Transition Is a Process: Pair-to-Video Change Detection Networks for Very High Resolution Remote Sensing Images

Manhui Lin, Guangyi Yang, and Hongyan Zhang, *Senior Member, IEEE*

***Abstract*— As an important yet challenging task in Earth observation, change detection (CD) is undergoing a technological revolution, given the broadening application of deep learning. Nevertheless, existing deep learning-based CD methods still suffer from two salient issues: 1) incomplete temporal modeling, and 2) space-time coupling. In view of these issues, we propose a more explicit and sophisticated modeling of time and accordingly establish a pair-to-video change detection (P2V-CD) framework. First, a pseudo transition video that carries rich temporal information is constructed from the input image pair, interpreting CD as a problem of video understanding. Then, two decoupled encoders are utilized to spatially and temporally recognize the type of transition, and the encoders are laterally connected for mutual promotion. Furthermore, the deep supervision technique is applied to accelerate the model training. We illustrate experimentally that the P2V-CD method compares favorably to other state-of-the-art CD approaches in terms of both the visual effect and the evaluation metrics, with a moderate model size and relatively lower computational overhead. Extensive feature map visualization experiments demonstrate how our method works beyond making contrasts between bi-temporal images. Source code is available at https://github.com/Bobholamovic/CDLab.**

***Index Terms*— Change detection, deep learning, convolutional neural network, remote sensing, very high resolution image.**

I. Introduction

CHANGE detection (CD) aims at identifying changes occurring between two or more images acquired in the same geographical area at different times [1] and has long been a topic of immense interest in remote sensing. A typical CD model accepts a bi-temporal image pair and predicts a change map that delineates the change type at each pixel, expressed as either *change* or *no-change* in a binary CD problem. CD plays a role in a wide array of applications including damage assessment [2], urban studies [3], ecosystem monitoring [4], agricultural surveying [5], and resource management [6].

Recently, the growing availability of very high resolution (VHR) multi-spectral imagery has further spurred the development of CD, pushing this area of study to the forefront of automated Earth observation research [7], [8]. In this regard, extensive efforts toward machine learning-based CD are opening up new avenues for monitoring land cover change at a fine scale [9]. In spite of the enrichment of training data with the increasing spatial resolution, the CD methods based on machine learning are still hindered by two major impediments. First, the objects with the same semantic concept can show distinct spectral characteristics at different times and different spatial locations. This is caused by the complexity of objects present in the scene and different imaging conditions at the bi-temporal phases [9]. Second, the definition of change varies across applications and can be in part dependent on the subjective consciousness of humans [10], which increases the risk of finding pseudo changes in the detection results from the CD method. Fortunately, recent advances in deep learning technology [11] have emerged as a promising new solution to the CD problem. The deep learning architectures, which are particularly valuable for exploiting discriminative and generalizable features, soon became the prevalent tool when establishing CD models.

According to [12], the general architecture of deep learning-based CD (DLCD) methods can be divided into two categories: 1) single-stream frameworks, and 2) double-stream frameworks. In a single-stream framework, an early-fusion strategy is applied to the bi-temporal image pair before any further processing step, such that the image segmentation algorithms can be conveniently borrowed to solve the CD problem with only slight modifications. For example, in [13], the difference maps of the pre-event and post-event images were first computed, and the multivariate morphological reconstruction operators were applied to remove noise. The filtered result was then processed by a symmetric fully convolutional neural network (CNN) to localize the changed areas. Unlike the single-stream frameworks, in a double-stream framework, Siamese and pseudo-Siamese encoders are commonly used to deal with the two temporal inputs. Different CD approaches then focus on how the different features extracted from the dual encoders are consumed to produce the change map [14], with one [15] or more [16] decoders. For instance, the authors of [17] proposed to fuse the bi-temporal features by means of using a temporal-symmetric transformer, which guarantees

Manuscript received 30 March 2022; revised 8 October 2022; accepted 16 November 2022. Date of publication 7 December 2022; date of current version 16 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61871298 and Grant 42071322, and in part by the Natural Science Foundation of Hubei Province under Grant 2020CFA053. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mai Xu. *(Corresponding author: Hongyan Zhang.)*

Manhui Lin and Hongyan Zhang are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China (e-mail: zhanghongyan@whu.edu.cn).

Guangyi Yang is with the School of Electronic Information, Wuhan University, Wuhan 430072, China.

Digital Object Identifier 10.1109/TIP.2022.3226418

the symmetry of deep features in time order and enjoys strong representation ability.

Although existing DLCD methods have shown dominant advantages in various scenarios, there still remain two limitations that need to be overcome. First, few researchers have oriented their effort toward a specialized modeling of time. Regardless of when the bi-temporal features are combined in the CD framework (e.g., early-fusion or late-fusion), point-to-point differencing and channel-wise concatenation are two fusion operators that have been most frequently adopted. These two operators, however, essentially imply a compare-and-contrast operation on the input image pair, whether explicit or implicit. We argue that neither point-wise differencing nor simple concatenation account for sufficient consideration on the temporal dimension. This is one of the reasons why existing approaches fail to precisely locate the real changes and immensely suppress the pseudo ones. Second, the challenges posed by the coupling of spatial and temporal domains have not received enough attention in algorithm design. For the single-stream approaches, a joint single-path structure is employed, such that this coupling exists throughout the encoding-decoding pipeline. For the double-stream approaches, the encoders harness only spatial features, while the decoder has to deal with both spatial and temporal features. We argue that the spatiotemporal coupling makes it hard for the network to concentrate on one aspect at a time, thereby adversely impacting the training of the CD models.

