Transformer-based Multimodal Change Detection with Multitask Consistency Constraints

Biyuan Liu[a], Huaixin Chen[a], Kun Li[b], Michael Ying Yang[b*]
[a] University of Electronic Science and Technology of China
[b] University of Twente
lby9469@gmail.com, huaixinchen@uestc.edu.cn, {k.li, michael.yang}@utwente.nl

***Abstract*—Change detection plays a fundamental role in Earth observation for analyzing temporal iterations over time. However, recent studies have largely neglected the utilization of multimodal data that presents significant practical and technical advantages compared to single-modal approaches. This research focuses on leveraging pre-event digital surface model (DSM) data and post-event digital aerial images captured at different times for detecting change beyond 2D. We observe that the current change detection methods struggle with the multitask conflicts between semantic and height change detection tasks. To address this challenge, we propose an efficient Transformer-based network that learns shared representation between cross-dimensional inputs through cross-attention. It adopts a consistency constraint to establish the multimodal relationship. Initially, pseudo-changes are derived by employing height change thresholding. Subsequently, the $L2$ distance between semantic and pseudo-changes within their overlapping regions is minimized. This explicitly endows the height change detection (regression task) and semantic change detection (classification task) with representation consistency. A DSM-to-image multimodal dataset encompassing three cities in the Netherlands was constructed. It lays a new foundation for beyond-2D change detection from cross-dimensional inputs. Compared to five state-of-the-art change detection methods, our model demonstrates consistent multitask superiority in terms of semantic and height change detection. Furthermore, the consistency strategy can be seamlessly adapted to the other methods, yielding promising improvements.**

***Index Terms*—change detection, multimodal, height change, multitask consistency, Transformer-based.**

I. Introduction

The field of change detection is undergoing a significant evolution characterized by higher temporal frequencies [1], finer-grained analyses [2], and increased dimensionality [3], [4]. Recent advancements in Earth observation techniques have enabled daily change detection [1] and fine-grained analysis spanning up to nine distinct change categories [2]. Moreover, there have been exciting breakthroughs beyond traditional 2D change detection [3], [4]. However, a predominant number of prevailing developments still center around single-modal and 2D change detection, such as introducing contrastive metrics for learning class-distinct features [5], [6], leveraging multitask consistency for semi-supervised training [7], [8], and adopting attention mechanisms to model long-range context [9], [10].

Some noteworthy examples demonstrate the incorporation of multimodal data into change detection task offers both practical flexibility and technical advantages. Combining optical images with synthetic aperture radar (SAR) [11], [12] alleviates weather-related and atmospheric restrictions. Using point cloud data from Lidar and photogrammetry [13] for detecting 3D changes is obviously less constrained in input pairs formation, resulting in enhanced flexible application. Furthermore, the different imaging modalities may be complementary for enhancing the change detection in some extreme conditions (e.g. flooding [14] and burned areas [15]). In our particular context, merging Digital Surface Models (DSM) with aerial imagery incorporates vital vertical data, enhancing the granularity of change detection. Our approach surpasses the conventional 2D semantic change detection methods by providing a more nuanced understanding of spatial changes (see Figure 8).

Due to the scarcity of bi-temporal 3D data, existing methods for high dimensional change detection often rely on multi-source 3D data, requiring manual modality alignment before change detection, such as dense image matching [16] and artificial feature selection [13], [17], which are time-consuming processes that risk information loss. In MTBIT [18], the use of bi-temporal 2D images to infer changes in building height overlooks the benefits of multimodal data integration, as exhaustively discussed in section 2. Consequently, there is an ongoing gap in research on change detection beyond 2D that fully leverages the potential of multimodal data.

To address these gaps and tackle the inherent challenges, we have developed a multi-temporal dataset named Hi-BCD. This dataset comprises pairs of pre-event Digital Surface Models (DSM) representing height data and post-event aerial images. It is designed for detecting multi-category semantic and height changes in buildings across three cities in the Netherlands. Through extensive benchmarking of the state-of-the-art change detection methods, including convolutional neural network (CNN) based and Transformer-based [19] methods, we discover the potential multitask conflicts between semantic change detection (classification task) and height change detection (regression task). In Figure 2, the multitask branches have brought great impact on each other. The performance of semantic change detection branch declined due to the height branch, while the height change detection branch conversely gained some improvements due to semantic hints. Therefore, we propose a novel Transformer-based pipeline that learns shared representation from images and DSM data via cross-attention (see Figure 1). It is equipped with an explicit multitask consistency strategy, which involves the mapping from continuous height change to discrete pseudo change with

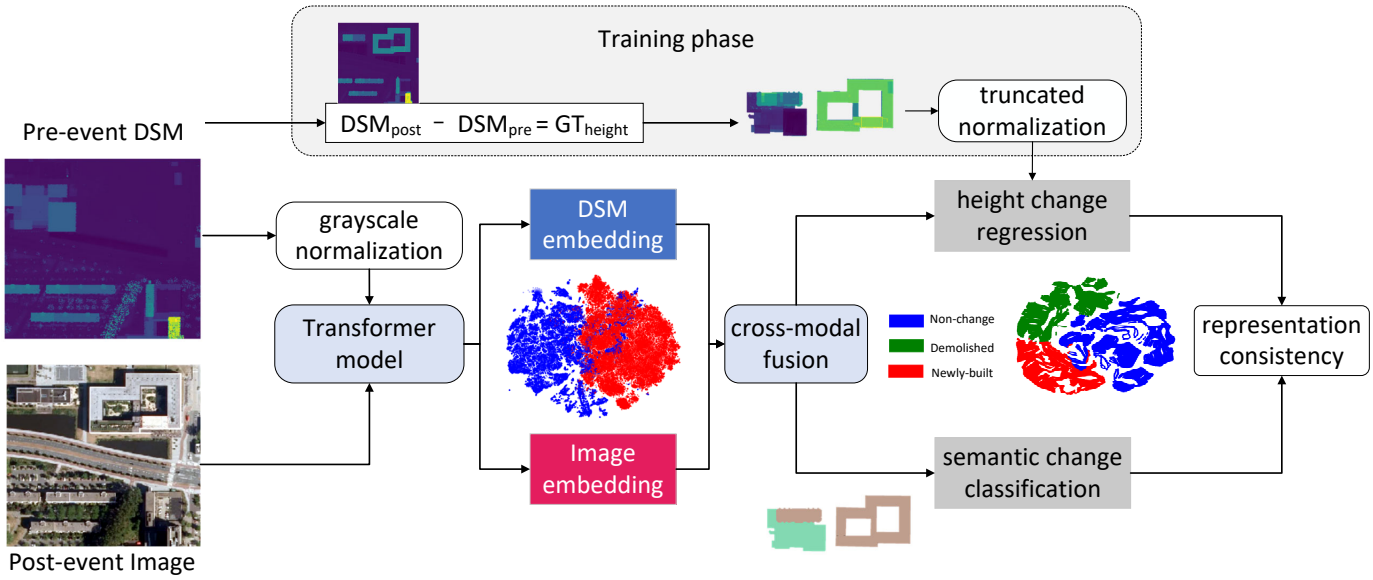

Fig. 1. The conceptual pipeline showing how multimodal image and DSM data are utilized for detecting height and semantic changes simultaneously.

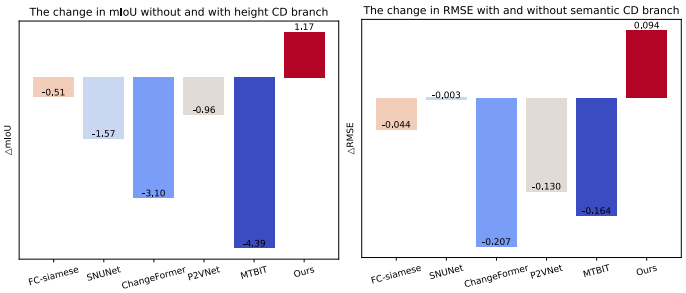

Fig. 2. The performance change of semantic (left) and height (right) change detection in a single-task and multitask manner, which implies the multitask conflicts between 2D semantic and height change detection.

a soft-thresholding module. Then the pixel-wise similarity is maximized between pseudo change and real semantic change, enabling the information interaction of multimodal change maps. The contributions of this paper are summarized as follows:

- We propose an efficient and light Transformer-based network that fuses feature of cross-dimensional modalities via parallelly arranged cross-attention modules.
- We reveal the potential multitask conflicts in state-of-the-art methods while simultaneously handling semantic and height change detection. We propose a multitask consistency constraint that quantifies the similarity between semantic and pseudo change obtained through height change thresholding for alleviating multitask interference.
- We build a multimodal DSM-to-image buiding change detection dataset called Hi-BCD, with generously sized high-resolution tiles. It enables the detection of 2D semantic and 3D height changes simultaneously from cross-dimensional modalities.
- The experiment in Hi-BCD demonstrates that our method outperforms existing methods with consistent semantic and height change detection results. Additionally, the proposed consistency strategy can be easily employed to enhance the other methods.

The rest of this paper is organized as follows: Section 2 provides a brief overview of single-modal and multimodal change detection methods. Section 3 introduces the proposed multimodal change detection network and multitask consistency. Section 4 describes our dataset. In Section 5, we perform a comparison with some current state-of-the-art convolutional neural network (CNN) based and Transformer-based change detection models, which reveals the multitask conflicts and demonstrates the superiority of our method. We conduct ablation studies about the influence of multitask consistency in semantic and height change detection, providing a better understanding of our model. Section 6 draws conclusions.

II. Related Work

Single-modal change detection. The most remarkable achievements occur in the field of single-modal 2D image change detection, where large-scale data are available for utilization. These studies have improved the accuracy and efficiency with superior training metrics [5], [6], densely-connected structure [20], [21], enhanced local and global context aggregation [9], [22], [27], light-weight components [23], and decoupled change modeling [24]. Some latest studies also focus on detecting multi-class changes for in-depth scene understanding [2], [25]. In [26], Qi et al. introduced a grid-based method that categorizes grids into one of three change patterns: significant increase, significant decrease, or roughly unchanged, marking a substantial advancement beyond traditional binary outcome approaches (changed or unchanged).

It is a significant trend that also challenges to detect the volumetric or vertical information in real applications such as quantitative estimation of changes in urban areas, forest biomasses, and land morphology [18], [28]. As multi-view imaging and aerial laser scanning (ALS) technologies continue to advance, an increasing amount of DSM-to-DSM [3] and

cloud-to-cloud differencing methods [28], [29] have emerged. However, it is a strong hypothesis that multi-temporal data are available, especially for 3D data, leaving a barrier to the wide applicability of these methods.

Multimodal change detection. A large number of multimodal change detection studies focus on detecting changes between optical and SAR images to alleviate the restriction of weather and atmosphere. Due to the scarcity of multimodal data, the recent studies tend to build the pixel-wise or graph correlation in an unsupervised manner without considering the deep features, such as the energy-based model [11], coupled dictionary learning [30], Markov random field model [14], change vector analysis [31], and graph representation learning [12], [32], [33]. Sun et al. [33] presents an iterative robust graph combined with Markovian co-segmentation, focusing on structure consistency for enhanced detection accuracy. The [34], [35] further this by employing structure cycle consistency and an improved nonlocal patch-based graph to address noise and sensor differences, showing superior performance across multiple datasets and scenarios. These methods signify advancements in unsupervised change detection without the need for labeled data. Regarding these deep learning approaches, the majority employ an explicit image translation process [36]–[38]. Conversely, some methods prefer a single-modal change detection framework that projects two heterogeneous images into a common latent space [39], [40].

Some recent efforts are delving into change detection beyond 2D with multimodal data. In [41], the Siamese CNN was employed to detect changes between point clouds obtained from ALS and dense image matching. In [17], the features including color, shape, and elevation maps are manually extracted for change detection between the point cloud and image. In [16], a multi-source point cloud processing network was devised to detect genuine 3D changes. Yet, most of these methods require a time-consuming pre-processing step for modality alignment, which potentially results in information loss. Conversely, we directly handle multimodal data across different dimensions inspired by cross-attention mechanisms [42]. In MTBIT [18], it attempted to infer a change map represented by DSM from bi-temporal 2D images. Unlike MTBIT, we propose to directly deal with DSM-to-image multimodal inputs for various reasons: 1) Estimating height from single view image remains an ill-posed problem. The introduced pre-temporal DSM provides abundant context priors about vertical information near the change areas. 2) The ground truth elevation change is essentially generated with bi-temporal DSMs, either in MTBIT or our method. Therefore, it is underutilization of a considerable amount of 3D information in MTBIT. 3) For change detection that spans over a long time, there may exist a significant resolution gap in multi-temporal images (e.g., 2.0m vs. 0.25m as shown in Figure 8(a)(c)). On the contrary, the DSM data derived from point cloud in our dataset allows for at least 4 point records at a $0.25\times0.25m^2$ grid.

