European Journal of Remote Sensing

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tejr20

Towards zero-shot learning in 3D change detection: improving generalization with custom augmentations and evaluation

Riccardo Contu, Valerio Marsocci, Virginia Coletta, Roberta Ravanelli & Simone Scardapane

To cite this article: Riccardo Contu, Valerio Marsocci, Virginia Coletta, Roberta Ravanelli & Simone Scardapane (2025) Towards zero-shot learning in 3D change detection: improving generalization with custom augmentations and evaluation, European Journal of Remote Sensing, 58:1, 2453468, DOI: 10.1080/22797254.2025.2453468

To link to this article: https://doi.org/10.1080/22797254.2025.2453468

Published online: 11 Feb 2025.

EUROPEAN JOURNAL OF REMOTE SENSING
2025, VOL. 58, NO. 1, 2453468
https://doi.org/10.1080/22797254.2025.2453468

Towards zero-shot learning in 3D change detection: improving generalization with custom augmentations and evaluation

Riccardo Contu[a], Valerio Marsocci[b], Virginia Coletta[a], Roberta Ravanelli[a,c] and Simone Scardapane[d]

[a]Department of Civil, Edil and Environmental Engineering, Sapienza University of Rome, Rome, Italy; [b]Geomatics Research Group, KU Leuven, Gent, Belgium; [c]Geomatics Unit, Department of Geography, University of Liège, Liège, Belgium; [d]Department of Information, Electronics, and Telecommunications Engineering, Sapienza University of Rome, Rome, Italy

ABSTRACT

Detecting and quantifying changes is crucial for monitoring transformations of the Earth's surface. It is thus essential to employ methods that can effectively retrieve both 2D and 3D changes over time. The MultiTask Bitemporal Images Transformer (MTBIT) was recently introduced to tackle 2D and 3D Change Detection (CD) tasks using bi-temporal optical images. Despite strong performances on existing benchmarks, MTBIT shows some limitations, i.e. a pronounced tendency to overfit the training distribution and difficulty in inferring extreme values, which motivates the need for improvements. We hence propose a new set of custom augmentations, applied individually or in specific combinations, to discern intricate geometries small structures and subtle terrain changes. Furthermore, to address conventional evaluation metrics' limitations we introduce the true positive RMSE (tpRMSE) metric which provides a more comprehensive understanding of MTBIT efficacy. The most successful augmentation combination reduces cRMSE to 5.88 m and tpRMSE to 5.34 m, from 6.33 m and 5.60 m of the baseline, respectively. Finally, a first zero-shot learning experiment is carried out on a new small dataset, achieving promising improvements towards domain generalization. In summary, the proposed contributions enhance the practical utility and reliability of MTBIT in real-world applications, addressing critical challenges in the field of Remote Sensing CD.

ARTICLE HISTORY
Received 29 February 2024
Revised 6 June 2024
Accepted 8 January 2025

KEYWORDS
3D change detection; data augmentation; deep learning; remote sensing; 3D change detection metrics; zero-shot learning

Introduction

Remote Sensing Change Detection (RSCD) refers to methods and algorithms employed to identify changes among scenes of the same area captured in different epochs (Hayet et al., 2020). Remote sensing (RS) data are indeed an essential source for large-scale and regular estimation of changes affecting Earth's surface (Hafner, Ban, et al., 2022), whose analysis constitutes a preliminary step for understanding relationships and interactions between human actions and natural phenomena (Lu et al., 2004).

In particular, 3D Change Detection (3D CD) comprehends all the methods able to retrieve the quantitative 3D (volumetric/elevation) changes between two or more epochs and not simply the 2D (footprint) extent of the areas affected by the changes (Coletta et al., 2022; Marsocci, Coletta, et al., 2023; Qin et al., 2016). 3D CD has versatile applications in fields such as environmental monitoring, urban planning, ecology, and civil engineering, for tasks ranging from urban development assessment to ecological monitoring and disaster evaluation (Qin et al., 2016).

In this context, the applications of Artificial Intelligence (AI) are continuously increasing, ranging from the initial processing of the imagery to high-level data understanding and knowledge extraction (Zhang & Zhang, 2022). In particular, an ever-increasing number of researchers are currently employing Deep Learning (DL) for RSCD tasks, due to its better performance over conventional CD methods (Bai et al., 2022), with significant advantages in automation and accuracy (M. Xu et al., 2023).

For instance, the impressive results of Vision Transformers (Vits) (Dosovitskiy et al., 2020) and Diffusion Models (Yang et al., 2024) in solving computer vision tasks have recently led researchers in employing these two families of architectures also for RSCD (Bandara et al., 2022; Chen, Qi, et al., 2022; C. Chen et al., 2021; Marsocci, Gonthier, et al., 2023). Moreover, Vision Foundation Models (VFMs), such as Segment Anything Model (SAM) (Kirillov et al., 2023) and Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021), have lately emerged and gained great attention thanks to their ability to leverage the general knowledge gained in "metascale" datasets (Ding et al., 2024) to solve specific downstream tasks in computer vision applications through their fine-tuning on small specialized datasets. Only recently, VFMs have been applied

CONTACT Riccardo Contu riccardo.contu@uniroma1.it Department of Civil, Edil and Environmental Engineering, Sapienza University of Rome, Via Eudossiana 18, 00184 Rome, Italy

to the 2D CD of very high-resolution RS images (K. Chen et al., 2023; Ding et al., 2024; Dong et al., 2024).

Recently, the MultiTask Bitemporal Images Transformer (MTBIT) (Marsocci, Coletta, et al., 2023), a neural network belonging to the family of ViTs (Dosovitskiy et al., 2020), has been proposed to simultaneously solve the 2D and 3D CD tasks from bi-temporal optical images, thus without the need to rely on elevation data during the inference phase. MTBIT showed promising performances in the 3D CD task, laying the foundations for the development of DL algorithms able to automatically produce 3D CD maps (raster maps containing the quantitative changes in elevation) together with standard 2D CD maps (raster maps identifying the horizontal extent of the areas affected by the elevation changes) (Figure 1) (Marsocci, Gonthier, et al., 2023).

Despite robust performances in in-domain inference tasks, MTBIT's capacity to adapt to new, out-of-domain data remains limited, primarily due to a pronounced tendency of the network to overfit the training distribution (Marsocci, Coletta, et al., 2023). This phenomenon limits the ability of the model to perform effectively in zero-shot scenarios, where no specific domain information is provided. Specifically, zero-shot learning in computer vision is a paradigm where a model is trained to recognize scenes, objects and/or changes (when talking about CD more specifically (Chen, Yokoya, et al., 2023; Zheng et al., 2024)) it has never been seen during the training phase. Unlike traditional supervised learning, where models are trained on a predefined set of categories or a specific domain, zero-shot learning enables a CV system to generalize to new, unseen classes or domains at inference time. To address these limits and to enhance the practical utility of the network, in this work we present our efforts to increase the generalization performances (i.e. in- and out-of-domain performance) of MTBIT through the proposal of a new set of custom augmentations.

It is well known that the performances of a DL network strongly depend on the amount and on the quality of the available training data, which, however, are often onerous to collect (Fuentes Reyes et al., 2023; M. Xu et al., 2023). Moreover, the production of annotated bi-temporal 2D/3D datasets is especially problematic, both for the double effort required for labeling the dataset and for the difficulties in finding multimodal data available in the same area in – as much as possible – contemporary epochs (Fuentes Reyes et al., 2023).

A significant effort has already been made in this regard to build the 3DCD dataset (Coletta et al., 2022; Marsocci, Gonthier, et al., 2023), on which MTBIT was originally trained. The 3DCD dataset consists of a relatively limited number of images, but, at the moment of writing, it remains one of the only two benchmarks (the other being SMARS (Fuentes Reyes et al., 2023)) providing 2D/3D multimodal and multitemporal data suitable for training and evaluating DL algorithms in RS 2D/3D CD, and the only one including real data. Given the aforementioned difficulties with the expansion of the 3DCD dataset, this study hence focuses on the proposal of a new set of custom augmentations to improve the generalization capabilities of MTBIT.

Data augmentation is a technique commonly used in DL to artificially expand the dataset by applying various transformations to the existing training data to produce new and modified versions of the original training samples. Data augmentation increases, thus, the variability of the training set, helping the network to generalize better to unseen data and to be more robust to the overfitting phenomenon (M. Xu et al., 2023; Zhu et al., 2023a). Despite the demonstrated effectiveness of such augmentation techniques, few contributions developed specific custom augmentations, especially to address CD.

Specifically, we implemented several augmentation strategies to help MTBIT predict subtle terrain changes and difficult-to-detect geometries such as, but not limited to, small newly constructed or demolished buildings and low road embankments. To our knowledge, besides some standard augmentation methods like the ones available in (Buslaev et al., 2020; PyTorch, n.d.), there are no augmentation strategies specifically conceived for the 3D CD task.

Furthermore, in this work, we also address the evaluation of MTBIT CD performances, a fundamental

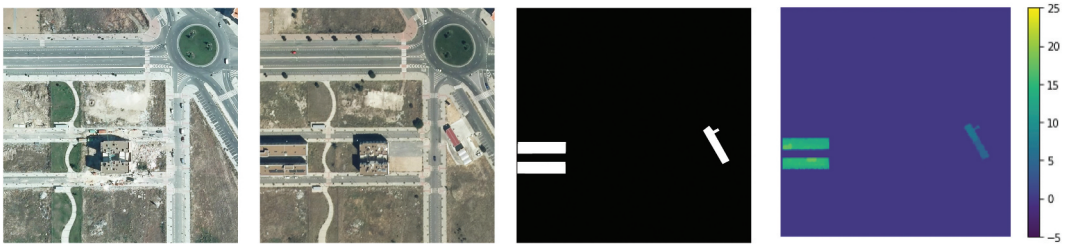


Figure 1. An example from the Avila dataset: (a) 2010 RGB image, (b) 2017 RGB image, (c) 2D CD map (the pixels affected by a change in elevation have a value of 1 and are shown in white), (d) 3D CD map: the color bar is expressed in meters.

step to correctly assess the effects of the implemented set of custom augmentations. The limitations of the standard Root Mean Squared Error (RMSE) metric were already reported in (Marsocci, Coletta, et al., 2023), where the change RMSE (cRMSE) was proposed as a new 3D CD metric. In the cRMSE, the error is computed considering all the pixels of the predicted 3D CD map, and then, it is normalized with respect to the number of pixels affected by an actual change (i.e. the number of ground truth pixels affected by a change). Nonetheless, even if the cRMSE is more sensitive to the detected 3D changes than the standard RMSE, it still gives a partial understanding of the actual performance of the network on the 3D CD task. For this reason, we propose a new metric, namely the true positive RMSE (tpRMSE), where the error is computed considering just the pixels affected by an actual elevation change and then it is normalized with respect to the number of changing pixels.

Contributions of the work

The main contributions of this work can be summarized as follows:

- we propose a new custom set of augmentations, specifically implemented for the 3D CD task, capable of improving the in- and out-of-domain performances of MTBIT;
- we introduce a new metric, the tpRMSE, for a more complete evaluation of MTBIT performances on the 3D CD task;
- we test the generalization capabilities of MTBIT in a zero-shot scenario, with and without the application of the proposed set of augmentations, assessing its performances on the 3D CD Avila dataset, a new test set with respect to the 3DCD dataset (Coletta et al., 2022) used to train the model.

