Adversarial Instance Augmentation for Building Change Detection in Remote Sensing Images

Hao Chen, Wenyuan Li and Zhenwei Shi*, *Member, IEEE*

***Abstract*—Training deep learning-based change detection (CD) models heavily relies on large labeled datasets. However, it is time-consuming and labor-intensive to collect large-scale bi-temporal images that contain building change, due to both its rarity and sparsity. Contemporary methods to tackle the data insufficiency mainly focus on transformation-based global image augmentation and cost-sensitive algorithms. In this paper, we propose a novel data-level solution, namely Instance-level change Augmentation (IAug), to generate bi-temporal images that contain changes involving plenty and diverse buildings by leveraging generative adversarial training. The key of IAug is to blend synthesized building instances onto appropriate positions of one of the bi-temporal images. To achieve this, a building generator is employed to produce realistic building images that are consistent with the given layouts. Diverse styles are later transferred onto the generated images. We further propose context-aware blending for a realistic composite of the building and the background. We augment the existing CD datasets and also design a simple yet effective CD model - CDNet. Our method (CDNet + IAug) has achieved state-of-the-art results in two building CD datasets (LEVIR-CD and WHU-CD). Interestingly, we achieve comparable results with only 20% of the training data as the current state-of-the-art methods using 100% data. Extensive experiments have validated the effectiveness of the proposed IAug. Our augmented dataset has a lower risk of class imbalance than the original one. Conventional learning on the synthesized dataset outperforms several popular cost-sensitive algorithms on the original dataset. Our code and data will be made publicly available.**

***Index Terms*—High-resolution optical remote sensing image, Convolutional neural networks, Building change detection, Adversial instance augmentation, Synthetic data.**

I. Introduction

CHANGE detection (CD) based on remote sensing (RS) images is the process of identifying differences in RS images at different times in the same geographical location [1]. Nowadays, the availability of very high-resolution (VHR) satellite data (e.g., WorldView-3, QuickBird, and Gaofen-2) and aerial data is opening up new avenues for urban monitoring at a fine scale. Specifically, the detailed spatial information provided by the VHR optical RS images makes it possible to detect small objects, such as buildings at the instance level. Identifying the change of buildings has a wide range of applications in urban planning [2], illegal construction detection [3], and disaster assessment [4]. Information extraction based on RS images is still mainly based on manual visual interpretation. Automatic building CD technology can reduce considerable labor costs and time consumption, which has raised increasing attention [2, 5–9].

Supervised deep learning techniques have achieved great success in information extraction on RS images due to its powerful ability to learn high-level feature representation [10–13] The prosperity of deep learning technology is inseparable from large labeled datasets. Unfortunately, in the remote sensing image building CD task, it is hard to collect effective bi-temporal images because of the rarity and sparsity of the positive class (see Fig. 1 (left)). Annotating a large-scale CD dataset is also time-consuming and labor-intensive. Existing building CD datasets [6, 14] usually only cover very small regions and limited image conditions. A deep learning-based CD model lacks sufficient generalization ability to be applicable to new RS images that contain building objects of different appearances or that are obtained from different imaging conditions, if not preparing new training data through heavy work. Especially, when only a small amount of training data available, the CD model is prone to overfitting or presenting poor performance on the change category. Therefore, it is of great value to develop an automated method to synthesize new change detection data that contains plenty of target changes.

Contemporary methods to improve the generalization ability of the CD model under a small data regime are mainly focusing on transformation-based image augmentation (e.g., flip and rotation) [6], transferring a pre-training model (e.g., from ImageNet) [15], or adjusting the optimization objectives (e.g., the weighted loss) [16–18]. Different from existing methods, we propose a novel data-level solution to improve building CD performance. Our synthesis method, namely Instance-level change Augmentation (IAug), can generate new CD data that contains changes involving spatially densely distributed and color-diverse buildings by leveraging generative adversarial training and image blending. For ease of implementation, we augment the samples from the existing CD dataset with synthesized building targets. To this end, we aim to synthesize effective and realistic CD samples (see Fig. 1 (right)) by making full use of existing bi-temporal images and building targets. The motivation of our method lies in two aspects:

Firstly, the building change in the real scene is usually rarely and sparsely distributed. Conventional learning algorithms on such an imbalanced dataset including rare classes may bias towards dominant classes while exhibiting poor performance

The work was supported by the National Key R&D Program of China under the Grant 2019YFC1510905, the National Natural Science Foundation of China under the Grant 61671037 and the Beijing Natural Science Foundation under the Grant 4192034. *(Corresponding author: Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn))*

Hao Chen, Wenyuan Li and Zhenwei Shi are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China.

Fig. 1. An example of a CD sample in the LEVIR-CD dataset [6]. (Left) An original CD sample. (Right) The synthesized CD sample using our instance augmentation method.

in minority classes. Superimposing more changed targets into the image can increase the number of the positive class, and reduce the risk of class imbalance.

Secondly, the redundant information in the no-change areas (e.g., bare land, grassland) may bring limited gain to the performance of a CD model. We could blend building targets of diverse patterns on these areas to further enhance the discrimination power of the CD model.

The key idea of IAug is to blend the external building instance onto an appropriate position of one of the bi-temporal images. In the framework of IAug, there are two main components: building object generation and CD sample synthesis. The procedure of IAug is demonstrated in Fig. 2 (top).

The first stage is to generate building targets. We propose a building synthesis approach for controllable shape and appearance. Firstly, a semantic building generator is employed to generate realistic images conditioned on the input semantic layouts. The GAN-generated images are semantically aligned with the input labels. Then, a color transfer method is proposed to control the style of the generated building image. Through the transfer process, we could generate images with more diverse styles. In the second stage, we blend the generated building instances on each sample in the existing building CD dataset to synthesize new samples. Context information (e.g., shadow) is an important clue for building detection [19]. Therefore, we propose a context-aware approach, i.e., extract the context surrounding a building and then utilize it for realistic and effective image composition. Furthermore, we employ several blending strategies to obtain diverse results, which can prevent the CD model from overfitting one composition mode.

To inspect the effectiveness of the proposed IAug, we also design a simple yet effective CD network (CDNet), which consists of a feature extractor (deep siamese fully convolutional networks) and change classifier (shallow fully convolutional networks), as shown in Fig. 2 (bottom). We adopt a late-fusion difference strategy to fuse the bi-temporal information, that is, to difference the high-level bi-temporal features to obtain the feature difference image (FDI). We prefer to employ "difference" rather than "concatenate" because the "concatenate" operation introduces asymmetry to the model, which is contrary to the task of symmetric CD (i.e., when swapping the chronological order of the bi-temporal images, the prediction result remains the same).

The contribution of this work can be summarised as follows:

- We propose a synthesis framework, namely IAug, to efficiently synthesize new CD samples that contain changes involving plenty and diverse buildings. We augment the existing CD dataset by leveraging generative adversarial training and image blending. Our synthesized dataset can also reduce the risk of class imbalance.
- We propose a building synthesis method towards controllable shape and style. As far as we know, we are the first to use GAN-based synthesized image data to manipulate the CD samples in a controllable manner. Furthermore, our context-aware instance augmentation could synthesize realistic and effective CD samples.
- We have reproduced several state-of-the-art (SOTA) CD methods on both the LEVIR-CD and WHU-CD datasets, and our method (IAug + CDNet) obtains the best results. Our method achieves comparable results with only 20% of the training data as the current SOTA methods using 100% data.

II. Related Work

A. Building Change Detection Methods

Some progress has been made in building change detection for high-resolution optical RS images.

Many early attempts extract handcrafted features that contain spatial/contexture information of buildings in the texture-rich images. Spatial features such as gray level co-occurrence matrices [3, 20, 21], wavelets [21], and morphological features [2, 3, 20, 22–24], are employed as a complement of spectral features to suppress false alarms. The morphological building index (MBI) [25] is widely employed for indicating the presence of buildings. For example, Huang et al. [2] leverage MBI, spectral variation, and shape conditions to identify the building change in an unsupervised manner.

Traditional building CD methods that rely on handcrafted features have limited performance due to insufficient feature discrimination. Recently, deep learning techniques, especially deep convolutional neural networks (CNN), which automatically learn hierarchical abstract image representations, have been successfully applied in building CD [5, 7, 8, 26, 27]. When giving sufficient training samples, deep learning-based methods show superior performance than traditional counterparts [11]. Our paper falls into the deep learning-based approach.

