A Novel Approach to Incomplete Multimodal Learning for Remote Sensing Data Fusion

Yuxing Chen, *Graduate Student Member, IEEE* Maofan Zhao, Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—The mechanism of connecting multimodal signals through self-attention operation is a key factor in the success of multimodal Transformer networks in remote sensing data fusion tasks. However, traditional approaches assume access to all modalities during both training and inference, which can lead to severe degradation when dealing with modal-incomplete inputs in downstream applications. To address this limitation, we propose a novel approach to incomplete multimodal learning in the context of remote sensing data fusion and the multimodal Transformer. This approach can be used in both supervised and self-supervised pre-training paradigms. It leverages the additional learned fusion tokens in combination with modality attention and masked self-attention mechanisms to collect multimodal signals in a multimodal Transformer. The proposed approach employs reconstruction and contrastive loss to facilitate fusion in pre-training, while allowing for random modality combinations as inputs in network training. Experimental results show that the proposed method delivers state-of-the-art performance on two multimodal datasets for tasks such as building instance / semantic segmentation and land-cover mapping when dealing with incomplete inputs during inference.**

***Index Terms*—Data Fusion, Multimodal, Transformer, Remote Sensing.**

I. Introduction

REMOTE sensing becomes more and more important in various Earth Observation (EO) tasks. With the increasing availability of multimodal RS data, researchers now can develop more diverse downstream applications. Despite the abundance of multimodal remote sensing data, each modality captures only certain specific properties and, therefore, cannot thoroughly describe the observed scenes. Thus the use of single-mode data results in limitations in many applications. Multimodal RS data fusion addresses these limitations [1]. For instance, synthetic aperture radar (SAR) provides physical structure information, while LiDAR collects both structure and depth information [2]. Meanwhile, multispectral (MS) and hyperspectral (HS) sensors measure radiation reflectance across different wavelengths of the electromagnetic spectrum. By merging the complementary information present in multimodal data, it is possible to improve the accuracy and reliability of many data analysis tasks, such as change detection [3] and land-cover mapping [4]. To integrate the complementary information provided by different sensors and remote sensing products (e.g., Land-Use and Land-Cover), traditional methods [5] exploit handcrafted features based on domain-specific knowledge and fusion strategies that often are not able to capture all the information present in the data.

Due to the growth of artificial intelligence methodologies, deep learning shows great potential in modelling the complex relationships between different modality data and is widely used in remote sensing data fusion tasks. Among the others, there are three main multimodal RS data fusion scenarios, SAR-optical [6]–[9], LiDAR-optical [2], [10]–[12], and image-map [13], [14], where the deep Convolutional Neural Networks (CNNs) and Transformer networks are widely used. Nevertheless, deep CNNs methods assume that all modalities are available during training and inference, which can be a limiting factor in practical applications, as data collection processes may miss some data sources for some instances. In such cases, existing multimodal data fusion methods may fail to deal with incomplete modalities, leading to severe degradation in performance. The approach used in this situation is called incomplete multimodal learning and aims at learning methods that perform inference which is robust to any subset of available modalities. A simple strategy for incomplete multimodal learning using CNNs is to synthesize the missing modalities using generative models. For instance, Generative Adversarial Networks (GANs) can effectively overcome the problems arising from missing or incomplete modalities in building footprint segmentation [15]. Another set of methods explores knowledge distillation from the present modality to incomplete modalities. In this context, Kampffmeyer et al. [16] proposed to use an additional network, the hallucination network, for mitigating missing data modalities in the testing of urban land-cover classification tasks. The network takes a modality as input that is assumed to be available during both training and testing, trying to learn a mapping function from this modality to the missing one.

Although promising results are obtained, such methods have to train and deploy a specific model for each subset of missing modalities, which is complicated and often unreliable in downstream tasks. Moreover, all these methods require complete modalities during the training process. Recent incomplete multimodal learning methods for downstream tasks focus on learning a unified model, instead of a bunch of distilled networks. In this context, the modality-invariant fusion embedding across different modalities may contribute to more robust performance, especially when one or more modalities are missing. As a competitive multimodal data fusion model, Transformer does not need to access all modalities in the network training and inference thanks to its flexibility and

Y. Chen, M. Zhao and L. Bruzzone are with the Department of Information Engineering and Computer Science, University of Trento, 38122 Trento, Italy (e-mail: yuxing.chen@unitn.it; mfzhao1998@163.com; lorenzo.bruzzone@unitn.it). M. Zhao is also with the Aerospace Information Research Institute, Chinese Academy of Sciences, and also with the University of Chinese Academy of Sciences, Beijing 100049, China.

Corresponding author: L. Bruzzone

sequence modelling strategy, which can be effective in both scenarios: with and without missing modalities. Current works exploited Transformers for multimodal RS data fusion in a complete fusion scenario, such as lidar and hyperspectral data fusion [17]. For incomplete multimodal data fusion, MBT [18] and Zorro [19] propose to fuse audio and video data using learnable tokens in the Transformer network. However, the definition of a dedicated Transformer for incomplete multimodal learning in remote sensing tasks has not been addressed yet and the existing multimodal RS data fusion methods do not allow missing data in the training process. Moreover, Ma et al. [20] point out that the vanilla Transformer tends to be overfitted on one modality input.

Another limitation in the technique is that most multimodal data fusion methods are based on the supervised learning paradigm. Supervised approaches are task-specific and have limitations to be generalized to other tasks. Moreover, training on a large amount of multimodal data is cost expensive and collecting an adequate number of labeled data for each task is challenging for end-users. Thus, the research community usually relies on a few fine-tuning steps on a pre-trained model to adapt a network to a specific task. Pre-training without supervision has gained a lot of attention as it is more general and does not require labeled data. The self-supervised learning method for SAR-optical feature fusion [3] is an example of such an approach. However, this pre-training approach needs to access all modalities during network training.

In order to address the aforementioned issue, this paper proposes to exploit Transformer to build a unified model for incomplete multimodal learning for remote sensing tasks, which can be used in both the supervised and self-supervised pre-training paradigms. This is achieved by using additional learned fusion tokens for multimodal signal collection in the network. However, only using the additional learned fusion token cannot capture enough information from other modality tokens. In this context, we use a modality attention block to further distill different modality information to fusion tokens. Using this technique, the proposed approach can leverage reconstruction and contrastive loss to build fusion across the different modalities in pre-training. Moreover, it can use a random modality combination training strategy in supervised training. This makes the learning and inference feasible also when incomplete modality data are given as input.

The three main contributions of this paper consist in: (1) we propose to use modality attention and masked self-attention in multimodal Transformer to build additional fusion tokens across different modalities, which enable both contrastive and mask-reconstruction pre-training for incomplete multimodal inputs; (2) based on the proposed approaches, we use the random modality combination training strategy in downstream tasks, which ensures task performance with incomplete inputs on inference; (3) we benchmark our approach on two datasets: the public DFC2023 track2 and the created quadruplet dataset, obtaining results that show that the proposed approach can be pre-trained on a large-scale remote sensing multimodal dataset in a self-supervised manner. The proposed approach achieves state-of-the-art performance when compared with the vanilla multimodal Transformer [18] on RS.

The rest of this paper is organized as follows. Section II presents the related works on multimodal RS data fusion, multimodal masked autoencoder and multimodal Transformer. Section III introduces the proposed approach by describing the network architecture, modality attention, masked self-attention, mask-reconstruction pre-training and contrastive pre-training as well as the random modality combination training strategy. The descriptions of the datasets, network setup, experimental settings and downstream tasks are given in Section IV. Experimental results obtained on building instance/semantic segmentation and LULC (Land-use Land-cover) mapping tasks as well as the ablation studies are illustrated in Section IV. Finally, Section V concludes the paper.

II. Related Works

A. Multimodal RS Data Fusion

In recent years, deep learning methods have been widely used in multimodal RS data fusion, including LiDAR-optical [2], [10]–[12], SAR-optical [6], [7], [7]–[9], and image-map fusion [13], [14]. In the case of LiDAR-optical data fusion, Paisitkriangkrai et al. [21] propose fusing optical and LiDAR data through concatenating deep and expert features as inputs to random forests. Several advanced techniques have subsequently been developed, with the aim of enhancing feature extraction ability. Audebert et al. [22] suggest the use of deep fully convolutional networks to investigate the early and late fusion of LiDAR and multispectral data. Similarly, Chen et al. [23] employ a two-branch network to separately extract spectral-spatial-elevation features, followed by a fully connected layer to integrate these heterogeneous features for final classification. Other novel fusion strategies are also introduced, such as the use of a cross-attention module [24], a reconstruction-based network [25], and a graph fusion network [26]. A recent study proposes a multimodal Transformer network to fuse LiDAR and hyperspectral images for classification [17]. Similar to LiDAR-optical fusion, many researchers also develop the Digital Surface Model (DSM) and optical fusion methods, where the DSM can be acquired by stereo-optical images. Also, SAR-optical data fusion widely adopts deep learning methods. For example, Kussul *et al.* [9] explore the deep CNNs in SAR-optical fusion for LULC classification and demonstrate their superiority with respect to traditional MLP classifiers. A recent study proposes a deep learning architecture, namely TWINNS, to fuse Sentinel-1 and Sentinel-2 time series data in land-cover mapping [8]. Similarly, Adrian *et al.* [7] use a 3-dimensional deep learning network to fuse multi-temporal Sentinel-1 and Sentinel-2 data for mapping ten different crop types, as well as water, soil and urban area. Map data, such as topography, land use, road and census data, may be combined with remotely sensed data to improve the accuracy of image classification, object recognition, and change detection. For example, Sun et al. [27] present a method of data fusion of GIS and RS using a neural network with unchanging data memory structure based on users' aim. Xu et al. [14] perform road extraction based on satellite images and partial road maps using a two-branch partial to complete network.

