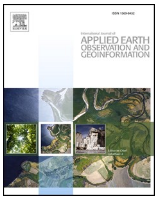
International Journal of Applied Earth Observation and Geoinformation 128 (2024) 103758

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Automated deformation detection and interpretation using InSAR data and a multi-task ViT model

Mahmoud Abdallah [a,c], Samaa Younis [c], Songbo Wu [a,b], Xiaoli Ding [a,b,*]

[a] *Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China*
[b] *Research Institution for Land and Space, The Hong Kong Polytechnic University, Hong Kong, China*
[c] *Public Works Department, Mansoura University, Mansoura, Egypt*

ARTICLE INFO

Keywords:
InSAR
Deformation detection and interpretation
Machine learning
Vision transformer model

ABSTRACT

Many geological hazards are associated with ground deformations. Prompt and accurate detection and interpretation of ground deformation is therefore vital to geohazard mitigation. Multitemporal Interferometric Synthetic Aperture Radar (MT-InSAR) is an effective geodetic technique for monitoring ground deformation. However, accurate computation and interpretation of deformation using InSAR are often hindered by various errors and a lack of expert knowledge. We present a new advanced deep learning model based on a multi-task vision transformer (MT-ViT) to automatically detect, locate, and interpret deformation using single interferograms. To address the issue of limited training data in InSAR applications, the proposed model utilizes pretrained weights from optical images and transfers them to a simulated InSAR dataset. Then real interferograms are used to fine-tune the weights in the network. An overall loss function is designed, which considers the classification and localization losses in the model. The effectiveness of the proposed model is demonstrated using both simulated and real InSAR datasets that contain either coseismic or volcanic deformation. The experimental results from the model are also compared with the state-of-the-art convolutional neural network (CNN) based techniques. The results show significant improvement in both the accuracy of the results and the computational efficiency over the CNN-based approaches. The MT-ViT model achieved 99.4 % classification accuracy, 54.1 % mean intersection over union (IOU), and 0.9 km localization accuracy. A comprehensive evaluation of the hyperparameters in training the MT-ViT model was carried out, which will inform future research in this direction. The research results highlight the promising capabilities of MT-ViT in near real-time deformation monitoring and automated deformation interpretation.

1. Introduction

Geohazards, including earthquakes, volcanic eruptions, landslides, and land subsidence present significant risks to the public and critical infrastructures. The impact of such events has been widely reported, highlighting the importance of proper mitigating of the geohazards (Loughlin et al., 2015). Ground deformation is often observable prior to the occurrence of a geohazard (Cicerone et al., 2009). Prompt monitoring and interpretation of such deformation can enable the provision of early warnings and the implementation of useful measures for geohazards (Loughlin et al., 2015; Ma et al., 2020). Interferometric Synthetic Aperture Radar (InSAR), particularly multi-temporal InSAR (MT-InSAR) has been widely used for ground deformation monitoring and geohazard identification (Sun et al., 2015; Anantrasirichai et al., 2018; Wu et al., 2020). The technique tends to serve as a near real-time deformation monitoring tool, especially with the availability of modern SAR data offering higher revisit frequency and wider coverage, such as the Sentinel-1(Silva et al., 2021). However, automatic and prompt processing of large InSAR datasets remains a considerable challenge (Anantrasirichai et al., 2019a; Silva et al., 2021). Additionally, the lack of expert knowledge in both InSAR and geotechnical engineering also hinders the timely interpretation of ground deformation (Anantrasirichai et al., 2018; Rouet-Leduc et al., 2021). To address those issues, various methods have been developed to use MT-InSAR to retrieve deformation measurements and different physical models have been employed to understand geohazard characteristics (Ansari et al., 2017; Wang et al., 2020). Although those methods are very useful, analyzing many MT-InSAR interferograms for comprehensive ground deformation

* Corresponding author.
E-mail address: xl.ding@polyu.edu.hk (X. Ding).

https://doi.org/10.1016/j.jag.2024.103758
Received 9 January 2024; Received in revised form 25 February 2024; Accepted 8 March 2024
Available online 16 March 2024

analysis is still time-consuming and challenging.

The advent of machine learning (ML) techniques such as convolutional neural networks (CNNs) and multi-task learning (MTL) has revolutionized InSAR analysis. For instance, one notable application is the detection of volcanic deformation signals through the integration of InSAR measurements time series with independent component analysis (ICA) (Ebmeier, 2016; Gaddes et al., 2018, 2019). Encoder-decoder architectures have been used to identify subtle deformation (Sun et al., 2020). CNNs have been employed to analyze cumulative displacement to identify invisible deformation signals in single Sentinel-1 interferograms (Anantrasirichai et al., 2019b). CNNs have been utilized for binary classification of deformation associated with volcanic eruptions and earthquake events, achieving 86 % classification accuracy (Anantrasirichai et al., 2019a; Brengman and Barnhart, 2021). The accuracy of the classification of volcanic deformation was improved to 95 % by modifying the fully connected layers by the VGG16 model developed by Visual Geometry Group Lab of Oxford University, with the detection error of being approximately 2 km (Gaddes et al., 2021). In addition, multi-task learning (MTL) has also been developed and integrated with CNNs to reduce computational costs and improve training efficiency by training a single model to perform multiple related tasks simultaneously (Ruder, 2017; Zhang and Yang, 2018).

Despite the promising results achieved by CNNs in the deformation detection and interpretation of InSAR data, the extensive training data required for these models poses a challenge. However, publicly shared InSAR results often have limited training data available. Researchers have explored alternative approaches such as using pretrained models (Anantrasirichai et al., 2018; Gaddes et al., 2021), simulating a large InSAR dataset (Brengman and Barnhart, 2021), and self-supervised learning techniques (Bountos et al., 2022). On the other hand, CNN-based techniques are also often limited in the accuracy of deformation detection, consistently falling below 91 % due to the lack of consideration for global contextual information (Bountos et al., 2022) which is crucial for achieving higher accuracy.

To overcome those two problems, we propose an improved ML approach, based on the vision transformer model (ViT) instead of CNN, which utilizes self-attention mechanisms to capture global contextual information, a technique successfully used in natural language and computer vision tasks (Vaswani et al., 2017; Dosovitskiy et al., 2020; Wang et al., 2022). To enhance the reliability of training with limited InSAR data, the proposed method is operated as a two-stage model that transfers pre-trained weights from a model trained with optical images to that trained with a simulated InSAR dataset, followed by fine-tuning the model with real SAR interferograms. The model incorporates classification and localization losses to derive an overall loss function and includes a pooling layer for simultaneous classification and localization of deformations. The effectiveness of the MT-ViT model is evaluated using both volcanic and coseismic datasets, demonstrating the excellent capability of the proposed model in simultaneously detecting, locating, and interpreting ground deformation patterns based on single interferograms. A comprehensive investigation into the impact of hyperparameter tuning on the performance of the MT-ViT model is conducted to provide a reference for various InSAR applications using ViT model-based deformation interpretation. To facilitate practical use, a desktop application integrating the trained MT-ViT model, named SARViT, has been developed and is freely accessed by the public.

The rest of the paper is organized as follows: the architecture of the model and the training procedure will be presented in Section 2. Section 3 will present the experiments and results. A comprehensive discussion of the hyperparameters for the MT-ViT will be provided in Section 4 followed by the conclusions in Section 5.

2. Methodology

We will introduce in this section the architecture of the MT-ViT model and the training procedure, covering the data preparation, loss function, evaluation metrics, data manipulation, and fitting parameters for training and validation.

2.1. Architecture of MT-ViT model

The architecture of the MT-ViT is illustrated in Fig. 1 (a). It consists of patch embedding, a transformer encoder, and a pooling layer. The patch embedding includes breaking down the input interferogram into fixed-sized patches. The patches are then flattened, normalized, and linearly projected (Dosovitskiy et al., 2020). A learnable positional embedding is fed with each patch to inform the model about the position of the patches in the projection sequence (Dosovitskiy et al., 2020). A trainable class embedding is added at the beginning of each sequence to embed the classification characteristics.

Each transformer encoder contains a multi-head attention block and a multilayer perceptron neural network (MLP) as shown in Fig. 1 (b). The multi-head attention layer is composed of the concatenation of several self-attention modules that can process everything simultaneously as shown in Fig. 1 (c). The self-attention module calculates a set of attention weights for each input feature vector based on the relationships between all the feature vectors in the input, allowing the model to capture the long-range dependencies and relationships between the features. In the attention module, the embedded patches are fed as query, key, and value sequences. The query and key representation undergo a dot product matrix multiplication (i.e., MatMul) to produce a scoring matrix that represents how much attention is between two embedded patches as shown in Fig. 1 (d) and Eq. (1). The score matrix is scaled down by a scale factor to ensure more stability in the gradient calculation as the multiplication can have exploding effects. The attention score is converted into probability through a softmax function to drive the value matrix to extract the important information and suppress the irrelevant information. The MLP applies a non-linear transformation to the attended feature vectors (Dosovitskiy et al., 2020).

$$Attention(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V are the query, key, and value matrices, and d_k is the key scale factor.

A pooling layer is introduced after the transformer encoder as shown in Fig. 1 (a). Different pooling layers have been proposed including, e.g., Separate, Average, and Flatten layers. The Separate layer divides the embedding into class and position, computes the mean of the embeddings, and assigns the output of the class embedding to the class head and the positional embedding to the localization head. The Average layer computes the mean of the embeddings and assigns the average embeddings to the classification head and the localization head. The Flatten layer unrolls the whole embeddings and assigns the flattened representation to both the classification head and the localization heads. The output of the pooling layer is the input double MLP networks, the first with hidden neurons representing the number of classes and the second with hidden neurons representing the parameters of the bounding box.

2.2. Simulation of training data

A large simulated InSAR dataset composed of 100,000 interferograms per class was simulated to train the MT-ViT model (i.e., 300,000 and 400,000 interferograms for volcanic and coseismic deformation, respectively). Each of the interferograms contained unwrapped phases from a linear combination of deformation signal, stratified atmospheric errors, turbulent atmospheric errors, and orbital errors. The volcanic deformation sources included sill and dyke while the coseismic deformation sources were dip-slip, strike-slip, and thrust faults. The deformation signal for both the volcanic and coseismic activities was obtained from the forward path of the Okada model with different source

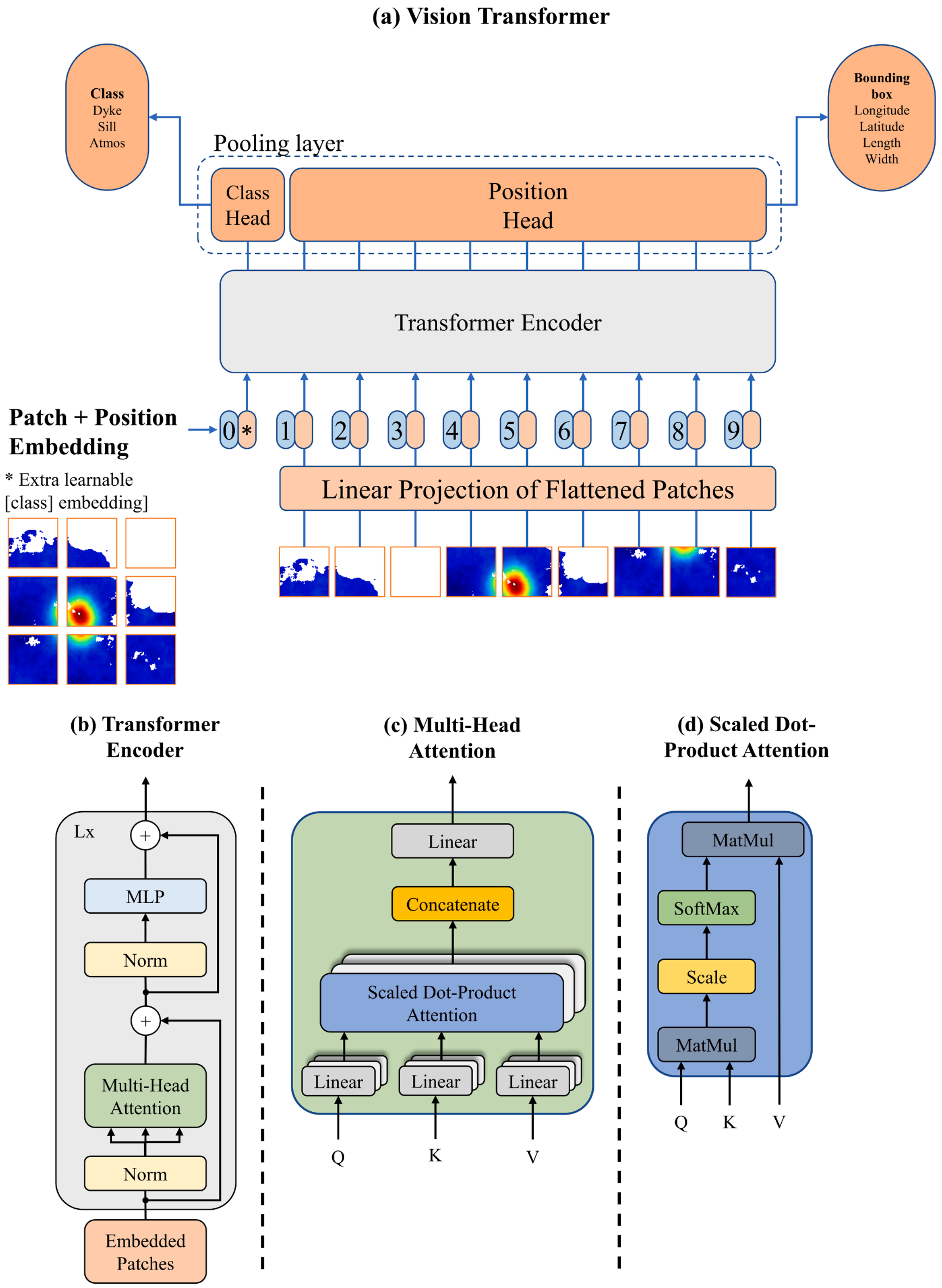


Fig. 1. Architecture of the MT-ViT model proposed for InSAR deformation detection, location, and interpretation. (a) Vision transformer. (b) Transformer encoder. (c) Multi-head attention mechanism. (d) Mathematical operation to calculate the attention between two patches.

parameters shown in Table 1 (Okada, 1985). The simulated deformation was projected to the radar line of sight (LOS) direction using the incident angle varying between 31 and 46 degrees and heading angles of −12 and 192 degrees for the ascending and descending tracks of the Sentinal-1 satellite, respectively. The topographic related atmospheric errors were obtained from randomly scaling the elevation acquired from the digital elevation model (DEM) over the target area. Fig. 2 shows the spatial distribution of the DEMs used to simulate the tropospheric errors for both volcanic and coseismic datasets. The turbulent atmospheric errors were obtained from the spatial correlation between pixels and randomly changing the correlation length (Lohman and Simons, 2005). The orbital errors were simulated using a linear ramp equation (Ebmeier, 2016; Sun et al., 2020). Finally, the interferograms thus derived were randomly masked to simulate the decorrelation associated with the Sentinel-1 interferograms. We chose a signal-to-noise ratio threshold of 2.0 to –guarantee clear deformation. For volcanic deformation, the

Table 1
The parameters were applied to the Okada model to simulate displacement.

