Boosting Binary Object Change Detection via Unpaired Image Prototypes Contrast

Mingwei Zhang, Qiang Li, *Member, IEEE*, Yuan Yuan, *Senior Member, IEEE,* and Qi Wang, *Senior Member, IEEE,*

***Abstract*—Binary object change detection aims to monitor the evolution of the object of interest in a fixed region. Constructing a relevant dataset for deep learning models is strenuous. In the existing datasets, there is usually an imbalance between changed and unchanged samples, as well as a restricted diversity within the changed samples. Aiming at that, some methods utilize unpaired images used for object segmentation to generate pseudo-bitemporal images for change detection. However, due to the existence of the domain gap between different data sources, the model obtained by these methods can not well generalize to the real bitemporal images. Inspired by them but to avoid the domain difference, we explore how to directly use the unpaired images within a real change detection dataset to complement changed samples. In detail, a concise metric-based framework is designed, which consists of two branches, a projector and a predictor. The framework obtains the change map by computing the distance between the bitemporal embedding outputted by the projector. Meanwhile, instructed by an indirect semantic supervision module (ISSM) specially designed, the predictor can generate the semantic confidence map distinguishing the pixels in an image into two categories. Based on the output of the framework, an unpaired image prototype contrast module (UIPCM) is proposed. It enriches the diversity of the change samples for training by combining the prototypes in unpaired images at the feature level, leading to alleviating the imbalance between changed and unchanged samples. Besides, a dual margin contrastive loss (DMCL) is adopted during training. It can reduce the constraint on the consistency of bitemporal embedding in unchanged regions. The benefits and the superiority of the proposed method are demonstrated on two well-recognized datasets. The code is available at https://github.com/ptdoge/UIPC.**

***Index Terms*—Object Change Detection, Indirect Semantic Supervision, Unpaired Image Prototype Contrast**

I. Introduction

OBJECT change detection aims to detect the change of the object of interest on the earth surface through bitemporal images, which are acquired from one area on two periods. Different from the general binary change detection, it not only provides the position of the change but also gives the type of the change. It can be finely divided into two tasks: binary object change detection and semantic change detection. The former focuses on one specific type of change, e.g., building change detection. The latter requires relevant algorithms to describe the ”from-to” types of changes, e.g., land cover change detection. Owing to the generated fine-grained object change information, object change detection has been used for disaster rescue [1], [2], ecological protection [3], and urban development [4].

With the availability of large amounts of annotated data, many object change detection methods based on the convolutional neural network (CNN) [5]–[7] or the vision Transformer (ViT) [8]–[10] have been designed. Notably, for the annotated data used for binary object change detection, only the state of the pixel in the bitemporal images is usually given [11], [12], i.e., change or non-change. Given a pair of bitemporal images, the methods developed for binary object change detection obtain only one change map (CM), which suggests whether the object of interest has been changed. Meanwhile, for the annotated data used for semantic change detection, in addition to the state label, the semantic class of the pixel is also provided [13]–[15]. Therefore, the approaches designed for semantic change detection are usually based on a multi-task learning framework [13], [16]–[18], where the main task is to detect the changed regions and the sub-task is to predict the semantic category of the pixel, i.e., land cover mapping. No doubt simultaneously learning land cover types is critical for accurate semantic change detection. In contrast, high-accuracy binary object change identification can be achieved with only given state labels. Recent studies show that introducing the knowledge of the object segmentation task can further improve its accuracy.

For example, Liu *et al.* [19] adopt the model pre-trained on building extraction datasets to transfer building-relevant knowledge, which significantly improves the accuracy of building change detection. Zheng *et al.* [20] employ the strategy of joint semantic prediction and change detection to enhance the object change embedding during the training. However, these methods require fine-grained pixel-level annotations of the objects of interest in each image. Employing extra data to assist binary object change detection brings an increased annotation burden, which is labor-intensive and time-consuming. Thus, towards binary object change detection, *is it possible to predict the semantics of pixels in bitemporal images with only state labels, thereby improving the representation ability of the model to the object of interest?* In this work, a new perspective is given to answer the question with certainty.

For binary object change detection, there is usually a fundamental assumption that the bitemporal pixels own different categories in the changed areas [21]. One is the class to which the object of interest belongs. The other is unified

Mingwei Zhang is with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi’an 710072, P.R. China (e-mail: dlaizmw@gmail.com). (*Corresponding author: Qi Wang*)

Qiang Li, Yuan Yuan, and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi’an 710072, P.R. China (e-mail:liqmges@gmail.com, y.yuan1.ieee@gmail.com, crabwq@gmail.com).

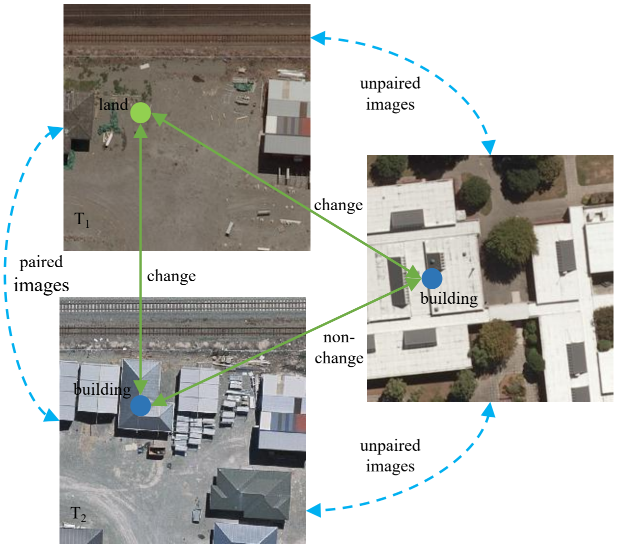

Fig. 1. A change detector is usually trained by paired bitemporal images, i.e., the images acquired in the same geographical location at different times. In this work, the potential of unpaired images is explored for change detection, where the objects with different semantics can construct the changed samples. The unpaired images refer to the images with different contents and geographical locations.

