OCA: Object-Based Change Augmentation for Few-Shot Building Change Detection in Very High-Resolution Remote Sensing Images

Chenxiao Zhang, Peng Yue, *Senior Member, IEEE*, Francesca Cigna, and Deodato Tapete

***Abstract*— Change detection (CD) based on multitemporal remote sensing imagery is a crucial step for various Earth observation applications. While deep learning (DL) has revolutionized CD, its data-driven nature demands substantial labeled images for supervised model training, which is costly and time-consuming. This article addresses the challenge of limited training samples by proposing a novel object-based change augmentation (OCA) method. Unlike conventional image-level augmentation methods that can introduce irrelevant contextual dependencies, OCA decomposes the augmentation process into few-shot object classification and foreground–background pasting, thereby generating in-distribution synthetic images with increased change diversity. An object-based training strategy is developed to create a high-confidence binary classifier for pseudosemantic segmentation, facilitating the copy–paste operation. Experimental results on the very-high-resolution remote sensing images demonstrate the superior performance of OCA compared to existing augmentation- and generation-based methods. A comprehensive analysis of parameter sensitivity, adaptability to varying training data volumes, and compatibility with diverse CD methods validates its robustness. This approach provides a practical and effective solution for few-shot CD scenarios, advancing the applicability of DL-based CD methods in training data-limited environments. Codes and data are available: https://github.com/openrsgis/OCA**

***Index Terms*— Change detection (CD), few-shot learning, remote sensing, semisupervised learning.**

I. Introduction

CHANGE detection (CD) refers to identifying the absence and/or presence of land surface modifications from multitemporal remote sensing images. CD has been widely used in various applications including postdisaster damage assessment [1], urban expansion [2], and war monitoring [3]. In the recent decade, CD has significantly shifted from handcrafted feature learning-based [4] to deep feature learning-based methods [5], attributed to the strong representation ability of deep neural networks.

The data-driven nature of deep learning (DL) methods heavily relies on a large amount of high-quality training samples. Labeling the multiscale building changes by visually comparing the bitemporal images in a top-down view requires substantial human effort and time. Factors such as scene illumination intensity, seasonal change, and shooting angle must be carefully taken into consideration to avoid mislabeling. To address the above challenges, two mainstream approaches have been investigated: augmentation-based methods and generation-based methods. Data augmentation significantly improves model generalizability to unseen data by enforcing the model to encounter more diverse features and avoids overfitting on limited training data [6]. Conventional data augmentation adopts geometrical transformations, such as image flipping, cropping, and rotation, and some nongeometrical transformations, such as noise injection and color conversion to create variations of existing data. However, image-level transformations do not decouple the foreground-changed instances with the background nonchanged land surface. Consequently, the model may rely on unnecessary context information while focusing less on instance features, which hampers learning the intrinsic characteristics of changed instances. For example, garden surroundings may be mistakenly considered an important factor for changed buildings discrimination, which is not applicable in factory regions where no garden appears. The other approach, i.e., generation-based methods, adopts either generative adversarial neural (GAN) or denoising diffusion probabilistic models to generate synthetic images and, while promising, faces significant challenges in requiring extensive pretraining on large-scale datasets, which may not always be available. Moreover, the training process is notoriously unstable, often resulting in unpredictable images and making it difficult for CD model training. More critically, the inherent domain gap between the generated images and the current task remains a major obstacle, hindering the transfer of generated change patterns to real-world change scenarios.

This article focuses on the data augmentation research by proposing a novel object-based change augmentation (OCA) method. OCA decomposes the change image augmentation task into two manageable subtasks, i.e., the few-shot

Received 24 February 2025; revised 7 May 2025; accepted 19 June 2025. Date of publication 30 June 2025; date of current version 3 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42201396 and Grant 42425108, in part by the Natural Science Foundation of Hubei Province under Grant ZRMS2022000666, and in part by China National Postdoctoral Program for Innovative Talents under Grant BX2021223. *(Corresponding author: Peng Yue.)*

Chenxiao Zhang and Peng Yue are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: zhangchx@whu.edu.cn; pyue@whu.edu.cn).

Francesca Cigna is with the Institute of Atmospheric Sciences and Climate (ISAC), National Research Council (CNR), 00133 Rome, Italy (e-mail: f.cigna@isac.cnr.it).

Deodato Tapete is with Italian Space Agency (ASI), 00133 Rome, Italy (e-mail: deodato.tapet@asi.it).

Digital Object Identifier 10.1109/TGRS.2025.3584092

object classification task and the foreground–background pasting task. It significantly improves data volume and change diversity by performing instance-level augmentation while maintaining the same data distribution as the current CD task. The major contributions of this work are summarized as follows.

1) A novel and efficient OCA method is introduced to address the challenge of limited training data in few-shot CD scenarios, specifically for very-high-resolution remote sensing imagery. Unlike existing approaches, OCA does not require pretraining on large-scale datasets and generates synthetic images that maintain the same data distribution as the target data. This greatly improves its adaptability over various datasets and CD models.
2) An object-based training strategy is proposed to build a high-confident binary classifier based on which the pseudosemantic segmentation maps can be acquired. Combined with a simple copy–paste operation, new images with diverse change patterns are acquired.
3) A thorough performance evaluation of OCA across various benchmark datasets, CD models, temporal exchange, and training scenarios demonstrates its effectiveness in very-high-resolution datasets.

The rest of this article is organized as follows. Section II provides a comprehensive review of related work, encompassing both fundamental and advanced image augmentation techniques, as well as image and instance generation methods relevant to CD. Section III details the proposed methodology, including the architecture and implementation of OCA. Section IV presents a thorough experimental evaluation, comparing the performance of OCA against various augmentation and generation-based approaches across different CD models and data configurations, along with detailed discussions. Finally, Section V summarizes the key findings of this work and provides a comprehensive analysis of the limitations of OCA, including future research directions for its adaptability to more complex CD scenarios.

II. Related Work

A. Few-Shot Learning Methods in Remote Sensing

Labeling remote sensing images for deep model training requires substantial labors and times compared to natural image labeling work, especially when labeling changes on bitemporal images where pseudochanges mix with real changes. Few-shot learning has emerged as a promising solution to overcome the challenge of limited labels. It can be broadly categorized into data augmentation-based methods and prior-knowledge-based methods [7]. Data augmentation seeks to expand the current training dataset into a more diverse dataset using image manipulation or generative models. In contrast, prior-knowledge-based methods focus on transferring knowledge from pretrained models or old tasks to new tasks, among which transfer learning [8] and metric learning represent two major strategies.

In the remote sensing community, few-shot learning has more been studied for image classification tasks. Tang et al. [9] propose a metric learning-based few-shot learning approach for hyperspectral image classification by leveraging spatial–spectral information to improve classification accuracy under limited training samples. Zhu et al. [10] develop a meta visual prompt tuning strategy by inserting learnable prompt tokens into a frozen pretrained vision Transformer model input space, enabling efficient few-shot remote sensing image classification. Qiu et al. [11] employ vision-language models such as CLIP [12] as feature extractors for few-shot remote sensing image classification, achieving significant performance improvement through a simple two-step classification process.

While few-shot remote sensing CD remains relatively underexplored compared to image classification or scene classification tasks, some initial work demonstrates the potential of few-shot learning for CD. For example, to alleviate the heavy labeling work on hyperspectral image CD, BiG-FSLF [13] proposes a cross-domain adaptation approach by transferring knowledge from very-high-resolution CD datasets, where labels are more economical to obtain, to hyperspectral image CD tasks, where label acquisition is more expensive. In the context of forest CD, FRSCD_DAFS [14] proposes a forest CD approach based on a data augmentation strategy and a few-shot learning approach. It first generates forest fragment images using a generative adversarial network and then pastes the generated fragments into the original CD dataset to achieve effective data augmentation. It further enhances CD performance by incorporating a metalearning module into the CD model.

B. Remote Sensing CD Methods

Early CD methods in the remote sensing community explored the direct comparison of pixel values at low spatial resolution images [15]. As pixel-based analysis brings the pepper-and-salt phenomenon, object-based methods taking regions as analysis units have been widely explored [16]. However, as those early methods rely on manually crafted features, they suffer from a low ability of semantic understanding, especially when applied to high spatial resolution images.

Recent CD methods have significantly shifted from manually crafted features to DL-based features. Early CD methods employ convolutional neural networks (CNNs) as model backbone and have achieved significant performance improvement compared with traditional methods [17], [18], [19]. For example, [5] first proposes a basic CNN architecture for CD, i.e., Siamese CNNs for CD including FC early fusion, FC Siamese concatenation, and FC Siamese difference. To further alleviate the gradient vanishing problem, [20] proposes a deeply supervised CNN with spatial–channel attention modules. Some later work extends the binary CD model by designing multitask learning architectures, such as [21] and [22], to predict both binary change maps and the corresponding semantic labels of changed pixels.

