°	1.6 × 1.6 km
Time Data	scene acquisition month	-	1.6 × 1.6 km
Reference Data	SIC	40 × 40 m	0.4 × 0.4 km
	SOD	40 × 40 m	0.4 × 0.4 km
	FLOE	40 × 40 m	0.4 × 0.4 km

Ice Sheet, which serves as the source domain, and the center of the Greenland Ice Sheet, which represents the target domain. The source domain subset contains 197 complete data files, while the target domain subset comprises 315 such files.

In the workflow for model development and evaluation, we initially employ an SSL strategy, leveraging the complete Ai4Arctic dataset, encompassing both source and target domain files, to pretrain the model in the absence of any labeled data. Subsequently, during the supervised learning phase, we train the model exclusively on the 197 files from the source domain and evaluate its performance on the 315 files from the target domain.

A. Multimodal Input Data

Each file encompasses a comprehensive set of remote sensing and ancillary parameters, including: dual-polarization Sentinel-1 Extra Wide (EW) imagery, The Advanced Microwave Scanning Radiometer 2 (AMSR2) passive microwave radiometer measurements, numerical weather prediction (NWP) variables derived from the ECMWF Reanalysis v5 (ERA5) reanalysis dataset, as well as ice chart data conforming to the World Meteorological Organization (WMO) sea ice classification schema, provided by either the Greenland Ice Service or the Canadian Ice Service. The resolution of each data is shown in Table I. It is important to note that, in selecting input data, we opted to utilize a subset of the multimodal data sources available in the Ai4Arctic dataset. Our primary focus is on evaluating the domain adaptability of the model, rather than investigating the intricate relationships between different input parameters. Accordingly, we have adopted the optimal input data configuration as recommended in the prior literature [9], [57]. For the Sentinel-1 EW data, AMSR2 data, ERA5 data, and temporal data, pixel values are normalized within the range of $[-1, 1]$. Missing values (NaN) in the SAR are replaced with -1. In addition, the data for each patch's longitude and latitude information are normalized to the range of [0, 1] to prevent any spatial information leakage. The normalized longitude and latitude data are integrated into the model, analogous to the relative position encoding described in [58] within the Transformer architecture, thereby providing potential advantages to the model.

The specific data are described as follows.

1) Sentinel-1 EW Data: The dataset comprises dual-polarimetric Sentinel-1 (available at https://dataspace.copernicus.eu/) SAR backscatter observations, including measurements in both the horizontal–horizontal (HH) and horizontal–vertical (HV) polarizations. Accompanying this radar data are the corresponding latitude and longitude grid information, organized in a 21-by-21 matrix. Crucially, no geographic projection or orthorectification has been applied to the data. The resolution of HH and HV polarization images is 40 m.

2) AMSR2 Data: The dataset consists of AMSR2 brightness temperature data, which has been resampled to match the geometry of Sentinel-1. The pixel spacing of the resampled data is 2 × 2 km. The AMSR2 data are obtained from JAXA (available at: https://earth.jaxa.jp/en/data/index.html). Specifically, it includes dual-pol (dual-polarization) AMSR2 brightness temperature data at 18.7 and 36.5 GHz.

3) ERA5 Data: The dataset comprises ERA5 weather forecast (NWP) parameters that have been resampled to align with the geometry of Sentinel-1, (available at: https://cds.climate.copernicus.eu/). The resampling process employed Gaussian weighted interpolation, resulting in a pixel resolution of 2 × 2 km. The included parameters are 10-m wind speed, 2-m air temperature, total column water vapor, and total column cloud liquid water.

4) Time Data: The temporal information associated with each pixel corresponds to the acquisition month of the SAR scene, denoted by a digital code such as "1" for January. The image dimensions of the SAR data match those of the AMSR2 dataset.

B. Reference Data

The sea ice chart parameters include the SIC, SOD, and FLOE.

1) SIC: The SIC parameter represents the areal fraction of a given spatial domain that is covered by sea ice, typically expressed as a percentage ranging from 0% (open water) to 100% (complete ice cover) in discrete 10% increments.

2) SOD: The SOD parameter can be regarded as a categorical representation of sea ice type, which serves as a

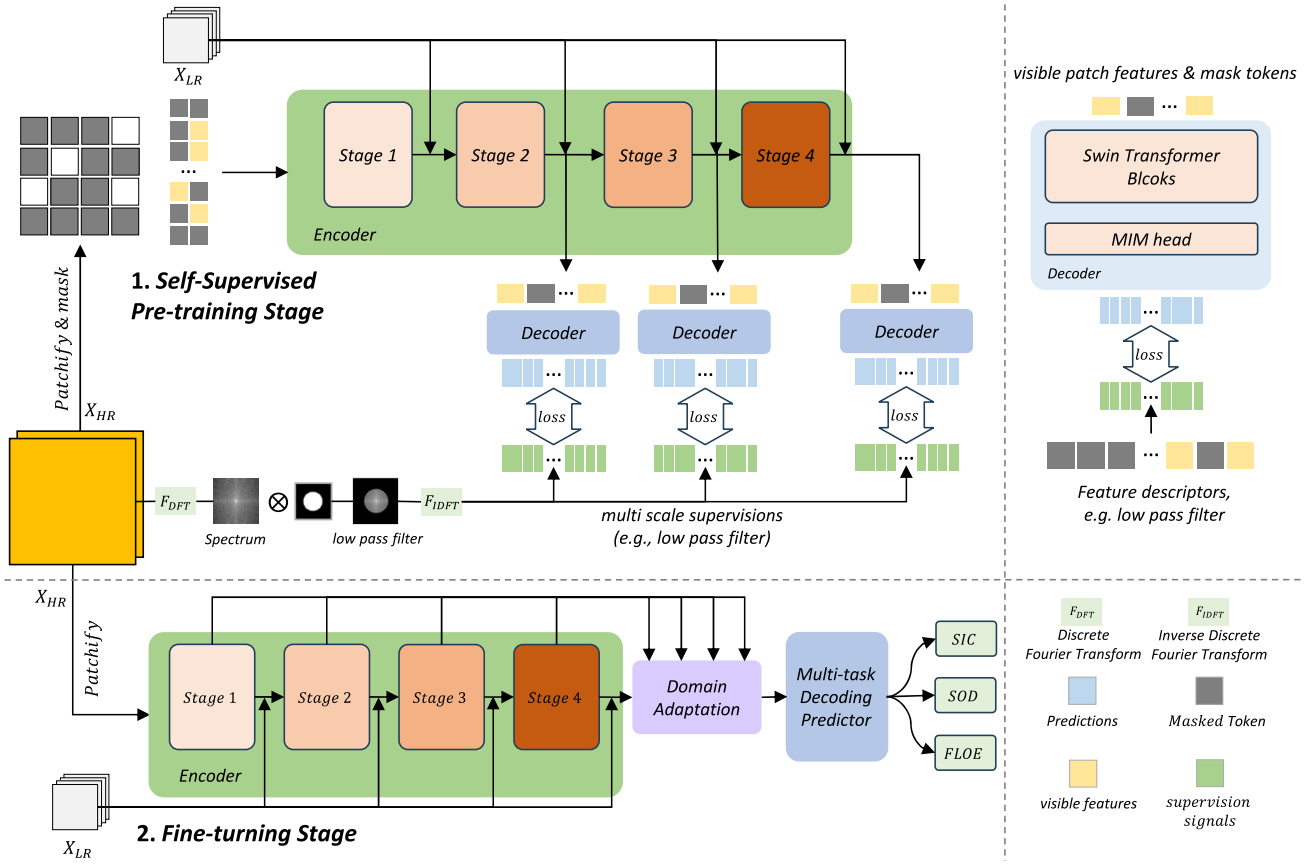

Fig. 2. Overall architecture of MLFMIM is depicted in this figure. The input HR image is divided into a series of nonoverlapping patches, where the masked patches are replaced with learnable mask tokens. The encoded LR image features are then concatenated with the visible patches. The encoder can be selected from the Transformer or CNN families. The output features of the encoder can be expressed as Stages$\{1, 2, 3, 4\}$. In the decoding stage, three decoders are introduced, corresponding to the predictions of Stage 2–Stage 4, respectively. To train the model, low-pass filtering is applied in the Fourier domain to the real HR images to construct the generated targets, and the MSE is used to measure the difference between the predicted and real images.

proxy indicator for the physical thickness of the sea ice cover, and thus, the relative ease of traversability. This parameter classifies the sea ice into five discrete categories: 0—Open Water, 1—New Ice, 2—Young Ice, 3—Thin First-Year Ice, 4—Thick First-Year Ice, and 5—Old Ice, where the latter class refers to sea ice that has persisted for more than one year.

3) FLOE: The FLOE parameter provides a categorical characterization of the spatial continuity and areal extent of the sea ice cover, with discrete classifications ranging from 0 to 6. The FLOE classes are defined as: 0—Open Water, 1—Cake Ice, 2—Small Floes, 3—Medium Floes, 4—Big Floes, 5—Vast Floes, and 6—Bergs, the latter of which encompasses variants of icebergs and glacier ice fragments.

III. Method

A. Architecture

The convolutional neural network (CNN) has been widely recognized for its progressive and superior performance in learning high-dimensional feature representations from radar remote sensing images [1], [9], [30], [59], [60], which can be attributed to the efficacy of its hierarchical feature extraction architecture. In this study, we first proposed the SSL method MLFMIM for multimodal data. The overall training process of MLFMIM is illustrated in Fig. 2. The multimodal encoder is pretrained using MLFMIM, to achieve more stable and robust feature representations for subsequent cross-scene tasks [55]. Following this, we propose the MFDA, which consists of three key modules: the multimodal encoder, the domain adaptation module, and the multi-task decoding predictor. The MLFMIM self-supervised pretraining method is employed to initialize the weights of the multimodal encoder within the MFDA framework. Subsequently, the weights of the pretrained model are fine-tuned through supervised learning. The MFDA framework comprises three key modules: the multimodal encoder, the domain adaptation module, and the multi-task decoding predictor. The domain adaptation module is specifically designed to bridge the gap between the feature representations of the source and target domains through adversarial learning methods. This approach enables comprehensive exploration and extraction of domain-invariant semantic features from multimodal remote sensing data, facilitating the transfer of these robust features across different domains.

B. Self-Supervised Model

In previous studies, a masking strategy has been employed to partition the input image into nonoverlapping blocks and randomly mask subsets of these blocks. In the MAE [61], the Transformer is utilized as the decoder, where only the visible blocks are provided as input to the encoder, and mask markers

are appended to the decoder to reconstruct the masked blocks. On the other hand, SimMIM adopts a fully connected layer as the decoder and feeds both the mask markers and visible patches to the encoder. Mathematically, the masking process of the MIM approach can be defined as $x_{\text{mask}} = x \odot (1 - \mathbb{M}) + T \odot \mathbb{M}$, where $\mathbb{M}$ represents a random occlusion mask, and T denotes the learnable mask mark.

To adapt the MIM approach for multimodal and multiresolution data, we mask the higher resolution (HR) data while leveraging the low-resolution (LR) data as a contextual cue, which is then integrated with the masked features. Simultaneously, we aim to enable the model to effectively capture and represent the LR physical characteristics, thereby facilitating the decomposition of complex relationships. By implementing this strategy, we can better handle data with disparate resolutions and improve the overall performance of the model.

While the pixel-level reconstruction employed by SimMIM has shown promising performance on natural image datasets, its direct adaptation to multimodal remote sensing data faces several challenges. Initially, the raw pixel values in remote sensing imagery frequently exhibit significant low-level spatial redundancies and are deficient in high-level semantic features. This characteristic can impede the model's capacity to discern meaningful scene-level abstractions. Furthermore, the inherent differences between Earth observation data and natural RGB images render the original single-layer decoder architecture of SimMIM less effective in comprehensively modeling the complex feature characteristics of remote sensing data, potentially impeding model convergence. Fortunately, recent studies have demonstrated that incorporating low-frequency information as the prediction target can help accelerate the convergence of masking-based models in hyperspectral remote sensing applications [55]. Moreover, the reconstruction of features across multiple scales has been demonstrated to enhance the model's multilevel semantic comprehension of the remote sensing scene [56].

To address these challenges, we propose a novel multimodal MIM framework tailored for remote sensing applications. Specifically, we use low-pass filtered signals as the target to capture the low-frequency information in the remote sensing data and avoid being interfered by the redundant spatial information of the original pixel values and the inherent noise of SAR images, which is beneficial for the model to learn high-level semantic features. In addition, we introduce a local multiscale reconstruction method to explicitly guide multiple lower layers of the model, effectively modeling remote sensing image features at different scales. This overcomes the problem that a single decoder is difficult to capture complex features and promotes rapid convergence of the model. Compared with previous works that focus on accelerating the encoding process through asymmetric encoder–decoder strategies or reducing input blocks, our approach explicitly leverages the characteristics of multimodal remote sensing data to enhance the model's learning efficiency.

The architecture of MLFMIM is depicted in Fig. 2. It consists of three primary components: the training target generator $[G(\cdot)]$, the reconstruction encoder (RE), and the reconstruction decoder (RD). The key benefit of the proposed method is the use of simple low-pass filtering for training target construction and reconstruction from partial observations, which helps to mitigate the noise in the generated features and facilitates the learning of more effective representations. Specifically, RE extracts latent representations from the unmasked feature segments, while RD reconstructs low-pass filtered representations of the masked data based on these latent representations. Furthermore, the introduction of hierarchical constraints is shown to accelerate the convergence of the model, thereby improving the efficiency of the pretraining process.

