Multi-Modal Data Fusion for Land-Subsidence Image Improvement in PSInSAR Analysis

KODAI SHIMOSATO AND NORIMICHI UKITA, (Member, IEEE)
Graduate School of Engineering, Toyota Technological Institute, Nagoya 468-8511, Japan
Corresponding author: Norimichi Ukita (ukita@toyota-ti.ac.jp)

This work was partly supported by JSPS KAKENHI Grant Number 19K12129.

ABSTRACT
There are three popular methods to understand the land subsidence: leveling, Global Navigation Satellite System, and Interferometric Synthetic Aperture Radar (InSAR) analysis using SAR images. While both leveling and the Global Navigation Satellite System can measure the amount of land subsidence only at specific points, InSAR analysis can observe a wide area in short time intervals. In terms of accuracy, however, InSAR analysis is inferior to leveling; centimeter/millimeter order (InSAR/PSInSAR analysis) vs. millimeter order (leveling). Among all observation errors in InSAR analysis, a tropospheric delay error has a large adverse effect on the measurement. It is difficult to suppress this tropospheric delay error by conventional methods because they try to remove error at each pixel independently in an InSAR image. However, geometrically-neighboring regions/pixels should be naturally correlated. Our proposed method employs such a neighboring relationship in a convolutional neural network (CNN). Our CNN is designed to improve InSAR analysis by mutually incorporating the InSAR image and the tropospheric delay error, which are estimated by any conventional methods. Experimental results demonstrate that our proposed method can reduce the mean error compared with a conventional method: from 10.3mm to 6.80mm.

INDEX TERMS
PSInSAR analysis, tropospheric delay, deep neural networks, multi-modal data fusion.

I. INTRODUCTION
The ground always shifts vertically and horizontally. One of the critical shifts is a land subsidence. Once the land subsidence occurs, the ground hardly returns to its original state. It is essential to understand the land subsidence and to take a countermeasure against it as early as possible, because it may cause huge negative impacts on our lives such as building collapse and lifeline damages.

We have three popular methods to measure the land subsidence: leveling, Global Navigation Satellite Systems (GNSSs), and Interferometric Synthetic Aperture Radar (InSAR) analysis. The functional properties of these three methods are summarized in Table 1. For leveling, a huge human effort is required in order to measure a wide area. The GNSSs including Global Positioning System (GPS), Global Navigation Satellite System (GLONASS), Galileo, and Quasi-Zenith Satellite System (QZSS) measure the crustal alteration so that ground-installation devices receive signals transmitted from artificial satellites. SAR is a form of high spatial-resolution micro-wave radar. SAR is transmitted from a satellite towards the ground surface. Its reflection intensity and phase are captured by the same satellite. From the phase shift between those captured at different capturing times, we can compute the ground subsidence. As summarized in Table 1, while both the leveling and GNSS (1) require special measurement devices and (2) measure the amount of the land subsidence only at specific points, InSAR analysis can observe a wide area in short time intervals with no ground-installation device. These advantages motivate us to employ InSAR analysis for various real-world applications (e.g., [1]–[3]). Although these advantages enrich the value of InSAR analysis, its accuracy is inferior to the leveling process.\

This paper proposes convolutional neural networks (CNNs) for improving InSAR analysis. Machine learning such as CNNs allows us to stochastically rectify conventional InSAR analysis, based on a number of training data. In this work, the training data is a set of InSAR analysis results with larger error and their ground-truth data measured by leveling with

1While the quality of leveling measurements changes among different leveling datasets, high-quality leveling measurements (i.e., 0.1 millimeter-order resolution [4]) were used in our experiments as described in Section IV.
less error. In accordance with a standard supervised machine learning manner, a CNN is trained so that they can rectify the result of InSAR analysis so that it is equal to the ground-truth data measured by leveling. Once the CNN is trained properly, it can rectify any new InSAR analysis results, which are not included in the training data. In addition, we focus on reducing the error associated with a tropospheric delay, which is considered to be adverse on the measurement by InSAR analysis.

The contributions of this paper are as follows:

- Our proposed method achieves millimeter-order accuracy in land-subsidence measurements by the CNN that fuses multi-modal data (i.e., land-subsidence and tropospheric-delay images) obtained by conventional InSAR analysis.
- We explore several types of multi-modal fusion approaches for our proposed CNN in order to evaluate the appropriateness of each approach.

![Optical RGB image](Google) ![SAR image](ESA)

**FIGURE 1.** Optical and SAR images captured by satellite-mounted cameras. Each pixel value in this SAR image shows the reflection intensity of SAR. The reflection intensity is, in general, higher in metals and minerals such as buildings and rocks. For example, pixel values in a part of an island shown in the lower left region in this SAR image are higher, because this island is an airport with landing fields.

II. RELATED WORK

A. InSAR ANALYSIS AND ITS EXTENSIONS

Different from an optical image (e.g., left image in Fig. 1), a SAR image (e.g., right image in Fig. 1) consists of the intensity and phase of radar signals. Assume that an orbiting satellite captures SAR images of the same ground surface at different times, as shown in Fig. 2. The distance between the satellite and the ground surface is expressed by the following equations [5]–[7] where \( r_1 \) and \( r_2 \) are the distances in the first and second capturing times:

\[
2r_1 = n_1\lambda + \Delta x_1, \quad (1)
\]

\[
2r_2 = n_2\lambda + \Delta x_2, \quad (2)
\]

where \( n_i, \lambda, \) and \( \Delta x_i \) denote the number of waves, the wave length, and the fraction (\( 0 \leq \Delta x_i < \lambda \)), respectively, \( i \) denotes the capturing time (i.e., \( i \in \{1, 2\} \)). If two SAR images are captured at the completely-same location, the difference between \( r_1 \) and \( r_2 \) excepting their first terms (i.e., the integer multiple of the wave length), which is expressed by the following equation, is determined only by the land subsidence:

\[
2 (r_2 - r_1) = \Delta x_2 - \Delta x_1 = \frac{\lambda (\phi_1 - \phi_2)}{2\pi}, \quad (3)
\]

where \( \phi_i \) denotes the phase of a radar signal. In the phase shift included in Eq. (3) (i.e., \( \phi_1 - \phi_2 \)), unknown scaling factors of the terrain are vanished. As a result, \( \phi_1 - \phi_2 \) represents the distance difference in the order of the radar wave length. Since Eq. (3) is computed without the integer multiple of the wave length, the phase unwrapping algorithm [8]–[10] is achieved for obtaining the distance difference, \( r_2 - r_1 \), between the two capturing times. Eq. (3) is computed in all pixels of the paired SAR images. The phase shift values compose an initial InSAR image.

This initial InSAR image still contains several types of erroneous components. Each pixel value, \( \phi = \phi_1 - \phi_2 \), in the initial InSAR image is called a residual phase and is expressed as a sum of five components, including an ideal phase shift, \( \phi_{def} \), and four erroneous components, as follows [11]:

\[
\phi = \phi_{def} + \phi_{atm} + \phi_{orb} + \phi_{DEM} + \phi_n \quad (4)
\]
• $\phi_{def}$ denotes an ideal phase shift caused only by the ground subsidence. Estimating $\phi_{def}$ is the goal of InSAR analysis.

• $\phi_{atm}$ is the estimation error caused by the atmospheric phase. $\phi_{atm}$ includes the estimation error of tropospheric delay, which is called a tropospheric delay error.

• In reality, two SAR images are captured at slightly-different locations, as shown in Fig. 2. This difference in the capturing locations produces $\phi_{orb}$. Since an InSAR image inevitably contains stripe noise, the Digital Elevation Model (DEM) is used for removing this stripe noise. This process produces $\phi_{DEM}$.

• $\phi_p$ denotes noise.

With the initial InSAR image, these four erroneous components are estimated and removed in the further InSAR analysis. While readers can refer to [12], for example, for more details, removing these erroneous components is disturbed by noisy values in SAR images. Noise in the SAR images strongly depends on ground objects that reflect radar signals. That is, a noise level of multiple SAR image captures is roughly determined in each pixel, if the SAR images are spatially aligned so that each pixel captures the same ground surface.

For removing the above erroneous components robust to noise, only pixels with the stable amplitudes of reflected radar signals are used in Persistent Scatterer InSAR (PSInSAR) [5], [13]. Given a sufficient number of spatially-aligned SAR images, a metric of this stability is provided by the standard deviation, $\sigma_p$, of the amplitudes in $p$-th pixel. In addition to $\sigma_p$, the mean $\mu_p$ of the amplitudes can also be used for this evaluation metric. This is because it is known that, with microwave radar, more stable and higher reflections are reflected from metals and minerals such as buildings and rocks [13]. With $\sigma_p$ and $\mu_p$, $p$-th pixel is selected by PSInSAR analysis if $\frac{\sigma_p}{\mu_p}$ is lower than a threshold in (e.g., typically $\frac{\sigma_p}{\mu_p} < 0.25$) [13]. In each selected pixel, $\mu_p$ is considered to be a pixel value in the PSInSAR image.

While PSInSAR can achieve millimeter-order accuracy in exceptionally-good conditions [14], the unknown and complex atmospheric density (in particular, water vapor pressure in the air) makes $\phi_{atm}$ larger. Other additional data (e.g., numerical meteorological models [15]) and the relationship between the phase and topography can reduce $\phi_{atm}$. Machine-learning methods [16] are also effective for reducing $\phi_{atm}$. However, tropospheric delay correction is still a difficult problem.

### B. InSAR Analysis Using Machine Learning

In [17]–[20], local areas in InSAR images are classified for detecting important time-series changes such as volcano deformations and ground deformations. Low-level processing with InSAR images such as phase filtering [21], phase unwrapping [22], and denoising [23] can be also improved by CNNs. In order to utilize time-series InSAR images, recurrent networks are also useful [24].

For fully utilize the performance of machine learning-based methods, a small amount of training data in InSAR analysis is often a bottleneck. This problem is alleviated by self-supervised learning [25] and synthetic data [18], [26].

In contrast to the aforementioned previous methods, our proposed method fuses a land-subsidence image, which is acquired by PSInSAR analysis, with a tropospheric-delay error map for further improving the accuracy of the land-subsidence image. How to fuse the land-subsidence image with the tropospheric-delay error map via CNNs is a technical contribution of our paper.

#### C. Multi-Modal Image Fusion Using CNNs

Multi-modal data fusion is widely used in various computer vision problems. Recently, as well as most computer vision methodologies, multi-modal data fusion is achieved by CNNs. CNNs allow us to fuse complex multi-modal data. Such a CNN is appropriate for fusing the land-subsidence image with the tropospheric-delay error map that is affected by complex atmospheric factors.

Typical network architectures proposed for multi-modal data fusion are shown in Fig. 3. In CNNs shown in (a), multi-modal data are first fused and then fed into encoder-decoder networks consisting of convolution layers. In (b), encoders extract features before multi-modal fusion. A network shown in (c) employs an additional path (colored by magenta in the figure) connecting multi-modal paths for information exchange between these multi-modal paths even before fusion. In (d), since one input modality is equal to the output modality, the path corresponding to this modality is regarded as dominant. Towards this dominant path, features extracted from other modalities are fused. In the example shown in (d), the first input modality colored by green is equal to the output modality.