In recent years, motivated by the long-range relation modeling in the Transformer architecture, the self-attention mechanism has been widely explored for CD. They either adopt a full Transformer-based architecture, such as Siamese

Transformer-based ChangeFormer [23] and Swinsunet [24], or hybrid CNN-Transformer architectures, such as BiT [25] and FTANet [18]. Similarly, Transformer-based semantic CD models have also been explored. For example, [26] designs a hybrid residual-block and Transformer-block architecture as the backbone of the image encoder, which is effective in obtaining both local and global information.

Although Transformer-based CD further pushes CD performance by exploring wider spatial reception, it comes with the problem of being heavy computation resource-intensive. Recently, the state space model [27] offers a new solution to capture long-range dependencies in a more efficient way. The state space model is good at long sequence modeling while maintaining low computational resources, based on which Mamba is proposed to enhance the state space model framework with selective state space processing, which allows Mamba to selectively focus on highly related contexts. For example, [28] proposes a visual Mamba backboned encoder for dual temporal image feature extraction and implements three architectures tailored for binary CD, semantic CD, and building damage assessment. A similar work [29] also explores a vision Mamba backboned framework for binary CD.

More recently, with the development of multimodal models, CD tasks have been more combined with natural language understanding tasks, realizing natural language outputs instead of image outputs. Liu et al. [30] pioneer the remote sensing CD captioning research by replacing the final image decoder with a caption decoder, outputting semantic descriptions of changed regions. Hoxha et al. [31] explore two different approaches, feature-fusion and image-fusion frameworks, with RNNs or SVMs to generate textual descriptions of detected changes. Chg3Cap [32] proposes a three-component architecture with a Siamese CNN image encoder, a self-attention-based change encoder, and a caption decoder. Inspired by the Mamba model, a recent work [33] innovates the Mamba's state space model architecture with spatial difference-aware SSM and temporal-traversing SSM, achieving efficient joint spatial–temporal modeling for remote sensing image change captioning while maintaining linear computational complexity. Compared with the change map derivation methods, change captioning methods focus more on providing user-friendly interpretations of images instead of quantifying changes.

C. Basic Data Augmentation Methods

Basic data augmentation approaches rely on fundamental operators including geometric manipulators, such as random cropping, flipping, rotation, and resizing, and some nongeometric manipulators, such as random noise injection and color space transformation. They have been proven very effective in a variety of tasks including land cover classification [34], scene classification [35], and object detection [36]. They are also highly reproducible, reliable, and fast with only a few lines of code. Similarly, they have also been widely utilized in CD [24], [25], [37], [38]. Notably, [20] imposes image rotation and flipping on the dual temporal images while conducting noise injection, image blurring, and image smoothing only on singular temporal images to enlarge the appearance gaps. However, as we have discussed before, they cannot decouple the co-occurrence of foreground changes with background nonchanges, which makes them suffer in augmenting the diversity of change patterns. The lack of advanced augmentation methods severely hampers generalizability in remote sensing CD.

D. Advanced Data Augmentation Methods

Advanced data augmentation approaches can be roughly categorized into two classes: image mixing and generative augmentation methods. Image mixing augmentation methods adopt a cut-and-paste idea to result in new images, which has gained popularity in image classification and object detection. Mixup [39] proposes to blend two images with a blending factor at each pixel, resulting in blending images retaining features of both images. Cutout masks a random region and fills the masked region with 0 or 255, which improves model robustness on image classification tasks [40]. CutMix [41] alleviates the problem of information loss in Cutout by pasting the mask region with a patch from other images. CutMix has also been introduced into the remote sensing tasks including semantic segmentation [42] and object detection [43]. Instead of randomly cutting and pasting, ResizeMix [44] proposes a saliency map-guided method to identify high-valuable patches on the source image and better positions on the target image, which proves effective in image classification. The above methods are suitable for semantic segmentation as they utilize image-level labels but are not available for pixel-level tasks including instance and semantic segmentation. ClassMix [45] first trains a segmentation model to get pseudolabels of two random images. Then, an object clipped from one image is pasted on the other image to generate new training samples in a semisupervised manner. Context decoupling augmentation (CDA) [46] first proposes to decouple the co-occurrence of foreground objects with backgrounds for instance segmentation. Similarly, it trains a semantic segmentation model to collect object instances with good segmentation. Then, the object instances are randomly pasted on other images to result in instance segmentation samples. Instead of training a segmentation model to collect object instances, ObjectAug [47] utilizes labels from the training set to get instances with fine boundaries. Notably, image inpainting [48] is introduced to restore the masked regions. Experiments on natural and medical image datasets with classical CNN models show an average gain of 3% on mean intersection over union (MIoU) [49]. Generative augmentation methods typically employ GAN [50] networks to generate synthetic images such as real-world objects. For example, based on a large building instance segmentation dataset, IAug [51] trains a GAN-based generator for building instance generation and a segment model for candidate region selection, respectively. The generated building instances and candidate regions are then composed to result in a synthetic dataset. However, affected by the variations in shooting environments and geography impacts, buildings from different datasets, especially those collected at different geolocations, may exhibit large appearance variations. Therefore, the selection of an additional building dataset should

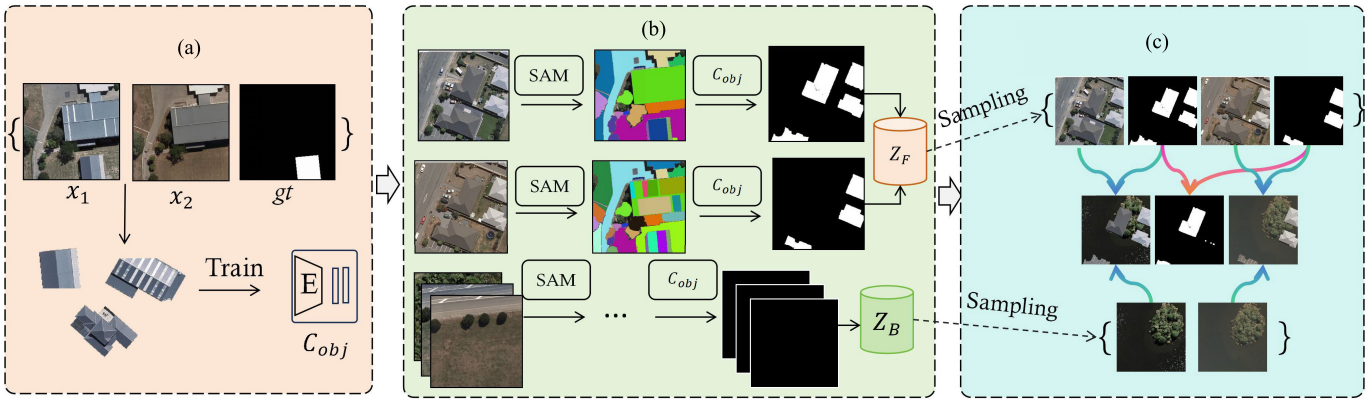

Fig. 1. Workflow of OCA. (a) Few-shot object-based classification. (b) Object-based pseudosegmentation. (c) Foreground–background.

be carefully decided to avoid the potential out-of-distribution problem.

E. Data Generation-Based Methods

Data generation-based methods utilize generative models, such as generative adversarial networks, variational autoencoders, and diffusion models to create synthetic or augmented images for various purposes. The remote sensing society has also explored its availability on CD. Some research adopts a supervised learning schema for cross-modality transformation. For example, Li et al. [52] propose an SAR-to-optical image transformation method based on CycleGAN to generate synthetic SAR or optical images. In such a way, the domain gap between the two modalities can be significantly alleviated, which, finally, improves supervised CD. In fact, the cross-modality transformation inherently aligns the two domains, instead of creating new change patterns. Correspondingly, approaches aiming to simulate new change patterns are proposed. Changen [53] proposes to generate change image pairs using a generic probabilistic model, improving CD performances while maintaining the realism of generated images. To further improve the diversity of change patterns on synthetic images and alleviate the heavy reliance on labeled image pairs, Changen2 adopts a diffusion model to automatically generate new images given any boundaries as a condition, which further pushes the capabilities of zero-shot CD. Interestingly, ChangeDiff proposes a two-stage approach to generate any semantic change pairs given only text prompts. It first generates a map layout based on a text prompt. Then, an optical image is generated given the map layout using ControlNet [54]. Notably, when used for data augmentation, ChangeDiff requires extensive semantic change labels, which is not applicable for few-shot binary CD. Generation-based methods are promising and show excellent performance in producing realistic images comparable to real-world ones. Moreover, they encompass huge potential in simulating diverse change patterns. However, pretrained generative models struggle to control the domain gap between generated images and the target datasets. While some methods utilize target datasets to generate images with the desired style, their unsupervised nature, relying only on weak conditions such as outlines, lacks semantic input. This results in generated objects with ambiguous semantic meanings and unreliable change labels.

III. Methodology

Fig. 1 illustrates the overall workflow of OCA. Initially, given the absence of explicit building labels, change labels are leveraged as prompts to segment potential building and nonbuilding objects from each bitemporal image pair. Subsequently, these segmented objects are used to train a binary object classification model, which, in turn, generates pseudobuilding labels for each image. Based on these labels, image pairs containing building instances are categorized into a "foreground zoo," while image pairs without building instances are put into a "background zoo." Finally, the change image pairs are generated by randomly combining foreground building segments with background images from these two zoos. Sections III-A–III-C provide a detailed explanation of each stage in this process.