as the other class. The state label on the changed position can indirectly represent the semantic label of the pixel, but it does not explicitly indicate which image of the bitemporal images the object of interest is in. To this end, a simple indirect semantic supervision module (ISSM) is designed to distinguish the two classes of pixels, i.e., the object of interest and others. *Why discern the types of pixels?* First, as in the previous studies, learning to predict the class of the pixel aims to enhance the distinction of the object embedding. Second, as shown in Fig. 1, if the semantic classes of the pixels can be judged, both the changed and unchanged samples can be massively produced via the interaction between the unpaired image pixels. Therefore, it is beneficial to alleviate the negative impacts of the imbalance between the changed and unchanged samples, e.g. the weak generalization of the model. Especially, the generated samples can be beyond the limitation of the same space at different times. This promising idea is first explored in the STAR [20], which aims to relieve the burdens of collecting bitemporal images and labeling changed pixels. It adopts single temporal building detection datasets to construct unpaired images for building change detection, which breaks the constraint that change representations must be learned from the bitemporal images corresponding to the same geographical area. However, it is founded on the refined annotation of large amounts of object instances, which also requires expensive labeling costs. Meanwhile, due to the domain gap between the data used for object segmentation and that to identify changes, the generalization of the model obtained through cross-domain learning is weak.

To utilize the benefits of the relation between the unpaired images to binary object change detection, this work explores how to generate changed samples via the unpaired images and the state labels in the given change detection dataset, which can explicitly avoid introducing the domain difference of different data sources. In detail, a metric-based framework is proposed, which includes two branches, a projector, and a predictor. The distance between the bitemporal features outputted by the projector is used to generate the CM. The output of the predictor is a semantic confidence map (SCM), which divides the pixels in an image into two categories with the constraint of the ISSM. On this basis, an unpaired image prototype contrast module (UIPCM) is designed. First, the prototype representation of the pixels corresponding to the object of interest and the others are obtained by aggregating the feature from the projector, where the categories of the pixels are indicated by the SCM from the predictor. Then, the change representations are generated by the interaction of the unpaired image prototypes. Different from constructing pseudo bitemporal images in previous studies, UIPCM newly generates the change embedding at the feature level, which indirectly complements the changed samples. In summary, the contributions made by this article are as follows:

- A concise metric-based binary object change detection framework is proposed to help model the semantic relation between the pixels of the unpaired images in a given change detection dataset, where only state labels of the pixels are provided. It consists of a projector branch and a predictor branch, which are responsible for representation extraction and semantic prediction respectively.
- An indirect semantic supervision module is developed, which is no-parametric. It can make the predictor in the proposed framework distinguish the pixels in bitemporal images into two classes without requiring the annotations of all object instances. Meanwhile, it can facilitate the enhancement of the representation discrimination of the object of interest.
- An unpaired image prototype contrast module is proposed. Based on the output of the proposed framework, it enriches object change representation via the interaction between the prototype embedding of the pixels with different classes in unpaired images, which does not introduce the domain gap of different data sources. The accuracy of the framework is effectively improved by it.

II. Related Work

A. Object Change Detection

Object change detection consists of two types, binary object change detection and semantic change detection. The representative task in the former is building change detection, for which there are many methods specially designed. Zhang *et al.* [22] propose a framework compatible with multiple attention mechanisms, aiming to fully exploit deep features to meet the needs of building change detection under multiple modalities and complex scenes [23], [24]. Liu *et al.* [25] propose a dual-task constrained Siamese network, which contains a building change detection sub-network and two building extraction sub-networks. This approach can help the model learn more discriminative object-level features, thus obtaining a more complete CM. Chen *et al.* [26] focus on the accuracy of edge prediction and the integrity of the identified changed areas to propose an edge-guided network EGDENet. Similarly, Bai *et al.* [27] propose an edge-guided recurrent convolutional neural network EGRCNN. Different from only studying the specific object of interest in binary object change detection, semantic change detection aims to identify multiple types of change. The relevant works mainly explore how to efficiently utilize the semantic information provided by the given pixel-wise semantic labels for the accuracy improvement of change identification. Ding *et al.* [28] propose a semantic change transformer to model the semantic transitions between the bi-temporal images. Besides, they design a CNN architecture in [29], where semantic temporal features are merged in the change detection unit. Zheng *et al.* [16] propose a deep multi-task encoder-transformer-decoder model, where change representation is learned from obtained bitemporal semantic representation. Inspired by the utilization of semantic information in the above works but differently, we investigate the benefits of semantic analysis with only the given state label towards binary object change detection.

B. Synthetic Data Change Detection

Aiming at the difficulty of collecting bitemporal images with changed objects of interest, there have been some data synthetic methods proposed recently. For example, Zheng *et al.* [20] design a single-temporal binary object change detection method, which uses single-temporal unpaired images with fine building instance annotations to train a building change detector. Similarly, Minseok *et al.* [21] employ spatially uncorrelated images with semantic labels to construct pseudo-bitemporal images. Meanwhile, they further present how to use single images to generate more diverse and realistic paired images. Hou *et al.* [30] propose a stable prototype-guided single-temporal supervised learning framework. This framework introduces a synthetic method of pseudo-bitemporal images that simulates object changes, background interference, seasonal variation, and lighting changes in real scenes. These methods effectively alleviate the problem that bitemporal images with changed samples are difficult to collect. Due to the existence of the domain difference between the generated data and the real bitemporal images, the accuracy of the obtained models is lower than the methods that directly adopt labeled bitemporal images for training. Furthermore, they ignore that image semantic annotation is still costly. Nonetheless, these methods inspire us to investigate the help of unpaired images for binary object change detection. In detail, this work explores how to effectively complement changed samples by using the state labels and spatially unpaired images in a change detection dataset, where the domain gap is avoided naturally.

C. Prototype Learning

Prototype Learning is a classical strategy in pattern recognition for classification or regression, etc. At early, the representative methods include K-nearest-neighbor (KNN) [31], learning vector quantization (LVQ) [32], and K-means clustering [33]. Later, prototype learning is combined with deep learning, which brings many new perspectives. Yang *et al.* [34] introduce convolutional prototype learning as a novel framework to enhance the robustness of CNNs by replacing the softmax layer with prototypes. Similarly, Zhou *et al.* [35] propose a non-parametric semantic segmentation method using non-learnable prototypes, which shows superior performance over parametric methods across various network architectures. Meanwhile, Zhou *et al.* [36] employ the region embedding prototypes to achieve semantic aggregation for weakly supervised semantic segmentation. Besides, prototype learning has shown remarkable potential in few-shot and zero-shot learning [37]–[39]. Fu *et al.* [40] address problems like hubness and domain shift on zero-shot learning by introducing a semantic class prototype graph. Liu *et al.* [41] explore the generalized few-shot semantic segmentation by learning projection onto orthogonal prototypes. In this work, we investigate an image prototype contrast strategy for binary object change detection, where the image prototype is obtained by aggregating the pixel representation with the same category in an image.