![Different network architectures for multi-modal data fusion.](image)
These InSAR images are used for PSInSAR analysis using InSAR images between one SAR image and all other images. SNAP [42], [43] to get initial InSAR images. With SNAP, SAR images (i.e., tropospheric delay images in our work. First of all, multiple operations indicated by “F” in the figure are employed. For example, multi-modal data are concatenated in [27], while the element-wise add operation is used in [38].

As described at the end of Section I, our proposed method fuses multi-modal data (i.e., land-subsidence and tropospheric-delay images obtained by conventional InSAR analysis) for millimeter-order accuracy.

III. PROPOSED METHOD

Our strategy for improving an InSAR image is to reduce the tropospheric delay by a CNN with a tropospheric-delay image. We expect that the distribution of the tropospheric delay values provides us critical information about where and how tropospheric delay error is large. Section III-A introduces how we obtain a land-subsidence image and its corresponding tropospheric-delay image from a pair of SAR images captured by a satellite. With these land-subsidence and tropospheric-delay images, our proposed network for tropospheric delay correction is proposed in Section III-B.

FIGURE 4. Land subsidence image generation from N SAR images. After SNAP achieves InSAR analysis, its output images (i.e., N − 1 InSAR images) are fed into StaMPS/MTI [39], [40] for PSInSAR analysis. Its output is one PSInSAR image. The snap2stamps software is used for converting the SNAP format to the StaMPS/MTI format. For further improvement, precise tropospheric delays are estimated by TRAIN [41]. The estimated tropospheric delays are used for correcting the PSInSAR image. Our proposed method requires (a) the corrected PSInSAR image and (b) the estimated tropospheric delay image.

A. PSInSAR FOR LAND-SUBSIDENCE AND TROPOSPHERIC-DELAY IMAGES

Figure 4 shows the process flow for generating InSAR and tropospheric-delay images in our work. First of all, multiple SAR images (i.e., N SAR images in the figure) are fed into SNAP [42], [43] to get initial InSAR images. With SNAP, InSAR images between one SAR image and all other images are computed. As a result, N − 1 InSAR images are obtained. These InSAR images are used for PSInSAR analysis using Stanford Method for Persistent Scatterers / Multi-Temporal InSAR (StaMPS/MTI) [39], [40], [44] in order to acquire one PSInSAR image. Inconsistency between data formats of SNAP and StaMPS/MTI is resolved by the snap2stamps software [12], [45]. Long-range phase-shift computation is done by Snapfu [10], [46] in StaMPS/MTI. Further improvement is achieved by tropospheric-delay correction by TRAIN [41], [47]. In TRAIN, a tropospheric delay is corrected under the assumption that a tropospheric delay error depends on the altitude at each point. With the tropospheric-delay image estimated by TRAIN, the PSInSAR image is improved.

Finally, this improved PSInSAR image is transferred to its land-subsidence image. The phase shift included in the PSInSAR image is captured along the radar direction, while the land subsidence is measured along the vertical axis. Based on the angle between them, StaMPS/MTI computes the land-subsidence image. In our experiments, θ was given by the European Space Agency (ESA). These land-subsidence and tropospheric-delay images, each of which is expressed as a 8-bit gray scale image, are used in our proposed method, which is described in what follows.

B. TROPOSPHERIC-DELAY CORRECTION WITH FUSION OF LAND-SUBSIDENCE AND TROPOSPHERIC-DELAY IMAGES

The network architecture of our proposed method is shown in Fig. 5. Two input images (i.e., land-subsidence and tropospheric-delay images) are fed into the network that produces the improved land-subsidence image. As with general CNNs, our network employs a combination of convolution and pooling layers for representing global contexts of input images. We used max pooling layers. Since each pooling operation reduces the size of a feature map, it is enlarged to its original size by unpooling layers [48]. Batch normalization layers [49] allow us to make the training process efficient and stable. A Rectified Linear Unit (ReLU) is used for an activation function. At the last layer, 1 × 1 convolution reconstructs the output image, as done in many image reconstruction methods such as image style transfer [50] and image super resolution [51], [52].

The basic architecture is based on Fig. 3 (d) with a dominant path, because the dominant path is appropriate for our problem where one of the input modalities is equal to the output modality (i.e., land-subsidence image). As with FuseNet [38], features extracted from another modality (i.e., tropospheric-delay image) are fused with the dominant path by the element-wise add indicated by yellow layers in Fig. 5.

1) DENSE SKIP CONNECTIONS FOR FUSING GLOBAL AND LOCAL FEATURES

Different from FuseNet, on the other hand, our network is augmented with skip connections between a pair of encoding and decoding layers. Through each skip connection, features extracted from the encoder and decoder are concatenated in the concatenation layer colored by gray in Fig. 5. Such skip
connections are proposed for simultaneously understanding global and local features in UNet [53].

2) LOSS FUNCTIONS
For training our network, ground-truth subsidence values are required in order to compare them with the output land-subsidence image. These ground-truth values are measured by leveling. The measured points are spatially aligned with the land-subsidence images based on their latitudes and longitudes. These spatially-aligned outputs and ground-truth measurements are compared for training the network. This training is achieved with the following Mean Squared Error loss (indicated by “MSE loss (l)”) with the ground-truth values measured by leveling. The global similarity between input and output land-subsidence images is kept by “MSE loss (p)”:

\[
MSE_l = \frac{1}{N_l} \sum_{i=1}^{N_l} (q_{gt,l} - p_{out,l})^2, \tag{5}
\]

where \(q_{gt,l}\) and \(p_{out,l}\) denote “the measurement value of \(l\)-th ground-truth leveling point” and “the value of its spatially-aligned pixel (denoted by \(s(l)\)) in the output land-subsidence image,” respectively. The number of the points in each land-subsidence image is \(N_l\).