1) Training Target Generator: In the target generator part, we introduce a low-pass filter to calculate the frequency-domain information of each channel by performing a discrete Fourier transform (DFT) on the image. For an input image $I \in \mathbb{R}^{H \times W}$, the DFT is mathematically defined as follows:

$$F_{\text{DFT}}(I)_{(u,v)} = \sum_{h=1}^{h=H} \sum_{w=1}^{w=W} I(h, w) e^{-2\pi i \left(\frac{uh}{H} + \frac{vw}{W}\right)} \tag{1}$$

where (H, W) represents the spatial size, (u, v) and (h, w), respectively, denote the frequency spectrum and spatial coordinates. $F_{\text{DFT}}(I)$ represents the frequency representation of the image. To retain the low-frequency components of the image in the spectral dimension, an ideal low-pass filter (F_{LPF}) is defined as follows:

$$F_{\text{LPF}(u,v)} = \begin{cases} 1, & \left(\frac{(u-u_c)^2}{\left(\frac{H}{4}\right)^2}\right) + \left(\frac{(v-v_c)^2}{\left(\frac{W}{4}\right)^2}\right) \leq 1 \\ 0, & \text{otherwise (O.W.)} \end{cases} \tag{2}$$

where u_c and v_c represent the center coordinates of the frequency spectrum. We control the amount of high-frequency components filtered out from the spectrum based on the dimensions $H \times W$ of the image, with a default setting of 1/4 of the central region. Subsequently, the filtered spectrum undergoes an inverse DFT denoted by F_{IDFT} to generate the final reconstructed image. Therefore, the generation objectives of low-pass filtering can be defined as follows:

$$G(I) = F_{\text{IDFT}}\left(F_{\text{LPF}(u,v)} \otimes F_{\text{DFT}}(I)_{(u,v)}\right) \tag{3}$$

where F_{DFT} and F_{IDFT} can be efficiently computed using the fast Fourier transform and $\otimes$ denotes elementwise multiplication.

2) RE and Decoder: Let $X_{\text{HR}} \in \mathbb{R}^{H \times W \times C_1}$ be the HR data, and $X_{\text{LR}} \in \mathbb{R}^{(H/4) \times (W/4) \times C_2}$ denote the LR data, where H and W represent the height and width of the images, and C_1 and C_2 represent the number of channels in the HR and LR data, respectively. X_{HR} is then encoded using a patch embedding [62] to obtain the feature representation $x^0_{\text{HR}} \in \mathbb{R}^{(H/4) \times (W/4) \times 64}$. Subsequently, x^0_{HR} undergoes a masking operation, resulting in the masked feature representation, which can be formulated as follows:

$$x^{\text{mask}}_{\text{HR}} = x^0_{\text{HR}} \odot (1 - \mathbb{M}) + T \odot \mathbb{M} \tag{4}$$

where $\mathbb{M}$ represents the random occlusion mask and T denotes the learnable mask mark.

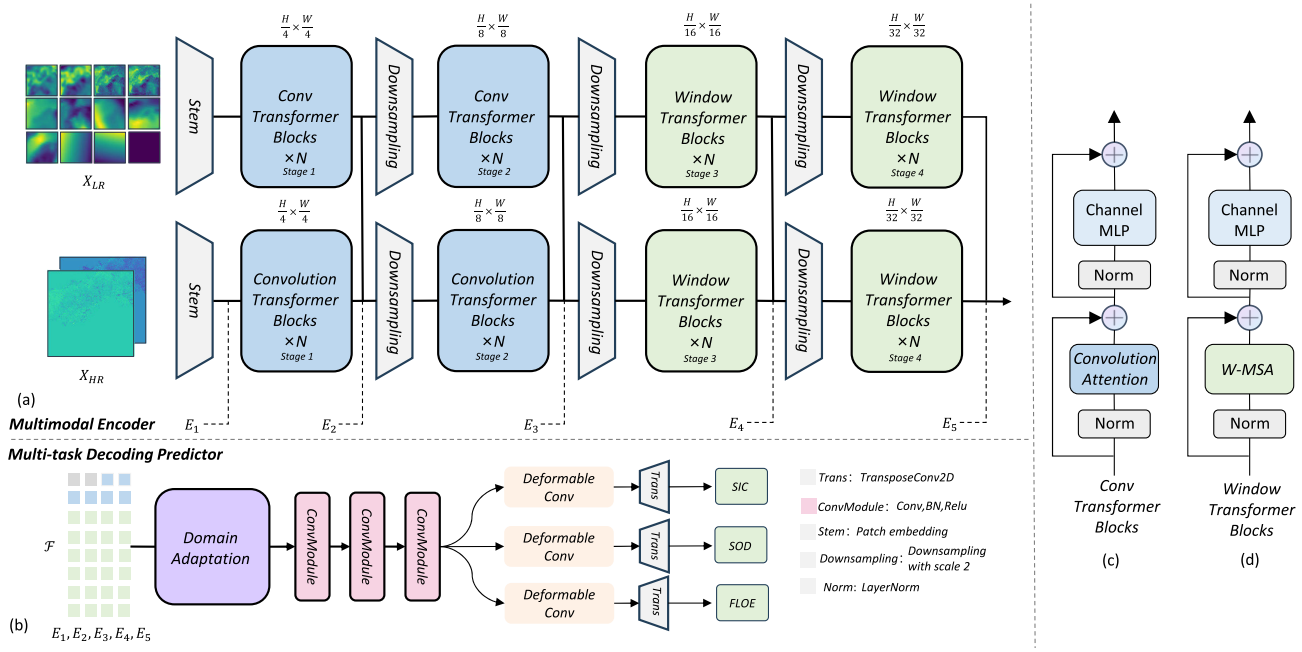

Fig. 3. Overall description of the proposed MFDA framework. X_{LR} represents low-resolution data and X_{HR} represents high-resolution data, the stem represents patch embedding [62], and the downsampling module utilizes the convolutional downsampling. The feature maps from the multimodal encoder, which fuses features from different stages, are concatenated and provided as input to the predictor. Subsequently, after DA and a series of convolutional modules, various deformable convolution and transposed convolutional layers are employed to upsample the feature maps and obtain the results for the three tasks. (a) Multimodal Encoder. (b) Multi-task Decoding Predictor. (c) Detail of Convolution Transformer Block. (d) Detail of Window Transformer Block.

X_{LR} is similarly encoded using a patch embedding to obtain the feature representation $x_{\mathrm{LR}}^0 \in \mathbb{R}^{(H/4)\times(W/4)\times 64}$. $x_{\mathrm{HR}}^{\mathrm{mask}}$ and x_{LR}^0 are then fed into the two branches of the encoder, which extracts multiscale features denoted by $\{x_{\mathrm{HR}}^i\}_{i=1}^4$ and $\{x_{\mathrm{LR}}^i\}_{i=1}^4$, respectively. For the i^{th} stage of the HR branch, we fuse it with the corresponding feature from the LR branch, represented as $x_{\mathrm{HR}}^i = x_{\mathrm{HR}}^i + x_{\mathrm{LR}}^i$. The output of the RE can be represented as $\{x_{\mathrm{HR}}^i\}_{i=2}^4 = f_{\mathrm{en}}(X_{\mathrm{HR}}, X_{\mathrm{LR}})$, where f_{en} denotes the transformation performed by the encoder blocks.

The hierarchical features $\{x_{\mathrm{HR}}^i\}_{i=2}^4$ obtained from f_{en} are passed to the RD. Subsequently, the RD utilizes the function $G(\cdot)$ to guide the reconstruction of features at various levels. Since the decoder is only utilized for output prediction during the pretraining stage, it can leverage a variety of architectural choices, such as a series of Vision Transformer [63], Swin Transformer [62], or convolutional layers. Although numerous studies have demonstrated that employing a lightweight decoder architecture can be sufficient to learn generalizable representations, it remains challenging to use a lightweight decoder for multimodal tasks. This is due to the difficulty in effectively reconstructing features from disparate modalities. In the field of remote sensing, architectures capable of achieving both local and global information propagation have been widely adopted. Therefore, this study employs a two-layer Swin Transformer as the decoder architecture.

3) Reconstruction Loss: During the pretraining phase, SSL techniques are employed, utilizing unlabeled training data. The loss function is defined as the mean squared error (MSE) between the mask data and the reconstructed mask data [55], which can be represented by the following formula:

$$\mathcal{L}_{\mathrm{MLFMIM}} = \sum_{l\in\kappa} w_l \cdot \frac{1}{N_l} \sum_{i=1}^{N_l} \mathbb{M}_i^l \cdot \left(G\left(y_i^l\right) - \hat{y}_i^l\right)^2 \tag{5}$$

where κ denotes the set of selected layers, w_l represents the coefficient of each local loss, and $\mathbb{M}_i^l$ is calculated by upsampling/downsampling the initial mask $\mathbb{M}$. N_l indicates the number of samples, $G(y_i^l)$ represents the expected output of the ith sample at the lth layer, and $\hat{y}_i^l$ denotes the predicted output of the ith sample at the lth layer. These local losses guide the patches on multiple selected layers to perform the semantic interactions at different scales, which not only accelerates the learning of multiple layers but also facilitates multiscale semantic understanding of the input.

C. MFDA

The model architecture of MFDA is illustrated in Fig. 3. Our MFDA comprises three main components: multimodal encoder (E), domain adaptor (DA) [consisting of discoverer (D), and corrector (C)], and multi-task decoding predictor (P).

1) Multimodal Encoder: The multimodal encoder is composed of two parallel branches specifically designed for feature extraction, with the gradual integration of LR data features into the corresponding HR branches. Each branch of the encoder comprises Stem, Convolutional Transformer Blocks, Downsampling, and Window Transformer Blocks.

In the context of the LR branch, we extract hierarchical features individually denoted by $\{E_i^{\mathrm{lr}}\}_{i=1}^4$, which are subsequently incorporated into the HR branch. The hierarchical feature representation for the HR branch is defined as $\{E_i^{\mathrm{hr}}\}_{i=1}^4$. To ensure information fusion at each stage, these features are fused, resulting in the fused features of the HR branch and LR branch expressed as: $E_i = E_i^{\mathrm{hr}} + E_i^{\mathrm{lr}}$, where E_i^{hr} and E_i^{lr} correspond to the feature maps from the respective stage. For each branch, the stem and downsampling consist of convolutional layers with downsampling aligned to the convolutional kernel size and stride. Convolutional

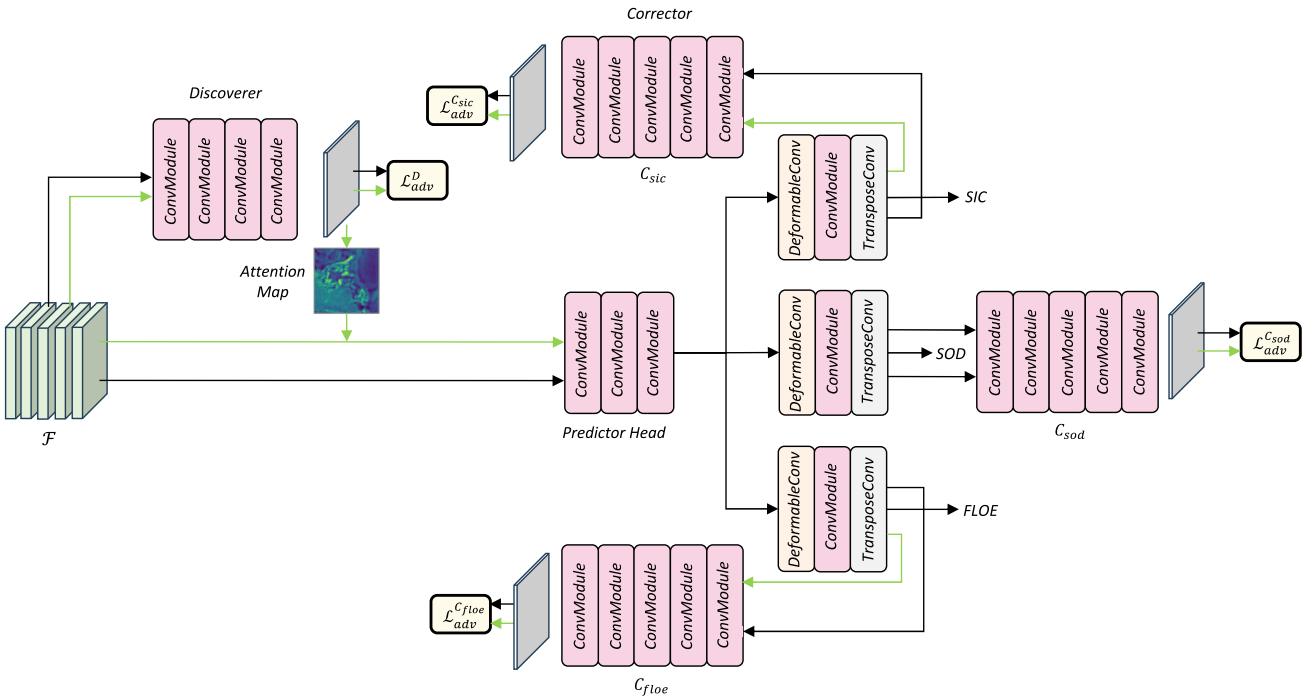

Fig. 4. Details of domain adaptation model. It is composed of discoverer (D) and corrector (C) (including C_{sic}, C_{sod}, and C_{floe}), which are used to better align the gap between the target domain and the source domain.

Transformer blocks and window Transformer blocks are essential components in MFDA. Although Transformer has shown good performance in sea ice prediction and extraction, it increases model parameters and computational complexity due to its global information extraction [64], [65]. Local information have been emphasized in remote sensing tasks such as classification and semantic segmentation [20], [66]. Utilizing a convolution-like architecture in the upper layers retains local relevant features and keeps the model compact. Recent studies, such as [67] and [68], confirm the importance of Transformer architecture design. A more general architecture was adopted, incorporating Transformer-style convolution modules in Stages 1 and 2, along with relative position encoding to enhance the model's inductive bias and emphasize its local modeling capability. The W-MSA and SW-MSA variants within the Swin Transformer model have demonstrated effectiveness in sea ice classification and ice-water separation tasks, exhibiting lower quadratic complexity than the standard MSA [62]. Consequently, MSA was replaced with W-MSA/SW-MSA in Stages 3 and 4 of the model, enabling it to possess both local and global modeling capabilities, thereby rendering it suitable for MTL of sea ice.