Related works

Unimodal and multimodal change detection

3D CD has recently gained great attention due to its capability of providing valuable information about volumetric dynamics (Fuentes Reyes et al., 2023; Marsocci, Gonthier, et al., 2023; Qin et al., 2016). 3D CD comprehends all the methods able to retrieve the quantitative 3D (volumetric/elevation) changes that occur among two or more epochs by analyzing remote sensing data captured at different times. In this context, the possibility of increasing the data sources for a timely and accurate monitoring of changes is invaluable. Under these circumstances, multimodality has attracted a growing interest in the RS community, as it enables the retrieval of a greater amount of – heterogeneous – information (Chen, Song, et al., 2023; Ghasem Abdi et al., 2017). Different multimodality approaches have been proposed moieez 2023, employing Digital Surface Models (DSMs) (Rottensteiner et al., 2012), aerial images (Emmanuel Fundisi et al., 2022), RADAR images (Emmanuel Fundisi et al., 2022; Hafner, Nascetti, et al., 2022), including also land cover maps (Hong et al., 2023) and other types of ancillary data (e.g. street views (Deuser et al., 2023), vector data (Audebert et al., 2017), geographical coordinates (Marsocci, Coletta, et al., 2023)). However, none of them considers multi-temporal data, preventing the application of CD techniques. Indeed, while unimodal CD (also known as homogeneous CD), i.e. the task of detecting changes from multitemporal images collected by the same kind of sensors has been widely studied over the past decades, and relatively few studies have been conducted on multimodal CD (Chen, Yokoya, et al., 2022; K. Chen et al., 2023). In multimodal CD, the pre-change and post-change images are collected by sensors with different acquisition modes, like in the case of multispectral and SAR sensors. Multimodal images are thus characterized by different statistical distributions, channel numbers, and noise levels (Chen, Yokoya, et al., 2023) and in this scenario it is challenging to achieve accurate CD results using methodologies developed specifically for unimodal imagery (H. Chen, Yokoya, et al., 2022). A few promising approaches have been recently proposed for multimodal CD, from data fusion (Hafner, Ban, et al., 2022) to the exploitation of modality-invariant structural relationships (Y. Sun et al., 2021, 2022, H. Chen, Yokoya, et al., 2022), and some of them have also been successfully applied to unimodal CD (Y. Sun et al., 2022). Nevertheless, a review of multimodal CD approaches is beyond the scope of this paper, since our task can be categorized as a unimodal CD multi-task application: our bi-temporal input images are indeed captured with the same – optical – modality (see Section 3 for the details) and are used to produce both the 2D and 3D CD maps.

On the other hand, numerous studies address unimodal 2D (or footprint) CD using various approaches (Bai et al., 2022; Bandara & Patel, 2022; Bing Liu et al., 2023; Chen, Yokoya, et al., 2022; Daudt et al., 2018; Shuwen Xu et al., 2020), ranging from diffusion models (Bandara et al., 2022) to self-supervised learning (Leenstra et al., 2021), and, more recently, leveraging the latent knowledge of VFMs (Chen, Song, et al., 2023; Ding et al., 2024). Moreover, the available datasets are growing both in number (Shafique et al., 2022) and features (Shi et al., 2022), ranging across diverse possible applications such as disaster assessment (Baldi et al., 2005; Ghuffar et al., 2013), coastal-line monitoring, urban development (Malpica et al., 2013; Menderes et al.,

2015; Taneja et al., 2011) and crop monitoring (Bendig et al., 2013; Liu et al., 2022).

Concerning 3D CD, two alternative approaches are generally used to retrieve 3D changes: the DSM difference and the cloud-to-cloud distance (Coletta et al., 2022; Qin et al., 2016; Shirowzhan et al., 2019; Stilla & Xu, 2023). MTBIT employs the first method: 3D CD maps are DSM difference maps. 3D CD maps convey the elevation change information in the same raster format as 2D CD maps, thus they can be easily integrated into standard 2D CD pipelines, whereas the unordered and irregular natures of point cloud data makes the extraction of the information they carry more difficult (Coletta et al., 2022).

Another line of research involves generating synthetic datasets, useful for obtaining large-scale datasets without the need for expensive ground truth (such as LiDAR data) (Song et al., 2023). Among these recent works, SMARS (Fuentes Reyes et al., 2023) stands out, offering an extensive cross-domain synthetic dataset with 2D and DSMs, as well as land cover maps.

However, in terms of algorithms for resolving the 3D CD task, MTBIT (Marsocci, Gonthier, et al., 2023) remains at the time of writing the only effective method, albeit with the previously highlighted limitations.

Data augmentation for remote sensing

Data augmentation has proven to be a useful technique in the majority of DL applications (Shorten & Khoshgoftaar, 2019). Indeed, data augmentation helps prevent overfitting, even when dealing with a limited number of images in the dataset (Oubara et al., 2022). Moreover, in recent years, it has been crucial for specific techniques, such as self-supervised learning (Marsocci et al., 2021).

Also in RS, this practice has gained significant adoption in recent years (Hao et al., 2023; Lalitha & Latha, 2022; Oubara et al., 2022; M. Xu et al., 2023), with data augmentation approaches that can vary depending on the task (Xingrui Yu et al., 2017) or on the methodology (Lv et al., 2021). For example, data augmentation has been used in the field of image scene classification with Convolutional Neural Networks (CNNs) (Shawky et al., 2020). In (Haut et al., 2019), random occlusion data augmentation was used for hyperspectral image classification. By combining three-dimensional models of aircraft, with remote sensing images (Yan et al., 2019), produced simulated images to improve the positive sample information of the dataset. Moreover, Xiao et al., 2021 integrates ship simulation data and utilizes a neural-style transfer network to create an extensive set of transformed remote sensing ship images.

For CD, some approaches based on mosaic simulation and haze image simulation (Wang et al., 2023), generative methods (Zhu et al., 2023b), and text descriptors (Chen et al., 2021) have been recently proposed. In addition, several data augmentation techniques have been proposed for zero-shot transfer, from data-warping, to simulation-based methods, and deep-generative-model-based methods (B. Xu et al., 2022) (X. Sun et al., 2021). However, to our knowledge, specific augmentations for 3D CD are not yet available, even though they are fundamental to improving generalization and performance, as we will show in the remainder of this paper.

The 3DCD datasets

3DCD dataset

MTBIT was trained on the 3DCD dataset (Coletta et al., 2022) which covers the area of Valladolid in the 2010 and 2017 epochs. The 3DCD dataset is composed of 472 pairs of RGB images cropped from optical aerial orthophotos with a size of 400 × 400 pixels and a Ground Sampling Distance (GSD) of 0.5 m, 472 2D CD maps in raster format (GSD 0.5 m) and 472 3D CD maps in raster format (GSD 1 m) (Coletta et al., 2022). More details about the 3DCD dataset can be found in (Coletta et al., 2022; Marsocci, Coletta, et al., 2023).

Avila dataset

The Avila dataset was used to evaluate the generalization potential of the proposed augmentation techniques. This dataset is composed of just 7 pairs of RGB images, cropped from optical orthophotos captured in 2010 and 2017, and the 7 corresponding 2D (GSD 0.25 m) and 3D CD (GSD 0.25 m) maps in raster format (Figure 1). The bi-temporal images are characterized by a size of 800 × 800 pixels and a GSD of 0.25 m, each one covering an area of 200 m × 200 m. The approach adopted to produce this small dataset was the same employed by the authors for the 3DCD dataset (Coletta et al., 2022).

Methodology

MTBIT

In (Marsocci, Coletta, et al., 2023), the MTBIT architecture was introduced to solve the 2D CD and 3D CD tasks simultaneously from bi-temporal optical images. Specifically, MTBIT belongs to the family of ViTs (Dosovitskiy et al., 2020). ViTs have recently emerged as a DL paradigm which enables networks that employ them to retrieve and integrate global contextual information through a self-attention mechanism, i.e. the interaction between input sequences that help the network to determine which region it should pay more attention to (Asadi Shamsabadi et al., 2022).

MTBIT architecture is composed of four parts (Figure 2):

(1) a Siamese semantic tokenizer, which generates a set of tokens (i.e. visual words) for each one of the input bitemporal optical images. It is made of a ResNet18 backbone, which extracts the features, and a spatial attention tokenizer, which converts the extracted features in semantic tokens;

(2) a Transformer-based encoder, which encodes positional and context information in the token space. It consists of multi-head self-attention layers, with positional encodings, followed by Multi-Layer Perceptron (MLP) blocks;

(3) a Transformer-based decoder that extracts feature maps projecting the tokens in the pixel space. It is made of multi-head cross-attention and MLP layers. This configuration helps to combine different contexts and spatial information at different levels;

(4) a pair of prediction heads, made of fully convolutional networks, that allow for a double mapping from the refined feature maps to the 2D and 3D CD maps.

As for loss (L), MTBIT uses a weighted combination of a 2D loss (L_{2D}) and a 3D loss (L_{3D}) (Marsocci, Gonthier, et al., 2023):

$$L = \alpha \cdot L_{2D} + \beta \cdot L_{3D} \quad (1)$$

where α and β are fixed weights ($\alpha = 1$ and $\beta = 3$). In particular, MTBTIT employs a binary cross-entropy (BCE) as L_{2D} and the mean squared error (MSE) as L_{3D}. More details about MTBIT can be found in (Marsocci, Coletta, et al., 2023).

Augmentation techniques

This work focuses on improving the performances of the MTBIT (Marsocci, Gonthier, et al., 2023) network by applying a custom set of data augmentation strategies that go beyond standard approaches like random horizontal flip, random geometric transformation, random Gaussian noise, and random radiometric transformation (i.e. brightness, sharpening, blurring, contrasting, saturation).

In particular, we propose the set of augmentations listed below, capable of enhancing the details of the bitemporal optical images or the generality of the 3D CD maps, and focus on the peculiarities of the 3D CD task (e.g. shadows and elevation changes). They are:

- Change-Guided Random Crop (Figure 3a): this technique performs a random crop (256 × 256 pixels) on both the bi-temporal images maintaining at least one pixel in the area corresponding to the changed area in the 2D CD mask;
- Random Crop or Resize (Figure 3b): this technique randomly chooses to resize or crop the images to 256 × 256 pixels with a 50% probability. This technique was already employed in a few other cases (Caron et al., 2021); we propose this augmentation as an effective means to combine different details (at different spatial scales) of the two bi-temporal images. In this way, the model can focus on different aspects of the images through the different epochs (two in our case);
- CutMix (Figure 3c,d): this technique randomly cuts out an arbitrary number of patches from the first-epoch image and pastes them onto the second-epoch image and vice versa (considering the corresponding positions of the first-epoch image). This augmentation forces the model to focus on the invariant features of the bitemporal representations. In fact, by mixing parts