In addition to \(MSE_l\), the following Mean Squared Error loss (indicated by “MSE loss (p)”) in Fig. 5) is also used for making the output image similar to its input image:

\[
MSE_p = \frac{1}{N_p} \sum_{i=1}^{N_p} (p_{in,i} - p_{out,i})^2, \tag{6}
\]

where \(p_{in,i}\) and \(p_{out,i}\) denote pixel values of \(i\)-th pixel in the input and output land-subsidence images, respectively. \(N_p\) denotes the number of pixels in each land-subsidence image.

Note that the above loss functions (5) and (6) impose two different conditions on a pixel corresponding to each leveling point. On such pixels, the loss function (5) should be more important in order to learn the values measured by leveling. In other words, the loss function (6) should be neglected at each leveling point. Therefore, pixels around the leveling points in \(MSE_p\) are low-weighted by \(W(i)\) as follows:

\[
MSE'_p = \frac{1}{N_p} \sum_{i=1}^{N_p} W(i) (p_{in,i} - p_{out,i})^2, \tag{7}
\]

where \(W(i)\) has lower values around the leveling points. \(W(i)\) is expressed by the 2D Gaussian mixture as shown in Fig. 6 where the dimension of the yellow rectangle is equal to the image size. Initially, all weights in \(W(i)\) are 1, and then the 2D Gaussian mixture is subtracted from this initial \(W(i)\). Each 2D isotropic Gaussian distribution is centered at one of the leveling points and is normalized so that the maximum value is 1. Therefore, \(0 \leq W(i) \leq 1\) where \(W(i) = 0\) at each leveling point.

With the above two loss functions (5) and (7), a loss for training our proposed network is defined as follows:

\[
L = MSE_l + \alpha MSE'_p, \tag{8}
\]

where \(\alpha\) denotes a weight constant.
3) RANDOM PATCH CROPPING FOR A SMALL NUMBER OF GROUND-TRUTH DATA

Unlike standard image recognition problems, a SAR image is huge. For example, 640 × 480 pixels in many images of the COCO dataset [54] vs. around 20,000 × 10,000 pixels in SAR images. This huge image cannot be directly located in a GPU, which is required for practical deep learning. For resolving this problem, we crop the huge image to small patch images, each of which is independently fed into the network. In our experiments, the size of each patch image is 1024 × 1024 pixels.\(^2\) This size is the maximum size that can be loaded to a 24GB NVIDIA Titan RTX GPU for an effective mini-batch size. We empirically found that a smaller patch image (e.g., 512 × 512 pixels) makes the training difficult to converge successfully. This might be because the smaller patch cannot cover a meaningful variation of altitudes observed in the tropospheric-delay image, while the altitudes give a large impact on tropospheric-delay error.

In the inference process, each image is divided into non-overlapping patches with no margin, and then all of the patches are independently fed into the network. By tiling the outputs estimated from these input patches, the rectified land-subsidence image is acquired. Since the locations of these patches are determined independently of those of leveling points in the inference process, the leveling points are located randomly in each patch. Therefore, leveling points should be randomly located in each patch also in the training process. To this end, in the training process, each patch is cropped randomly so that it contains at least one leveling point.

While the aforementioned random cropping process extracts only one patch around each leveling point in the training process, the performance of general machine learning including CNNs can be improved by increasing training data (i.e., cropped patches in our case). Since the amount of SAR images and leveling data is limited,\(^3\) the aforementioned one-patch cropping process might result in insufficient training data. To fully utilize a limited number of available training data, data augmentation [55] is widely employed. The data augmentation process increases the training data so that the original given training data is slightly changed as long as the image context is unchanged. For example, images are cropped, flipped, rotated, blurred, and degenerated with noise. Based on the above discussion, not only one but also several patches are randomly located and cropped from each leveling point, as illustrated in Fig. 7.

For the training process with the patches, our loss function (8) is modified so that this loss is computed in each patch.

\(^2\)1024 × 1024 pixels captures around 10 × 10 km in the land.

\(^3\)See Section IV-A for details.

IV. EXPERIMENTAL RESULTS

A. DATASET

1) LEVELING MEASUREMENT DATA

Official land subsidence data collected by Ministry of Environment, Japan [67], is used for our experiments [56]–[66]. This dataset covers 33 regions in Japan. While other common measurement data that are publicly available are not precise sufficiently compared with InSAR/PSInSAR analysis as shown in Table 1, the leveling process employed in this data collection provides 0.1 millimeter-order resolution [4]. Therefore, this dataset of the leveling points can be regarded as ground-truth data for our supervised learning strategy.

Table 2 shows the statistics of leveling points used in our experiments. The land-subsidence values of the leveling points between two continuous years (e.g., between 2010 and 2011) are compared with the subsidence maps of InSAR analysis between the same years. While this dataset contains 1,025 leveling points, each of which was measured on around
TABLE 2. Statistics of leveling points used in our experiments. The first, second, third, fourth, and fifth columns show (1) where the points are measured, (2) which year the measurement began, (3) which year the measurement ended, (4) the measurement period of the points, and (5) the number of the points measured in each area. The points in Ibaraki prefecture are used as test data, while all other points are used in the training process.

| Prefecture/City | Begin | End | Years | # of points |
|----------------|-------|-----|-------|-------------|
| Sapporo [56]   | 2014  | 2017| 4     | 12          |
| Yamagata [57]  | 2014  | 2017| 4     | 6           |
| Ibaraki [55]   | 2015  | 2017| 3     | 142         |
| Satsuma [59]   | 2014  | 2018| 5     | 98          |
| Chiba [60]     | 2015  | 2018| 4     | 214         |
| Tokyo [61]     | 2014  | 2018| 5     | 84          |
| Niigata [62]   | 2015  | 2018| 4     | 112         |
| Ishikawa [63]  | 2014  | 2017| 4     | 23          |
| Yamanashi [64] | 2014  | 2018| 5     | 13          |
| Gifu [65]      | 2014  | 2018| 5     | 48          |
| Aichi [66]     | 2014  | 2018| 5     | 203         |
| Mie [65]       | 2014  | 2018| 5     | 32          |
| Saga [66]      | 2014  | 2017| 4     | 38          |
| Total          | –     | –   | –     | 1,023       |

January 1st, some of them were not measured multiple times. For InSAR analysis, at least two SAR images are required in the same region. Therefore, among all leveling points in the above data, we used only points that were observed twice or more. In these points, 2,339 land-subsidence values were collected in total.