Source	Length (km)	Width (km)	Depth (km)	Slip (m)	Opening (m)	Strike (°)	Dip (°)	Rake (°)
Sill	2–6	2–6	1.5–3.5	0	0.2–1.0	0–360	0–5	0
Dyke	0–10	0–6	2–5	0	0.1–0.7	0–360	75–90	0
Normal	10–50	5–30	0–20	0.5–5.0	0	0–360	10–60	−90
Thrust	10–50	5–30	0–20	0.5–5.0	0	0–360	10–90	90
Strike-slip	10–50	5–30	0–20	0.5–5.0	0	0–360	50–90	0, 180

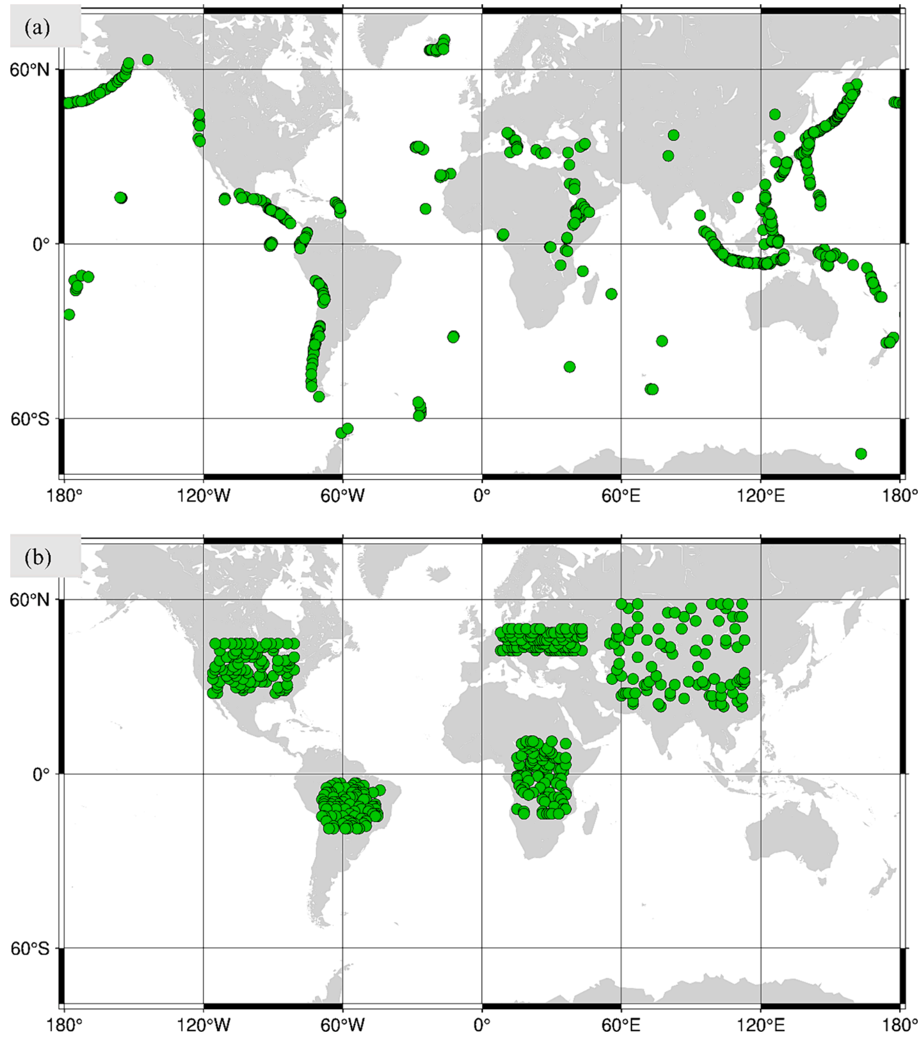


Fig. 2. The spatial distribution of the chosen DEMs to randomly create the topographic correlated atmospheric errors. Each green circle shows the center of the extracted DEM. (a) Volcanic locations. (b) Coseismic locations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

range was set from 0.05 m to 0.5 m, while for coseismic deformation, it was set from 0.05 m to 1.5 m. The label of each interferogram was the deformation source used to create the displacement. The center of the rectangular fault coincided with the center of the bounding box. The length and the width of the bounding box were extended till reaching the pixel that had at least 20 % of the maximum displacement. The interferograms were built based on the SyInterferoPy generator with additional features (Gaddes et al., 2019).

2.3. Real SAR data

The publicly available SAR volcanic dataset called VolcNet was also used for training and testing the MT-ViT model. The dataset contained a time series of interferograms of the most vulnerable volcanic areas in the world as shown in Fig. 3. The number of interferograms was about 500,000 where only 0.2 % were dyke volcanos. Therefore, the dataset had unbalanced labels as shown in Fig. 3. The duration and magnitude of the deformation in each time series were also included. All the possible combinations of the SAR acquisitions were used to create the interferograms. Fig. 4 (a) shows an example of the generated interferograms of the Wolf Volcano between 06/05/2016 and 23/06/2016. Each interferogram was labeled with the existing deformation signal and the deformation pixels were surrounded by a black bounding box. Fig. 4 (b) shows the duration, magnitude, and type of the volcanic signals over the same period as that of the Wolf Volcano.

In addition, interferograms containing coseismic deformation were

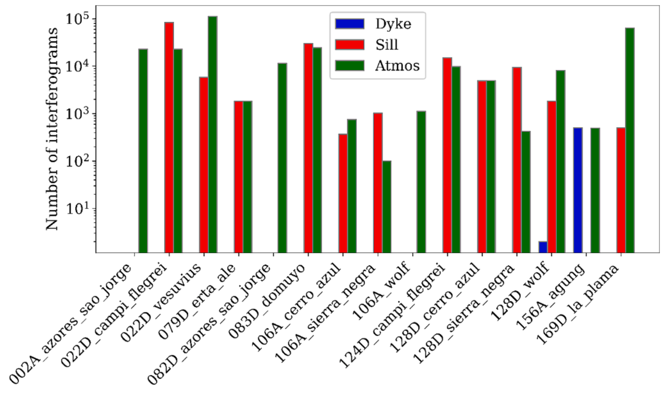


Fig. 3. Number of interferograms generated from each of the time series of SAR acquisitions.

also created for some of the famous earthquakes during 2016–2021 as shown in Fig. 5. Temporal baselines of 12, 24, and 36 days were chosen to create all the possible interferograms for both the ascending and descending tracks. The interferograms were processed using the ISCE software package (version 2.5, (Roseu et al., 2012)), the three-arc-second SRTM DEM (Farr et al., 2007), and precise orbit ephemeris. The phase unwrapping process was carried out using the minimum cost flow method (Costantini, 1998). Twenty-seven interferograms were processed while 9 were only noise and the remaining 18 contained coseismic deformation (e.g., 5 dip-slip, 5 strike-slip, and 8 thrust fault interferograms). Each interferogram was classified based on the fault parameters according to the report of the United States Geological Survey (USGS, 2022). Each interferogram was created by cropping a region of 280 × 280 pixels around the deformation pattern.

The center of the bounding box was placed at the pixel with the maximum displacement while the length and width of the bounding box were extended to pixels with 20 % of the maximum displacement for both the volcanic and coseismic deformation.

Data augmentation (DA) techniques were used to generate unique real interferograms with the associated bounding box while maintaining the same spatial resolution. The augmented real dataset is composed of 30,000 and 4,000 unique interferograms for volcanic and coseismic deformations, respectively. Brengman and Barnhart, (2021) utilized 32 interferograms and data augmentation to produce 5,184 augmented interferograms for training and 1,152 for validation, which were then used to fine-tune the ResNet model.

2.4. *Loss functions and metrics*

The MT-ViT model integrates two separate heads, namely the classification head and the localization head, to interpret the type of the deformation signal and its location, respectively. The classification head returns the probability that the deformation is related to a particular type of deformation source through a softmax activation function while the localization head returns the length, width, and center of the bounding box that contains the deformation through a linear activation function. The cross entropy (CE) loss function was chosen for the classification head and the mean square error (MSE) loss function was chosen for the localization head (Gaddes et al., 2021). The overall accuracy (OA) and area under the curve (AUC) (Bradley, 1997) were used as the assessment metrics for the classification head while mean absolute error (MAE), and intersection over union (IoU) were taken for the localization head. Table 2 summarizes the mathematical expression of the loss and metric functions used in the training process.

The total loss is the linear weighted combination of the CE and MSE losses,

$$\mathscr{L} = \lambda_{CE} * \mathscr{L}_{CE} + \lambda_{MSE} * \mathscr{L}_{MSE} \tag{2}$$

where λ_{CE} is the weight factor for the CE loss while, λ_{MSE} is the weight factor for the MSE loss. λ takes a value from 1.0, 5.0, 10.0, and 20.0 to allow an optimal combination of $\mathscr{L}_{CE}$ and $\mathscr{L}_{MSE}$.

2.5. *Data manipulation*

To make the data compatible with the MT-ViT algorithms, the unwrapped phase of the pixels in each interferogram was scaled to ranges from −1 to 1 while the masked pixels were filled with 0 s, and the interferograms were transformed into one-channel grayscale tensors. The pre-trained weights of the ViTs were adjusted to align with the interferogram by averaging the weight values over the channel dimension. In addition, different batch sizes (8, 16, 32, and 64) and patch sizes (16 × 16 and 32 × 32) were tested to identify the optimal values. A detailed discussion of this point will be presented in Section 4.

The training of the ViT model took place in two stages. The first was to train the model on a huge dataset that had a typical sample size of 10 million. The second was to fine-tune the trained weights on a desired dataset that has fewer samples and variety (Steiner et al., 2021). To overcome the challenge of only limited InSAR training data available, we made use of the pre-trained weights of ViTs obtained with ImageNet21K (i.e., 14 million images and 21,843 classes) (Wightman, 2013) in the first stage. We added a transfer stage to project the pre-trained weights from the optical images to the InSAR data domain. In the second stage, the fully connected part of the MT-ViT model was fine-tuned on a small real dataset of SAR interferograms, with 80 % used for training and 20 % for validation. Weight factors for the CE and MSE loss functions were initially set to 1.0. Different weight factors (1.0, 5.0, 10.0, and 20.0) were then explored to optimize the model performance. A learning rate of 0.00002 was employed to prevent weight disturbance, and early stopping was implemented to prevent overfitting and to reduce the unnecessary training epochs. The Adam optimizer was used to minimize the overall loss. The training was carried out using a graphics processing unit (GPU) with 16 GB of memory and implemented in the PyTorch framework.

3. Experimental results

In this section, we conducted an experiment to compare the performance of CNN and the proposed ViT model in interpreting and locating volcanic and coseismic deformation.

3.1. *Model training*

The MT-ViT model was trained separately with volcanic and coseismic datasets by considering the different sizes of the affected regions and the magnitudes of the deformation. The areas of coseismic deformation need to extend up to tens of kilometers to encompass the entire deformation areas (Zhao et al., 2021). Therefore, the spatial resolution for the volcanic deformation and the coseismic deformation were approximately 92 m and 490 m, respectively.

To evaluate the proposed two-stage training strategy, we started by training with a simulated volcanic dataset and a Small ViT architecture with a patch size of 16 × 16 and an Average pooling layer. Then, fine-tuning the model with a real dataset resulted in suboptimal performance with an OA and IoU below 60 % and 30 %, respectively, which is mainly due to the sub-performance to capture the attention features of the interferograms from the limited size of the simulated dataset. As the comparison, in the first stage, we used the pre-trained weights of ViTs obtained with the ImageNet21K dataset. In the transfer stage, each model was adjusted with a simulated dataset to adopt the pretrained attention for the InSAR domain. After the first epoch, the OA is significantly improved up to 96.7 % while the IoU did not exceed 5 % because the ViTs were constructed for image classification only. After the training procedure, all the results have been summarized in Table 3.