B. Multimodal Masked Autoencoder

The Multimodal Masked Autoencoder (MultiMAE) [28] is a novel self-supervised learning algorithm that demonstrates state-of-the-art performance on various vision benchmarks. Instead of relying on a contrastive objective, the MAE utilizes a pretext task that involves reconstructing masked patches of each input modality. It is based on a standard single-modal ViT and modality-specific encoders. The encoder is equipped with 2-D sine-cosine positional embeddings following the linear projection. MultiMAE does not make use of modality-specific embeddings, as the bias term in each linear projection is sufficient. MultiMAE employs a separate decoder for each task that is responsible for reconstructing the masked-out tokens from the visible tokens. The input to each decoder is a full set of visible tokens from all different modalities, including the learnable modality embeddings with 2-D sine-cosine positional embeddings. The input is then followed by MLPs and Transformer blocks. Only the masked tokens are considered in the loss calculation.

Suppose one of the input modalities is a tensor of dimensions $I \in R^{C\times H\times W}$, where H, W are the height and width of the image, respectively, and C is the number of channels. The input data is initially divided into non-overlapping patches $S \in R^{L\times P^2C}$, where P is the height and width of the patch, and $L = (H/P) \times (W/P)$ is the number of patches. These patches are then transformed into a sequence of embedded patch tokens $S' \in R^{L\times D}$, using a patch embedding function $f_p : R^{P^2C} \rightarrow R^D$. A fraction p_m of the sequence tokens is randomly masked, and the remaining visible tokens are fed into an encoder, which is a Vision Transformer (ViT). Due to the lack of positional information, additional positional embeddings are then added to patch embeddings to capture the spatial location of the patches. Each modality-specific decoder is composed of multiple transformer blocks that are trained for all tokens, where the masked tokens are replaced as the initialized learnable tokens. Each modality-specific decoder produces a modality-specific reconstruction, which is compared to the corresponding modality data using mean-squared error (MSE) loss, computed only on masked patches. Positional encoding allows the transformer to encode positional information. The positional encoding is:

$$\text{Encode}(k, 2i) = \sin\frac{k}{\Omega^{\frac{2i}{d}}}, \text{Encode}(k, 2i+1) = \cos\frac{k}{\Omega^{\frac{2i}{d}}} \quad (1)$$

Here, k is the position, i is the index of feature dimension in the encoding, d is the number of possible positions, and Ω is a large constant. The position is defined as the index of the patch along the x or y axis. Therefore, k ranges from 0 to H/P or W/P. This encoding provides two unique dimensions, one for x and one for y coordinates, which are concatenated for the final encoding representation.

The mask sampling strategy employed in MultiMAE plays a crucial role in achieving predictive coding across different modalities. This sampling strategy ensures that most modalities are represented to similar degrees. MultiMAE adopts a symmetric Dirichlet distribution to select the proportion of tokens per modality λ ($\lambda_i \sim Dir(\alpha)$), where $\sum \lambda_i = 1, \lambda > 0$. The concentration parameter $\alpha > 0$ controls the sampling. For simplicity and better representation parameter α is set to 1 in MultiMAE.

C. Multimodal Transformer

The self-attention blocks of Transformers build a natural bridge among multimodal signals in a unified architecture. Differently from the CNNs that use one network for each modality, the Transformer only use the same main architecture for all modalities with a modal-specific projector. Transformers integrate input tokens from all modalities into a single representation, while CNNs fuse features of each modality through concatenation or tensor fusion. However, such explicit integration requires the presence of all modalities during training, which undermines the pipeline in case of a missing modality. In contrast, Transformers use self-attention to embed a holistic multimodal representation and handle the absence of modalities by applying a mask on the attention matrix. Thus, multimodal Transformers are more adaptable to deal with modal-incomplete inputs. In addition, an easy-to-train model is vital for multimodal learning. The training load of a regular multimodal backbone increases as more modalities are added. This happens because the backbone typically contains separate sub-models for each modality, which must be trained individually. Instead, Transformers process modalities altogether in a single model, significantly reducing the training load.

However, Transformer models exhibit significant deterioration in performance with modal-incomplete inputs, especially in the context of multimodal inference where Transformer models tend to overfit the dominating modalities. To overcome this challenge, MBT [18] builds a multimodal architecture for video and audio, by using an additional fusion token to force information among different modalities to pass through by using cross-attention. However, the representation of each modality can also access to the others in MBT, which means they are not independent. In [19], a modality-aware masking mechanism is used in all attention operations to isolate the allocation of latent representations of individual modalities, which leads to a representation that is partially unimodal (i.e., part of the representation attends to a single modality) and partially multimodal (i.e., part of the representation attends to all modalities), thereby allowing for the use of contrastive learning.

III. Methodology

In this section, we describe the proposed incomplete multimodal fusion architecture with additional learned fusion tokens, modality attention and masked self-attention. This is done using as an illustration case, an optical-SAR-Digital Elevation Model(DEM)-MAP data fusion example. Then, we introduce the details of both pre-training using reconstruction and contrastive losses, as well as those of training using random modality combination on downstream tasks (see Fig. 1).

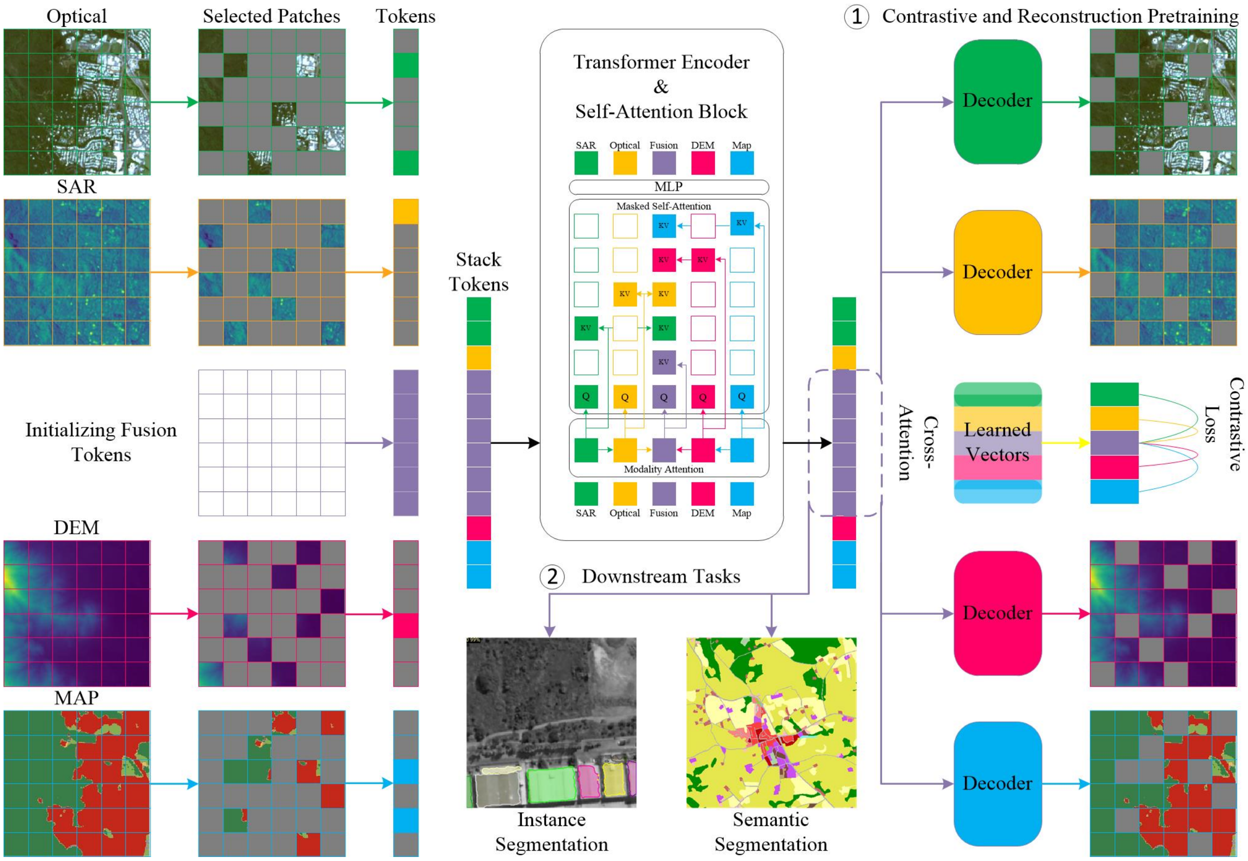


Fig. 1. Overview of the proposed framework. The inputs to our model are optical images, SAR images, DEM and Maps. Each of those inputs is patched using a 2D convolution and projected to feature vectors. All inputs are concatenated with a set of learnable fusion tokens and added to the position embedding. Next, we process these inputs through the Transformer Encoder, where the modality attention and the masked self-attention strategy are applied. (1) In pre-training, task-specific decoders reconstruct the masked patches by using the output fusion tokens. Meanwhile, the global vectors of each modality and fusion tokens are output using cross-attention, which allows for the use of contrastive loss between each modality and corresponding fusion tokens. (2) In the supervised training, the proposed framework can be trained on a specific downstream task by using a random modality combination training strategy.