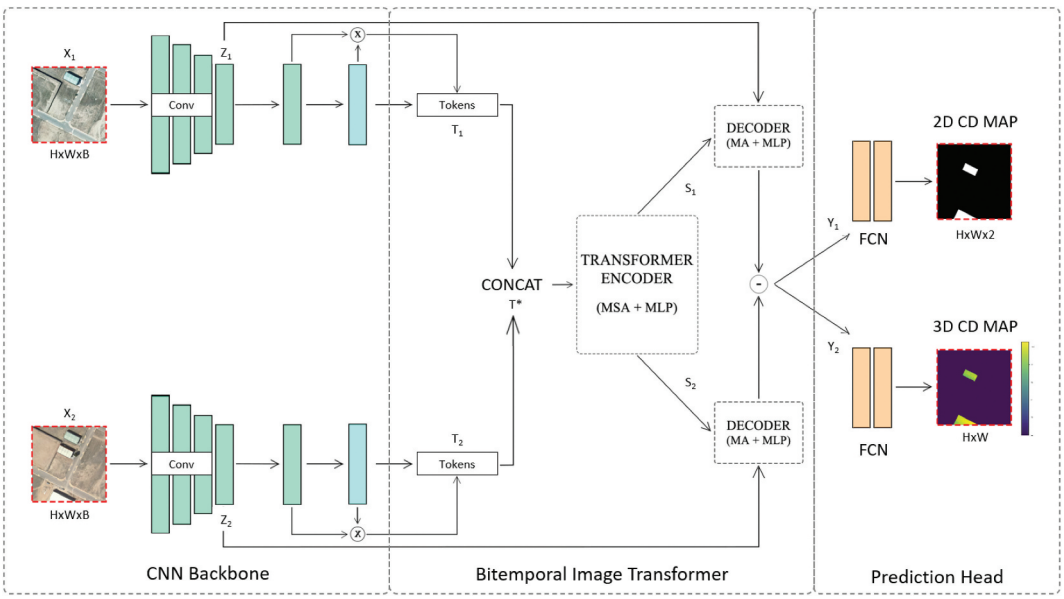
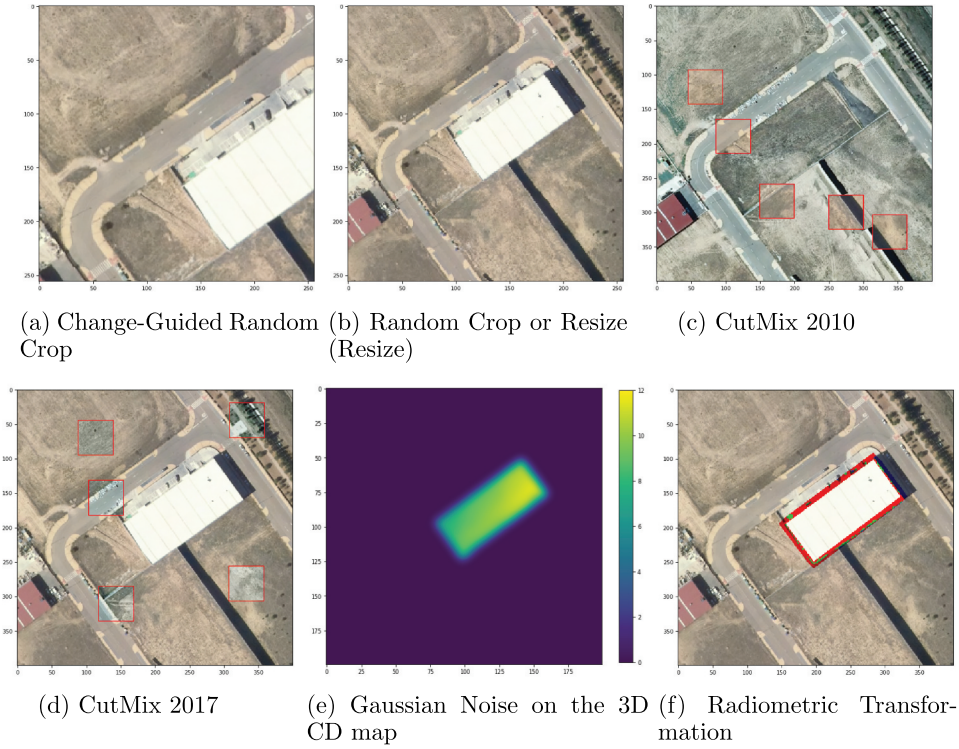


Figure 2. Schematic representation of the MultiTask bitemporal image transformer (MTBIT) network architecture.



(a) Change-Guided Random Crop (b) Random Crop or Resize (Resize) (c) CutMix 2010

(d) CutMix 2017 (e) Gaussian Noise on the 3D CD map (f) Radiometric Transformation

Figure 3. The proposed set of custom augmentations. The color bar in (e) is expressed in meters.

of the bi-temporal images, the model is enforced to grasp local dependencies that are strong also across different epochs, where each image continues to differ from itself;

- Gaussian Noise on 3D CD maps (Figure 3e): this technique applies a standard Gaussian Noise by convolving the 3D CD map with a Gaussian kernel with an arbitrary σ value. As the authors observed in (Marsocci, Coletta, et al., 2023), MTBIT is not able to infer the extreme values of the elevation changes, due to the distribution of the changed pixels. The injection of noise into the 3D CD map can help the model to achieve a better generalization;

- Radiometric transformation on the buffer zone (Figure 3f): this technique applies a radiometric transformation on the input bi-temporal images in correspondence to the borders of the changed areas on the 2D CD maps. The radiometric transformation acts on the brightness, contrast, saturation, and HUE of the border pixels, and their values can be set arbitrarily. We consider this augmentation as counter-evidence. The shadows are indeed fundamental to retrieving information on the elevation changes. The application of the radiometric transformations at the border of the areas affected by changes makes the model struggle to reconstruct the elevation changes.

Change detection metrics

Different CD metrics were considered to evaluate the performances of MTBIT. Specifically, we observed that the 3D CD metrics proposed in (Marsocci, Coletta, et al., 2023) provide only a partial understanding of the actual performances of the network on the 3D CD task.

The limitations of the standard Root Mean Squared Error (RMSE) metric (Equation 4) – too influenced by the presence of a large number of zeros (no changes) in the 3D CD maps – were already reported in (Marsocci, Gonthier, et al., 2023), where the change RMSE (cRMSE) was proposed as a new 3D CD metric.

The cRMSE (Equation 5) is an RMSE where the error is computed considering all the pixels of the predicted 3D CD map and then it is normalized with respect to the number of the pixels affected by an actual change (i.e. the number of ground truth pixels affected by a change). Nonetheless, even if the cRMSE is more sensitive to the actual 3D changes than the standard RMSE, from its evaluation we cannot be entirely certain that the model accurately infers the elevation changes between the two considered epochs, i.e. the elevation changes in which we are most interested. For this reason, we introduce the true positive RMSE (tpRMSE) metric.

The tpRMSE metric (Equation 6) is an RMSE where the error is computed considering only the pixels of the predicted 3D CD map affected by an actual elevation change (i.e. the pixels which in the ground truth 3D CD map are characterized by $\Delta H \neq 0$), and then they are normalized with respect to the number of these really changed pixels. The use of the tpRMSE, along with the cRMSE, allows for a better evaluation of the 3D CD performances since it considers only the

pixels affected by an actual change rather than considering all the pixels regardless of their class (change or no change).

The computation details of the three 3D CD metrics considered in this study are reported in Algorithm 1.

Algorithm 1: Difference among the 3D CD metrics

Input: predicted 3DCD maps $\widehat{\Delta H}$; ground truth ΔH
/* squared error */
for i *in* ΔH; // iterate on pixels
do
  SE $\leftarrow \widehat{\Delta H}_i - \Delta H_i$; // squared error
  if $\Delta H_i \neq 0$; // filter only changed pixels
  then
    TPSE $\leftarrow \widehat{\Delta H}_i - \Delta H_i$; // true positive s.e.
    $n_c += 1$; // count the changed pixels
  end
end
/* mean squared error */
MSE $\leftarrow$ SE/$len(\Delta H)$;
cMSE $\leftarrow$ SE/n_c;
tpMSE $\leftarrow$ TPSE/n_c;
/* root mean squared error */
RMSE, cRMSE, tpRMSE $\leftarrow$ $sqrt$(MSE), $sqrt$(cMSE), $sqrt$(tpMSE);

As regards the 2D CD metrics, we considered the standard segmentation metrics, i.e. the Intersection over Union (IoU, Equation 2) and the F1-score (F1, Equation 3), referred to as the change class (i.e. the pixels with value 1 in the 2D CD maps).

The considered CD metrics are formally expressed as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{2}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\Delta H}_i - \Delta H_i\right)^2} \tag{4}$$

$$\text{cRMSE} = \sqrt{\frac{1}{n_c}\sum_{i=1}^{n}\left(\widehat{\Delta H}_i - \Delta H_i\right)^2} \tag{5}$$

$$\text{tpRMSE} = \sqrt{\frac{1}{n_c}\sum_{i=1}^{n_c}\left(\widehat{\Delta H}_i^C - \Delta H_i^C\right)^2} \tag{6}$$

where TP stands for True Positives, TN for True Negatives, FP for False Positive and FN for False Negatives, $\widehat{\Delta H}_i$ for predicted pixels, ΔH_i for ground truth pixels, $\widehat{\Delta H}_i^C$ for predicted pixels, ΔH_i^C for respective ground truth changed pixels. The term n_c refers to the pixels that have a value of 1 (change class) in the ground truth 2D CD map and a value different from 0 ($\Delta H \neq 0$) in the ground truth 3D CD map.

One interesting aspect of the 3D CD metrics considered in this study is that both the tpRMSE and cRMSE are preferable to the standard RMSE since they are more sensitive to actual elevation changes and the errors in their inference. Most of the pixels in the ground truth 2D and 3D CD maps are indeed characterized by an absence of change ($\Delta H = 0$), so they are easier to detect than the minority of pixels affected by an actual elevation change ($\Delta H \neq 0$).

Experimental results

For each experiment, five independent tests were carried out; the corresponding mean and standard deviation are reported in all the tables summarizing the results. We follow the same procedure also for the baseline, i.e. the MTBIT architecture (Marsocci, Coletta, et al., 2023) trained using just the resize augmentation, needed to feed the network with 256×256 pixel images: this is the reason why the results are slightly different from those reported in (Marsocci, Gonthier, et al., 2023).

For the training phase, a single Tesla T4 16 GB GPU was used. We trained MTBIT for 300 epochs with the AdamW optimizer (Loshchilov & Hutter, 2019) and a learning rate of 0.0001. Furthermore, we normalized the 3D CD maps between −1 and 1, with a min-max normalization among $-25\ m$ and $30\ m$, using a TanH as the last non-linear layer of the proposed network. For the 2D loss (i.e. BCE), we set the weights to 0.05 for no-change pixels and 0.95 for change ones. Table 1 summarizes the hyperparameters selected for the augmentations.

For the zero-shot experiments, we simply took the models trained on the training dataset (i.e. the 3DCD dataset – Valladolid, Section 3.1) with the proposed set of custom augmentations and with just the resize augmentation (baseline) and performed inference in a zero-shot scenario on the Avila dataset (Section 3.2).

Single augmentation techniques

First, the proposed augmentation techniques were individually applied to assess their single contribution to the improvement of the CD metrics.

The principal objective of each proposed augmentation is to lower the values of the cRMSE and tpRMSE metrics with respect to the baseline and to achieve a better fitting of the distribution of the inferred elevation changes to the Ground Truth distribution.

The Change-Guided Random Crop technique was applied to enhance the training process by enabling

Table 1. Augmentation hyperparameters. For major details on the selection of the values of the hyperparameters, see Section 5.3.