A. Few-Shot Object-Based Classification

For ease of discussion, we denote each binary CD triplet as $\{x_1, x_2, y\}$ where x_1 and x_2 represent the prechange and postchange images and y represents the change label corresponding to each image pair. The mask of each change instance is acquired as $M^k \in \{0, 1\}^{HW}$, where k presents the kth objects, 0 indicates nonchanges, and 1 indicates changes. For single image tasks, for example, instance segmentation, one can easily obtain the object instance by multiplying the object mask with the image as follows:

$$I^k = M^k \odot I \tag{1}$$

where I^k is the kth object and I is the targeting image. However, as no building label is given in the building CD (BCD), the targeting image I with building presence needs to be identified. Here, we propose an object-based method to identify and collect building instances. The segment anything model (SAM) is a large-pretrained model for image segmentation. It supports multiscale prompts (i.e., points and boxes) as model inputs to easily segment the objects of interest.

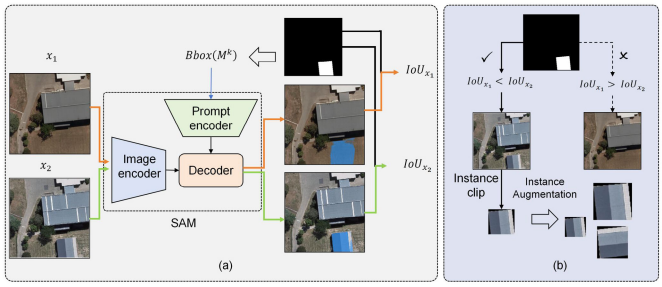

Fig. 2. Segment building and nonbuilding instances. (a) SAM-based building instance segmentation using bounding box prompts from building masks. (b) IoU-based target image selection and building/nonbuilding instance extraction for classification.

To identify potential buildings, the Bbox of each mask is first acquired as follows:

$$\text{Bbox}\left(M^k\right) = \left(\min_x, \min_y, \max_x, \max_y\right). \tag{2}$$

As shown in Fig. 2(a), x_1 and x_2 are fed into the image encoder to get the corresponding image embeddings. The bounding box of each change label is fed into the prompt encoder to get prompt embeddings. The decoder outputs the segmented objects by querying the image embeddings with prompt embeddings. In this way, candidate building instances S_{x_1}, S_{x_2} are acquired. We then compute the Intersection over Union (IoU) between change masks and each of the segmented objects and get the targeting image as follows:

$$I_{x_i} = \text{IoU}(m, \text{SAM}(m, x_i)) \tag{3}$$

$$I = \begin{cases} x_1, & \text{if } I_{x_1} > I_{x_2} \\ x_2, & \text{if } I_{x_1} < I_{x_2}. \end{cases} \tag{4}$$

Now, we have the target image with building presence [e.g., postchange image x_2 in Fig. 2(b)]. We mask the targeting image with an object mask using (1). Finally, we can get the segmented building instances. We then build a binary object classifier C_{obj} and train it with the acquired building and nonbuilding segments.

B. Object-Based Pseudosegmentation

As illustrated in Fig. 1(b), building identification within an image is achieved through an object classification workflow. This workflow first segments the image into a set of nonoverlapping, semantically meaningful patches. Subsequently, an object classifier categorizes each segment as either a building or a nonbuilding object. This process results in a pseudosegmentation map representing the building instances within the image. Given the absence of explicit building segmentation labels, this initial segmentation step is performed in an unsupervised manner, again leveraging SAM to segment images into meaningful land cover objects. It is important to note that SAM can generate a set of potentially overlapping patches for each input point prompt. To address this, we identify any overlapping patches within the generated set and segment them into independent, nonoverlapping patches. The object classifier, denoted as C_{obj}, acquired in the previous step is then applied to categorize each segment. The algorithmic details are described in the following:

$$M(x, y) = C_{\text{obj}}(s_i), \quad (x, y) \in s_i, \ s_i \in S \tag{5}$$

where $M(x, y)$ is the category value assigned to the pixel (x, y) that belongs to the segment s_i in the pseudobuilding segmentation map M. The segment s_i is a part of the set of all segments S obtained from SAM. To reduce false positives, prior knowledge is incorporated into the filtering process. First, to guarantee accuracy, only predictions with high confidence are used. Second, shape-based constraints are introduced. For instance, a maximum aspect ratio is imposed to eliminate elongated, road-like detection. In our experiments, we further refine the results using a maximum aspect ratio, along with minimum and maximum area thresholds. Finally, images from both training and testing sets with identified buildings are put into the foreground zoo Z_F, while those without buildings are put into the background zoo Z_B.

C. Foreground–Background Composition

Pseudosegmentation maps of all images are acquired in the last step. A straightforward CD approach is postclassification subtracting the prechange segmentation map from the postchange one. However, as only high-confidence predictions are retained in the pseudosegmentation map, a significant number of buildings are omitted, thereby leading to a low recall in the postclassification results.

To address this issue, we propose to generate high-confidence change images by composing the foreground change instance from foreground zoo Z_F with the background images from zoo Z_B. As shown in Fig. 1(c), we randomly sample from the foreground zoo Z_F and the background zoo Z_B to get foreground building instances and background surroundings. These sampled components are then combined to produce the final synthetic change maps. The operation is presented as follows:

$$I_i' = I_i \cdot M_i + B_j, \quad I_i \in Z_F, \ B_j \in Z_B \tag{6}$$

where foreground buildings are acquired by masking the original RGB image I_i with the building segmentation mask M_i, effectively isolating the buildings and removing the background. Then, a randomly selected background image with no buildings B_j is added pixelwise to the extracted foreground building, creating the composite synthetic image.

IV. Experiments and Discussion

A. Dataset

We evaluated the effectiveness of OCA across binary CD and semantic CD tasks using two datasets. The dataset used for binary CD is the WHUBCD dataset [55]. The WHUBCD dataset contains two aerial images with a size of 32 507 × 15 354 pixels covering an area of 20.5 km^2 in the city of Christchurch, New Zealand, where a 6.3-magnitude earthquake occurred in February 2011 and destroyed buildings that were rebuilt in the following years. This dataset concentrates on the binary changes of the building category, especially on detecting the reconstructed buildings following the earthquake event. The spatial resolution is 0.075 m/pixel. The dataset encompasses a total of 28 873 manufactural buildings including residential houses and factory buildings. As the two types of buildings vary a lot in appearance, we manually removed

the factory buildings from the dataset and used only residential houses. The bitemporal images were then clipped into 4053 small patches among which 50 patches were selected for model training and the rest 4003 were used for model testing. As the number of training samples significantly impacts the CD performance, we explore its influence in the experiments by progressively reducing the training set by ten images at each step. We report model performances with 10, 20, 30, 40, and 50 samples in Section IV-D. Moreover, to validate the impact of different image sizes, we constructed a new dataset with image sizes of 512 × 512 and performed a comparison of OCA with the synthetic pretraining and instance augmentation methods.

In addition, to further explore the availability of OCA in detecting other land covers, we conducted experiments using a semantic CD dataset: LEVIR-MCI [56]. This dataset has 10 077 bitemporal images, each with a size of 256 × 256 with a spatial resolution of 0.5 m/pixel. The dataset focuses on the change of constructed areas including the change of both buildings and roads. In addition to providing the masks of those changed objects, the dataset also indicates the semantic information of the changed objects. Notably, the semantic information is exclusively provided within the change mask, without category indications in either of the temporal images. This dataset requires CD models to rely more heavily on their CD capabilities rather than simply comparing semantic labels between two temporal images to identify changes. On the one hand, the lack of an explicit semantic prior of bitemporal images increases the challenge of the task and better reflects real-world applications, where the semantic labels of bitemporal images are often unavailable for use. On the other hand, its spatial resolution (0.5 m/pixel) is significantly lower than that of the WHUBCD dataset (0.075 m/pixel). This lower spatial resolution results in a loss of details and a large number of small-sized objects, which poses a challenge for the SAM model.