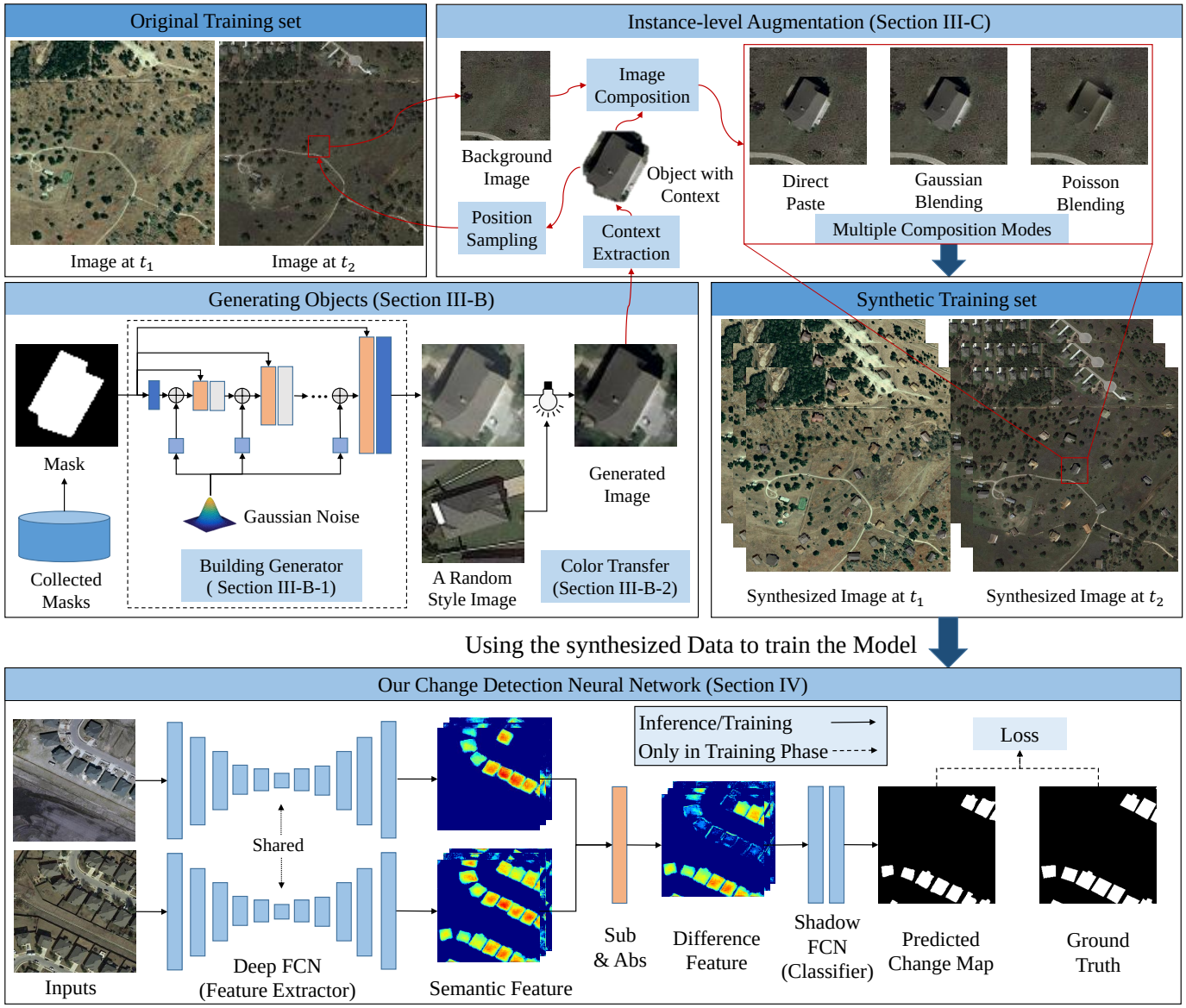

Fig. 2. Illustration of our proposed method. (Top) Procedure of synthesizing CD samples. (Bottom) Change detection model.

B. Deep learning based remote sensing image change detection

Here, we provide a brief review of extant DL-based supervised CD methods for optical RS images. In general, there are two main streams for using CNNs for CD [28].

The post-classification method [7, 27, 29] has two steps. First, a CNN/FCN is trained to separately classify the bi-temporal images, and then their classification results are compared to obtain the change category. For example, Ji et al. [7] used an FCN for pixel-wise building segmentation, and then fed the two binary building maps into a change detection network to obtain the change map. Because of requiring the semantic labels for both temporal images, this kind of method is impractical in the condition of only the change label available.

Another approach trains CNNs to directly generate the change result from the bi-temporal images. Many early attempts [30–32] model the CD task as a similarity detection process by splitting bi-temporal images into many patch-level pairs and applying a CNN on each pair to obtain its center prediction. The Pixel-level approach [5, 6, 8, 18, 33–41] uses FCNs to directly produce a high-resolution change map from the two inputs, which is usually more efficient and reliable than the patch-level approach. Existing FCN-based methods for fusing the two temporal information can be roughly divided into image-level and feature-level. Image-level fusion [33–35] concatenates the bi-temporal images as a single input to a semantic segmentation network. Feature-level fusion [5, 6, 8, 18, 33, 36–41] fuses the bi-temporal features from the middle of neural networks and then make decisions based on the fused features.

We conclude that the recent advances in RS CD are mainly focusing on addressing the three issues: 1) enhance feature discrimination power, 2) small labeled data, 3) imbalanced

CD samples.

Many recent works aim to improve the feature discrimination power of the neural networks, such as designing multi-level feature fusion structures [5, 6, 8, 36, 38], introducing attention modules [5, 8, 36], and self-attention mechanism [6, 41], or combining GAN-based optimization objectives [38, 39, 42, 43]. For instance, Hou et al. [38] introduced the GAN loss into CD to model the distribution of the two input images and the change map such that the CD network could generate more desirable results. Reducing the domain discrepancy of the two temporal images in the image-level can indirectly enhance network discrimination power. Fang et al. [39] used CycleGAN [44] as a preprocessing method to translate the bi-temporal images into one single domain so that the irrelevant appearance differences can be reduced and the real change can be highlighted.

To solve the small labeled data issue, transfer learning [15, 28], semi-supervised learning [45, 46] and active learning [47] have been adopted in recent work. We will latter discuss the class imbalance of CD in Section II-C.

The main purpose of our paper is to explore synthesizing data for enhancing the CD performance. Moreover, we design a simple yet effective symmetric siamese FCN for CD as our change detection network. We argue that symmetric structure is important for binary CD, which has rarely been discussed in recent work (see Sec. IV). The most similar CD work to us is [31]. They used the DIRSIG [48] generated imagery to train a patch-level CD network. The DIRSIG simulation environment could generate imagery from a constructed 3D scene model and illumination condition. However, constructing 3D scenes are so time-consuming and laborious that it is difficult to scale to large scenes. Also, there still remains a domain gap between the generated images and real-world images. Different from previous works, we synthesize the instance changes by leveraging the advanced GAN techniques. And we blend the generated instances onto the real CD samples. Our approach has many advantages: 1) generate realistic composite CD samples, 2) able to control the number and shape of changed instances, 3) easy to scale to a large dataset, 4) alleviate the small labeled data and class imbalance issues.

C. Class imbalance in change detection

The class imbalance phenomenon in remote sensing image change detection is severe due to the intrinsic low-frequency of change in the real-world. The targets of interest usually only occupy a much smaller number of pixels than the background [49]. In other words, the number of pixels that belong to the change class is much less than that of no-change. Naive machine learning algorithms on such imbalanced data have a bias toward the no-change class and tend to ignore the change class. Many studies [6, 8, 16, 17, 17, 18, 28, 50] have been performed to solve the class imbalance on the CD task.

One type of method is to over-sample the change examples such that the same number of positive and negative samples are selected for training the learner [50]. A more common way is to use weighted losses of different versions for enforcing the learner paying more attention to the change examples in the training phase. For instance, weighted cross-entropy loss [16, 17, 28], weighted dice loss [17] and weighted contrastive loss [6, 18] have been explored in recent CD works. Liu et al. [8] proposed a weighted focal loss that reshaped the original focal loss and added different weights in a non-linear form to different classes. Different from existing CD works, we leverage the advanced image generation and composition techniques to synthesize new samples of the change class. To the best of our knowledge, we are the first to blend the GAN-generated targets onto the bi-temporal sample to augment the number of instance changes.

III. Instance-level augmentation for synthesizing change detection samples

In this section, we firstly give an overview of the procedure for synthesizing CD data, then introduce a semantic object synthesis method towards controllable shape and appearance. Lastly, we present context-aware instance augmentation to synthesize CD samples.

A. overview

The real-world RS building CD task exhibits an imbalance in class distribution, wherein the number of the no-change class is much more than that of change. It is time-consuming and laborious to collect large amounts of bi-temporal images that contain changes of buildings. We present an automated synthesis method to effectively synthesize efficient CD data based on the existing CD dataset by leveraging additional building targets with semantic segmentation labels. We term our synthesis method as instance-level change augmentation (IAug). Our synthesizing procedure has two main steps:

1) **Object image generation**. We train a semantic building generator to generate an object image of controllable shape and size by specifying the input semantic mask. Then we transform the style of the generated image by matching its color distribution to that of a random style image. In this way, we can obtain collections of generated object/mask pairs. For more details see Section III-B.
2) **CD sample synthesis**. For each sample from the original CD dataset, we sequentially blend (with multiple modes) a certain number of object instances (with context) on any appropriate positions (via position sampling) of any one of the bi-temporal images. More details of our position sampling, context-aware blending, and multiple composition modes are later discussed in Section III-C.

B. Object Image Generation

1) Semantic Building Generator: Instead of directly using the cropped object images from the building segmentation dataset for image composition, we train a conditional generative adversarial network (GAN) on the collected object samples to generate building images. Our GAN-based generation approach has two main advantages: 1) accurate object mask. As shown in Fig. 3, the object mask from the original building labeling dataset is not accurately aligned with the building roof. The object cropped by the incorrect mask may

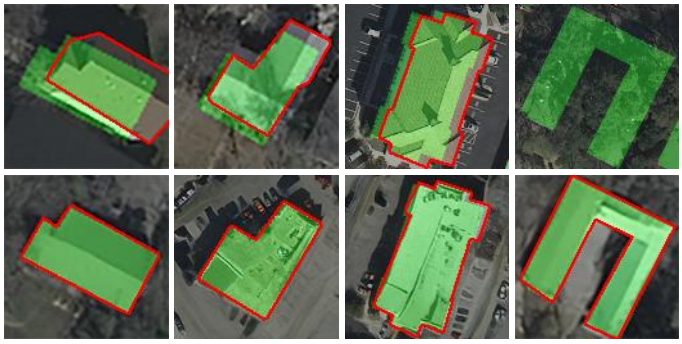

Fig. 3. Comparison of semantic consistency between the original image and the generated image. (Top) Some selected misaligned labels/wrong labels in the Inria dataset [51]. (Bottom) GAN-generated images. Our generated objects are well aligned with the label. (Green denotes the object mask and red denotes the outline of the building)

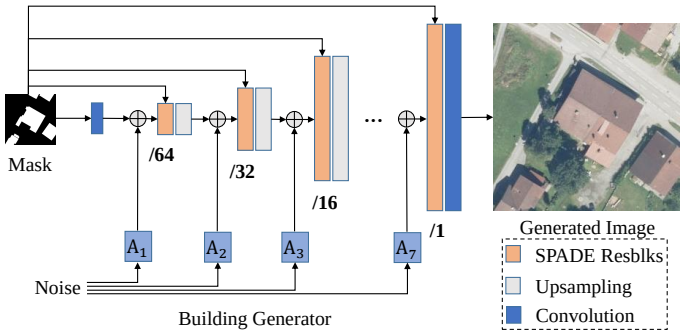

Fig. 4. Architecture of our building generator.

bring in a negative impact on the CD model. On the contrary, our approach could generate the object well aligned with the semantic mask. 2) controllable geometry characteristics. The building object in different datasets may have different geometry characteristics (shape and size) due to both the variance of buildings across different regions and the difference in camera conditions. To generate objects with similar geometry characteristics as in the target CD dataset, we could feed the generator the semantic mask cropped from the target dataset.