A. Network Architecture

The main architecture of the proposed approach is a ViT with modality-specific patch projection layers for each input modality. In detail, patches of each modality are projected to tokens using a specific linear projection for each modality. In this work, we use a 2D convolution to extract 16×16 patches and project them to the input dimension D. Next, position embeddings are added to the projected vectors so that the model is able to localize and distinguish each embedded patch. In addition to the multimodal input data, the learnable fusion tokens are introduced as one of the inputs. Differently to the bottleneck fusion tokens in MBT [18] and Zorro [19], we use the spatial tokens for dense downstream tasks, which have the same number of tokens of full input patches. In order to get local features, we add 2D sine-cosine positional embeddings on the spatial fusion tokens and use the modality attention to aggregate all modality information to fusion tokens. Then the projected patches together with the learnable tokens are concatenated into a sequence of tokens and given as input to the same Transformer encoder with masked self-attention. Since all our input data have a 2D structure, we add 2D sine-cosine positional embeddings after linear projection. Following the setting of MultiMAE, we do not consider any modality-specific positional embedding.

B. Modality Attention

We employ a modality attention mechanism to seamlessly integrate diverse modality input embeddings into learned fusion tokens for enhancing the feature learning capabilities. The modality fusion block is constituted by a succession of transformer layers, each comprising Multi-Headed Cross Attention (MCA), Layer Normalization (LN), and Multilayer Perceptron (MLP) blocks. Let us consider a multimodality input $z^l = [z^l_o, z^l_s, z^l_d, z^l_m]$, encompassing an optical token, a SAR token, a DEM token, and a map token, alongside a fusion token z^l_f. We denote a transformer layer within the fusion block as $z^{l+1}_f = Transformer([z^l_f, z^l])$, expressed as:

$$\begin{aligned} z^l_f &= MCA(LN([z^l_f, z^l])) + z^l_f \\ z^{l+1}_f &= MLP(LN(z^l_f)) + z^l_f \end{aligned} \tag{2}$$

Here, the MCA operation performs dot-product attention, with queries as linear projections of the fusion token and keys/values as linear projections of each modality token. In instances where a modality is absent, we substitute the initialized mask token z_{mask} to account for the different number of input modalities at each location due to the use of masking.

C. Masked Self-Attention

Masked self-attention is the key block of multimodal Transformer in contrastive pre-training. Using masked attention, we force part of the representation to attend only to itself, while other parts can attend to the whole representation. In the considered illustration case, the main goal of this approach is to split the representation into five parts: a part which only focuses on Optical tokens, a part which focuses on SAR tokens, a part which focuses on DEM tokens, a part which focuses on MAP tokens, and the fusion tokens which consider the whole representation. In this architecture, the self-attention in each layer and the cross-attention in the last layer both used this masking strategy. Here we introduce the masking binary tensor m that specifies which vectors can access each other. Entries of the masking matrix are $m_{i,j} = 1$ if information can flow from latent j to latent i. Versus, we set $m_{i,j} = 0$. The mask is applied to the standard attention output operation, which performs on keys k, values v and queries q, can be expressed as:

$$o_i = \sum_j \frac{m_{ij}\exp\left(\frac{q_i^\top k_j}{\sqrt{d_k}}\right)}{\sum_{\{j', m_{ij'}=1\}} \exp\left(\frac{q_i^\top k_{j'}}{\sqrt{d_k}}\right)} \cdot v_j \tag{3}$$

where the d_k is the dimension of k vector. In order to keep the performance of a single modality when other modalities are absent, the modality-specific representation can not access the fusion representation or other modalities. This explicitly prevents the information of the fusion stream from leaking into the unimodal representation. This is the key to preserve pure streams that correspond to single modalities. Thus, after applying this mask, the specific output o_s, o_o, o_d, o_m only contains information coming from the SAR, optical, DEM, MAP inputs, respectively. The fusion output o_f access all outputs in the model.

D. Reconstruction Pre-training

In order to train our network in an MAE way, we use a separate decoder for each generation task. The input to each decoder is the spatial tokens output from the cross attention. Following the same setting of MAE, we use shallow decoders with a low dimensionality, which consists of two Transformer blocks. MultiMAE mask across different modalities ensures the model develops predictive coding across different modalities besides different spatial patches. According to MultiMAE, we set a constant number of visible tokens at 512, which corresponds to 1/2 of all tokens in our experiment (learned fusion tokens and four modality inputs with 256 × 256 image size and 16 × 16 patch size). The proportion of tokens per modality λ are sampled from a symmetric Dirichlet distribution $(\lambda_{Optical}, \lambda_{SAR}, \lambda_{DEM}, \lambda_{MAP}) \sim Dir(\alpha)$, where $\lambda_{Optical} + \lambda_{SAR} + \lambda_{DEM} + \lambda_{MAP} = 1, \lambda \geq 0$. For simplicity and better representation of any possible sampled task, we use a concentration parameter $\alpha = 1$. As shown in Fig. 1, we adopt reconstruction loss (l_2 distance mean squared error) to recover the pixel color and l_1 loss for height information following MultiMAE and using cross-entropy loss (l_{ce}) on land-cover map reconstruction:

$$\begin{aligned}&L_{DEM} = l_1(Dec(o_f), DEM)\\ &L_{SAR\_Optical} = l_2(Dec(o_f), SAR) + l_2(Dec(o_f), Optical)\\ &L_{MAP} = l_{ce}(Dec(o_f), MAP)\end{aligned} \tag{4}$$

E. Contrastive Pre-training

We also add the class token for each modality input data and an additional global class token for the learned fusion tokens. To integrate information from the encoded visible tokens of other modalities, we add a single cross-attention layer using these tokens as queries that cross-attend to the encoded tokens of the last self-attention layer. We utilize the standard cross-attention operation and produce five different outputs: the vector outputs for each modality and their corresponding fusion vector outputs. This design opens the possibility to use contrastive learning among different modalities and fusion tokens. For a better multimodality alignment, we propose to use extra contrastive loss between each modality-specific output and the fusion vector. Specifically, given the optical vector output $z_o = CA(z_o, o_o)$ and the corresponding fusion output $z_{f\_o} = CA(z_{f\_o}, o_{f\_o})$, where CA is the cross-attention operation, o_{f_o} is the fusion tokens on the unmasked optical token positions, the contrastive loss can be formulated as:

$$L_c(z_o, z_{f\_o}) = -\mathop{\mathbb{E}}_{S}\left[\log \frac{e^{sim(z_o^i, z_{f\_o}^i)/\tau}}{\sum_{j=1}^{N} e^{sim(z_o^i, z_{f\_o}^j)/\tau}}\right] \tag{5}$$

where sim is a similarity function (i.e., cosine similarity), S is a set that contains $N-1$ negative samples and one positive sample. This equation introduces the loss for Optical-FUSION contrastive training. In order to contrast the output of all modalities, we define a contrastive loss between unimodal representations and their corresponding fusion representations. Thus, we can write the full loss as:

$$\begin{aligned}L =& L_{DEM} + L_{SAR\_Optical} + L_{MAP} + L_c(z_{f\_o}, z_o)\\ &+ L_c(z_{f\_s}, z_s) + L_c(z_{f\_d}, z_d) + L_c(z_{f\_m}, z_m)\end{aligned} \tag{6}$$

F. Random Modalities Combination

Besides the network design, the training strategy is vital to the performance of modal-incomplete inputs. The research in [20] finds that the Transformer models tend to overfit the dominating modalities in a task. To improve the robustness of the proposed approach against modal-incomplete data, we propose to leverage a random modality combination training strategy. Thanks to the proposed approach, we can randomly choose the different modality combinations or unimodal data in pre-training or supervised training on downstream tasks. During pre-training, multimodal inputs undergo random masking, yielding diverse modality combinations at each patch position. The modality attention block effectively integrates the extant modalities into fusion tokens and adapts to the absence of input modalities. This combination of random masking