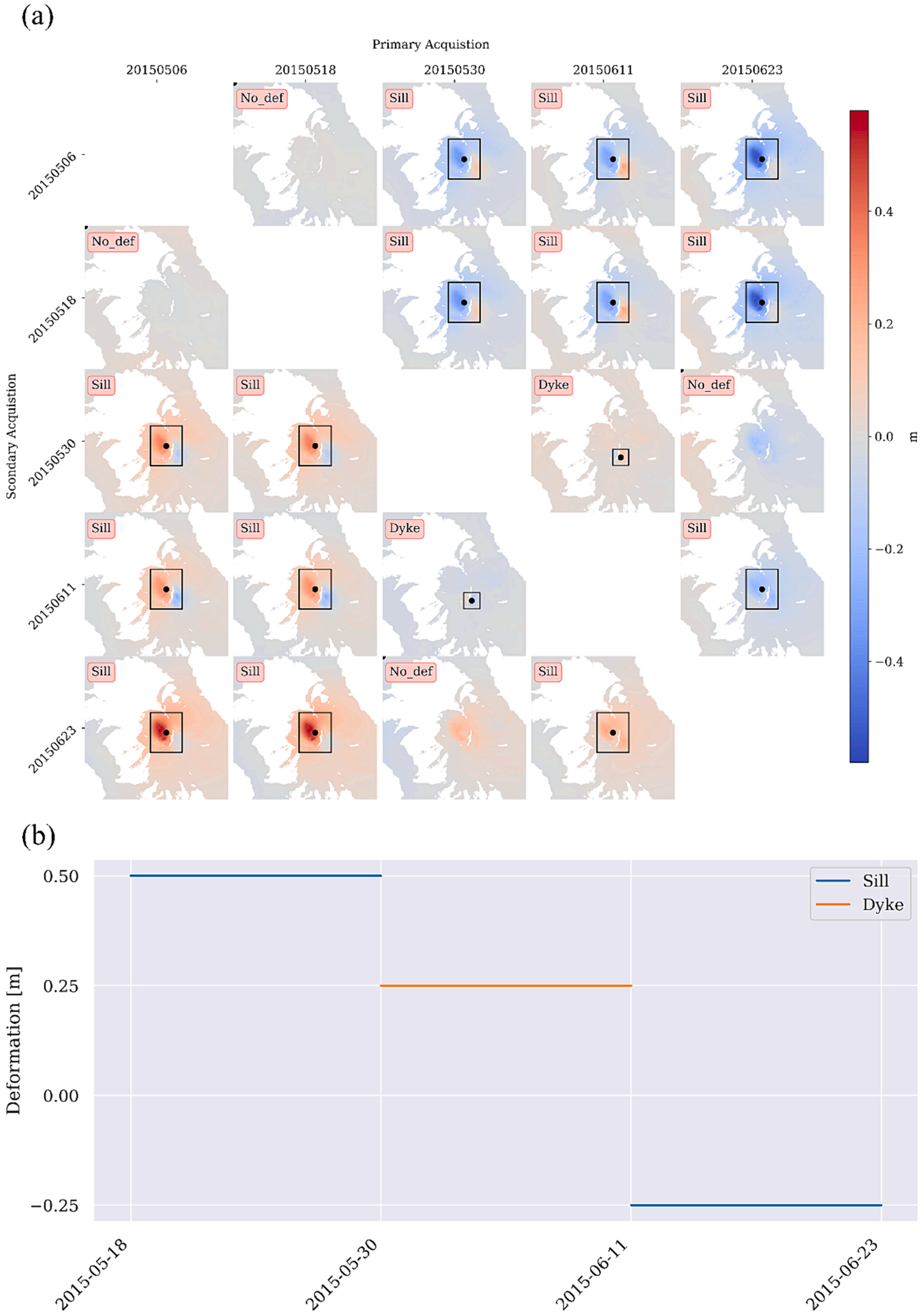


Fig. 4. Examples of the generated interferograms of the Wolf Volcano. (a) All the possible interferometric combinations of the SAR acquisitions of Track 128D over Wolf Volcano during 06/05/2016–23/06/2016. Displacements were present in the bound boxes. (b) The magnitude and type of the volcanic signals.

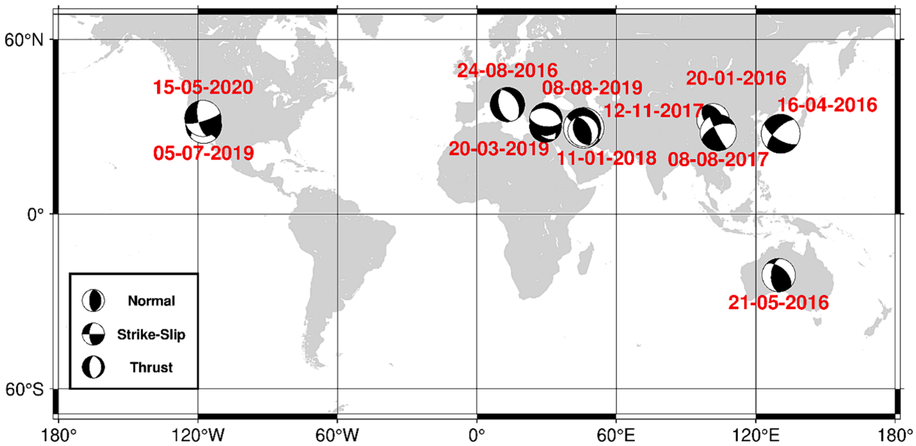


Fig. 5. Distribution of some strong earthquakes i.e., above 5.2 (2016–2020) for experiments for detection of coseismic deformation based on the source parameters from USGS moment tensor reports (USGS, 2022).

Table 2
Loss functions and assessment metrics for classification and localization heads.

Output head	Final layer	Losses	Metrics
Classification	SoftMax	$\mathscr{L}_{CE} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} p_i^{(j)} \log\left(\widehat{p}_i^{(j)}\right)$	$OA = \frac{T_P + T_N}{T_P + F_P + T_N + F_N}$
Localization	Linear	$\mathscr{L}_{MSE} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{b}\left(y_i^{(j)} - \widehat{y}_i^{(j)}\right)^2$	$MAE = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{b}\left\|y_i^{(j)} - \widehat{y}_i^{(j)}\right\|$
			$J(A,B) = \frac{\|A \cap B\|}{\|A \cup B\|}$

In the equations above, n is the number of training samples; c is the number of classes (i.e., 3 for volcanic signals); p is the one-hot encoded vector for each interferogram in a binary mode; $\widehat{p}$ is the predicted probability; b is the number of bounding box parameters (i.e., 4); y is the ground truth parameters; $\widehat{y}$ is the predicted parameters; T is the number of the correctly classified cases; F is the number of the misclassified cases; P is number of the positive samples; N is number of the negative samples; J is the Jaccard similarity coefficient; A is the area of the predicted bounding box; and B is the area of the ground-truth bounding box.

Table 3
Training results of MT-ViT model based on simulated and real InSAR datasets.

Deformation	Dataset	OA (%)	AUC (%)	IoU (%)	MAE (pixel)	Epochs required
Volcanic	Simulated	99.6	99.8	54.2	2.5	17
	Real	99.4	99.8	54.1	3.5	10
Coseismic	Simulated	99.7	99.9	69.5	1.1	52
	Real	99.0	99.8	67.7	0.8	50

With the proposed two-stage training strategy, the best OA and IoU were 99.4 %, and 54.1 %, respectively, and the lowest validation loss was achieved after 17, and 11 epochs respectively for the simulated and real volcanic datasets.

Based on the previous study by (Gaddes et al., 2021), simulating a more realistic InSAR dataset can reduce misclassification caused by strong atmospheric signals. Table 3 shows that there are no significant differences between the results from the real and simulated datasets, indicating the correctness of simulating and using both the coseismic and volcanic signals in this study. When comparing different deformation characteristics, such as volcanic and coseismic deformation, the IoU is increased from 54.1 % to 67.7 %, indicating the model performed better in locating the coseismic deformations compared to volcanic deformation The larger affected regions and associated deformation of coseismic events contribute to this performance difference. The localization accuracy for the coseismic deformation was better than that for the volcanic deformations, the localization error remained the same, i.e., under 1 km. The slight differences in deformation localization may have been caused by the boundary calculation process and the augmentation of the bounding boxes. Brengman and Barnhart, (2021) reported 85 % overall accuracy when employing two different channels, such as wrapped and unwrapped phase data, which might not align due to unwrapping errors. They encountered a training failure when incorporating the declaration, leading to a decrease in accuracy from 99.7 % to 93.6 % during transfer learning. To remedy this issue, we substituted the decorrelation from random pixels with zeros, which enhanced result stability.

For a visualization analysis, we have illustrated some examples of the results of volcanic deformation detection with the simulated interferograms in Fig. 6 (a–c), (d–f), and (g–i), which contain volcanic dyke, sill/point deformation, and atmospheric signals. The model distinguished the atmospheric signals correctly in both classification and localization. The model was able to correctly classify and locate the deformation signals. The deformation signals are successfully classified and located with a discrepancy between the predicted deformation area and the ground truth. After fine-tuning the model with the real dataset, the model could correctly classify both the atmospheric artifacts and the deformation signals in the real interferograms, as shown in Fig. 7. The predicted classification head and localization head were highly consistent, demonstrating the improvement compared with the VUDL-NET-21 model, which predicts atmospheric signals while a bounding box surrounds some other pixels (Gaddes et al., 2021). The volcanic dyke signal can be accurately identified by the proposed method even using a low coherence interferogram as shown in Fig. 7 (c).

To evaluate the capability of the MT-ViT model in distinguishing volcanic eruption from atmospheric delays, real InSAR datasets covering various volcanoes were analyzed. Some experimental results are illustrated in Fig. 8. It can be observed that all the volcano deformations were accurately located with high precision, except for the deformation regions over the Cerro Azul Volcano, as highlighted in Fig. 8 (d). This misidentification may be attributed to the model being misled by atmospheric delay. For further examination, detailed experimental results for one volcano, named Wolf, are presented in Fig. 9. All the

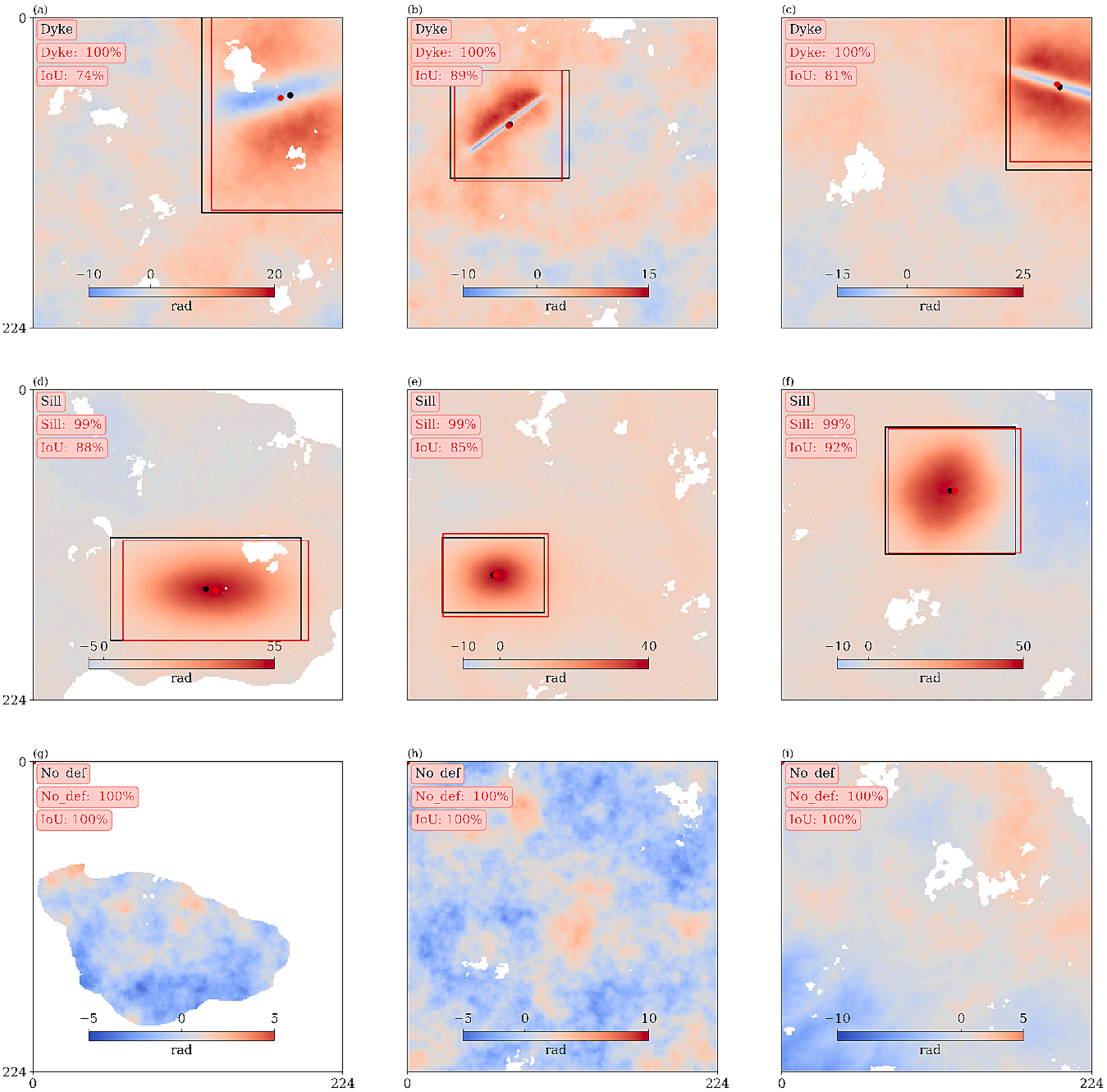


Fig. 6. Results of detection and localization from simulated InSAR dataset. The first row shows three interferograms that contained a volcanic dyke. The second row are interferograms that include sill. The third row are interferograms that had atmospheric signals. The ground truth signals, and their bounding boxes are in black. The predicted signals IoU, and the bounding boxes are in red. The bigger the bounding box is, the more significant the localization errors were. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

interferometric pairs generated by the SAR acquisitions are displayed, indicating the accuracy of the classification over all the interferograms except for three of them. These include the interferogram of 2015/05/06–2015/05/18, which is incorrectly interpreted as a sill/point deformation, the interferograms generated with 2015/05/18 and 2015/05/06, identified as atmospheric signals because the eruption started on 25 May (2015) and the interferograms generated with 2015/05/30 and 2015/06/23, which were identified as sill signals. De Novellis et al., (2017) pointed out a second phase of deformation occurring over Wolf Volcano in June - July during the eruption activity. Fig. 4 shows that this magnitude deformation (i.e., 5 cm) tends to be of the sill/point type. Therefore, the proposed model is also able to detect and locate the post-event deformation.