III. Methodology

A. Method Overview

The overall pipeline of the proposed binary object change detection framework is shown in Fig. 2(a), which includes a feature extractor (FE), a projector, and a predictor. The feature extractor adopts the ResNet-18 [42] model pre-trained on the ImageNet [43] dataset. It extracts multi-level features $\{F^j, j \in \{1,2,3,4,5\}\}$ from the inputted image, where the size of the feature is $1/2^i$ of that of the original image respectively. High-level features own rich semantic information and a large receptive field. Meanwhile, shallow-level features retain many texture details. The features from the second to fifth levels are fed into the projector and the predictor. The projector aims to generate fused discriminative features for change identification. As in our previous works [44], [45], the features are first transformed into the same size in the channel and spatial dimension. Then they are concatenated for fusion, which effectively utilizes the complementary information at different levels. Differently, the predictor aims to distinguish the semantic categories of the pixels, i.e., the object of interest or others. Thus, it first obtains the fused feature as that in the projector. The fused feature is then further mapped into a SCM, where the value of the pixels ranges from 0 to 1.

The proposed framework is responsible for the representation extraction and the semantic prediction of the inputted image simultaneously. On this basis, the UIPCM is introduced to utilize the representation F from the projector and the SCM from the predictor for the performance improvement of this framework. Especially, the predictor and the UIPCM are only introduced during training. This framework generates the CM by the binarization of the distance map (DM) computed from the bitemporal representation (F_{T_1}, F_{T_2}) of the bitemporal images outputted by the projector, which can be formulated as

$$\mathrm{DM} = (\|F_{T_1} - F_{T_2}\|_2)_{\uparrow_4}, \tag{1}$$

$$\mathrm{CM}(i,j) = \begin{cases} 1, & \text{if } \mathrm{DM}(i,j) > T \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where i and j indicate the i-th row and j-th column in the map. T is the given threshold for binarization.

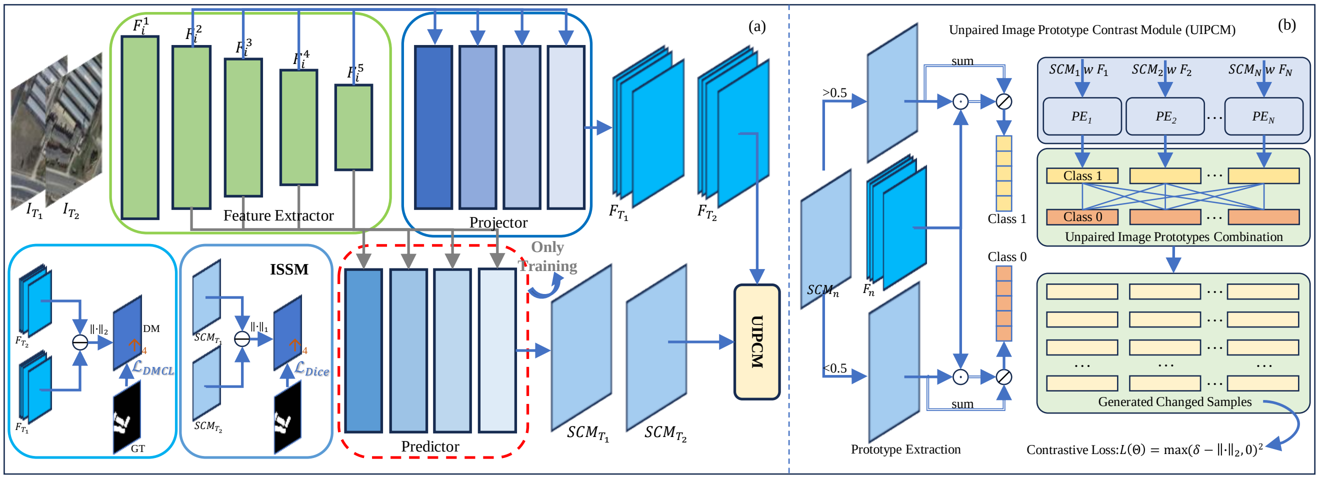

Fig. 2. (a) The overall pipeline of the proposed binary object change detection framework. The framework includes a projector branch and a predictor branch, where the change map is obtained by comparing the bitemporal embedding (F_{T_1}, F_{T_2}) of the projector. Meanwhile, the predictor aims to yield the SCM that can discern the pixels with the same class, which is achieved under the effect of the ISSM. The predictor is employed only for training. (b) The schematic of the UIPCM. First, it extracts the image prototypes by utilizing the output of the projector and that of the predictor. Then, it combines the prototypes in different images for the generation of the fresh change representation. Finally, the magnitude of the generated representation is optimized by contrastive metric to improve the performance of the proposed framework.

B. Indirect Semantic Supervision Module

Recent studies show introducing semantic prediction can help improve the performance of the model for binary object change detection. However, previous methods usually require extra pixel-wise annotation of the object of interest to supervise the learning of the semantic prediction branch. To avoid introducing the increased burden of data annotation, we explore how to use the state label of the bitemporal pixels to learn their semantic categories. The relation between the two can be described as follows

$$c_i = p_i^1 \oplus p_i^2, \tag{3}$$

where $c_i \in \{0,1\}$ indicates the ground truth (GT) state of the bitemporal pixels, i.e., change or non-change. $p_i \in \{0,1\}$ represents the semantic category of the pixel i. $\oplus$ is the xor operation. The above process can be further transformed as

$$c_i = \|p_i^1 - p_i^2\|_1. \tag{4}$$

The ISSM is designed based on this inspired transformation. Its structure is shown in Fig. 2(a). According to Eq. 4, the absolute difference between the SCM_{T_1} and the SCM_{T_2} outputted by the predictor can be regarded as the change confidence map of the bitemporal images,