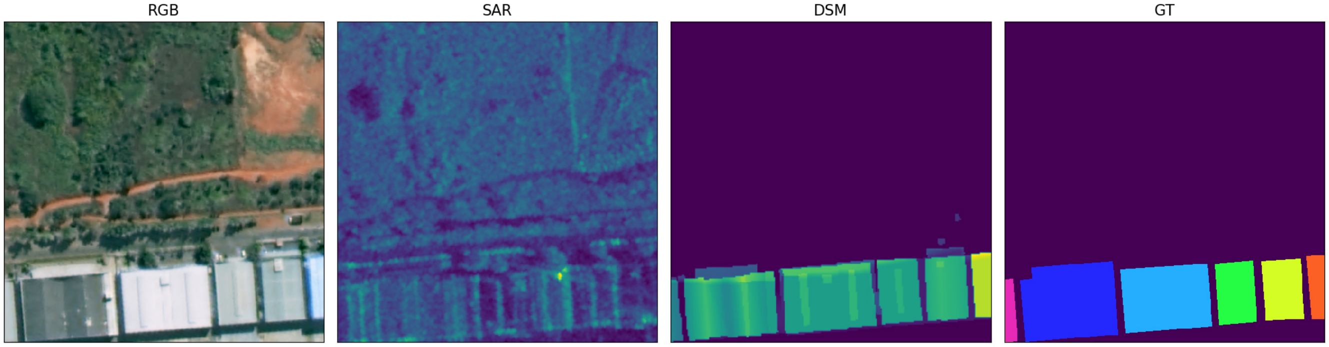


Fig. 2. Example of DFC2023 track2 data sample containing RGB and SAR images, DSM and ground truth.

and modality attention confers robustness upon the network, particularly when confronted with localized multimodal input absence. During supervised training on downstream tasks, PatchDropout is employed as a form of data augmentation. Furthermore, the selection of modalities during network training is randomized, encompassing unimodal input, modal-complete input, and modal-incomplete input scenarios. The integration of masked self-attention and additional learnable fusion tokens serves to maintain unimodal performance and accommodates the absence of entire modalities. The proposed methodology distinguishes itself by unifying all modalities through the incorporation of extra learned tokens, thereby substantially mitigating the impact of modal-incomplete inputs.

IV. Experiments

In this section, we evaluate the proposed approach in multiple settings. We first introduce the multimodal dataset used in this work. Then, we present the details of both pre-training and training on downstream tasks, as well as the evaluation procedures. Finally, we ablate the performance of the modal-complete and the modal-incomplete inputs to show the proposed approach's flexibility.

A. Description of Experiments

In order to showcase the proposed approach across the different modalities, we train the proposed approach in both a completely supervised paradigm and a fine-tuning paradigm with pre-trained weights. Many works have pointed out that the pre-trained big model on multimodal data can be beneficial on downstream tasks [29]. The pre-trained model can be then used for arbitrary downstream tasks with the fine-tuning of the task-specific decoder. Hence we can train a giant model on a large multimodal data set with as many modalities as possible. The pre-trained model can strengthen the ability to extract features that are only trained on a few or single modality data. In this section, we provide the details of the self-supervised pre-training and the supervised training on downstream tasks as well as the multimodal datasets.

B. Description of Datasets

We train and evaluate the performance of the proposed approach on two multimodal datasets for two downstream tasks, namely building instance / semantic segmentation and LULC mapping.

1) DFC2023 track2 - Building instance / semantic segmentation: The first data set is the track 2 dataset of DFC2023, which comprises a combination of RGB images, SAR images, and DSM data having a sample size of 256 × 256 pixels. It consists of 5332 triplet samples for supervised training and 1335 for evaluation, where RGB images have three channels, whereas both SAR images and DSM have one channel. While the objective of the original task is building height estimation, this study simplifies it as building instance / semantic segmentation. The dataset consists of images obtained from GaoJing-1, GaoFen-2 and GaoFen-3 satellites, with spatial resolutions of 0.5 m, 0.8 m and 1 m, respectively. Normalized Digital Surface Models (nDSMs) are used as a reference in Track2 and are created from stereo images captured by GaoFen-7 and WorldView-1 and -2 with approximately 2 m ground sampling distance (GSD). The dataset was collected from seventeen cities across six continents and hence is highly diverse in terms of landforms, building types and architecture. The labels of building instance segmentation adopt the MS COCO format and are provided in a JSON file. A sample of the labels is shown in Fig. 2 for illustration.

2) Quadruplet Dataset - Land-Use Land-Cover (LULC) mapping: The second dataset considers diverse data sources obtained from Google Earth Engine (GEE) platform, encompassing Sentinel-1, Sentinel-2, LiDAR DEMs and Dynamic World LULC maps, with a sample size of 256 × 256 pixels (see Fig. 3 and Fig. 4). The dataset comprises 37 regions across various landscapes and LULC classes in France and Australia. It consists of 5340 quadruplet samples for training and 783 quadrupled samples for evaluation, where the Sentinel-1 images have two channels (VV and VH polarization channels), the Sentine-2 images have four channels (RGB and NIR bands), and the LiDAR DEMs and the Dynamic World LULC maps are both with one channel. The Sentinel-1 mission provides data from a dual-polarization C-band SAR instrument and produces the calibrated and ortho-corrected S1 GRD products. We download the data from the COPERNICUS/S1_GRD category on GEE, resampling it into 10 m resolution and using dual-band VV+VH. Similarly, we download the Sentinel-2 data from the COPERNICUS/S2_SR_HARMONIZED category, which provides multispectral imaging with 13 spectral

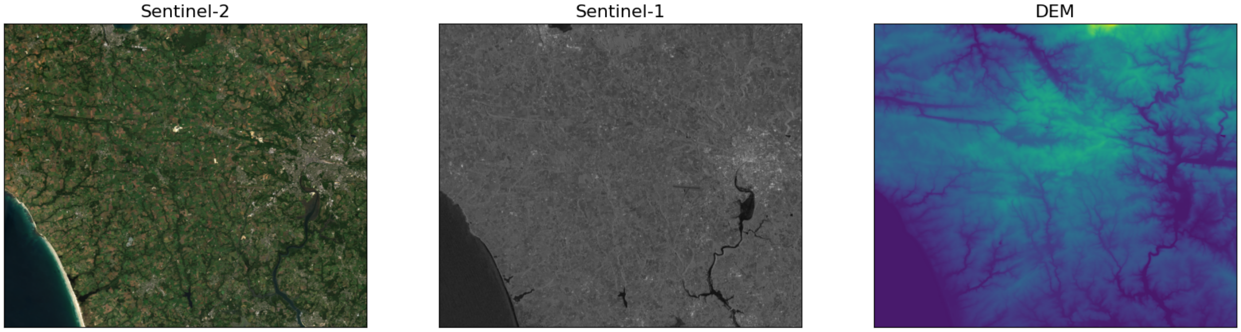


Fig. 3. Example of Quadruplets Data Set containing Sentinel1, Sentinel-2 and DEM data.

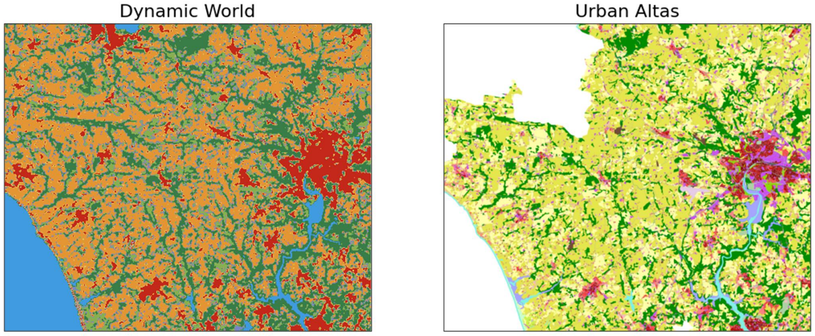


Fig. 4. Example of Dynamic World Map and European Urban Atlas data.

bands suitable for large-scale LULC mapping. We resample the Sentinel-2 data into 10 m resolution, and use the RGBN bands in this work. Two types of LiDAR DEMs are provided in this research. In France, we utilize the RGE ALTI dataset, which is a digital elevation model created using airborne lidar, with a pixel size of 1 m. We resample this dataset to 10 meters, with a vertical accuracy that ranges from 0.2 m to 0.5 m and an average accuracy of 7 m in steep slope areas. In Australia, we use a digital elevation model 5 m grid derived from 236 individual LiDAR surveys conducted between 2001 and 2015. We compile and resample the available 5 m resolution LiDAR-derived DEMs using a neighbourhood-mean method to create 10 m resolution datasets for each survey area, which we used in this work. The Dynamic World MAP (DNW) dataset comprises globally consistent, 10 m resolution, near real-time land-use and land-cover predictions derived from Sentinel-2 imagery. It features ten bands that include estimated probabilities for each of the nine LULC classes (water, trees, grass, crops, shrub and scrub, flooded vegetation, built-up area, bare ground, and snow & ice). It also has a class ”label” band indicating the class with the highest estimated probability, which makes it suitable for multi-temporal analysis and custom product creation. Lastly, we utilize the labeled class-reference from the UrbanAtlas 2018 database containing 27 LULC classes as the label of this dataset. The dataset provides integer rasters with index labels. We create raster maps with 10 m resolution that geographically match the Sentinel-1/-2 images using the open-data vector images freely available on the European Copernicus program website.