Within the HR branch, the Convolutional Transformer Blocks involve the definition of input features as E_{i-1} and output features as E_i^{hr}, which can be expressed as follows:

$$E_i' = E_{i-1} + \text{ConvAtt}(\text{BN}(E_{i-1})) \tag{6}$$

$$E_i^{\text{hr}} = E_i' + \text{CMLP}\left(\text{BN}\left(E_i'\right)\right) \tag{7}$$

$$\text{ConvAtt}(x) = \text{Conv}_{\text{pw2}}\left(\text{Conv}_{\text{dw}}\left(\sigma\left(\text{Conv}_{\text{pw1}}(x)\right)\right)\right) \tag{8}$$

where $\text{Conv}_{\text{pw}}(\cdot)$ represents the pointwise convolution, $\text{Conv}_{\text{dw}}(\cdot)$ represents the depthwise convolution, $\sigma(\cdot)$ represents the activation function, $\text{BN}(\cdot)$ represents the batch normalization, and $\text{CMLP}(\cdot)$ consists of two layers of convolution with an activation function in between. In the HR branch, for window Transformer blocks, we define the input features as E_{i-1} and the output features as E_i^{hr}, which can be expressed as follows:

$$E_{i-1}' = E_{i-1} + \text{W-MSA}(\text{LN}(E_{i-1})) \tag{9}$$

$$E_{i-1}'' = E_{i-1}' + \text{MLP}\left(\text{LN}\left(E_{i-1}'\right)\right) \tag{10}$$

$$E_i' = E_{i-1}'' + \text{SW-MSA}\left(\text{LN}\left(E_{i-1}''\right)\right) \tag{11}$$

$$E_i^{\text{hr}} = E_i' + \text{MLP}\left(\text{BN}\left(E_i'\right)\right) \tag{12}$$

where LN refers to layer normalization, MLP refers to the multilayer perceptron, and W/SW-MSA is derived from the Swin Transformer architecture proposed in [62]. Subsequently, a sequence of convolutional and upsampling modules is applied to E_0 and $\{E_i\}_{i=1}^4$, followed by their concatenation to form fusion features $\mathcal{F}$. This process can be expressed as follows:

$$\mathcal{F} = \text{Cat}\left(W_E^0 E_0, \left\{\text{Up}_{\times 2^{2i-2}}\left(W_E^i E_i\right)\right\}_{i=1}^4\right) \tag{13}$$

where $\text{Up}\times 2^{2i-2}$ denotes the upsampling operation, $\times 2^{2i-2}$ represents the scale factor of upsampling, and W_S^i represents the learnable parameter.

2) Domain Adaptation: Fig. 4 shows the detail of the domain adaptation module. Generative adversarial networks (GANs) have gained significant popularity in addressing pixel-level alignment and knowledge transfer challenges across diverse source and target domains. Taking inspiration from the previous works [39], [40], [41], [42], [43], [44], [69], [70], [71], we introduced a domain adaptation module to tackle the domain offset issue after merging features. Our

approach involves training a cross-domain segmentation network using randomly selected images from both the source and target domains through adversarial training. To achieve improved alignment between the target and source domains, we employ two types of convolution-based discriminant networks, namely, discoverer (D) and corrector (C) (including C_{sic}, C_{sod}, and C_{floe}). These networks facilitate effective alignment by addressing the representation offset between the domains. Specifically, D generates pixel-level confidence scores, enabling local correction of the representation offset by reweighting intermediate features. Concurrently, the multi-task category C aims to enhance global semantic alignment by considering three distinct label distributions during the final prediction stage. This global semantic alignment operation can be viewed as a soft constraint, bringing different task categories closer and promoting similarity among the same categories across the domains.

Let X_s and X_t be the sets of input images for the source and target domains, respectively. In addition, let $\mathcal{F}_s$ and $\mathcal{F}_t$ be the fusion features extracted from the source and target domains, respectively, using the encoder E. To align the feature distributions of $\mathcal{F}_t$ and $\mathcal{F}_s$, we optimize the discriminator D using the adversarial loss function $\mathcal{L}_{\text{adv}}^{D}$. This comparison provides an aligned confidence score for each position in $\mathcal{F}_t$, resulting in the formation of an attention map denoted by α. Notably, α is then utilized to reweight $\mathcal{F}_t$, resulting in a new feature map denoted by $\hat{\mathcal{F}}_t = \alpha \otimes \mathcal{F}_t$. Subsequently, pixel-level predictions p_t^{sic}, p_t^{sod}, and p_t^{floe} are obtained through P, with a heightened focus on poorly aligned areas. Subsequently, we employ the loss function $\mathcal{L}_{\text{adv}}^{C}$ to optimize the discriminator C, which facilitates adversarial learning by comparing the predictions p_t and p_s. The above procedures can be expressed as follows:

$$p_t^{\text{sic}}, p_t^{\text{sod}}, p_t^{\text{floe}} = P(\mathcal{F}_t) \tag{14}$$

$$p_s^{\text{sic}}, p_s^{\text{sod}}, p_s^{\text{floe}} = P(\mathcal{F}_s). \tag{15}$$

3) Multi-Task Decoding Predictor: Regarding the target domain features $\hat{\mathcal{F}}_t$ obtained from the DA module, we employ the convolutional group module for feature representation modeling, followed by upsampling. The convolutional group module comprises three consecutive convolutional layers, where each layer is accompanied by a batch normalization layer and a rectified linear unit (ReLU) function. In addition, we introduce three sets of deformable convolutions at the model's final stage. In comparison with conventional fixed grid convolutions, deformable convolutions [72] adapt to irregular shapes and structures by learning spatial transformations, enabling more precise capture of target features. This convolutional approach allows the model to dynamically adjust its receptive field, thereby enhancing its adaptability to the diverse and complex input data. The incorporation of deformable convolutions significantly improves the model's ability to adapt to changes in the shape of sea ice.

4) Loss Function: The overall loss function mainly consists of three loss terms

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \eta \mathcal{L}_{\text{adv}}^{D} + \mu \mathcal{L}_{\text{adv}}^{C} \tag{16}$$

where η and μ are the hyperparameters that have been set to 0.5. The first term on the right-hand side of (16) can be expressed as follows:

$$\begin{aligned}\mathcal{L}_{\text{seg}}(y, \hat{y}) = {} & \lambda_1 \mathcal{L}_{\text{ce}}(y_{\text{sic}}, \hat{y}_{\text{sic}}) \\ & + \lambda_2 \mathcal{L}_{\text{ce}}(y_{\text{sod}}, \hat{y}_{\text{sod}}) \\ & + \lambda_3 \mathcal{L}_{\text{ce}}(y_{\text{FLOE}}, \hat{y}_{\text{FLOE}})\end{aligned} \tag{17}$$

where $\mathcal{L}_{\text{ce}}$ represents the label smoothing cross-entropy loss function [73]. Also, the second term on the right-hand side of (16) can be expressed as follows:

$$\begin{aligned}\mathcal{L}_{\text{adv}}^{D}(D) = {} & \min_{D} \mathbb{E}_{X_s \sim p(X_s)} \left[(D(\mathcal{F}_s) - 0)^2 \right] \\ & + \mathbb{E}_{X_t \sim p(X_t)} \left[(D(\mathcal{F}_t) - 1)^2 \right]\end{aligned} \tag{18}$$

$$\mathcal{L}_{\text{adv}}^{D}(E) = \min_{E} \mathbb{E}_{X_t \sim p(X_t)} \left[(D(E(\mathcal{F}_t)) - 0)^2 \right] \tag{19}$$

where $\mathcal{F}_s$ represents the feature map extracted from the input sample X_s in the source domain, while $\mathcal{F}_t$ represents the feature map extracted from the input sample X_t in the target domain. The terms $\mathbb{E}_{X_s \sim p(X_s)}$ and $\mathbb{E}_{X_t \sim p(X_t)}$ denote the expectation of sampling from the data distribution of the source domain and target domain, denoted by $p(X_s)$ and $p(X_t)$. Also, the third term on the right-hand side of (16) can be expressed as follows:

$$\begin{aligned}\mathcal{L}_{\text{adv}}^{C}(C) = {} & \min_{C} \mathbb{E}_{X_s \sim p(X_s)} \left[\left(C\left(p_s^{\text{sic}}, p_s^{\text{sod}}, p_s^{\text{floe}}\right) - 0 \right)^2 \right] \\ & + \mathbb{E}_{X_t \sim p(X_t)} \left[\left(C\left(p_t^{\text{sic}}, p_t^{\text{sod}}, p_t^{\text{floe}}\right) - 1 \right)^2 \right]\end{aligned} \tag{20}$$

$$\begin{aligned}\mathcal{L}_{\text{adv}}^{C}(E, P) = {} & \min_{E} \mathbb{E}_{X_t \sim p(X_t)} [(C(\{P_{\text{sic}}, P_{\text{sod}}, P_{\text{floe}}\} \\ & \circ E(X_t)) - 0)^2]\end{aligned} \tag{21}$$

where $P_\kappa \circ E(X_t)$ represents the category mapping of features extracted from target domain input samples, the parameters p_s^{sic}, p_s^{sod}, p_s^{floe} from (14) and the parameters p_t^{sic}, p_t^{sod}, p_t^{floe} from (15) are incorporated into (20).

IV. Experiments

A. Experimental Platform Parameter Settings

The experiments were conducted on a desktop equipped with an NVIDIA GeForce RTX 4060Ti 16-GB GPU using PyTorch 1.10. For the training of both MLFMIM and MFDA, we employed the Adam optimizer [74] as the initial optimizer, with an initial learning rate of 5×10^{-4} [55]. The MLFMIM model was trained for 400 epochs with a batch size of 32. The MFDA model was trained for 200 epochs with a batch size of 32. To assess the efficacy of the model, we employed the $F1$-score ($F1$) and overall accuracy (OA) as evaluation metrics.

B. Determination of MFDA Parameters

The current study investigates the optimal configuration of the MFDA. The experimental outcomes are delineated in Table II, corroborating the findings with previous studies that have dissected the effects of varying stage layer stacking on model efficacy [62], [66]. This experiment aims to determine the optimal architectural parameters of MFDA through experimental evaluation. Specifically, the study explores the influence of the number of channels and the layers across

TABLE II

IMPACT OF DIFFERENT HYPERPARAMETERS ON THE MODEL PERFORMANCE WAS INVESTIGATED, WHILE MLFMIM AND DA WERE NOT INCORPORATED. IN THE MFDA, THE NUMBER OF CHANNELS IN THE FIRST STAGE WAS SET TO C, AND THE CHANNEL NUMBERS IN THE SUBSEQUENT STAGES INCREASED IN THE MANNER OF [C, 2C, 4C, 8C]. LAYERS DENOTE THE NUMBER OF STACKED BLOCKS AT EACH STAGE OF THE NETWORK, AND GRAY REPRESENTS THE BEST RESULT

C	Layers	SIC		SOD		FLOE	
		F1 (%)	OA (%)	F1 (%)	OA (%)	F1 (%)	OA (%)
16	[2,2,2,2]	73.3	71.5	68.6	69.4	72.7	74.6
16	[2,2,6,2]	74.1	72.0	68.1	70.6	75.3	76.5
16	[3,4,8,3]	73.7	72.0	67.8	69.3	68.3	70.6
32	[2,2,2,2]	75.6	72.7	67.9	69.6	72.8	74.7
32	[2,2,6,2]	76.3	73.1	70.0	70.7	75.6	75.8
32	[3,4,8,3]	76.4	73.3	68.8	71.4	71.6	73.2
64	[2,2,2,2]	74.0	72.0	68.2	70.2	70.4	72.4
64	[2,2,6,2]	76.2	73.1	70.2	71.4	74.8	75.6
64	[3,4,8,3]	73.3	71.5	68.6	69.4	72.7	74.6

different stages on model performance. A total of nine experimental combinations were designed to investigate the impact of basic channel numbers of {16, 32, and 64}, as well as the layers of {[2, 2, 2, 2], [2, 2, 6, 2], and [3, 4, 8, 3]}. The findings suggest that a larger number of channels can lead to an excessive number of parameters, increase model complexity, and potentially result in overfitting, thereby reducing the model's generalization performance. A smaller number of channels may not be able to fully capture the rich features of the input data, limiting the model's expressiveness. A medium-sized number of channels, such as 32 channels, appears to provide a better balance between the performance and model complexity.

The findings indicate that when the stage stacking configuration in MFDA is an even number, such as [2, 2, 2, 2] or [2, 2, 6, 2], the model performance is superior to the odd stage stacking configuration, such as [3, 4, 8, 3]. The shifted window attention mechanism proposed in Swin Transformer [62] appears to function more effectively under the even stage configuration. Specifically, Stage 3 and Stage 4 utilize this mechanism to promote cross-window feature communication and fusion, thereby enhancing the model's receptive field and expression ability. However, an excessively low number of stage stacks, such as [2, 2, 2, 2], may limit the model's expression ability and prevent it from fully learning the complex features in the data. Therefore, this study selects the [2, 2, 6, 2] layers configuration as a more reasonable compromise. This configuration retains the advantages of the Swin Transformer's shifted window mechanism while increasing the stacking depth of Stage 3, thereby enhancing the model's expressive capability.

C. Efficacy of MTL

To investigate the efficacy of MTL, we performed three task experiments. The experimental results are presented in Table III. Notably, the multi-task model achieved a marked enhancement in $F1$-scores, with increments of 0.9%, 2.4%, and 4.6% for the respective tasks, thereby outperforming the single-task model. These results demonstrate that MTL yields improved performance. This enhancement can be attributed to the inherent correlations among various sea ice parameters and the capacity of MTL to bolster the model's robustness against noise and uncertainty. When the training dataset for sea ice is limited, the feature representation achieved through MTL outperforms that of single-task learning, thereby enhancing the model's generalization ability. This approach leverages positive transfer between tasks, wherein the learning of one task actively facilitates the learning of other related tasks, thus improving the overall training and inference performance. For instance, in a specific region, higher SIC is typically associated with larger floe sizes. In an MTL framework, the model can learn these correlations, resulting in simultaneous performance improvements across all tasks. By integrating multiple related tasks and reusing features, MTL establishes implicit constraints between tasks, offering a more effective solution compared to traditional single-task learning. Furthermore, through information sharing and optimization processes, MTL not only enhances model performance but also effectively mitigates the risk of overfitting, thereby further improving the model's generalization ability.