Multitask learning. In multi-task learning (MTL), a single model is trained to simultaneously predict outcomes for multiple tasks, leveraging data across these tasks to achieve better performance than if each task were learned independently [44], [47], [48]. Unfortunately, MTL often causes performance degradation compared to single-task models [49]. A main reason for such degradation is gradients conflict [43], [44]. The ***model-level*** multitask optimization involves addressing multitask conflicts through gradient manipulation. These per-task gradients may have conflicting directions or a large difference in magnitudes, with the largest gradient dominating the update direction. Various heuristics have been introduced for manipulating the task-specific gradient, such as the uncertainty of the tasks [50], the norm of the gradients [45], equal cosine similarities [51], and Pareto optimal [46]. The ***Task-level*** multitask learning typically establishes correlations among multiple tasks using specific transformations [50], [52], [53] or by integrating multiple tasks that are inherently consistent [54]–[56]. For instance, [53] learns a transformation between semantic segmentation and depth feature spaces. Zhu et al. [52] explicitly measure the border consistency between segmentation and depth and minimize it in a greedy manner by iteratively supervising the network towards a locally optimal solution. Kendall et al. [50] models the uncertainty of segmentation and depth to re-weight themselves in the loss function. In the context of change detection, the auxilary task to predict the segmentation boundaries of bi-temporal inputs is widely adopted [54]–[56], where the learned boundary representation can be shared with the change detection branch. The auxiliary constraint is usually beneficial, as it introduces inductive bias through the inclusion of related additional information. Nonetheless, as reported in [57], it can sometimes hamper the performance of each task. Certain studies have delved into the multitask relationship by considering multitask consistency. In [2], three consistency metrics including binary, change area, and no-change area consistency are used for evaluation. In [58], the consistency between bi-temporal semantic labels and the change labels is exploited to enhance semi-supervised generalization.

Zhu et al. [52] underscore the unique challenge of correlating semantic maps with depth maps due to their significant but complex relationship. We explore a straightforward yet potent task-level transformation to navigate the intricacies of multitask conflicts. We highlight the connection between negatively changed height, often associated with demolished buildings, and positively changed height, as observed in newly constructed areas.

III. Method

A. Problem definition and data preprocessing

1) Multimodal change detection problem: Figure 1 and 3 depict the pipeline of multimodal change detection problem. We aim to detect both height change and semantic change with pre-event DSM and post-event image. In training phase, the ground truth of height change is obtained by subtracting the multi-temporal DSMs and performing truncated normalization as section III-A2. The floating ranged DSM pixels are normalized to grayscale before inputting them to the embedding model. The Transformer-based embedding model $\mathcal{T}$, which is detailed in section III-B, learns the DSM and

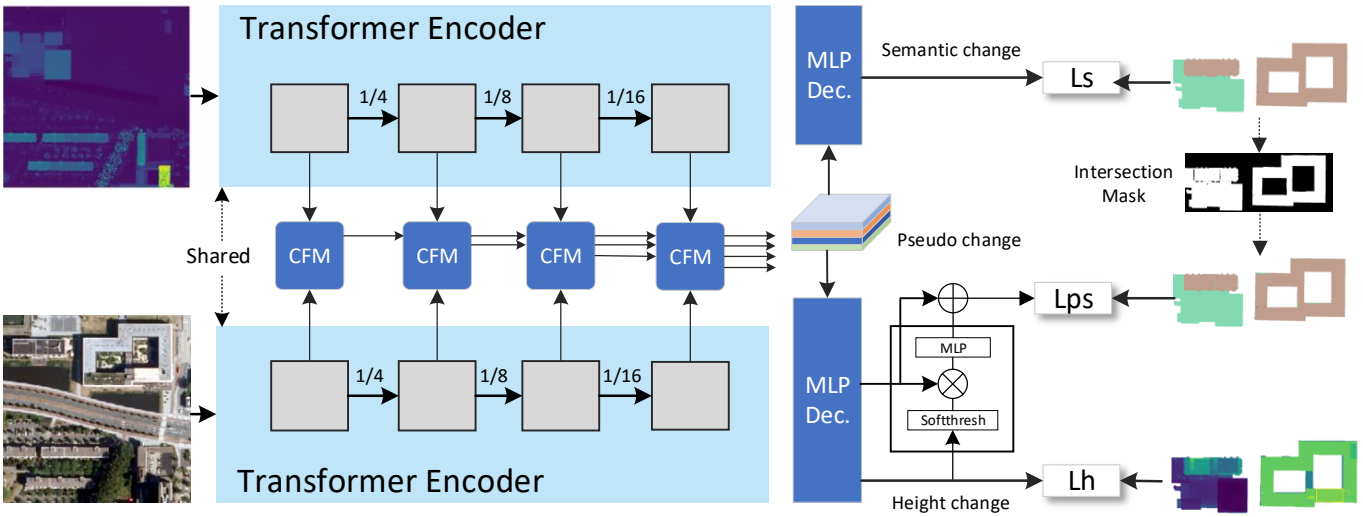

Fig. 3. Our Transformer-based multimodal change detection pipeline is named MMCD. It consists of the pyramid backbone with four Transformer layers, the cross-modal fusion module (CFM), and the multi-layer perception (MLP) decoder. The multitask consistency acts as an explicit constraint for enhancing multimodal correlation.

image embedding with shared parameters. It can formulated as

$$X_H = \mathcal{T}(DSM_{pre}), X_I = \mathcal{T}(image_{post}). \quad (1)$$

Following the backbone, the cross-model fusion process aims to augment the representation of a specific modality by integrating cues from other modalities, as elaborated in Section III-B. Building on the representation consistency between height change and semantic change, we introduce an explicit consistency constraint. This constraint is designed to ensure that both tasks mutually enhance each other's performance, as detailed in Section III-C.

2) Data preprocessing: **Truncated normalization for height changes**: Height change detection leverages the $L2$ regression loss, with practical implementations [18], [59] incorporating a Tanh activation layer. This layer normalizes the final decoded layer's output to a range of -1 to 1 for enhanced training stability. Therefore, the ground truth height should be rescaled to $[-1, 1]$ during training. The normalization parameters are obtained by truncating the distribution to include 99.5% of the height change values from the training set, optimizing the model's focus on dominant height changes. Specifically, the truncated rescaling range is $[-27.29, 87.26]$ (meters) in our training set.

Gray-scale Normalization for input DSMs: In practical implementation [60], [61], a normalization process is necessary during training image data. For gray-scale RGB images, the integer pixel value ranges from 0 to 255, while the digital surface model could be negative or positive floating values. This causes gradient fluctuation during training and makes it hard to converge [62]. To this end, we first rescale the height values in DSM into gray-scale as follows:

$$Height_{rescaled} = \frac{(Height - min)}{max - min} \times 255 \quad (2)$$

where $min = -10$ and $max = 40$. These two hyperparameters are also determined by truncating nearly 99% of the height value ranges. Then the standard normalization is applied for height values and gray-scale image values. It modifies the data of each channel so that the mean is zero and the standard deviation is one.

B. MMCD: Transformer-based multimodal change detection network

Efficient Pyramid backbone. The Transformer network [19], increasingly dominant in recent multimodal data processing, encounters significant challenges when managing bi-temporal inputs that are both high resolution. The architecture, while revolutionary for its attention mechanisms and scalability, struggles with the computational and memory demands posed by large-scale inputs. We aim to develop an efficient and lightweight backbone architecture, considering the typically large data volumes in remote sensing. Therefore, we employ an efficient Transformer block with sequence reduction as our backbone, which is detailed in [63], [64]. Furthermore, we minimize the embedding dimensions to achieve a more compact model size, effectively halving the complexity compared to ChangeFormer [64] (see Table III). The pyramid features output from height and image branches are denoted as $\mathrm{X}_{\mathrm{H}}^{n}$ and $\mathrm{X}_{\mathrm{I}}^{n}$, where $n \in \{1, 2, 3, 4\}$.

Cross-modal fusion. As mentioned in [52], despite the high relevance between depth (or height) data and gray-scale images, establishing the definitive relationship between them is challenging. Inspired by the widely used cross-attention mechanism [42], which considers all the multimodal features (such as text, images, audio, or video) as sequences, we compute attention scores by calculating the dot product between the query from one modality and the keys from another modality, followed by a softmax function to normalize the scores. It allows the model to dynamically focus on specific parts of an image based on the context provided by the other modality, and vice versa. The designed cross-modal fusion module is shown in Figure 4(a). It parallelly takes the feature embedding

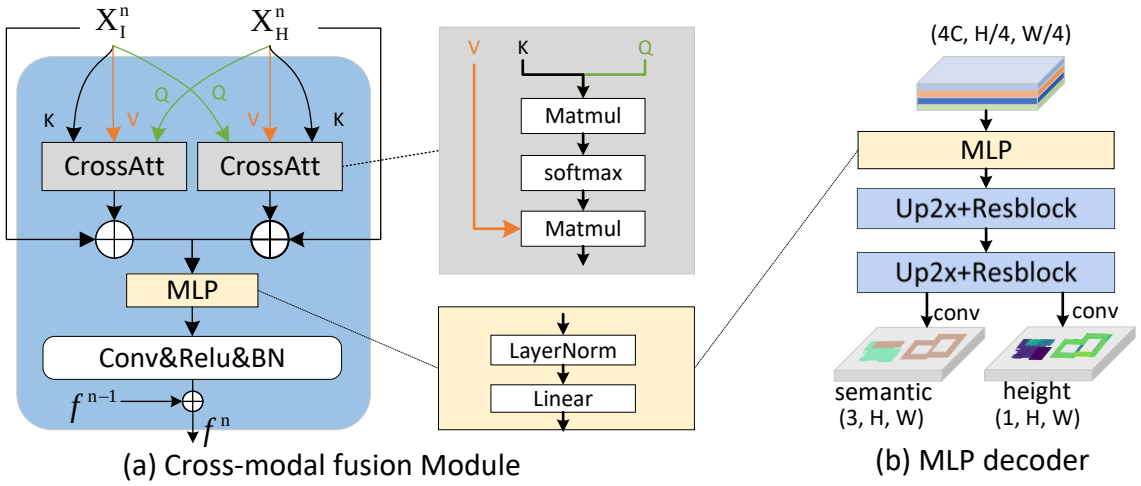

Fig. 4. The structure of feature fusion module and decoder in our method.

from one modality as query, and the embedding from the other modality as key and value. The standard self-attention component is

$$\text{Attention}(X) = \text{softmax}(QK)V = \text{softmax}\Big((W_q X)(W_k X)^T\Big) W_v X, \tag{3}$$

where W_q, W_k and $W_v \in R^{C\times d}$ are learnable matrices, $X \in \mathbb{R}^{d\times C}$ is input sequence, and the softmax is $e^{x_i}/\sum_{j=1}^{N} e^{x_j}$. The C is sequence length and d is embedding dimension. The left cross-attention block in Figure 4(a) can be formulated as

$$\text{Cross-attention}(\mathrm{X}_\mathrm{H}^n, \mathrm{X}_\mathrm{I}^n) = \text{softmax}\left((W_q \mathrm{X}_\mathrm{H}^n)(W_k \mathrm{X}_\mathrm{I}^n)^T\right) W_v \mathrm{X}_\mathrm{I}^n. \tag{4}$$

This layer allows each modality to query (seek information from) the other modality's representations. The attention scores $\left((W_q \mathrm{X}_\mathrm{H}^n)(W_k \mathrm{X}_\mathrm{I}^n)^T\right)$ are used to create a weighted sum of the value vectors from the attended modality, allowing the model to focus more on the relevant features. This results in a richer, contextually informed representation that combines information from both modalities. The multimodal feature spaces before and after the cross-attention is depicted in Figure 14. The CFM uses two symmetrically arranged cross-attention operators for capturing mutual relationships between features derived from the DSM and image branches. Next, they are merged through MLP and the convolutional unit, which is then pixel-wise added to the previous feature layer f^{n-1} to obtain f^n.

MLP decoder. Figure 4(b) depicts the streamlined structure of our MLP decoder. We use non-parameterized up-sampling, while ChangeFormer [64] utilizes learnable transposed-convolution that incurs higher computation cost. Furthermore, the residual convolutional blocks [65] are utilized to enhance the local relations during the up-sampling process. The final semantic and height change maps are generated through 3×3 convolution block following the approach outlined in MTBIT [18].