B. CD Models

The backbone of C_{obj} is the pretrained ResNet50. We modified the channels of the last two fully connected layers to accommodate binary prediction. C_{obj} was trained for 200 epochs with a batch size of 20. Six CD models, including SIAM [5], IFN [20], BIT [25], ELGCNet [17], FTANet [18], and MambaBCD [28], were utilized to evaluate the generalizability of synthetic images in regard to different architectures. Table I illustrates model details including backbones, architectures, network parameter numbers, and flops. SIAM adopts a simple Siamese Unet architecture with a lightweight model design. IFN further improves CD performance based on SIAM by integrating spatial and channel attention into the Siamese architecture with the VGG16 backbone; it has a much larger parameter volume of 50 M and requires more computational resources with 82G Flops. Recently, BIT proposed a hybrid architecture that combines the vision Transformer and the CNN model, achieving high performance at certain CD scenarios. ELGC-Net introduces an efficient CD framework with a local–global context aggregator module that captures enhanced contextual information through pooled-transpose attention and depthwise convolution. For the implementation of ELGCNet, we revised the final output channel from 1 to 3 while maintaining the other model architectures and parameters the same as those provided by the authors. FTANet employs a lightweight MobileNetV2 as an encoder backbone. It takes a Transformer and an INN feature extractor to extract high-frequency local and global information and fuses the two features by a spatial attention module. For FTANet, we utilize the same configuration as the open-sourced codes. MambaCD differs from the existing network architectures by incorporating the VMamba structure as the encoder backbone. For the implementation of MambaCD, considering its heavy computation requirements and our limited computation resources, we take VSSM_tiny as its backbone and set the patch size in VSSM to 10. For model training, we train the models with change maps with the highest resolutions (i.e., change map in the last layer).

TABLE I
CD MODELS

Model	*Architecture*	*Backbone*	*Params*	*Flops*
SIAM (*ICIP, 2018*)	CNN	Res18	1.35 M	4.7G
IFN (*ISPRSJ,2020*)	CNN+Attention	VGG16	50 M	82G
BIT (*TGRS, 2021*)	CNN+Vision Transformer	Res18	11 M	25G
ELGCNet (*TGRS, 2024*)	CNN+ attention	DWConv + transpose attention	10.5M	188G
FTANet (*TGRS, 2025*)	CNN+Vision Transformer	MobileNetV2	4.9M	6.8G
MambaCD (*TGRS, 2024*)	Mamba	VSSM_tiny	408M	18M

C. Benchmark Methods

For fair comparisons, simple and advanced image augmentation methods and image augmentation-based methods are selected as benchmark methods.

1) Basic Image Augmentation Methods: Image-level random flipping and rotation are performed at each training iteration. It creates robust inputs by simple image operations with low computation cost. Given the rotation-invariant property of convolutional operations, such basic image augmentation can significantly boost the performance of CNN-based models.

2) Advanced Image Augmentation Methods: Two copy–paste-based methods were utilized for advanced image augmentation: Cutout and CutMix. For Cutout, we randomly mask a rectangular region in each image pair and fill it with zeros. CutMix builds upon this by generating a similar random mask, where pixel values of 0 and 1 indicate removal and retention, respectively. We then randomly select two image pairs. The first pair is masked with the 1-valued regions to extract a foreground, while the second is masked with the 0-valued regions to obtain a background. These masked images are then combined to create a new, composite training sample.

3) Synthetic Data Pretraining Methods: Image generation methods offer a promising approach for zero- or few-shot CD tasks. Synthetic image pairs are first trained on a specific CD model, and then, the pretrained model is utilized to fine-tune the model on target CD datasets, which is referred to as synthetic data pretraining and can significantly improve CD model performance. In this work, two synthetic datasets generated with different approaches were utilized for benchmark comparison. The Changen2-S1-15K dataset [57] is a state-of-the-art (SOTA) synthetic remote sensing CD dataset generated with diffusion models. It comprises 15 459 synthetic bitemporal image pairs with building change labels. The other dataset is the SyntheWorld [58] that is generated using image rendering techniques. Specifically, the first subdataset with an image size of 512 × 512 containing 10 000 images was utilized for model pretraining. While Changen2-S1-15K is more realistic, its low spatial resolution may limit its application to high-resolution images. In contrast, SyntheWorld has high spatial resolution but appears less realistic due to the image rendering techniques. For experiments, we first pretrain the three benchmark CD models on the two datasets and then fine-tune them on the target dataset. For model fine-tuning, as synthetic datasets have different spatial resolution and image styles from the target dataset, we employ histogram matching and image downsampling to align the source and target datasets; in such a way, we guarantee a fair comparison.

4) Instance Generation Methods: We also compare the performance of OCA with the instance augmentation method, IAug [51]. IAug generates new building instances based on a GAN network that is pretrained on a large building instance dataset, AIRS. It should be noted that IAug requires two well-pretrained models as bases: one is used for building segmentation on the target dataset, and the other is used for generating new building instances. In the experiments, we directly use the pseudobuilding label of target images from OCA as the pasting guidance of IAug. We used the building labels provided by the authors and generated the corresponding building instances by two pretrained building generation models, i.e., the Inria model and the Airs model. All the other parameters align with the code released by the authors.

D. Quantitative Comparison

1) Performance Evaluation With Limited Data: Table II reports the BCD performance of OCA and other augmentation-based methods including random flipping and rotation, Cutout, and Cutmix, given only ten training images in the WHUBCD dataset. Five metrics are used to describe their CD performances, i.e., overall accuracy (OA) and mean IoU (MIoU) on the change and nonchange categories, $F1$ score, precision (P), and recall (R) on the change category. Overall, OCA consistently demonstrates superior performance, particularly combined with the IFN model. When trained on IFN given only ten training samples, OCA achieves the highest scores across all metrics, including an OA of 98.51%, an MIoU of 84.25%, and an $F1$ score of 82.38%. This suggests that OCA effectively leverages instance augmentation to improve model generalization and robustness when training data is extremely scarce. Notably, FTANet achieves competing performance against IFN with much fewer model parameters, validating the robustness of the augmented image generated by OCA on lightweight CD models.

TABLE II
PERFORMANCE COMPARISON WITH AUGMENTATION-BASED METHODS ON THE WHUBCD DATASET. RED INDICATES THE BEST. BLUE INDICATES THE SECOND BEST

Method	*Model*	*OA*	*MIoU*	*F1*	*P*	*R*
Baseline	IFN	97.23	75.70	70.35	61.23	82.67
	BIT	92.89	54.97	29.30	24.20	37.11
	SIAM	53.20	28.95	11.69	6.32	77.95
	ELGCNet	92.4	59.64	42.65	30.44	71.19
	FTANet	60.77	33.84	15.45	8.45	90.27
	MambaCD	94.37	61.52	44.71	36.67	57.26
Random Flipping	IFN	97.33	76.77	72.05	61.66	86.66
	BIT	95.20	52.90	19.20	29.00	14.35
	SIAM	72.12	40.71	18.54	10.49	79.87
	ELGCNet	95.69	67.77	57.12	47.25	72.21
	FTANet	76.09	44.1	22.99	13.18	89.86
	MambaCD	93.79	61.51	45.44	34.9	65.1
Random Rotation	IFN	88.23	55.86	38.63	24.36	93.21
	BIT	94.00	57.14	33.84	30.12	38.61
	SIAM	63.72	32.16	1.37	0.77	6.33
	ELGCNet	93.39	61	44.71	33.47	67.32
	FTANet	86.11	52.22	31.66	19.67	81.02
	MambaCD	93.78	58.39	37.56	31.24	47.09
Cutout	IFN	97.47	77.52	73.15	63.23	86.76
	BIT	96.03	48.01	0.00	0.00	0.00
	SIAM	86.33	51.40	28.87	18.19	69.85
	ELGCNet	96.00	67.96	57.18	49.75	67.23
	FTANet	86.15	53.34	34.78	21.39	92.97
	MambaCD	93.61	61.25	45.05	34.21	65.93
Cutmix	IFN	96.11	71.21	63.42	50.65	84.80
	BIT	96.03	48.01	0.00	0.00	0.00
	SIAM	71.06	40.4	19.49	10.95	88.17
	ELGCNet	94.9	64.9	51.91	41.49	69.32
	FTANet	75.21	43.46	22.43	12.81	90.24
	MambaCD	93.98	61.38	44.87	35.27	61.63
OCA	IFN	**98.51**	**84.25**	**82.38**	**77.86**	87.47
	BIT	97.24	76.42	71.55	60.56	87.41
	SIAM	97.98	79.95	76.54	71.01	83.00
	ELGCNet	96.5	73.13	66.56	53.66	**87.63**
	FTANet	**98.29**	**82.67**	**80.32**	**73.95**	**87.88**
	MambaCD	96.16	70.67	62.36	51.09	80.01

While Cutout and Cutmix also show certain improvements over the baseline, their performance is unstable when trained with different CD models. For instance, Cutout exhibits slight performance increases on IFN, ELGCNet, FTANet, MambaCD, and SIAM models while a significant performance drop on the BIT model, indicating method sensitivity to specific

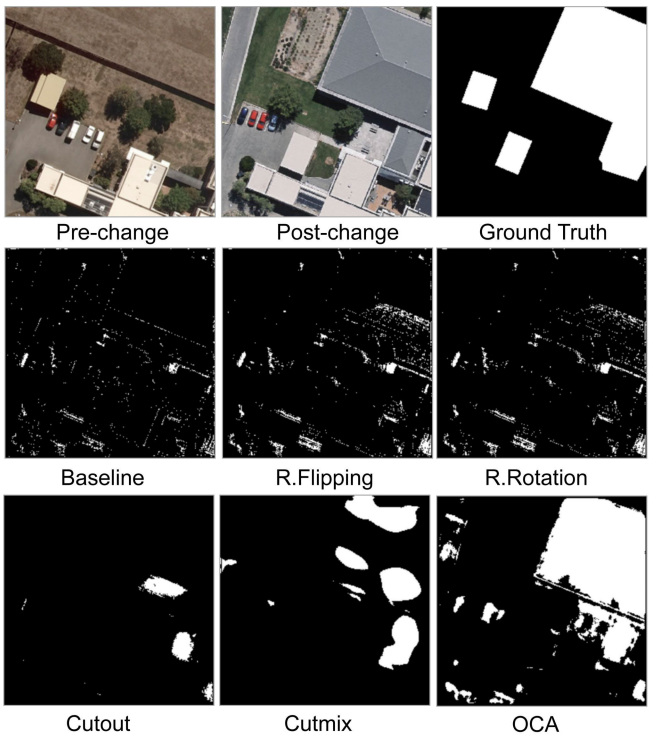

Fig. 3. Example results on dense BCD of OCA and benchmark methods.