TABLE III

COMPARISON BETWEEN MULTI-TASK AND SINGLE-TASK LEARNING, GRAY REPRESENTS THE BETTER RESULT

Task	SIC		SOD		FLOE	
	F1 (%)	OA (%)	F1 (%)	OA (%)	F1 (%)	OA (%)
Single Task	75.6	72.7	67.6	69.5	71.0	71.9
Multi Task	76.3	73.1	70.0	70.7	75.6	75.8

D. Parameter Sharing Experiment

In this study, we examined the performance difference between soft parameter sharing and hard parameter sharing methods for sea ice classification tasks. Specifically, we configured the model's decoder according to different sharing strategies. For the soft parameter sharing approach, we used three independent decoder parameter sets to learn the distinct classification tasks of SIC, SOD, and FLOE. In contrast, for the hard parameter sharing method, we employed a single-shared decoder and achieved MTL by connecting three independent classification heads.

The comparative experimental outcomes are provided in Table IV. In the evaluation of OA across the SIC, SOD, and FLOE tasks, the UNetHard model demonstrated superior performance over the UNetSoft model, with respective OA improvements of 2.1%, 0.3%, and 0.5%. In addition, the MFDAHard model surpassed the MFDASoft model in the SOD task by an OA margin of 3.6%. These findings indicate that in sea ice applications, soft parameter sharing methods perform less effectively than hard parameter sharing in MTL. The advantages of hard parameter sharing lie in its simplicity and efficiency, particularly when addressing highly correlated tasks [75]. By sharing representations, hard parameter sharing mitigates the risk of overfitting to a single task and enhances training and inference efficiency by reducing the number of parameters. Given the inherent similarities among different

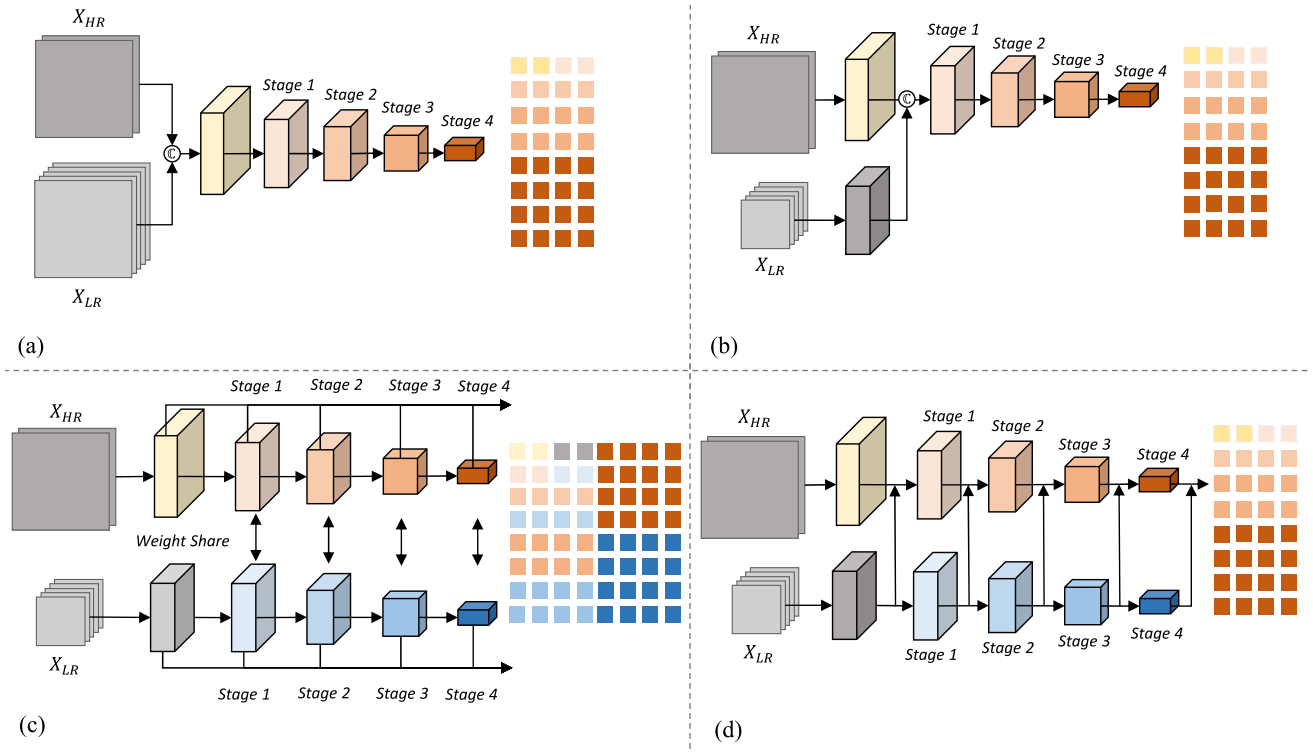

Fig. 5. Overall description of the proposed data fusion paradigm (CNN). (a)–(d) Correspond to (1)-(4) CNN-based fusion paradigms.

TABLE IV
COMPARISON RESULTS BETWEEN DIFFERENT PARAMETER SHARING, GRAY REPRESENTS OUR METHOD

Parameter Sharing	SIC		SOD		FLOE	
	F1 (%)	OA (%)	F1 (%)	OA (%)	F1 (%)	OA (%)
UNetSoft	72.6	70.8	68.2	70.0	75.9	76.3
UNetHard	75.8	72.9	67.5	70.3	75.9	76.8
MFDASoft	76.6	73.7	72.6	74.1	76.1	77.4
MFDAHard	78.0	73.7	75.8	77.7	79.8	79.7

sea ice classification tasks—such as the correlation where a higher SIC corresponds to a higher floe ratio in a specific area—hard parameter sharing can better leverage these task similarities, thus improving the effectiveness of MTL for sea ice classification. In contrast, soft parameter sharing tends to increase the number of model parameters, which can complicate training and diminish the model's generalization performance. Consequently, the subsequent experiments in this study adopt a hard parameter sharing configuration.

E. Fusion Paradigm Selection

In this work, we investigate the impact of different fusion paradigms on multimodal data fusion for semantic segmentation. Existing deep learning models for this task can be broadly categorized into CNN-based and Transformer-based fusion methods. For the CNN-based approaches, the choice of fusion level has a significant impact on the overall performance. Conversely, for the Transformer-based fusion methods, the self-attention mechanism can be leveraged to fuse multimodal features at the token level, which may offer advantages over the CNN-based techniques. However, fusion at different levels still plays a key role in determining the performance of Transformer-based fusion methods as well.

For the CNN-based fusion, we conducted an evaluation of four distinct training paradigms, as delineated in Fig. 5. To delineate these paradigms, we introduce a symbolic representation for input modalities and network components. Distinct colors denote the Encoder Blocks/Convolutions of each branch, with the final blocks representing feature maps from various stages.

Specifically, different paradigms depicted in the corresponding images can be described as follows.

1) Direct resampling of multimodal data to a unified resolution, followed by concatenation at the model input

$$\mathcal{F} = f_{\text{en}}(X_{\text{HR}} \parallel X_{\text{LR}}). \tag{22}$$

2) Independent convolutional processing of modalities, subsequent concatenation, and feeding into a shared encoder

$$\mathcal{F} = f_{\text{en}}(W_{\text{HR}} X_{\text{HR}} \parallel W_{\text{LR}} X_{\text{LR}}). \tag{23}$$

3) Separate convolutional layers for modalities, followed by a shared encoder with weight sharing

$$\mathcal{F} = f_{\text{en}}(W_{\text{HR}} X_{\text{HR}}),\ f_{\text{en}}(W_{\text{LR}} X_{\text{LR}}). \tag{24}$$

4) Initial separate convolutional processing, fusion of LR and HR feature maps, and progressive integration of LR features into the HR branch via an independent encoder

$$\mathcal{F} = f_{\text{en}}(W_{\text{HR}} X_{\text{HR}}) + f_{\text{en}}(W_{\text{LR}} X_{\text{LR}}). \tag{25}$$

For the Transformer-based fusion methods, different paradigms are shown in Fig. 6. We adopt methodologies

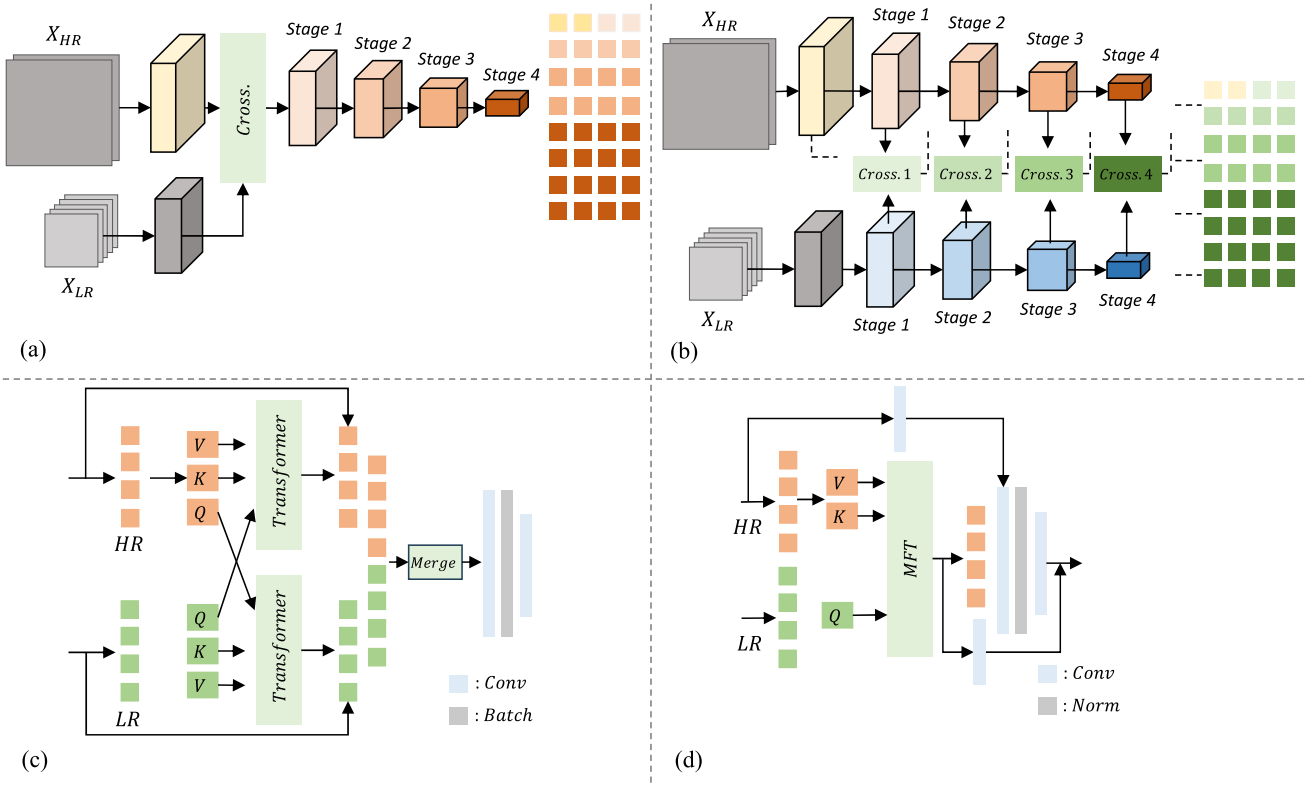

Fig. 6. Overall description of the proposed data fusion paradigm (Transformer). (a) and (b) Correspond to (1)-(2) Transformer-based fusion paradigms. (c) Detail of CrossFormer. (d) Detail of MFT.

from the CrossFormer [76] and MFT [46] frameworks. The schematic employs color coding to differentiate the Encoder Block/Convolution/Transformer components across various branches, with the final blocks indicating the aggregated feature maps from successive stages. The notation Cross. in the figure signifies the CrossBlock1/CrossBlock2 fusion operations.

Different fusion methods can be described as follows.