Semantic image synthesis refers to the task of generating realistic images conditioning on the input semantic layout [52, 53]. In our setting, we aim to generate a photorealistic image given a semantic label map, which has two classes: building and others.

Here, we present our building generator, which is based on a SOTA conditional GAN (GauGAN [53]). GauGAN is a generator network composed of several SPADE residual blocks (Resblks) with upsampling. The semantic layout is fed into each block to better preserve semantic information through the whole depth of the generator. We make a little modification on the original GauGAN to produce more diverse and higher quality synthesis results. Concretely, we introduce explicit learnable noises into each stage of generation. Previous works [54, 55] have shown that adding such noise could improve the quality of synthesized images.

Fig. 4 illustrates the architecture of our building generator G, which starts from a downsampled semantic mask, transforms it into a photorealistic image in a progressive manner via seven SPADE Resblks [53]. We add a noise inserting layer before each SPADE Resblk. To achieve this, we generate seven single-channel images consisting of uncorrelated Gaussian noise. These noise images are fed into each noise inserting layer respectively. In a noise inserting layer, the noise image is broadcasted to feature maps using learnable per-channel scaling factors A, then the generated noise maps are added to the original feature maps. Except for the last Resblk, an upsampling layer via bilinear interpolation is added after each Resblk. Therefore, our generator has six upsampling layers in total. At first, the input semantic mask is downsampled to $h/64 \times w/64$ from $h \times w$, then pass through the generator to output a synthesized image of size $h \times w$.

The architecture of our discriminator D follows the one used in the GauGAN [53]. It takes the concatenation of the semantic label map and the image as input.

Given a training set including pairs of corresponding images $\{(\mathbf{s_i}, \mathbf{x_i})\}$, where $\mathbf{s_i}$ is a semantic label map and $\mathbf{x_i}$ is a corresponding real image, our conditional GAN learns to generate new data with the same statistics as the real images conditioned on the input semantic label maps via the following minimax game:

$$\min_G \max_D \mathcal{L}_{\mathrm{GAN}}(G, D), \tag{1}$$

where we employ a hinge loss as our GAN loss:

$$\begin{aligned} \mathcal{L}_{GAN} = & E_{(\mathbf{s},\mathbf{x})\sim p_{\mathrm{data}}(\mathbf{s},\mathbf{x})}[\max(0, 1 - D(\mathbf{s}, \mathbf{x}))] \\ & + E_{\mathbf{s}\sim p_{\mathrm{data}}(\mathbf{s})}[\max(0, 1 + D(\mathbf{s}, G(\mathbf{s})))]. \end{aligned} \tag{2}$$

To further improve the performance, we employ the discriminator feature matching loss $\mathcal{L}_{\mathrm{F}}$ [52] and the perceptual loss $\mathcal{L}_{\mathrm{P}}$ [56]. Therefore, our full objective function is the weighted sum of the GAN loss, feature matching loss, and perceptual loss, which is given by:

$$\begin{aligned} \mathcal{L} = & \min_G \max_D \mathcal{L}_{\mathrm{GAN}}(G, D) \\ & + \lambda_F \mathcal{L}_{\mathrm{F}}(G, D) + \lambda_P \mathcal{L}_{\mathrm{P}}(G, D), \end{aligned} \tag{3}$$

where λ_F, λ_P control the importance of the three terms. We follow Pix2PixHD [52] to set $\lambda_F = 10, \lambda_P = 10$.

We use Adam solver with $\beta_1 = 0, \beta_2 = 0.9$ to train our generator for 100 epochs. The initial learning rate of 2×10^{-4} is used for the generator and discriminator. We keep the same learning rate for the first 50 epochs and linearly decay it to 0 over the remaining 50 epochs.

2) Object Style Transfer: We observe that the GAN-generated building images have a different appearance (i.e., color distribution) from those in the building CD dataset. Formally, let source S be a set of GAN-generated images, and target T be a set of building images from the original CD dataset. In other words, there exists a domain shift between S and T. Unpaired image-to-image translation methods, such as CycleGAN [44] could be utilized to fill in the domain gap. However, in our setting, we do not have enough object samples from the original CD dataset to train CycleGAN. Therefore, we resort to a simple yet effective non-learning approach to match the color distribution of the two image sets. We transfer the style of a random reference image onto

Fig. 5. Examples of color transfer. The color of the target image is transferred to the source image. The target images are cropped from the LEVIR-CD dataset [6] and the source images are generated by our building generator.

Fig. 6. An example of composite images to illustrate the importance of context. Superimpose the building target (a) onto the image (b) with different modes to obtain the image composition: no context (c), shadow (d), and shadow + neighborhood (e).

Fig. 7. Comparison of different image composition modes. (a) Direct paste. (b) Gaussian blending. (c) Poisson blending.

the object image. By doing so, we can obtain various styles of transformed images, whose distribution is closer to that of the target domain. Fig. 5 illustrates some selected examples of object style transfer.

To achieve this, we employ a color transfer (CT) method [57] which can change one image's color characteristics to accord with another in the three-dimension color space directly. Considering an image as a set of points in the RGB space, we can fit this cluster using a 3-dimensional Gaussian distribution. The key of this method is to calculate a transformation matrix, which moves data points of the source cluster by scaling, rotation, and translating, such that the transformed cluster has the same mean and covariance as the target one.

We make a little modification on the original CT method towards a more faith color transfer in the building areas. Our hypothesis is that the pixels that belong to the same kind of category in one image follow the Gaussian distribution. The non-building areas in the image may contain various kinds of objects (e.g., shadow, tree, grass, soil, and road), and it may not be suitable to describe all the pixels in these areas with a single Gaussian distribution. Therefore, we use the pixels that belong to the building area, instead of all the pixels in the image, to calculate the transformation matrix. Then we apply this matrix to translate the source images. Because we have the semantic masks that indicate the building category for both the source image and the target image, it is easy to implement this idea. Please note that although the non-building area in the transformed image may present an unnatural appearance due to the calculated transform matrix may not be suitable for objects of all kinds, its impact can be reduced by context extraction and image blending in the follow-up process.

C. CD Sample Synthesis

Here, we present context-aware instance-level augmentation to synthesize CD samples. Our synthesis method has three main steps:

- **Object Context Extraction**. The mere presence of the context surrounding the building is a critical cue for object recognition. For realistic and effective image composition, we cut the building area as well as its nearby context (i.e., shadow and neighborhood pixels) from the object image. We display an example of the composition results in Fig. 6 to show the importance of the context information for realistic image synthesis. Our shadow extraction algorithm is given in Section III-C1. More context information (neighborhood pixels) could be obtained by the subtraction between the dilated object mask and the original object mask.
- **Position Sampling**. A naive method is to uniformly sample a position in the image for inserting the object. To avoid inserting the object overlapped with the existing object in the image, we use a reference mask R to guide the sampling process. The reference mask records the areas of existing objects. We use rejection sampling [58] to avoid sampling the positions within the object areas. The reference mask is updated after a new object insertion. The initial reference mask is the union of two temporal label maps. The label map can be generated by feeding the original background image to a semantic segmentation model (UNet) [59], which has been trained on a building segmentation dataset.
- **Image Composition**. We utilize three different image composition methods to prevent the CD model from overfitting one composition mode. Fig. 7 shows the composition results of different image blending methods.

To sum up, we give the overall process of synthesizing CD samples. First, for each sample (bi-temporal images B_1, B_2, a change mask L) in the CD dataset D, we randomly sample N objects from the generated building object dataset D_{object}. Then we extract the mask with context M_c for each object image I. Rejection sampling is used to sample an appropriate position from the reference mask R, such that the inserted object is not overlapped with existing objects. Finally, each

object is superimposed on an image of either temporal. Here, we use three different blending modes to obtain three groups of augmented samples. Details of Instance-level augmentation for CD sample synthesis are shown in Algorithm 1.

Algorithm 1: Instance-level Augmentation for CD Sample Synthesis.

Input: $D_{object} = \{(I^k, M^k)|k = 1 : K\}$ (a set of building samples with a size of K)
Input: N (the number of instances to blend on each CD sample)
Input: $D = \{(B_1^j, B_2^j, L^j)|j = 1 : J\}$ (the original CD training set with a size of J)
Input: $R = \{(R^j|j = 1 : J\}$ (the referenced label set with a size of J)
Output: D_{aug} (augmented CD training set)

Initialize $D_{aug} \leftarrow \emptyset$
// iterate each sample in D
for j *in* $1 : J$ **do**
  // perform image composition for each blending mode
  for *mode in* {*'Direct', 'Gaussian', 'Poisson'*} **do**
    $B_1, B_2, L, R \leftarrow B_1^j, B_2^j, L^j, R^j$
    // random sample N instances from D_{object}
    for i *in* $1 : N$ **do**
      sample $(I^i, M^i) \sim D_{object}$
      $M_s^i \leftarrow$ShadowExtract(I^i, M^i)
      $M_c^i \leftarrow$dilation(M_s^i)
      // sample either image from two temporals
      sample $B_t \sim \{B_1, B_2\}$
      // sample an effective position from R
      while *True* **do**
        $h, w \leftarrow sizeof(M^i)$
        $H, W \leftarrow sizeof(R)$
        sample $x \sim$ Uniform$(0, H - h)$
        sample $y \sim$ Uniform$(0, W - w)$
        if $M \cap R[x : x + h, y : y + w]$ *is None* **then**
          break
        end
      end
      // blend the object and B_t, update L, R
      $B_t[x : x + h, y : y + w] \leftarrow$ ImageComposite($B_t[x : x + h, y : y + w], I^i, M_c^i$,mode)
      $L[x : x + h, y : y + w] \leftarrow M^i$
      $R[x : x + h, y : y + w] \leftarrow M_c^i$
    end
    $D_{aug} \leftarrow D_{aug} \cup (B_1, B_2, L)$
  end
end

1) Shadow Extraction: Based on the observation of the generated object image, we conclude that the shadow area has three important attributes, in terms of brightness, shape, and location.