For fair comparison where training and test data have no overlap, measurements in Ibaraki prefecture are used as test data. All other measurements are used in the training process. In total, 2,339 land-subsidence values are split to 2,055 training data and 284 test data; 284 is the number of the land-subsidence values obtained from three-year leveling measurements in Ibaraki.

As described in Section III-B, a SAR image is cropped to small patch images around each leveling point. The number of the leveling points included in each patch image is 4.49 on average. The maximum, minimum, and variance are 22, 1, and 16.23, respectively.

2) SAR IMAGES

We use charge-free SAR images collected by ESA Sentinel–1 [68]. Sentinel–1 observes the leveling points included in our dataset introduced in Section IV-A1 with a period of 12 days. Since all the leveling points were observed on around January 1st, we use SAR images that were captured between December 25th and January 5th for making a pair of the SAR image and leveling measurements included in this SAR image. We assume that the time difference between the leveling and SAR measurements (i.e., ±5 days at most) is sufficiently short in contrast to a year (i.e., time interval between two leveling measurements) so that the SAR measurements are almost synchronized with the leveling measurement. Note that, while a SAR image is asynchronous with the leveling points, the land-subsidence image and the tropospheric-delay error map (i.e., two inputs of the proposed network) are synchronized because they are estimated from the SAR image. This is important because the tropospheric-delay error map is affected by the weather changing from moment to moment; if these two inputs are asynchronous, the changing weather makes it difficult to fuse the two inputs. On the other hand, a short time-difference between the two inputs and the leveling points is acceptable under the assumption that the land subsidence is not caused in such a short time (i.e., ±5 days at most). In total, 346 SAR images are collected for our dataset. In this SAR image collection, 10 SAR images are collected on average for each region. These 10 SAR images allow us to achieve PSInSAR for estimating InSAR images robust to noise.

As described in Section III-A, SAR and InSAR images were generated by SNAP [42]. From nine SAR images, eight InSAR images are generated. While these SAR images consist of reflection intensity values, each pixel value in the InSAR images is a phase shift.

B. EVALUATION

All estimation results are evaluated by root mean square error (RMSE), which is expressed as follows:

\[
\text{RMSE} = \sqrt{\frac{1}{N_t} \sum_{l=1}^{N_t} (q_{gt,l} - p_{out,l})^2},
\]

(9)

where \(q_{gt,l}\) and \(p_{out,l}\) denote the ground-truth land-subsidence value of \(l\)-th leveling point and its spatially-aligned pixel (denoted by \(s(l)\)) in the output images, respectively. \(N_t\) is the number of leveling points in all images; in our experiments, \(N_t = 284\), which is the total number of the test data of the land subsidence values.

C. PARAMETERS

All parameters were determined empirically so that RMSE (9) is minimized with all training data.

\(\alpha = 1.5\) in the loss function (8). \(\sigma^2\) of the Gaussian distribution for defining \(W(i)\) in the loss function (7) is 200.

As mentioned before, we used StaMPS/MTI [39], [40] for PSInSAR analysis. Its parameters used in our experiments are shown in Table 6.

D. CNN TRAINING

Our CNN shown in Fig. 5 was trained by the Adam optimizer [69]. The learning rate and the size of the mini-batch were 1.0e-5 and 8, respectively. The training process was halted when the number of epochs reached 100.

E. RESULTS

RMSE computed from PSInSAR images obtained by the baseline method shown in Fig. 4 is 10.3.

In the first experiment, we evaluate the effectiveness of multi-modal data fusion. In order to validate the effectiveness of the tropospheric-delay image for multi-modal data fusion, the network path for the tropospheric-delay image is removed from our proposed network shown in Fig. 5. That is, the
TABLE 3. Effectiveness of multi-modal data fusion with the tropospheric-delay image (TD image). Estimated results were evaluated with Root Mean Square Error (RMSE) [mm]. This evaluation was done with the test data (i.e., the data of Ibaraki prefecture), which was excluded from the training data. Since this criterion is an error metric, a smaller value means a better performance. The best score is colored by red. The baseline result obtained by PSInSAR is shown in the leftmost row.

|                | PSInSAR (Baseline) | Unimodal CNN (Without TD image) | Our multi-modal fusion CNN (With TD image) |
|----------------|--------------------|---------------------------------|-------------------------------------------|
| RMSE           | 10.3               | 7.15                            | 6.80                                      |

TABLE 4. Performance comparison among different multi-modal fusion network architectures. (a) The original UNet [53], (b) Modified UNet that accepts two images, (c) ACNet [34], (d) Our proposed network with no skip connection. (a’), (b’), (c’), and (d’) are augmented from (a), (b), (c), and (d) by skip connections.

|                | No skip connection | With skip connections          |
|----------------|--------------------|--------------------------------|
|                | (a) UNet           | (a’) UNet                      |
|                | (b) UNet²          | (b’) UNet²                     |
|                | (c) ACNet          | (c’) ACNet                     |
|                | (d) Ours           | (d’) Ours                      |
| RMSE           | 7.97               | 6.81                           |
|                | 7.41               | 6.91                           |
|                | 7.89               | 7.05                           |
|                | 7.73               | 6.80                           |

TABLE 5. Ablation study (1). The effectiveness of the proposed randomly-shifted cropping is demonstrated.