1) Initial feature extraction is conducted via convolutional layers for each input modality, succeeded by the application of CrossBlock (denoted as "Cross.") for feature fusion across modalities

$$\mathcal{F} = f_{\text{en}}(\text{Cross.}(W_{\text{HR}}X_{\text{HR}}, W_{\text{LR}}X_{\text{LR}})). \tag{26}$$

2) Features are extracted for each modality using separate encoders postconvolutional processing. Subsequently, the Cross. operation is applied to integrate features from different modalities and stages

$$\mathcal{F} = \{\text{Cross.}i(f_{\text{en}}(W_{\text{HR}}X_{\text{HR}})_i, f_{\text{en}}(W_{\text{LR}}X_{\text{LR}})_i)\}_{i=1}^{4}. \tag{27}$$

Let the input feature as z_{k-1} and the output as z_k. The forward process of Transformer [46], [77], [78] can be expressed as follows:

$$z'_k = z_{k-1} + \text{MSA}(\text{LN}(z_k)) \tag{28}$$

$$z_k = z'_k + \text{MLP}\big(\text{LN}\big(z'_k\big)\big). \tag{29}$$

We denote the input features at Stage i are z_{HR} and z_{LR}, then map z_{HR} to Q_h, K_h, and V_h, and z_{LR} to Q_l, K_l, and V_l. Subsequently, we perform the cross-attention mechanism to compute the output representations. Finally, convolution is applied, and the features are projected to the same dimension. The forward process can be expressed as follows:

$$z'_{\text{HR}} = z_{\text{HR}} + \text{MSA}\left(\text{LN}\left(\underbrace{z_{\text{LR}}, z_{\text{HR}}, z_{\text{HR}}}_{Q_l, K_h, V_h}\right)\right) \tag{30}$$

$$z'_{\text{LR}} = z_{\text{LR}} + \text{MSA}\left(\text{LN}\left(\underbrace{z_{\text{HR}}, z_{\text{LR}}, z_{\text{LR}}}_{Q_h, K_l, V_l}\right)\right) \tag{31}$$

$$\hat{z}_{\text{HR}} = z'_{\text{HR}} + \text{MLP}\big(\text{LN}\big(z'_{\text{HR}}\big)\big) \tag{32}$$

$$\hat{z}_{\text{LR}} = z'_{\text{LR}} + \text{MLP}\big(\text{LN}\big(z'_{\text{LR}}\big)\big) \tag{33}$$

$$z_o = W_i^o\big(\hat{z}_{\text{HR}} \parallel \hat{z}_{\text{LR}}\big) \tag{34}$$

where W_i^o refers to learnable matrix. In addition, the forward process of MFT can be expressed as follows:

$$z' = z_{\text{HR}} + \text{MSA}\left(\text{LN}\left(\underbrace{z_{\text{LR}}, z_{\text{HR}}, z_{\text{HR}}}_{Q_l, K_h, V_h}\right)\right) \tag{35}$$

$$z_o = W_i^o\big(z' + \text{MLP}\big(\text{LN}\big(z'\big)\big) + z_{\text{LR}}\big). \tag{36}$$

The results of different paradigms are presented in Table V. It should be noted that we did not evaluate MLFMIM and DA, which allows us to examine the differences between the various fusion paradigms. For the CNN-based paradigm, Paradigm d performs best on the SIC task. Compared to Paradigms a–c, the $F1$ score is higher by 4.2%, 4.1%, and 2%, respectively. However, the $F1$ and OA scores of Paradigm d

TABLE V
COMPARATIVE EVALUATION OF DATA FUSION PARADIGMS ACROSS CNN-BASED AND TRANSFORMER-BASED, MLFMIM, AND DA ARE NOT USED, AND GRAY REPRESENTS OUR METHOD. IN THE CNN-BASED METHODS, PARADIGMS A–D DENOTE THE DISTINCT FUSION STRATEGIES. IN THE TRANSFORMER-BASED APPROACHES, AC AND BC REFER TO THE PARADIGMS A AND B OF THE CROSSFORMER MODULE WITHIN THE CROSS. FUSION TECHNIQUE. IN ADDITION, AD AND BD REPRESENT THE PARADIGMS A AND B OF THE MFT FUSION APPROACH IN THE CROSS. FRAMEWORK

Method		SIC		SOD		FLOE	
		F1 (%)	OA (%)	F1 (%)	OA (%)	F1 (%)	OA (%)
CNN	a	72.1	70.7	68.6	68.9	70.4	71.9
	b	72.2	70.2	68.5	66.5	73.6	74.7
	c	74.3	72.3	70.9	72.1	68.4	70.5
	d	76.3	73.1	70.0	70.7	75.6	75.8
Transformer	ac	76.2	72.9	68.4	69.0	76.5	76.8
	ad	73.0	71.1	69.1	69.7	72.5	73.6
	bc	71.9	70.5	69.4	71.1	77.4	78.5
	bd	73.5	71.4	70.1	69.9	75.3	76.5

on the SOD task are lower than Paradigm c by 0.9% and 1.4%, respectively. This suggests that for both CNN-based and Transformer-based methods, it is preferable to process HR and LR data separately. Using a shared-weight feature extraction module to process HR and LR data simultaneously can lead to confusion in the feature extraction module, resulting in decreased model accuracy. For the Transformer-based paradigm, the $F1$ score of Paradigm ac on SIC is better than ad, bc, and bd by 3.2%, 4.3%, and 2.7%, respectively, but its performance on SOD is the worst. Paradigm bc has the best $F1$ score on FLOE, which is better than ac, ad, and bd by 0.9%, 4.9%, and 2.1%, respectively. This suggests that using CrossFormer and MFT to replace the splicing and addition operations can improve the model, but the introduction of Transformer modules in shallow fusion will slow down the training and cause confusion between different tasks. The strategy of progressive fusion using Paradigm c achieves the best results, as it separates the feature extraction modules for HR and LR data and considers the interaction between different layers. Therefore, we choose this paradigm as our baseline method.

F. MIM Ablation Experiment

To explore the potential of the MIM method on multimodal data, we investigated the impact of various improvements on MLFMIM. The experimental results are presented in Table VI. We define the baseline models, which include SimMIM [54] and LocalMIM [56]. LocalMIM outperforms SimMIM, particularly on the SOD task, where the F1 score and OA are higher by 2% and 3%, respectively. This suggests that the multistage constraint method is effective. Given the versatility of MLFMIM, we explored replacing the original linear decoder with a two-layer MLP or Transformer/Swin Transformer decoder. We found that the Swin Transformer decoder is more efficient, especially in improving the FLOE index, where the $F1$ score and OA reached 77.1% and 78.1%, respectively. For the prediction target, considering that remote sensing data are typically processed by low-pass and high-pass filtering, we compared the original pixels, low-pass filtering, and high-pass filtering as targets. Using high-pass filtering resulted in a gradient effect, causing the model to degrade, particularly in the $F1$ index of SIC, which was 3.3% and 4.3% lower than the RGB and low targets, respectively. In addition, as shown in Fig. 7, we visualized the reconstruction of different features. We found that using low-pass filtering as the training target is more effective. It is challenging for the model to reconstruct the edges, details, and noise in the original pixels and high-pass filtering, and it is difficult to learn useful features.

G. Reconstruction Masking Analysis

Table VII presents the experimental results using different masking ratios for the MLFMIM. The best performance was achieved at a masking ratio of 60%. Notably, the MLFMIM method was found to be effective even at low masking ratios, highlighting the effectiveness of its independent reconstruction approach. This improvement is attributed to the encoder's ability to learn robust feature representations, thereby mitigating the impact of multimodal data noise. From a masking ratio of 20%–60%, the model's performance gradually increased as the masking ratio increased. Even at higher masking ratios, such as 75%–90%, the model was still able to effectively learn features. However, extremely high masking ratios, such as 95%, posed challenges for the model in learning effective features due to significant information loss and ambiguity.

H. Comparative Experiment

The proposed method was benchmarked against a diverse set of representative deep learning techniques for semantic segmentation, including CNN-based architectures, such as UNet [33], PSPNet [79], LinkNet [80], DeepLabv3 [81], MMSeaIce [9], and UNet++ [82], as well as Transformer-based models including SwinUNet [9], SwinUper [62], UniFormer [58], and PoolFormer [67]. In addition, DA approaches, specifically AdaptSegNet [40] and DAST [71], were included in the comparative evaluation. For a fair comparison, the network configurations were kept consistent with the settings reported in the original publications. Where necessary, minor parameter adjustments were made to accommodate the requirements of the multimodal remote sensing data segmentation task.

The experimental results are presented in Table VIII. The proposed method demonstrates significant advantages over the baseline techniques. While U-Net performs well within a single region, it exhibits suboptimal performance in cross-region retrieval tasks, particularly in the classification of different ice types, achieving the lowest $F1$ score of 67.5%. LinkNet is less effective than U-Net in extracting FLOE features, likely due to the absence of deconvolution operations. Both the PSPNet and DeepLabV3 models demonstrated their respective strengths across the three evaluation tasks. The PSPNet model achieved the best overall results in the SIC and FLOE tasks, with $F1$ scores exceeding those of DeepLabV3 by 2.9% in

TABLE VI
COMPARISON OF DIFFERENT DATA FUSION PARADIGM, MLFMIM, AND DA ARE NOT USED, AND GRAY REPRESENTS THAT WE CHOOSE

Method		SIC		SOD		FLOE	
		F1 (%)	OA (%)	F1 (%)	OA (%)	F1 (%)	OA (%)
Baseline	SimMIM	75.7	72.6	73.3	74.5	73.1	74.4
	LocalMIM	76.3	73.6	75.3	77.5	74.1	75.4
Decoder	Linear	76.5	73.7	74.2	75.6	74.4	75.3
	2-layer-MLP	76.0	73.8	76.1	77.2	76.0	77.2
	2-layer-Transformer	76.9	73.8	75.8	77.7	75.2	76.7
	2-layer-Swin	76.8	73.8	75.9	77.7	77.1	78.1
Target	RGB	77.0	73.3	73.7	74.7	75.4	76.2
	Low	78.0	73.7	75.8	77.7	79.8	79.7
	High	73.7	72.3	73.7	75.8	74.4	75.9

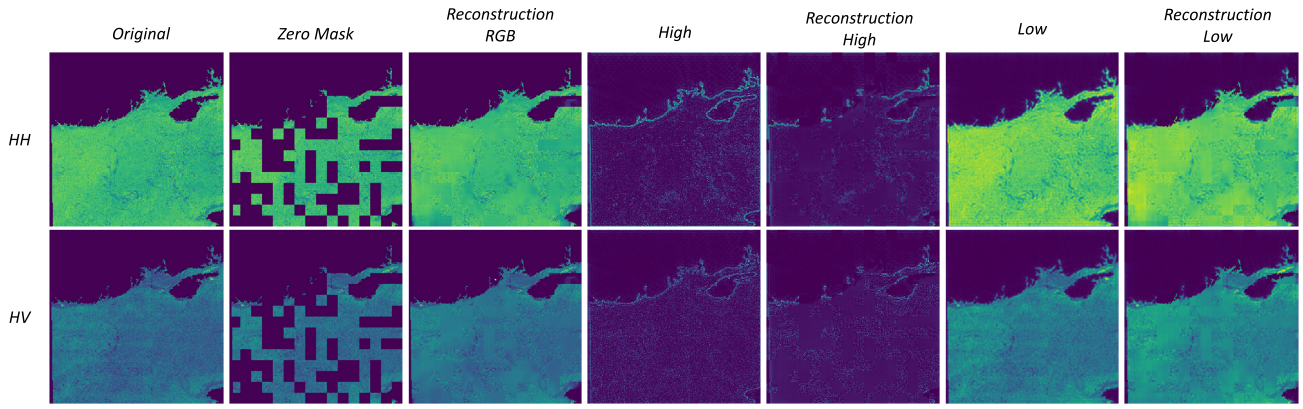

Fig. 7. Reconstruction visualization results of different targets.

TABLE VII
COMPARISON OF DIFFERENT RATES, AND GRAY REPRESENTS OUR METHOD

Rate	SIC		SOD		FLOE	
	F1 (%)	OA (%)	F1 (%)	OA (%)	F1 (%)	OA (%)
20%	76.6	73.1	73.7	75.2	76.7	77.7
40%	76.7	73.6	74.9	76.2	77.3	78.3
60%	78.0	73.7	75.8	77.7	79.8	79.7
75%	77.0	73.4	74.3	76.1	79.2	79.0
90%	77.3	74.1	74.0	75.7	77.0	78.0
95%	74.3	72.1	72.0	71.1	73.3	73.1

SIC and 2.7% in FLOE. Notably, the MMSeaIce model, as a state-of-the-art (SOTA) approach, performed well across all three tasks; however, it exhibited suboptimal performance in FLOE and struggled to distinguish between different types of sea ice in the cross-scene SOD task. In the case of UNet++, its densely connected architecture exhibited a tendency to overfit within a single region, rendering it less suitable for cross-scenario classification tasks. In addition, for the Swin Transformer model, the UNet-like architecture proved inadequate, as evidenced by its poor performance in the SOD task, where it achieved an $F1$ score of only 66.9%, consistent with findings from the previous literature [9].

Compared with the combination of the Swin Transformer and U-Net architectures, integrating the Swin Transformer with the UperNet [83] approach demonstrated better suitability for cross-scene classification tasks. Specifically, the $F1$-scores on the SOD and FLOE benchmarks were 1.6% and 2.5% higher, respectively, than the SwinUNet configuration. Regarding the UniFormer model, the architecture was primarily designed for video segmentation tasks; thus, the decoder design may need to be reconsidered for optimal performance on cross-scene classification problems. For the PoolFormer model, the results confirm the superiority of the Transformer-based architecture. In the context of sea ice classification, the use of complex spatial representation acquisition methods may not provide significant benefits, and architectural adjustments could lead to substantial performance improvements. Regarding the DA-based methods, such as AdaptSegNet, the $F1$ and OA on the SOD task reached 67.7% and 68.9%, respectively, which are weaker than the performance of DeepLabV3 and other benchmark models. However, these DA-based approaches exhibited improved results on the FLOE task, surpassing DeepLabV3 by 1.6% higher $F1$-scores. The DAST model achieved an $F1$ of 76.3% on the SIC task, outperforming DeepLabV3 and other methods, but underperformed on the SOD and FLOE tasks. This disparity can be attributed to the fact that these DA-based techniques were primarily designed for RGB image inputs, and they fail to adequately capture the role of local and global features in remote sensing imagery, rendering them less effective in

TABLE VIII
COMPARATIVE EXPERIMENTS ON AI4ARCTIC DATASET, GRAY REPRESENTS OUR PROPOSED METHOD

Method	SIC		SOD		FLOE		Params (M)	FLOPs (G)
	F1 (%)	OA (%)	F1 (%)	OA (%)	F1 (%)	OA (%)		
UNet	75.8	72.9	67.5	70.3	75.9	76.8	0.490	2.699
PspNet	76.0	72.8	69.1	70.8	75.0	75.8	0.977	2.139
LinkNet	75.6	72.8	68.3	69.5	73.8	74.9	5.804	3.463
Deeplabv3	73.1	71.3	72.3	72.5	72.3	74.0	15.939	17.352
UNet++	74.2	72.3	68.5	71.2	75.1	75.9	9.738	15.598
MMSeaIce	74.2	72.9	69.6	71.9	76.6	77.9	0.523	10.735
SwinUnet	75.6	72.7	66.9	68.4	69.9	71.0	50.057	4.635
SwinUPer	74.3	72.1	68.5	69.9	72.4	73.4	30.506	9.200
UniFormer	71.7	70.5	68.2	69.5	71.5	73.2	17.409	6.332
PoolFormer	74.4	71.7	66.8	68.0	71.3	72.6	23.415	7.943
AdaptSegNet	73.9	71.9	67.7	68.9	73.9	75.2	14.865	17.360
DAST	76.3	72.6	69.6	70.0	68.9	70.1	7.046	67.804
MFDA	78.0	73.7	75.8	77.7	79.8	79.7	9.421	4.746

MTL scenes. In contrast, the MFDA achieved the best overall results across the SIC, SOD, and FLOE tasks, particularly excelling on the SOD task. This can be attributed to the pretraining of the MLFMIM encoder, which enabled the model to learn common features across diverse scene images, and the subsequent introduction of domain adaptation techniques, which further harmonized the feature representations across different domains.