- Brightness. The brightness of a shadow area is usually smaller than that in other areas due to its low spectral reflectance. Most shadow regions could be extracted by pre-defined thresholding.
- Location. The shadow is usually near a building structure. We could exclude the dark pixels that are far away from the building or is inside the building region. Moreover, a building instance has at least four sides. We observe that in most images more than two sides of the building cast shadows and the centroid of the shadow structure is in the building region. We could use these properties to choose the real shadow area from candidate areas.
- Smoothness. The shadow is usually continuous and contains no holes.

Based on the above recognition, we propose a simple yet effective method to extract the shadow structure surrounding the building. Given an object image I and its mask M, we want to obtain the mask with shadow M_S. Here, we give the three main steps to obtain M_S.

- Thresholding. We segment from I a group of dark pixels, each of whose average intensity is below a pre-defined threshold t.
- False alarm removal. We use a morphological dilation operation to obtain the region surrounding a building structure. $M_e = M \oplus E - M$, where M_e denotes the region surrounding the building, the structuring element E has a size of $e \times e$. We could adjust the e to control the expanding size. We remove the dark pixels that do not belong to this region. Then we obtain all the connected components in this region as candidate shadow areas. For each candidate shadow, if its centroid is in the building area, we assign this component as a shadow area, otherwise remove it.
- Hole filling. We fill the holes in the remaining shadow areas to give the final shadow mask. We merge the shadow mask with M to get M_S.

Fig. 8 shows some selected examples of shadow extraction. We can observe that the presence of the shadow area of the building makes the building image more realistic.

2) Image Compostion: Let I be an object image, M_c be its mask with context, B_{patch} be a cropped patch from the background image, and C be the composite image. Here, we give three methods to calculate C.

- Direct paste. The masked object is directly placed on the selected position in the background image. The composite image can be calculated by an alpha blending: $C = M_c \cdot I + (1 - M_c) \cdot B_{patch}$.
- Gaussian blending. Similarly, we use alpha blending to composite images. The difference is that we blur the object mask by a Gaussian filter to alleviate the composition artifact.
- Poisson blending. We use Poisson blending [60] to make a composition that looks seamless and natural. Note that the color of the object may be adjusted to make it harmonious with the background.

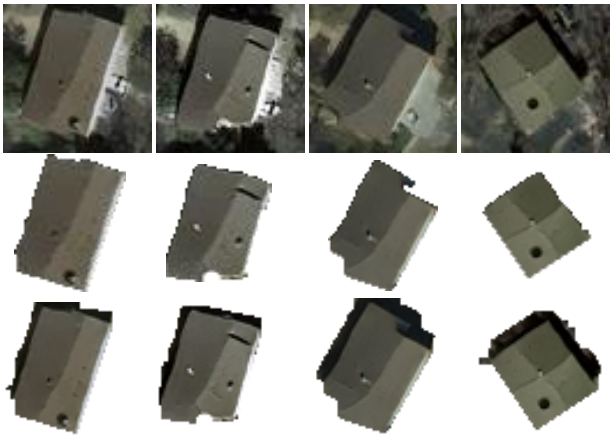

Fig. 8. Examples of shadow extraction results. (Top) building images. (Middle) building targets. (Bottom) building targets with shadows.

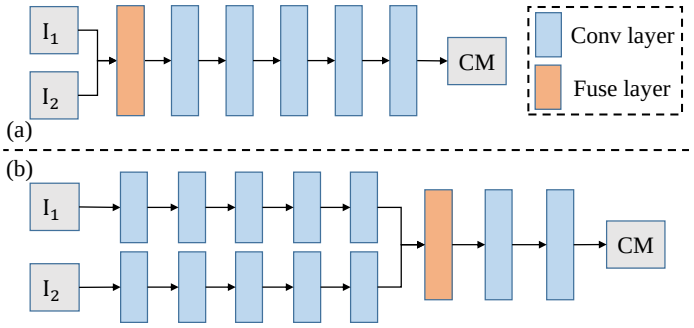

Fig. 9. Illustration of different neural networks architectures for merging the bi-temporal information. (a) Early fusion. (b) Late fusion.

IV. Change Detection Neural Network

In this section, we elaborate on the proposed model for building change detection. First, we will introduce the overall architecture of the model, and then explain its detailed implementation.

A. Model

Given two registered images captured at different times, our goal is to obtain a pixel-level binary mask where each location indicates a change category (building change or not). To extract the change information, we employ Deep Fully Convolutional Networks (DFCN), which could learn complex image representations with multiple levels of abstraction. Building a DFCN that can predict a pixel-level change mask for a given image pair is straightforward: the network must feed the two input images through several convolutional layers, and generates an output map where each location assigns a possibility to each of the change categories.

Note that the change detection network must handle two input image patches. It is important to discuss when and how to merge the information of the two patches. We split this issue into two parts: "when to merge" and "how to merge". "When to merge" cares about in which stage to merge the information of the two input patches. For this part, we introduce the early-fusion mode and the late-fusion mode, as shown in Fig. 9. In the early-fusion mode, the two input patches are merged before fed into DCNN. While in the late-fusion mode, the two patches are first fed to DCNN to generate high-level image representations and then merged in the feature space. "how to merge" the information of the two patches is critical for the consequent change decision process. Here, we give two operations: Concat and Sub. The Concat operation concatenates the features (or images) of the two patches, which preserves all the information of the two patches. While the Sub operation calculates the element-wise absolute distance between the features (or images) of the two patches.

In our work, we consider the combination of the late-fusion mode and the Sub operation. The binary CD task has inherently 'symmetric characteristics', which means that the change detection result remains the same if we shuffle the order of the input two images. To facilitate network learning, instead of forcing the networks to learn the 'symmetric characteristics', we directly design a symmetric CD network that naturally has this property. Our symmetric change detection structure is invariant to the order change of the two input patches. To simplify the explanation, we note the Concat operation as $[x, y]$, and the Sub operation as $|x - y|$, where x, y are two different scalars. Then we have $[x, y] \neq [y, x]$, $|x-y| = |y-x|$. Therefore, we prefer the Sub operation to the Concat operation. Directly performing Sub operation in the raw-image space may induce many false alarms and lose much significant information, due to the radiometric difference between the bi-temporal images caused by variations in imaging conditions (.e.g., illumination). Therefore, we adopt the late-fusion mode instead of the early-fusion mode.

The structure of our CD networks (CDNet) is illustrated in Fig. 2 (bottom). We have a relatively complex pixel-level image feature extractor (deep FCN) that happens in parallel for both patches, and a distance metric to calculate the Feature Difference Images (FDI) between the two patches, followed by a relatively simple classifier (shadow FCN) that receives the FDI as input to give the change probability maps. We wish to learn a semantic feature embedding for each image pixel, such that semantically similar pixels are close to each other, and semantically different ones are far apart in the embedding space. In this way, the distance between the bi-temporal pixels in the embedding space indicates semantic change information. Then, the change information could be easily detected from the FDI by a simple classifier.

Let $I = (I_1, I_2)$ be a bi-temporal image pair, $f : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{H\times W\times C}$ be the feature extractor and $g : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{H\times W\times 2}$ be the classifier, where H, W are the image height and width respectively, and C is the channel dimension of the image embedding. Given a bi-temporal image pair I, our change detection networks generate the predicted change probability maps $P \in \mathbb{R}^{H\times W\times 2}$, which is given by

$$P = \sigma(g(D)) = \sigma(g(|f(I_1) - f(I_2)|)), \tag{4}$$

where FDI $D \in \mathbb{R}^{H\times W\times C}$ is the element-wise absolute distance between the two feature maps and $\sigma(\cdot)$ denotes a softmax function pixel-wisely operated on the channel dimension of the score maps from the output of the classifier.

In the inference phase, the prediction mask $M \in \mathbb{R}^{H \times W}$ is computed by a pixel-wise Argmax operation on the channel dimension of P.

In the training phase, let B represents the minibatch size, where b indexes the bth minibatch. Given input samples $\{(I^b, Y^b)|b \in \{1, 2, ..., B\}\}$ in a minibatch, change detection networks generate the change probability maps $\{P^b\}$. We denote $P^b_{hw} = [P^b_{hw0}, P^b_{hw1}]$ as a length-2 vector for the pixel located at (h, w) of P^b, where P^b_{hw0} and P^b_{hw1} is the probability of no-change and change, respectively. Ground truth $Y^b \in \{0, 1\}^{H \times W}$ provide the change category for each location of the bth sample, where 0 and 1 indicate no-change and change respectively. The loss function of the change detection networks is defined as follows:

$$L = \frac{1}{B \times H \times W} \sum_{b=1}^{B} \sum_{h=1,w=1}^{H,W} l(P^b_{hw}, Y^b_{hw}), \tag{5}$$

where $l(P^b_{hw}, y) = -log(P^b_{hwy})$ is the cross-entropy loss, and Y^b_{hw} is the label for the pixel at location (h, w) in the bth batch.