|                | No random crop | Random crop |
|----------------|----------------|-------------|
| RMSE           | 7.26           | 6.80        |

TABLE 6. Parameters in our StaMPS process.

| Parameter       | Value  |
|-----------------|--------|
| max_topo_err [m]| 10     |
| filter_grid_size[m]| 50     |
| clap_win         | 32     |
| clap_alpha       | 1      |
| clap_beta        | 0.3    |
| density_rand     | 20     |
| unwrap_grid_size [m]| 100   |
| unwrap_gold_n_win| 32     |
| unwrap_time_win  | 180    |

TABLE 7. Hyperparameters used in each method.

|                  | (a) | (a’) | (b) | (b’) | (c) | (c’) | (d) | (d’) |
|------------------|-----|------|-----|------|-----|------|-----|------|
| Learning rate    | 1e-06 | 1e-06 | 1e-06 | 1e-06 |
| Batch size       | 8    | 8    | 8    | 8    |

For validating the superiority of our proposed network for the fusion of land-subsidence and tropospheric-delay images, several variations of multi-modality fusion networks are evaluated. Table 4 shows the results (i.e., RMSEs) of all the variations. In the table, (a), (b), (c), and (d) are the original UNet [53] that accepts one input image, the modified UNet that accepts two images (i.e., land-subsidence and tropospheric-delay images), ACNet [34], and our proposed network. Since the original UNet (a) accepts only one input image, the land-subsidence and tropospheric-delay images are concatenated channel-wise to be merged into one image. That is, given a $W \times H$ pixel land-subsidence image and a $W \times H$ pixel tropospheric-delay image, a $W \times H \times 2$ image is fed into UNet in (a). The detailed architecture of (b) is shown in Fig. 8. Note that all of (a), (b), (c), and (d) have no skip connections that fuse global and local features. On the other hand, (a’), (b’), (c’), and (d’) are augmented from (a), (b), (c), and (d) by skip connections. The hyperparameters used in each variant are shown in Table 7.

Compared with PSInSAR as the baseline, all deep network-based methods with multi-modal data fusion get better results. However, several fusion methods are inferior to the unimodal CNN, which is shown in Table 3. This might be because it is not easy to train CNNs with more parameters successfully.

Among the four variations of network architectures, our proposed networks (d) and (d’) outperform other networks both with and without skip connections. Furthermore, the skip connections can improve the performance, i.e., (d) 7.73 vs. (d’) 6.80. These results demonstrate that our proposed method (d’) is the best among all variations shown in Table 4.

In what follows, the results of our ablation studies are shown for demonstrating advantages in our algorithm and parameter choices. Table 5 shows the results obtained by our method with and without randomly-shifted patch cropping in the training process. We can see that the randomly-shifted patch cropping scheme can decrease RMSE.

TABLE 8. Ablation study (2). The effect of the receptive field is evaluated. While a wide variety of receptive fields are evaluated, its size has less impact on the performance.

| Receptive field | 68 | 140 | 284 | 572 |
|-----------------|----|-----|-----|-----|
| RMSE            | 7.01 | 6.80 | 7.18 | 7.32 |

In a CNN, the size of its receptive field is important in determining how much the CNN captures global context information. As shown in Table 8, a wide range of receptive...
Ablation study (3). The effectiveness of data augmentation using randomly-shifted patch cropping is validated.

| Input images | x1 | x2 | x3 | x4 | x5 | x6 |
|--------------|----|----|----|----|----|----|
| RMSE         | 7.34 | 7.17 | 7.17 | 7.06 | 6.80 | 7.18 |

Table 9 shows the effects of the data augmentation process described in Section III-B. We tested the augmentation processes varying the number of randomly-cropped patches around each leveling point (denoted by $N_a$ where $N_a \in \{1, 2, 3, 4, 5, 6\}$). While the best result is obtained with $N_a = 5$, the performance is decreased with $N_a = 6$. This performance drop might be caused because excessive data augmentation saturates useful information. Furthermore, increasing randomly-shifted patch cropping over-spreads the appearance variation of patch images. This makes it difficult to train the network properly.

While the lowest error is obtained with $N_a = 5$, all other experiments shown in this section (i.e., Tables 3, 4, 5, 8, and 10) are done with $N_a = 4$ for efficient training; more training data requires more training time. Note that, for fair comparison, data augmentation with $N_a = 4$ was executed for other methods as well as for our proposed method, which gets the best result.

Table 10 shows results acquired with and without residual learning [33], which is effective for deep networks. In our network shown in Fig. 5, a residual path begins right after the input land-subsidence image and ends right before the output land-subsidence image. The output of the last $1 \times 1$ convolution layer (indicated by the purple layer in the figure) is elementwise-added to the input land-subsidence image via the residual path in order to reconstruct the output land-subsidence image. While the performance is not gained by this residual learning, its performance drop is small. These results show that our network is (1) sufficiently deep for improving the land-subsidence image as shown in Table 3 but (2) sufficiently trained without overfitting because the network is not too deep (i.e., if the network is too deep, residual learning might be able to improve the performance).

For visual evaluation, input and output images of our proposed network are shown in Fig. 9. The input and output images are shown on the left and right sides in the same row. In the upper example (i.e., (a) success case), the global subsidence values are similar between the input and output images. This demonstrates that the MSE loss (7) works correctly. Based on the fact that (1) the input and output images are similar as shown in Fig. 9 (a) and (2) the subsidence values of the leveling points in the output are improved as shown in Tables 3, it is expected that the output image is estimated more accurately than the input image. In contradiction to the success case, in the lower example (i.e., (b) failure case), it can be seen that (1) the input and output images are dissimilar and (2) the ground-truth value overlaid at the center is dissimilar to its neighboring pixels in the land-subsidence image modified by our method. In this failure case, a large difference between the ground-truth leveling value and its neighboring values in the land-subsidence image estimated by PSInSAR. For resolving this problem, more additional cues might be required in addition to the tropospheric delay error map. Otherwise, the leveling value might be possibly
errorneous due to measurement error or some other reasons. In such a case, our proposed method requires an explicit mechanism for removing the errorneous training data.