We present a visual comparison of the segmentation results obtained from various deep learning methods across different geographic locations and time periods. The results for the freezing season are illustrated in Figs. 8–11. The U-Net and LinkNet models exhibit a tendency for minor classification errors, with their delineation of ice-water boundaries appearing relatively ambiguous. Conversely, the PSPNet and DeepLabV3 models struggle to accurately capture the variability in SIC; however, their segmentation outputs tend to be more coherent and natural compared with U-Net and LinkNet, which may be attributed to the effects of the global pooling mechanism. The MMSeaIce model demonstrates strong performance in segmenting SIC and effectively captures floes, yet it exhibits significant misclassification results for SOD, likely due to the pronounced variations in sea ice types across different scenarios. The UNet++ architecture, with its nested encoding–decoding structure, excels at capturing multiscale feature information, yet it frequently produces incoherent misclassification artifacts when processing cross-scene images. The SwinUNet model tends to underestimate the presence of sea ice and encounters challenges in accurately representing the morphological characteristics of various ice types. Both SwinUper and UniFormer perform comparably in the ice-water separation task; however, the UniFormer model exhibits more pronounced errors in simulating the dynamic changes of the ice-water boundary, often failing to distinctly differentiate between ice cubes of varying sizes and confusing small ice cubes, pancake ice, and medium-sized ice cubes. In contrast, the PoolFormer architecture effectively captures boundary changes between different ice regions but often underestimates the density and size of the ice. The results of AdaptSegNet exhibit significant fluctuations along the ice-water separation boundary, with SIC gradually increasing from 10% to 100%, yet it often underestimates the overall SIC. Nevertheless, AdaptSegNet achieves some correct classifications in the SOD task, and the changes in the separation boundary are also reflected in the floe task. The DAST model does not demonstrate a clear trend in the ice-water separation boundary for the SIC task; however, it does exhibit a trend of change in the SOD and floe tasks, and its performance aligns more closely with the ground truth than with the actual situation, particularly when compared with CNN or Transformer-based methods. The introduction of deformable convolution [72] in the MFDA model enhances its performance in the FLOE task, enabling it to effectively capture the size characteristics of ice blocks and clearly delineate the boundaries between different ice types; however, it still exhibits limitations in ice-water separation.

The results for the melting season are presented in Fig. 12. CNN-based methods, such as U-Net, tend to underestimate SIC in the ice-water separation zone and produce inaccurate estimates in areas with high SIC boundaries. Transformer- and UperNet-based methods, including SwinUper, struggle to effectively classify results across varying densities; however, SwinUNet demonstrates improved performance in addressing this issue. The integration of Transformer architectures with U-shaped structures effectively delineates boundaries of different densities. In the case of the DA model, AdaptSegNet can accurately partition SIC over a wide range, although it still exhibits instances of misclassification. DAST tends to overestimate SIC, particularly in the intermediate regions. The MFDA model achieves the highest accuracy in SIC, providing overall close classification results, yet it still faces challenges in capturing the nuances of the ice-water melting process at the boundary. In the SOD task, methods based on U-shaped structures are prone to generating ambiguous classification results, whereas Transformer- and UperNet-based methods tend to produce more coherent outcomes. DA-based methods, such as AdaptSegNet, often yield incorrect classifications in areas where open water meets land boundaries. The MFDA model achieves the best overall performance in the SOD task; however, it occasionally misclassifies melted water as old ice in open water regions. In the floe classification task, U-shaped structure-based methods exhibit smoother boundaries between different categories of floes. In contrast, Transformer- and UperNet-based methods tend to focus on

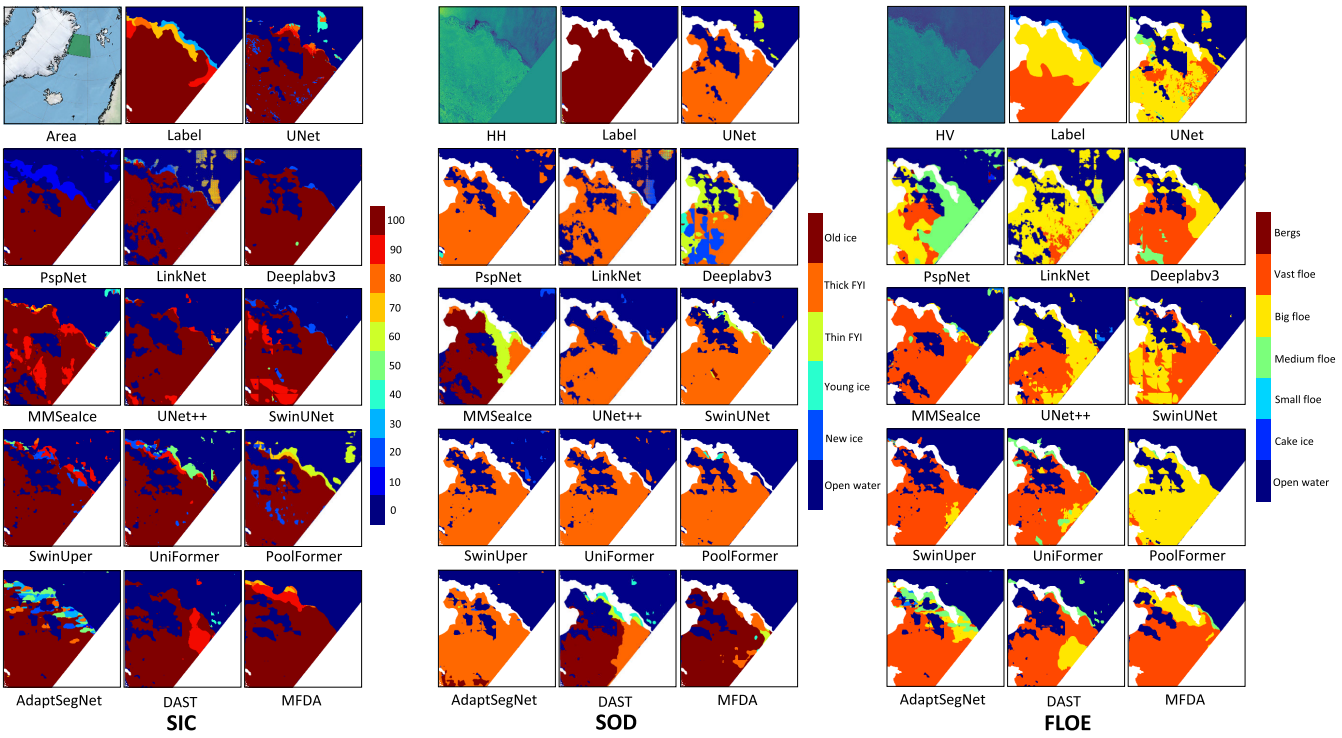

Fig. 8. This figure presents the sea ice mapping results in the SAR scene (ID: 20201104T171455_dmi) acquired through various models, located near the Greenland Sea. The corresponding methodology and image label are indicated below. Unmarked land or regions in the figure are depicted in white shading.

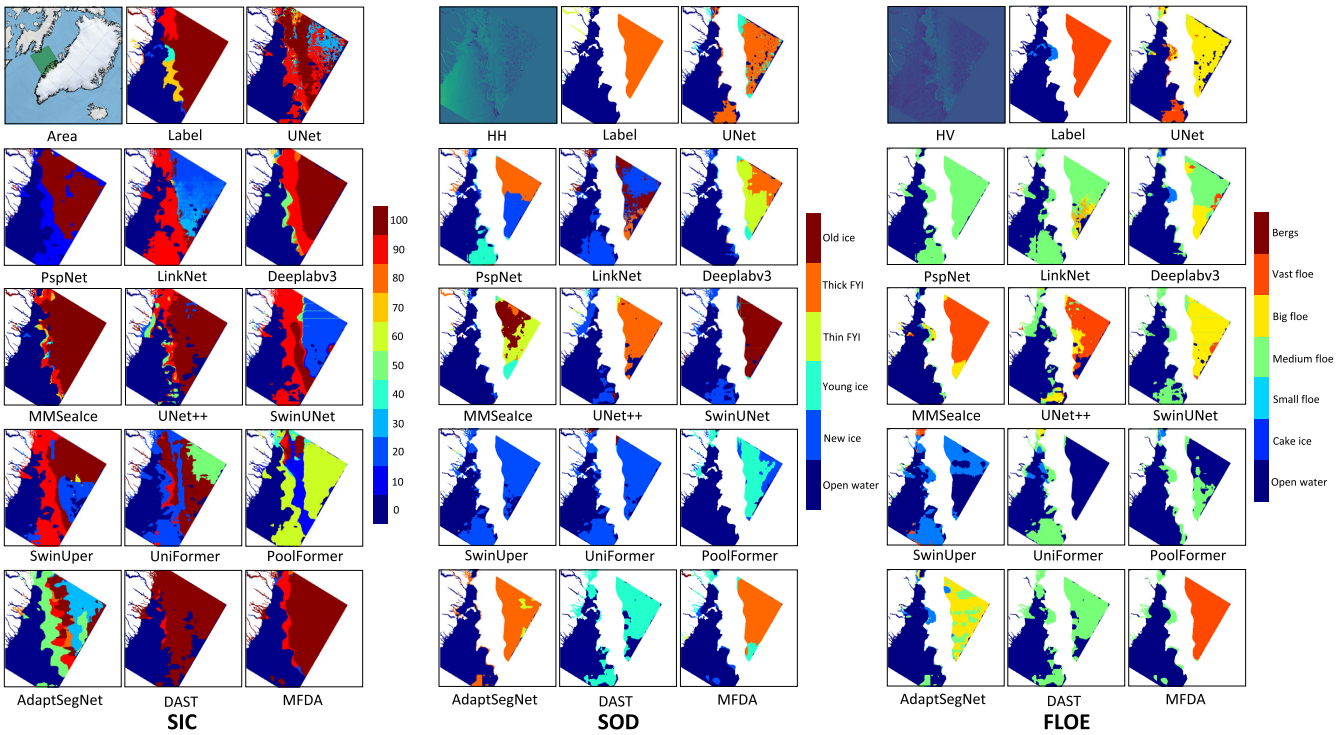

Fig. 9. This figure presents the sea ice mapping results in the SAR scene (ID: 20190330T102526_dmi) acquired through various models, located at the junction of Baffin Bay and the Labrador Sea. The corresponding methodology and image label are indicated below. Unmarked land or regions in the figure are depicted in white shading.

the global category of floes but struggle to differentiate between vast floes and big floes. The PoolFormer model faces challenges in recognizing the size of ice due to its lack of an attention mechanism. Although DA-based methods account for differences across various scenes, they do not adequately consider the morphological characteristics of different

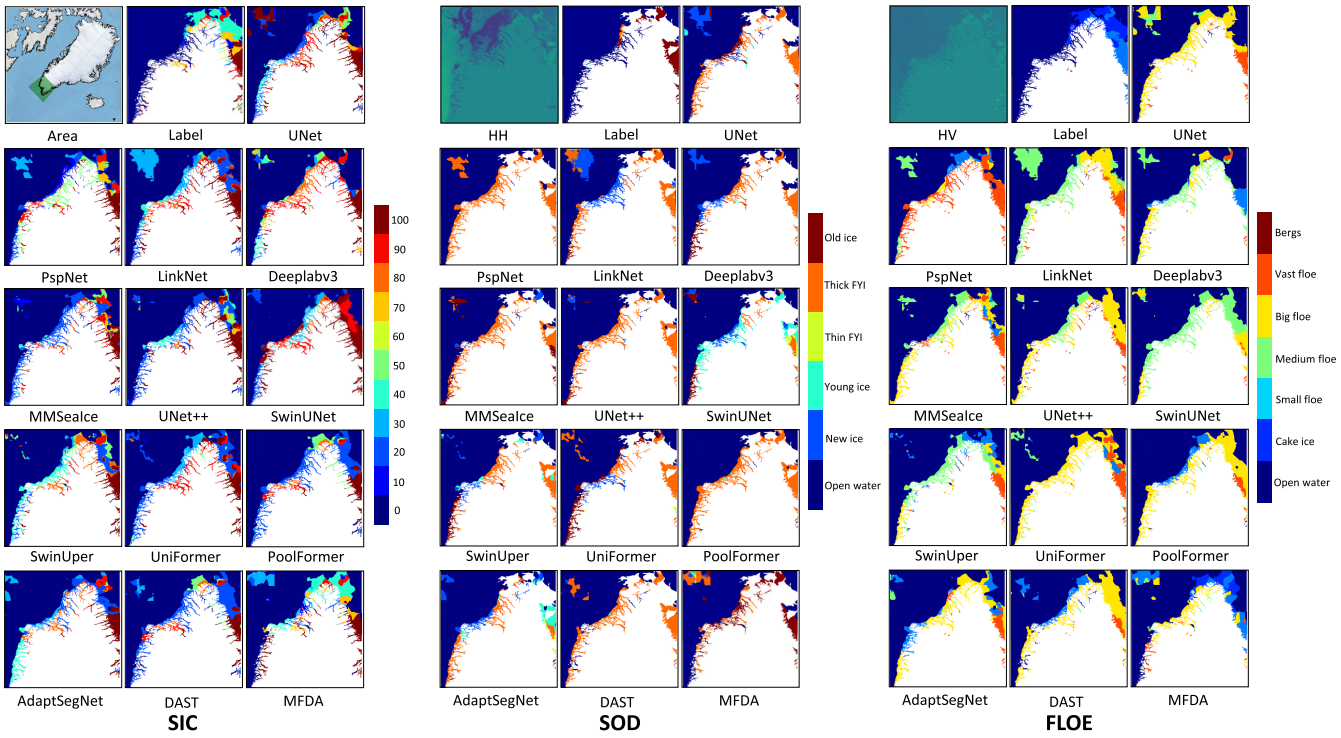

Fig. 10. This figure presents the sea ice mapping results in the SAR scene (ID: 20210402T203557_dmi) acquired through various models, located near the Labrador Sea. The corresponding methodology and image label are indicated below. Unmarked land or regions in the figure are depicted in white shading.