$$\hat{c}_i = \|\hat{p}_i^1 - \hat{p}_i^2\|_1, \tag{5}$$

where $\hat{p}_i$ indicate the semantic confidence of the pixel i. Without bells and whistles, the c_i is utilized to instruct the learning of the $\hat{c}_i$. From Eqs. 4 and 5, if c_i is equal to 1, the $\hat{p}_i^1$ and $\hat{p}_i^2$ will tend to 0 or 1 respectively. The $\hat{p}_i^1$ and $\hat{p}_i^2$ in unchanged regions (i.e., $c_i = 0$) will tend to 0 or 1 simultaneously. With the convergence of the model during training, it can be inferred that the SCM generated by the predictor can progressively distinguish the pixels in a given image into two classes, and the representation ability of the feature extractor to the object of interest is constantly enhanced. Notably, there are two solutions to make $\hat{c}_i$ be 1 in Eq. 5, $\hat{p}_i^1 = 1$, $\hat{p}_i^2 = 0$ or $\hat{p}_i^1 = 0$, $\hat{p}_i^2 = 1$. Therefore, different from semantic learning by given semantic labels, the exact category of a pixel can not be indicated in the SCM unless the SCM is visualized. Nevertheless, the SCM obtained by indirect supervision can demonstrate which pixels are of the same category, which is enough to distinguish the pixels of the object of interest and those of others.

C. Unpaired Image Prototype Contrast Module

Based on the SCM predicted by the predictor, we can split the pixels in an image into two classes, which are the object of interest and the others respectively. Thereafter, these pixels with different classes can be used to generate new changed samples, which can help alleviate the adverse impacts of the imbalance between changed samples and unchanged ones for binary object change detection. In detail, the UIPCM is proposed to explore the benefits. First, as shown in Fig. 2(b), the prototype representations of the two types of pixels in an image are extracted. In detail, according to the SCM, the pixel whose value is larger than 0.5 is regarded as class 1, otherwise as class 0. That is formulated as follows

$$\operatorname{sgn}(SCM(i,j)) = \begin{cases} 1, & SCM(i,j) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\operatorname{sgn}(\cdot)$ is used to identify the class of a pixel. Considering the representation dependency among the pixels with the same category in an image, their central feature is viewed as their prototype, which can be formulated as follows,

$$PE_n^1 = \frac{\sum_{i,j} F_n(i,j) * \operatorname{sgn}(SCM_n(i,j))}{\sum_{i,j} \operatorname{sgn}(SCM_n(i,j))}, \tag{7}$$

$$PE_n^0 = \frac{\sum_{i,j} F_n(i,j) * (1 - \operatorname{sgn}(SCM_n(i,j)))}{\sum_{i,j} (1 - \operatorname{sgn}(SCM_n(i,j)))}, \tag{8}$$

where F_n is the feature of the image n outputted by the projector, and PE_n^1 represents the prototype of the pixels predicted as class 1 in the image n. n refers to the n-th image in the N images (i.e., $N/2$ pairs of bitemporal images) within the current training batch. The two types of prototype representations can effectively reveal the characteristics of the object of interest and the others. Then, we combine the unpaired image prototypes to generate change embedding,

$$\hat{F}_{k,l} = PE_k^1 - PE_l^0,\ \ k,l \in \{1,2,\dots,N\}, \tag{9}$$

where $\hat{F}_{k,l}$ indicates the newly generated change representation. According to Eq. 9, there are N^2 change representations constructed in a training batch. Obviously, due to the diversity of the object of interest and the others, the diversity of change samples can be enriched. Meanwhile, the imbalance between the changed samples and unchanged samples can be reduced. To sum up, the relation between unpaired images is effectively utilized in UIPCM. Finally, the magnitude of the generated change embedding is optimized by contrastive metric,

$$\mathcal{L}_{UIPC} = \frac{1}{N^2}\sum_{k,l}\max(\delta - \|\hat{F}_{k,l}\|_2, 0)^2, \tag{10}$$

where δ is a given threshold.

D. Loss Function

Most metric-based methods adopt batch balance contrastive loss (BCL) to optimize the proposed models. It forces a model to make the bitemporal representation consistent in regions other than the changed object of interest. However, since there are always some irrelevant changes existing, making the bitemporal embedding there simply consistent easily makes the model fall into a local optimum. To alleviate the above problem, the dual margin contrastive loss (DMCL) [46] is employed to optimize the DM generated by the proposed framework,

$$\begin{aligned}\mathcal{L}_{DMCL} = \frac{1}{2}[\frac{1}{|\mathcal{B}_0|}\sum_{i\in\mathcal{B}_0}(1-c_i)\max(d_i - m_1, 0)^2 + \\ \frac{1}{|\mathcal{B}_1|}\sum_{i\in\mathcal{B}_1}c_i\max(m_2 - d_i, 0)^2],\end{aligned} \tag{11}$$

where d_i indicates the distance between the bitemporal features corresponding to the pixel i. $\mathcal{B}_0$ and $\mathcal{B}_1$ represent the set of the unchanged pixels and that of the changed pixels in the current training batch respectively. m_1 and m_2 are the margin values. The DMCL allows the difference in bitemporal representation in unchanged regions. It degenerates into the BCL when m_1 is equal to 0. Besides, the Dice loss [47] is employed to optimize the output $\hat{C}$ of the ISSM,

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_i \hat{c}_i c_i + \alpha}{\sum_i \hat{c}_i + \sum_i c_i + \alpha}, \tag{12}$$

where α represents the smooth factor. The values in the output of the predictor will indirectly converge to 0 or 1 by the supervision to $\hat{C}$, thus the pixels in an image can be distinguished into two classes. The overall loss used to train the proposed framework is formulated as,

$$\mathcal{L} = \mathcal{L}_{DMCL} + \lambda_1\mathcal{L}_{Dice} + \lambda_2\mathcal{L}_{UIPC}, \tag{13}$$

where the λ_1 and the λ_2 are the balanced coefficients. In the overall loss, the UIPC loss aims to make the magnitude of generated change embedding larger than a given threshold. This motivation is the same as that in the DMCL. In addition, it would increase the Euclidean distance between the two types of prototypes from different images, which can help enhance the discrimination between the features of the interested object and the others. Therefore, it can be inferred that UIPCM can boost the performance of the framework by it.