3) Downstream Tasks: We evaluate the proposed approach against state-of-the-art methods on two downstream tasks: building instance / semantic segmentation, and LULC mapping. In particular, the evaluation is performed on the supervised learning and the fine-tuning paradigms. For these two downstream tasks, we replace the pre-trained decoders with randomly initialized Mask2Former [30]. Mask2Former incorporates masked attention to discern localized features and forecast outputs for panoptic, instance, and semantic segmentation within a unified framework. The model predicts binary masks associated with global class labels, thereby streamlining tasks related to semantic and panoptic segmentation and yielding notable empirical results. At the core of Mask2Former lies a specialized Transformer decoder equipped with predefined queries. This decoder integrates a masked attention operator, strategically extracting localized features by confining cross-attention within the foreground region of the predicted mask for each query, as opposed to encompassing the entirety of the feature map. In the following, we give an overview of the two tasks.

Building Instance / Semantic Segmentation: We follow the Mask2Former but replace the backbone with the proposed network. In the supervised experiments, we train the whole network from scratch using a random modality combination strategy. In the fine-tuning experiments, we consider two strategies, one is to update the network on the pre-trained ViT-T backbones trained only using reconstruction loss, and the other is to update the whole network on the pre-trained ViT-T backbones trained using reconstruction and contrastive losses. We train our model on DFC2023 track2 train split and report the validation accuracy on the validation split. Along with the results of building instance segmentation, we also provide the binary building semantic segmentation results.

Land-Use Land-Cover Mapping: We still use the Mask2Former with the proposed backbone on the quadruplet dataset to generate LULC maps. However, we consider 7 classes merged from the semantic hierarchy defined by UrbanAtlas. For that, we extract 7 semantic classes by taking the argmax of the prediction head. The same training strategy as that of the building instance segmentation is used in this task. We train our model on 10 (5340 samples) cities and report the validation accuracy on the other 2 (783 samples) cities.

4) Architecture Details: The proposed approach uses a ViT-T as the main structure and consists of 4 and 5 input adapters with a patch size of 16×16 pixels for the pre-training in the two different tasks. Differently from the standard MultiMAE, we add the learnable fusion tokens as input by using an additional input adapter to add 2D sine-cosin position encoding. The fusion tokens are as many as the number of patched inputs of each modality.

After adding the position encodings, the fusion tokens with all modality inputs are given as input to a modality attention block. In self-attention, we use the masked algorithm to avoid the fusion information leak to a single modality. In order to get the global features of each modality and the corresponding fusion tokens, we use an additional cross-attention layer to map the patch embeddings into the vector output. Then an auxiliary contrastive loss is added between each modality output vector and the corresponding fusion output vector.

For mask-reconstruction pre-training, we follow the same setting of the MultiMAE decoder but without positional embeddings and cross-attention layer. The fusion tokens are projected into the decoder dimension by using a linear projection layer and then added to a learned modality embedding. After this, two Transformer blocks and a linear projector are used to project and reshape it to form an image or a map.

For the two downstream tasks, we adopt the same settings from Mask2Former. For the pixel decoder, we use 2 MS-DeformAttn layers applied to feature maps with resolution 1/8, 1/16 and 1/32, and use a simple upsampling layer with lateral connection on the final 1/8 feature map to generate the feature map of resolution 1/4 as the per-pixel embedding. We use the Transformer decoder with 4 layers and 100 queries for instance segmentation, 2 queries for binary building semantic segmentation and 9 queries for LULC mapping. We use the binary cross-entropy loss and the dice loss for the mask loss. The final loss is a combination of mask loss and classification loss. For instance segmentation, we use the standard AP@50 (average precision with a fixed IoU of 0.5) metric. For semantic segmentation, we use the mIoU (mean Intersection-over-Union) metric.

5) Training Details: For pre-training, we train our model for 1600 epochs on 6667 triplet data on the DFC2023 track2 data set and 6123 quadruplet data on the quadruplet data set, individually. We use the AdamW optimizer with a base learning rate of 1e-4 and weight decay of 0.05. We warm up training for 40 epochs, starting from using cosine decay. We set the batch to 40 using a single Nvidia RTX 3090. All data are resized to 256×256. The number of non-masked tokens given to the encoder is set to half of all tokens on the two data sets. For the second dataset, where we use the land-cover map as an additional modality input with 64-dimensional class embeddings.

For instance segmentation and semantic segmentation using Mask2Former, we use AdamW optimizer and the step learning rate schedule. We use an initial learning rate of $1e^{-4}$ and a weight decay of 0.05. A learning rate multiplier of 0.1 is applied to the backbone with the pre-training and not in the supervised learning. We decay the learning rate at 0.9 and 0.95 fractions of the total training steps by a factor of 10. We train our models for 50 epochs with a batch size of 10 in both the building segmentation task and the building instance segmentation task, and 30 epochs with a batch size of 30 in the LULC mapping task.

Concerning the training strategy involving random modality combinations at each iteration, we systematically adjust the selection of input modalities and the spatial random mask, as required by the constraints imposed by the sample feature size in the mini-batch gradient descent process. The selection of input modalities adheres to a uniform distribution, and the spatial random mask employs a symmetric Dirichlet distribution to determine the proportion of tokens associated with each modality.

C. Experimental Results

1) Multimodal Comparison: We evaluate the proposed approach with the two paradigms, one is supervised from scratch, and the other is fine-tuning with pre-trained weights. Considering no dedicated Transformer for incomplete multimodal remote sensing data fusion, we compare the proposed approach against a technique that uses origin self-attention and the same number of learnable fusion tokens, termed MultiViT, on modal-complete and modal-incomplete inputs for building instance/semantic segmentation and LULC mapping tasks. The results reported in Tables I and II reveal that the proposed approach outperforms MultiViT in building instance/semantic segmentation tasks when evaluated with modal-complete inputs. Similarly, in the context of the LULC mapping task, the performance of the proposed approach excels over that of MultiViT. With regards to modal-incomplete inputs, the proposed approach performs impressively well on all modal-incomplete inputs and single modality inputs for both tasks due to the joint use of the modality attention block and the masked self-attention as well as the random modality combination training strategy. For building instance/semantic segmentation, there is a visible dominance of RGB images over all other modalities, followed by DSM, while SAR images make the slightest contribution to the task, even causing noise. In this situation, MultiViT completely overfits on dominant modality inputs and fails on the task with single modality inputs when evaluated with modal-incomplete inputs. Similarly, for LULC mapping, Sentinel-2 images along with the dynamic world map have a significant influence on the task, followed by Sentinel-1 images and DEM. The proposed approach achieves the best performance with a mIoU of 0.278 with modal-complete inputs, whereas MultiViT overfits on dynamic world

TABLE I
Quantitative evaluations of proposed approach versus MultiViT with complete and incomplete multimodality inputs on the DFC2023 track2 dataset. Results are reported on AP@50 for instance segmentation and mIoU for semantic segmentation and consider the supervised result (sup.) and the fine-tuning result with the mask-reconstruction pre-trained weights (Fine. w/G) as well as the fine-tuning results with both mask-reconstruction and contrastive pre-trained weights (Fine. w/G&C).

Multimodal Input	Sup. MultiViT		Sup. Propsed		Fine. w/ G.		Fine. w/ G. & C.	
	ins.	sem.	ins.	sem.	ins.	sem.	ins.	sem.
SAR, RGB, DSM	0.071	0.818	**0.234**	0.857	0.221	0.852	0.208	**0.858**
SAR, RGB	0.018	0.536	**0.214**	0.821	0.211	0.809	0.196	**0.822**
SAR, DSM	0.050	0.707	**0.134**	0.783	0.094	0.782	0.076	**0.784**
RGB, DSM	0.062	0.749	**0.233**	0.854	0.215	0.848	0.203	**0.855**
SAR	0.003	0.392	**0.032**	**0.577**	0.024	0.555	0.029	0.573
RGB	0.016	0.492	**0.212**	0.814	0.206	0.800	0.196	**0.814**
DSM	0.015	0.686	**0.110**	**0.763**	0.071	0.757	0.051	0.758

TABLE II
Quantitative evaluations of proposed approach versus MultiViT with complete and incomplete multimodality inputs on the quadruplets dataset. The results are reported in terms of mIoU values and consider the supervised result (sup.) and the fine-tuning result with the mask-reconstruction pre-trained weights (Fine. w/G) as well as the fine-tuning results with both mask-reconstruction and contrastive pre-trained weights (Fine. w/G&C).