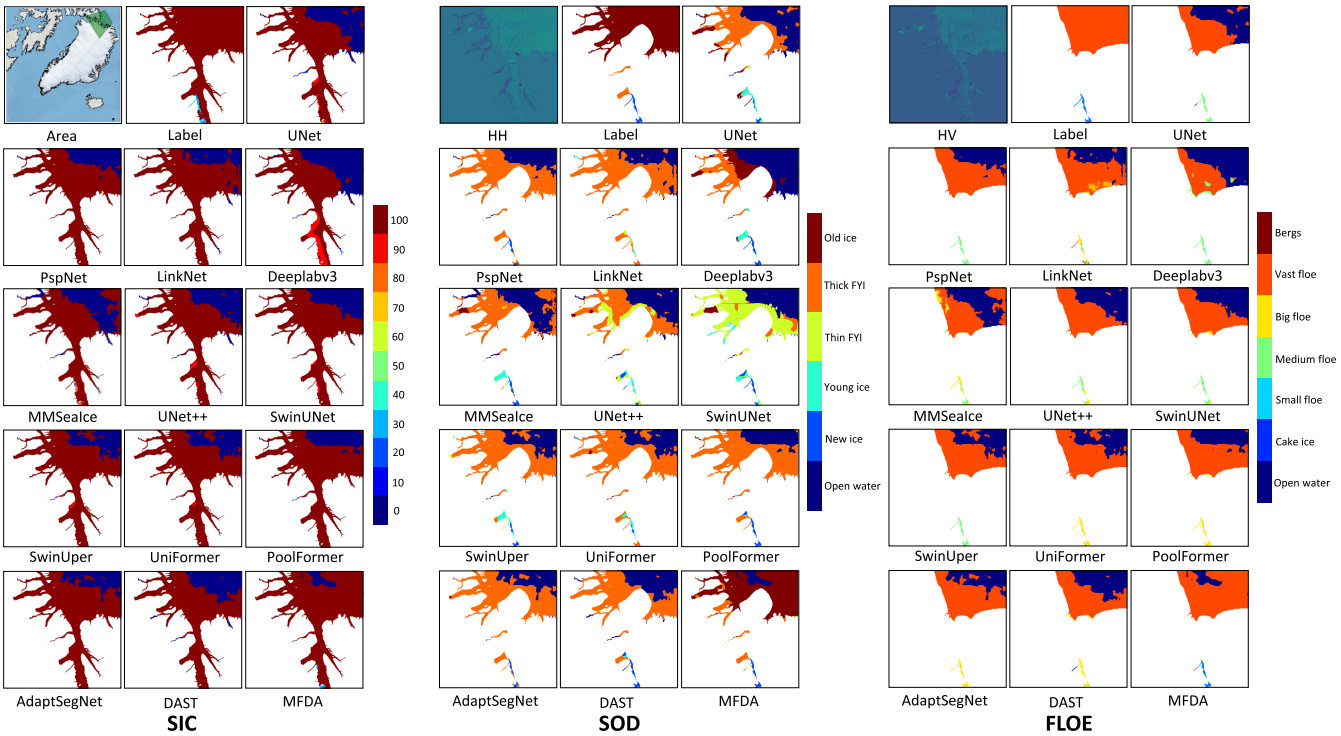

Fig. 11. This figure presents the sea ice mapping results in the SAR scene (ID: 20211214T115920_dmi) acquired through various models, located at the junction of Baffin Bay and the Arctic Ocean. The corresponding methodology and image label are indicated below. Unmarked land or regions in the figure are depicted in white shading.

floe categories. While MFDA demonstrates the best overall results in the FLOE classification task, it still encounters misclassification issues near open water, which may be attributed to ice fragmentation.

I. Module Ablation Experiment

We conducted ablation experiments to analyze the contributions of different modules, and the results are presented in Table IX. The combination of Convolutional Transformer

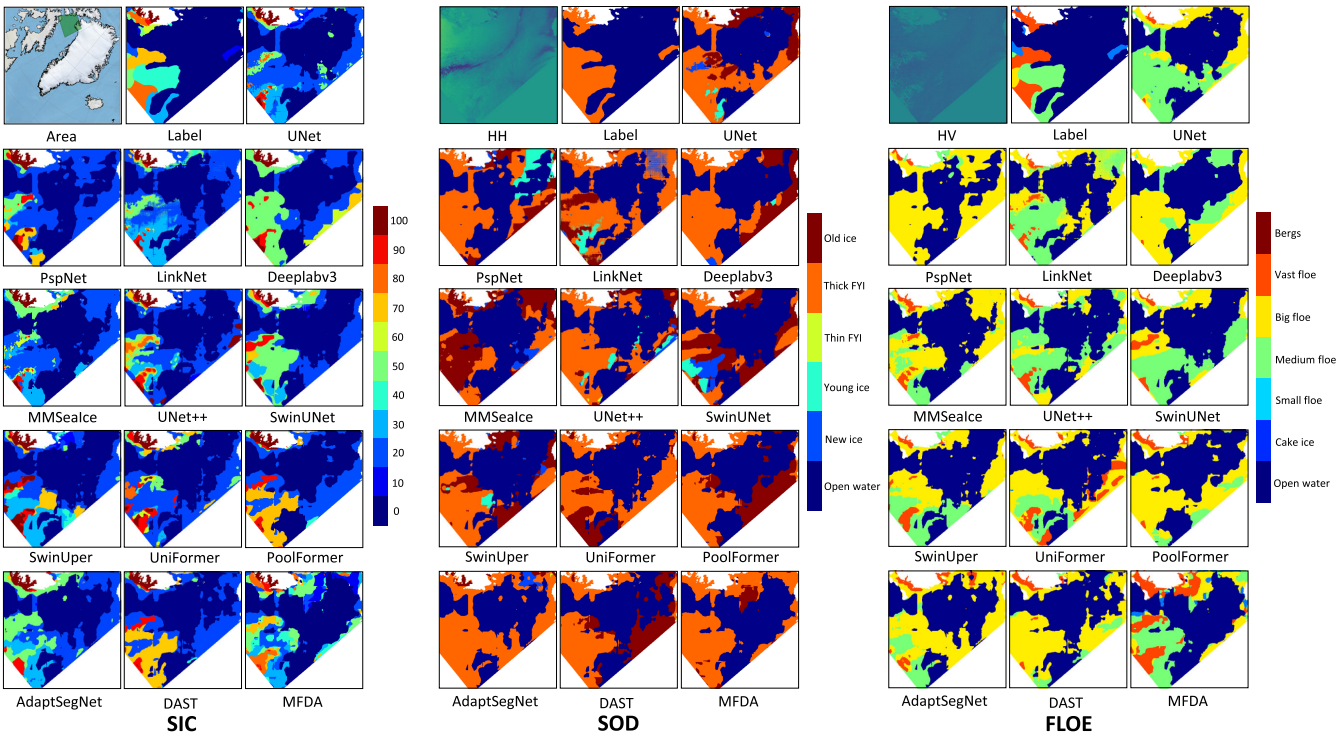

Fig. 12. This figure presents the sea ice mapping results in the SAR scene (ID: 20210706T115305_dmi) acquired through various models, located near the Baffin Bay. The corresponding methodology and image label are indicated below. Unmarked land or regions in the figure are depicted in white shading.

TABLE IX

ABLATION EXPERIMENTS OF DIFFERENT MODULES, WHERE CNN REPRESENTS THE COMPLETE USE OF CONVULATIONAL TRANSFORMER AS THE BACKBONE, TRANSFORMER REPRESENTS THE COMPLETE USE OF WINDOW TRANSFORMER AS THE BACKBONE, AND MIX REPRESENTS THE MIXTURE OF THE TWO. DISCOVERER AND CORRECTOR COME FROM DA, AND DEFCONV REPRESENTS DEFORMABLE CONVOLUTION

Method	CNN	Transformer	Mix	Discoverer	Corrector	DefConv	SIC		SOD		FLOE	
							F1 (%)	OA (%)	F1 (%)	OA (%)	F1 (%)	OA (%)
a	✓	✗	✗	✓	✓	✓	75.7	73.1	71.5	73.2	74.3	76.2
b	✗	✓	✗	✓	✓	✓	75.4	72.8	70.5	72.7	73.3	75.2
c	✗	✗	✓	✓	✓	✓	78.0	73.7	73.7	75.0	79.8	79.7
d	✗	✗	✓	✗	✗	✗	76.3	73.1	70.0	70.7	75.6	75.8
e	✗	✗	✓	✗	✗	✓	74.9	73.1	73.2	74.7	74.3	75.5
f	✗	✗	✓	✓	✗	✓	76.2	73.5	73.8	75.4	77.7	78.6
g	✗	✗	✓	✗	✓	✓	76.1	72.9	75.2	76.1	73.5	74.7
h	✗	✗	✓	✓	✓	✗	77.0	73.3	73.7	74.7	75.4	76.2

and Window Transformer significantly improved the model performance compared with CNN and Transformer baselines, with an $F1$-score increase of 0.023 and 0.026 for the SIC task, and an $F1$-score increase of 0.055 and 0.065 for the FLOE task. The discoverer and corrector components played a key role in the classification of SOD and FLOE. Removing the discoverer component reduced the model performance by 0.019 and 0.063 in $F1$-score for the SIC and FLOE tasks, respectively, but increased the $F1$-score for the SOD task by 0.015. Removing the corrector component reduced the $F1$-score for the SIC and FLOE tasks by 0.018 and 0.021, respectively, but it still promoted the SOD task performance, which may be related to the inherent error bias in the polynya classification of sea ice. The introduction of deformable convolution also improved the model, especially for the FLOE task, where it increased the $F1$-score and OA by 0.044 and 0.035, respectively. This is because deformable convolution is effective at capturing shape relationships, which are highly correlated with the FLOE task. Through the careful combination of these different modules, the MFDA achieved the best overall results.

J. Feature Visualization

Fig. 13 presents a visualization of the selected encoder output features from the proposed MFDA framework, with and without the application of DA. A comparative analysis was conducted on the first 16 feature maps. The results indicate that the feature maps generated with the inclusion of data augmentation exhibit a more refined and detailed appearance, capturing finer-grained characteristics such as object edges, contours, and textural structures. In contrast, the feature maps produced without data augmentation demonstrate a comparatively inferior visual quality. This visual comparison

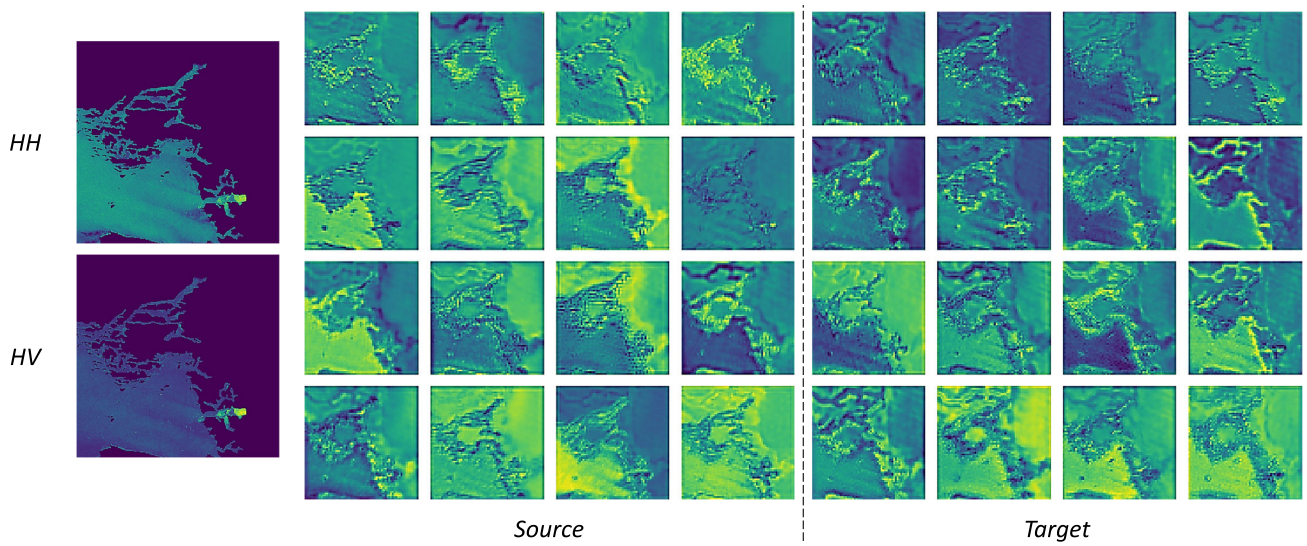

Fig. 13. Visualization of feature maps was conducted for the input and output of the DA module. The source represents the feature maps without the application of the DA module, while the target denotes the feature maps obtained after the application of the DA module.

provides empirical evidence supporting the effectiveness and importance of incorporating DA within the MFDA framework.

V. Conclusion

Sea ice classification and segmentation based on SAR has long been an indispensable and important research topic in the field of polar remote sensing. The exploration and commercial development of polar regions are inextricably linked to the rapid advancement of sea ice retrieval remote sensing technology. However, the well-designed and specialized segmentation methods developed thus far are largely only applicable to sea ice in relatively homogeneous single regions, and these methods are often tailored to a single retrieval task, without considering the inherent relationships between different sea ice parameters. This limitation seriously hinders the application deployment in areas without readily available reference data, as route planning and management tasks, such as policy making and risk assessment, require research on multi-task sea ice retrieval in diverse scenes.