In summary, we present the implementation process of the proposed method in Algorithm. 1, where the Θ represents the parameters of the framework. D indicates the set of the bitemporal images in a given change detection dataset. Especially, the UIPCM is not introduced until the current epoch is larger than the set warmup epoch.

Algorithm 1: The Implementation of Our Method

```
Initialize $epoch \leftarrow 1$;
Initialize $t \leftarrow 0$;
Initialize $\Theta_0$ via the pretrianed weights on ImageNet;
while $epoch \le max\_epoch$ do
    for $I_{T1}, I_{T2} \in D$ do
        $t \leftarrow t + 1$;
        for $i \in \{1,2\}$ do
            $F_i^1, F_i^2, F_i^3, F_i^4 \leftarrow \mathrm{FE}(I_{Ti})$;
        end
        for $i \in \{1,2\}$ do
            $F_{Ti} \leftarrow \mathrm{Projector}(F_i^1, F_i^2, F_i^3, F_i^4)$;
            $SCM_{Ti} \leftarrow \mathrm{Predictor}(F_i^1, F_i^2, F_i^3, F_i^4)$;
        end
        $\mathrm{DM} \leftarrow (\|F_{T1} - F_{T2}\|_2)_{\uparrow 4}$;
        $\mathcal{L}_{DMCL} \leftarrow \mathrm{DMCL}(DM, GT)$;
        $\mathcal{L}_{Dice} \leftarrow \mathrm{Dice}(|SCM_{T1} - SCM_{T2}|_1, GT)$;
        $\mathcal{L} \leftarrow \mathcal{L}_{DMCL} + \lambda_1\mathcal{L}_{Dice}$;
        if $epoch > warmup\_epoch$ then
            $\mathcal{L}_{UIPC} \leftarrow \mathrm{UIPCM}(F_T, SCM_T)$;
            $\mathcal{L} \leftarrow \mathcal{L} + \lambda_2\mathcal{L}_{UIPC}$;
        end
        $\Theta_t \leftarrow Backward(\mathcal{L}, \Theta_{t-1})$;
    end
    $epoch \leftarrow epoch + 1$;
end
```

IV. Experiments

A. Dataset and Evaluation Metric

To demonstrate the effectiveness and superiority of the proposed method, we execute experiments on two well-recognized building change detection datasets LEVIR-CD [11] and WHU-CD [48], where the building is the object of interest. The LEVIR-CD consists of 637 pairs of images with the size of 1024×1024, which are collected from the Google Earth platform. The WHU-CD includes two pairs of high-resolution large-scale images for training and testing, which own rich details. The images in the two datasets are cropped into the patches with the size of 256×256 for training, testing, and

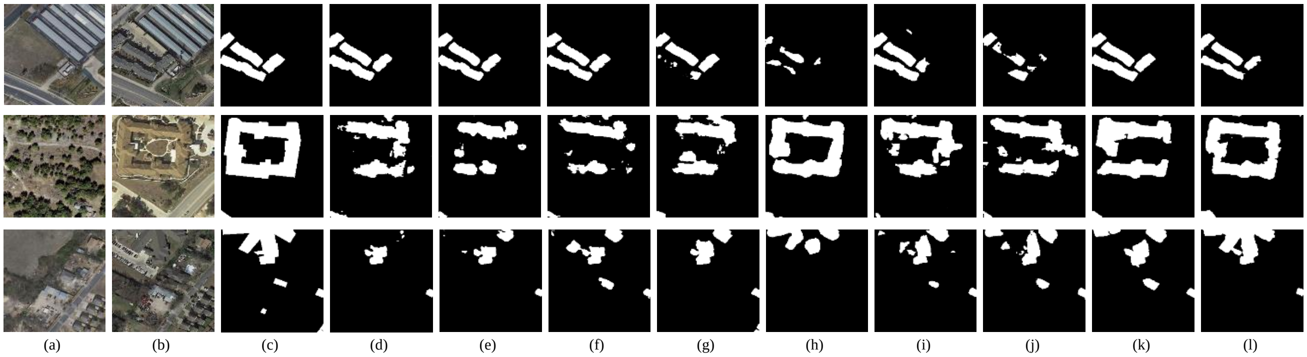

Fig. 3. The visualized comparison on the LEVIR-CD dataset. (a) Pre-temporal image. (b) Post-temporal image. (c) Ground Truth. (d) FC-EF. (e) FC-Siam-diff. (f) FC-Siam-conc. (g) SNUNet. (h) DSAMNet. (i) STANet. (j) ChangeFormer. (k) EGRCNN. (l) Ours.

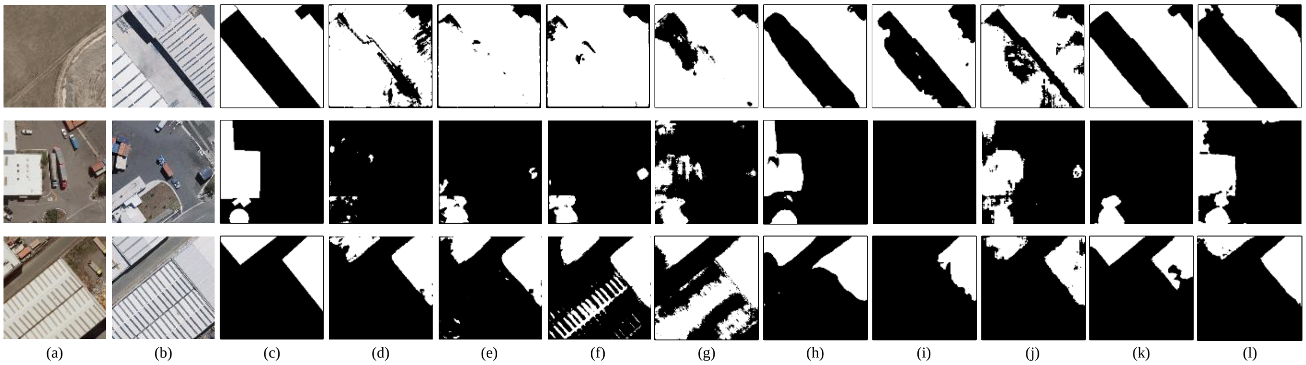

Fig. 4. The visualized comparison on the WHU-CD dataset. (a) Pre-temporal image. (b) Post-temporal image. (c) Ground Truth. (d) FC-EF. (e) FC-Siam-diff. (f) FC-Siam-conc. (g) SNUNet. (h) DSAMNet. (i) STANet. (j) ChangeFormer. (k) EGRCNN. (l) Ours.

validation. The patches for training are obtained by random cropping, while those for validation and testing are by cropping without overlapping. As with most previous methods, the following five metrics are used to objectively evaluate the performance of different methods: precision (P), recall (R), intersection over union (IoU), F1-score (F1), and overall accuracy (OA).