To tackle these issues, our conceptual idea is to explicitly model the change process, i.e. the transition, and thus naturally convert the change detection problem into a transition understanding task. To give a literal explanation, in our context, the concept of *transition* slightly distinguishes from *change* in that the latter refers to the discrete events of state alteration, whereas the former puts more emphasis on the continuous process (i.e. a time period) of becoming different. Guided by this insight, we build a new CD framework, named pair-to-video change detection (P2V-CD). First, we seek to augment temporal information by transforming the input bi-temporal image pair into a video clip, which in turn translates the CD task into a dense classification problem with video data as input. Then, we construct two decoupled encoders to handle the spatial and temporal streams separately, in a way that fundamentally alleviates the negative effects brought by the spatiotemporal coupling. The dual-encoder design not only addressed the spatiotemporal coupling issue, but also forms a more efficient architecture for dense classification of an input video. Apart from that, lateral connections are inserted in the model to promote the positive influence propagated between the two encoders. Finally, the deep supervision technique is further introduced, which enforces the learning of more fruitful and expressive features. We believe that our work not only provides a new paradigm for detecting change without explicitly contrasting the bi-temporal images, but also marks a first step toward the interconnection of CD and video understanding.

The main contributions of this paper are summarized as follows:

1) The CD problem is framed as a video understanding task for the first time. We build a CD model based on this idea, featuring a novel pair-to-video transformation and a dual-branch encoder.

2) We explicitly consider the spatiotemporal coupling issue in CD encoders and propose a new network design to decouple the spatial and temporal dimensions of the CD problem.

3) The proposed P2V-CD method achieves state-of-the-art results both qualitatively and quantitatively on three public CD datasets, SVCD, LEVIR-CD, and WHU. The model size and computational complexity of P2V-CD are moderate.

The remainder of this paper is organized as follows. Section II reviews the relevant literature on CD and video understanding. Section III gives the details of our proposed P2V-CD framework. In Section IV, the experimental setting is descried, and the qualitative and quantitative results are reported. Finally, conclusions are drawn in Section V.

II. Related Work

A. Modeling of Time in Change Detection

One crucial issue in CD is modeling the temporal correlation between images of the same location on different dates [18]. Since the early stage of CD, substantial effort has been devoted to the exploitation of discriminative features [19], [20] and the alignment of domain spaces [21], [22]. Yet, the dedicated modeling of time has long been overlooked. Although deep learning has provided a powerful tool to create more complicated CD models, most DLCD approaches inherit the practices of conventional CD methods [23], [24] and use straightforward strategies such as channel concatenation and point-wise differencing when building temporal relations [25], [26], [27]. We argue that such simple strategies cannot accommodate the key characteristics of the CD problem. Nevertheless, there are also approaches that focus on the temporal dimension, which include two major categories: recurrent neural network (RNN) [28] based methods and attention-based methods.

In RNN-based methods, the multi-temporal features are represented as a sequence that flows through the RNN cells and updates the hidden states. Mou et al. [18] took a first step to combining CNNs and RNNs for the purpose of CD. They investigated three types of RNN structures using two datasets and found that the RNNs built with the long short-term memory (LSTM) block achieved the top performance. Chen et al. [29] proposed SiamCRNN, which is based on the same idea with [18], but with more ad hoc designs for heterogeneous multi-source image pairs. The RNN-based methods in [18] and [29] are able to adaptively learn the temporal dependencies between the bi-temporal images. However, the input sequence of the RNN in both methods technically consists of only two images. This may question the applicability of the LSTM in CD tasks, since the LSTM was originally designed for processing longer sequences [30]. At the same time, short sequences do not provide a dense temporal context for an accurate detection. In view of these shortcomings, Papadomanolaki et al. [31] proposed to use additional images captured between the provided two dates. In their work, a UNet-like architecture that integrates fully convolutional LSTM blocks was engaged to model the temporal relationship of spatial feature representations.