To address this challenge, we studied the problem of cross-scene multi-task sea ice retrieval and proposed a unified solution. First, we investigated the fusion of multimodal data. We developed the MFDA, which utilizes a unified CNN and Transformer architecture to solve MTL based on the multimodal data fusion. Subsequently, we introduced a novel self-supervised method to enable the encoder component to learn a common representation of multimodal data across different scenes. Finally, we incorporated a DA module between the common representation and the multi-task prediction head to narrow the gap in data representation between different scenes. Extensive experiments conducted on the Ai4Arctic dataset demonstrate that the proposed MFDA model achieves SOTA segmentation performance, outperforming other CNN, Transformer, and DA-based methods across a wide range of key evaluation metrics.

In future work, we aim to extend the scope of our research to the Antarctic region in order to more comprehensively investigate the problem of cross-scene large-scale sea ice retrieval. In recent years, diffusion models have demonstrated remarkable performance in remote sensing tasks, such as image denoising, cloud removal, and image generation, which has garnered significant attention from the academic and industrial communities. Going forward, we will continue to explore the application potential of diffusion models in SAR denoising and generation, and assess their feasibility as data augmentation tools. Simultaneously, we will further investigate the application of conditional diffusion models to accurately estimate sea ice density, with the aim of improving cross-scene sea ice parameter retrieval.

Acknowledgment

The authors are grateful to the reviewers for their constructive feedback and insightful recommendations on this article. They would also acknowledge the providers of the AI4Arctic dataset and the organizers of the AutoICE competition.

References

[1] N. Wunderling, M. Willeit, J. F. Donges, and R. Winkelmann, "Global warming due to loss of large ice masses and Arctic summer sea ice," *Nature Commun.*, vol. 11, no. 1, p. 5177, Oct. 2020.

[2] C. Eayrs, D. Holland, D. Francis, T. Wagner, R. Kumar, and X. Li, "Understanding the seasonal cycle of Antarctic sea ice extent in the context of longer-term variability," *Rev. Geophys.*, vol. 57, no. 3, pp. 1037–1064, Sep. 2019.

[3] T. R. Andersson et al., "Seasonal Arctic sea ice forecasting with probabilistic deep learning," *Nature Commun.*, vol. 12, no. 1, p. 5124, Aug. 2021.

[4] H.-W. Lee, M.-I. Roh, and K.-S. Kim, "Ship route planning in Arctic ocean based on POLARIS," *Ocean Eng.*, vol. 234, Aug. 2021, Art. no. 109297.

[5] K. Kusahara, G. D. Williams, R. Massom, P. Reid, and H. Hasumi, "Spatiotemporal dependence of Antarctic sea ice variability to dynamic and thermodynamic forcing: A coupled ocean–sea ice model study," *Climate Dyn.*, vol. 52, nos. 7–8, pp. 3791–3807, Apr. 2019.

[6] M. Sudmanns et al., "Big Earth data: Disruptive changes in Earth observation data management and analysis?" *Int. J. Digit. Earth*, vol. 13, no. 7, pp. 832–850, Jul. 2020.

[7] W. Han et al., "A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 87–113, Aug. 2023.

[8] W. Huang et al., "Sea ice extraction via remote sensing imagery: Algorithms, datasets, applications and challenges," *Remote Sens.*, vol. 16, no. 5, p. 842, Feb. 2024.

[9] X. Chen et al., "MMSeaIce: A collection of techniques for improving sea ice mapping with a multi-task model," *Cryosphere*, vol. 18, no. 4, pp. 1621–1632, Apr. 2024.

[10] K. Kortum, S. Singha, G. Spreen, N. Hutter, A. Jutila, and C. Haas, "SAR deep learning sea ice retrieval trained with airborne laser scanner measurements from the MOSAiC expedition," *Cryosphere*, vol. 18, no. 5, pp. 2207–2222, May 2024.

[11] J. Yin et al., "Integrating remote sensing and geospatial big data for urban land use mapping: A review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, Dec. 2021, Art. no. 102514.

[12] Y. Han, Y. Liu, Z. Hong, Y. Zhang, S. Yang, and J. Wang, "Sea ice image classification based on heterogeneous data fusion and deep learning," *Remote Sens.*, vol. 13, no. 4, p. 592, Feb. 2021.

[13] X.-M. Li, Y. Sun, and Q. Zhang, "Extraction of sea ice cover by Sentinel-1 SAR based on support vector machine with unsupervised generation of training data," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3040–3053, Apr. 2021.

[14] P. Roy, M. Behera, and S. Srivastav, "Satellite remote sensing: Sensors, applications and techniques," *Proc. Nat. Acad. Sci., India A, Phys. Sci.*, vol. 87, pp. 465–472, Dec. 2017.

[15] A. Bhardwaj, L. Sam, A. Bhardwaj, and F. J. Martín-Torres, "LiDAR remote sensing of the cryosphere: Present applications and future prospects," *Remote Sens. Environ.*, vol. 177, pp. 125–143, May 2016.

[16] X. Sun, X. Zhang, W. Huang, Z. Han, X. Lyu, and P. Ren, "Sea ice classification using mutually guided contexts," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4204019.

[17] J. E. Box et al., "Key indicators of Arctic climate change: 1971–2017," *Environ. Res. Lett.*, vol. 14, no. 4, 1971, Art. no. 045010.

[18] J. E. Walsh et al., "Extreme weather and climate events in northern areas: A review," *Earth-Sci. Rev.*, vol. 209, Oct. 2020, Art. no. 103324.

[19] M. T. Hosain, J. R. Jim, M. F. Mridha, and M. M. Kabir, "Explainable AI approaches in deep learning: Advancements, applications and challenges," *Comput. Electr. Eng.*, vol. 117, Jul. 2024, Art. no. 109246.

[20] Y. Chen, Q. Yan, and W. Huang, "MSSFF: Advancing hyperspectral classification through higher-accuracy multistage spectral–spatial feature fusion," *Remote Sens.*, vol. 15, no. 24, p. 5717, Dec. 2023.

[21] Y. Ren, X. Li, X. Yang, and H. Xu, "Development of a dual-attention U-Net model for sea ice and open water classification on SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[22] W. Song et al., "Automatic sea-ice classification of SAR images based on spatial and temporal features learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 9887–9901, Dec. 2021.

[23] I. de Gelis, A. Colin, and N. Longépé, "Prediction of categorized sea ice concentration from Sentinel-1 SAR images based on a fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5831–5841, 2021.

[24] Y. Huang, Y. Ren, and X. Li, "Deep learning techniques for enhanced sea-ice types classification in the Beaufort sea via SAR imagery," *Remote Sens. Environ.*, vol. 308, Jul. 2024, Art. no. 114204.

[25] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856.

[26] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102926.

[27] S. Sengupta et al., "A review of deep learning with special emphasis on architectures, applications and recent trends," *Knowl.-Based Syst.*, vol. 194, Apr. 2020, Art. no. 105596.

[28] T. Zhang et al., "Deep learning based sea ice classification with Gaofen-3 fully polarimetric SAR data," *Remote Sens.*, vol. 13, no. 8, p. 1452, Apr. 2021.

[29] B. Dowden, O. D. Silva, W. Huang, and D. Oldford, "Sea ice classification via deep neural network semantic segmentation," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11879–11888, May 2021.

[30] X. Chen, K. A. Scott, M. Jiang, Y. Fang, L. Xu, and D. A. Clausi, "Sea ice classification with dual-polarized SAR imagery: A hierarchical pipeline," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2023, pp. 224–232.

[31] R. Huang, C. Wang, J. Li, and Y. Sui, "DF-UHRNet: A modified CNN-based deep learning method for automatic sea ice classification from Sentinel-1A/B SAR images," *Remote Sens.*, vol. 15, no. 9, p. 2448, May 2023.

[32] K. Kortum, S. Singha, and G. Spreen, "Robust multiseasonal ice classification from high-resolution X-band SAR," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408512.

[33] A. Kucik and A. Stokholm, "AI4SeaIce: Selecting loss functions for automated SAR sea ice concentration charting," *Sci. Rep.*, vol. 13, no. 1, p. 5962, Apr. 2023.

[34] A. Stokholm, A. Kucik, N. Longépé, and S. M. Hvidegaard, "AI4SeaIce: Task separation and multistage inference CNNs for automatic sea ice concentration charting," *EGUsphere*, vol. 2023, pp. 1–25, Jun. 2023.

[35] A. Stokholm, T. Wulf, A. Kucik, R. Saldo, J. Buus-Hinkler, and S. M. Hvidegaard, "AI4SeaIce: Toward solving ambiguous SAR textures in convolutional neural networks for automatic sea ice concentration charting," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4304013.

[36] M. Patel, X. Chen, L. Xu, Y. Chen, K. A. Scott, and D. A. Clausi, "Region-level ground truth can produce pixel-level segmentation for sea ice concentration," *J. Comput. Vis. Imag. Syst.*, vol. 9, no. 1, pp. 88–89, 2023.

[37] J. Buus-Hinkler et al., "AI4Arctic sea ice challenge dataset, DTU [code and data set]," 2022, doi: 10.11583/DTU.c.6244065.v2.

[38] M. Patel, X. Chen, L. Xu, Y. Chen, K. A. Scott, and D. A. Clausi, "Region-level labels in ice charts can produce pixel-level segmentation for sea ice types," 2024, *arXiv:2405.10456*.

[39] F. Qi, X. Yang, and C. Xu, "A unified framework for multimodal domain adaptation," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 429–437.

[40] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[41] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2019, pp. 2517–2526.

[42] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2507–2516.

[43] M. Zhang, X. Zhao, W. Li, Y. Zhang, R. Tao, and Q. Du, "Cross-scene joint classification of multisource data with multilevel domain adaption network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp. 11514–11526, Aug. 2023.

[44] M. Kim, S. Joung, S. Kim, J. Park, I.-J. Kim, and K. Sohn, "Cross-domain grouping and alignment for domain adaptive semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 1799–1807.

[45] Z. Xiong, S. Chen, Y. Wang, L. Mou, and X. X. Zhu, "GAMUS: A geometry-aware multi-modal semantic segmentation benchmark for remote sensing data," 2023, *arXiv:2305.14914*.

[46] Y. Chen, Q. Yan, and W. Huang, "MFTSC: A semantically constrained method for urban building height estimation using multiple source images," *Remote Sens.*, vol. 15, no. 23, p. 5552, Nov. 2023.

[47] Z. Gao et al., "Joint learning of semantic segmentation and height estimation for remote sensing image leveraging contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5614015.

[48] D. Muhtar, X. Zhang, P. Xiao, Z. Li, and F. Gu, "CMID: A unified self-supervised learning framework for remote sensing image understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607817.

[49] T. Zhang et al., "Object-centric masked image modeling based self-supervised pretraining for remote sensing object detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5013–5025, 2023.

[50] Z. Li, H. Chen, J. Wu, J. Li, and N. Jing, "SegMind: Semisupervised remote sensing image semantic segmentation with masked image modeling and contrastive learning method," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4408917.

[51] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.

[52] X. Cao, H. Lin, S. Guo, T. Xiong, and L. Jiao, "Transformer-based masked autoencoder with contrastive loss for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5524312.

[53] J. Lin, F. Gao, X. Shi, J. Dong, and Q. Du, "SS-MAE: Spatial–spectral masked autoencoder for multisource remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531614.

[54] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9653–9663.

[55] Y. Chen and Q. Yan, "LFSMIM: A low-frequency spectral masked image modeling method for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[56] H. Wang, Y. Tang, Y. Wang, J. Guo, Z.-H. Deng, and K. Han, "Masked image modeling with local multi-scale reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2122–2131.

[57] X. Chen et al., "A comparative study of data input selection for deep learning-based automated sea ice mapping," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 131, Jul. 2024, Art. no. 103920.

[58] K. Li et al., "UniFormer: Unifying convolution and self-attention for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12581–12600, Oct. 2023.

[59] Q. Yan and W. Huang, "Sea ice remote sensing using GNSS-R: A review," *Remote Sens.*, vol. 11, no. 21, p. 2565, Nov. 2019.

[60] Q. Yan and W. Huang, "Sea ice sensing from GNSS-R data using convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 10, pp. 1510–1514, Sep. 2018.

[61] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[62] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[63] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[64] I. Sudakow, V. K. Asari, R. Liu, and D. Demchev, "MeltPondNet: A swin transformer U-Net for detection of melt ponds on Arctic sea ice," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8776–8784, 2022.

[65] N.-C. Ristea, A. Anghel, and M. Datcu, "Sea ice segmentation from SAR data by convolutional transformer networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2023, pp. 168–171.

[66] Y. Chen, P. Liu, J. Zhao, K. Huang, and Q. Yan, "Shallow-guided transformer for semantic segmentation of hyperspectral remote sensing imagery," *Remote Sens.*, vol. 15, no. 13, p. 3366, Jun. 2023.

[67] W. Yu et al., "MetaFormer baselines for vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 896–912, Feb. 2024.

[68] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2022.

[69] J. Choi, T. Kim, and C. Kim, "Self-ensembling with GAN-based data augmentation for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6830–6840.

[70] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7178–7193, Oct. 2020.

[71] F. Yu, M. Zhang, H. Dong, S. Hu, B. Dong, and L. Zhang, "DAST: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10754–10762.

[72] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[73] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[74] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[75] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese, "Which tasks should be learned together in multi-task learning?" in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9120–9132.

[76] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: cross-modal fusion for RGB-x semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, Dec. 2023.

[77] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[78] Y. Chen and Q. Yan, "Unlocking the potential of CYGNSS for pan-tropical inland water mapping through multi-source data and transformer," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 133, Sep. 2024, Art. no. 104122.

[79] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[80] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[81] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[82] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. 18th Int. Workshop Deep Learn. Med. Image Anal.*, vol. 11045, Granada, Spain. Cham, Switzerland: Springer, Sep. 2018, pp. 3–11.

[83] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.