B. Implementation Details and Parameter Settings

We use the AdamW [49] optimizer and set the initial learning rate to 0.0001. We train 200 epochs on the two datasets with the batch size set to 8. The learning rate is halved every 40 epochs. The threshold T in Eq. 2 is set to 2. The m_1 and m_2 in the DMCL are 1 and 2 respectively. The δ in Eq. 10 is set to 2. The λ_1 and λ_2 in the overall loss are all 0.1. In particular, considering that the UIPCM is dependent on the output of the predictor, it is not introduced until the framework has been trained for 10 epochs during training.

C. Comparison with the Advanced Methods

Eight recent advanced methods are selected for comparison. FE-EF [50], FC-Siam-diff [50], and FC-Siam-conc [50] are three classical change detection architectures based on the full convolutional network. The difference between them is the extracted strategy of the change information. FC-EF directly extracts the change features from the concatenated bitemporal images. FC-Siam-diff and FC-Siam-conc acquire the change representation at the feature level by the subtraction and concatenation of the bitemporal features respectively. SNUNet [51] is designed based on a Siamese network and the nested UNet, which effectively maintains the texture and structure information from the shallow features at the high-level features. DSAMNet [52] and STANet [11] are the two metric-based models. The former introduces spatial attention and channel attention for performance improvement. The latter proposes a multi-scale spatial-temporal attention mechanism to alleviate the negative effect of the registration error. ChangeFormer [10] is a representative architecture based on the ViT. EGRCNN [27] is a building change detection method, which attempts to introduce the edge detection of the building of change to improve its accuracy. These methods are implemented following their original details.

Figs. 3 and 4 exhibit the qualitative comparison results on the LEVIR-CD and the WHU-CD datasets, where our method generates relatively complete and precise results for the changed buildings with different types, shapes, and scales shown. In contrast, the predicted results of the compared methods are either relatively coarse or large amounts of missed pixels. The quantitative results are reported in Tab. I. Compared with the given methods, the proposed framework achieves comparable performance on the LEVIR-CD dataset and the best accuracy on the WHU-CD dataset. Notably, the framework does not introduce sophisticated modules in the network model as in compared methods. It concisely and efficiently utilizes the multi-scale features extracted via the ResNet-18 model to obtain bitemporal representation for

TABLE I
EXPERIMENTAL RESULTS ON LEVIR-CD AND WHU-CD DATASETS (%). THE BOLD INDICATES THE BEST PERFORMANCE.

Method	LEVIR-CD					WHU-CD				
	P	R	F1	IoU	OA	P	R	F1	IoU	OA
FC-EF [50]	86.09	83.05	84.77	73.57	98.47	82.84	78.02	80.36	67.16	98.62
FC-Siam-diff [50]	90.05	83.48	86.64	76.43	98.69	73.27	82.86	77.77	63.62	98.29
FC-Siam-conc [50]	90.46	83.84	87.02	77.03	98.73	65.41	86.32	74.42	59.27	97.85
SNUNet [51]	90.69	88.93	89.80	81.49	98.97	83.95	88.95	86.38	76.02	98.99
DSAMNet [52]	81.28	88.68	84.81	73.63	98.38	71.22	**92.28**	80.40	67.22	98.37
STANet [11]	85.01	**91.38**	88.07	78.69	98.74	88.59	85.18	86.86	76.76	99.07
ChangeFormer [10]	91.03	87.77	89.37	80.78	98.94	88.22	79.86	83.83	72.16	98.89
EGRCNN [27]	88.58	91.13	89.84	81.55	98.95	90.92	89.41	90.16	82.08	99.29
Ours	**91.89**	88.10	**89.95**	**81.74**	**99.00**	**94.22**	89.04	**91.56**	**84.43**	**99.41**

TABLE II
ABLATION STUDIES ON LEVIR-CD AND WHU-CD DATASETS (%). THE BOLD INDICATES THE BEST PERFORMANCE.

Method	LEVIR-CD					WHU-CD				
	P	R	F1	IoU	OA	P	R	F1	IoU	OA
Baseline	92.07	85.31	88.56	79.47	98.88	94.34	81.19	87.27	77.42	99.14
Baseline+DMCL	**92.18**	86.25	89.11	80.37	98.93	94.00	86.23	89.95	81.73	99.30
Baseline+DMCL+ISSM	90.92	88.09	89.48	80.97	98.95	**95.31**	86.28	90.57	82.76	99.35
Baseline+DMCL+ISSM+UIPCM	91.89	**88.10**	**89.95**	**81.74**	**99.00**	94.22	**89.04**	**91.56**	**84.43**	**99.41**

change detection. As for the proposed modules, they are introduced only during training. Nevertheless, from the visualization comparison and the quantitative evaluation, the proposed framework is still competitive.

D. Ablation Studies

In this section, the effectiveness of the ISSM and the UIPCM is explored, and the help of the DMCL is illustrated. The quantitative results of ablation studies are reported in Tab. II. The first row refers to the baseline model trained by the BCL. When the DMCL is used to train the model, its performance is enhanced significantly, especially in the WHU-CD dataset. Due to the variation of the season and the effect of human activities, there are many irrelevant changes other than the building, e.g., the moving of vehicles, or the new construction of a road. Meanwhile, the DMCL allows the differences between the bitemporal features in the above-mentioned regions, which can maintain the diversity of the extracted features with different land cover types to alleviate the over-fitting of the model. Therefore, the introduction of the DMCL is helpful. Thereafter, the incorporation of the ISSM further improves the accuracy of the model. Fig. 5 shows some SCMs, which reveal the positions of the buildings in the given images relatively accurately. The shown buildings in Fig. 5 are complex and diverse, which credibly illustrates the efficacy of the ISSM to help the predictor distinguish semantics. Finally, the UIPCM is injected, which is dependent on the SCM from the predictor and the dense embedding from the projector. According to Tab. II, the accuracy enhancement of the model on the two datasets illustrates its effectiveness. Especially, the recall increases by 2.76% on the WHU-CD dataset, which indicates more missed pixels or instances are detected. This indirectly confirms the benefit of the UIPCM in enriching the diversity of changed samples for training. To sum up, the three improvements are indeed useful.