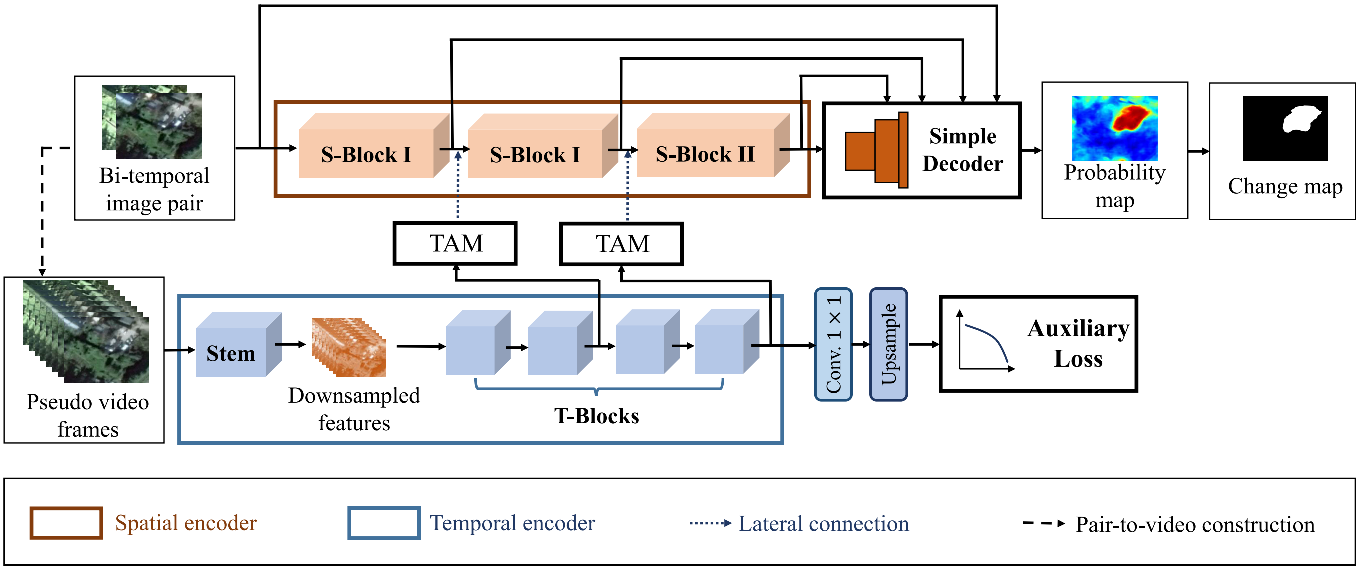


Fig. 1. Overall framework of P2V-CD. The input of the spatial encoder is the original bi-temporal image pair, and the input of the temporal encoder is the constructed pseudo video frames. The lateral connections are added between the encoders to enable the exchange of the spatiotemporal information. Finally, the decoder takes as input the output features of the third S-Block in the spatial encoder. TAM stands for temporal aggregation module in the figure.

Nonetheless, the auxiliary time series data can be costly to acquire, and the RNN architecture processes a multi-temporal image sequence successively, which prevents parallelization and incurs long training times [32].

With the explosively increasing application of attention mechanisms in computer vision [33], [34], [35], some researchers set out to model the temporal relationships in CD tasks using attention-based approaches. Chen and Shi [36] were some of the first to introduce the attention mechanism to the CD field. In [36], they designed two versions of spatiotemporal attention modules and demonstrated the effectiveness of attention-based feature enhancement. Yet, the high computational complexity remains an issue in this method, since all possible pixel pairs in the two periods of images were involved in the calculation of the attention map. Subsequently, Diakogiannis et al. [14] cast their focus on more complex similarity metrics and proposed a series of networks called CEECNet. The fractal Tanimoto function [37] was used to exploit the relative information that exists in the two inputs. Although the attention-based temporal fusion strategies are more advanced than concatenation and differencing, they are still limited to the compare-and-contrast operations between two images. We argue that for further improvement of detection performance, more sophisticated modeling of temporal structures need to be explored. This forms one motivation for our work.

B. Video Understanding

Video understanding refers to the recognition and localization (in space and time) of different actions or events appearing in a video [38]. In recent years, video understanding has leveraged the strong capacity of deep learning architectures to capture complex synergies across time, and it has realized progress in both efficiency and accuracy [39]. Despite the intrinsic defect of being computationally inefficient, 3D CNNs have been one of the commonly used deep structures for video understanding [40], [41], [42], as they have a higher degree of parallelism than RNNs and the ability to account for long-term dependencies [32]. The two-stream network [43] is a most representative method of video understanding using 3D convolutions. In [43], two 3D CNNs with identical structures are respectively used to deal with the spatial stream (a single frame) and the temporal stream (multi-frame optical flow) of the input video, and the softmax scores are combined by late fusion. A similar approach is the SlowFast network [44], which consists of a slow pathway operating at a low frame rate and a fast pathway operating at a high frame rate, to capture the motion at fine temporal resolution. We stress that both [43] and [44] adopted a dual-branch architecture that decouples space and time in the video. This inspired us to devise space-time decoupled encoders, in order to address the CD problem in a more explicit manner.

Although there are vast differences in fundamental objectives and general methodologies, the CD and video understanding tasks both involve two basic dimensions, space and time, and pay paramount attention to temporal reasoning. It is therefore of interest to find out how these two fields connect with each other.

III. Pair-to-Video Networks for Change Detection

A. Overall Framework

The overall architecture of the proposed P2V-CD network is shown schematically in Fig. 1. Unlike the existing CD methods, P2V-CD explicitly separates the spatial and temporal dimensions of the CD problem by viewing and processing the input data from both perspectives. That is, the proposed model treats the bi-temporal images both as a pair and as a video. In the spatial aspect, the concatenation of the