Model variations. Figure 3 depicts the final model architecture, while the variations of it, such as "only semantic cd branch", "only height cd branch", and "semantic + height cd branch" in Table V, are briefly depicted Figure 5.

C. Multitask consistency by predicting the pseudo semantic change

We bridge the gap between height change detection and semantic change detection tasks by imposing an explicit consistency constraint, facilitated through the prediction of an auxiliary pseudo-change map. Our approach addresses the inherent discrepancy between semantic and height changes—the former being categorical with discrete values, and the latter represented by continuous floating-point values. As shown in Figure 6(a)(b), by adopting zero as a threshold for height change, we obtain a classification map termed pseudo change that includes three classes: 0 (the background), 1 (positive height change) and the -1 (negative height change), which differs from the semantic change (Figure 6(c)). The hard thresholding can be formulated as

$$T_h(x) = \begin{cases} 1, x > 0 \\ 0, x = 0 \\ -1, x < 0. \end{cases} \tag{5}$$

Since it is not differential, we adopt a soft thresholding function as follow

$$T_s(x) = 2 \times sigmoid(\frac{x}{t}) - 1, \tag{6}$$

where $sigmoid = 1/(1+e^{-x})$, and t is a positive temperature parameter for controlling the sharpness of the transition around zero. In our experiment, we set $t = 0.5$. Smaller t leads to a more accurate approximation to $T_h(x)$. This can be implemented with a sigmoid and an MLP layer for introducing strong prior to the pseudo-change branch. The pseudo change highly overlaps with semantic change but is not totally the same as shown in Figure 6(d). Therefore, only overlapped areas are considered when measuring the consistency between them. The objective is to minimize the following objective function

$$\mathcal{L}_{\text{consistency}} = \min_{\text{Pred}_{sc}, \text{Pred}_{psc}} (GT_{psc} \cap GT_{sc}) * |Pred_{psc} - Pred_{sc}|, \tag{7}$$

where $|\cdot|$ is kind of distance. GT_{sc} and GT_{psc} are ground truth of semantic and pseudo change. $Pred_{sc}$ and $Pred_{psc}$

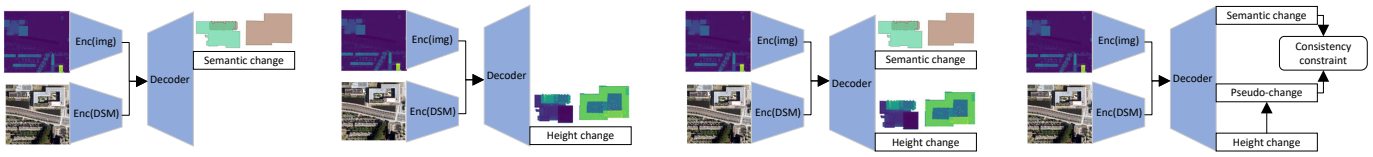

Fig. 5. The network variations of our method, arranged from left to right, include: the only semantic change detection branch, only height change detection branch, the multitask branch, and the multitask branch with consistency constraint.

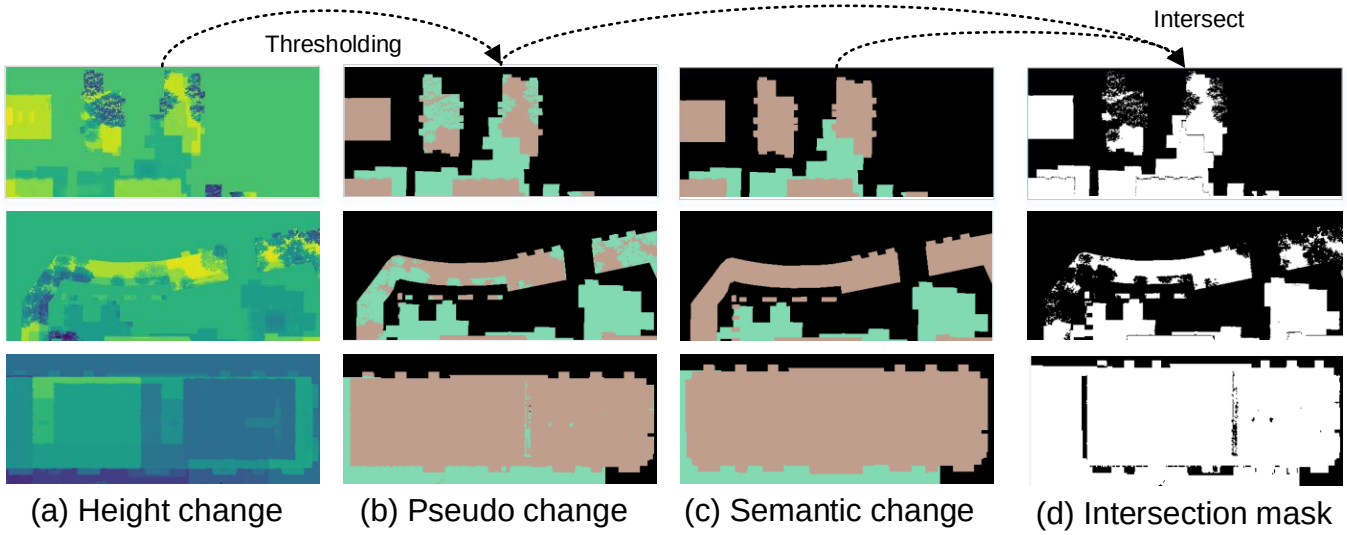

Fig. 6. The inconsistency of multimodal change labels. (a) height change; (b) pseudo change by classifying the zero height as unchanged, positive height as newly-built and negative height as demolished regions; (c) semantic change; (d) intersection mask between height change and pseudo change, where the intersection rate of training, validation, and testing set are 79.71%, 89.23% and 90.03% respectively.

are corresponding model prediction. In practical implementation, the semantic and pseudo change branches are separately supervised with GT_{sc} and GT_{psc} respectively, which leads to consistency in their overlapping regions.

D. Loss function

We employ the weighted cross-entropy loss for both the semantic and pseudo change detection branches and utilize mean-square error ($L2$ loss) for the height change detection branch as [18], which are denoted as $\mathcal{L}_{height}$, $\mathcal{L}_{\text{pseudo}}$, and $\mathcal{L}_{\text{semantic}}$ respectively. The final training loss is

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{pseudo}} + \lambda_2 \cdot \mathcal{L}_{\text{height}} + \lambda_3 \cdot \mathcal{L}_{\text{semantic}} \quad (8)$$

where λ_1, λ_2, and λ_3 are fixed loss weights, which are 0.2, 0.2, and 0.6 in our experiment setting.

IV. Hi-BCD: A Multimodal Dataset for Building Change Detection between Height Map and Optical Image

Existing datasets and limitations. Table I presents a concise comparison of existing change detection datasets, encompassing both single-modal and multimodal datasets. A substantial volume of bi-temporal 2D image datasets available, supporting high-resolution, high-frequency, and multi-class change analysis. Despite the abundance of single-modal datasets, limited research has ventured beyond 2D change detection using multimodal data. The **Shuguang** dataset used in [12] contains a pair of SAR and optical images for detecting 2D construction change. The constraint of a small sample size dictates that most earlier methods can only be developed from an unsupervised standpoint. The **multimodalCD** [13] dataset incorporates multi-view image-based and ALS-based point clouds, which are transformed into DSMs to detect binary 2D changes exclusively. However, it confines its focus to 2D changes within small tile sizes, even though it contains beyond-2D information. The **3DCD** [18] dataset employs bi-temporal images to identify not only binary building changes but also variations in height. However, the resolution of height change is merely half that of 2D changes, and it exclusively

TABLE I
The main details of typical existing change detection datasets. First three rows: single-modal datasets. Last four rows: multimodal datasets. Note that the 3DCD [18] employs bi-temporal images to infer changes across multiple modalities.

Name	N. images	Tile size	Resolution	CD map	Classes
LEVIR-CD [66]	637	1024 × 1024	0.5 m	2D	2
Hi-UCD [2]	40800	512 × 512	0.1 m	2D	9
DynamicEarthNet [1]	54750	1024 × 1024	3 m	2D	7
Shuguang [12]	1	921 × 593	-	2D	1
multimodalCD [13]	3615	100 × 100	0.1 m	2D	2
3DCD [18]	472	400 × 400	0.5 / 1m	2D / 3D	1
Hi-BCD (ours)	1500	1000 × 1000	0.25 m	2D / 3D	2

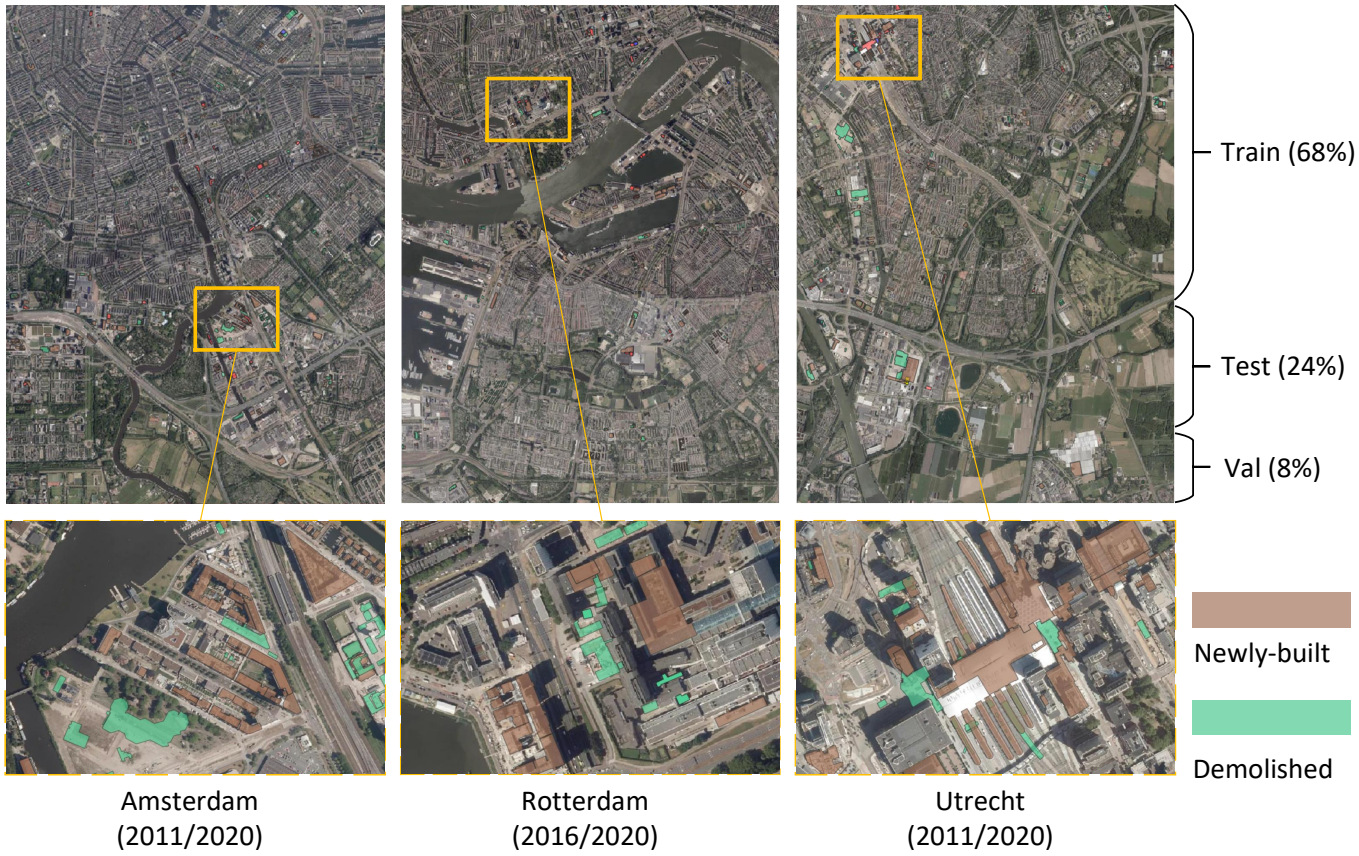

Fig. 7. Hi-BCD dataset. It encompasses three cities in the Netherlands and provides two types of building changes. The dates of early and late periods are denoted under each city. For each city, 68%, 24%, and 8% of the tile are used for training, testing, and validation, respectively.

classifies binary changes. To establish a new foundation for change detection, we introduce the Hi-BCD dataset. It leverages multimodal inputs to simultaneously output semantic and height changes, providing generously sized tiles and high-resolution 2D and 3D semantic change maps.