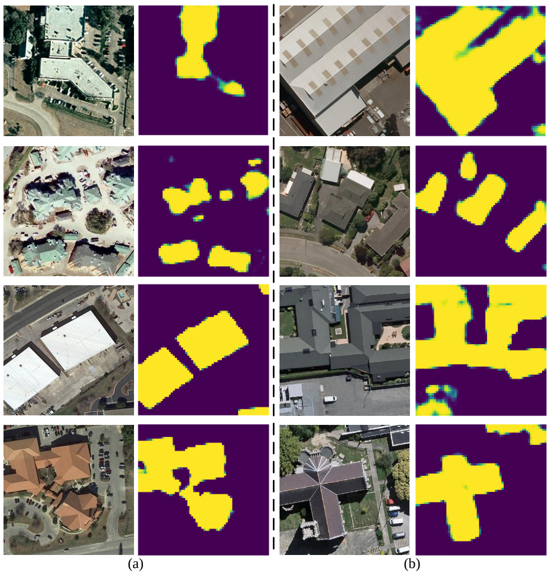

Fig. 5. The visualization of SCMs in some representative scenes on the LEVIR-CD dataset (a) and the WHU-CD dataset (b).

TABLE III
THE EXPLORATION ABOUT THE COEFFICIENT λ_1 (%). ($\lambda_2 = 0.1$)

λ_1	P	R	F1	IoU	OA
0.1	94.22	89.04	91.56	84.43	99.41
0.3	94.61	87.74	91.04	83.56	99.38
0.5	93.45	89.95	91.67	84.62	99.41
0.7	94.33	88.37	91.25	83.91	99.39

E. Hyperparameter Analysis

The hyperparameters λ_1, λ_2, and m_1 in the proposed method are investigated in this section. First, we analyze the effect of the λ_1 and λ_2 on the accuracy of the model on the WHU-CD dataset. Second, the role of the m_1 is explored on the two datasets.

Balanced coffecients λ_1 and λ_2: The λ_1 and λ_2 are employed to tune the proportion of the Dice loss and the

UIPC loss in the overall loss. Tab. III reports the effect of λ_1 on the performance of the framework, where the λ_2 is equal to 0.1. When λ_1 is 0.1 and 0.5, the model performance is close and better than that when λ_1 is 0.3 and 0.7. To ensure higher precision, the value of λ_1 is set to 0.1. In addition, Tab. IV shows the impact of different λ_2 on the performance of the framework when setting λ_1 to 0.1. When λ_2 is 0.1, the overall performance of the model is the best, and competitive precision and recall are achieved. Therefore, λ_1 and λ_2 are both set to 0.1. Besides, it can be found from Tab. IV that the introduction of the UIPCM improves the performance of the model, which reaffirms its effectiveness.

TABLE IV
THE EXPLORATION ABOUT THE COEFFICIENT λ_2 (%). ($\lambda_1 = 0.1$)

λ_2	P	R	F1	IoU	OA
–	95.31	86.28	90.57	82.76	99.35
0.1	94.22	89.04	91.56	84.43	99.41
0.3	93.11	89.10	91.06	83.59	99.37
0.5	94.27	88.50	91.29	83.98	99.39
0.7	93.13	89.77	91.42	84.20	99.39

The margin m_1 in DMCL: This work adopts the DMCL to train the proposed framework. m_2 is set to 2, which is the same as the BCL utilized by previous works [44], [45]. We focus on exploring the role of the margin m_1. Tab. V shows the experimental results on the two datasets, where UIPC indicates the introduction of the ISSM and UIPCM to the baseline model. As reported in Tab. V, the performance of the model is effectively improved in both cases when m_1 is not 0. $m_1 \neq 0$ means that the bitemporal features in changed areas without including the object of interest are no longer forced to be completely consistent during the training process, which alleviates the overfitting of the model and increases the diversity of bitemporal features. If the m_1 is close to 0, it increases the risk of model overfitting. In addition, it can be noticed that if m_1 is close to m_2, the performance degrades. This can be attributed to the fact that the discrimination between the changed and the unchanged representation is weakened. To alleviate the disadvantages caused by m_1 being too large or too small, the m_1 is set to half of m_2 based on the experimental results. Meanwhile, the results in Tab. V further demonstrate that the introduction of the ISMM and UIPCM is helpful, without being affected by the variation of m_1.

TABLE V
THE EXPLORATION ABOUT THE MARGIN m_1 ON THE LEVIR-CD AND WHU-CD DATASETS (%).

m_1	UIPC	LEVIR-CD		WHU-CD	
		F1	IoU	F1	IoU
0.	✗	88.56	79.47	87.27	77.42
	✓	89.27	80.62	89.32	80.69
0.5	✗	89.11	80.36	89.17	80.46
	✓	89.44	80.90	90.14	82.05
1.0	✗	89.11	80.37	89.95	81.73
	✓	89.95	81.74	91.56	84.43
1.5	✗	89.26	80.61	89.54	81.07
	✓	89.58	81.13	90.58	82.77

V. CONCLUSION

In this paper, we explore the utilization of unpaired images for boosting binary object change detection. A concise metric-based framework is proposed. It is made of a projector branch and a predictor branch. They obtain the dense embedding and the semantic confidence map of the input image respectively. Among them, the predictor is induced to distinguish the pixels in an image into two classes (the object of interest and the others) under the effect of an indirect semantic supervision module. Founded on the framework, we propose an unpaired image prototype contrast module. First, depending on the semantic confidence map and the dense embedding from the framework, the image prototypes are extracted. Then, the diversity of the change embedding for training is enriched by combining the prototypes in unpaired images to generate change representation, leading to improving the performance of the framework. Besides, we introduce a dual margin contrastive loss. It is helpful to alleviate the over-fitting of the framework, thus improving the accuracy of the obtained change map. The effectiveness of the proposed method is demonstrated by the extensive experiments executed on two well-recognized building change detection datasets. Meanwhile, compared with the advanced methods, the proposed method achieves competitive performance.