B. Implement Details

Feature Extractor: Note that we need to obtain a change map with the same size as the input images. To achieve this, it is essential to generate high-resolution semantic feature maps. However, the high-level features in DFCN are accurate in semantics but coarse in location, while the low-level features contain fine details but lack semantic information. Therefore, we fuse the low-level, fine appearance information, and the high-level, coarse semantic information to balance the inherent tension between semantics and location. Here, we employ a UNet [59] structure to extract pixel-level image representations. As illustrated in Fig. 10, the networks have an encoder-decoder structure with the same amount of downsampling and upsampling. The encoder follows the ResNet-18 [61] structure that has 5 stages each with downsampling of a stride of 2. The decoder also has 5 stages each with upsampling by a factor of 2. Each of the first 4 stages in the decoder consists of upsampling the high-level feature maps, followed by concatenation with the corresponding low-level feature maps of the same size from the encoder, and two 3×3 convolutions, each followed by a ReLU and a BatchNorm. The configuration of the convolutional layers [1] in Conv Block1, Decoder Block1-4, Conv Block2 are [(512, 3×3, 1) $\times$ 2], [(256, 3×3, 1) $\times$ 2], [(128, 3×3, 1) $\times$ 2], [(64, 3×3, 1) $\times$ 2], [(32, 3×3, 1) $\times$ 2], [(16, 3×3, 1) $\times$ 2], respectively. In this way, we fuse the upsampled high-level features with the low-level features to obtain finer semantic feature representations.

Classifier: Benefiting from the high-resolution and high-level semantic change features extracted by the deep feature extractor, a very shallow FCN can be employed here for change discrimination. The classifier consists of two 3×3 convolutional layers ([(16, 3×3, 1) $\times$ 2]) and a 1×1 convolutional layer ([(2, 1×1, 1) $\times$ 1]).

[1]The configuration of convolution layers is "[(Number of the Filters, Size, Stride) $\times$ number of convolution layers]". The batch normalization (BN) and ReLU layers are omitted for simplicity.

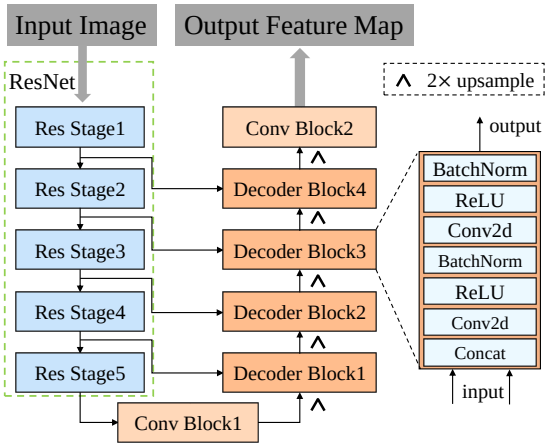

Fig. 10. Structure of the feature extractor.

V. Experimental Results and Analysis

A. Change Deteciton Datasets

To evaluate the effectiveness of our method, we employ two VHR RS image building CD datasets: LEVIR-CD[6] and WHU Building CD dataset[14].

LEVIR-CD. LEVIR-CD is a public large scale building CD dataset, which contains 637 pairs of bi-temporal images, each size of 1024×1024. These images have a 0.5 m spatial resolution. The dataset has over 31K changed building instances. We follow the default dataset split [6]: 445/64/128 for training/validation/testing. Considering GPU memory capacity limitation, we cut images into small patches of size 256×256 with no overlap. Therefore, we obtain 7120/1024/2048 pairs of patches for training/validation/testing respectively.

WHU-CD. WHU Building CD dataset includes one pair of optical RS images with a size of 32507×15354 and 0.075 m spatial resolution. Similar to LEVIR-CD, we crop images into small patches of size 256×256 with no overlap. In this way, we collect 7620 pairs of patches. As the data provider has not given a dataset split suggestion, here we randomly split the dataset into three parts: 6096/762/762 pairs for training/validation/testing respectively.

B. Synthesis Details

1) Synthesizing Building Objects: To train our building generator, we collect building samples from two public RS building labeling datasets: Inria building dataset[51] and AIRS (Aerial Imagery for Roof Segmentation)[62]. The Inria dataset consists of 1800 pairs of bi-temporal aerial RGB images, each size of 5000×5000 and spatial resolution of 0.3 m. Inria contains more than 210K building instances. The AIRS dataset contains aerial images covering the area of Christchurch city in New Zealand (at 7.5 cm resolution, with RGB bands). AIRS includes more than 230K building instances.

We collect 27041/69694 training samples from these two datasets respectively. Each sample including a semantic map

TABLE I
SUMMARY OF ALL THE TRAINING DATASETS. N DENOTES THE NUMBER OF BUILDING INSTANCES BLENDED ON EACH SAMPLE. N=0 REPRESENTS THE ORIGINAL TRAINING SET.

Training data	N	Imbalanced Ratio
LEVIR-CD	0	21.09
LEVIR_IR15	15	15.42
LEVIR_IR10	50	9.80
LEVIR_IR5	120	5.36
WHU-CD	0	22.26
WHU_IR15	1920	14.67
WHU_IR10	4562	9.85
WHU_IR5	10434	5.36

and a corresponding image (size of 256×256), is cropped from the provided image/label map in the existing building dataset. Each sample has a building instance at its center and contains at least one complete building instance. We train our building generators on these two training sets respectively. To match the resolution between the CD dataset and the building dataset, the building generator trained on the Inria dataset is for the LEVIR-CD dataset and the one trained on the ARIS dataset is for the WHU-CD dataset.

In the inference phase, the building generator produces building images conditioned on the input semantic label maps, which are cropped from the label maps of the original CD dataset. Each semantic mask is centered on a building object. We use masks size of $16 \times 16 \sim 64 \times 64$ to generate building images for LEVIR-CD, and use masks size of $64 \times 64 \sim 256 \times 256$ for WHU-CD. To this end, the generated images have similar geometric characteristics as those in the original building CD dataset.

Then, for each GAN-generated image, we randomly choose a target building image from the CD dataset and perform CT to transfer its color information onto the source image.

2) Synthsizing CD Training Samples: For the LEVIR-CD dataset, N_1 instances are blended onto either temporal images (size of 1024×1024) of each sample from the training set. For the WHU-CD dataset, as all the data is one pair of large-size images, we first blend N_2 instances onto these images (size of 32507×15354) to obtain an augmented image pair, and then we cut it into pairs of small patches (size of 256×256) and select the corresponding 6096 pairs as the augmented training set. Note that we only augment the training set of the CD dataset without changing the validation set and the testing set.

Based on the LEVIR-CD and WHU-CD datasets, we construct several synthesized training sets with different imbalance ratios [63] by using different numbers of augmented instances on each CD sample. Here, the imbalance ratio is the proportion of the number of pixels belong to the no-change class to the number of the change class. The summary of the original training sets and corresponding synthesized training sets are listed in Table I. Some selected samples from these datasets are shown in Fig. 11. More details on the discussion of the hyperparameter N are given in Section V-G.

C. Experimental Setup

We make a comparison of our method with several state-of-the-art CD methods:

- FC-EF [33]: Image-level fusion method, where the bi-temporal images are concatenated as a single input to a fully convolutional network.
- FC-Siam-Di [33]: Feature-level fusion method, where a Siamese FCN is employed to extract multi-level features and feature difference is used for fusion of the bi-temporal information.
- FC-Siam-Conc [33]: Feature-level fusion method, where a Siamese FCN is employed to extract multi-level features and feature concatenation is used to fuse the bi-temporal information.
- DTCDSCN [8]: Multi-scale feature concatenation method, where a deep siamese FCN is trained using a weighted focal loss and two additional semantic segmentation decoders are trained under the supervision of the label maps of each temporal. We omit the semantic segmentation decoders for a fair comparison.
- STANet [6]: Metric-based siamese FCN based method, which integrates the spatial-temporal attention mechanism to obtain more discriminative features.

We implement the above CD networks using their public codes with default hyperparameters.

To further verify the effectiveness of the proposed IAug, we make a comparison with some popular cost-sensitive techniques for addressing the class imbalance on the imbalanced CD dataset. For a fair comparison, all these comparison methods are based on the same baseline CD network.

- CDNet: our baseline CD network is trained on the original CD training set using conventional cross-entropy loss.
- CDNet (W): Using a weighted cross-entropy loss to train the CDNet on the original CD training set. The weight assigned to each incorrect example is inversely proportional to the number of representatives of that class.
- CDNet (F): Using the focal loss [64] to train the CDNet on the original CD training set.
- CDNet (WF): Using a weighted focal loss [8] to train the CDNet on the original CD training set.
- CDNet (D): Using a dice loss [17] to train the CDNet on the original CD training set.
- CDNet+IAug: Training the CDNet on our synthesized CD training set using conventional cross-entropy loss.

Our models are implemented on a PyTorch deep learning framework [65] and trained using a single NVIDIA Tesla V100 GPU. In the training phase, the inputs of the CDNet are images of 256×256 pixels with data augmentation, including flip, rescale, crop, and gaussian blur. The stochastic gradient descent (SGD) with momentum is applied for training. The initial learning rate is set to 0.01, the momentum and the weight decay is set to 0.99 and 0.0005, respectively. "Poly" learning rate policy [66] is used to polynomially decay the learning rate during iteration. The decay coefficient is set to 0.9. After completing training 100 epochs, the learning rate drops to zero. The batch size is set to 8. After each training epoch, the validation data is used to evaluate the performance