Multimodal Input	Sup. MultiViT	Sup. Proposed	Fine. w/ G.	Fine. w/ G. & C.
S1, S2, DEM, DNW	0.248	0.278	0.275	**0.280**
S1, S2, DEM	0.063	**0.265**	0.261	0.262
S1, S2, DNW	0.248	0.280	0.274	**0.281**
S1, DEM, DNW	0.211	0.251	0.252	**0.262**
S2, DEM, DNW	0.219	**0.277**	0.275	0.276
S1, S2	0.064	**0.265**	0.260	0.262
S1, DEM	0.055	0.224	**0.227**	0.225
S1, DNW	0.219	0.238	0.251	**0.263**
S2, DEM	0.058	0.228	0.257	**0.249**
S2, DNW	0.226	0.276	0.274	**0.277**
DEM, DNW	0.175	0.232	0.231	**0.248**
S1	0.076	0.215	**0.219**	0.218
S2	0.056	0.230	**0.258**	0.250
DEM	0.028	0.045	0.017	**0.046**
DNW	0.180	0.236	0.235	**0.250**

maps, and performs slightly better when the dynamic world map is present but fails when it is not present in the inputs.

In the context of the fine-tuning paradigm, the proposed approach is assessed through two distinct pre-training methods: one that employs mask-reconstruction pre-training and another that combines mask-reconstruction and contrastive pre-training. The outcomes of the evaluation for both tasks are presented in Table I and Table II. As one can see, different tasks show controversial results. Specifically, in the case of the building instance segmentation task, the training-from-scratch model demonstrates superior performance compared to all other models. The fine-tuning outcome related to mask-reconstruction is ranked as the second-best, while the fine-tuning result involving both mask-reconstruction and contrastive pre-training exhibits comparatively diminished results. In the building semantic segmentation task, the results of the training-from-scratch model and the fine-tuning on both mask-reconstruction and contrastive pre-training achieve comparable performance. This performance surpasses that observed in the fine-tuning result solely based on the mask-reconstruction pre-training. In contrast, for the land-cover mapping task, the fully finetuned model, incorporating both mask-reconstruction and contrastive pre-training, is the top-performing model among all the models listed in the tables. This demonstrates the potential of mask-reconstruction and contrastive pre-training in augmenting downstream LULC tasks. By comparing two fine-tuning results, it becomes evident that the inclusion of contrastive pre-training yields further enhancements in performance compared to the exclusive utilization of mask-reconstruction pre-training.

For the single modality input, our goal is not to show state-of-the-art performance in this setting, as we are trying to solve the dramatic degradation of unimodal inference with a multimodal backbone. Here we show the ability of the proposed approach to produce meaningful unimodal outputs when fed with unimodal data. To do this, we only input one modality and neglect other modality inputs. As we can see on both datasets (Table I and Table II), the MultiViT suffers significant degradation from missing of modalities and completely fails to work on the non-dominated modalities. In contrast, the proposed approach using the random modality combination strategy achieves high performance also when only one modality is available. This is due to the fact that in the proposed models, some capacity is allocated to each modality specifically and the model is able to produce unimodal outputs. Besides the quantitative analysis, we also provide a visual qualitative comparison. Fig. 5 and Fig. 6 show the results of building instance / semantic segmentation and LULC mapping, respectively. For building instance / semantic segmentation, similarly to Table I, the proposed approach with a supervised paradigm achieves the best performance followed by the results of fine-tuning. The MultiViT achieves the worst

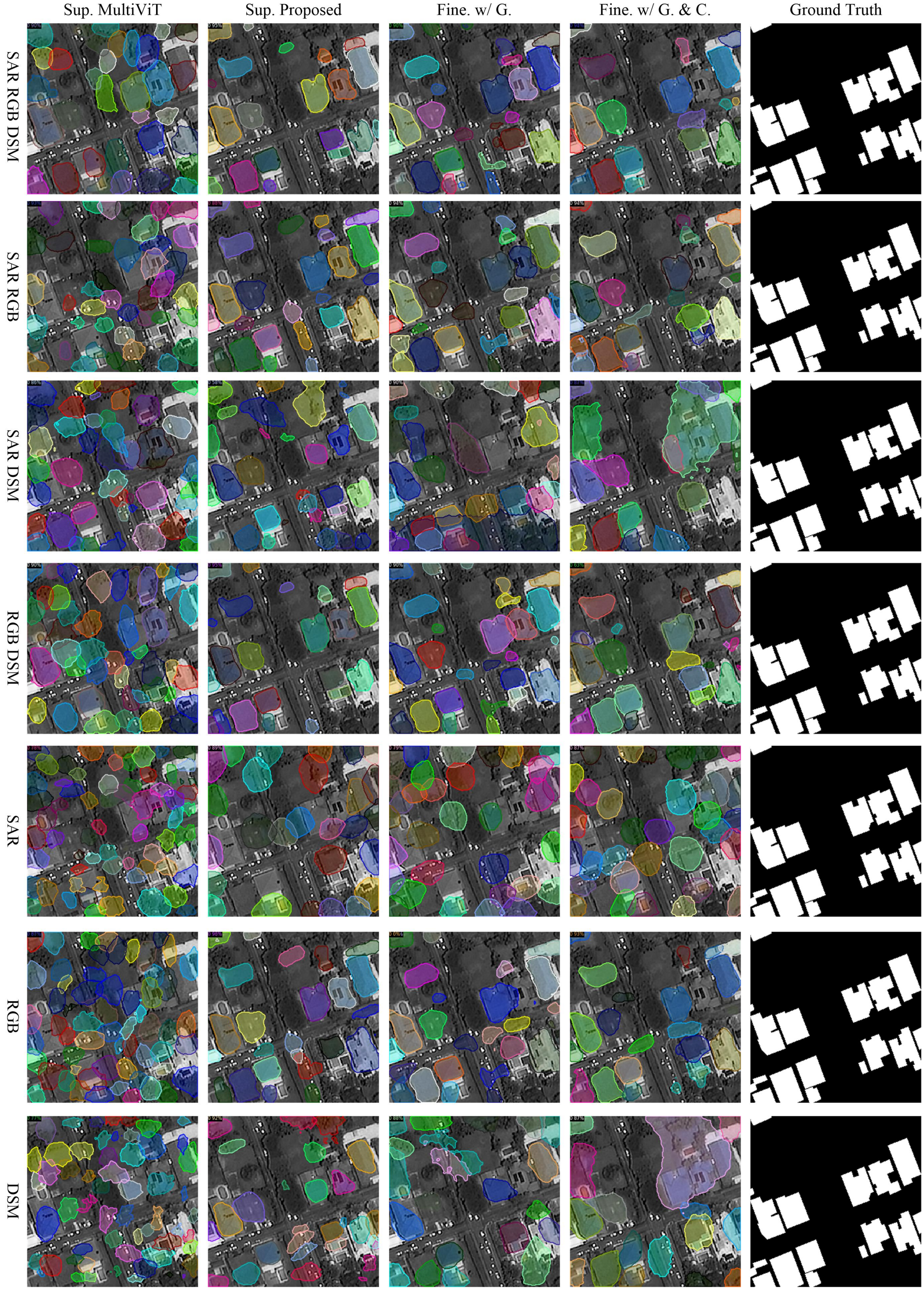


Fig. 5. Results of proposed approaches in the supervised and the two fine-tuning paradigms versus MultiViT on DFC2023 track2 dataset and consider the supervised result (sup.) and the fine-tuning result with the mask-reconstruction pre-trained weights (Fine. w/G) as well as the fine-tuning results with both mask-reconstruction and contrastive pre-trained weights (Fine. w/G&C).