Augmentation Type	Figure 5	Hyperparameters
Baseline	a	–
Change-Guided Random Crop	b	256×256 crop
Random Crop or Resize	c	$p = 0.5$
Cut Mix	d	5 patches of 50×50
Gaussian Noise	e	$\sigma = 3$ m
Rad. Tranf. on Buffer Zone	f	3 pixels wide buffer

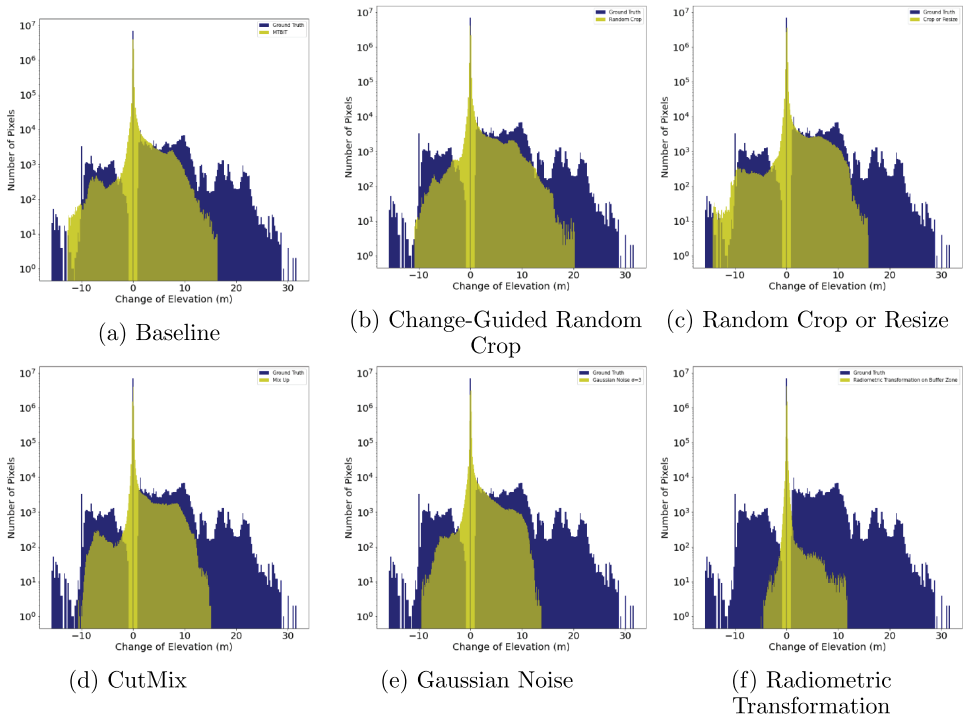
MTBIT to focus on regions affected by actual changes in elevation. In the case of this augmentation, each crop always includes at least one pixel with a value of 1 (change class) in the 2D CD map. As expected, this technique leads to a reduction in the 3D CD metrics, with the cRMSE decreasing by approximately 20 cm and the tpRMSE by around 15 cm (Table 2b). The histogram of the inferred elevation changes shows in this case a better fit with the Ground Truth distribution, particularly around 20 m, where the baseline falls short (Figure 4b). However, there may be a slight deterioration of the inference performances in some cases, where the 3D CD maps are not accurately predicted (Figure 5b).

The Random Crop or Resize strategy, which involves cropping or resizing images randomly with a 50% probability, also aligns with expectations and significantly impacts the detection of negative elevation change values. This method manages to cover a broader range of values than the baseline, as it is possible to observe in the histogram (Figure 4c). Notably, the cRMSE is reduced by 35 cm, reaching a relatively low value of 5.50 ± 0.10 m. Also, visual results (Figure 5c) demonstrate improvements with respect to the baseline (Figure 5a), especially in outlining challenging geometries. The combination of different resolutions and areas of focus appears to enhance the learning process of the network.

The CutMix technique aims to direct the network's focus to relevant areas of the image, which, in this study, are the areas affected by elevation changes. Unlike cropping strategies, CutMix involves mixing

Table 2. Validation results of the different augmentation techniques applied to the 3DCD dataset using the MTBIT architecture.

		2D CDmetrics		3D CDmetrics		
Augmentation type	Figure 5	F1(%)	IoU(%)	RMSE(m)	cRMSE(m)	tpRMSE(m)
Baseline	a	62.80 ± 0.40	45.82 ± 0.40	1.18 ± 0.02	6.33 ± 0.06	5.60 ± 0.08
Change-Guided Random Crop	b	58.73 ± 0.72	41.58 ± 0.73	1.17 ± 0.02	6.15 ± 0.10	**5.45** ± 0.06
Random Crop or Resize	c	**63.52** ± 0.81	**46.55** ± 0.87	**1.16** ± 0.03	**6.00** ± 0.07	5.50 ± 0.10
CutMix	d	62.63 ± 1.14	45.61 ± 1.17	1.18 ± 0.02	6.15 ± 0.06	5.47 ± 0.03
Gaussian Noise	e	61.74 ± 0.55	44.61 ± 0.56	1.18 ± 0.01	6.16 ± 0.03	5.51 ± 0.06
Rad. Transf. on Buffer Zone	f	12.51 ± 2.94	6.59 ± 2.13	1.54 ± 0.03	7.45 ± 0.07	7.31 ± 0.09



(a) Baseline (b) Change-Guided Random Crop (c) Random Crop or Resize

(d) CutMix (e) Gaussian Noise (f) Radiometric Transformation

Figure 4. Comparison between the distribution of the elevation change values in the ground truth (in blue) and the distribution of the predicted elevation change values obtained with the different augmentations reported in Table 2 (in yellow).

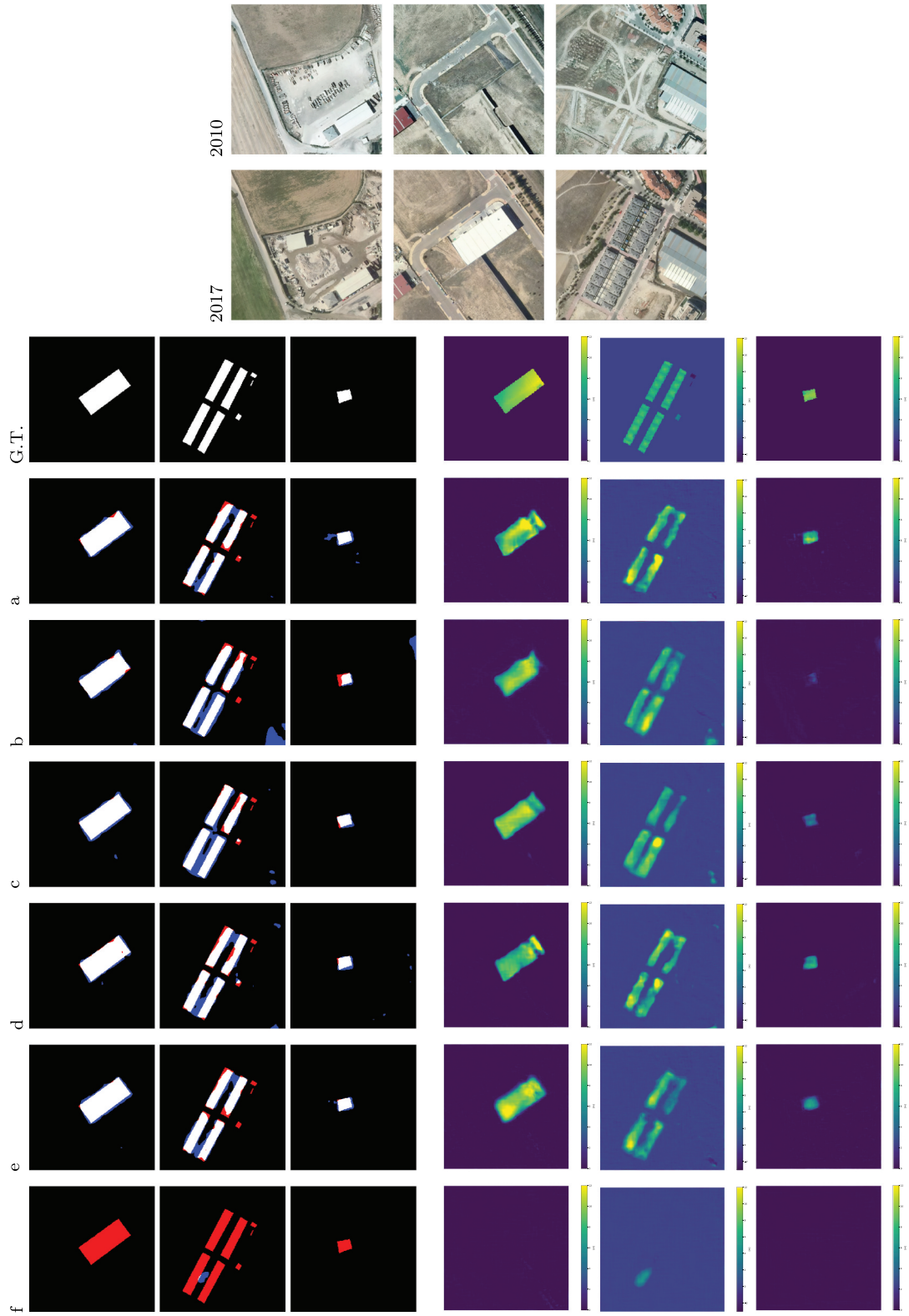


Figure 5. Visualisation of 2D CD and 3D CD maps for the augmentations described in Table 2. For the 2D change predictions, in black TN, in white TP, in red FN, in blue FP.

images in non-changing areas by adding noise. Through experimentation, the best configuration involved cropping five patches of 50 × 50 pixels from both the bi-temporal images (Section 5.3). The CutMix method excels in predicting pixel values (Figure 4d) in the low elevation range. Furthermore, the cRMSE metric is reduced in this case by approximately 20 cm (Table 2d). Visual results are comparable to those of the baseline (Figure 5a).

The Gaussian Noise technique applies a standard Gaussian noise to the 3D CD maps by convolving the considered 3D CD map with a Gaussian kernel with an arbitrarily chosen σ value. After trial and error tests, a σ value of 3 m was found to be the most effective

(Section 5.3). The comparison of the histogram of the inferred elevation changes with the baseline does not reveal significant differences in the case of this augmentation, but the cRMSE and tpRMSE metrics show improvements of 20 cm and 10 cm, respectively (Table 2e). Visual results are also promising (Figure 5e).

The radiometric transformation around the buffer zone manipulates the areas of the bi-temporal input images around newly constructed or demolished buildings (areas corresponding to – elevation – changes in the 2D CD maps), representing shadow regions in the images. The network relies on these shadows to infer the elevation of buildings. This augmentation is applied to verify the learning process adopted by the network. The results obtained with this technique are indeed notably worse if compared to the other augmentations and even worse than the baseline (Table 2f), as expected. The performance decline can be attributed to the perturbation of the crucial shadow areas around the buildings, which are essential for assessing elevation variations.

Finally, it is possible to better comprehend the peculiarities of the tpRMSE metric (Equation 6) by considering the first and the third predictions shown in Figure 5a, i.e. the baseline predictions. In these two cases, the RMSE is equal to 0.81 m and 0.56 m, respectively. However, these low values are mainly due to the limited extent of the areas affected by an elevation change. Conversely, for the same two predictions, the tpRMSE values are, respectively, 2.77 m and 6.04 m, highlighting the efficacy of the proposed metric in describing the errors in the 3D CD task.

In summary, the implemented augmentation techniques (except the radiometric transformation, as expected) contribute to the 3D CD metric improvement, and they also have a positive impact on the 2D CD metrics.

Combined augmentation techniques

In this analysis, we applied the most promising augmentation techniques of the previous analysis in combination with each other to evaluate if their combined use can lead to better CD performances than their individual use or the use of the baseline.

Specifically, we combined:

- CutMix with Gaussian Noise: we combined a CutMix strategy in the input bi-temporal images with 50 × 50 pixel crops with a Gaussian Noise in the 3D CD map with a $\sigma = 3$ m;
- Crop or Resize with Gaussian Noise: we combined a Random Crop or Resize with a 50% probability of cropping or resizing in the input bi-temporal images with a Gaussian Noise in the 3D CD map with a $\sigma = 3$ m.