Dataset overview. As depicted in Figure 7, our study area involves three cities in the Netherlands, including Amsterdam, Rotterdam, and Utrecht. The dates of the pre-temporal and post-temporal periods are indicated beneath each city in Figure 7. We can observe significant variation in the capture dates of the pre-temporal DSM, which reflects the extensive updating period of high-dimensional data. This hinders the application of high-dimensional change detection with dual-temporal 3D data. The DSM data is generated from point clouds using grid sampling and strictly orthogonal projections, whereas the aerial images are ortho-photographs. The data volume for each city comprises five hundred of 1000×1000 DSM-to-image pairs with a ground sampling distance of 0.25 meter (Figure 8(b)(c)). The corresponding multi-class 3D and 2D changes are shown in Figure 8(e)(f). The vertical accuracy is about 0.15 meters [3]. For each city tile, 68%, 24%, and 8% are allocated for training, testing, and validation respectively. Two types of change including 'newly-built' and 'demolished' are defined in the dataset. More details about change objects, pixels, and samples are provided in Table II. Figure 9 portrays the cumulative frequency of height for the two types of change.

TABLE II
THE MAIN DETAILS OF THE HI-BCD DATASET, INCLUDING CHANGED OBJECTS, PIXELS, AND SAMPLE AMOUNT OF THREE CITIES IN THE NETHERLANDS.

Attribute	Category	Amsterdam	Rotterdam	Utrecht
changed objects	newly-built	389	510	458
	demolished	251	229	187
changed pixels	amount	6.625M	5.139M	7.73M
	$prop_{/total}$	1.3%	1.0%	1.5%
samples (size: $1k \times 1k$)	total	500	500	500
	with change	40.8%	34.2%	43%

Annotation procedure. We build the dataset based on AHN[1] (Actueel Hoogtebestand Nederland), the nationwide elevation data project in the Netherlands. Specifically, the early-period elevation data is obtained by rasterizing point clouds from AHN3 (2011-2019), while the aerial images[2] with the close date to AHN4 (2020-2022) are used as late period. The annotation procedure is as follows:

1) **Change definition.** The construction and demolition of buildings are annotated based on the difference map between AHN3 and AHN4. Since the capture date of aerial images does not precisely align with that of the point cloud, we have focused our annotations on multi-class building changes that

[1]http://www.ahn.nl/
[2]https://www.beeldmateriaal.nl/over-beeldmateriaal

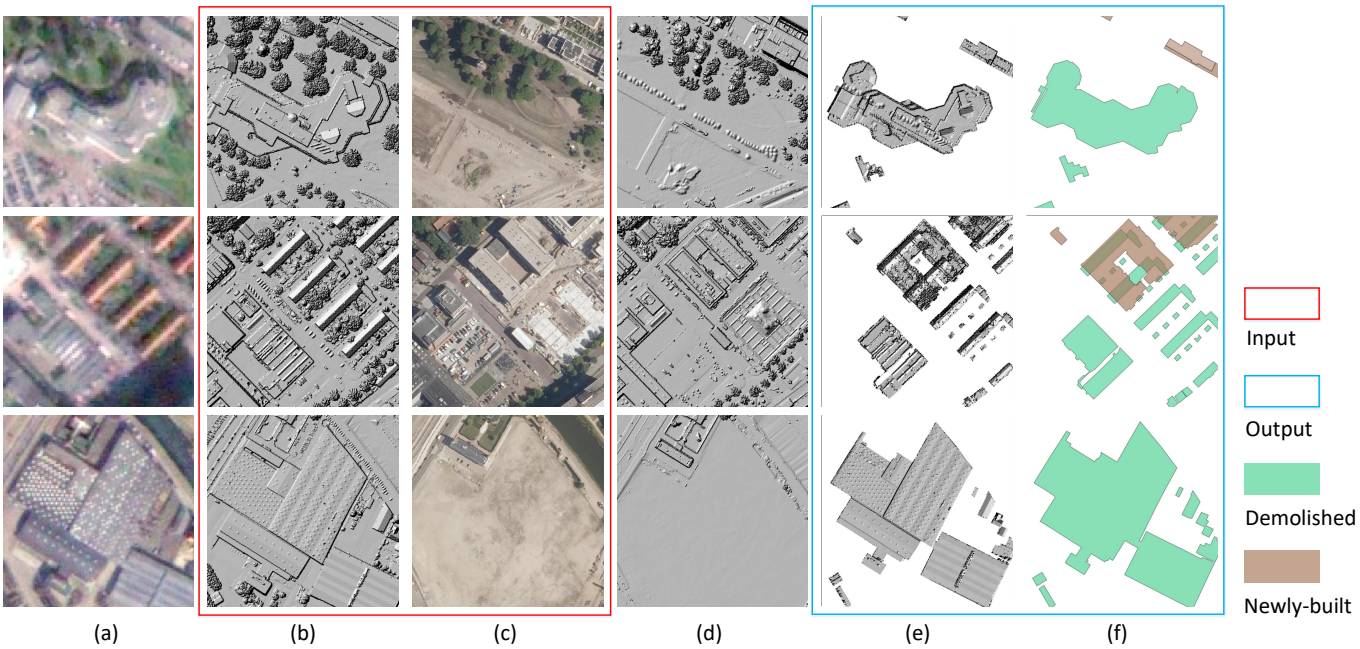

Fig. 8. Examples of Hi-BCD dataset, where the DSMs are displayed in hillshade manner (A widely used visual technique to give a three-dimensional appearance from DSM data). The (a) and (b) are pre-temporal images (**2.0m** resolution) and DSMs (**0.25m** resolution). The (c) and (d) are post-temporal images and DSMs (all **0.25m** resolution). The (e) and (f) are height and semantic changes between the multimodal temporal inputs.

remain relatively stable within a one-year time frame, while excluding highly dynamic changes such as trees and vehicles. Note that the demolished and newly-built buildings do not highly correspond to the negative and positive height change values respectively, due to tree occlusions and the penetration of low-reflectivity surfaces such as glass roofs, as depicted in Figure 6. We define a mask that indicates whether the building changes are highly relevant to their height change values, which is

$$M(i,j) = \begin{cases} 1, \text{highly relevant} \\ -1, \text{otherwise} \end{cases} \tag{9}$$

2) **Edge situation.** The second row in column (f) of Figure 8 illustrates a complex scenario where various building changes intersect. This suggests a sequence of events where a building is first demolished and subsequently replaced by a new one. For such a situation, the change type is determined based on the elevation difference, where pixels with a positive change in height are categorized as newly built, while pixels with a negative change in height are associated with demolished building changes. While we have provided a simplified representation by aggregating these overlapping changes into single-type changes, there remains an opportunity for future research to delve into finer sub-situations, describing the entire evolution process.

3) **Change label generation.** Based on the above-mentioned change definition and edge situation, the change map (CM) is formulated as

$$CM(i,j) = \begin{cases} \text{demolished}, \Delta H \cdot M(i,j) < 0 \text{ and building in early period} \\ \text{newly-built}, \Delta H \cdot M(i,j) > 0 \text{ and building in late period} \\ \text{background}, otherwise. \end{cases} \tag{10}$$

where ΔH is the elevation difference between the bi-temporal DSM data, i.e., $\Delta H = \text{DSM}_{AHN4} - \text{DSM}_{AHN3}$. The def-

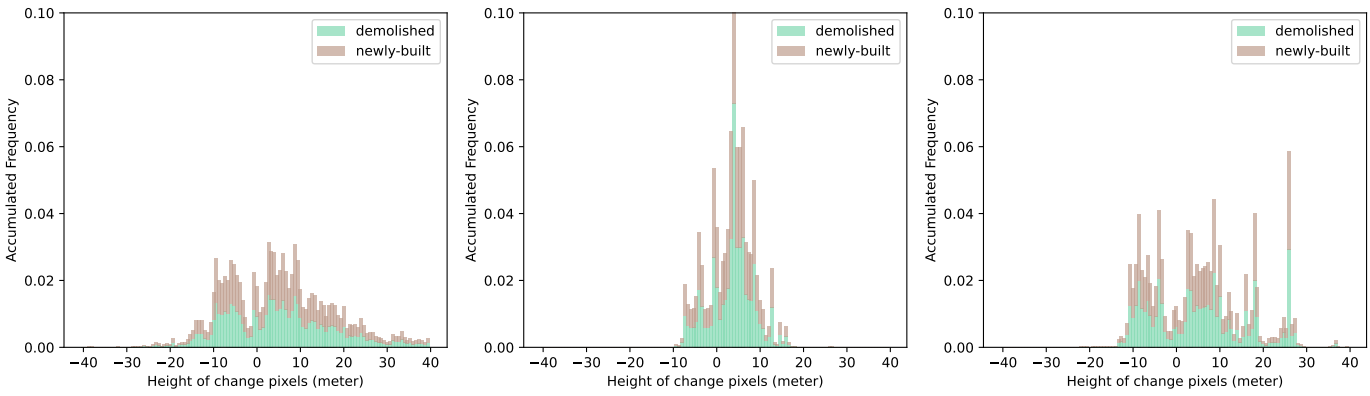

Fig. 9. The cumulative frequency of height values for two types of change pixels.

inition of the first two cases implies a sub-situation of bi-temporal buildings. It may be a little misaligned with the building in image data due to viewpoint distortion. The labels for 3D height changes correspond to the masked regions in the AHN4-to-AHN3 difference map.

4) **Tile splitting.** The original tile size of each city is 25000×20000 with a pixel granularity of 0.25 meter, which is subsequently split into 500 pairs of 1000×1000 sized samples. Note that we retained samples that do not include changes to better reflect the data distribution of the real-world scenario.

Challenges of Hi-BCD dataset. The fundamental challenge of our dataset is to learn the representation of multi-class elevation changes from multimodal and cross-dimensional inputs. There are some inherent misalignments between the multimodal inputs. 1) The pixels hold diverse implications as the DSMs rasterize the height dimension from Lidar point clouds and represent the absolute land elevation, while the images reflect the intensity of visible light. They exhibit vastly different numerical ranges, where the height range is $[-8.24, 183.64]$ for the original DSM and $[-99.55, 134.21]$ for the changes in the training set, while the image values are in grayscale $[0, 255]$. 2) Their distribution differs a lot as the DSMs exhibit similar height in the ground regions, while the images portray different colors and textures for various land covers. 3) There exists geometry misalignment due to viewpoint distortion of aerial images although they utilize the same coordinate system. Furthermore, severe change-unchange imbalance can be observed in Table II. Additionally, there is inconsistency between the semantic change labels and the height change labels as shown in Figure 6.

V. Experiments and Results

A. Experiment setting

Implementation details. The Tanh function normalizes height outputs to $[-1, 1]$. The elevation scale of training set $[-27.29, 87.26]$ covering 99.5% of pixels is used for denormalization. We set class weights of 0.05, 0.95, and 0.95 to the background, demolished, and newly-built areas for weighted cross-entropy loss. All the models are pre-trained in LEVIR-CD [66] and then trained for 300 epochs with equivalent batch size of 8. At ease of multi-scale downsampling, the original tile size of 1000×1000 is adjusted to 1024×1024 during training. More details can be found here[3].

Metrics. For semantic change detection evaluation, we used the mean intersection over union (mIoU) and F1-score as denoted in [18]. For the height change detection, we keep consistent with relevant research [18], [59], [67], including the following metrics:

- Root Mean Square Error (RMSE): $\sqrt{\frac{1}{n}\sum(H_r - H_e)^2}$
- Mean Average Error (MAE): $\frac{1}{n}\sum|H_r - H_e|$
- Root Mean Square Error (cRMSE): $\sqrt{\frac{1}{n}\sum(H_r - H_e)^2}$
- Average relative error (cRel): $\frac{1}{n}\sum\frac{|H_r - H_e|}{H_r}$
- Mean normalized cross correlation (ZNCC): $\frac{1}{N}\sum_i^N \frac{(H_{ri}-\mu_{H_r})(H_{ei}-\mu_{H_e})}{\sigma^{H_r}\sigma^{H_e}}$

[3]https://github.com/qaz670756/MMCD

where H_r denotes the reference height, H_e denotes the estimated height, and N denotes the estimated pixel count. μ and σ are the mean values and standard deviations of H_r and H_e, respectively. The cRMSE and cRel indicate that only changed areas are considered. The ZNCC quantifies the spatial correlation between output and ground truth, while the other metrics measure the degree of absolute errors at each pixel in meters. Moreover, we include the million parameter count (MParams) and Giga Floating-Point Operations (GFLOPs) as metrics to compare the model complexity [9], [68].