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3387837

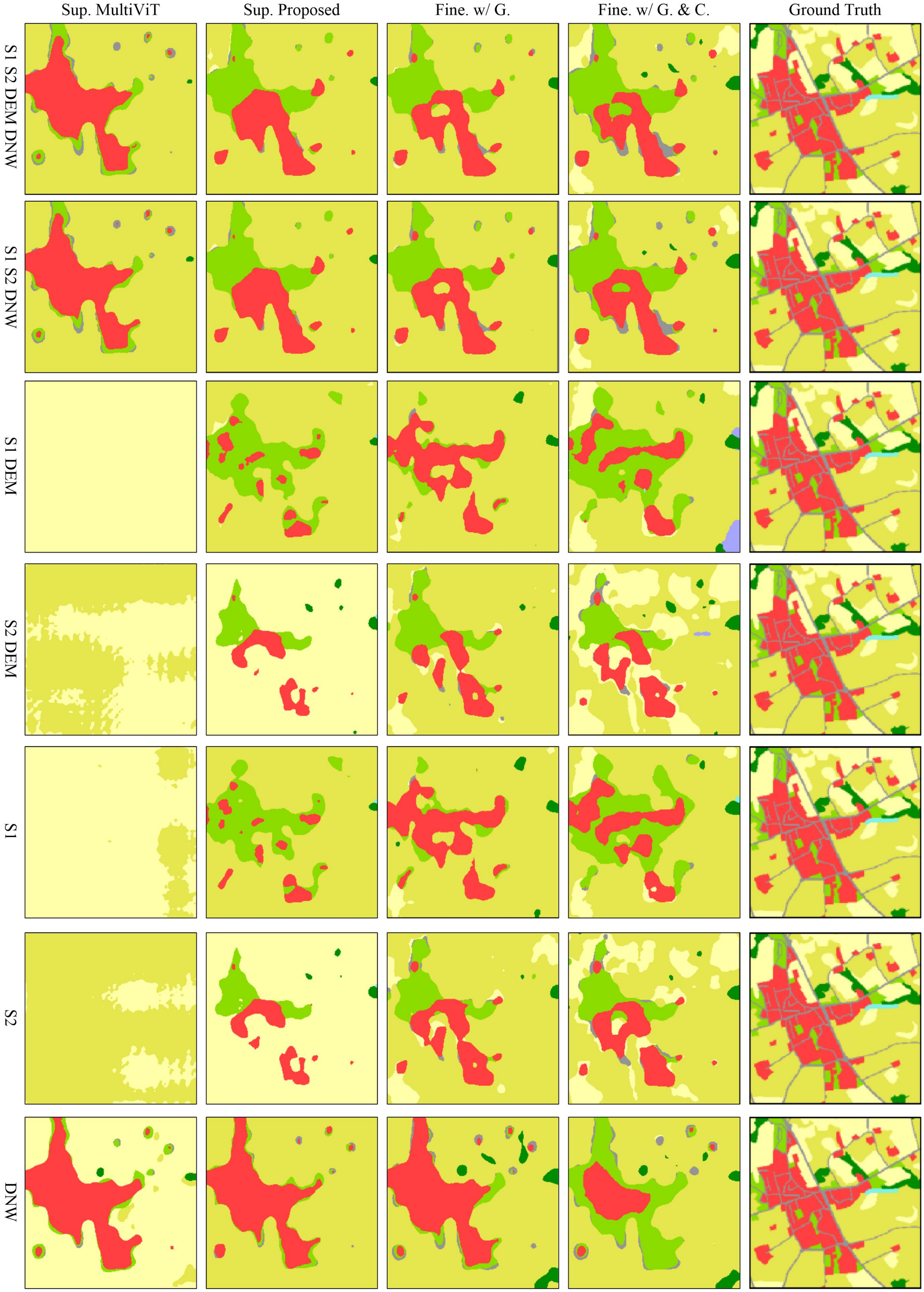


Fig. 6. Results of proposed approaches in the supervised and the two fine-tuning paradigms versus MultiViT on the quadruplets dataset and consider the supervised result (sup.) and the fine-tuning result with the mask-reconstruction pre-trained weights (Fine. w/G) as well as the fine-tuning results with both mask-reconstruction and contrastive pre-trained weights (Fine. w/G&C).

TABLE III

Quantitative evaluations of the proposed approach on the different settings of masked self-attention (w/o Mask), random modality combination training strategy (w/o Random), and modality attention (w/o Attention) with complete and incomplete multimodality inputs on the DFC2023 track2 dataset. Results are reported in terms of AP@50 for instance segmentation and mIoU for semantic segmentation.

Multimodal	w/o Mask		w/o Random		w/o Attention		w/ all	
Input	ins.	seg.	ins.	seg.	ins.	seg.	ins.	seg.
SAR, RGB, DSM	0.218	0.840	0.215	**0.858**	0.118	0.752	**0.234**	0.857
SAR, RGB	**0.218**	0.795	0.173	0.734	0.081	0.642	0.214	**0.821**
SAR, DSM	0.097	0.773	0.080	0.725	0.071	0.714	**0.134**	**0.783**
RGB, DSM	0.212	0.839	0.195	0.831	0.115	0.737	**0.233**	**0.854**
SAR	0.029	0.545	0.002	0.407	0.009	0.470	**0.032**	**0.577**
RGB	**0.217**	0.791	0.144	0.659	0.083	0.614	0.212	**0.814**
DSM	0.078	0.755	0.065	0.707	0.063	0.697	**0.110**	**0.763**

TABLE IV

Quantitative evaluations of the proposed approach on the different settings of masked self-attention (w/o Mask), random modality combination training strategy (w/o Random), and modality attention (w/o Attention) with complete and incomplete multimodality inputs on the quadruplets dataset. The results are reported in terms of mIoU.

Multimodal Input	w/o Mask	w/o Random	w/o Attention	w/ all
S1, S2, DEM, DNW	0.278	0.265	0.239	**0.278**
S1, S2, DEM	0.262	0.240	0.178	**0.265**
S1, S2, DNW	0.279	0.266	0.237	**0.281**
S1, DEM, DNW	**0.256**	0.223	0.234	0.251
S2, DEM, DNW	0.270	0.240	0.239	**0.277**
S1, S2	0.261	0.232	0.179	**0.266**
S1, DEM	0.220	0.149	0.150	**0.225**
S1, DNW	**0.254**	0.223	0.232	0.239
S2, DEM	**0.232**	0.185	0.147	0.228
S2, DNW	0.273	0.251	0.236	**0.276**
DEM, DNW	**0.238**	0.194	0.219	0.233
S1	**0.219**	0.145	0.144	0.215
S2	**0.231**	0.179	0.149	0.230
DEM	0.032	0.044	0.032	**0.045**
DNW	**0.238**	0.193	0.220	0.236

performance, especially with the modal-incomplete inputs. For the LULC mapping task, the fine-tuning with contrastive and mask-reconstruction pre-trained weights outperforms other approaches, while MultiViT exhibits reliable performance only with DNW input.

In addition to the performance of the proposed approach on different modality combinations, an in-depth analysis of individual modalities and their combination for each task is conducted based on the outcomes derived from the proposed supervised learning framework. Concerning building instance/semantic segmentation tasks, optical images prominently contribute as the primary modality, followed by DSM data, while SAR images exhibit a comparatively smaller impact. In the context of building instance segmentation, SAR images provide limited beneficial information, and similar results are obtained by the exploration of various modality combinations. The simultaneous integration of SAR, optical and DSM data obtains optimal performance, with the joint usage of optical and DSM data yielding comparable results. Conversely, joint deployments of SAR either with optical or DSM data result in a suboptimal performance. For the LULC mapping task, DNW maps emerge as the most significant contributor, with Sentinel-2 images exhibiting a similar performance to DNW maps. In contrast, Sentinel-1 images contribute less significantly, and DEM fails to provide essential information. The joint use of DNW maps and Sentinel-2 images outperforms individual deployments, surpassing outcomes achieved without their integration. Notably, the combined usage of Sentinel-1/2 images and DNW maps achieves the highest performance, even surpassing the integration of all four modalities. In some cases the use of a singular modality may introduce noise, potentially impacting the overall performance of multimodal data fusion. The proposed approach, emphasizing incomplete multimodal remote sensing data fusion, not only advances the understanding of modality contributions but also facilitates a judicious selection of the most appropriate modality combination during inference.

2) Ablation Studies: To ensure robust performance in the presence of modal-incomplete inputs, an exhaustive analysis on how the various strategies influence the effectiveness of the proposed approach is undertaken. Despite the good performance of the proposed approach on different modality combinations outlined in the final results, the use of a training strategy involving random modality combinations serves to mitigate overfitting on dominant modalities in which its impact on the performance of modal-complete inputs remains ambiguous.

Incorporating masked self-attention avoids information flow from one modality to the other, thereby preserving modality-specific information through the network, as highlighted in the final results. This proves particularly advantageous for unimodal inputs, contributing to a better performance with the modal-incomplete inputs. Masked self-attention is mainly used in contrastive pre-training to maintain the independence of each modality, especially when dealing with text and images. Meanwhile, masked self-attention is not mandatory in mask-reconstruction pre-training and supervised training. Concurrently, the utilization of masked self-attention introduces a constraint on the interaction between disparate modalities, which warrants a more in-depth ablation study within the framework of supervised training to furnish insights into its potential benefits in this specific context.

Furthermore, modality attention assumes a pivotal role in assimilating information from the current modality into additional fusion tokens for each patch token, thereby enhancing the meaningfulness of the representations encoded by the extra fusion tokens. The efficacy of modality attention requires further validation through dedicated ablation studies, aligning with the detailed analysis of individual modalities and their combination presented in the final results. To evaluate the generalizability of the proposed components, all ablations were performed on both tasks: the building instance / semantic segmentation and LULC mapping on the supervised paradigm, reinforcing the comprehensive analysis of modalities and their combinations conducted in the final results.