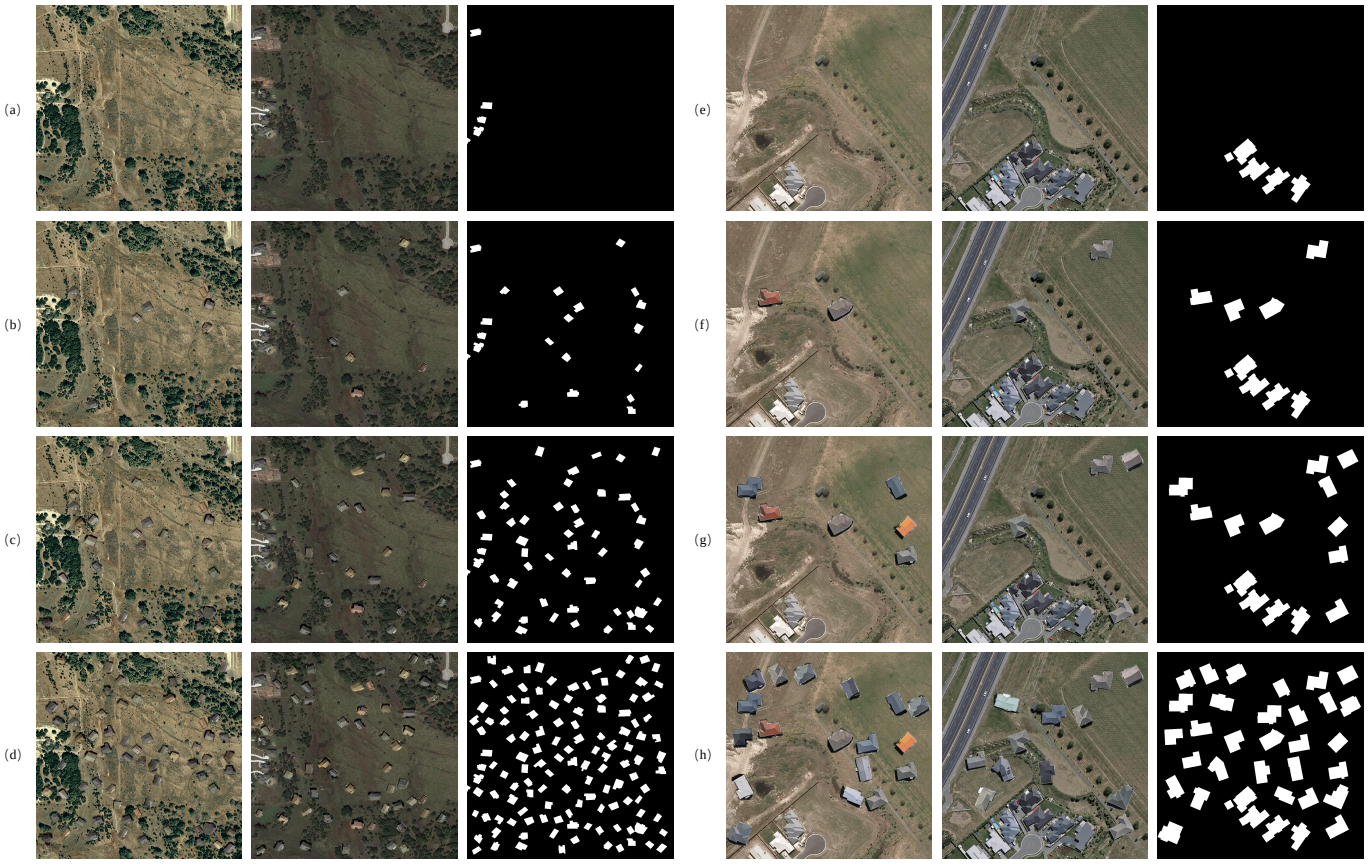

Fig. 11. Illustration of the synthesized samples from several training sets of different imbalanced ratios. (a) LEVIR-CD, (b) LEVIR_IR15, (c) LEVIR_IR10, (d) LEVIR_IR5, (e) WHU-CD, (f) WHU_IR15, (g) WHU_IR10, (h) WHU_IR5.

of the model. The best model on the validation set is saved as the final training result.

D. Evaluation Metrics

To evaluate the proposed approach quantitatively, we use the F1-score with regard to the change category as the evaluation indices. F1-score is calculated by the precision and recall of the test. Let TP, FP, FN represent the number of true positive, false positive, and false negative respectively. F1-score is computed by the following formula:

$$F1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}}, \tag{6}$$

where the precision and recall are defined as follows:

$$\begin{aligned} \text{precision} &= \text{TP} / (\text{TP} + \text{FP}) \\ \text{recall} &= \text{TP} / (\text{TP+FN}). \end{aligned} \tag{7}$$

E. Overall Comparison

1) Comparison with other SOTA CD methods.

To verify the effectiveness of the proposed method (CDNet + IAug), the compared CD networks are trained on the original CD datasets and the proposed CDNet is trained on our synthesized CD datasets. Here, we use LEVIR_IR5 and WHU_IR5 as our synthesized CD training datasets. To further compare the performance of each method under conditions of different data amounts, we set a variety of data conditions: 5%, 20%, and 100%. These percentages represent the proportion of training data used in each data regime.

The overall quantitative results of all the compared methods on the two test sets are listed in Table II. The results show that the proposed method outperforms other CD approaches with respect to the F1-score in every data regimes on the two datasets. It is worthwhile to mention that our method using only 20% of the training data could achieve comparable results as the state-of-the-art methods using 100% of the training data.

Table II also shows that our CDNet (w.o. IAug) delivers comparable or even better performance than other methods on the two datasets. It may attribute to the effectiveness of our designed network structure (e.g., symmetric structure, ResNet backbone, high-resolution and high-level feature differencing, and shallow change classification networks). Additionally, we can observe that our IAug introduces more significant improvements to the F1 score against CDNet under a small data regime (5% and 20%) than under a big data regime (100%). It indicates that our IAug can effectively alleviate the small labeled data problem.

To fairly compare the model efficiency, we test all the methods on a computing server equipped with an Intel Xeon Silver 4214 CPU and an NVIDIA Tesla V100 GPU. Table III lists the number of model parameters (Params.), floating-point operations per second (FLOPs), and GPU inference time of

the compared methods. The input to the model has a size of $256 \times 256 \times 3$. The reported time is the average of the inference time of the model for 100 random inputs. The results show that the proposed method outperforms the recent DTCDSCN and STANet with smaller model parameters and less computational cost.

2) Comparison with other class-sensitive algorithms

The proposed IAug can be viewed as a data-level solution to amuse the class imbalance that occurs when conventional learning on an imbalanced CD dataset. To verify the effectiveness of the proposed IAug, we make a comparison with four class-sensitive algorithms for solving the class imbalance on the CD task. These methods employ different types of weighted losses so that the learner can pay more attention to the examples of the change class in the training phase. Different from these algorithmic solutions, the proposed IAug belongs to a data solution. We have not compared with other data solutions, such as a naive over-sampling method, because it is not easy to implement in a pixel-level prediction task, and also it is somehow equivalent to using a weighted loss.

The comparison results are shown in Table IV. It can be seen that compared with other methods, the proposed approach has achieved the best performance in the two building CD datasets. Compared to the baseline, our IAug could consistently improve the CD performance on the two datasets. It indicates that our data-level solution is more effective in reducing class imbalance and could introduce more stable improvement to the performance of the CD model compared with other algorithmic solutions.

It is not a surprising result because our IAug can not only alleviate class imbalance but also introduces additional instance-level supervision information to the CD model. We cannot emphasize the importance of data too much. As Goodfellow et al. have stated that: "it is often much better to gather more data than to improve the learning algorithm" [67].

3) Generalizability of IAug

To further verify the effectiveness of the proposed IAug, we train several state-of-the-art CD networks on the original and synthetic datasets respectively. Table V reports the results of these methods on the LEVIR-CD and WHU-CD test sets. Quantitative results have shown that our proposed IAug can consistently improve the performance of these CD networks on the two test sets.

F. Ablation Studies

Here, we conduct ablation experiments of the four components of our IAug: semantic object generator (GAN), object color transfer (CT), context-aware blending (CB), multiple composition modes (MCM). The baseline is trained on the original CD dataset. We incrementally add the above four components to synthesize corresponding training sets to evaluate their respective gains to the performance of the model. All these experiments are performed on the LEVIR-CD dataset using CDNet. We also set three data regimes: 5%, 20%, and 100% to evaluate the performance of CDNet trained using different proportions of the training set. Note that we blend no more than 120 GAN-generated building targets on each sample from the LEVIR-CD dataset to synthesize new training sets.

As shown in Table VI, quantitative results illustrate that the four components of IAug bring considerable performance improvements across the different data settings. In a small data regime, the improvement inducing by each component is much significant. When using 5% of the training data, the contributions of these four components to the model performance increment are 6%, 3%, 4%, and 1%, respectively. GAN improves the performance the most because it largely increases the number of effective buildings of change so as to improve the model discrimination ability to the rare class. The ablation on CT indicates that blending building instance with a similar appearance as the target domain can achieve better performance in the target dataset. The experimental results with regard to CB imply that the context surrounding the building may be a critical cue for object recognition. MCM further boosts the model performance because the diverse samples via composition modes can improve the generalization ability of the CD model and avoid the network from overfitting a single composition mode.

G. Class Imbalance Analysis

The real-world RS building CD task exhibits imbalanced class distributions. As shown in Table I, the imbalanced ratios of the original LEVIR-CD and WHU-CD datasets are both higher than 20. In this work, we have constructed several synthesized training sets with different imbalance ratios (near 15, 10, and 5 respectively) by superimposing different numbers of building instances on each CD sample. Here, we analyze the impact of the number (i.e., N) of augmented instances on the performance of the change detection model. The same CDNet is applied to each synthesized training set to evaluate the CD performance. The trained models are evaluated on the corresponding LEVIR-CD and WHU-CD test sets, respectively.

The performance of the model trained on different training sets is shown in Table VII. The results on the training sets with different imbalanced ratios indicate that the smaller the imbalanced ratio of the training data set, the better the model performance. Fig. 12 depicts the change of the F1-score associated with the number (N) of building instances blended on each CD sample. It can be seen that the performance of the model improves when N increases. We also observe that the marginal benefit on the model performance is declining with the number of instances increasing. For example, there is only a 0.2% improvement in model performance when doubling the instance amount to 120 (from LEVIR_IR10 to LEVIR_IR5). Moreover, there is an upper limit for N when the image does not have more space to superimpose more building instances. Therefore, we set the maximum number N=120 for LEVIR-CD and N=10434 for WHU-CD. In this case, the imbalance ratio of the corresponding synthetic dataset is close to 5.

VI. Conclusion

In this paper, instance-level change augmentation is proposed to efficiently synthesize effective building CD samples by leveraging generative adversarial training and image

TABLE II
PRECISION (PREC), RECALL (REC), AND F1 OF DIFFERENT METHODS ON LEVIR-CD AND WHU-CD TEST SETS. THE HIGHEST CLASSIFICATION ACCURACY IN EACH DATA REGIME IS MARKED IN BOLD.