Compared methods. Limited research, such as [18], has concurrently detected semantic and height changes. To provide a benchmarking, we follow the structure of MTBIT [18] that maintains the original change detection network while slightly modifying the decoder with an additional height change detection branch. Among the selected methods, the FC-Siamese [22] is the first fully convolution-based change detection architecture widely used for comparison. The SNUNet [20] is a state-of-the-art CNN-based method featuring a densely connected backbone. The ChangeFormer [64] is a Transformer-based method that yields promising results in most change detection benchmarks. The P2VNet [24] models the change process in a novel multi-frame transition perspective. The MTBIT extends the Transformer-based BIT [9] for simultaneously detecting semantic and height changes.

B. Comparison with state-of-the-arts

In this section, we evaluate semantic and height change detection to explore how these two tasks influence each other. Note that semantic change detection refers to multi-class 2D change detection in our context. The model operates in a multitask situation when performing joint semantic and height change detection. Otherwise, it is in a single-task setting when addressing only one of these tasks.

Semantic change detection. In Table III, our method achieves competitive semantic results across both single-task and multitask settings. Specifically, by employing a single semantic change detection branch, we achieve close results to ChangeFormer with only half the model complexity. Besides that, the model with the largest number of parameters (ChangeFormer) and highest computational cost (P2VNet) achieved the second and third best results, respectively. When augmented with our consistency-enhanced height prediction branch, the metric numbers exhibit continued improvement. On the contrary, the other methods that attached to the height change detection branch without consistency constraint, show a notable degradation. This suggests that the added height change detection branch hinders the learning of the semantic branch, where a similar phenomenon is also observed in prior works [44], [52]. Different optimization objectives among multiple tasks can lead to potential mutual interference during feature optimization. With the help of the consistency constraint, our method prevents performance degradation due to interference from the height change detection branch.

Height change detection. Table IV reveals an intriguing pattern: Among the 30 metric results obtained with a height change detection branch, 23 of them gain improvements when

TABLE III
THE SEMANTIC CHANGE DETECTION PERFORMANCE BEFORE AND AFTER ATTACHING THE HEIGHT PREDICTING BRANCH. THE METHODS WITH ∗ ARE ORIGINALLY DESIGNED WITH A HEIGHT BRANCH. THE COLORS RED, GREEN, AND BLUE INDICATE THE TOP THREE RESULTS.

Method	Year	only semantic cd branch				semantic + height cd branch				Complexity (two branches)	
		IoU_D ↑	IoU_N↑	mIoU↑	F1-score↑	IoU_D ↑	IoU_N↑	mIoU↑	F1-score↑	MParams↓	GFLOPs↓
FC-Siamese [22]	2018	27.77	29.18	28.48	44.33	27.60	28.34	27.97	43.72	1.552	92.908
SNUNet [20]	2021	25.47	26.07	25.77	40.98	20.77	22.97	21.87	35.87	3.012	220.696
ChangeFormer [64]	2022	47.17	36.67	41.92	58.88	38.89	41.45	40.17	57.31	29.75	340.165
P2VNet [24]	2022	37.41	30.61	34.00	50.65	35.63	30.46	33.04	49.62	5.425	527.442
*MTBIT [18]	2023	37.44	29.97	33.71	50.31	34.40	27.04	30.72	46.88	15.2	154.72
*Ours	2023	43.76	39.10	41.43	58.55	44.29	40.90	42.59	59.72	11.659	168.893

TABLE IV
THE HEIGHT CHANGE DETECTION PERFORMANCE WITHOUT AND WITH SEMANTIC BRANCH. THE UNDERLINED NUMBERS INDICATE A DECREASE WITH THE ATTACHED SEMANTIC CHANGE DETECTION BRANCH COMPARED TO ITS SINGLE-BRANCH COUNTERPART. THE METHODS WITH ∗ ARE ORIGINALLY DESIGNED WITH A HEIGHT BRANCH. THE COLORS RED, GREEN AND BLUE INDICATE THE TOP THREE.

Method	only height cd branch					semantic + height cd branch				
	RMSE↓	MAE↓	cRMSE↓	cRel↓	cZNCC↑	RMSE↓	MAE↓	cRMSE↓	cRel↓	cZNCC↑
FC-Siamese	1.505	0.446	8.622	1.838	0.274	1.461	0.309	8.995	1.506	0.373
SNUNet	1.574	0.498	9.397	2.988	0.186	1.671	0.779	9.216	2.704	0.265
ChangeFormer	1.658	0.313	8.404	2.110	0.308	1.343	0.402	8.204	2.485	0.394
P2V-CD	1.392	0.399	9.037	1.937	0.261	1.408	0.305	8.932	1.463	0.377
*MTBIT	1.530	0.475	9.441	1.780	0.179	1.457	0.400	8.563	1.987	0.345
*Ours	1.273	0.397	8.317	2.711	0.379	1.267	0.290	8.281	1.900	0.394

working with a semantic branch, while only 7 of them show a decrease. This phenomenon underscores the positive impact of learning the shared representation through the implicit hints from semantic change. As also denoted in [52], the object boundaries are easier to capture from the semantic map compared to the depth map. Incorporating a dedicated branch for predicting pseudo changes from the height map establishes a clear correlation between semantic and height changes. This enhancement is evident in the improved performance of height change detection across five different metrics. With only half the model complexity of ChangeFormer, our model reaches the top in most metrics. Note that this enhancement, facilitated by a consistency constraint, can be conveniently adapted to other methods, yielding promising improvements as showcased in Table VI.

Qualitative Results. Figure 10 depicts the visual comparison of semantic and height change detection for the top three methods. The tendency could be observed in the first four columns of Figure 10. In the last four columns, collapsed results are evident for models lacking semantic hints in single-task setting, with ours being the exception. Note that the single-task visual results of our method are derived from the height change detection branch that is augmented with multitask consistency.

Figure 11 presents a comparative analysis of height prediction, illustrating the distribution patterns across various methods. In terms of the height change value range, other change detection methods generally tend to underestimate, whereas our method delivers a more accurate range. Regarding the overall distribution, most methods exhibit a single peak around zero, except for ChangeFormer, MTBIT, and our approach, which align more closely with the actual distribution that features at least two significant clusters. However, for all methods, the dominant portion of height outputs clusters near zero. This reflects the background to changed areas imbalance. Addressing this problem remains a direction for future research.

Figure 12 depicts the large-scale change detection results. Figure 13 and 14 visualize the multimodal feature from the encoder and decoder layers. The baseline of our model is the one that does not have the multitask consistency constraint (MC). Given that our Transformer backbone is a streamlined variant of ChangeFormer, we include ChangeFormer's results for comparative analysis. Our baseline differs from ChangeFormer in having fewer parameters and incorporating a cross-modal fusion module. In Figure 13, the multimodal features from DSM (blue points) and image (red points) are roughly separated into two parts with some overlapping. From left to right, the red cluster is more and more compact. In Figure 14, the three distinct classes are represented with specific colors in the output: demolished buildings are marked in red, newly-built buildings in blue, and the background in green. Our method, when compared to both ChangeFormer and our baseline model that did not incorporate consistency constraints, demonstrates an improvement in the feature space representation. Specifically, it shows better intra-class consistency and inter-class separation.

C. Ablation study

This section explores the impact of the proposed multitask consistency for semantic and height change detection. Initially, we demonstrate that the implicit information, shared at the backbone stage, yields benefits for height change detection but introduces challenges for semantic change detection.

From the results of row 1 and row 3 in Table V, an evident decline in semantic change detection can be observed. Con-

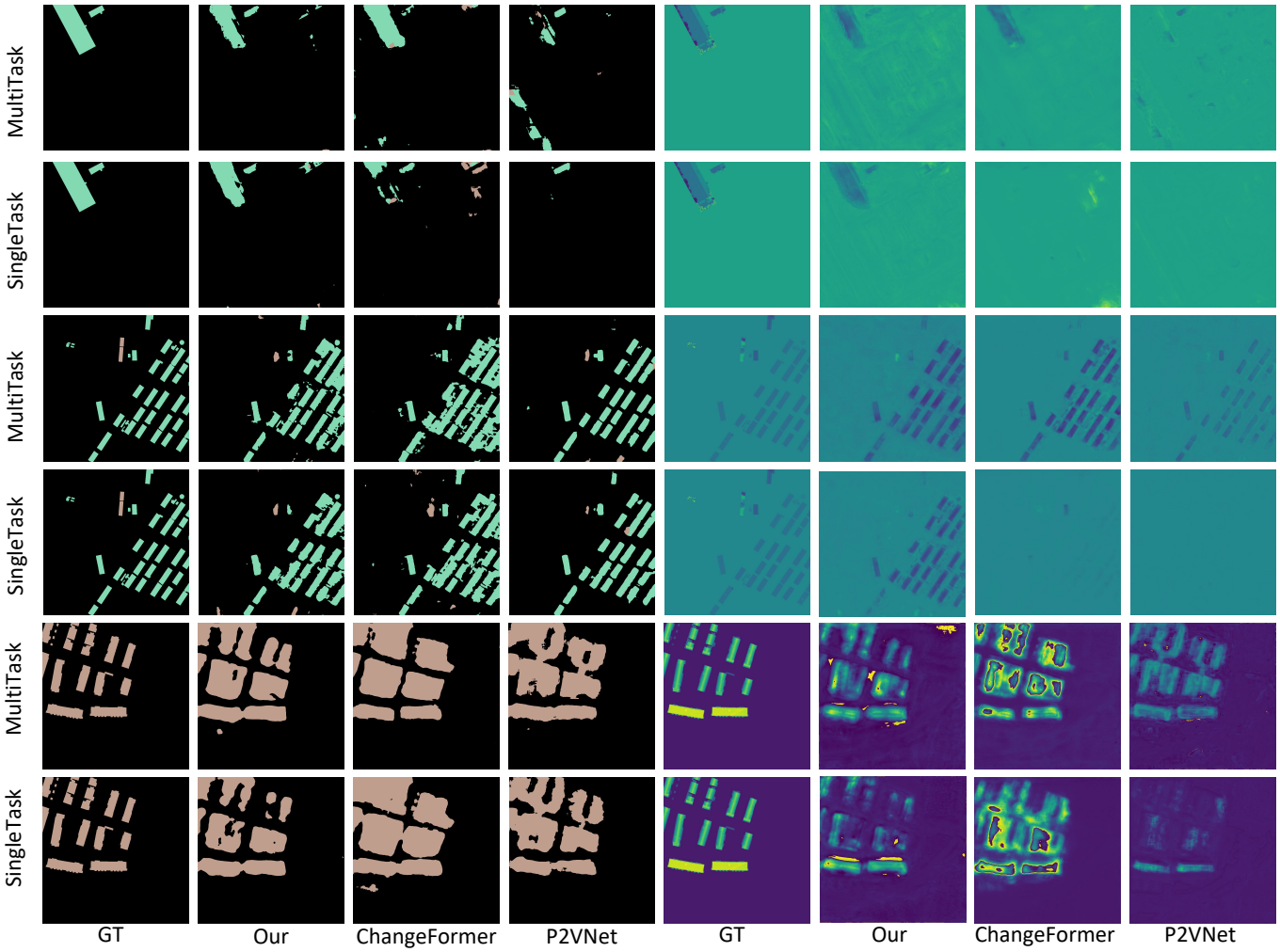

Fig. 10. Visual comparison of semantic (first four columns) and height (last four columns) changes for the top-three methods in single-task (only semantic or height branch) and multitask change detection settings. Note that height changes of our model are from the consistency augmented height branch corresponding to row 3 of Table V.

versely, the height metric results from rows 2 and 3 demonstrate the utility of semantic hints for estimating in enhancing the height changes estimation, even with only implicit shared information in the common backbone. Figure 15(b)(c)(e)(f) illustrates some visual examples wherein the semantic branch exhibits improved recovery of height changes. However, it tends to introduce additional noise in the background regions. From the odd-numbered rows of columns (c) and (f), we can observe that the attention regions corresponding to the semantic output closely resemble height changes. This suggests a strong coupling between their learned representations, which is the reason why the inclusion of the height branch poses challenges for semantic change detection, resulting in suboptimal performance in both tasks, as demonstrated in Figure 16(b)(c)(e)(f). To alleviate the problem, we designed the pseudo-change branch for two purposes:

1) In the absence of semantic change hints, it serves as a sub-complete semantic map to assist in the height change

TABLE V
ABLATION STUDY ABOUT THE IMPACT OF MULTITASK CONSISTENCY (+MC) ON HEIGHT CHANGE PREDICTING BRANCH (+3D) AND SEMANTIC CHANGE DETECTION (+2D) IN SINGLE-TASK OR MULTITASK SCENARIOS. THE '-' SYMBOL INDICATES THAT THE ABLATED MODEL DOES NOT CONTAIN THE CORRESPONDING COMPONENT TO GET THE METRIC OUTPUTS.