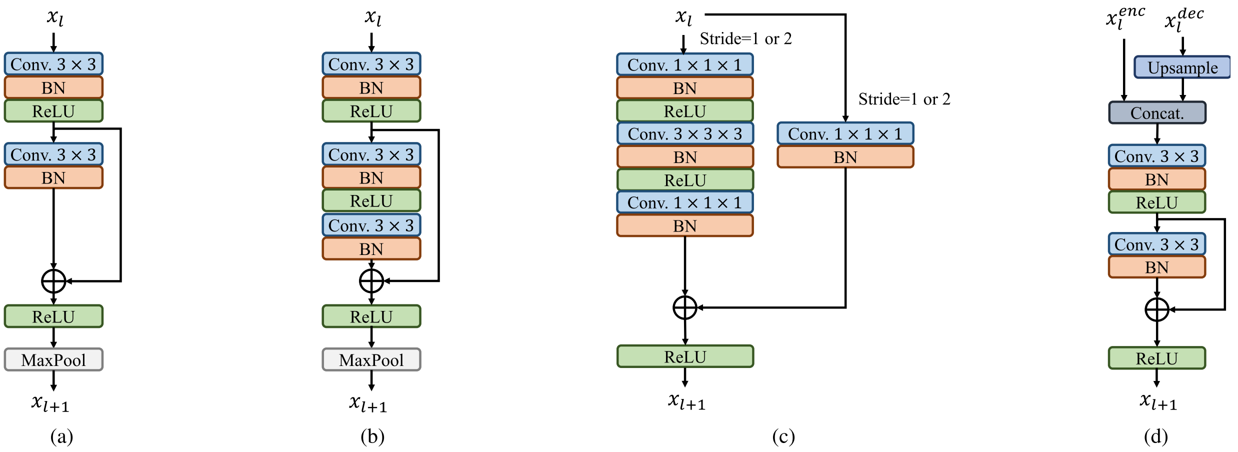


Fig. 2. Micro-topology of P2V-CD: (a) the type-one spatial block (S-Block I), (b) the type-two spatial block (S-Block II), (c) the temporal block (T-Block), and (d) the decoding block. In (a), (b), and (c), x_l denotes the input features of the current block, and x_{l+1} denotes the output features of the current block. In (d), x_l^{enc} denotes the input encoding features, x_l^{dec} denotes the input decoding features, and x_{l+1} denotes the output features.

bi-temporal images is used as the input, and a spatial encoder is responsible for capturing the spatial context that helps locate changed areas. We build the spatial encoder with a series of spatial blocks (S-Blocks). In the temporal aspect, we first construct a pseudo video frame sequence from the input image pair through pair-to-video transformation, in order to get a finer view of the temporal data. Then, a temporal encoder, consisting of a stem and four temporal blocks (T-Blocks), is designated to excavate the temporal information regarding change. Note that in the stem, the constructed video is spatially downsampled to reduce spatial information, such that the temporal encoder can be more focused on the temporal dimension. Deep supervision is imposed on the output of the fourth T-Block, in order to enhance the discriminative capacity of the temporal features. Furthermore, we fuse the feature maps of deeper layers in the two encoders through lateral connections. In specific, the feature maps from the second and the fourth T-Blocks in the temporal encoder are first transformed by a temporal aggregation module (TAM) and then forwarded to the second and the third S-Blocks in the spatial encoder, respectively.

We claim that the motivation of the dual-branch encoder design is two-fold: First, due to the richer temporal information, a video understanding model for CD purposes are more prone to the spatiotemporal coupling issue than conventional CD models that take two images as input. By decoupling the spatial and temporal dimensions, the dual-branch encoder can help boost detection performance. Second, CD models typically require heavier decoders than video understanding models in order to perform dense classification. A simple transfer use of video understanding architectures in the CD tasks (e.g. a UNet-like structure that connects two mirrored video understanding networks) calls for a considerable amount of computational cost, and the cost is sometimes unaffordable. The dual-encoder designs differs from pure 3D networks in that the more efficient 2D convolutional layers are used to process the high-resolution input data, while the more complex 3D convolutional layers are used to process downsampled input. This makes the network efficient for practical use.

To parse the information on change from the spatial and temporal clues captured by the encoders, a simple yet efficient decoder is further devised. The decoder hierarchically chains a convolutional layer (the *bottom* convolution) and four basic decoding blocks to predict a probability map in a progressive fashion. The details of the decoding block are shown in Fig. 2 (d). The input of the decoding block is the features of the previous decoding block or the bottom convolution (i.e. the decoding features), and the features from the spatial encoder at the corresponding scale (i.e. the encoding features). The decoding feature maps are first upsampled to fit the size of the encoding feature maps. Afterward, we associate the encoding and decoding features by concatenating them in the channel dimension, and the fused features are resolved by a stack of two convolutional layers. The rectified linear unit (ReLU) activation function is used to add nonlinearity, and the batch normalization (BN) [45] layers are interleaved to stabilize the training process. We also use a residual connection [46] in the second convolutional layer. It is worth mentioning that the original bi-temporal images are transmitted to the final decoding block, which helps preserve the spatial details of the image. Finally, the probability map produced by the decoder is binarized with a threshold of 0.5 to obtain the change map. As described above, we use a much simpler and more compact decoder than that adopted in [25]. We argue that with the decoupled spatial and temporal encoders and the dedicated temporal modeling, the CD problem can be settled without a sophisticated decoder.

B. Pair-to-Video Construction

One key idea behind the proposed P2V-CD method is to frame the CD problem as a transition understanding task. To achieve this goal, a crucial step is to construct a video, or more concretely, a frame sequence, from the given pair of bi-temporal images. The pioneering work in [31] also proposed to explicitly reveal the procedure of temporal evolution, where additional time series images, acquired at dates between the acquisition time of the original images, were gathered for use. However, a major concern is that not all the data required in the construction of the transition video can be collected in an economical and less laborious manner at high quality for any intermediate time. In view of this, we propose an alternative strategy to generate the frame sequence directly from the bi-temporal images, without the need for the manually captured external data. The constructed pseudo video comprises N frames in total, and the n-th ($0 \leq n < N$) frame $\mathbf{F}_n$ is deduced using the original images $\mathbf{I}_1$ and $\mathbf{I}_2$, along with the frame index n. In this sense, the generic formula to determine $\mathbf{F}_n$ is:

$$\mathbf{F}_n = \phi\left(\mathbf{I}_1, \mathbf{I}_2, n\right). \tag{1}$$

Note that the mapping ϕ in (1) is constrained by exactly two boundary conditions, i.e. $\phi\left(\mathbf{I}_1, \mathbf{I}_2, 0\right) = \mathbf{I}_1$ and $\phi\left(\mathbf{I}_1, \mathbf{I}_2, N-1\right) = \mathbf{I}_2$. It is challenging to decide ϕ, due to the ill-posed nature of the problem of generating a longer image sequence with just two source images. In this work, we propose to perform linear interpolation in the temporal dimension to simulate the intermediate transition states. Mathematically, the n-th interpolated frame can be calculated by:

$$\mathbf{F}_n = \mathbf{I}_1 + \frac{n}{N-1}\left(\mathbf{I}_2 - \mathbf{I}_1\right) \tag{2}$$