V. CONCLUDING REMARKS

This paper proposes a method for improving the accuracy of a land-subsidence image acquired by PSInSAR analysis. Our method employs a PSInSAR scheme for obtaining the land-subsidence image as the baseline. This baseline is improved by being fused with the tropospheric-delay image through our proposed CNN. Our CNN achieves this multi-modal data fusion as well as global and local feature fusion.

Future work includes the usage of other modalities (e.g., atmospheric water vapor and numerical meteorological models) that are known to be essential modalities for PSInSAR analysis. While our proposed method is applicable to SAR data observed anywhere anywhere in theory, it is possible that region-and-time specific training is required for maintaining the performance independently of region and time.

REFERENCES

[1] P. Milillo, D. Perissin, J. T. Salzer, P. Lundgren, T. Lacava, G. Milillo, and C. Serio, “Monitoring dam structural health from space: Insights from novel InSAR techniques and multi-parametric modeling applied to the Pertusillo dam Basilicata, Italy,” Int. J. Appl. Earth Obs. Geoinf., vol. 52, pp. 221–229, Oct. 2016.

[2] X. Liu, Q. Chen, J. Zhao, Q. Xu, R. Luo, Y. Zhang, Y. Yang, and G. Liu, “The spatial response pattern of coseismic landslides induced by the 2008 Wenchuan earthquake to the surface deformation and Coulomb stress change revealed from InSAR observations,” Int. J. Appl. Earth Obs. Geoinf., vol. 87, May 2020, Art. no. 102030.

[3] Z. Yao, X. Hu, X. Liu, W. Xie, Y. Dong, H. Qiu, Z. Chen, Y. Shi, X. Xu, M. Huang, and J. Zhuang, “A machine learning-based pulmonary venous obstruction prediction model using clinical data and CT image,” Int. J. Comput. Assist. Radiol. Surgery, vol. 16, no. 4, pp. 609–617, Apr. 2021.

[4] Benchmark Leveling, Geographical Survey Institute. Accessed: Oct. 15, 2021. [Online]. Available: https://www.gsi.go.jp/sokuchikijun/suijun-survey.html

[5] K. Ishitsuka and T. Matsuoka, “Accuracy evaluation of persistent scatterer interferometry using synthetic datasets,” IEEE Trans. Geosci. Remote Sens., vol. 230, Sep. 2019, Art. no. 111608.

[6] N. Anantrasirichai, J. Biggs, F. Albino, P. Hill, and D. Bull, “Application of machine learning to classification of volcanic deformation in routinely generated InSAR data,” J. Geophys. Res., Solid Earth, vol. 15, pp. 6592–6606, Apr. 2018.

[7] C. M. J. Brengman and W. D. Barnhart, “Identification of surface deformation in InSAR using machine learning,” Geochim. Geophys. Geosyst., vol. 22, no. 3, Mar. 2021, Art. no. e2020GC009204.

[8] N. Anantrasirichai, J. Biggs, K. Kelevezi, Z. Sadeghi, T. Wright, J. Thompson, A. M. Achim, and D. Bull, “ Detecting ground deformation in the built environment using sparse satellite InSAR data with a convolutional neural network.” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 4, pp. 2940–2950, Apr. 2021.

[9] S. Mukherjee, A. Zimmer, X. Sun, P. Ghuman, and I. Cheng, “An unsupervised generative neural approach for InSAR phase filtering and coherence estimation,” IEEE Trans. Geosci. Remote Sens. Lett., early access, Jul. 31, 2020, doi: 10.1109/LGRS.2020.3010504.

[10] S. Fica, F. Calvanese, G. Scarpa, and P. Rizzoli, “A CNN-based coherence-driven approach for InSAR phase unwrapping,” IEEE Geosci. Remote Sens. Lett., early access, Oct. 21, 2020, doi: 10.1109/LGRS.2020.3029565.

[11] S. Mukherjee, A. Zimmer, N. K. Kottayil, X. Sun, P. Ghuman, and I. Cheng, “CNN-based InSAR denoising and coherence metric,” in Proc. IEEE SENSORS, Oct. 2018, pp. 1–4.

[12] O. L. Stephenson, T. Kohnne, E. Zhan, B. E. Cahill, S.-H. Yun, Z. E. Ross, and M. Simons, “Deep learning-based damage mapping with InSAR coherence time series,” IEEE Trans. Geosci. Remote Sens., early access, Jun. 29, 2021, doi: 10.1109/TGRS.2021.3084209.

[13] N. I. Bountos, I. Papoutsis, D. Michail, and N. Anantrasirichai, “Self-supervised contrastive learning for volcanic unrest detection,” IEEE Geosci. Remote Sens. Lett., early access, Aug. 19, 2021, doi: 10.1109/LGRS.2021.3104506.

[14] S. Fica, G. Gobbi, P. Rizzoli, and L. Buzzzone, “Φ-Net: Deep residual learning for InSAR parameters estimation,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 5, pp. 3917–3941, May 2021.

[15] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, “Indoor semantic segmentation using depth information,” in Proc. Int. Conf. Learn. Represent., 2012.

[16] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning hierarchical features for scene labeling,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[17] C. Couppie, C. Farabt, L. Najman, and Y. LeCun, “Indoor semantic segmentation using depth information,” in Proc. Int. Conf. Learn. Represent., 2012.

[18] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Indoor semantic segmentation using depth information,” in Proc. Int. Conf. Learn. Represent., 2012.