Settings			Semantic metrics				Height metrics				
+2d	+3d	+MC	IoU_D	IoU_N	mIoU	F1score	RMSE	MAE	CRMSE	cRel	ZNCC
✓			43.76	39.10	41.43	58.55	-	-	-	-	-
	✓		-	-	-	-	1.460	0.301	8.289	2.075	0.311
✓	✓		39.05	37.31	38.18	55.26	1.367	0.358	8.875	1.922	0.311
	✓	✓	-	-	-	-	1.273	0.397	8.317	2.711	0.379
✓	✓	✓	44.29	40.90	42.59	59.72	1.267	0.290	8.281	1.900	0.394

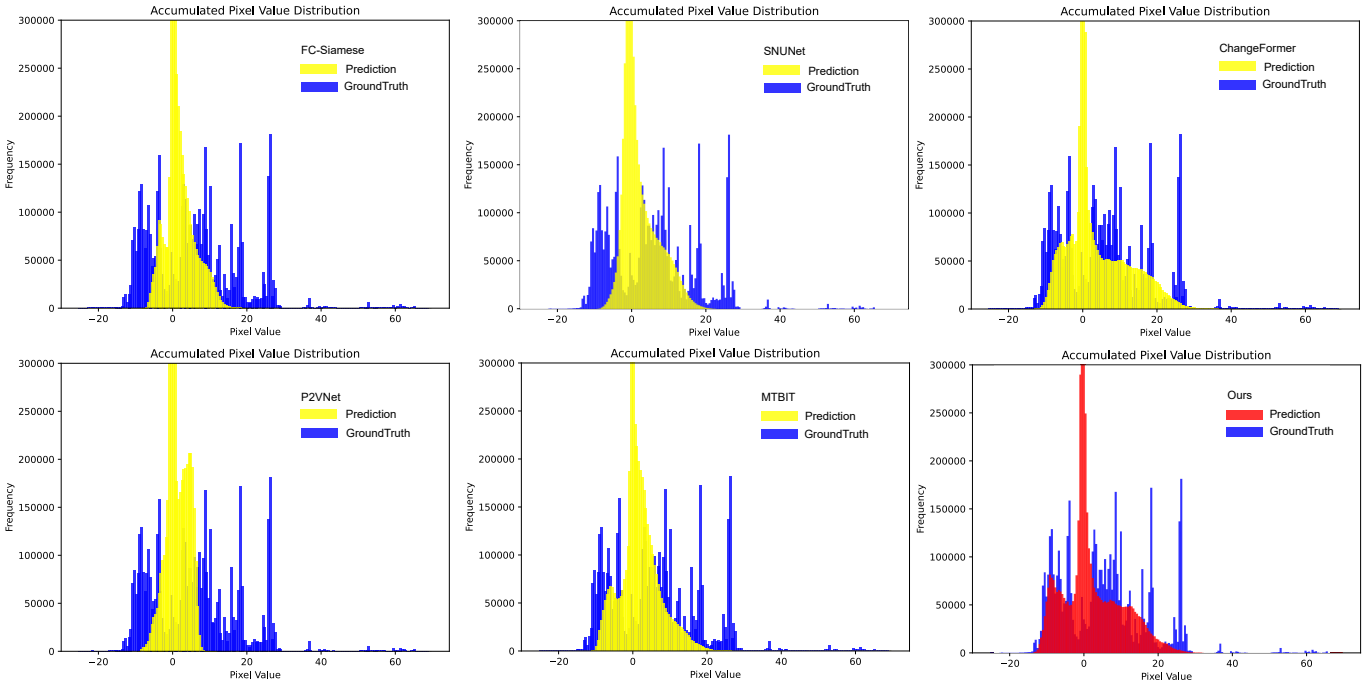

Fig. 11. The comparison of predicted height distribution, where blue denotes ground truth, yellow and red denote the other methods and ours.

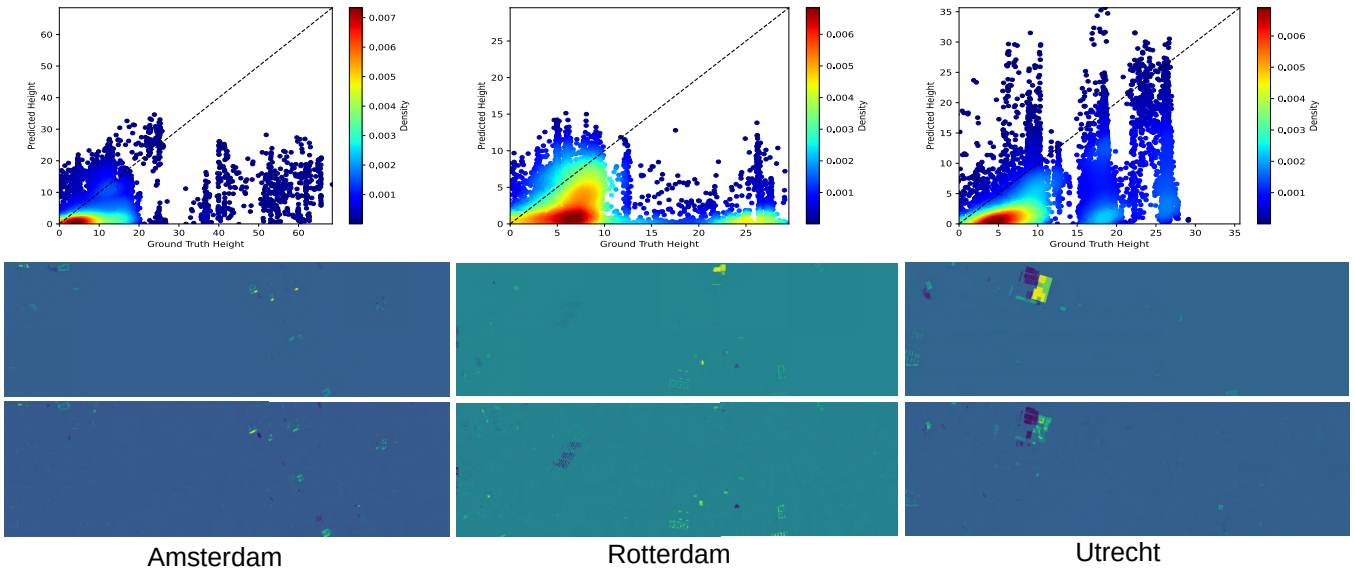

Fig. 12. Large scale visualization of height change predictions in three cities (please enlarge for more details). First row: the scatter between ground truth and predicted heights with density coloring map. Second row: ground truth of the whole test set. Third row: the predictions of our model.

detection, resulting in an augmented height detection branch. Comparing row 2 to row 4 in Table V, we found that the semantic hint from the pseudo change branch is even better than the original semantic branch. We speculate that it was attributed to the explicit soft thresholding process, which brings stronger prior about multitask relationship than the limited hints from shared backbone. The odd-numbered rows of the final column in Figure 15 and 16 provide a visualization of learned representation in the soft thresholding layer, which accurately locates some edge cases that were missed by the original multitask scenario.

2) With both semantic and height change branches in place, the pseudo change branch fosters a consistent relationship between the two tasks. Interestingly, the final row of Table V demonstrates that our consistency strategy has not only mitigated the multitask conflicts but also encouraged further improvements via explicit multitask interaction, as shown in Figure 15(d)(g) and 16(d)(g). Furthermore, our consistency strategy can be seamlessly applied to the other change detection methods and yields promising improvements, as shown

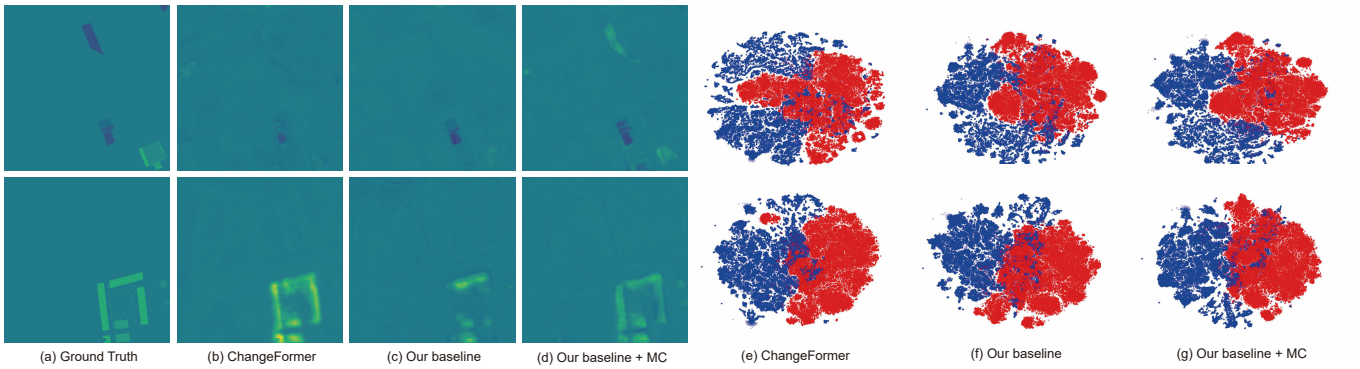

Fig. 13. The feature space of the DSM and image branches visualized by t-SNE. The red indicates an image feature point and the blue indicates a DSM feature point.

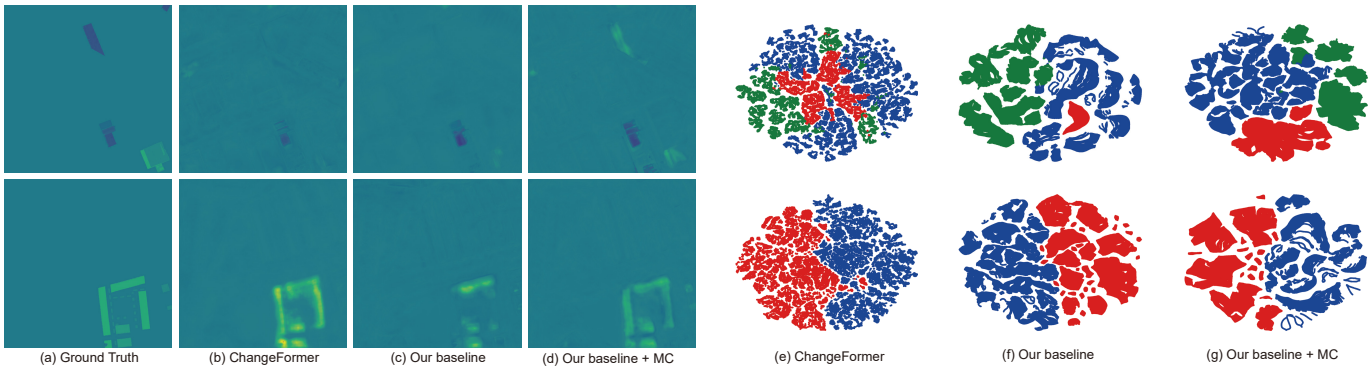

Fig. 14. The feature space of last decoder layer visualized by t-SNE. The red indicates a new-built building pixel, green indicates a demolished building pixel, and blue indicates an unchanged pixel.