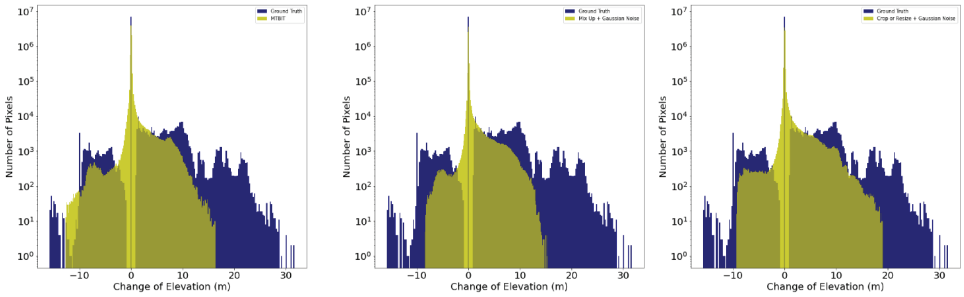
Table 3 reports the CD metrics achieved with the aforementioned augmentation combinations. As shown in Table 3, we achieved superior CD performance for both the combinations with respect to the baseline and the individual custom augmentations.

Specifically, when combining CutMix with Gaussian Noise – the first combination –, we observed a reduction of 35 cm in the cRMSE when compared to the baseline (an improvement of nearly 6%), along with a 15 cm reduction in the tpRMSE. In terms of 2D CD metrics, we observed approximately a 3% increase in the F1 score and IoU. Furthermore, when comparing this combination with the single augmentations, we noticed improvements in both the 3D and 2D CD metrics (Table 3b). However, the advantages of this combination are not evident from the analysis of the histogram of the elevation changes (Figure 6b). Nevertheless, a visual comparison of the ground truth CD maps with the predicted ones reveals excellent results (Figure 7b).

The second combination, Crop or Resize + Gaussian Noise, emerges as the most promising strategy: it significantly reduces the cRMSE value by 45 cm, achieving an improvement of about 7% with respect to the baseline. Moreover, the tpRMSE also decreased by approximately 30 cm. Concerning the 2D CD metrics, this combination did not achieve the same level of improvement as the 3D CD metrics: both the F1 score and IoU remain stable (Table 3c). The comparison of the elevation change distribution between the baseline and the one obtained with this combination shows significant improvements in both negative and positive elevation change values, especially in the range between 10 m and 20 m (Figure 6c). Visual evaluations of the predicted CD maps show substantial improvements on the baseline, particularly in challenging geometries (Figure 7c).

Table 3. Validation results of the combinations of different augmentation techniques applied to the 3D CD dataset using the MTBIT architecture.

		2D CDmetrics		3D CDmetrics		
Augmentation type	Figure 7	F1(%)	IoU(%)	RMSE(m)	cRMSE(m)	tpRMSE(m)
Baseline	a	62.80 ± 0.40	45.82 ± 0.40	1.18 ± 0.02	6.33 ± 0.06	5.60 ± 0.08
CutMix + Gaussian Noise	b	**65.67** ± 0.92	**49.73** ± 1.00	1.17 ± 0.01	5.98 ± 0.05	5.45 ± 0.07
Crop or Resize + Gaussian Noise	c	65.16 ± 0.52	48.97 ± 0.53	**1.16** ± 0.02	**5.88** ± 0.08	**5.34** ± 0.09



(a) Baseline (b) CutMix + Gaussian Noise (c) Crop or Resize + Gaussian Noise

Figure 6. Comparison between the distribution of the elevation change values in the ground truth (in blue) and the distribution of the predicted elevation change values obtained with the different augmentations reported in Table 3 (in yellow). The plot in a) refers to the results of the baseline, while b) and c) refer to the results achieved with the two different custom combined augmentations.

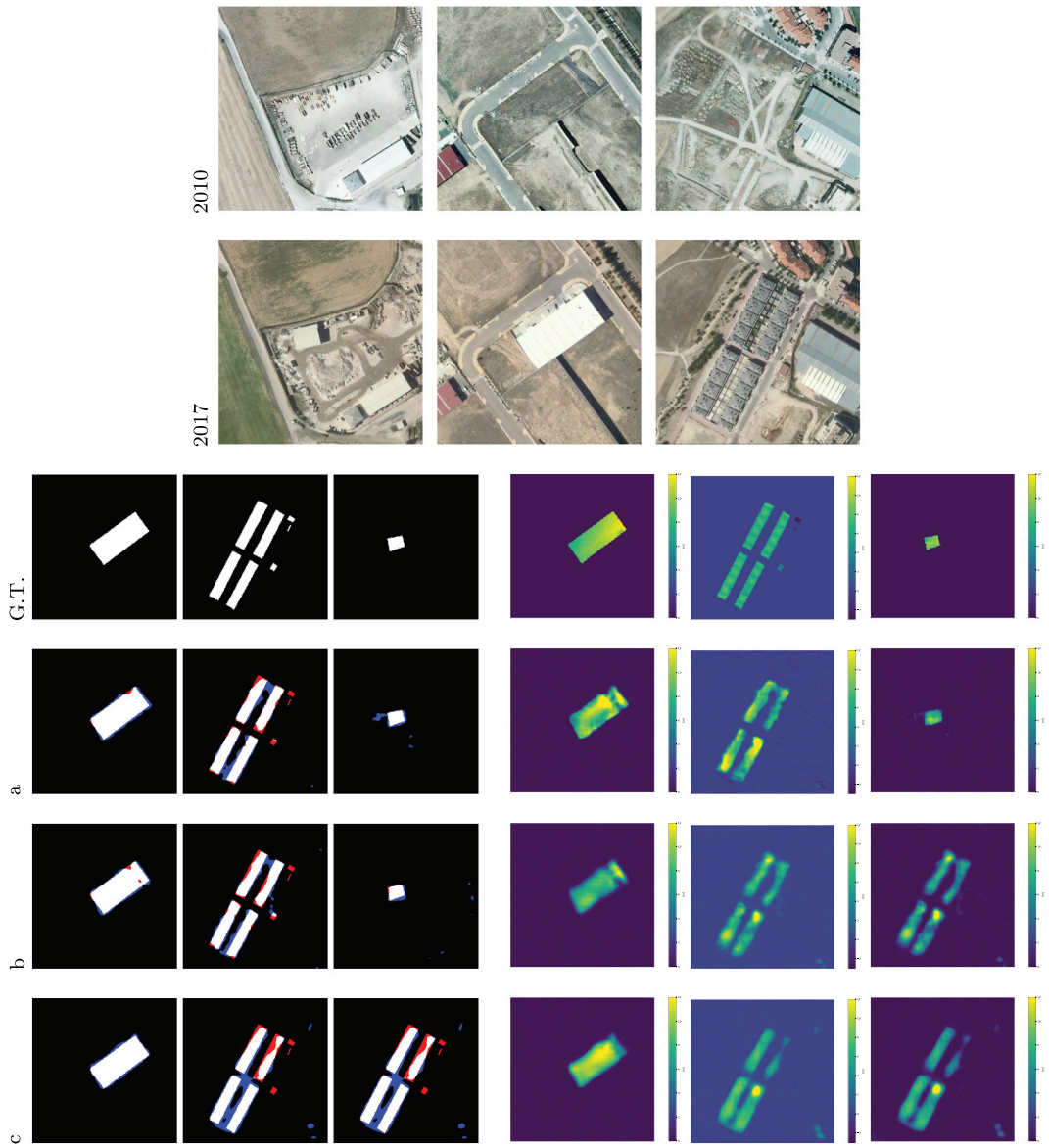


Figure 7. Visualisation of 2D CD and 3D CD maps for the augmentations described in Table 3. For the 2D change predictions, in black TN, in white TP, in red FN, in blue FP.

Ablation studies

As highlighted in the previous sections, the Gaussian noise on the 3D CD map, the CutMix, and the Crop or Resize are the best among the proposed augmentations.

Besides the Crop or Resize strategy, the efficiency of the other two augmentations relies in one case on the sigma value used for smoothing the 3D CD map and in the other case on the size and the number of crops that from one input image get pasted on the other. Tables 4 and 5 offer an overview of the ablation studies carried out on the sigma value for the Gaussian noise and on the number and size of crops for the CutMix strategy.

Investigating the augmentation performance on an additional 2D/3D CD architecture

To further evaluate the efficiency of the proposed set of custom augmentations and to assess their generality, we applied the combined augmentations introduced in Section 5.2 to an additional 2D/3D CD architecture.

Among the networks tested in (Marsocci, Coletta, et al., 2023), we selected the SUNet (Siamese ResU-Net) model, an architecture able to achieve performances comparable to the ones of MTBIT, but with more trainable parameters. SUNet is made by a Siamese encoder, a convolutional decoder and a pair of prediction heads; skip connections – like in UNet – are used to enhance the features (Marsocci, Gonthier, et al., 2023). In particular, we assessed the effects of the combined augmentations on SUNet18, the configuration of SUNet which employs two ResNet18 as encoders.

Table 6 provides the values of the evaluation metrics, proposed in Section 4.3, for SUNet18 with the two different combined augmentations in comparison with the corresponding baseline, where just the 256 × 256 resize augmentation is applied. The results show how the augmentation strategies work well also in the case of SUNet18, generally improving the values of the metrics with respect to the baseline and thus enhancing the performances of the network. Finally, comparing the results of SUNet18 (Table 6) with the corresponding results of MTBIT (Table 3), we can see that the overall improvement in the case of SUNet18 is lower than the improvement of MTBIT, which remains the best network for the 3D CD task also after the application of the proposed set of custom augmentations.

Moreover, to deeply understand the behavior of the investigated augmentations, we report the training loss curves for the best experiments with MTBIT and with SUNet18 in Figure 8. For MTBIT, the performance enhancements provided by the augmentations, already discussed in Section 5.2 and attested by the improvements of the evaluation metrics, are visible also in Figure 8, where the training loss values of the augmented MTBIT versions decrease more rapidly than the corresponding baseline. Regarding SUNet18, the metric improvement reported in Table 6 is not so visible in the training loss curves. The Cut Mix strategy in combination with the Gaussian Noise on the 3D masks shows a trend almost comparable to the baseline, especially after the first 100 epochs. The Crop/Resize strategy in combination

Table 4. Ablation studies on sigma values (Gaussian noise on the 3D CD map).

	2D CDmetrics		3D CDmetrics		
Sigma value (m)	F1(%)	IoU(%)	RMSE(m)	cRMSE(m)	tpRMSE(m)
1	59.90 ± 0.58	43.57 ± 0.60	1.25 ± 0.01	6.30 ± 0.01	5.65 ± 0.04
3	**61.74 ± 0.55**	**44.61 ± 0.56**	**1.18 ± 0.01**	**6.16 ± 0.03**	**5.51 ± 0.06**
4	60.23 ± 0.14	43.09 ± 0.28	1.24 ± 0.02	6.28 ± 0.04	5.60 ± 0.14

Table 5. Ablation study on number and size of cuts for the CutMix strategy.

		2D CDmetrics		3D CDmetrics		
Cuts number	Size [pixels]	F1(%)	IoU(%)	RMSE(m)	cRMSE(m)	tpRMSE(m)
10	25×25	60.22 ± 1.04	43.09 ± 1.07	**1.26** ± 0.02	6.35 ± 0.12	5.65 ± 0.12
5	**50×50**	**62.63 ± 1.14**	**45.61 ± 1.17**	**1.18 ± 0.02**	**6.15 ± 0.06**	**5.47 ± 0.03**

Table 6. Test results of the combinations of different augmentation techniques applied to the 3DCD dataset using the SUNet18 architecture.