We first validate the importance of the random modality combination training strategy on downstream tasks in a supervised paradigm. As shown in Tables III and IV, the model without the modality random combination training strategy experiences severe degradation with modal-incomplete inputs and even without an improvement on the result of modal-complete inputs. In addition, we test the effect of the modality attention by removing it from the proposed network. The corresponding results show a significant drop in performance, indicating that the modality attention enables superior interaction of the fusion token with each modality and facilitates learning more discriminative features for downstream tasks. For masked self-attention, we show the supervised results without masked self-attention for both tasks (see Table III and IV). In the first row, we remove the masked self-attention blocks while keeping the random modality combination training strategy, which results in a comparable or even worse performance with respect to the proposed approach. This is probably because even masked self-attention hinders the interaction between different modalities; however, the use of masked attention helps to maintain unimodal performance and benefits the whole training process. The benefits of the use of masked self-attention also can be found in pre-training. Compared with the mask-reconstruction pre-training, the use of masked self-attention in the combination pre-training helps to avoid the information flow from one modality to the other. As one can observe (see the semantic segmentation results in Tables I and II), the unimodal inference performs close to the modal-incomplete inputs as the modality streams are more independently treated. In contrast, the results without contrastive pre-training tend to overfit dominant modalities and are relatively poor on other modalities. Moreover, lower performances are observed on one single modality.

V. Conclusion

In this work, we have introduced an incomplete multimodal learning framework for multimodal remote sensing data fusion which can be used in both supervised training and self-supervised pre-training paradigms. Unlike previous multimodal remote sensing data fusion approaches, the proposed approach enables the training and inference of models with modal-incomplete inputs. By using the modality attention mechanism and masked self-attention, we are able to pre-train the network using contrastive and reconstruction losses in the MultiMAE framework, and also to train the network from scratch or finetune the model on downstream tasks using a random modality combination strategy. This strategy allows the network to maintain high performance even when dealing with modal-incomplete inputs or a single modality in the inference stage.

We evaluated our model on two multimodal remote sensing datasets, demonstrating flexibility in network training and inference, and state-of-the-art performance when presented with modal-incomplete inputs. It is worth noting that this study focused solely on different modality raster data.

In future work, we plan to optimize the computational efficiency of the proposed approach and incorporate diverse modalities of data, such as text and vector data, into the proposed framework.

References

[1] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen *et al.*, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1, pp. 6–39, 2019.

[2] C. Paris and L. Bruzzone, "A three-dimensional model-based approach to the estimation of the tree top height by fusing low-density lidar data and very high resolution optical images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 467–480, 2014.

[3] Y. Chen and L. Bruzzone, "Self-supervised change detection in multiview remote sensing images," *arXiv preprint arXiv:2103.05969*, 2021.

[4] Chen, Yuxing and Bruzzone, Lorenzo, "An approach based on contrastive learning and vector quantization to the unsupervised land-cover segmentation of multimodal images," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 4811–4814.

[5] D. Deus, "Integration of alos palsar and landsat data for land cover and forest mapping in northern tanzania," *Land*, vol. 5, no. 4, p. 43, 2016.

[6] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The sen1-2 dataset for deep learning in sar-optical data fusion," *arXiv preprint arXiv:1807.01569*, 2018.

[7] J. Adrian, V. Sagan, and M. Maimaitijiang, "Sentinel sar-optical fusion for crop type mapping using deep learning and google earth engine," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 215–235, 2021.

[8] D. Ienco, R. Interdonato, R. Gaetano, and D. H. T. Minh, "Combining sentinel-1 and sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 11–22, 2019.

[9] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778–782, 2017.

[10] S. T. Seydi, H. Rastiveis, B. Kalantar, A. A. Halin, and N. Ueda, "Bdd-net: An end-to-end multiscale residual cnn for earthquake-induced building damage detection," *Remote Sensing*, vol. 14, no. 9, p. 2214, 2022.

[11] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, "Feature extraction for classification of hyperspectral and lidar data using patch-to-patch cnn," *IEEE transactions on cybernetics*, vol. 50, no. 1, pp. 100–111, 2018.

[12] A. Diab, R. Kashef, and A. Shaker, "Deep learning for lidar point cloud classification in remote sensing," *Sensors*, vol. 22, no. 20, p. 7868, 2022.

[13] H. Kemper and G. Kemper, "Sensor fusion, gis and ai technologies for disaster management," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 1677–1683, 2020.

[14] Q. Xu, C. Long, L. Yu, and C. Zhang, "Road extraction with satellite images and partial road maps," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[15] B. Bischke, P. Helber, F. Koenig, D. Borth, and A. Dengel, "Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation," in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2018, pp. 1–6.

[16] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Urban land cover classification with missing data modalities using deep convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 6, pp. 1758–1768, 2018.

[17] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *arXiv preprint arXiv:2203.16952*, 2022.

[18] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 200–14 213, 2021.

[19] A. Recasens, J. Lin, J. Carreira, D. Jaegle, L. Wang, J.-b. Alayrac, P. Luc, A. Miech, L. Smaira, R. Hemsley *et al.*, "Zorro: the masked multimodal transformer," *arXiv preprint arXiv:2301.09595*, 2023.

[20] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 177–18 186.

[21] S. Paisitkriangkrai, J. Sherrah, P. Janney, V.-D. Hengel *et al.*, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 36–43.

[22] N. Audebert, B. Le Saux, and S. Lefèvre, “Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 20–32, 2018.

[23] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, “Deep fusion of remote sensing data for accurate classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1253–1257, 2017.

[24] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, “Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 92–93.

[25] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, “Deep encoder–decoder networks for classification of hyperspectral and lidar data,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.

[26] X. Du, X. Zheng, X. Lu, and A. A. Doudkin, “Multisource remote sensing data classification with graph fusion network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10 062–10 072, 2021.

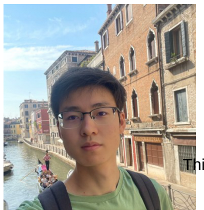
[27] Y. Sun and X. Li, “Data fusion of gis and rs by neural network,” in *2010 International Conference on Intelligent Control and Information Processing*. IEEE, 2010, pp. 648–651.

[28] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, “Multimae: Multi-modal multi-task masked autoencoders,” in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*. Springer, 2022, pp. 348–367.

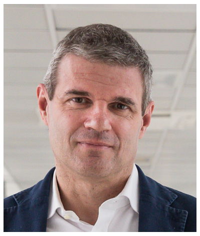
[29] M. Singh, Q. Duval, K. V. Alwala, H. Fan, V. Aggarwal, A. Adcock, A. Joulin, P. Dollár, C. Feichtenhofer, R. Girshick *et al.*, “The effectiveness of mae pre-pretraining for billion-scale pretraining,” *arXiv preprint arXiv:2303.13496*, 2023.

[30] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

Yuxing Chen received the M.S. degree in geodesy and surveying engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree with the University of Trento, Trento, Italy. His research self-supervised learning and its application in remote sensing.



Maofan Zhao is currently pursuing the Ph.D. degree in the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. From 2021 to 2023, he was a Visiting Ph.D. Student with the University of Trento, Trento, Italy. His research interests include remote sensing image analysis, deep learning and urban remote sensing.



Lorenzo Bruzzone (S’95-M’98-SM’03-F’10) received the M.S. degree (summa cum laude) in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively.

He is currently a Full Professor of telecommunications with the University of Trento, Trento, Italy, where he teaches remote sensing, radar, and digital communications. He is the Founder and the Director of the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento. He is the Principal Investigator of many research projects, including Radar for Icy Moon Exploration instrument in the framework of the JUpiter ICy moons Explorer mission of the European Space Agency. He has authored or co-authored 218 scientific publications in referred international journals (157 in the IEEE journals), more than 290 papers in conference proceedings, and 21 book chapters. He has edited or co-edited 18 books or conference proceedings and 1 scientific book. His papers have been cited more than 25 000 times, h-index 74. His research interests include remote sensing, radar and SAR, signal processing, machine learning, and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects.

Dr. Bruzzone was a Distinguished Speaker of the IEEE Geoscience and Remote Sensing Society from 2012 to 2016. He was invited as a Keynote Speaker in more than 30 international conferences and workshops. He is a member of the Permanent Steering Committee of this series of workshops. He has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing since 2003 and a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society since 2009. He is the Co-Founder of the IEEE International Workshop on the Analysis of Multitemporal Remote-Sensing Images (MultiTemp) series. He was a recipient of the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (first place), Seattle, in 1998, and many International and National Honors and Awards including the recent IEEE GRSS 2015 Outstanding Service Award, and the 2017 IEEE IGARSS Symposium Prize Paper Award. He has been the Founder of the IEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE for which he has been the Editor-in-Chief from 2013 to 2017. He was a Guest Co-Editor of many special issues of international journals. He is currently an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.