[19] C. Farabt, C. Couprie, L. Najman, and Y. LeCun, “Learning hierarchical features for scene labeling,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[20] C. Couppie, C. Farabt, L. Najman, and Y. LeCun, “Indoor semantic segmentation using depth information,” in Proc. Int. Conf. Learn. Represent., 2012.

[21] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Indoor semantic segmentation using depth information,” in Proc. Int. Conf. Learn. Represent., 2012.

[22] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Indoor semantic segmentation using depth information,” in Proc. Int. Conf. Learn. Represent., 2012.

[23] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Indoor semantic segmentation using depth information,” in Proc. Int. Conf. Learn. Represent., 2012.
K. Shimosato, N. Ukita: Multi-Modal Data Fusion for Land-Subsidence Image Improvement in PSInSAR Analysis

Leveling Measurement Data Provided by Yamagata. Accessed: Oct. 15, 2021. [Online]. Available: https://www.pref.yamagata.lg.jp/taiki-sui/65_020.html

Leveling Measurement Data Provided by Tokyo. Accessed: Oct. 15, 2021. [Online]. Available: https://www.metro.tokyo.jp/xjigyo/tech/start/03-joyouhou/suijun/suijun.html

Leveling Measurement Data Provided by Niigata. Accessed: Oct. 15, 2021. [Online]. Available: https://www.pref.niigata.lg.jp/seikyotsu/kankyoenergy/050051.html

Leveling Measurement Data Provided by Ibaraki. Accessed: Oct. 15, 2021. [Online]. Available: https://www.pref.ibaraki.jp/sekai/kan/yokou/kogai/kogai/jiban.html

Leveling Measurement Data Provided by Saitama. Accessed: Oct. 15, 2021. [Online]. Available: https://www.pref.saitama.lg.jp/a0505/901-20091217-288.html

Leveling Measurement Data Provided by Chiba. Accessed: Oct. 15, 2021. [Online]. Available: https://www.pref.chiba.lg.jp/suisho/jiban/chinka/torinkumi/seikaomote.html

Leveling Measurement Data Provided by Kanagawa. Accessed: Oct. 15, 2021. [Online]. Available: https://www.env.go.jp/water/jiban/seikatsukankyo/kantai/kogai/kogai/jiban.html

Leveling Measurement Data Provided by Tokai. Accessed: Oct. 15, 2021. [Online]. Available: https://www.gsi.go.jp/chubu/toukai3kenjiban.html

Leveling Measurement Data Provided by Saga. Accessed: Oct. 15, 2021. [Online]. Available: https://www.pref.saga.lg.jp/list00017.html

Ground Subsidence Regions in Japan in 2017. Accessed: Oct. 15, 2021. [Online]. Available: https://www.em.go.jp/water/jiban/gaikyo/gaikyo29.pdf

Copernicus Open Access Hub. Accessed: Oct. 15, 2021. [Online]. Available: https://scihub.copernicus.eu

Norimichi Ukita (Member, IEEE) was born in Japan, in 1973. He received the B.E. degree in engineering from Toyota Technological Institute, in 2020. His research interests include remote sensing and image enhancement methods.

Kodai Shimosato was born in Japan, in 1996. He received the B.S. degree in engineering from Toyota Technological Institute, in 2020. His research interests include remote sensing and image enhancement methods.

[57] Leaping Measurement Data Provided by Yamagata. Accessed: Oct. 15, 2021. [Online]. Available: https://www.pref.yamagata.lg.jp/taiki-sui/65_020.html

[58] Leaping Measurement Data Provided by Ibaraki. Accessed: Oct. 15, 2021. [Online]. Available: https://www.pref.ibaraki.jp/sekai/kan/yokou/kogai/kogai/jiban.html

[59] Leaping Measurement Data Provided by Saitama. Accessed: Oct. 15, 2021. [Online]. Available: https://www.pref.saitama.lg.jp/a0505/901-20091217-288.html

[60] Leaping Measurement Data Provided by Chiba. Accessed: Oct. 15, 2021. [Online]. Available: https://www.pref.chiba.lg.jp/suisho/jiban/chinka/torinkumi/seikaomote.html

[61] Leaping Measurement Data Provided by Kanagawa. Accessed: Oct. 15, 2021. [Online]. Available: https://www.env.go.jp/water/jiban/seikatsukankyo/kantai/kogai/kogai/jiban.html

[62] Leaping Measurement Data Provided by Tokai. Accessed: Oct. 15, 2021. [Online]. Available: https://www.gsi.go.jp/chubu/toukai3kenjiban.html

[63] Leaping Measurement Data Provided by Saga. Accessed: Oct. 15, 2021. [Online]. Available: https://www.pref.saga.lg.jp/list00017.html

[64] Leaping Measurement Data Provided by Ground Subsidence Regions in Japan in 2017. Accessed: Oct. 15, 2021. [Online]. Available: https://www.em.go.jp/water/jiban/gaikyo/gaikyo29.pdf

[65] Leaping Measurement Data Provided by Copernicus Open Access Hub. Accessed: Oct. 15, 2021. [Online]. Available: https://scihub.copernicus.eu

[66] Leaping Measurement Data Provided by Norimichi Ukita (Member, IEEE) was born in Japan, in 1973. He received the B.E. degree in engineering from Okayama University, in 1996 and 1998, respectively, and the Ph.D. degree in informatics from Kyoto University, in 2001. From 2001 to 2016, he was an Assistant Professor (from 2001 to 2007) and an Associate Professor (from 2007 to 2016) with the Nara Institute of Science and Technology. He was a Visiting Research Scientist at Carnegie Mellon University, from 2007 to 2009. In 2016, he became a Professor with the Toyota Technological Institute.

[67] Leaping Measurement Data Provided by Dr. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Represent., vol. 3, no. 1, pp. 48–63, May 2015.