An example of the constructed pseudo video frames can be found in the first row of Fig. 9. Obviously, (2) assumes a uniform transition from $\mathbf{I}_1$ to $\mathbf{I}_2$, which is favored for its simplicity. Technically, the P2V construction is equivalent to the process of sampling frames from a virtual video that depicts a linear transition of all pixels, with a temporal resolution (or frame rate) proportional to $\frac{1}{N}$. Also, it is clear from (2) that the output frame sequence of the P2V construction advantageously prevents the loss of information by including the bi-temporal input as a subset. Since real-world transitions can be rather complex processes, in most cases, (2) tends to give less faithful approximations to the intermediate states of the real change. However, we will show in Section IV that the linear interpolation strategy is sufficient for P2V-CD to achieve competitive detection results. This leads to the speculation that, the key player in the P2V-CD method is the explicit modeling of transition, rather than an accurate estimation (or observation) on the true transition occurring on the ground.

C. Spatial and Temporal Blocks

As illustrated in Section III-A, the two encoders of the proposed P2V-CD network are explicitly devised to focus on different dimensions of the data, which drives the expertise for the design of the building blocks.

1) Spatial Blocks: The spatial encoder emphasizes the spatial structure of the input image, for which we use two types of S-Blocks, namely S-Block I and S-Block II, as shown in Fig. 2 (a) and (b), respectively. Both types of S-Block start with a conv-BN-ReLU-conv-BN sequence and end with a pooling layer, except that S-Block II contains one more convolutional layer (also, the corresponding BN and ReLU layers). The max-pooling layer is leveraged for downsampling, making it possible to capture a multi-scale spatial context by stacking the S-Blocks. Further, a shortcut connection is added between the output of the first ReLU layer and the output of the last normalization layer in the block. For S-Block II, the entire block can be seen as a standard ResBlock [46] plus a preceding convolutional layer. However, for S-Block I, please note that the residual connection skips only one convolutional layer and differs from the original design in [46] that skips two convolutional layers. The reason is two-fold: On the one hand, we want to keep the basic block lightweight by reducing the use of convolutional layers. On the other hand, the residual connection in S-Block I is used to preserve spatial details, rather than to circumvent the degradation problem [46] in deep networks.

2) Temporal Blocks: The building block of the temporal encoder, i.e. the T-Block, is illustrated in Fig. 2 (c). The T-Block is analogous to the 3D ResBlock introduced in [47], where 3D convolutional layers with a kernel size of 3 in the temporal dimension are utilized to contextualize the features related to temporal events. We choose this architecture for its simplicity and effectiveness. Meanwhile, taking into account the distinct role of the temporal encoder compared with the spatial encoder, we propose two bespoke designs in addition. First, a conv-BN-ReLU stem is inserted at the beginning of the temporal encoder. The kernel size of the input convolutional layer is set to $3 \times 9 \times 9$, and the stride is set to $1 \times 4 \times 4$, yielding $4\times$ downsampled feature maps to be processed by the consequent T-Blocks. This design endows the temporal encoder with the capability to efficiently exploit the temporal information, while paying less attention to the spatial details. The spatial clues, on the other hand, can be provided by the other encoder in a less redundant manner. Second, we adopt a bottleneck structure in the T-Block instead of the basic ResBlock structure. In this formalism, two $1 \times 1 \times 1$ convolutional layers are used for first reducing and then restoring the number of channels, which lowers the computational cost [46] and further allows us to economically build wider and deeper networks. It is also important to note that the temporal encoder, despite its name, actually encodes both temporal and spatial features through 3D convolutions. This does not conflict with the design of spatiotemporal decoupling, because the temporal encoder pays much less attention to the spatial dimensions than the temporal dimension. We observed in our experiments that the fully-decoupled version, where all 3D kernels in the temporal encoder have a size of 1 in the spatial dimensions, achieved worse detection results. This possibly implies the significance for the temporal encoder to incorporate a certain amount of spatial information.

3) Instantiation: We now give the detailed specification of the building blocks in the P2V-CD network. As described in Section III-A, the spatial encoder consists of three S-Blocks, and the temporal encoder is composed of a stem and four T-Blocks. The number of the output channels in each S-Block was set to 32, 64, and 128, respectively. We selected the S-Block I type for the first two S-Blocks and the S-Block II type for the last S-Block. The temporal encoder should contain more layers and more convolutional filters than its spatial counterpart, because the video frames provide a richer source of data than the original bi-temporal image pair. Hence, setting the temporal stem to output 64-channel features, we use 256,

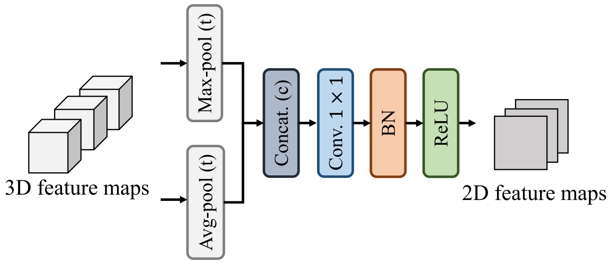


Fig. 3. Illustration of the temporal aggregation module. We first aggregate the temporal information using global max-pooling and global average pooling. Then, the aggregated features are transformed to 2D feature maps, with the help of a conv-BN-ReLU sequence. In the figure, (t) means that the operation is performed in the temporal dimension, and (c) means that the operation is performed in the channel dimension.