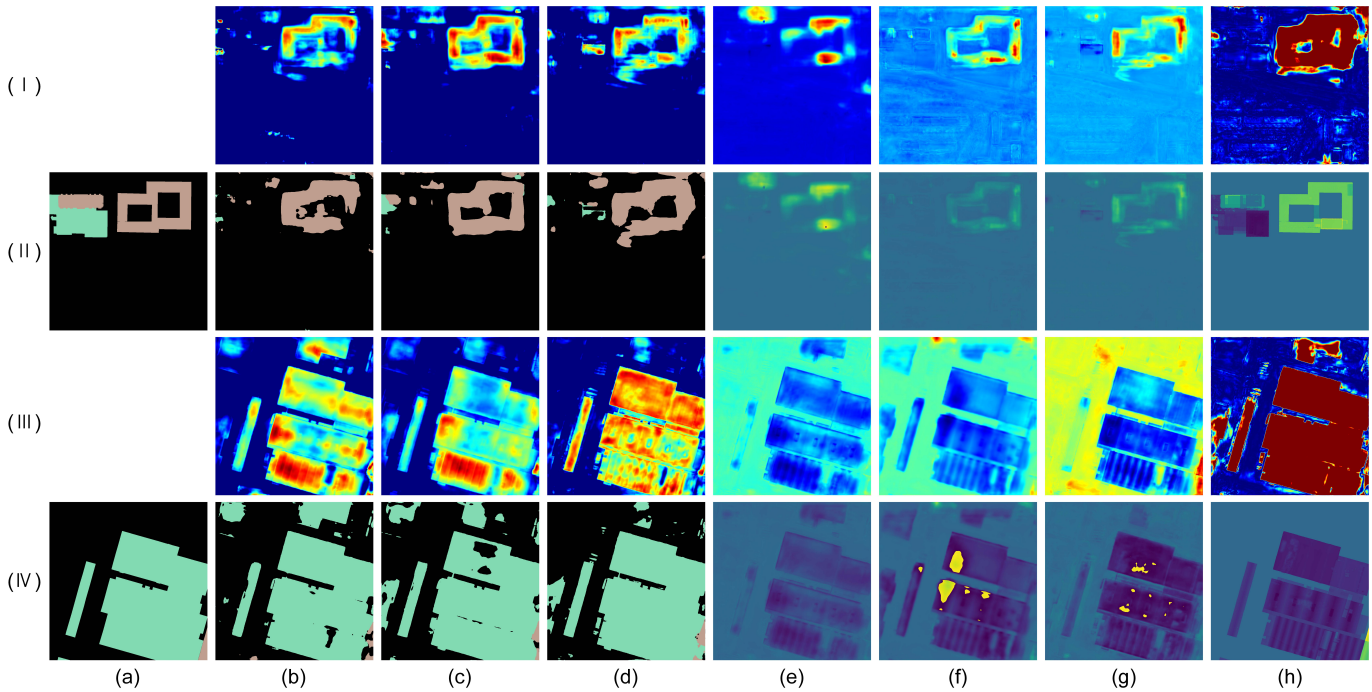

Fig. 15. Visual examples highlighting the improvement via multitask interaction. (a) semantic ground truth; (b)(e) single-task results and corresponding attention maps; (c)(f) multitask results and corresponding attention maps; (d)(g) multitask results and corresponding attention maps with consistency constraint; (h) attention maps for soft thresholding layer (odd-numbered rows) and ground truth height changes (even-numbered rows).

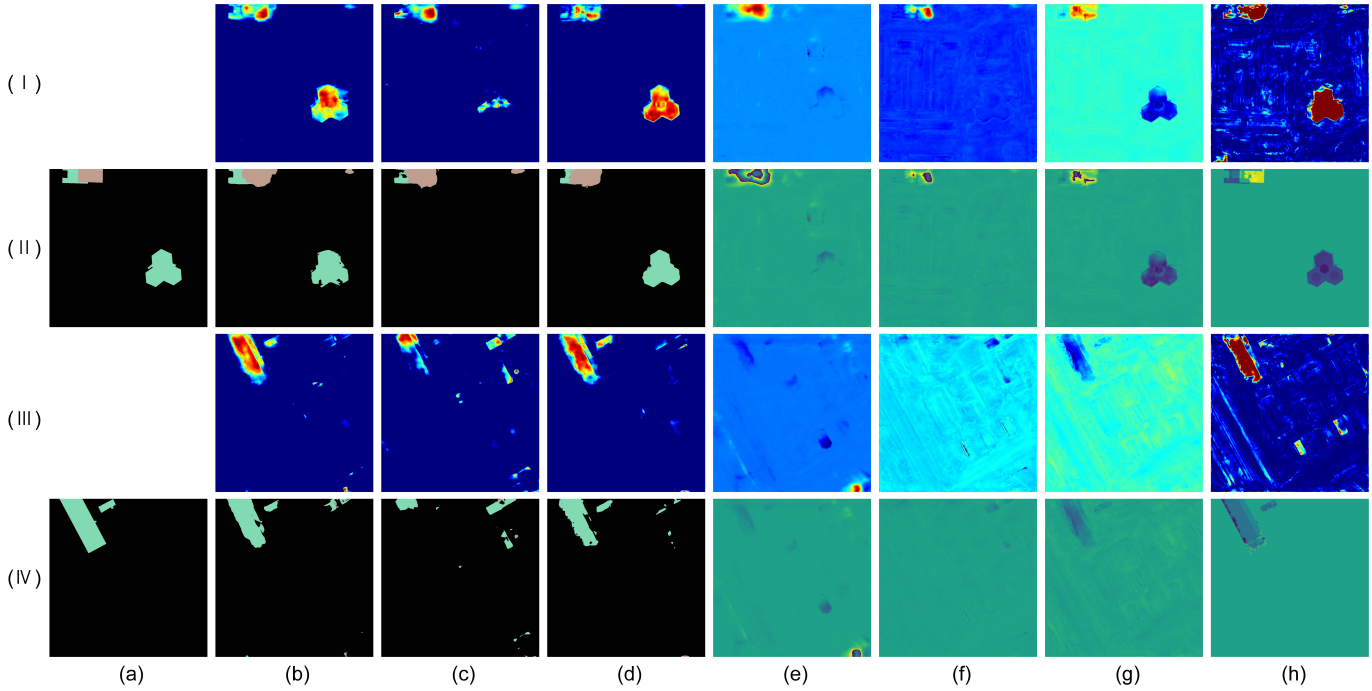

Fig. 16. Visual examples highlighting the decline due to multitask branches. (a) semantic ground truth; (b)(e) single-task results and corresponding attention maps; (c)(f) multitask results and corresponding attention maps; (d)(g) multitask results and corresponding attention maps with consistency constraint; (h) attention maps for soft thresholding layer (odd-numbered rows) and ground truth height changes (even-numbered rows).

in Table VI.

To further investigate the impact of our consistency constraint on height change detection, We adjusted the temperature parameter of equation 6. Figure 17 shows that a smaller temperature value leads to a sharper transition near zero and greater overlap between the ground truth and predicted height change. This is because the sharp transition near zero suppresses background noise where height is unchanged, allowing more attention to changed regions. However, this parameter acts as a double-edged sword; a too-small value means that background noise is more likely to be mistaken for a change target. Therefore, in our experiments, we set t to 0.5 as the final setting.

By setting the λ_3 parameter of the semantic branch to 0.6, we explore various parameter combinations within Table VII. In the first row, equating the loss weights for the height and semantic change branches results in a degradation of semantic change detection performance, with minimal improvement observed in the height change metrics. Conversely, increasing λ_1 for the pseudo change branch enhances the semantic change detection metrics, underscoring the task's emphasis, albeit at the cost of diminished height change metrics. When λ_1 to λ_3 are equalized, there's a notable improvement in height metrics compared to the configuration in row 2. Nonetheless, this increased weight on the pseudo change branch continues to adversely impact the height change detection performance relative to the baseline established in row 1. For further details, please consult Table 5 in our revised manuscript.

D. Discussion

The experiment results strongly emphasize the multitask conflicts between 2D semantic change detection and 3D height estimation. Specifically, the semantic change exhibits a distinct boundary that aids in pinpointing changes compared to height change estimation. A similar phenomenon has been documented in [52], which highlights that object boundaries are more readily discerned from segmentation labels than from depth maps. From Figure 15 and 16, we could observe the

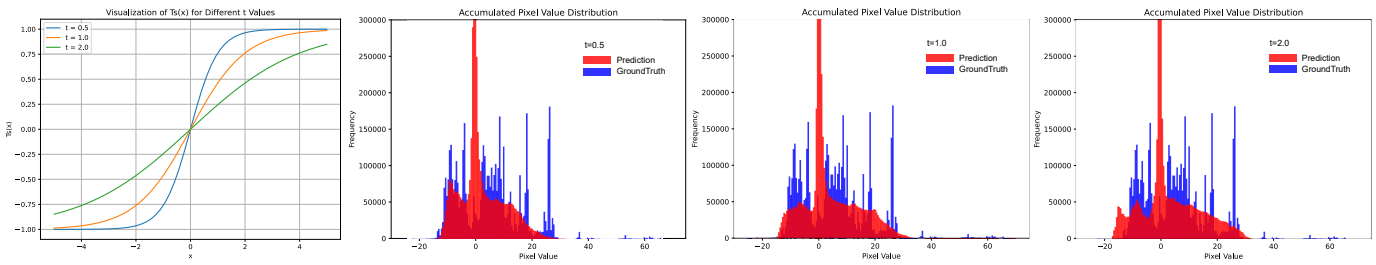

Fig. 17. The impact of different t values on height change detection.

TABLE VI
THE IMPACT OF OUR CONSISTENCY STRATEGY ON THE OTHER METHODS, INCLUDING ONE CNN-BASED AND TWO TRANSFORMER-BASED CHANGE DETECTION MODELS. THE **bold** FONT INDICATES BETTER RESULTS.

Settings	Height metrics					Semantic metrics	
	RMSE	MAE	cRMSE	cRel	ZNCC	mIoU↑	F1-score↑
FC-siamese	1.461	**0.309**	8.995	**1.506**	0.373	27.97	43.72
+multitask consistency	**1.398**	0.353	**8.655**	1.873	**0.395**	**28.85**	**44.78**
ChangeFormer	1.343	0.402	8.204	2.485	0.394	40.17	57.31
+multitask consistency	**1.317**	**0.297**	**8.085**	**1.825**	**0.447**	**41.37**	**58.39**
MTBIT	1.457	0.400	**8.563**	1.987	**0.345**	30.72	46.88
+multitask consistency	**1.373**	**0.343**	8.788	**1.639**	0.281	**32.44**	**48.90**

TABLE VII
THE INFLUENCE OF DIFFERENT COMBINATIONS OF HYPERPARAMETERS.

Hyperparameters	semantic metrics		height metrics	
	mIoU	F1-score	RMSE	MAE
$\lambda_1 = 0.2, \lambda_2 = 0.6, \lambda_3 = 0.6$	41.42	58.53	1.263	0.289
$\lambda_1 = 0.6, \lambda_2 = 0.2, \lambda_3 = 0.6$	43.24	60.22	1.302	0.388
$\lambda_1 = 0.6, \lambda_2 = 0.6, \lambda_3 = 0.6$	42.38	59.49	1.289	0.301
$\lambda_1 = 0.2, \lambda_2 = 0.2, \lambda_3 = 0.6$	42.59	59.72	1.267	0.290

inherent consistency between height change and the activation map of semantic change within a single testing example. This suggests that the decreased performance of the multitask setting is mainly caused by the height branch. As denoted in [44], jointly addressing semantic segmentation and depth estimation tasks tends to yield suboptimal results compared to single-task settings.

Our proposed multitask consistency constraint links the height change branch and pseudo-change branch via soft-thresholding. By minimizing the disparity between pseudo-change and semantic change, we enable gradient interaction from the semantic change map to the height change map. This establishes a coherent objective for the multitask branches and ultimately enhances the performance of both tasks, as evidenced in our experiments.

Our approach incorporates DSM data derived from point cloud data. Looking ahead, we aim to enhance change detection between point cloud and image data. This will involve refining representation learning methods that learn visual representations from multimodal data (point clouds and images) without extensive labeled datasets [69], [70]. Additionally, the techniques of multi-scale feature fusion [71] between point cloud and image data is pivotal for effective multimodal change detection.

VI. CONCLUSION

The prevailing direction in change detection research is toward achieving higher frequency, finer granularity, and increased dimensionality. However, there exists a noticeable gap in the literature about multimodal and cross-dimensional change detection. In this paper, we presented a novel pipeline for detecting height and semantic change simultaneously from DSM-to-image multimodal data. We revealed that the leading change detection methods, including CNN-based and Transformer-based methods, struggled with the conflicts of multitask change detection. We proposed a Transformer-based network equipped with multitask consistency constraint, which achieves the best semantic and height change detection performance with limited model complexity. We found that the consistency strategy with a small temperature parameter is able to suppress the background noise and leads to sharper results in change regions. It can be also seamlessly employed to the other change detection methods and produce promising improvements. Our dataset and model are poised to become foundational benchmarks for future research within the remote sensing change detection community. This work paves the way for a range of intriguing research avenues, such as the exploration of more fine-grained semantic categorizations based on varying scales of height changes, and the integration of currently dominant large pre-trained models into our framework.

However, several unresolved questions persist, including both data and algorithmic aspects. Firstly, we have initiated a dataset expansion plan, necessitating enhanced automation in our annotation workflow to accommodate large-scale applications. Furthermore, as aforementioned in Section V-B, we need to delve deeper into addressing the substantial class imbalance issue with various strategies, including refining training metrics, implementing data augmentation techniques, and optimizing the architecture. Additionally, leveraging the power of state-of-the-art large pre-trained models could lead to a more efficient training process and improved generalization.

REFERENCES

[1] A. Toker, L. Kondmann, M. Weber, M. Eisenberger, A. Camero, J. Hu, A. P. Hoderlein, Ç. Şenaras, T. Davis, D. Cremers, et al., Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21158–21167.