Similarly, an experiment using simulated interferograms with coseismic deformation was displayed in Fig. 10. The IoU reached about 70 % and the localization errors were 1.1 km. the localization accuracy was nearly the same as for real interferograms as shown in Fig. 11. It is shown that increasing the number of real interferograms is necessary for more robust fine-tuning, but the interferometric process and accurate labeling are labor-intensive.

3.2. Comparative experiments

To verify the effectiveness of the proposed method, we have conducted comparison experiments with Gaddes et al., (2021), who updated the fully connected part of the VGG16 model and used transfer learning to maintain the same weights of the convolutional part to create a MT-CNN. The comparison results are summarized in Table 4. In terms of the number of trainable parameters, the fully connected part is almost four times that of the convolutional part due to the feature difference between the interferograms and the optical images. In contrast, the MT-ViT has approximately one-third of the free parameters of the MT-CNN. Regarding the accuracy of the classification and localization, MT-ViT has shown improvements of 4 % and 50 %, respectively. The

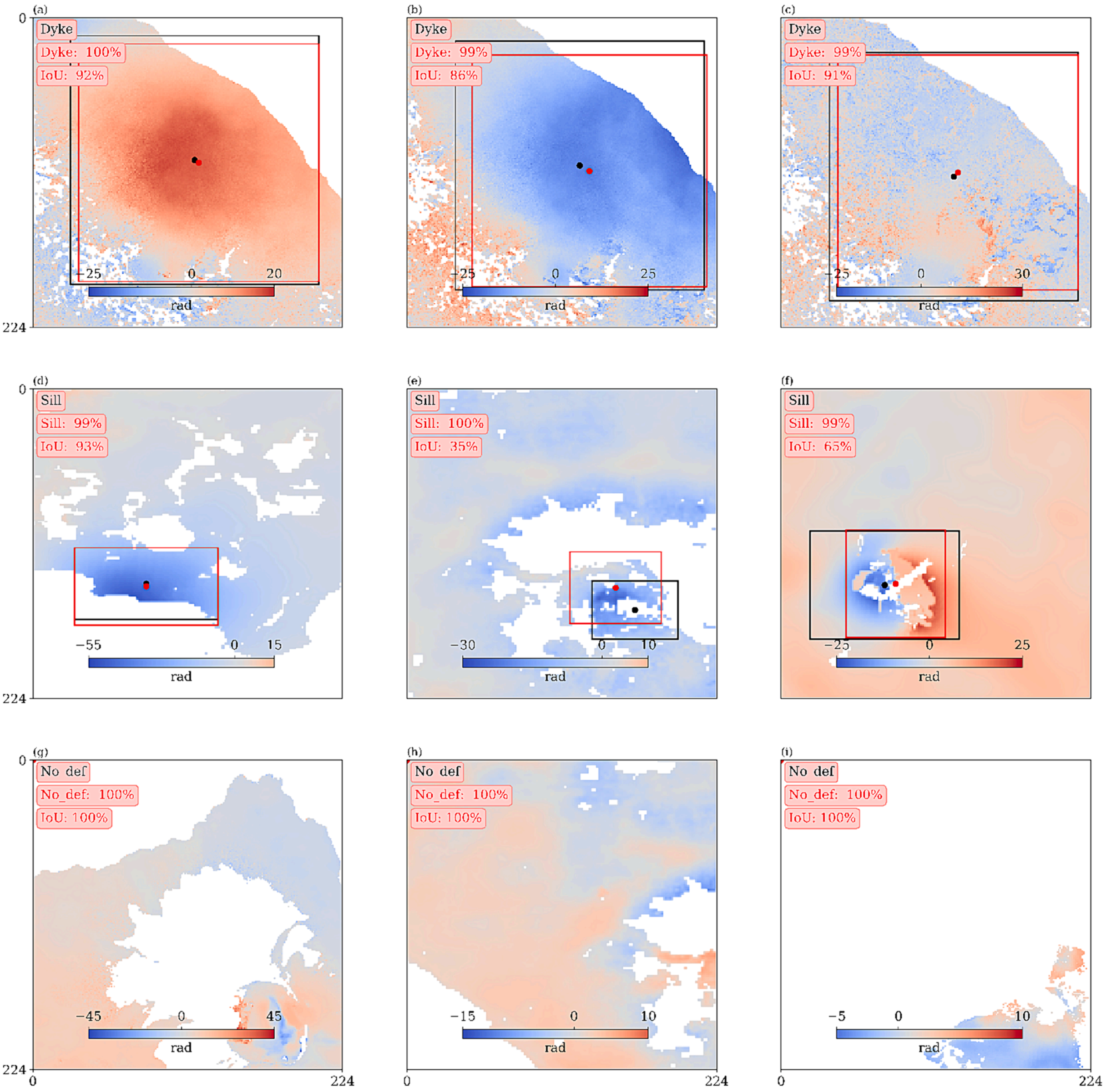


Fig. 7. Results of classification and localization of volcanic deformation from a sample of the real InSAR dataset (VolcNet). The first row shows three interferograms that contained a volcanic dyke. The second row are interferograms that include a sill. The third row are interferograms that had atmospheric signals. The ground truth signals, and the bounding boxes are in black. The predicted signals, IoU and bounding boxes are in red. The bigger the bounding box is, the more significant the localization errors were. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

localization error was reduced by half, with 10 pixels corresponding to approximately 0.9 km when using a 90-m SRTM.

To assess the computational load, A ResNet model (He et al., 2016) was modified to perform multi-task processing and compared with the proposed MT-ViT model. More specifically, all the experiments utilize pre-trained weights from training using optical images without any adjustment. ResNet model had 13.5 million free parameters, of which 3,600 were in the fully connected part. it was retrained using simulated interferograms and fine-tuned with real interferograms. The fine-tuning process started with making only the fully connected part trainable. The fine-tuning process of ResNet-34 is shown in Fig. 12. At each step, we changed the status of the connected residual blocks to be trainable until reaching the first convolutional block. The comparison results are summarized in Table 5. The CNN model outperformed the ViT model when the simulated data was used, while its performance degraded when the real data was used. The classification and localization accuracies were improved by 1 %, and 12 %, respectively, when the MT-ViT model and real data were used. It is worth noting that, as shown in Fig. 12, the IoU is not affected by the trainable parameters used and is always below 50 %.

3.3. Desktop application for using MT-ViT model

Based on the training conducted in this study, we have developed a desktop application with Qt framework and shared it as a free application. It can be accessed through: https://github.com/m-elhussieny/GeohazardLab. The schematic design of the application is shown in Fig. 13. The user needs to select the type of the deformation (i.e., coseismic, or volcanic) and the input data (i.e., real, or synthetic). In addition to the ViT architecture, it also supports the ResNet model. The user can select interferograms from a local disk. The application can display an interferogram under study, the classification label and the bounding box surrounding a deformation area.

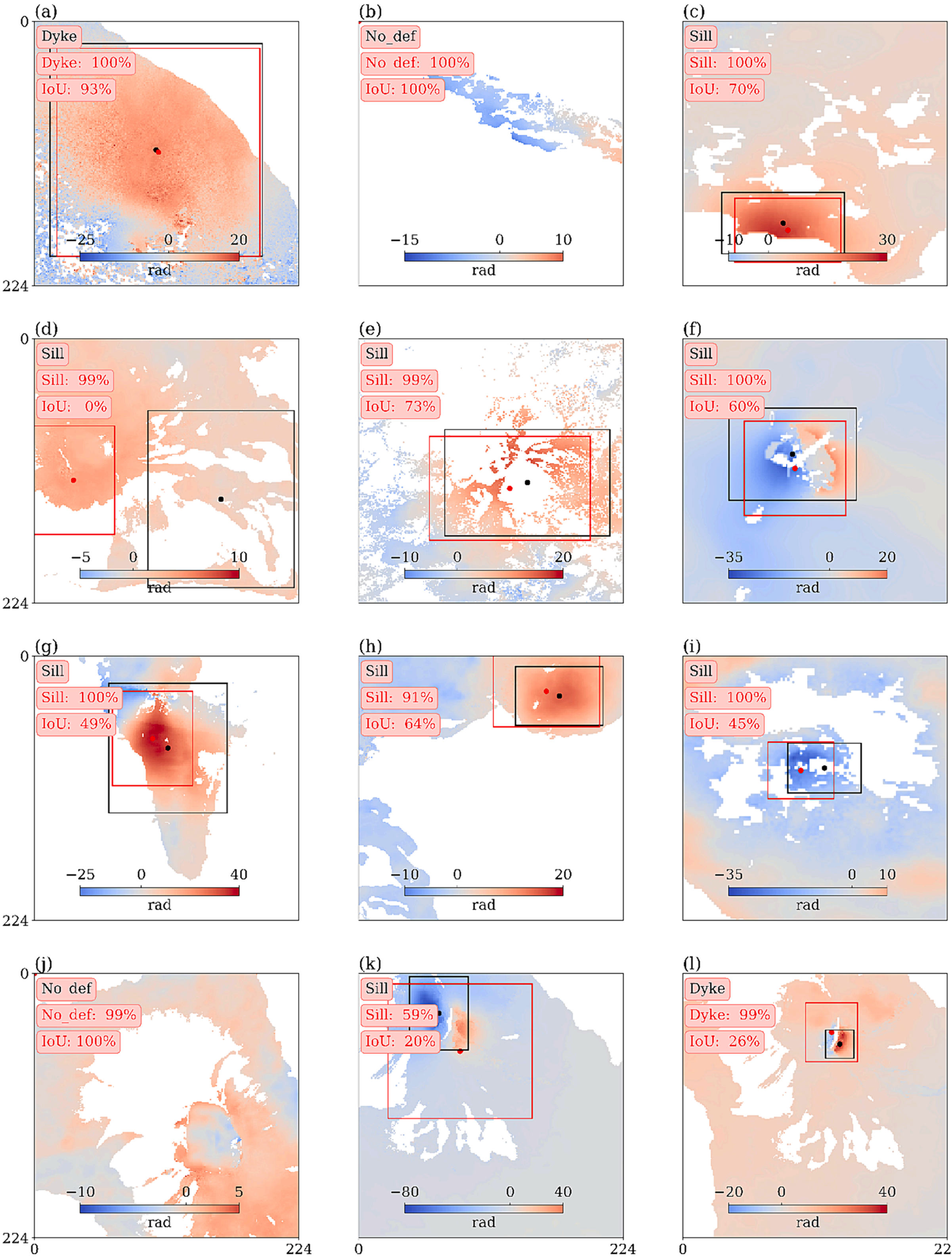


Fig. 8. Results of classification and localization from a sample of the real InSAR dataset (VolcNet). (a) Agung volcano. (b) Azores Sao Jorge volcano. (c) Campi Flegrei volcano. (d) Cerro Azul volcano. (e) Domuyo volcano. (f) Erta Ale volcano. (g) La Plama volcano. (h) Sierra Negra volcano. (i) Vesuvius volcano. (j) Wolf volcano (k) Wolf volcano. (l) Wolf volcano.

4. Discussion

We will in this section analyze based on the VolcNet dataset crucial factors that affect the proposed model, including architectural choice, patch size, pooling layer, batch size, and weight factors.