	LEVIR-CD			**WHU-CD**		
	5%	20%	100%	5%	20%	100%
	Prec / Rec / F1	Prec / Rec / F1	Prec / Rec / F1	Prec / Rec / F1	Prec / Rec / F1	Prec / Rec / F1
FC-EF [33]	0.785 / 0.417 / 0.545	0.764 / 0.739 / 0.751	0.869 / 0.802 / 0.834	0.269 / 0 / 0.001	0.714 / 0.604 / 0.654	0.716 / 0.673 / 0.694
FC-Siam-conc [33]	0.676 / 0.021 / 0.041	0.815 / 0.797 / 0.806	0.895 / 0.833 / 0.863	0.499 / 0.639 / 0.56	0.417 / 0.676 / 0.516	0.473 / 0.777 / 0.588
FC-Siam-diff [33]	**0.916** / 0.2 / 0.329	0.886 / 0.7 / 0.782	**0.92** / 0.768 / 0.837	0.488 / 0.496 / 0.492	0.487 / 0.462 / 0.548	0.609 / 0.736 / 0.666
DTCDSCN [8]	0.829 / 0.653 / 0.73	0.833 / 0.851 / 0.842	0.885 / 0.868 / 0.877	0.463 / 0.559 / 0.507	0.604 / 0.707 / 0.652	0.639 / 0.823 / 0.72
STANet [6]	0.755 / **0.726** / 0.74	0.802 / **0.864** / 0.832	0.838 / **0.91** / 0.873	0.709 / 0.672 / 0.69	0.764 / 0.756 / 0.76	0.794 / 0.855 / 0.823
Ours (CDNet)	0.89 / 0.525 / 0.661	**0.917** / 0.741 / 0.82	0.905 / 0.846 / 0.875	0.71 / 0.663 / 0.686	0.829 / 0.76 / 0.793	0.898 / 0.833 / 0.864
Ours (CDNet+IAug)	0.804 / 0.721 / **0.76**	0.901 / 0.851 / **0.875**	0.916 / 0.865 / **0.89**	**0.777** / **0.695** / **0.734**	**0.868** / **0.781** / **0.822**	**0.914** / **0.869** / **0.891**

TABLE III
COMPARISON ON MODEL EFFICIENCY. WE REPORT THE NUMBER OF MODEL PARAMETERS (PARAMS.), FLOATING-POINT OPERATIONS PER SECOND (FLOPS) AND GPU INFERENCE TIME. THE INPUT IMAGE TO THE MODEL HAS A SIZE OF $256 \times 256 \times 3$.

Model	Params.(M)	FLOPs (G)	Time (ms)
FC-EF [33]	1.35	1.78	7.17
FC-Siam-conc [33]	1.55	2.66	9.61
FC-Siam-diff [33]	1.35	2.36	10.10
DTCDSCN [8]	40.70	7.21	25.65
STANet [6]	16.93	6.58	23.15
Ours	14.33	5.50	13.81

TABLE IV
QUANTITATIVE RESULTS OF OUR METHOD AND SEVERAL CLASS-SENSITIVE ALGORITHMS ON LEVIR-CD AND WHU-CD TEST SETS. THE HIGHEST CLASSIFICATION ACCURACY IS MARKED IN BOLD.

	LEVIR-CD	**WHU-CD**
	Precision / Recall / F1	Precision / Recall / F1
CDNet	0.905 / 0.846 / 0.875	0.898 / 0.833 / 0.864
CDNet (W)	0.886 / **0.873** / 0.879	0.872 / 0.853 / 0.862
CDNet (F)	**0.921** / 0.799 / 0.856	0.907 / 0.801 / 0.851
CDNet (WF)	0.898 / 0.857 / 0.877	0.888 / 0.846 / 0.866
CDNet (D)	0.908 / 0.855 / 0.881	0.818 / 0.83 / 0.824
CDNet+ IAug	0.916 / 0.865 / **0.890**	**0.914** / **0.869** / **0.891**

TABLE V
RESULTS OF SEVERAL CD NETWORKS ON THE LEVIR-CD AND WHU-CD TEST SETS. "+ IAUG" DENOTES THE CD NETWORK IS TRAINED ON THE CORRESPONDING SYNTHESIZED TRAINING SET (LEVIR_IR5 OR WHU_IR5), OTHERWISE THE CD NETWORK IS TRAINED ON THE ORIGINAL TRAINING SET.

	LEVIR-CD	**WHU-CD**
	Precision / Recall / F1	Precision / Recall / F1
FC-EF [33]	0.869 / 0.802 / 0.834	0.716 / 0.673 / 0.694
+ IAug	0.892 / 0.855 / 0.873	0.834 / 0.714 / 0.769
FC-Siam-conc [33]	0.895 / 0.833 / 0.863	0.473 / 0.777 / 0.588
+ IAug	0.913 / 0.858 / 0.885	0.841 / 0.777 / 0.807
FC-Siam-diff [33]	0.92 / 0.768 / 0.837	0.609 / 0.736 / 0.666
+ IAug	**0.922** / 0.849 / 0.884	0.85 / 0.796 / 0.822
DTCDSCN [8]	0.885 / 0.868 / 0.877	0.639 / 0.823 / 0.72
+ IAug	0.896 / 0.882 / 0.889	0.819 / 0.865 / 0.842
STANet [6]	0.838 / 0.91 / 0.873	0.794 / 0.855 / 0.823
+ IAug	0.845 / **0.919** / 0.880	0.903 / 0.863 / 0.882
Ours (CDNet)	0.905 / 0.846 / 0.875	0.898 / 0.833 / 0.864
+IAug	0.916 / 0.865 / **0.890**	**0.914** / **0.869** / **0.891**

TABLE VI
ABLATION STUDIES OF OUR IAUG ON LEVIR-CD TEST SET. ABLATIONS ARE PERFORMED ON 1) BUILDING GENERATOR (GAN), 2) COLOR TRANSFER (CT), 3) CONTEXT-AWARE BLENDING (CB) AND 4) MULTIPLE COMPOSITION MODES (MCM).

	GAN	CT	CB	MCM	5%	20%	100%
baseline	×	×	×	×	0.661	0.820	0.875
IAug	✓	×	×	×	0.703	0.858	0.88
IAug	✓	✓	×	×	0.723	0.86	0.881
IAug	✓	✓	✓	×	0.750	0.862	0.884
IAug	✓	✓	✓	✓	**0.760**	**0.875**	**0.890**

TABLE VII
RESULTS OF THE PERFORMANCE OF THE MODEL TRAINED ON DIFFERENT DATASETS.

Training data	Precision	Recall	F1
LEVIR-CD	0.905	0.846	0.875
LEVIR_IR15	0.909	0.863	0.885
LEVIR_IR10	0.913	0.864	0.888
LEVIR_IR5	**0.916**	**0.865**	**0.89**
WHU-CD	0.898	0.833	0.864
WHU_IR15	0.912	0.852	0.881
WHU_IR10	0.911	0.862	0.886
WHU_IR5	**0.914**	**0.869**	**0.891**

blending. We introduce a GAN-based approach to generate realistic building images with controllable shape and appearance according to the specified input semantic label map and reference style. In this way, we can obtain various building images that are well aligned with the semantic layouts. Furthermore, we propose context-aware blending to synthesize

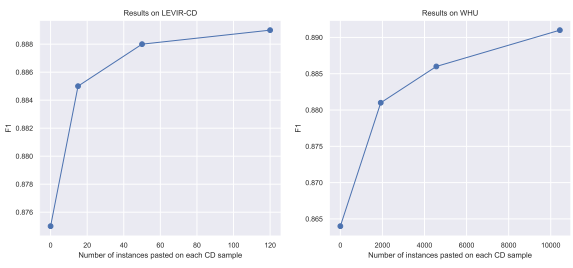

Fig. 12. The effect of the number of instances on model performance. The F1-score of each model is reported.

realistic CD samples. We also design a simple yet effective CD neural network (CDNet), which constructs high-resolution and high-level feature difference images for change analysis. Our method (CDNet + IAug) outperforms several state-of-the-art CD methods on two building CD datasets (LEVIR-CD and WHU-CD). Notably, we achieve comparable results with only 20% of the training data as the current SOTA methods using 100% data. Extensive experiments have validated the effectiveness of our proposed method. Our synthesized dataset can also reduce the risk of class imbalance. Conventional learning on the synthesized dataset outperforms several popular cost-sensitive algorithms on the original dataset.

Our paper may have some limitations. First, only the RGB channels of RS images are used in this work due to the constraint of the existing building CD datasets. Our method can be extended to multispectral images (more than RGB bands) upon the availability of the public building CD dataset that contains large-scale high-resolution bitemporal multispectral images. Second, the generated building image may not be faithful as that in the target domain. Due to the lack of explicit shadow masks (or the sun direction) to train the building generator, we can not control the shape and orientation of the generated shadow. Therefore the shadow that is implicitly learned by the generator may not be cast in the same direction as that in the target image. Moreover, there may remain style discrepancy between the generated building and the target one because only a simple color transfer approach is employed. Third, our rule-based shadow extraction method may falsely detect the dark land covers (e.g., dense vegetation) that are very close to and surround the building on multiple sides.

It deserves note that although this paper focuses on the building CD task, the proposed IAug can be extended to other semantic target change detection (e.g., roads, vegetation, rivers). The future work includes the modification of IAug for an extension to other semantic target CD, and an online instance change augmentation method, which is storage-friendly and can greatly increase the diversity of augmented instances and further enhances the generalization of the model. Also, more sophisticated techniques of domain adaptation (how to generate an image more suitable for the target domain) and image blending (how to make a more realistic image composition) can be explored in the future.

REFERENCES

[1] A. SINGH, "Review article digital change detection techniques using remotely-sensed data," *International Journal of Remote Sensing*, vol. 10, no. 6, pp. 989–1003, 1989. [Online]. Available: https://doi.org/10.1080/01431168908903939

[2] X. Huang, L. Zhang, and T. Zhu, "Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 7, no. 1, pp. 105–115, 2014.