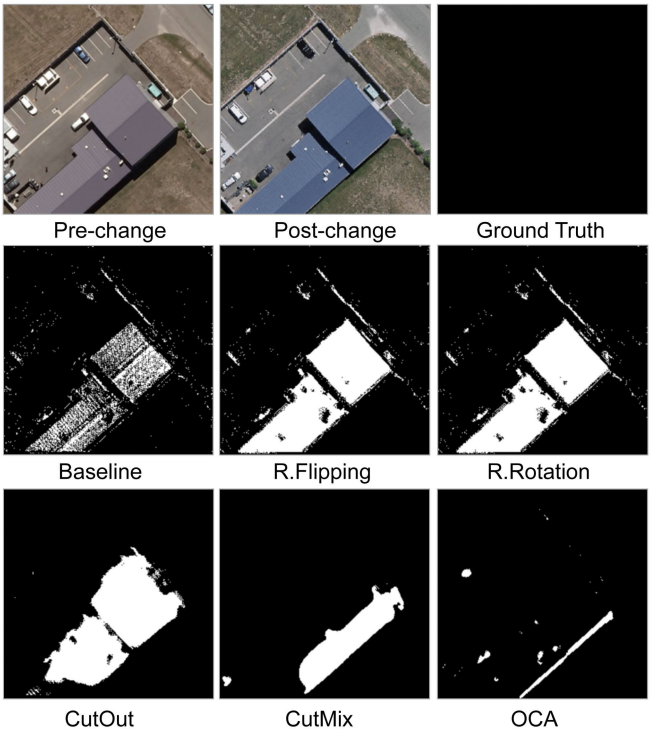

Fig. 4. Example results on pseudo-BCD of OCA and benchmark methods.

CD models. Similarly, Cutmix achieves higher performance than the baseline on the SIAM model, while it also fails to learn CD from the augmented images with an $F1$ score of 0 when trained on BIT, demonstrating its sensitivity to model selection. Interestingly, BIT fails to learn useful information from the augmented images of Cutout and Cutmix, which may be caused by its Transformer module because it considers the global context, while the Cutmix strategy destroys the global context, resulting in misleading results. For other CD models, Cutmix generally performs worse than Cutout. For example, the $F1$ of IFN on Cutmix is 10% lower than that on Cutout. This performance degradation is potentially due to buildings being mixed with other types of land cover in the images, which might confuse CD models.

The simple augmentation method, random flipping, shows comparable performance to the baseline on IFN and slightly improves on SIAM, highlighting its effectiveness on CD tasks. However, random rotation shows a completely different performance; it achieves better performance on BIT but worse performance on IFN and SIAM.

To summarize, OCA consistently surpasses all other methods, showing stable performance on all CD models. Specifically, compared with the baseline, it boosts the $F1$ of 12.03%, 42.25%, 64.85%, 24%, 64.87%, and 17.65 when combined with IFN, BIT, SIAM, ELGCNet, FTANet, and MambaCD, respectively. The results demonstrate the strong ability of OCA in few-shot CD.

Fig. 3 presents the results of OCA and benchmark methods on dense BCD. As we can see, OCA achieves the best result, with clear improvements in capturing large building instance changes. Notice how OCA better preserves the sharp edges and integrity of the large building change compared to Cutout and Cutmix and is less prone to producing scattered false positives compared to random rotation and flipping. The extensive missed detections in positive change regions demonstrate the limitations of benchmark methods in accurately simulating new buildings with diverse appearances that differ from those present in the training set. Fig. 4 presents the results on images with pseudochanges. OCA accurately identifies the pseudobuilding change introduced by color changes in the postchange image, while random rotation, flipping, Cutmix, and Cutout struggle to identify the pseudochanges with a large number of false positives as they primarily focus on simulating change pairs within the training set. This limited exploration of potential changes in the test set hinders their ability to generalize to unseen or pseudochanges. The comparative visualizations demonstrate that OCA is more effective in identifying large building changes with high object integrity and high ability in overcoming pseudochanges than the other data augmentation methods.

Table III reports the semantic CD results of OCA and other augmentation-based methods, given only eight training images from the LEVIR-MCI dataset. In this experiment, the simulation of multicategory object changes follows the same procedure as BCD. Specifically, changed buildings and roads are first segmented from the selected images. Based on the extracted objects, we train a building classifier and a road classifier, respectively. Then, we use the two classifiers to segment the potential building and road regions in the LEVIR-MCI test set, respectively. Finally, building and road segments are stacked together to get the final simulated change maps with changed buildings and roads labeled in different

TABLE III
PERFORMANCE COMPARISON WITH AUGMENTATION-BASED METHODS ON THE LEVIR-MCI DATASET. RED INDICATES THE BEST. BLUE INDICATES THE SECOND BEST

Method	*Model*	*OA*	*MIoU*	*F1_R*	*P_R*	*R_R*	*F1_B*	*P_B*	*R_B*
Baseline	IFN	86.70	45.43	30.17	46.59	19.44	52.53	67.40	41.86
	BIT	89.00	33.14	7.85	11.46	12.73	18.24	5.67	8.35
	SIAM	89.72	30.85	4.67	0.52	6.45	8.66	3.66	0.27
	ELGCNet	83.94	35.29	19.71	19.10	16.66	16.37	24.13	22.91
	FTANet	87.65	39.20	24.19	27.41	19.51	35.32	31.8	22.4
	MambaCD	91.07	30.87	2.27	0.68	14.05	15.64	1.24	0.35
Random Flipping	IFN	87.09	46.76	35.27	46.85	22.34	60.62	83.77	38.18
	BIT	86.76	35.95	20.62	17.08	16.95	23.36	26.33	13.46
	SIAM	89.45	30.88	5.09	0.77	6.50	8.90	4.19	0.40
	ELGCNet	90.39	39.45	24.28	23.86	26.69	40.60	22.26	16.90
	FTANet	90.47	46.28	35.31	42.02	**35.29**	44.10	35.32	40.12
	MambaCD	90.01	32.65	13.71	0.44	15.94	29.01	12.03	0.22
Random Rotation	IFN	81.16	37.17	26.28	22.59	15.31	57.93	**92.82**	14.03
	BIT	86.71	34.42	22.36	7.24	17.09	17.80	32.34	4.54
	SIAM	91.25	30.48	0.34	0.02	7.70	10.31	0.18	0.01
	ELGCNet	90.06	36.34	26.57	5.91	24.36	57.98	29.23	3.11
	FTANet	**91.11**	42.48	21.53	38.95	**34.65**	49.46	15.61	32.12
	MambaCD	90.55	33.07	14.89	0.74	21.13	51.61	11.50	0.37
Cutout	IFN	86.56	**48.86**	34.23	55.53	21.17	**74.72**	**89.35**	44.18
	BIT	90.07	35.71	17.33	13.69	19.80	48.11	15.40	7.98
	SIAM	91.30	30.44	0.01	0.00	5.09	5.99	0.00	0.00
	ELGCNet	88.32	41.87	26.96	34.8	20.10	50.58	40.95	26.53
	FTANet	**91.24**	47.61	**37.51**	44.41	41.92	51.00	33.94	39.32
	MambaCD	90.78	33.38	15.52	1.60	26	36.65	11.06	0.82
Cutmix	IFN	90.12	56.13	**44.15**	**66.56**	29.85	**68.53**	84.69	**64.71**
	BIT	88.90	37.17	20.53	19.73	20.44	32.27	20.62	14.21
	SIAM	91.30	30.45	0.02	0.09	10.27	44.56	0.01	0.04
	ELGCNet	89.08	43.03	32.67	32.85	24.22	51.53	50.16	24.11
	FTANet	89.64	47.29	36.69	45.95	31.80	46.39	43.36	45.51
	MambaCD	90.87	33.08	10.79	4.95	22.51	45.41	7.1	2.62
OCA	IFN	67.88	31.85	16.29	31.32	9.16	25.66	73.22	40.20
	BIT	70.6	33.32	16.1	34.34	9.21	28.67	64.08	42.8
	SIAM	90.55	**48.48**	34.87	49.69	34.53	43.71	35.23	57.56
	ELGCNet	63.80	29.90	17.12	29.27	9.68	19.86	74.25	55.57
	FTANet	79.50	42.43	24.77	**50.12**	14.73	42.56	77.7	**60.94**
	MambaCD	66.14	30.43	18.38	27.17	10.63	18.17	67.71	53.82

colors. In addition to OA and MIoU metrics utilized in the WHUBCD dataset, we also evaluate the semantic CD accuracy at each category using $F1$, R, and R on the building and road categories. Specifically, $F1\_R$, P_R, and R_R are the $F1$ score, precision, and recall on the road category. $F1\_B$, P_B, and R_B are the $F1$ score, precision, and recall on the building category.