REFERENCES

[1] L. Gong, Q. Li, and J. Zhang, "Earthquake building damage detection with object-oriented change detection," in *Proc. IEEE International Geoscience and Remote Sensing Symposium*, 2013, pp. 3674–3677.

[2] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sensing of Environment*, vol. 265, p. 112636, 2021.

[3] P. Barmpoutis, T. Stathaki, and V. Kamperidou, "Monitoring of trees' health condition using a UAV equipped with low-cost digital camera," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 8291–8295.

[4] F. Chaabane, S. Réjichi, C. Kefi, H. Ismail, and F. Tupin, "Anarchic urban expansion detection and monitoring with integration of expert knowledge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 8306–8309.

[5] G. Cheng, G. Wang, and J. Han, "ISNet: Towards improving separability for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[6] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[7] G. Wang, G. Cheng, P. Zhou, and J. Han, "Cross-level attentive feature aggregation for change detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[8] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[9] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, "An attention-based multiscale transformer network for remote sensing image change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 599–609, 2023.

[10] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 207–210.

[11] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[12] L. Shen, Y. Lu, H. Chen, H. Wei, D. Xie, J. Yue, R. Chen, S. Lv, and B. Jiang, "S2Looking: A satellite side-looking dataset for building change detection," *Remote Sensing*, vol. 13, no. 24, p. 5094, 2021.

[13] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Computer Vision and Image Understanding*, vol. 187, p. 102783, 2019.

[14] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, M. Pelillo, and L. Zhang, "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[15] S. Tian, Y. Zhong, Z. Zheng, A. Ma, X. Tan, and 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*, vol. 193, pp. 164–186, 2022.

[16] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 228–239, 2022.

[17] S. Tian, X. Tan, A. Ma, Z. Zheng, L. Zhang, and Y. Zhong, "Temporal-agnostic change region proposal for semantic change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 204, pp. 306–320, 2023.

[18] Y. Niu, H. Guo, J. Lu, L. Ding, and D. Yu, "SMNet: Symmetric multi-task network for semantic change detection in remote sensing images based on cnn and transformer," *Remote Sensing*, vol. 15, no. 4, p. 949, 2023.

[19] T. Liu, M. Gong, D. Lu, Q. Zhang, H. Zheng, F. Jiang, and M. Zhang, "Building change detection for vhr remote sensing images via local–global pyramid network and cross-task transfer learning strategy," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[20] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery," in *Proc. IEEE International Conference on Computer Vision*, 2021, pp. 15 193–15 202.

[21] M. Seo, H. Lee, Y. Jeon, and J. Seo, "Self-pair: Synthesizing changes from single source for object change detection in remote sensing imagery," in *Proc. IEEE Winter Conference on Applications of Computer Vision*, 2023, pp. 6374–6383.

[22] H. Zhang, G. Ma, and Y. Zhang, "Intelligent-BCD: A novel knowledge-transfer building change detection framework for high-resolution remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 5065–5075, 2022.

[23] A. López-Cifuentes, M. Escudero-Viñolo, J. Bescós, and Álvaro García-Martín, "Semantic-aware scene recognition," *Pattern Recognition*, vol. 102, p. 107256, 2020.

[24] W. Liu, X. Ma, Y. Zhou, D. Tao, and J. Cheng, "p-laplacian regularization for scene recognition," *IEEE Transactions on Cybernetics*, vol. 49, no. 8, pp. 2927–2940, 2019.

[25] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2021.

[26] Z. Chen, Y. Zhou, B. Wang, X. Xu, N. He, S. Jin, and S. Jin, "EGDE-Net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 203–222, 2022.

[27] B. Bai, W. Fu, T. Lu, and S. Li, "Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[28] L. Ding, J. Zhang, K. Zhang, H. Guo, B. Liu, and L. Bruzzone, "Joint spatio-temporal modeling for semantic change detection in remote sensing images," *arXiv preprint arXiv:2212.05245*, 2022.

[29] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-temporal semantic reasoning for the semantic change detection in hr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[30] S. Hou, G. Zhang, H. Cui, X. Li, Y. Chen, H. Li, H. Wang, and X. Ma, "Stable prototype-guided single-temporal supervised learning for change detection and extraction of building," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–22, 2023.

[31] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[32] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[33] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[34] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Robust classification with convolutional prototype learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3474–3482.

[35] T. Zhou, W. Wang, E. Konukoglu, and L. Van Gool, "Rethinking semantic segmentation: A prototype view," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2582–2593.

[36] T. Zhou, M. Zhang, F. Zhao, and J. Li, "Regional semantic contrast and aggregation for weakly supervised semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4299–4309.

[37] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[38] S. Chen, W. Hou, Z. Hong, X. Ding, Y. Song, X. You, T. Liu, and K. Zhang, "Evolving semantic prototype improves generative zero-shot learning," in *Proc. International Conference on Machine Learning*, vol. 202, 2023, pp. 4611–4622.

[39] C. Wang, S. Min, X. Chen, X. Sun, and H. Li, "Dual progressive prototype network for generalized zero-shot learning," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 2936–2948.

[40] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, "Zero-shot learning on semantic class prototype graph," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 2009–2022, 2018.

[41] S. Liu, Y. Zhang, Z. Qiu, H. Xie, Y. Zhang, and T. Yao, "Learning orthogonal prototypes for generalized few-shot semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 319–11 328.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[44] M. Zhang, Q. Li, Y. Miao, Y. Yuan, and Q. Wang, "Difference-guided aggregation network with multiimage pixel contrast for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[45] M. Zhang, Q. Li, Y. Yuan, and Q. Wang, "Edge neighborhood contrastive learning for building change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[46] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2021.

[47] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017, pp. 240–248.

[48] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on geoscience and remote sensing*, vol. 57, no. 1, pp. 574–586, 2019.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations*, 2015.

[50] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE International Conference on Image Processing*, 2018, pp. 4063–4067.

[51] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[52] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.