[3] X. Huang, Y. Cao, and J. Li, "An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images," *Remote Sensing of Environment*, vol. 244, p. 111802, 2020.

[4] J. Z. Xu, W. Lu, Z. Li, P. Khaitan, and V. Zaytseva, "Building damage detection in satellite imagery using convolutional neural networks," 2019.

[5] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "Pga-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection," *Remote Sensing*, vol. 12, no. 3, p. 484, 2020.

[6] 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.

[7] S. Ji, Y. Shen, M. Lu, and Y. Zhang, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sensing*, vol. 11, no. 11, p. 1343, 2019.

[8] 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*, pp. 1–5, 2020.

[9] M. Bouziani, K. Goïta, and D.-C. He, "Automatic change detection of buildings in urban environment from very high spatial resolution images using existing geodatabase and prior knowledge," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, pp. 143–153, 2010.

[10] J. Li, X. Huang, and J. Gong, "Deep neural network for remote-sensing image interpretation: status and perspectives," *National Science Review*, vol. 6, no. 6, pp. 1082–1086, may 2019.

[11] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sensing*, vol. 12, p. 1688, 2020.

[12] H. Chen, T. Shi, Z. Xia, D. Liu, X. Wu, and Z. Shi, "Learning to segment objects of various sizes in vhr aerial images," vol. 875, Beijing, China, 2018, pp. 330 – 340. [Online]. Available: http://dx.doi.org/10.1007/978-981-13-1702-6_33

[13] S. Lei, Z. Shi, and Z. Zou, "Coupled adversarial training for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3633–3643, 2019.

[14] 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. Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2019.

[15] A. Song and J. Choi, "Fully convolutional networks with multiscale 3d filters and transfer learning for change detection in high spatial resolution satellite images," *Remote Sensing*, vol. 12, no. 5, p. 799, 2020.

[16] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data," in *2019 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2019, Yokohama, Japan, July 28 - August 2, 2019*. IEEE, 2019, pp. 214–217.

[17] Z. Cao, M. Wu, R. Yan, F. Zhang, and X. Wan, "Detection of small changed regions in remote sensing imagery using convolutional neural network," vol. 502. IOP Publishing, jun 2020, p. 012017. [Online]. Available: https://doi.org/10.1088%2F1755-1315%2F502%2F1%2F012017

[18] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, pp. 1845–1849, 2017.

[19] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 11–28, jul 2016.

[20] P. Xiao, X. Zhang, D. Wang, M. Yuan, X. Feng, and M. Kelly, "Change detection of built-up land: A framework of combining pixel-based detection and object-based recognition," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 119, pp. 402–414, sep 2016.

[21] A. Lefebvre and T. Corpetti, "Monitoring the morphological transformation of beijing old city using remote sensing texture analysis," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 10, no. 2, pp. 539–548, 2017.

[22] N. Falco, M. D. Mura, F. Bovolo, J. A. Benediktsson, and L. Bruzzone, "Change detection in vhr images based on morphological attribute profiles," vol. 10, pp. 636–640, 2013.

[23] D. Wen, X. Huang, L. Zhang, and J. A. Benediktsson, "A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation," *IEEE Trans. Geosci. Remote. Sens.*, vol. 54, no. 1, pp. 609–625, jan 2016.

[24] Y. Tang, X. Huang, and L. Zhang, "Fault-tolerant building change detection from urban high-resolution remote sensing imagery," *IEEE Geosci. Remote. Sens. Lett.*, vol. 10, no. 5, pp. 1060–1064, sep 2013.

[25] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, pp. 161–172, 2012.

[26] Y. Sun, X. Zhang, J. Huang, H. Wang, and Q. Xin, "Fine-grained building change detection from very high-spatial-resolution remote sensing images based on deep multitask learning," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[27] K. Nemoto, R. Hamaguchi, M. Sato, A. Fujita, T. Imaizumi, and S. Hikosaka, "Building change detection via a combination of CNNs using only RGB aerial imageries," in *Remote Sensing Technologies and Applications in Urban Environments II*, T. Erbertseder, N. Chrysoulakis, Y. Zhang, and W. Heldens, Eds., vol. 10431, International Society for Optics and Photonics. SPIE, 2017, pp. 107 – 118. [Online]. Available: https://doi.org/10.1117/12.2277912

[28] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *TGRS*, pp. 1–15, 2020.

[29] R. Liu, M. Kuffer, and C. Persello, "The temporal dynamics of slums employing a cnn-based change detection approach," *Remote. Sens.*, vol. 11, no. 23, p. 2844, 2019.

[30] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IGARSS*, 2018. [Online]. Available: https://rcdaudt.github.io/publication/2018-08-22-urban-change-detection

[31] F. U. Rahman, B. Vasu, J. V. Cor, J. Kerekes, and A. E. Savakis, "Siamese network with multi-level features for patch-based change detection in satellite imagery," in *2018 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2018, Anaheim, CA, USA, November 26-29, 2018*. IEEE, 2018, pp. 958–962.

[32] M. Wang, K. Tan, X. Jia, X. Wang, and Y. Chen, "A deep siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images," *Remote Sensing*, vol. 12, no. 2, p. 205, 2020.

[33] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *ICIP*, 2018. [Online]. Available: https://github.com/rcdaudt/fully_convolutional_change_detection

[34] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved unet++," *Remote Sensing*, vol. 11, no. 11, p. 1382, 2019.

[35] P. P. de Bem, O. A. de Carvalho Junior, R. F. Guimarães, and R. A. T. Gomes, "Change detection of deforestation in the brazilian amazon using landsat data and convolutional neural networks," *Remote Sensing*, vol. 12, no. 6, p. 901, 2020.

[36] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS*, vol. 166, pp. 183–200, 2020.

[37] T. Bao, C. Fu, T. Fang, and H. Huo, "Ppcnet: A combined patch-level and pixel-level end-to-end deep network for high-resolution remote sensing image change detection," vol. PP, pp. 1–5, 2020.

[38] B. Hou, Q. Liu, H. Wang, and Y. Wang, "From w-net to cdgan: Bitemporal change detection via deep learning techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 1790–1802, 2020.

[39] B. Fang, L. Pan, and R. Kou, "Dual learning-based siamese framework for change detection using bi-temporal vhr optical remote sensing images," *Remote Sensing*, vol. 11, no. 11, p. 1292, 2019.

[40] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote. Sens. Lett.*, vol. 16, no. 2, pp. 266–270, 2019.

[41] 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.

[42] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," vol. XLII-2, 2018, pp. 565–571.

[43] W. Zhao, L. Mou, J. Chen, Y. Bo, and W. J. Emery, "Incorporating metric learning and adversarial network for seasonal invariant change detection," *IEEE Trans. Geosci. Remote. Sens.*, vol. 58, no. 4, pp. 2720–2731, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8937747

[44] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[45] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "Semicdnet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2020.

[46] F. Jiang, M. Gong, T. Zhan, and X. Fan, "A semisupervised gan-based multiple change detection framework in multi-spectral images," *IEEE Geosci. Remote. Sens. Lett.*, vol. 17, no. 7, pp. 1223–1227, 2020.

[47] J. Li, X. Huang, and X. Chang, "A label-noise robust active learning sample collection method for multi-temporal urban land-cover classification and change analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 163, pp. 1–17, may 2020.

[48] A. A. Goodenough and S. D. Brown, "Dirsig5: Next-generation remote sensing data and image simulation framework," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 11, pp. 4818–4833, nov 2017.

[49] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100–1111, 2018.

[50] R. Wang, J. Zhang, J.-W. Chen, L. Jiao, and M. Wang, "Imbalanced learning-based automatic sar images change detection by morphologically supervised pca-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 4, pp. 554–558, apr 2019.

[51] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *2017 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2017, Fort Worth, TX, USA, July 23-28, 2017*. IEEE, 2017, pp. 3226–3229.

[52] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 8798–8807. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_High-Resolution_Image_Synthesis_CVPR_2018_paper.html

[53] T. Park, M. Liu, T. Wang, and J. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2332–2341.

[54] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4401–4410. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html

[55] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "Sean: Image synthesis with semantic region-adaptive normalization," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5103–5112.

[56] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[57] X. Xiao and L. Ma, "Color transfer in correlated color space," in *Proceedings VRCIA 2006 ACM International Conference on Virtual Reality Continuum and its Applications, Chinese University of Hong Kong, Hong Kong, China, June 14-17, 2006*, H. Sun, Ed. ACM, 2006, pp. 305–309. [Online]. Available: https://github.com/hangong/Xiao06_color_transfer

[58] L. Martino and J. Míguez, "Generalized rejection sampling schemes and applications in signal processing," *Signal Process.*, vol. 90, no. 11, pp. 2981–2995, 2010.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., vol. 9351. Springer, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28

[60] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, 2003.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2016.90

[62] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 147, pp. 42–55, 2019.

[63] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[64] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[65] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library."

[66] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[67] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

Hao Chen received his B.S. degree from the Image Processing Center School of Astronautics, Beihang University in 2017. He is currently pursuing his doctorate degree in the Image Processing Center, School of Astronautics, Beihang University.His research interests include machine learning, deep learning and semantic segmentation.

Wenyuan Li received his B.S. degree from North China Electric Power University, Beijing, China in 2017. He is currently working toward his doctorate degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include deep learning, image processing, and pattern recognition.

Zhenwei Shi (M'13) received his Ph.D. degree in mathematics from Dalian University of Technology, Dalian, China, in 2005. He was a Postdoctoral Researcher in the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was Visiting Scholar in the Department of Electrical Engineering and Computer Science, Northwestern University, U.S.A., from 2013 to 2014. He is currently a professor and the dean of the Image Processing Center, School of Astronautics, Beihang University. His current research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi serves as an Associate Editor for the *Infrared Physics and Technology*. He has authored or co-authored over 100 scientific papers in refereed journals and proceedings, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Neural Networks, the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Geoscience and Remote Sensing Letters and the IEEE Conference on Computer Vision and Pattern Recognition. His personal website is http://levir.buaa.edu.cn/.