[2] S. Tian, Y. Zhong, Z. Zheng, A. Ma, X. Tan, L. Zhang, Large-scale deep learning based binary and semantic change detection in ultra high resolution remote sensing imagery: From benchmark datasets to urban application, ISPRS Journal of Photogrammetry and Remote Sensing 193 (2022) 164–186.

[3] M. Cserép, R. Lindenbergh, Distributed processing of dutch ahn laser altimetry changes of the built-up area, International Journal of Applied Earth Observation and Geoinformation 116 (2023) 103174.

[4] U. Stilla, Y. Xu, Change detection of urban objects using 3d point clouds: A review, ISPRS Journal of Photogrammetry and Remote Sensing 197 (2023) 228–255.

[5] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, X. Qiu, Change detection based on deep siamese convolutional network for optical aerial images, IEEE Geoscience and Remote Sensing Letters 14 (10) (2017) 1845–1849. doi:10.1109/LGRS.2017.2738149.

[6] S. Zagoruyko, N. Komodakis, Learning to compare image patches via convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4353–4361.

[7] Y. Quan, A. Yu, W. Guo, X. Lu, B. Jiang, S. Zheng, P. He, Unified building change detection pre-training method with masked semantic annotations, International Journal of Applied Earth Observation and Geoinformation 120 (2023) 103346.

[8] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, X. Huang, Semicdnet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images, IEEE Transactions on Geoscience and Remote Sensing 59 (7) (2020) 5891–5906.

[9] H. Chen, Z. Qi, Z. Shi, Remote sensing image change detection with transformers, IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–14.

[10] C. Zhang, L. Wang, S. Cheng, Y. Li, Swinsunet: Pure transformer network for remote sensing image change detection, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–13.

[11] R. Touati, M. Mignotte, An energy-based model encoding nonlocal pairwise pixel interactions for multisensor change detection, IEEE Transactions on Geoscience and Remote Sensing 56 (2) (2017) 1046–1058.

[12] H. Chen, N. Yokoya, C. Wu, B. Du, Unsupervised multimodal change detection based on structural relationship graph representation learning, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–18.

[13] Z. Zhang, G. Vosselman, M. Gerke, C. Persello, D. Tuia, M. Yang, Change detection between digital surface models from airborne laser scanning and dense image matching using convolutional neural networks, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4 (2019) 453–460.

[14] R. Touati, M. Mignotte, M. Dahmane, Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based markov random field model, IEEE Transactions on Image Processing 29 (2019) 757–767.

[15] G. Fodor, M. V. Conde, Rapid deforestation and burned area detection using deep multimodal learning on satellite imagery, arXiv preprint arXiv:2307.04916 (2023).

[16] Z. Zhang, G. Vosselman, M. Gerke, C. Persello, D. Tuia, M. Y. Yang, Detecting building changes between airborne laser scanning and photogrammetric data, Remote sensing 11 (20) (2019) 2417.

[17] H. Zhang, M. Wang, F. Wang, G. Yang, Y. Zhang, J. Jia, S. Wang, A novel squeeze-and-excitation w-net for 2d and 3d building change detection with multi-source and multi-feature remote sensing data, Remote Sensing 13 (3) (2021) 440.

[18] V. Marsocci, V. Coletta, R. Ravanelli, S. Scardapane, M. Crespi, Inferring 3d change detection from bitemporal optical images, ISPRS Journal of Photogrammetry and Remote Sensing 196 (2023) 325–339.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[20] S. Fang, K. Li, J. Shao, Z. Li, Snunet-cd: A densely connected siamese network for change detection of vhr images, IEEE Geoscience and Remote Sensing Letters 19 (2021) 1–5.

[21] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, G. Liu, A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images, ISPRS Journal of Photogrammetry and Remote Sensing 166 (2020) 183–200.

[22] R. C. Daudt, B. Le Saux, A. Boulch, Fully convolutional siamese networks for change detection, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 4063–4067.

[23] B. Liu, H. Chen, Z. Wang, Lsnet: Extremely light-weight siamese network for change detection in remote sensing image, arXiv preprint arXiv:2201.09156 (2022).

[24] M. Lin, G. Yang, H. Zhang, Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images, IEEE Transactions on Image Processing 32 (2022) 57–71.

[25] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, P. He, Scdnet: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery, International Journal of Applied Earth Observation and Geoinformation 103 (2021) 102465.

[26] B. Qi, Q. Kun, Z. Han, H. Wenjun, L. Zhili, X. Kai, Building change detection based on multi-scale filtering and grid partition, in: 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS), IEEE, 2018, pp. 1–8.

[27] D. A. J. Quispe, J. Sulla-Torres, Automatic building change detection on aerial images using convolutional neural networks and handcrafted features, International Journal of Advanced Computer Science and Applications 11 (6) (2020).

[28] R. Qin, J. Tian, P. Reinartz, 3d change detection–approaches and applications, ISPRS Journal of Photogrammetry and Remote Sensing 122 (2016) 41–56.

[29] I. de Gélis, S. Saha, M. Shahzad, T. Corpetti, S. Lefèvre, X. X. Zhu, Deep unsupervised learning for 3d als point clouds change detection, arXiv preprint arXiv:2305.03529 (2023).

[30] V. Ferraris, N. Dobigeon, Y. Cavalcanti, T. Oberlin, M. Chabert, Coupled dictionary learning for unsupervised change detection between multimodal remote sensing images, Computer Vision and Image Understanding 189 (2019) 102817.

[31] S. Chirakkal, F. Bovolo, A. Misra, L. Bruzzone, A. Bhattacharya, Unsupervised multiclass change detection for multimodal remote sensing data, in: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2022, pp. 3223–3226.

[32] Y. Sun, L. Lei, X. Tan, D. Guan, J. Wu, G. Kuang, Structured graph based image regression for unsupervised multimodal change detection, ISPRS Journal of Photogrammetry and Remote Sensing 185 (2022) 16–31.

[33] Y. Sun, L. Lei, D. Guan, G. Kuang, Iterative robust graph for unsupervised change detection of heterogeneous remote sensing images, IEEE Transactions on Image Processing 30 (2021) 6277–6291.

[34] Y. Sun, L. Lei, X. Li, X. Tan, G. Kuang, Structure consistency-based graph for unsupervised change detection with homogeneous and heterogeneous remote sensing images, IEEE transactions on geoscience and remote sensing 60 (2021) 1–21.

[35] Y. Sun, L. Lei, D. Guan, J. Wu, G. Kuang, L. Liu, Image regression with structure cycle consistency for heterogeneous change detection, IEEE Transactions on Neural Networks and Learning Systems (2022).

[36] L. T. Luppino, M. Kampffmeyer, F. M. Bianchi, G. Moser, S. B. Serpico, R. Jenssen, S. N. Anfinsen, Deep image translation with an affinity-based change prior for unsupervised multimodal change detection, IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–22.

[37] X. Li, Z. Du, Y. Huang, Z. Tan, A deep translation (gan) based change detection network for optical and sar remote sensing images, ISPRS Journal of Photogrammetry and Remote Sensing 179 (2021) 14–34.

[38] Y. Wu, J. Li, Y. Yuan, A. Qin, Q.-G. Miao, M.-G. Gong, Commonality autoencoder: Learning common features for change detection from heterogeneous images, IEEE transactions on neural networks and learning systems 33 (9) (2021) 4257–4270.

[39] R. Shao, C. Du, H. Chen, J. Li, Sunet: Change detection for heterogeneous remote sensing images from satellite and uav using a dual-channel fully convolution network, Remote Sensing 13 (18) (2021) 3750.

[40] M. Yang, L. Jiao, F. Liu, B. Hou, S. Yang, M. Jian, Dpfl-nets: Deep pyramid feature learning networks for multiscale change detection, IEEE Transactions on Neural Networks and Learning Systems 33 (11) (2021) 6402–6416.

[41] Z. Zhang, G. Vosselman, M. Gerke, D. Tuia, M. Y. Yang, Change detection between multimodal remote sensing data using siamese cnn, arXiv preprint arXiv:1807.09562 (2018).

[42] C.-F. R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 357–366.

[43] Z. Wang, Y. Tsvetkov, O. Firat, Y. Cao, Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models, arXiv preprint arXiv:2010.05874 (2020).

[44] B. Liu, X. Liu, X. Jin, P. Stone, Q. Liu, Conflict-averse gradient descent for multi-task learning, Advances in Neural Information Processing Systems 34 (2021) 18878–18890.

[45] Z. Chen, V. Badrinarayanan, C.-Y. Lee, A. Rabinovich, Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks, in: International conference on machine learning, PMLR, 2018, pp. 794–803.

[46] A. Navon, A. Shamsian, I. Achituve, H. Maron, K. Kawaguchi, G. Chechik, E. Fetaya, Multi-task learning as a bargaining game, arXiv preprint arXiv:2202.01017 (2022).

[47] O. Sener, V. Koltun, Multi-task learning as multi-objective optimization, Advances in neural information processing systems 31 (2018).

[48] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Transactions on Knowledge and Data Engineering 34 (12) (2021) 5586–5609.

[49] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, S. Savarese, Which tasks should be learned together in multi-task learning?, in: International Conference on Machine Learning, PMLR, 2020, pp. 9120–9132.

[50] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7482–7491.

[51] L. Liu, Y. Li, Z. Kuang, J. Xue, Y. Chen, W. Yang, Q. Liao, W. Zhang, Towards impartial multi-task learning, iclr, 2021.

[52] S. Zhu, G. Brazil, X. Liu, The edge of depth: Explicit constraints between segmentation and depth, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13116–13125.
[53] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, J. Yang, Pattern-affinitive propagation across depth, surface normal and semantic segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4106–4115.
[54] Z. Zheng, Y. Zhong, S. Tian, A. Ma, L. Zhang, Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection, ISPRS Journal of Photogrammetry and Remote Sensing 183 (2022) 228–239.
[55] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, M. Pelillo, L. Zhang, Asymmetric siamese networks for semantic change detection in aerial images, IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–18.
[56] Y. Deng, J. Chen, S. Yi, A. Yue, Y. Meng, J. Chen, Y. Zhang, Feature-guided multitask change detection network, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15 (2022) 9667–9679.
[57] R. C. Daudt, B. Le Saux, A. Boulch, Y. Gousseau, Multitask learning for large-scale semantic change detection, Computer Vision and Image Understanding 187 (2019) 102783.
[58] Q. Shu, J. Pan, Z. Zhang, M. Wang, Mtcnet: Multitask consistency network with single temporal supervision for semi-supervised building change detection, International Journal of Applied Earth Observation and Geoinformation 115 (2022) 103110.
[59] W. Zhao, C. Persello, A. Stein, Semantic-aware unsupervised domain adaptation for height estimation from single-view aerial images, ISPRS Journal of Photogrammetry and Remote Sensing 196 (2023) 372–385.
[60] Z. Zhang, G. Vosselman, M. Gerke, C. Persello, D. Tuia, M. Y. Yang, Detecting building changes between airborne laser scanning and photogrammetric data, Remote sensing 11 (20) (2019) 2417.
[61] X. Sun, W. Zhao, R. V. Maretto, C. Persello, Building polygon extraction from aerial images and digital surface models with a frame field learning framework, Remote sensing 13 (22) (2021) 4700.
[62] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, pmlr, 2015, pp. 448–456.
[63] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 568–578.
[64] W. G. C. Bandara, V. M. Patel, A transformer-based siamese network for change detection, in: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2022, pp. 207–210.
[65] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[66] H. Chen, Z. Shi, A spatial-temporal attention-based method and a new dataset for remote sensing image change detection, Remote Sensing 12 (10) (2020) 1662.
[67] H. A. Amirkolaee, H. Arefi, Height estimation from single aerial images using a deep convolutional encoder-decoder network, ISPRS journal of photogrammetry and remote sensing 149 (2019) 50–66.
[68] B.-Y. Liu, H.-X. Chen, Z. Huang, X. Liu, Y.-Z. Yang, Zoominnet: A novel small object detector in drone images with cross-scale knowledge distillation, Remote Sensing 13 (6) (2021) 1198.
[69] Y. Wu, J. Liu, M. Gong, P. Gong, X. Fan, A. Qin, Q. Miao, W. Ma, Self-supervised intra-modal and cross-modal contrastive learning for point cloud understanding, IEEE Transactions on Multimedia (2023).
[70] Y. Wu, X. Hu, Y. Zhang, M. Gong, W. Ma, Q. Miao, Sacf-net: Skip-attention based correspondence filtering network for point cloud registration, IEEE Transactions on Circuits and Systems for Video Technology (2023).
[71] Y. Wu, Q. Yao, X. Fan, M. Gong, W. Ma, Q. Miao, Panet: A point-attention based multi-scale feature fusion network for point cloud registration, IEEE Transactions on Instrumentation and Measurement (2023).