	2D CDmetrics		3D CDmetrics		
Augmentation type	F1(%)	IoU(%)	RMSE(m)	cRMSE(m)	tpRMSE(m)
Baseline	57.89	40.73	1.22	6.69	5.92
CutMix + Gaussian Noise	**60.95**	**43.83**	1.23	**6.41**	6.05
Crop or Resize + Gaussian Noise	57.99	39.86	**1.22**	6.48	**5.84**

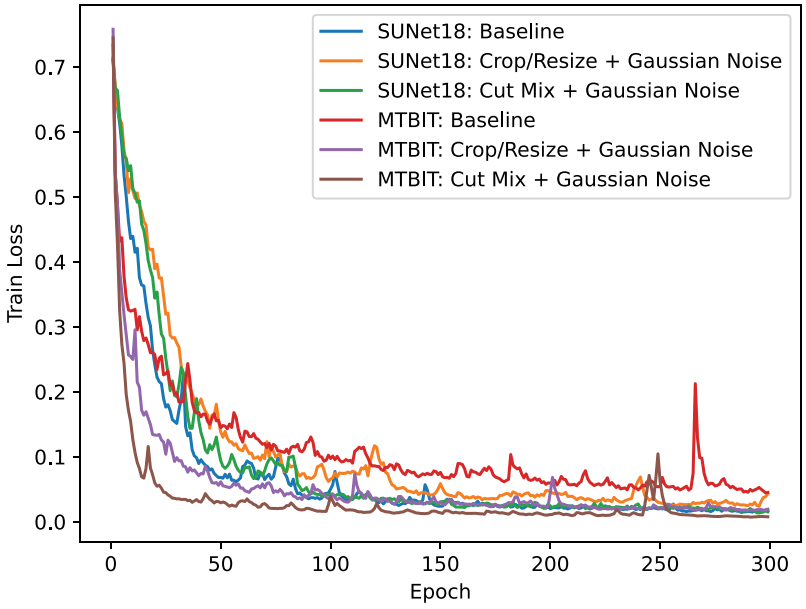


Figure 8. Loss curves comparison.

with the Gaussian Noise on the 3D masks seems to have a slower loss decrease that eventually reaches loss values comparable to the ones obtained using other augmentation strategies. In any case, there are two aspects that should be considered: i) the metrics show an overall improvement, thanks to the proposed augmentations; ii) the loss curve can be an interesting tool to understand how the training process evolves, but it is not exhaustive when judging the final performance of the network. In fact, smaller loss values can indicate an overfitting on the training set, a behavior that we want to avoid, also through the proposed augmentations.

Adding standard augmentation

Finally, we tested the proposed set of augmentation techniques in combination with the standard augmentations already used in (Marsocci, Coletta, et al., 2023): resizing to 256 × 256 (the only one already considered in the baseline, see Section 5), random horizontal flip, random geometric transformation (i.e. shifting, scaling and rotating), addition of random Gaussian noise on the input bitemporal images, and random radiometric transformation (i.e. brightness, sharpening, blurring, contrasting, saturation). These additional standard augmentations will be referred to as *Augmentation* in the remainder of the paper. To evaluate the results, we focused only on the 3D metrics: RMSE, cRMSE and tpRMSE.

Table 7b shows how the addition of standard augmentations alone led to small improvements with respect to the baseline, specifically, a reduction of 11 cm and of almost 10 cm in cRMSE and tpRMSE, respectively.

However, the metric improvements achieved with the application of the standard augmentations (Table 7) combined with the custom ones are lower than the ones obtained only using the set of combined custom augmentations introduced in this study (Table 3). One possible reason could be that the network gets overwhelmed or confused when exposed to various forms of diverse augmentation simultaneously. Thus, while data augmentation is essential, it must be carried out without interfering with the essential features of the bi-temporal images.

Exploring zero-shot capabilities

Table 8 presents the results of the application of the combined augmentations in a zero-shot scenario, i.e.

Table 7. Validation results of the different augmentation techniques with standard augmentation applied to the 3DCD dataset.

		3D CDmetrics		
Augmentation type	Figure 9	RMSE(m)	cRMSE(m)	tpRMSE(m)
Baseline	A	1.18 ± 0.02	6.33 ± 0.06	5.60 ± 0.08
Baseline with *Augmentation* (Marsocci, Gonthier, et al., 2023)	B	1.20 ± 0.02	6.22 ± 0.08	5.53 ± 0.10
Random Crop or Resize with *Augmentation*	C	1.17 ± 0.01	6.29 ± 0.12	5.39 ± 0.12
Mixup with *Augmentation*	D	1.21 ± 0.01	6.46 ± 0.13	5.56 ± 0.09
Change-Guided Random Crop with *Augmentation*	E	1.19 ± 0.01	6.39 ± 0.16	5.58 ± 0.10
Gaussian Noise with *Augmentation*	F	1.20 ± 0.02	6.24 ± 0.03	5.57 ± 0.07
CutMix with Gaussian Noise with *Augmentation*	G	1.17 ± 0.02	6.34 ± 0.04	5.52 ± 0.11
Crop or Resize with Gaussian Noise with *Augmentation*	h	1.18 ± 0.01	6.41 ± 0.04	5.54 ± 0.12

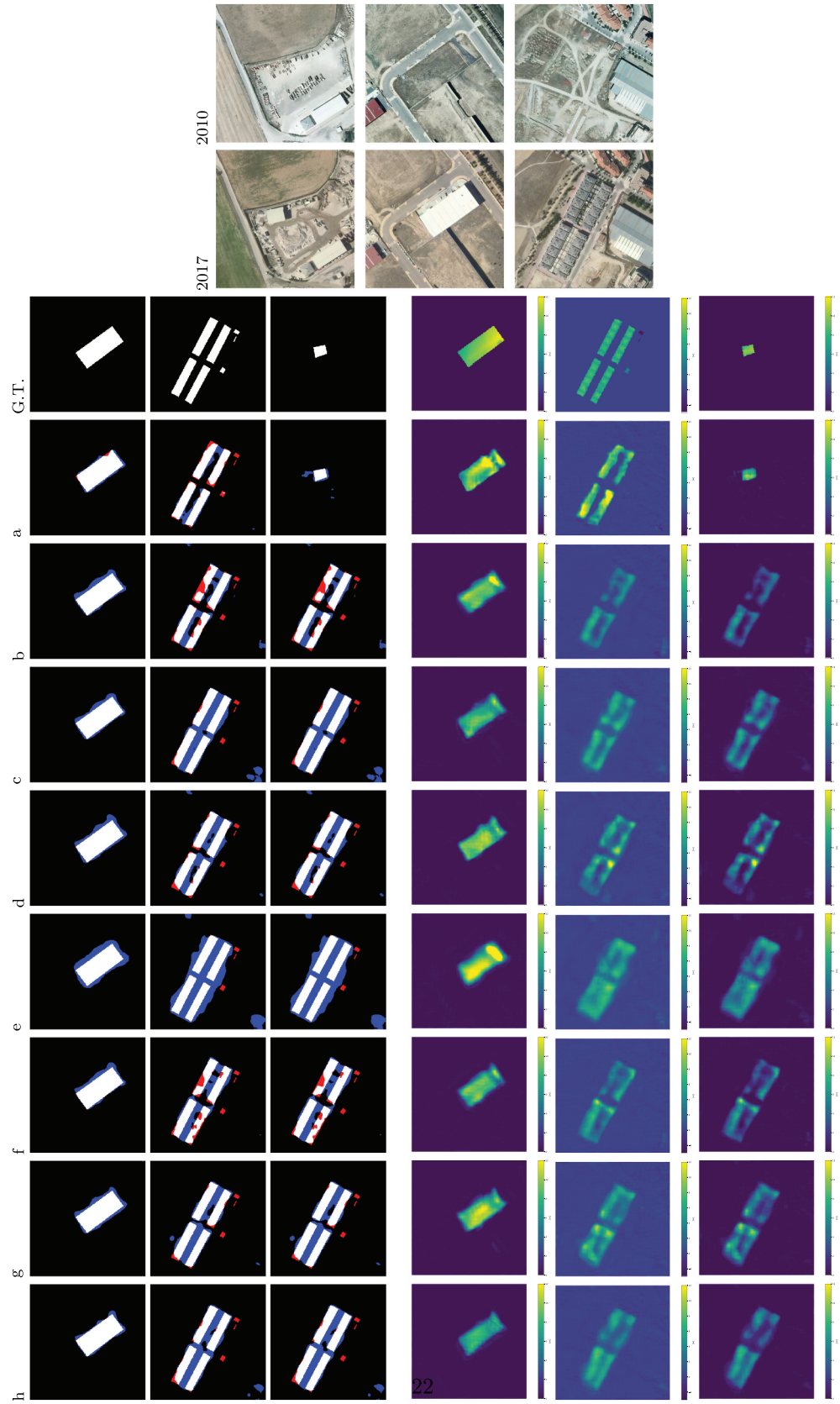


Figure 9. Visualisation of 2D CD and 3D CD maps for the augmentations described in Table 7.

performing inference on the new Avila dataset, to test the zero-shot capabilities of the proposed augmentation strategies. While the improvements are not substantial, they attest initial progress in the pursuit of domain generalization.

The combination of CutMix with Gaussian Noise on the 3D CD map demonstrates improvements in the 2D CD metrics (e.g. F1 scores of 40.60%) but not significant progress in the 3D CD metrics.

Table 8. Results of the combination of the proposed custom augmentation strategies on the new avila dataset.

	2D CD metrics		3D CD metrics		
Augmentation type	F1%)	IoU (%)	RMSE (m)	cRMSE (m)	tcRMSE (m)
Baseline	34.36 ± 2.31	22.58 ± 3.99	0.86 ± 0.02	7.74 ± 0.21	6.68 ± 0.15
CutMix + Gaussian Noise	40.60 ± 3.65	25.53 ± 2.84	0.80 ± 0.02	**7.07 ± 0.09**	6.79 ± 0.12
Crop or Resize + Gaussian Noise	**44.94 ± 2.16**	**29.01 ± 1.80**	**0.78 ± 0.01**	**7.07 ± 0.11**	**6.61 ± 0.14**

More substantial improvements are observed with Crop or Resize combined with Gaussian Noise on the 3D CD map. This combination yields the best 2D CD results (e.g. F1 score of 44.94%). Moreover, there are notable improvements in the 3D CD metrics, with the lowest RMSE (0.78 m), cRMSE (7.07 m), and tcRMSE (6.61 m) values among the presented strategies.

It is crucial to emphasize that these improvements, while promising, are preliminary steps toward achieving zero-shot learning. Further refinement and exploration of augmentation techniques, as well as model optimization, may contribute to more substantial advancements in the future.

Conclusions

Detecting changes in all their aspects (2D, 3D) are essential for monitoring transformations of the Earth's surface. In this work, we proposed a new set of custom augmentations to improve the performances of MTBIT (Marsocci, Coletta, et al., 2023), a deep learning architecture recently developed to simultaneously solve the 2D and 3D CD tasks from bi-temporal optical images.

Specifically, several data augmentation strategies (Change-Guided Random Crop, Crop or Resize, CutMix, Gaussian Noise, Radiometric Transformation on the buffer zone) were implemented and applied both individually and in specific combinations (CutMix + Gaussian Noise and Crop or Resize + Gaussian Noise) during the training on the 3D CD dataset. Furthermore, standard augmentations (e.g. horizontal and vertical flips) were applied to all the aforementioned techniques, too.

Finally, a new metric, the tpRMSE (True Positive RMSE), was introduced in addition to the 3D CD metrics employed in (Marsocci, Gonthier, et al., 2023) (RMSE and cRMSE) for a more comprehensive evaluation of MTBIT performances in the 3D CD task.

The results obtained are encouraging. The quantitative evaluation based on the 3D CD metrics shows that three of the four proposed augmentation strategies have a positive impact on the reduction of the cRMSE and tpRMSE values with respect to the baseline. Moreover, most of the proposed augmentation strategies show a better agreement of the predicted 3D change values with the ground truth histogram compared to the baseline.