4.1. Impact of model structure

The available architectures with ImageNet weights were use, including ViT (Dosovitskiy et al., 2020), ConViT (d'Ascoli et al., 2022), DeiT (Touvron et al., 2021), Swin (Liu et al., 2021), ResNet (He et al.,

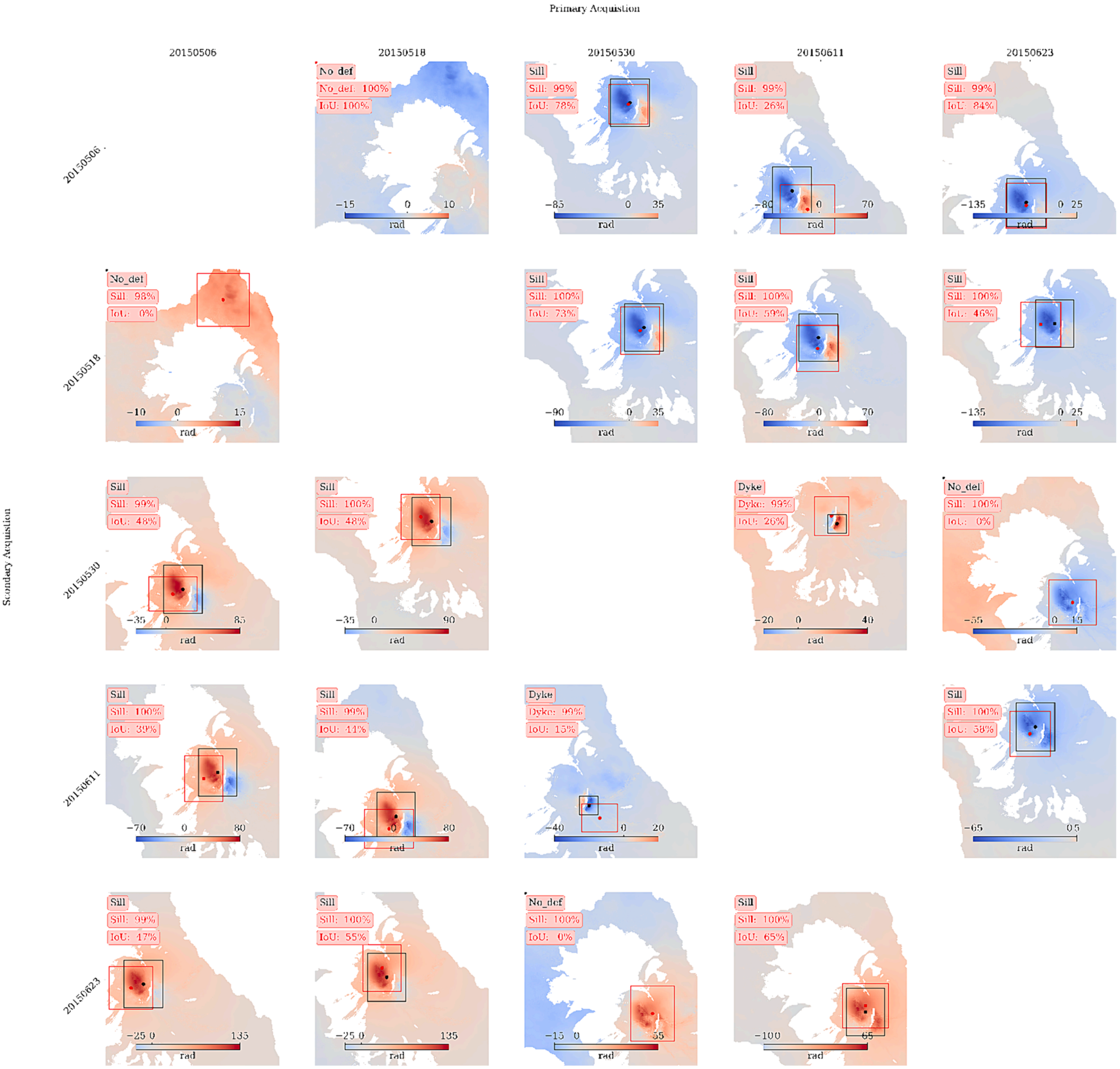


Fig. 9. Results of classification and localization using all possible combinations of the SAR acquisitions of Track 128D during 06/05/2016–23/06/2016 over Wolf Volcano.

2016), VGG (Simonyan and Zisserman, 2015), Inception (Szegedy et al., 2015), DenseNet (Huang et al., 2017), and ConvNext (Liu et al., 2022) as shown in Fig. 14. ViT and VGG architectures demonstrated similar performance in both stages. Consequently, ViT can serve as a classifier using only ImageNet weights without modification. However, the ResNet architecture was not suitable for this dataset. We observed an improvement in classification accuracy of 2 % to 8 % and an improvement in location accuracy of 25 % to 40 %. Qualitatively, the t-SNE plots in Fig. 15 clearly illustrate the impact of transferring ImageNet weights to the simulated data, enabling the model to accurately distinguish negative examples. Although self-supervised learning has shown great potential in enhancing feature extractor weights with unlabeled data, it requires a large batch size to include sufficient positive and negative examples. This demand for resources may not be available to us for this study (Bountos et al., 2022; Chattopadhyay et al., 2023).

4.2. Impact of model architecture

To test the impact of the model architecture, the publicly available ViT models including the Base, Small, and Tiny models are used, which has been summarized in Table 6 (Dosovitskiy et al., 2020; Steiner et al., 2021). All the models have the same layers (depths) with different embedding sizes and numbers of heads in the multi-headed attention. The Base model has four times the number of trainable parameters as that of the Small model, and sixteen times that of the Tiny model. Each model takes three-channel tensors as input where each interferogram is considered a one-channel tensor. Consequently, the pre-trained weights in the patch projection layer were averaged for the channel dimension. The results of comparison of the performances of the models are presented in Table 7. The OA was almost the same for all the models while the maximum difference in the IoU was 4.7 %. Interestingly, the Small model achieved the same OA as the Base model but a 5.5 % improvement in the MAE and a 9 % reduction in training time compared to the Base model.

4.3. Impact of patch size

We used the Small model with patch sizes of either 16 × 16 or 32 × 32 pixels. Increasing the patch size reduced the computational cost of patch-to-patch attention and led to an overall increase in the

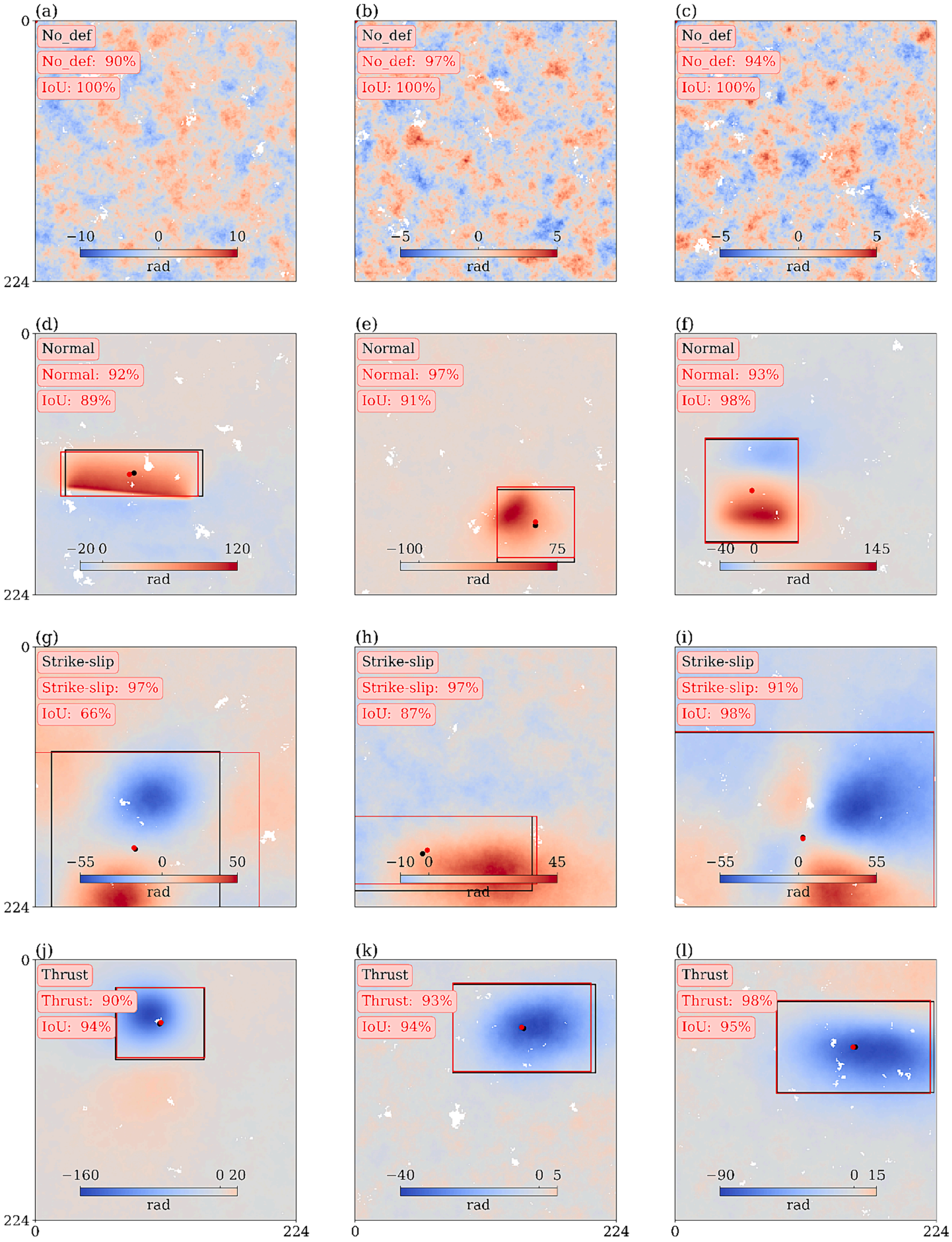


Fig. 10. Results of classification and localization when using simulated InSAR dataset. The first row are interferograms that had atmospheric signals. The second row shows interferograms that contained normal fault deformation. The third row gives interferograms that included strike-slip deformation. The fourth row contains interferograms that included thrust fault deformation. The ground truth signals, and their bounding boxes are in black. The predicted signals, IoU, and their bounding boxes are in red. The bigger the bounding box is, the more significant the localization errors were. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

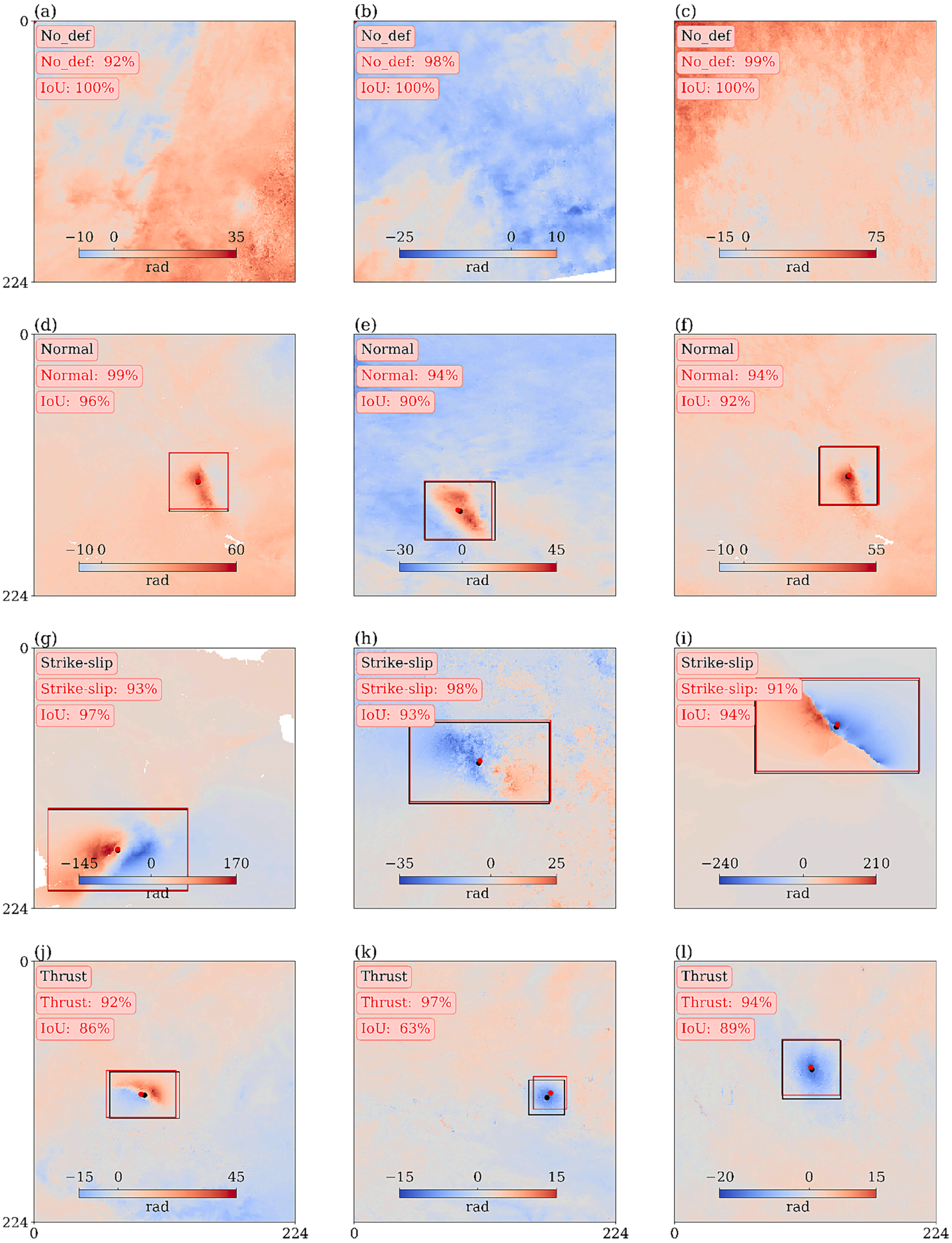


Fig. 11. Results of classification and localization using a sample of the real InSAR dataset. The first row shows interferograms that had atmospheric signals. The second row is interferograms that contain normal fault deformation. The third row gives interferograms that included strike-slip fault deformation. The fourth row contains interferograms that include thrust fault deformation. The ground truth signals, and their bounding boxes are in black. The predicted signals, IoU, and the bounding boxes of the deformation signals are in red. The bigger a bounding box is, the more significant the localization errors were. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Comparison between CNN and ViT for classifying and detecting volcanic eruption using real dataset.