For the multicategory CD task, we mainly focus on the $F1$ scores on each category, i.e., $F1\_R$ and $F1\_B$. It can be seen from Table III that the best performances are achieved by Cutout and Cutmix when applied to IFN, achieving an $F1\_B$ of 74.72% and an $F1\_R$ of 44.15%, respectively. As building changes account for the most change cases in LEVIR-MCI, Cutout also achieves the highest MIoU of 48.86% when applied to IFN. This indicates the effectiveness of Cutout and Cutmix in semantic CD tasks.

FTANet shows remarkable adaptability to various augmentation strategies, with the best OA gains of 91.24% with Cutout compared to the baseline of 87.65%. This suggests that lightweight models with frequency-temporal attention mechanisms respond particularly well to geometric transformations that preserve structural integrity.

The OCA method produces diverse benefits across different model architectures. It significantly improves the performance of SIAM, boosting MIoU from 30.85% to 48.48% and $F1\_R$ from 4.67% to 34.87%, showing the most substantial performance improvement among all the augmentation methods. However, OCA causes performance degradation when applied to IFN, BIT, ELGCNet, and MambaCD, with OA decreases of 18.82%, 18.4%, 25.26%, and 24.93%, respectively. As we can see from the table, the low OA scores are mainly attributed to the low R_R and R_B scores. The low recall on the changed instances indicates the poor usability of OCA on low spatial resolution images. The inadaptability mainly occurs in the instance segmentation stage, which uses SAM to segment images into objects. The low spatial resolution makes the segmented road and building instances very small and difficult to be recognized by the road and building classifiers.

Traditional augmentation methods such as random flipping and random rotation show moderate but generally positive

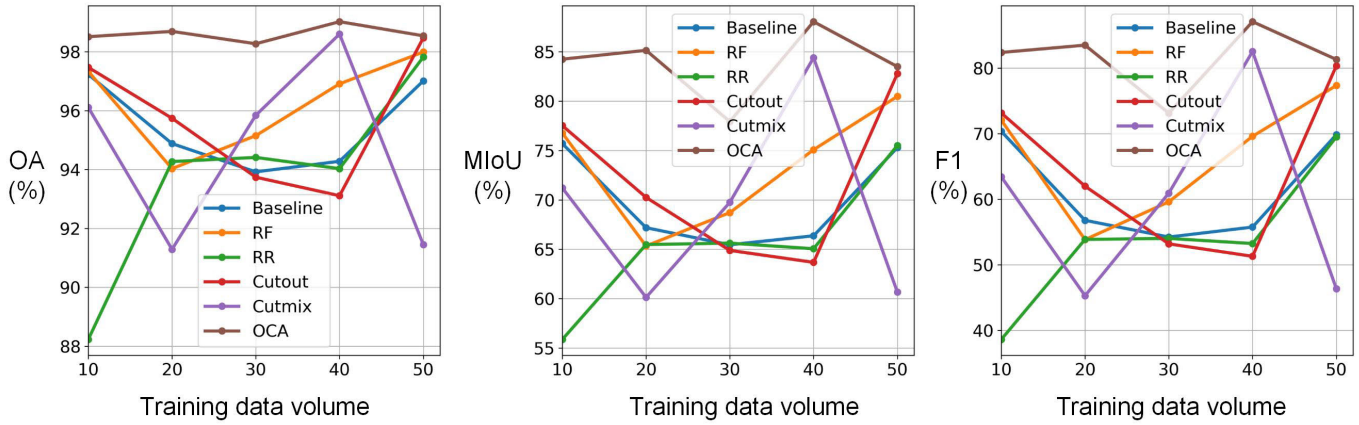

Fig. 5. Comparison of different image augmentation methods on IFN.

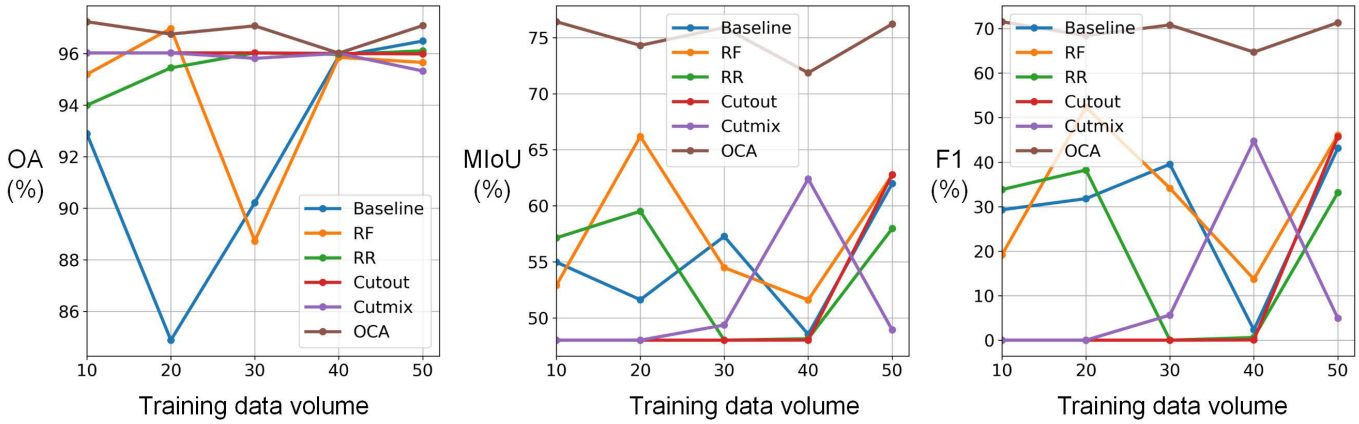

Fig. 6. Comparison of different image augmentation methods on BIT.

effects across most models. Particularly, random flipping improves FTANet's building detection ability substantially, raising its MIoU from 39.2% to 46.28%.

Interestingly, Transformer- and VMamba-based models such as BIT and MambaCD demonstrate less consistent improvements from augmentation techniques compared to CNN-based counterparts. This may be attributed to their architectural design that captures global contexts, which could be disrupted by certain local perturbations introduced by augmentation methods.

To conclude, no single augmentation strategy universally benefits all CD models. Especially in terms of MIoU, all the augmentation methods can only bring moderate improvements or even worse than the baseline. This indicates that existing vision-based augmentation methods, including OCA, are merely effective for semantic CD tasks, especially those semantic CD datasets with low spatial resolutions.

2) Performance Stability Across Varying Data Scales: Fig. 5 illustrates the performance landscapes of various augmentation methods, trained with IFN, across different training data volumes. Notably, OCA consistently achieves peak performance across all metrics and data volumes, significantly outperforming the baseline, random flipping, and rotation methods. Furthermore, the performance differences for OCA across varying training data volumes are minimal. For example, increasing the training set by 30 images only yields a 2.1% increase in $F1$ score and a 1.6% increase in MIoU, indicating its performance stability and robustness to changes in training data volume. Interestingly, the baseline, random flipping, Cutout, and Cutmix methods exhibit significant performance degradation when trained with 20 samples compared to ten samples. This decline may be attributed to category imbalance introduced by the additional ten samples. However, OCA, despite facing the same category imbalance challenge, demonstrates a performance improvement with 20 training samples, achieving a 1.11% increase in $F1$ score compared with ten training samples. While OCA also experiences performance degradation with 30 training samples, it still maintains the best overall performance compared to all other augmentation methods.

The figure shows that all the methods exhibit unstable performance across varying training samples. This suggests that simply adding more samples to the training data does not guarantee improved model generalization, especially in few-shot learning scenarios. Therefore, the selection of appropriate training samples emerges as a critical strategy in few-shot CD.