The only augmentation technique that resulted in a significant degradation of the metrics (both 2D and 3D) is the Radiometric Transformation on the buffer zone, as expected. This technique indeed applies a radiometric transformation on the input bi-temporal images in correspondence to the borders of the changed areas. Thus, the degradation in performance of this augmentation can be attributed to the perturbation of the shadow areas around elevated objects (e.g. buildings), which are essential for the network to learn the elevation changes.

Finally, the addition of standard augmentations to the proposed set of custom augmentations did not produce the desired results: the 3D CD metrics in this case were worse than the results obtained considering the proposed augmentations alone.

In summary, the most successful solution was the combination of Crop or Resize strategy with Gaussian Noise on the 3D CD map, which reduced cRMSE to 5.88 m and tpRMSE to 5.34 m, from 6.33 m and 5.60 m of the baseline, respectively. This combination also provided a significant improvement in the 2D CD metrics: the F1 score increased from 62.80% of the baseline to 65.16% and the IoU score from 45.82% to 48.97%. The qualitative assessment performed by visually comparing the predicted CD maps with ground truth maps also showed encouraging results. Nevertheless, artifacts characterized by small elevation values are still not accurately predicted.

Finally, some first zero-shot learning experiments were also carried out on a new small dataset, achieving not substantial but promising improvements towards domain generalization.

In conclusion, the proposed contributions enhance the practical utility and reliability of MTBIT in real-world applications, addressing critical challenges in the field of Remote Sensing CD. However, algorithmic improvements can in the future be considered using larger datasets, such as SMARS (Fuentes Reyes et al., 2023).

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Asadi Shamsabadi, E., Xu, C., Rao, A. S., Nguyen, T., Ngo, T., & da Costa, D. D. (2022). Vision transformer-based autonomous crack detection on asphalt and concrete surfaces. *Automation in Construction*, *140*, 104316. https://doi.org/10.1016/j.autcon.2022.104316

Audebert, N., Le Saux, B., & Lefèvre, S. (2017). Joint learning from earth observation and OpenStreetMap data to get faster better semantic maps. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1552–1560). https://doi.org/10.1109/CVPRW.2017.199

Bai, T., Wang, L., Yin, D., Sun, K., Chen, Y., Li, W., & Li, D. (2022). Deep learning for change detection in remote sensing: A review. *Geo-Spatial Information Science*, *8*(9), 1–27. https://doi.org/10.1080/10095020.2022.2085633

Baldi, P., Fabris, M., Marsella, M., & Monticelli, R. (2005). Monitoring the morphological evolution of the sciara del fuoco during the 2002–2003 stromboli eruption using multi-temporal photogrammetry. *Isprs Journal of Photogrammetry & Remote Sensing*, *59*(4), 199–211. https://doi.org/10.1016/j.isprsjprs.2005.02.004

Bandara, W. G. C., Nair, N. G., & Patel, V. M. (2022). *Ddpm-cd : Remote sensing change detection using denoising diffusion probabilistic models. arXiv preprint arXiv: 2206.11892.v2.* https://doi.org/10.48550/arXiv.2206.11892

Bandara, W. G. C., & Patel, V. M. (2022). A transformer-based siamese network for change detection. *IGARSS, 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium* (pp. 207–210). https://doi.org/10.1109/IGARSS46834.2022.9883686

Bendig, J., Bolten, A., & Bareth, G. (2013, 12). Uav-based imaging for multi-temporal, very high resolution crop surface models to monitor crop growth variability. *Photogrammetrie - Fernerkundung - Geoinformation*, *2013* (6), 551–562. https://doi.org/10.1127/1432-8364/2013/0200

Bing Liu, B., Yu, A., Zuo, X., Wang, R., Qiu, C., & Yu, X. (2023). Deep hierarchical transformer for change detection in high-resolution remote sensing images. *European Journal of Remote Sensing*, *56*(1), 2196641. https://doi.org/10.1080/22797254.2023.2196641

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, *11* (2), 125. https://doi.org/10.3390/info11020125

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). *Emerging properties in self-supervised vision transformers. arXiv preprint arXiv: 2104.14294.* https://doi.org/10.48550/arXiv.2104.14294

Chen, C., Ma, H., Yao, G., Lv, N., Yang, H., Li, C., & Wan, S. (2021). Remote sensing image augmentation based on text description for waterside change detection. *Remote Sensing*, *13*(10), 1894. https://doi.org/10.3390/rs13101894

Chen, H., Qi, Z., & Shi, Z. (2022). Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience & Remote Sensing*, *60*, 1–14. https://doi.org/10.1109/TGRS.2021.3095166

Chen, H., Song, J., Wu, C., Du, B., & Yokoya, N. (2023). Exchange means change: An unsupervised single-temporal change detection framework based on intra- and inter-image patch exchange. *Isprs Journal of Photogrammetry & Remote Sensing*, *206*, 87–105. https://doi.org/10.1016/j.isprsjprs.2023.11.004

Chen, H., Yokoya, N., & Chini, M. (2023). Fourier domain structural relationship analysis for unsupervised multimodal change detection. *Isprs Journal of Photogrammetry & Remote Sensing*, *198*, 99–114. https://doi.org/10.1016/j.isprsjprs.2023.03.004

Chen, H., Yokoya, N., Wu, C., & Du, B. (2022). Unsupervised multimodal change detection based on structural relationship graph representation learning. *IEEE Transactions on Geoscience & Remote Sensing*, *60*, 1–18. https://doi.org/10.1109/TGRS.2022.3229027

Chen, K., Liu, C., Li, W., Liu, Z., Chen, H., Zhang, H., & Shi, Z. (2023). *Time travelling pixels: Bitemporal features integration with foundation model for remote sensing image change detection.* https://doi.org/10.48550/arXiv.2312.16202

Chen, K., Zou, Z., & Shi, Z. (2021). Building extraction from remote sensing images with sparse token transformers. *Remote Sensing*, *13*(21), 4441. https://doi.org/10.3390/rs13214441

Coletta, V., Marsocci, V., & Ravanelli, R. (2022). 3DCD: A new dataset for 2D and 3D change detection using deep learning techniques. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII- B3-2022*, 1349–1354. https://doi.org/10.5194/isprs-archives-XLIII-B3-2022-1349-2022

Daudt, R. C., Le Saux, B., & Boulch, A. (2018). Fully convolutional siamese networks for change detection. *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 4063–4067). https://doi.org/10.1109/ICIP.2018.8451652

Deuser, F., Habel, K., & Oswald, N. (2023). *Sample4geo: Hard negative sampling for cross-view geo-localisation. arXiv preprint arXiv: 2303.11851.* https://doi.org/10.48550/arXiv.2303.11851

Ding, L., Zhu, K., Peng, D., Tang, H., Yang, K., & Bruzzone, L. (2024). Adapting segment anything model for change detection in vhr remote sensing images. *IEEE Transactions on Geoscience & Remote Sensing*, *62*, 1–11. https://doi.org/10.1109/tgrs.2024.3368168

Dong, S., Wang, L., Du, B., & Meng, X. (2024). Changeclip: Remote sensing change detection with multimodal vision-language representation learning. *Isprs Journal of Photogrammetry & Remote Sensing*, *208*, 53–69. https://doi.org/10.1016/j.isprsjprs.2024.01.004

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv. https://doi.org/10.48550/ARXIV.2010.11929

Emmanuel Fundisi, E., Tesfamichael, S. G., & Ahmed, F. (2022). A combination of sentinel-1 RADAR and Sentinel-2 multispectral data improves classification of morphologically similar savanna woody plants. *European Journal of Remote Sensing*, *55*(1), 372–387. https://doi.org/10.1080/22797254.2022.2083984

Fuentes Reyes, M., Xie, Y., Yuan, X., d'Angelo, P., Kurz, F., Cerra, D., & Tian, J. (2023). A 2D/3D multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection. *Isprs Journal of Photogrammetry & Remote Sensing*, *205*, 74–97. https://doi.org/10.1016/j.isprsjprs.2023.09.013

Ghasem Abdi, G., Samadzadegan, F., & Reinartz, P. (2017). A decision-based multi-sensor classification system using thermal hyperspectral and visible data in urban area. *European Journal of Remote Sensing*, *50*(1), 414–427. https://doi.org/10.1080/22797254.2017.1348914

Ghuffar, S., Székely, B., Roncat, A., & Pfeifer, N. (2013). Landslide displacement monitoring using 3D range flow

on airborne and terrestrial LiDAR data. *Remote Sensing*, *5* (6), 2720–2745. https://doi.org/10.3390/rs5062720

Hafner, S., Ban, Y., & Nascetti, A. (2022). Unsupervised domain adaptation for global urban extraction using sentinel-1 SAR and Sentinel-2 MSI data. *Remote Sensing of Environment*, *280*, 113192. https://doi.org/10.1016/j.rse.2022.113192

Hafner, S., Nascetti, A., Azizpour, H., & Ban, Y. (2022). Sentinel-1 and sentinel-2 data fusion for urban change detection using a dual stream U-Net. *IEEE Geoscience & Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2021.3119856

Hao, X., Liu, L., Yang, R., Yin, L., Zhang, L., & Li, X. (2023). A review of data augmentation methods of remote sensing image target recognition. *Remote Sensing*, *15*(3), 827. https://doi.org/10.3390/rs15030827

Haut, J. M., Paoletti, M. E., Plaza, J., Plaza, A., & Plaza, L. (2019). Hyperspectral image classification using random occlusion data augmentation. *IEEE Geoscience & Remote Sensing Letters*, *16*(11), 1751–1755. https://doi.org/10.1109/LGRS.2019.2909495

Hayet, S. S., Sally, E. G., Abdelmounaam, R., Samy, A.-A., & Nour, E. I. B. (2020). What is a remote sensing change detection technique? Towards a conceptual framework. *International Journal of Remote Sensing*, *41*(5), 1788–1812. https://doi.org/10.1080/01431161.2019.1674463

Hong, D., Zhang, B., Li, H., Li, Y., Yao, J., Li, C., Werner, M., Chanussot, J., Zipf, A., & Zhu, X. X. (2023). Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sensing of Environment*, *299*, 113856. https://doi.org/10.1016/j.rse.2023.113856

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., & Girshick, R. (2023). *Segment anything*. https://doi.org/10.48550/arXiv.2304.02643

Lalitha, V., & Latha, B. (2022). A review on remote sensing imagery augmentation using deep learning. *Materials Today: Proceedings* (Vol. 62. pp. 4772–4778). (International Conference on Innovative Technology for Sustainable Development). https://doi.org/10.1016/j.matpr.2022.03.341

Leenstra, M., Marcos, D., Bovolo, F., & Tuia, D. (2021). Self-supervised pre-training enhances change detection in sentinel-2 imagery. *International conference on pattern recognition* (pp. 578–590). https://doi.org/10.1007/978-3-030-68787-8˙426

Liu, M., Chai, Z., Deng, H., & Liu, R. (2022). A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, *15*, 4297–4306. https://doi.org/10.1109/JSTARS.2022.3177235

Loshchilov, I., & Hutter, F. (2019). 7th International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=Bkg6RiCqY7.