Study	Base model	Free/total parameters (million)	OA (%)	AUC (%)	IoU (%)	MAE (pixel)	Error (pixel/km)
Gaddes et al., (2021)	CNN	60.5/75.2	95.5	—	—	—	20/1.8
This study	ViT	21.7/21.7	**99.4**	**99.8**	**54.1**	**3.5**	**10/0.9**

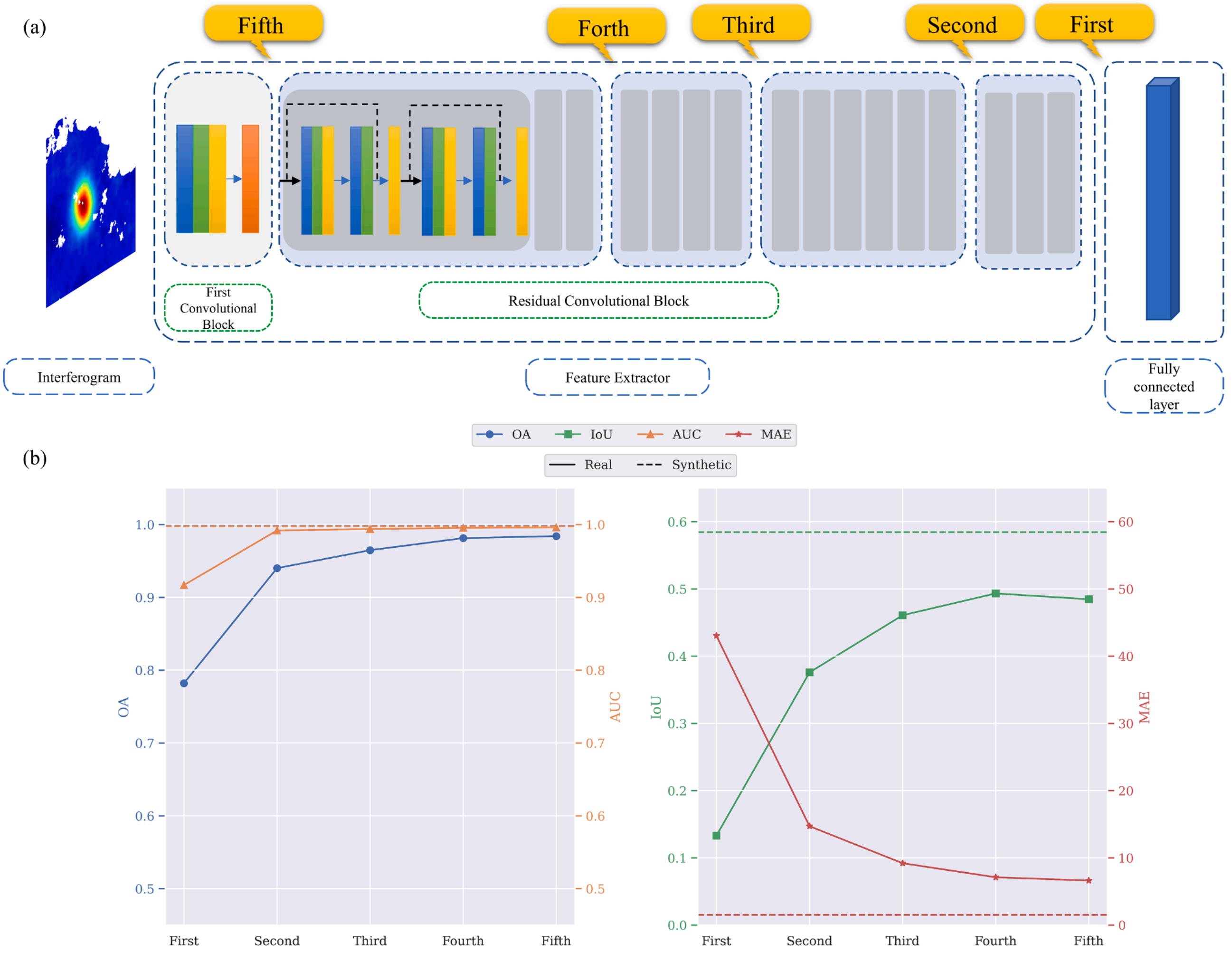


Fig. 12. The fine-tuning process of ResNet-34. (a) The schematic design of the ResNet-34. (b) The evaluation metrics at each step. The results of simulated data are shown in dash lines while the real data are in solid lines.

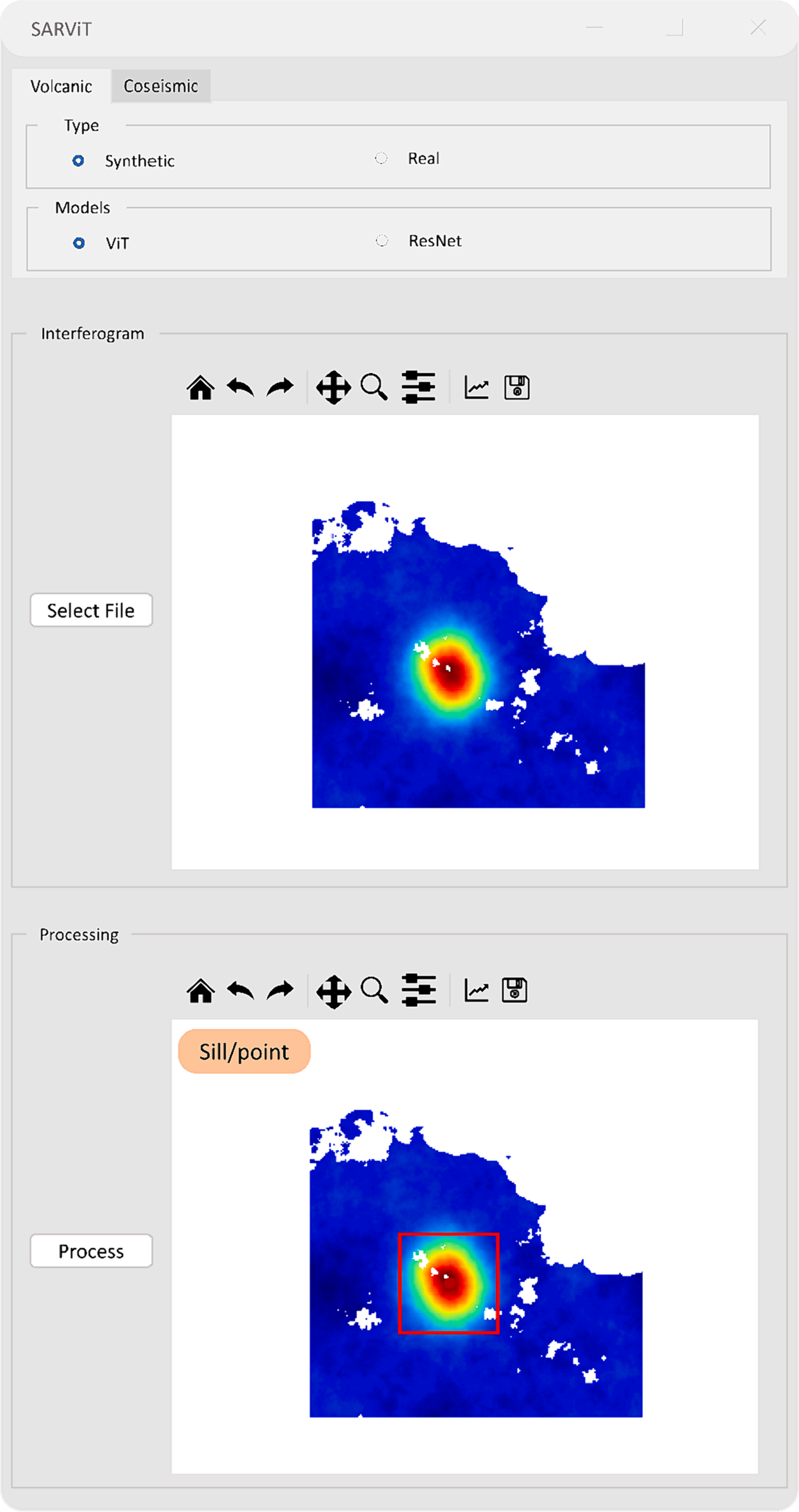
Table 5
Comparison between CNN and ViT used to classify and detect volcanic eruption.

Dataset	Base model	OA (%)	AUC (%)	IoU (%)	MAE (pixel)
Simulated	CNN	**99.9**	**99.8**	**58.5**	**1.5**
	ViT	98.9	99.7	54.2	2.5
Real	CNN	98.4	99.6	48.5	6.6
	ViT	**99.4**	**99.8**	**54.1**	**3.5**

computational cost. The main distinction between these two patch sizes is in the number of trainable parameters within the patch embedding layer and the patch projection layer. The shape of the projection filter is (embedding size, number of channels, patch size, patch size) whereas the shape of the embedded patches is (number of patches, embedding size). Utilizing a patch size of 16 × 16 to divide a 224 × 224 interferogram resulted in 14 × 14 patches, whereas a patch size of 32 × 32 produces 7 × 7 patches. Increasing the patch size from 16 × 16 leads to an increase in the trainable parameters (Table 6), but the results only showed a slight improvement in the AUC values (Table 8). The Small ViT model with a patch size of 16 × 16 exhibited a 5.3 % enhancement in the IoU, a 14.6 % improvement in the MAE, and a 62.9 % reduction in the training time.

4.4. *Impact of pooling layers*

To assess the impacts of different pooling layers, a pooling layer was added after the transformer encoder to adjust the representation flow. Three types of pooling layers were utilized, i.e., Flatten layer, Separate layer, and Average layer. Based on the results presented in Table 9, it was observed that the Flatten layer achieved the best performance for the classification task. The trainable parameters were approximately 185 times of those for the Separate and Average layers. On the other hand, the Average layer had the best performance for the localization task. Although the Flatten layer required the lowest number of epochs, the computational cost i.e., trainable parameters increased as presented



SARViT

Volcanic | Coseismic

Type

● Synthetic ○ Real

Models

● ViT ○ ResNet

Interferogram

Processing

Sill/point

Fig. 13. Schematic design of the desktop application.

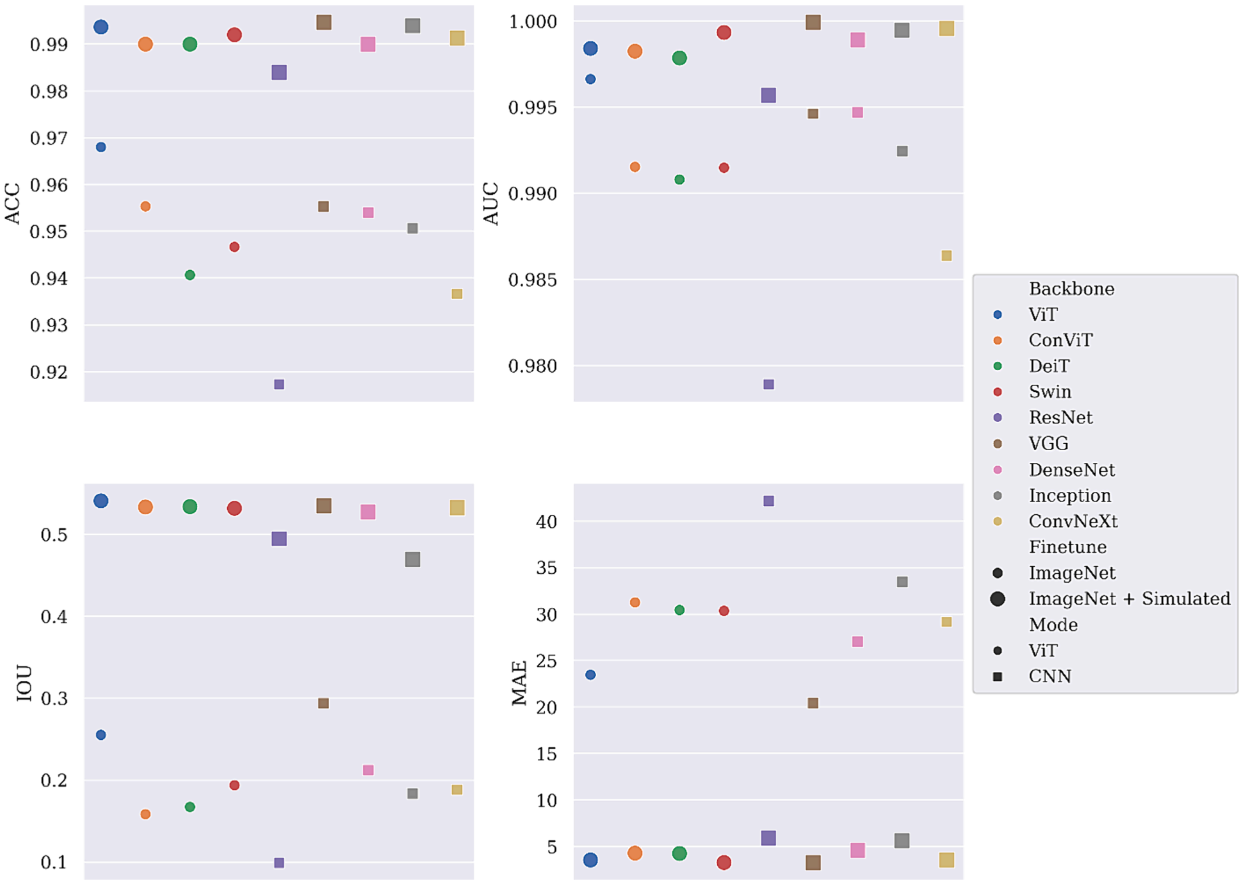


Fig. 14. Performance comparison of different architectures on the VolcNet dataset, utilizing ImageNet weights transferred to a simulated dataset. ViT-based models are denoted by circles, while CNN-based models are represented by squares. The smaller shapes indicate models fine-tuned with ImageNet weights, whereas the larger shapes represent models fine-tuned with ImageNet weights updated using simulated data.