256, 512, and 512 filters in the endmost convolutional layers of the four T-Blocks, respectively. In the third T-Block, the first convolutional layers of both the main branch and the residual branch have an input stride of 2 in all three dimensions, in order to perform downsampling spatially and temporally. We note that the output stride of both encoders is 8.

D. Lateral Connections

One basic intuition behind the proposed P2V-CD method is to decouple the two complementary aspects, space and time, of the CD task. Since a dual-encoder design is adopted to capture the spatial and temporal context, another central problem concerns how the two encoders interact with each other for mutual improvement. As pointed out in [44], inter-pathway information exchange is important in that one of the pathways (or encoders in our context) would otherwise be unaware of the representation learned by the other pathway, which could impair the prediction accuracy of the network. Since the transformation from 2D features (as in the spatial encoder) to 3D features (as in the temporal encoder) is non-trivial, we follow [44] and use unidirectional connections to fuse the features from the temporal encoder into the spatial encoder. It should be noted that although the connections are one-way, the gradient flow from the spatial encoder also in response affects the parameters of the temporal encoder during model training.

To this end, we develop a temporal aggregation module (TAM) that bridges the gap between the features of the two encoders, whose architecture is exhibited in Fig. 3. The design of the TAM is inspired by [48]. Suppose that the output features of the T-Block have C channels. First, the global average pooling and global max-pooling operations are applied to the 3D features along the temporal dimension. With temporal pooling, we expect to *summarize* the transition information for each pixel and thus find out more clues about where the change occurred. The pooled features, which can be seen as a temporal signature, are then concatenated in the channel dimension to generate an efficient feature descriptor with $2C$ channels. Finally, a 1×1 convolutional layer with C filters, in conjunction with a BN layer and the ReLU activation function, executes a point-to-point transformation on the aggregated feature maps. The output of the TAM is fused into the spatial encoder by concatenation. It is necessary to point out that the output feature maps of the TAM do not always match the corresponding feature maps of the S-Block in spatial resolution. An upsampling layer is used to resize the TAM features to address mismatches.

E. Loss Function

As illustrated in Fig. 1, a 1×1 convolutional layer and an upsampling layer are added to the end of the second TAM, which yields a side output with the same size as the ground-truth labels. This intermediate probability map is then used to compute the auxiliary loss, allowing the network parameters of the temporal encoder to be trained in a deeply-supervised [49] manner. We resort to the deep supervision technique for two reasons: On the one hand, the deep supervision mitigates the vanishing gradient issue and thus improves the convergence of deep networks [50]. On the other hand, by receiving direct feedbacks from the ground-truth change maps, the hidden layers are regulated to learn more discriminative feature representations [51], which relieves the burden on the decoder and further boosts the detection performance.

With the deep supervision imposed on the temporal encoder, the overall loss function is formulated as:

$$L = l\left(\mathbf{P}_{final}, \mathbf{R}\right) + \lambda l\left(\mathbf{P}_{inter}, \mathbf{R}\right), \tag{3}$$

where l represents the specific loss function adopted for each output-labels pair, $\mathbf{P}_{final}$ and $\mathbf{P}_{inter}$ denote the output probability maps of the final layer and the intermediate layer, respectively, $\mathbf{R}$ stands for the reference labels, and λ is the weighting factor. In consideration of the imbalanced distribution of the change and no-change classes in the CD task, we choose the weighted binary cross-entropy (BCE) function for l in order to train more reliable classifiers. The weighted BCE loss function is defined as:

$$\begin{aligned} l_{BCE} = \frac{1}{H \times W} \sum_{i,j} \big[& -w_c R_{i,j} log\left(P_{i,j}\right) \\ & - w_u \left(1 - R_{i,j}\right) log\left(1 - P_{i,j}\right) \big], \end{aligned} \tag{4}$$

where H and W denote the height and width of the image, and w_c and w_u are the tradeoff coefficients for the change and no-change classes.

IV. Experiments and Discussion

A. Dataset Description and Experimental Setting

Our experiments involve three publicly available CD datasets, which are described as follows:

1) The Season-Varying Change Detection (SVCD) Dataset [52]: This general-purpose CD dataset is composed of eleven registered remote sensing image pairs obtained from Google Earth. The spatial resolution of the images ranges from $3\,cm$ to $100\,cm$ per pixel. After random cropping and rotation, 16000 pairs of 256×256 fragments were generated, of which 10000 pairs are used for training, 3000 for validation and 3000 for testing, respectively. At least one changed pixel is included in each of the generated fragment pairs. For this dataset, an object is annotated as changed only if its semantic class has been altered, whereas seasonal changes in the scene, such as the withering of leaves, are not taken into account in the ground-truth labels. This places a higher demand on the CD methods in terms of the accuracy of radiometric calibration and the utilization of semantic information.

2) The Learning, Vision, and Remote Sensing Change Detection (LEVIR-CD) Dataset [36]: This building CD dataset consists of 637 pairs of remote sensing image tiles with a size of 1024×1024 pixels. The spatial resolution is $0.5\,m$ per pixel, and 31333 individual changes that cover diverse types of buildings are contained in the dataset. We use the standardized train/validation/test split provided by the authors. Further, for ease of training and evaluating models on large raster images, we followed [14] and extracted 256 × 256 chips from the original tiles using a sliding window. The training chips were overlapped with a stride of 128 pixels, in order to create adequate samples, and the validation chips are non-overlapping with a stride of 256 pixels. During testing, a stride of 64 was adopted, and the final prediction score was averaged over the prediction scores of all inference windows that overlap on the given pixel. In contrast to the SVCD dataset, large quantities of tiles that contain purely negative samples reside in the LEVIR-CD dataset, which allows us to assess how a CD approach applies to real-world scenarios.