Lu, D., Mausel, P., Brondízio, E., & Moran, E. (2004). Change detection techniques. *International Journal of Remote Sensing*, *25*(12), 2365–2401. https://doi.org/10.1080/0143116031000139863

Lv, N., Ma, H., Chen, C., Pei, Q., Zhou, Y., Xiao, F., & Li, J. (2021). Remote sensing data augmentation through adversarial training. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, *14*, 9318–9333. https://doi.org/10.1109/JSTARS.2021.3110842

Malpica, J. A., Alonso, M. C., Papí, F., Arozarena, A., & Agirre, A. M. D. (2013). Change detection of buildings from satellite imagery and lidar data. *International Journal of Remote Sensing*, *34*(5), 1652–1675. https://doi.org/10.1080/01431161.2012.725483

Marsocci, V., Coletta, V., Ravanelli, R., Scardapane, S., & Crespi, M. (2023). Inferring 3D change detection from bitemporal optical images. *Isprs Journal of Photogrammetry & Remote Sensing*, *196*(23), 325–339. https://doi.org/10.1016/j.isprsjprs.2022.12.009

Marsocci, V., Gonthier, N., Garioud, A., Scardapane, S., & Mallet, C. (2023). Geomultitasknet: Remote sensing unsupervised domain adaptation using geo-graphical coordinates. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (Vol. 5. pp. 2075–2085). https://doi.org/10.1109/CVPRW59228.2023.00201

Marsocci, V., Scardapane, S., & Komodakis, N. (2021). MARE: Self-supervised multi-attention REsu-net for semantic segmentation in remote sensing. *Remote Sensing*, *13*(16), 3275. https://doi.org/10.3390/rs13163275

Menderes, A., Erener, A., & Sarp, G. (2015). Automatic detection of damaged buildings after earthquake hazard by using remote sensing and information technologies. *Procedia Earth and Planetary Science*, *15*(6), 257–262. (World Multidisciplinary Earth Sciences Symposium, WMESS 2015). https://doi.org/10.1016/j.proeps.2015.08.063

Moieez, H., Marsocci, V., & Scardapane, S. (2023). Continual self-supervised learning in earth observation with embedding regularization. *Igarss 2023 - 2023 ieee international geoscience and remote sensing symposium* (Vol. 5. pp. 5029–5032). https://doi.org/10.1109/IGARSS52108.2023.10283121

Oubara, A., Wu, F., Amamra, A., & Yang, G. (2022). Survey on remote sensing data augmentation: Advances, challenges, and future perspectives. In M.R. Senouci, S. Y. Boulahia, & M.A. Benatia (Eds.), *Advances in computing systems and applications* (Vol. 6, pp. 95–104). Springer International Publishing. https://doi.org/10.1007/978-3-031-12097-8˙9

PyTorch. (n.d.). *Albumentations*. https://albumentations.ai/docs/examples/pytorchsemanticsegmentation/

Qin, R., Tian, J., & Reinartz, P. (2016). 3D change detection – approaches and applications. *Isprs Journal of Photogrammetry & Remote Sensing*, *122*(3), 41–56. https://doi.org/10.1016/j.isprsjprs.2016.09.013

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., & Sutskever, I. (2021, July 18–24). Learning transferable visual models from natural language supervision. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (Vol. 139. pp. 8748–8763). PMLR. https://doi.org/10.48550/arXiv.2103.00020

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., & Breitkopf, U. (2012). The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, *I-3*(5), 293–298. https://doi.org/10.5194/isprsannals-I-3-293-2012

Shafique, A., Cao, G., Khan, Z., Asad, M., & Aslam, M. (2022). Deep learning-based change detection in remote sensing images: A review. *Remote Sensing*, *14*(4), 871. https://doi.org/10.3390/rs14040871

Shawky, O. A., Hagag, A., El-Dahshan, E.-S. A., & Ismail, M. A. (2020). Remote sensing image scene

classification using cnn-mlp with data augmentation. *Optik*, *221*(6), 165356. https://doi.org/10.1016/j.ijleo.2020.165356

Shi, Q., Liu, M., Li, S., Liu, X., Wang, F., & Zhang, L. (2022). A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Transactions on Geoscience & Remote Sensing*, *60*(6), 1–16. https://doi.org/10.1109/TGRS.2021.3085870

Shirowzhan, S., Sepasgozar, S. M., Li, H., Trinder, J., & Tang, P. (2019). Comparative analysis of machine learning and point-based algorithms for detecting 3D changes in buildings over time using bi-temporal lidar data. *Automation in Construction*, *105*(6), 102841. https://doi.org/10.1016/j.autcon.2019.102841

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Shuwen Xu, S., Liao, Y., Yan, X., & Zhang, G. (2020). Change detection in sar images based on iterative otsu. *European Journal of Remote Sensing*, *53*(1), 331–339. https://doi.org/10.1080/22797254.2020.1852606

Song, J., Chen, H., & Yokoya, N. (2023). *Syntheworld: A large-scale synthetic dataset for land cover mapping and building change detection*. https://doi.org/10.48550/arXiv.2309.01907

Stilla, U., & Xu, Y. (2023). Change detection of urban objects using 3D point clouds: A review. *Isprs Journal of Photogrammetry & Remote Sensing*, *197*(6), 228–255. https://doi.org/10.1016/j.isprsjprs.2023.01.010

Sun, X., Wang, B., Wang, Z., Li, H., Li, H., & Fu, K. (2021). Research progress on few-shot learning for remote sensing image interpretation. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, *14*(7), 2387–2402. https://doi.org/10.1109/JSTARS.2021.3052869

Sun, Y., Lei, L., Li, X., Sun, H., & Kuang, G. (2021). Nonlocal patch similarity based heterogeneous remote sensing change detection. *Pattern Recognition*, *109*(5), 107598. https://doi.org/10.1016/j.patcog.2020.107598

Sun, Y., Lei, L., Li, X., Tan, X., Kuang, G., Li, X., Zhang, B., & Plaza, A. (2022). Multisource data reconstruction-based deep unsupervised hashing for unisource remote sensing image retrieval. *IEEE Transactions on Geoscience & Remote Sensing*, *60*(5), 1–16. https://doi.org/10.1109/TGRS.2022.3231215

Taneja, A., Ballan, L., & Pollefeys, M. (2011). Image based detection of geometric changes in urban environments. *2011 International Conference on Computer Vision* (Vol. 6. pp. 2336–2343). https://doi.org/10.1109/ICCV.2011.6126515

Wang, Z., Liu, D., Wang, Z., Liao, X., & Zhang, Q. (2023). A new remote sensing change detection data augmentation method based on mosaic simulation and haze image simulation. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, *16*(7), 4579–4590. https://doi.org/10.1109/JSTARS.2023.3269784

Xiao, Q., Liu, B., Li, Z., Ni, W., Yang, Z., & Li, L. (2021). Progressive data augmentation method for remote sensing ship image classification based on imaging simulation system and neural style transfer. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, *14*(6), 9176–9186. https://doi.org/10.1109/JSTARS.2021.3109600

Xingrui Yu, X., Wu, X., Luo, C., & Ren, P. (2017). Deep learning in remote sensing scene classification: A data augmentation enhanced convolutional neural network framework. *GIScience & Remote Sensing*, *54*(5), 741–758. https://doi.org/10.1080/15481603.2017.1323377

Xu, B., Zeng, Z., Lian, C., & Ding, Z. (2022). Generative mixup networks for zero-shot learning. *IEEE Transactions on Nesural Networks and Learning Systems*, *7*. https://doi.org/10.1109/TNNLS.2022.3142181

Xu, M., Yoon, S., Fuentes, A., & Park, D. S. (2023). A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, *137*(3), 109347. https://doi.org/10.1016/j.patcog.2023.109347

Yan, Y., Zhang, Y., & Su, N. (2019). A novel data augmentation method for detection of specific aircraft in remote sensing rgb images. *Institute of Electrical and Electronics Engineers Access*, *7*(6), 56051–56061. https://doi.org/10.1109/ACCESS.2019.2913191

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., & Yang, M.-H. (2024). *Diffusion models: A comprehensive survey of methods and applications*. https://doi.org/10.48550/arXiv.2209.00796

Zhang, L., & Zhang, L. (2022). Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, *10*(2), 270–294. https://doi.org/10.1109/MGRS.2022.3145854

Zheng, Z., Zhong, Y., Zhang, L., & Ermon, S. (2024). Segment any change. *arXiv Preprint arXiv: 2402.01188*. https://doi.org/10.48550/arXiv.2402.01188

Zhu, S., Jing, W., Kang, P., Emam, M., & Li, C. (2023a). Data augmentation and few-shot change detection in forest remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, *16*(4), 5919–5934. https://doi.org/10.1109/JSTARS.2023.3285389

Zhu, S., Jing, W., Kang, P., Emam, M., & Li, C. (2023b). Data augmentation and few-shot change detection in forest remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, *16*(7), 5919–5934. https://doi.org/10.1109/JSTARS.2023.3285389

Appendix. Algorithms

In this Appendix, we report the pseudo-codes of the combined augmentations and the augmentations that achieved the best performance. Specifically, Algorithm 2 reports the CutMix + Gaussian Noise augmentation, while Algorithm 3 reports the Crop or Resize + Gaussian Noise augmentation.

Algorithm 2: CutMix+Gaussian Noise augmentation

Data: RGB image at time 0 im_{t0}, RGB image at time 1 im_{t1}, 2D ground truth $m2d$, 3D ground truth $m3d$
Input: Number of cuts N_C, size of the cut S_C, magnitude of the filtering σ

/* CutMix */
$h, w, c = shape(im_{t0})$; // take the shape of the images
for i *in* N^C*;* // iterate on cuts
do
  while *True* **do**
    $x, y \leftarrow randint(0, h - S_C), randint(0, w - S_C)$; // sampling where to cut
    $m2d^C \leftarrow m2d[y : y + S_C, x : x + S_C]$; // create 2D mask crop
    if $(count(m2d^C) == 1) > 0$
    tcp[r]check if there is a change* **then**
      $\text{im}_{t0}^C \leftarrow im_{t0}[y : y + S_C, x : x + S_C, :]$; // crop on t0 image
      $\text{im}_{t1}^C \leftarrow im_{t1}[y : y + S_C, x : x + S_C, :]$; // crop on t1 image
      $\text{im}_{t0}[y : y + S_C, x : x + S_C, :] \leftarrow im_{t1}^C$; // augmented t0 image
      $\text{im}_{t1}[y : y + S_C, x : x + S_C] \leftarrow im_{t1}^C$; // augmented t1 image
      break
    else
      sample another x and y
    end
  end
end
/* Gaussian Noise */
m3d $\leftarrow$ gaussian-filter(m3d, σ)

Algorithm 3: Crop or Resize+Gaussian Noise augmentation

Data: RGB image at time 0 im_{t0}, RGB image at time 1 im_{t1}, 2D ground truth $m2d$, 3D ground truth $m3d$
Input: w_C, h_C, minimum number of changed pixel n_C, magnitude of the filtering σ

/* Crop or Resize */
$h, w, c \leftarrow= shape(im_{t0})$; // take the shape of the images
$p \leftarrow N(0, 1)$ **if** $p > 0.5$ **then**
  /* Resize */
  $im_{t0} \leftarrow \text{resize}(im_{t0}, \text{size} = (w_C, h_C))$
  $im_{t1} \leftarrow \text{resize}(im_{t1}, \text{size} = (w_C, h_C))$
  $m2d \leftarrow \text{resize}(m2d, \text{size} = (w_C, h_C))$
  $m3d \leftarrow \text{resize}(m3d, \text{size} = (w_C, h_C))$
else
  /* Crop */
  while *True* **do**
    $x, y \leftarrow randint(0, h - h_C), randint(0, w - w_C)$; // sampling where to cut
    $m2d \leftarrow m2d[y : y + h_C, x : x + w_C]$; // create 2D mask crop
    if $(count(m2d) == 1) > n_C$*;* // check how many changed pixels
    then
      $\text{im}_{t0} \leftarrow im_{t0}[y : y + h_C, x : x + w_C, :]$
      $\text{im}_{t1} \leftarrow im_{t1}[y : y + h_C, x : x + w_C, :]$
      $m3d^C \leftarrow m3d[y : y + h_C, x : x + w_C, :]$
      break
    else
      resample another x and y
    end
  end
end
/* Gaussian Noise */
m3d $\leftarrow$ gaussian-filter(m3d, σ)