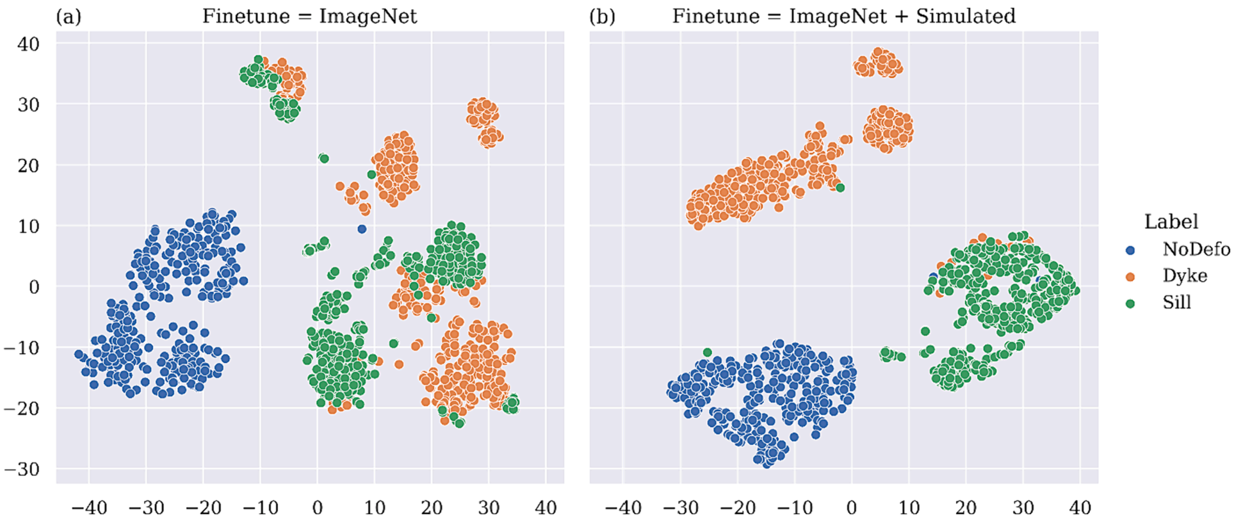


Fig. 15. T-sne comparison of the performance of the small vit model on the volcnet dataset, finetuned with imagenet weights, and imagenet weights updated using simulated data.

in Table 9. Considering the superior performance of the Average layer in the localization task, comparable performance in the classification task, and lower computational costs, the Average layer will be discussed below further.

4.5. *Impact of batch sizes*

It is commonly understood that larger batch sizes result in reduced computational time when utilizing parallel GPUs. This hover may also lead to poorer generalization. To evaluate the effects of batch sizes during the training process we employed batch sizes of 8, 16, 32, and 64

Table 6
Hyperparameters in publicly available ViT models.

Architecture	Patch size (pixel × pixel)	Embedding size	Depth	Heads	Expansion ratio	Free parameters (million)
Base	16 × 16	768	12	12	4	85.8
	32 × 32	768	12	12	4	87.5
Small	**16 × 16**	**384**	**12**	**6**	**4**	**21.7**
	32 × 32	384	12	6	4	22.5
Tiny	16 × 16	192	12	3	4	5.5

interferograms to train the Small ViT model with an Average pooling layer and a patch size of 16 × 16. Due to limitations in the GPU memory, we were unable to use batch sizes larger than 64 interferograms When the batch size was increased beyond 8 interferograms, both the IoU and MAE decreased to below 50 % and then started to rise. Compared to a standard batch size of 32, a batch size of 8 improved the model performance by 6.1 %, 16.7 %, and 65.5 %, respectively in the IoU, MAE, and training time as summarized in Table 10.

4.6. *Impact of weighting factors on classification and localization*

To enable the MT-ViT model to both classify and locate volcanic eruption signals, a combined loss function that consisted of a weighted linear combination of the CE loss and MSE loss was used. Table 11 demonstrates that when the weighting factor for MSE loss was fixed at 1.0, increasing the weighting factor for CE loss had a positive impact on both classification and localization performance. Conversely, when the weighting factor was fixed at 1.0 for CE loss, increasing the weighting factor for the MSE loss had a negative impact on the performance. By assigning weighting factors of 20.0 and 1.0 for CE and MSE losses, respectively, the performance for OA, AUC, IoU, MAE, and training time can be improved by 0.5 %, 0.4 %, 0.7 %, 0.0 %, and −20 %, respectively, compared to using equal weights. Our experimental findings for the optimal factor differ from those suggested in Gaddes et al., (2021) with the CNN model, which indicated a weighting factor of 1000.0 and 1.0 for the classification and localization losses, respectively. The recommendation to use 1000 to 1 weighting factor improved classification accuracy but degraded localization accuracy compared to using equal weights.

However, the MT-ViT was able to classify the deformation signal presented in the interferograms, the average IoU still needs to be improved. If the location of the deformation is the focus for the MT-ViT model, the average IoU can be further improved to about 63 %. Considering that the affected area by a volcano eruption may extend to tens of kilometers, a 0.9-kilometer positional error was taken as the acceptable accuracy, the proposed MT-ViT can serve as an automatic tool to locate the deformation.

4.7. *Visualization of attention maps*

We also validated the MT-ViT model using attention maps although using only the attention maps to explain a model behaviour is insufficient as the model is complex and has many layers (Chefer et al., 2021). The shape of the attention tensor is (number of heads, number of batches). The nearest neighbour's interpolation method was upsampling the patch attention to the original size of the input interferogram. Fig. 16 shows the attention maps of three simulated interferograms while Fig. 17 shows the attention maps of three real interferograms. It was expected that high attention values were correlated to the deformation pixels but in contrast, lower attention values were presented in other areas. The MT-ViT returned correctly not only the deformation class but also the location of the deformation. Figs. 16 and 17 (a), (b), (c), and (d) show that the attention patterns of the MT-ViT were related to the spatial distribution of the deformation pixels.

Table 7
Performances of different ViT architectures models based on the VolcNet dataset.

Architecture	OA (%)	AUC (%)	IoU (%)	MAE (pixel)	Best epoch
Base	**98.9**	**99.7**	**54.6**	3.7	11
Small	**98.9**	99.4	53.8	**3.5**	**10**
Tiny	98.5	99.4	49.9	5.1	11

Table 8
Performance of different patch sizes of the Small ViT model.

Patch size	OA (%)	AUC (%)	IoU (%)	MAE (pixel)	Best Epoch
16 × 16	**98.9**	99.4	**53.8**	**3.5**	**10**
32 × 32	**98.9**	**99.7**	51.1	4.1	27

Table 9
Performance of different pooling layers used in the Small ViT model.

Pooling layer	Trainable parameters (k)	OA (%)	AUC (%)	IoU (%)	MAE (pixel)	Number of epochs
Average	**2.7**	98.9	99.4	**53.8**	**3.5**	10
Flatten	500	**99.1**	**99.8**	53.3	4.3	**4**
Separate	**2.7**	97.9	99.3	53.0	4.1	11

4.8. *Limitations and future work*

The ViT model is constrained to a 224x224 pixel resolution for the interferogram even when using global average pooling due to the fixed-size patch embedding. For greater flexibility, consider ConViT or ConvNeXt, which integrate convolutional processes into the attention mechanism. Future improvements could involve the use of complex simulations to incorporate various deformation sources. Detecting low-slip earthquakes may require a specialized architecture to distinguish true signals from noise. The YOLO architecture could offer a suitable solution for multi-source detection and addressing misalignment between the classification and localization heads.

5. Conclusions

Prompt and accurate detection and interpretation of ground deformations are crucial in geohazard investigations. ML models have been applied to InSAR deformation interpretation based on individual or time series of interferograms, the performance of the models is limited in general by the scarcity of training data. We propose in this paper, an advanced ML model that utilizes a vision transformer architecture to effectively detect, locate, and interpret ground deformation signals based on single InSAR interferograms. We validated the model with both simulated and real SAR datasets, focusing on volcanic and earthquake deformations. The results obtained from the experiments demonstrated

Table 10
Performance of different batch sizes in training the Small ViT model.

Batch Size	OA (%)	AUC (%)	IoU (%)	MAE (pixel)	Best Epoch
8	**98.9**	99.4	**53.8**	**3.5**	**10**
16	98.7	99.4	48.5	5.5	14
32	**98.9**	**99.6**	50.7	4.2	29
64	98.7	**99.6**	51.1	4.2	59

Table 11
Impacts of different weighting factors on classification and localization losses.

Weighting factors		OA	AUC	IoU	MAE	Best
Classification	Localization	(%)	(%)	(%)	(pixel)	Epoch
1000.0	1.0	99.0	99.9	53.3	4.17	11
20.0	1.0	**99.4**	**99.8**	54.1	**3.5**	12
10.0	1.0	98.9	99.7	**54.2**	3.6	10
5.0	1.0	98.7	99.4	51.7	4.4	**9**
1.0	1.0	98.9	99.4	53.8	**3.5**	10
1.0	5.0	98.7	99.6	52.0	4.3	18
1.0	10.0	98.4	99.5	53.4	4.0	16
1.0	20.0	97.9	99.4	53.8	4.0	14

the effectiveness of the proposed ML model in deformation detection and interpretation. The following are the main conclusions of the study,

- The integration of multitask learning technique with vision transformer model (MT-ViT) was proposed to automatically detect, locate, and interpret deformation using single SAR interferograms. By considering the global features in the ViT model, the OA of deformation detection was improved from 95.5 % to 99.4 % compared to CNN networks that only focused on latent features.
- The experimental results demonstrated that the proposed MT-ViT model was able to detect deformation as small as 5 cm with an accuracy of above 99.0 %. The localization accuracy of the coseismic deformation was highly dependent on the spatial resolution of the SAR interferograms.
- The experimental results have shown that the Small architecture with 16 × 16 patch sizes, Average pooling layer, batch size of 8 interferograms, and 20:1 wt ratio between the classification loss and localization loss was the best option. The findings also indicated that increasing the trainable parameters was not always beneficial.

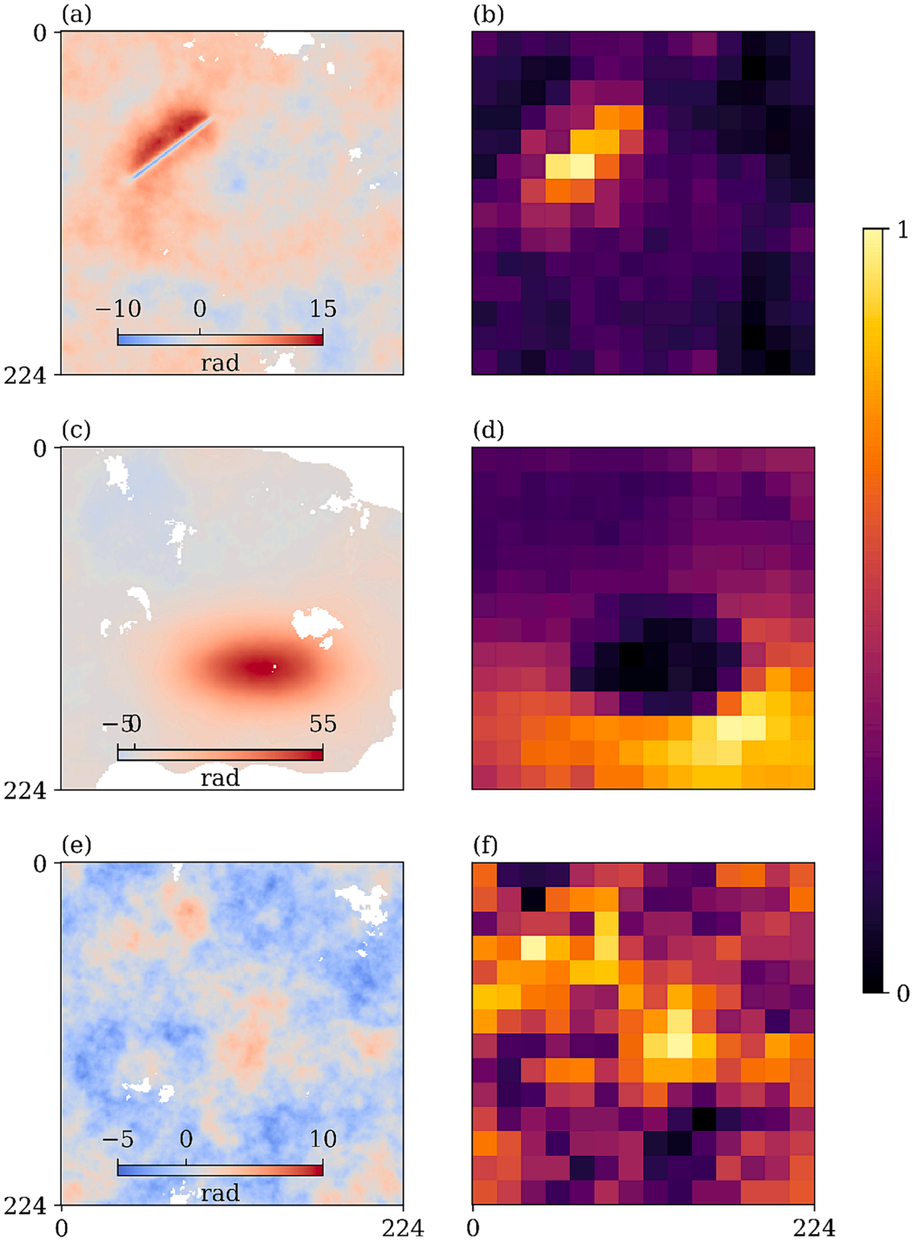


Fig. 16. Visualization of attention maps of three simulated interferograms. (a) Interferogram with volcanic dyke. (c) Interferogram with sill/point. (e) Interferogram with atmospheric signal. (b), (d), and (f) Averages of the final multi-headed attention blocks over the interferograms (a), (c), and (e), respectively.