Fig. 6 presents the results obtained with the BIT model across varying training data volumes. In general, all methods

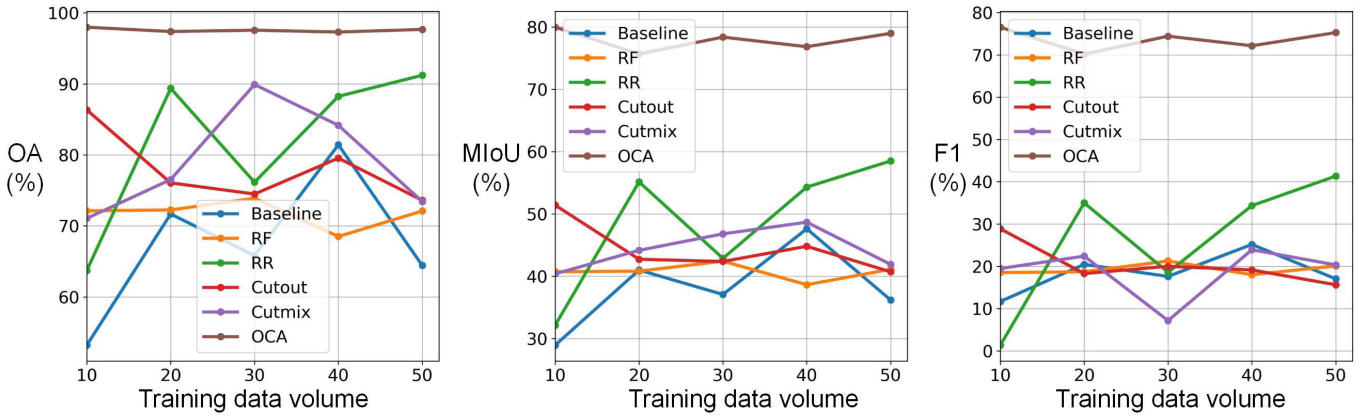

Fig. 7. Comparison of different image augmentation methods on SIAM.

perform worse than when trained with the IFN model (see Fig. 3). Notably, the baseline, random rotation, Cutout, and Cutmix methods all exhibit model unconvergence at different training volumes, resulting in performance drops to or near zero, particularly in the $F1$ score in Fig. 4. This instability suggests a greater training difficulty for Transformer architectures such as BIT, which rely on global context. As the cut-and-paste strategies of Cutout and Cutmix disrupt the global context, this leads to the observed unconvergence of BIT. In contrast, OCA demonstrates greater stability across different training data volumes, maintaining a relatively consistent OA and MIoU and achieving a noticeably higher $F1$ score compared to the other methods, especially at smaller training data volumes. This suggests that the high completeness of foreground changes and the background context facilitate model convergence within the BIT architecture.

Fig. 7 presents the performance landscapes of various augmentation methods when employed with the SIAM model. Once again, OCA consistently outperforms all other methods across different training data volumes. In contrast to the results obtained with the BIT model, the SIAM model shows a relative indifference to the benchmark augmentation methods. Specifically, random flipping, random rotation, Cutout, and Cutmix achieve $F1$ scores comparable to the baseline performance. Furthermore, the SIAM model exhibits a degree of insensitivity to variations in training data volume. This behavior may be attributed to the relatively lightweight and simple architecture of the CNN-based SIAM model, which relies more on sufficient high-quality training data. As OCA produces a large number of diverse change image pairs based on the test dataset, it achieves stable and high performance on all training data volumes.

Comparing augmentation methods across different CD models reveals that IFN consistently outperforms BIT and SIAM in all tested scenarios, suggesting its superior effectiveness. This may be attributed to the larger model size of IFN. The greater number of network parameters potentially contributes to improved model stability. However, given that the primary focus of our work lies on augmentation methods rather than CD model architectures, further investigation into this aspect is reserved for future research.

TABLE IV

PERFORMANCE COMPARISON WITH GENERATION-BASED METHODS. RED INDICATES THE BEST. BLUE INDICATES THE SECOND BEST

Method	*Model*	*OA*	*MIoU*	*F1*	*P*	*R*
Chang-en2	IFN	97.13	48.56	0.00	0.00	0.00
	BIT	97.13	48.56	0.00	0.00	0.00
	SIAM	97.13	48.56	0.00	0.00	0.00
Synthe-World	IFN	96.94	48.47	0.00	0.00	0.00
	BIT	96.94	48.47	0.00	0.00	0.00
	SIAM	96.94	48.47	0.00	0.00	0.00
IAug_inria	IFN	98.17	81.70	79.00	72.64	86.59
	BIT	96.56	72.25	64.92	54.61	80.02
	SIAM	93.56	62.82	48.81	35.66	77.34
IAug_airs	IFN	95.29	69.12	60.26	45.33	**89.86**
	BIT	**98.68**	**85.11**	**83.44**	**83.17**	83.71
	SIAM	96.51	56.18	27.40	**79.10**	16.57
OCA	IFN	98.00	71.82	62.71	70.71	56.33
	BIT	**98.79**	**83.24**	**80.74**	77.30	**84.52**
	SIAM	98.54	79.09	74.74	77.24	72.40

3) Impact of Image Size: In this section, we present performance results of OCA and synthetic data pretraining (Changen2 and SyntheWorld) and instance generation methods (IAug) to validate if the proposed method can be applied to images with various sizes and to compare its performance against SOTA generation-based methods. Specifically, all the benchmark methods were evaluated using ten labeled image pairs with a size of 512 × 512. We employed the two pretrained building instance generators (IAug_inria and IAug_airs) for change image simulation, and the corresponding results are presented in Table IV.

For the synthetic pretraining methods, CD models pretrained on Changen2 and SyntheWorld demonstrate a limited ability to transfer knowledge from the synthetic images to the target dataset. This is likely due to the limited number of fine-tuning samples and the substantial style variations between the source and target images. For the image generation-based methods, the IAug series achieved performance comparable to OCA and even surpassed it with specific CD models. When trained with IFN, IAug_inria attains the highest $F1$ score of 79% compared to 62.71% for OCA. Similarly, IAug_airs achieves the best $F1$ score of 83.44% when trained with

TABLE V
TEMPORAL EXCHANGE PERFORMANCE COMPARISON. RED INDICATES THE BEST. BLUE INDICATES THE SECOND BEST

Method	*Model*	*OA*	*MIoU*	*F1*	*P*	*R*
OCA	IFN	**98.51**	**84.25**	**82.38**	**77.86**	**87.47**
	BIT	97.24	76.42	71.55	60.56	**87.41**
	SIAM	97.98	79.95	76.54	71.01	83.00
OCA_RTE	IFN	**98.69**	**85.04**	**83.34**	**83.96**	82.73
	BIT	97.61	77.62	73.18	66.12	81.92
	SIAM	97.56	78.02	73.86	64.36	86.65

BIT, outperforming 80.74% for OCA. However, with the SIAM model, OCA achieves the best performance, reaching an $F1$ score of 74.74%, significantly exceeding 48.81% for IAug_inria and 27.4% for IAug_airs. Although OCA does not surpass IAug when trained with IFN and BIT, it exhibits consistent performance with the smallest performance gap to the best-performing method. These results validate the strong robustness of OCA to image size and its more stable and competitive performance compared to SOTA generation-based methods. Furthermore, it is important to note that IAug requires two additional pretrained models for new building instance generation, whereas OCA does not have this requirement.

4) Effectiveness of Temporal Exchange: Exchanging temporal images can simulate bidirectional temporal changes, thus further improving the diversity of change patterns in the temporal dimension. In this experiment, we randomly exchange the postchange background image with the prechange image to simulate either the pre-to-post or the post-to-pre change pattern.

As shown in Table V, when trained with IFN, random temporal exchange with OCA (OCA_RTE) gains a minor performance boost, an $F1$ score increase of 0.92%, and an MIoU increase of 0.79%. When trained with BIT, the performance gap between OCA_RTC and OCA is further enlarged. However, both OCA_RTC and OCA demonstrate lower performance compared to when trained with IFN. While the results on IFN and BIT suggest a slight performance boost brought by the temporal exchange strategy, training with SIAM presents a different trend. Specifically, OCA_RTC underperforms OCA with a significant drop of 2.68% in $F1$ score. Conclusively, the effectiveness of temporal exchange is model-dependent, offering marginal to moderate benefits for complex architectures, such as IFN and BIT. However, it shows performance degradation when implemented with a simple architecture, for example, the SIAM model.

V. CONCLUSION

Data scarcity has become a critical challenge in DL-based CD for remote sensing images. Conventional image augmentation techniques often introduce irrelevant contextual dependencies, while generation-based methods suffer from training instability and domain gaps. To overcome these limitations, we propose a novel OCA method, designed to decouple foreground change instances from background context. The key contribution lies in the two-stage approach: few-shot object classification and foreground–background pasting, enabling the generation of more effective and in-distribution synthetic training triplets. Quantitative comparisons against augmentation-based baselines, including random flipping, random rotation, Cutout, and Cutmix, demonstrated the consistent superiority of OCA across various CD models (IFN, BIT, SIAM, ELGCNet, FTANet, and ChangeMamba), particularly when trained with limited labels. Comparisons with SOTA generation-based methods (Changen2, SyntheWorld, and IAug) revealed the competitive performance of OCA while not requiring extensive pretraining or additional models.