3) The Wuhan University (WHU) Building Change Detection Dataset [53]: This dataset consists of two aerial images of size 32507×15354. The spatial resolution is $0.3\,m$ per pixel. Akin to the LEVIR-CD dataset, the WHU dataset focuses exclusively on building changes. As no official split was suggested in [53], we first cropped the original images into 256 × 256 chips with no overlap and then randomly split the chips into three subsets: 6096/762/762 for training/ validation/test, respectively. This is the same splitting strategy as was adopted in [9]. It should be stressed that the WHU dataset is smaller than the other two datasets, so the model trained on this dataset is more prone to overfitting. Therefore, it is observable on this dataset the sensitivity of a CD approach toward the data volume.

We use three metrics, precision, recall, and the F1 score, to quantitatively evaluate the performance of the CD models. These metrics are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{7}$$

where TP stands for the number of true positives, TN stands for the number of true negatives, FP stands for the number of false positives, and FN stands for the number of false negatives.

All models were implemented in PyTorch and the training was powered by a single NVIDIA TITAN RTX GPU. For all three datasets, the batch size was set to 8, and the Adam optimizer [54] was used to update the network parameters. Unless otherwise stated, we set the number of the constructed frames N to 8, and the weighting coefficient of the auxiliary loss λ to 0.4. The class-balancing weights of the BCE loss were fixed at 0.5 for the change and 0.5 for the no-change classes. For augmentation purposes, the training data were randomly flipped, shifted, and rotated by multiples of 90°, before being collated to form the mini-batch. For the SVCD dataset, the initial learning rate was 0.0004, and the models were trained for 260k iterations to achieve sufficient convergence. For the LEVIR-CD dataset, the learning rate initiated at 0.002, and we trained the models for 220k iterations. For the WHU dataset, the training started with a learning rate of 0.0004 and continued for 160k iterations. A step decaying strategy was applied to the learning rate schedule for all three datasets. The decay rate and the step size (in epochs) were set to 0.1 and 70 for the SVCD dataset, 0.2 and 60 for the LEVIR-CD dataset, and 0.2 and 105 for the WHU dataset. We used the validation step after each epoch to select the best model with the highest F1 score and reported its performance on the test subset.

B. Comparison With State-of-the-Art Methods

We compare P2V-CD against nine SOTA DLCD approaches, i.e., fully convolutional early fusion (FC-EF) [25], fully convolutional Siamese-concatenation (FC-Siam-conc) [25], fully convolutional Siamese-difference (FC-Siam-diff) [25], CDNet [55], the spatial-temporal attention neural network (STANet) [36], the bi-temporal image transformer (BIT) [9], L-UNet [31], the deeply supervised image fusion network (DSIFN) [51], and SNUNet [10]. Among them, FC-EF, FC-Siam-conc, FC-Siam-diff, and CDNet are four classic deep learning-based methods commonly used in the CD field. STANet is a late-fusion method that integrates the spatial-temporal attention mechanism to bolster the discrimination between the deep features of the bi-temporal images. As another attention-based method, BIT employs a tailored CNN in combination with a pair of transformer encoder and decoder to address the CD problem in an efficient manner. L-UNet is typical of the convolutional and recurrent approaches, where fully convolutional LSTM blocks are equipped in a deep neural network for end-to-end training. DSIFN and SNUNet are two recently proposed methods, both of which have been shown to be effective by obtaining competitive results on large-scale CD datasets.

In order to make an apples-to-apples comparison, we re-implemented all these comparative methods and reproduced the results in the same experimental environment. Also, slight modifications were made on some of the methods. Specifically, we used the BCE loss to train all models, and the weights and biases, except those of the pre-trained backbone in DSIFN, were randomly initialized for learning-from-scratch. The multi-task branches were removed from L-UNet, and we fed whole images to the network, instead of the cropped patches as in the original paper. For STANet, a BIT-style classifier was used in place of the original one, as we observed more stable gradient descent and higher evaluation indices by doing so. For each comparative method, we chose a set of optimal hyperparameters that maximized the F1 score on the validation subset.

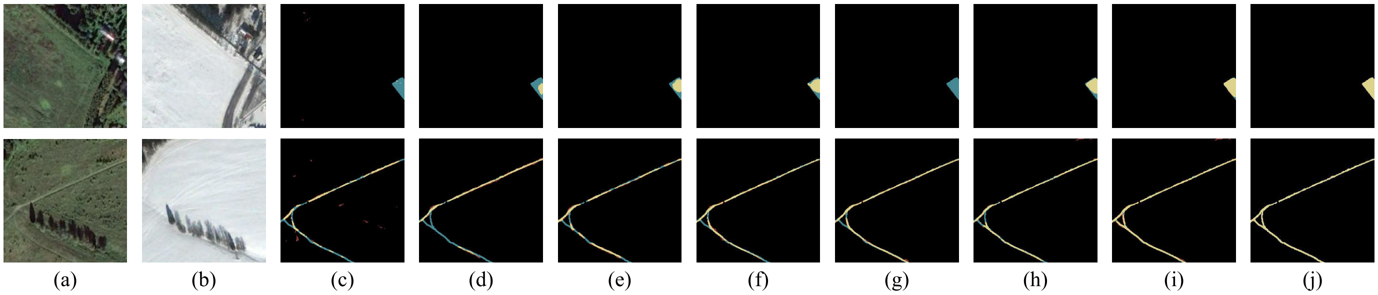


Fig. 4. Qualitative results of the different methods on the SVCD dataset: (a) the RGB image of the first date, (b) the RGB image of the second date, (c) CDNet, (d) STANet, (e) BIT, (f) L-UNet, (g) DSIFN, (h) SNUNet, (i) P2V-CD, and (j) the grount-truth labels. We highlight the TP areas in yellow, the FP areas in red, and the FN areas in blue. The black color denotes the TN areas. Please zoom in for a better view.