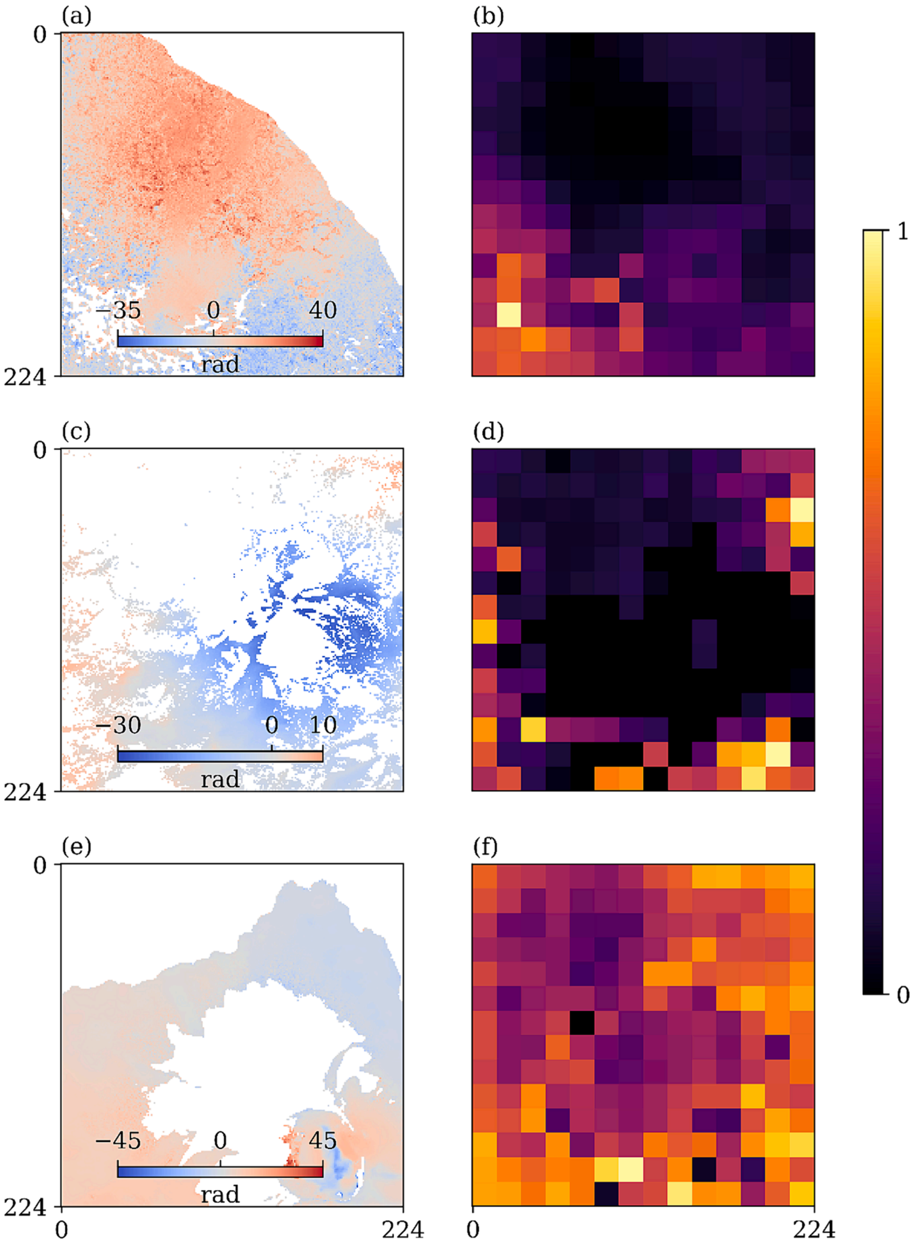


Fig. 17. Visualization of attention maps of three real interferograms. (a) Interferogram with volcanic dyke. (c) Interferogram with sill/point. (e) Interferogram with atmospheric signal. (b), (d), and (f) Averages of the final multi-headed attention blocks over the interferograms (a), (c), and (e), respectively.

CRediT authorship contribution statement

Mahmoud Abdallah: Writing – original draft, Visualization, Methodology, Formal analysis. **Samaa Younis:** Writing – original draft, Visualization, Methodology, Formal analysis. **Songbo Wu:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Xiaoli Ding:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank Mathew Gaddes for making the VolcNet dataset available at https://github.com/matthew-gaddes/VolcNet. All the experiments were carried out using the high-performance computing service (HPC) provided by the University Research Facility in Big Data Analytics (UBDA). The research was jointly supported by the Research Grants Council (RGC) of the Hong Kong Special Administrative Region (PolyU 152233/19E, PolyU 152318/22E, PolyU 152344/23E), the National Natural Science Foundation of China (Grants 42330717), and the University Grants Council of the Hong Kong Polytechnic University (1-BD2S).

References

Anantrasirichai, N., Biggs, J., Albino, F., Hill, P., Bull, D., 2018. Application of machine learning to classification of volcanic deformation in routinely generated InSAR data. J. Geophys. Res. Solid Earth 123, 6592–6606. https://doi.org/10.1029/2018JB015911.

Anantrasirichai, N., Biggs, J., Albino, F., Bull, D., 2019a. A deep learning approach to detecting volcano deformation from satellite imagery using synthetic datasets. Remote Sens. Environ. 230, 111179 https://doi.org/10.1016/j.rse.2019.04.032.

Anantrasirichai, N., Biggs, J., Albino, F., Bull, D., 2019b. The application of convolutional neural networks to detect slow, sustained deformation in InSAR time series. Geophys. Res. Lett. 46, 11850–11858. https://doi.org/10.1029/2019GL084993.

Ansari, H., Zan, F.D., Bameler, R., 2017. Sequential estimator: toward efficient InSAR time series analysis. IEEE Trans. Geosci. Remote Sens. 55, 5637–5652. https://doi.org/10.1109/TGRS.2017.2711037.

Bountos, N.I., Papoutsis, I., Michail, D., Anantrasirichai, N., 2022. Self-supervised contrastive learning for volcanic unrest detection. IEEE Geosci. Remote Sens. Lett. 19 https://doi.org/10.1109/LGRS.2021.3104506.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30, 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

Brengman, C.M.J., Barnhart, W.D., 2021. Identification of surface deformation in InSAR using machine learning. geochemistry. Geophys. Geosyst. 22, 1–15. https://doi.org/10.1029/2020GC009204.

Chattopadhyay, Soumitri, Ganguly, S., Chaudhury, S., Nag, S., Chattopadhyay, Samiran, 2023. Exploring Self-Supervised Representation Learning for Low-Resource Medical Image Analysis, pp. 1440–1444. <https://doi.org/10.1109/icip49359.2023.10222058>.

Chefer, H., Gur, S., Wolf, L., 2021. Transformer Interpretability Beyond Attention Visualization. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. pp. 782–791. <https://doi.org/10.1109/CVPR46437.2021.00084>.

Cicerone, R.D., Ebel, J.E., Britton, J., 2009. A systematic compilation of earthquake precursors. Tectonophysics 476, 371–396. https://doi.org/10.1016/j.tecto.2009.06.008.

Costantini, M., 1998. A novel phase unwrapping method based on network programming. IEEE Trans. Geosci. Remote Sens. 36, 813–821. https://doi.org/10.1109/36.673674.

d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L., 2022. ConViT: improving vision transformers with soft convolutional inductive biases. J. Stat. Mech. Theory Exp. https://doi.org/10.1088/1742-5468/ac9830.

De Novellis, V., Castaldo, R., De Luca, C., Pepe, S., Zinno, I., Casu, F., Lanari, R., Solaro, G., 2017. Source modelling of the 2015 Wolf volcano (Galápagos) eruption inferred from Sentinel 1-A DInSAR deformation maps and pre-eruptive ENVISAT time series. J. Volcanol. Geotherm. Res. 344, 246–256. https://doi.org/10.1016/j.jvolgeores.2017.05.013.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929.

Ebmeier, S.K., 2016. Application of independent component analysis to multitemporal InSAR data with volcanic case studies. J. Geophys. Res. Solid Earth 121, 8970–8986. https://doi.org/10.1002/2016JB013765.

Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., Alsdorf, D., 2007. The shuttle radar topography mission. Rev. Geophys. 45, 65–77. https://doi.org/10.1029/2005RG000183.

Gaddes, M., Hooper, A., Bagnardi, M., Inman, H., Albino, F., 2018. Blind signal separation methods for InSAR: the potential to automatically detect and monitor signals of volcanic deformation. J. Geophys. Res. Solid Earth 123, 10226–10251. https://doi.org/10.1029/2018JB016210.

Gaddes, M., Hooper, A., Bagnardi, M., 2019. Using machine learning to automatically detect volcanic unrest in a time series of interferograms. J. Geophys. Res. Solid Earth 124, 12304–12322. https://doi.org/10.1029/2019JB017519.

Gaddes, M., Hooper, A., Albino, F., 2021. Simultaneous classification and location of volcanic deformation in SAR interferograms using deep learning and the VolcNet database. JGR-Solid Earth Simult. https://doi.org/10.31223/X5CW2J.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem, 770–778. https://doi.org/10.1109/CVPR.2016.90.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 2017-Janua, 2261–2269. https://doi.org/10.1109/CVPR.2017.243.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2022-June, pp. 11966–11976. <https://doi.org/10.1109/CVPR52688.2022.01167>.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proc. IEEE Int. Conf. Comput. vis. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.

Lohman, R.B., Simons, M., 2005. Some thoughts on the use of InSAR data to constrain models of surface deformation: noise structure and data downsampling. Geochem. Geophys. Geosyst. 6 https://doi.org/10.1029/2004GC000841.

Loughlin, S., Sparks, S., Brown, S., Jenkins, S., Vye-Brown, C., 2015. Global Volcanic Hazards and Risk. Cambridge University Press.

Ma, P., Zhang, F., Lin, H., 2020. Prediction of InSAR time-series deformation using deep convolutional neural networks. Remote Sens. Lett. 11, 137–145. https://doi.org/10.1080/2150704X.2019.1692390.

Okada, Y., 1985. Surface deformation due to shear and tensile faults in a half-space. Int. J. Rock Mech. Min. Sci. Geomech. Abstr. 23, 128. https://doi.org/10.1016/0148-9062(86)90674-1.

Roseu, P.A., Gurrola, E., Sacco, G.F., Zebker, H., Dra. An fauzia rozani, 2012. The InSAR scientific computing environment. EUSAR 2012; 9th Eur. Conf. Synth. Aperture Radar 2012-April, 730–733.

Rouet-Leduc, B., Jolivet, R., Dalaison, M., Johnson, P.A., Hulbert, C., 2021. Autonomous extraction of millimeter-scale deformation in InSAR time series using deep learning. Nat. Commun. 12, 1–11. https://doi.org/10.1038/s41467-021-26254-3.

Ruder, S., 2017. An Overview of Multi-Task Learning in Deep Neural Networks. https://doi.org/10.48550/arXiv.1706.05098.

Silva, B., Sousa, J.J., Lazecky, M., Cunha, A., 2021. Deformation fringes detection in SAR interferograms using deep learning. Proc. Comput. Sci. 196, 151–158. https://doi.org/10.1016/j.procs.2021.11.084.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. 1–14. https://doi.org/10.48550/arXiv.1409.1556.

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L., 2021. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. https://doi.org/10.48550/arXiv.2106.10270.

Sun, J., Wauthier, C., Stephens, K., Gervais, M., Cervone, G., La Femina, P., Higgins, M., 2020. Automatic detection of volcanic surface deformation using deep learning. J. Geophys. Res. Solid Earth 125, 1–17. https://doi.org/10.1029/2020JB019840.

Sun, Q., Zhang, L., Ding, X., Hu, J., Liang, H., 2015. Investigation of slow-moving landslides from ALOS/PALSAR images with TCPInSAR: A case study of Oso, USA. Remote Sens. 7, 72–88. https://doi.org/10.3390/rs70100072.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. Proc. IEEE Comput. Soc. Conf. Comput. vis. Pattern Recognit. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. Proc. Mach. Learn. Res. 139, 10347–10357.

USGS, 2022. Earthquak Hazards Program [WWW Document]. <https://earthquake.usgs.gov/> (Accessed 8.1.22).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 2017-Decem, 5999–6009.

Wang, D., Zhang, Q., Xu, Y., Zhang, J., Du, B., Tao, D., Zhang, L., 2022. Advancing plain vision transformer towards remote sensing foundation model. IEEE Trans. Geosci. Remote Sens. 14, 1–15. https://doi.org/10.1109/TGRS.2022.3222818.

Wang, B., Zhao, C., Zhang, Q., Lu, Z., Li, Z., Liu, Y., 2020. Sequential estimation of dynamic deformation parameters for SBAS-InSAR. IEEE Geosci. Remote Sens. Lett. 17, 1017–1021. https://doi.org/10.1109/LGRS.2019.2938330.

Wightman, R., 2013. Pytorch image models (timm): Vit training details. [WWW Document]. <https://github.com/huggingface/pytorch-image-models>.

Wu, S., Yang, Z., Ding, X., Zhang, B., Zhang, L., Lu, Z., 2020. Two decades of settlement of Hong Kong International Airport measured with multi-temporal InSAR. Remote Sens. Environ. 248, 111976 https://doi.org/10.1016/j.rse.2020.111976.

Zhang, Y., Yang, Q., 2018. An overview of multi-task learning. Natl. Sci. Rev. 5, 30–43. https://doi.org/10.1093/nsr/nwx105.

Zhao, X., Wang, C., Zhang, H., Tang, Y., Zhang, B., Li, L., 2021. Inversion of seismic source parameters from satellite InSAR data based on deep learning. Tectonophysics 821, 229140. https://doi.org/10.1016/j.tecto.2021.229140.