However, we also acknowledge limitations in the current implementation of OCA. When applied to images with lower spatial resolution (e.g., 0.5 m in the LEVIR dataset [59] compared to 0.075 m in the WHU-CD dataset), the semantic segmentation model SAM struggled to effectively extract small-sized objects. In addition, we observe that the performance of SAM on building segmentation is demonstrably higher compared to other land cover objects because building instances often have regular shapes and moderate sizes. The direct applicability of OCA to CD tasks involving more irregularly shaped or variably sized land cover types requires further exploration and evaluation. In addition, the high computational cost of OCA is also a concern, primarily due to the time-consuming SAM-based unsupervised segmentation. This step becomes particularly burdensome when a large number of point prompts are used. These limitations highlight the need for lightweight SAM-based or alternative segmentation models specifically tailored to remote sensing images with various spatial resolutions. Future research should focus on developing or adapting segmentation techniques that offer robust performance across diverse land cover classes, potentially incorporating contextual information or leveraging transfer learning approaches.

REFERENCES

[1] 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 Sens. Environ.*, vol. 265, Nov. 2021, Art. no. 112636.
[2] Z. Zhu et al., "Understanding an urbanizing planet: Strategic directions for remote sensing," *Remote Sens. Environ.*, vol. 228, pp. 164–182, Jul. 2019.
[3] F. D. W. Witmer, "Remote sensing of violent conflict: Eyes from above," *Int. J. Remote Sens.*, vol. 36, no. 9, pp. 2326–2352, May 2015.
[4] D. Lu, P. Mausel, E. Brondízio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, 2004.
[5] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.
[6] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*.
[7] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, "Research progress on few-shot learning for remote sensing image interpretation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2387–2402, 2021.
[8] T. Liu et al., "Building change detection for VHR remote sensing images via local–global pyramid network and cross-task transfer learning strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4704817.
[9] H. Tang, Y. Li, X. Han, Q. Huang, and W. Xie, "A spatial–spectral prototypical network for hyperspectral remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 167–171, Jan. 2020.

[10] J. Zhu et al., "MVP: Meta visual prompt tuning for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5610413.

[11] C. Qiu et al., "Few-shot remote sensing image scene classification: Recent advances, new baselines, and future trends," *ISPRS J. Photogramm. Remote Sens.*, vol. 209, pp. 368–382, Mar. 2024.

[12] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2021, pp. 8748–8763.

[13] X. Wang, S. Li, X. Zhao, and K. Zhao, "BiG-FSLF: A cross heterogeneous domain few-shot learning framework based on bidirectional generation for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5516213.

[14] S. Zhu, W. Jing, P. Kang, M. Emam, and C. Li, "Data augmentation and few-shot change detection in forest remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5919–5934, 2023.

[15] P. L. Rosin, "Thresholding for change detection," *Comput. Vis. Image Understand.*, vol. 86, no. 2, pp. 79–95, May 2002.

[16] F. Bovolo, S. Marchesi, and L. Bruzzone, "A framework for automatic and unsupervised detection of multiple changes in multitemporal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2196–2212, Jun. 2012.

[17] M. Noman, M. Fiaz, H. Cholakkal, S. Khan, and F. S. Khan, "ELGC-Net: Efficient local–global context aggregation for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4701611.

[18] T. Zhu et al., "FTA-Net: Frequency-temporal-aware network for remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 3448–3460, 2025.

[19] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep Siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.

[20] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.

[21] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Understand.*, vol. 187, Oct. 2019, Art. no. 102783.

[22] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral–spatial–temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[23] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 207–210.

[24] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.

[25] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.

[26] 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 Sens.*, vol. 15, no. 4, p. 949, Feb. 2023.

[27] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2021, *arXiv:2111.00396*.

[28] H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya, "ChangeMamba: Remote sensing change detection with spatiotemporal state space model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4409720.

[29] J. N. Paranjape, C. De Melo, and V. M. Patel, "A mamba-based Siamese network for remote sensing change detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Feb. 2025, pp. 1186–1196.

[30] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5633520.

[31] G. Hoxha, S. Chouaf, F. Melgani, and Y. Smara, "Change captioning: A new paradigm for multitemporal remote sensing image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5627414.

[32] S. Chang and P. Ghamisi, "Changes to captions: An attentive network for remote sensing change captioning," *IEEE Trans. Image Process.*, vol. 32, pp. 6047–6060, 2023.

[33] C. Liu, K. Chen, B. Chen, H. Zhang, Z. Zou, and Z. Shi, "RSCaMa: Remote sensing image change captioning with state space model," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[34] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[35] X. Yu, X. Wu, C. Luo, and P. Ren, "Deep learning in remote sensing scene classification: A data augmentation enhanced convolutional neural network framework," *GISci. Remote Sens.*, vol. 54, no. 5, pp. 741–758, May 2017.

[36] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 3–22, Nov. 2018.

[37] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," 2022, *arXiv:2209.08290*.

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

[39] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[40] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[41] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," 2019, *arXiv:1905.04899*.

[42] M. B. Pereira and J. A. dos Santos, "ChessMix: Spatial context data augmentation for remote sensing semantic segmentation," in *Proc. 34th SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2021, pp. 278–285.

[43] X. Xu et al., "A data augmentation strategy combining a modified pix2pix model and the copy-paste operator for solid waste detection with remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8484–8491, 2022.

[44] J. Qin, J. Fang, Q. Zhang, W. Liu, X. Wang, and X. Wang, "ResizeMix: Mixing data with preserved object information and true labels," 2020, *arXiv:2012.11101*.

[45] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "ClassMix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1368–1377.

[46] Y. Su, R. Sun, G. Lin, and Q. Wu, "Context decoupling augmentation for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6984–6994.

[47] J. Zhang, Y. Zhang, and X. Xu, "ObjectAug: Object-level data augmentation for semantic image segmentation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[48] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[49] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[50] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[51] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603216.

[52] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 14–34, Sep. 2021.

[53] Z. Zheng, S. Tian, A. Ma, L. Zhang, and Y. Zhong, "Scalable multi-temporal remote sensing change data generation via simulating stochastic change process," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21818–21827.

[54] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3836–3847.

[55] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[56] C. Liu, K. Chen, H. Zhang, Z. Qi, Z. Zou, and Z. Shi, "Change-agent: Toward interactive comprehensive remote sensing change interpretation and analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5635616.
[57] Z. Zheng, S. Ermon, D. Kim, L. Zhang, and Y. Zhong, "Changen2: Multi-temporal remote sensing generative change foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 2, pp. 725–741, Feb. 2025.
[58] J. Song, H. Chen, and N. Yokoya, "SyntheWorld: A large-scale synthetic dataset for land cover mapping and building change detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 8272–8281.
[59] H. Chen and Z. Shi, "A spatial–temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, 2020.

Chenxiao Zhang received the Ph.D. degree in geographic information systems from Wuhan University, Wuhan, China, in 2020.

Since 2021, he has been an Associate Research Fellow at the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include geospatial data mining, remote sensing image interpretation, and geospatial intelligence.

Peng Yue (Senior Member, IEEE) received the bachelor's degree in surveying and mapping from Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 2000, and the master's and Ph.D. degrees in remote sensing from Wuhan University, Wuhan, in 2003 and 2007, respectively.

Since October 2015, he has been the Director of the Institute of Geospatial Information and Location-Based Services and the Deputy Director of the Department of Geographic Information Engineering, School of Remote Sensing and Information Engineering, Wuhan University. In June 2020, he was appointed as the Vice Dean of the School of Remote Sensing and Information Engineering. He is deeply involved in research on remote sensing big data, geospatial artificial intelligence, and spatiotemporal big data platforms and analysis. He has authored over 100 papers, including more than 60 SCI-indexed papers, and holds over 50 patents and software copyrights.

Dr. Yue is the first Chinese Scholar to chair IEEE Geoscience and Remote Sensing Society's Technical Committee and holds a leadership role in the OGC Technical Standards Working Group. In addition, he is the founding President of the OGC China Forum and serves as the Chair/Co-Chair of several international OGC standards working groups.

Francesca Cigna received the B.Sc. and M.Sc. degrees (cum laude) in environmental engineering from the University of Palermo, Palermo, Italy, in 2006 and 2007, respectively, and the Ph.D. degree in Earth sciences from the University of Florence, Florence, Italy, in 2011.

She was with Italian Space Agency (ASI), the British Geological Survey of the Natural Environment Research Council, and the University of Florence. Since 2021, she has been a Senior Researcher in Earth observation with the Institute of Atmospheric Sciences and Climate of the National Research Council (ISAC-CNR), Rome, Italy. Her research interests include satellite SAR and optical imagery, advanced InSAR and change detection methods, mapping and monitoring of natural and anthropogenic hazards and risks in urban and rural environments, shallow geological processes, land subsidence, hydrogeology, landscape archeology, and cultural heritage.

Deodato Tapete is a Researcher in Earth Observation and a Data Analyst at Italian Space Agency (ASI). He has worked on several national and international projects across different geographic areas, from Italy to Middle East and North Africa, the United Kingdom, and Latin America. He has recently cooperated with the Archeological Park of Colosseum and is currently involved in the Extraordinary Plan of Monitoring and Conservation of Immovable Cultural Heritage led by Italian Ministry of Culture.

Dr. Tapete is a member of the Computer Applications and Quantitative Methods in Archaeology (CAA) Scientific Committee and a member of the Aerial Archaeology Research Group and European Association of Archaeologists (EAA).