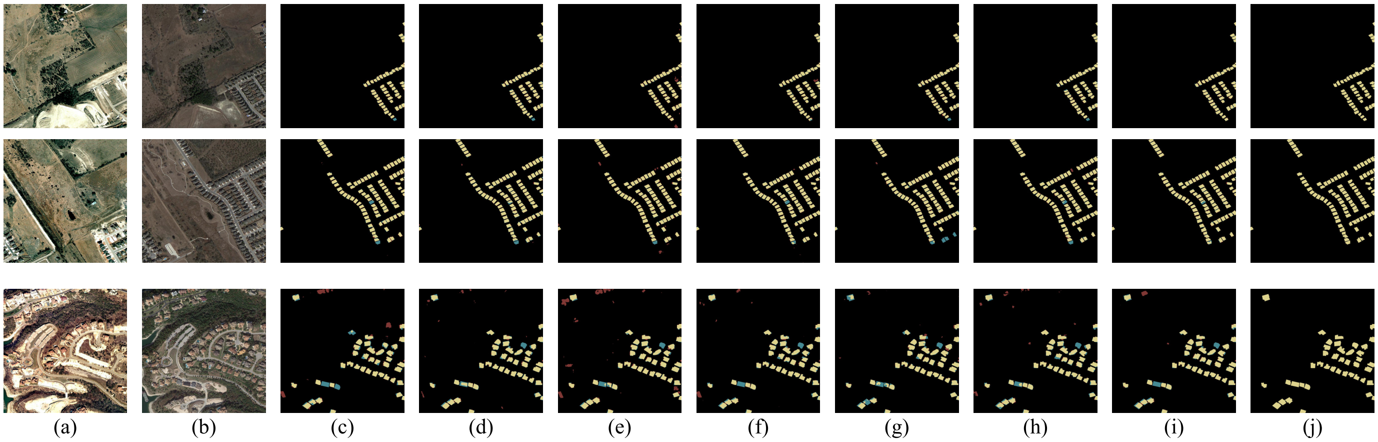


Fig. 5. Qualitative results of the different methods on the LEVIR-CD dataset: (a) the RGB image of the first date, (b) the RGB image of the second date, (c) CDNet, (d) STANet, (e) BIT, (f) L-UNet, (g) DSIFN, (h) SNUNet, (i) P2V-CD, and (j) the grount-truth labels. We highlight the TP areas in yellow, the FP areas in red, and the FN areas in blue. The black color denotes the TN areas. Please zoom in for a better view.

TABLE I
QUANTITATIVE RESULTS OF THE CD METHODS ON THE SVCD DATASET

Model	Precision	Recall	F1 score	Params (M)
FC-EF	0.8741	0.5180	0.6505	1.35
FC-Siam-conc	0.9249	0.5878	0.7188	1.55
FC-Siam-diff	0.9365	0.5432	0.6876	**1.35**
CDNet	0.9251	0.8777	0.9007	1.43
STANet	0.9517	0.9288	0.9401	16.97
BIT	0.9607	0.9349	0.9476	3.04
L-UNet	0.9652	0.9441	0.9545	8.45
DSIFN	0.9765	0.9485	0.9623	35.99
SNUNet	0.9809	0.9742	0.9775	12.03
P2V-CD	**0.9857**	**0.9826**	**0.9842**	5.42

TABLE II
QUANTITATIVE RESULTS OF THE CD METHODS ON THE LEVIR-CD DATASET

Model	Precision	Recall	F1 score	Params (M)
FC-EF	0.9064	0.7884	0.8433	1.35
FC-Siam-conc	0.9255	0.8440	0.8828	1.55
FC-Siam-diff	0.9121	0.8118	0.8590	**1.35**
CDNet	0.9152	0.8805	0.8975	1.43
STANet	**0.9338**	0.8658	0.8985	16.97
BIT	0.9080	0.8974	0.9027	3.04
L-UNet	0.9318	0.8864	0.9085	8.45
DSIFN	0.9245	0.8706	0.8967	35.99
SNUNet	0.9308	0.8990	0.9147	12.03
P2V-CD	0.9332	**0.9060**	**0.9194**	5.42

TABLE III
QUANTITATIVE RESULTS OF THE CD METHODS ON THE WHU DATASET

Model	Precision	Recall	F1 score	Params (M)
FC-EF	0.8065	0.7886	0.7974	1.35
FC-Siam-conc	0.7035	0.8764	0.7805	1.55
FC-Siam-diff	0.6788	0.8359	0.7492	**1.35**
CDNet	0.9249	0.8809	0.9023	1.43
STANet	0.9269	0.8899	0.9080	16.97
BIT	0.8491	0.8720	0.8604	3.04
L-UNet	0.7900	0.8938	0.8387	8.45
DSIFN	**0.9626**	0.8677	0.9127	35.99
SNUNet	0.8990	0.8682	0.8833	12.03
P2V-CD	0.9548	**0.8947**	**0.9238**	5.42

The quantitative results on the three datasets are exhibited in Table I, II, and III, respectively. The highest index or smallest model size in each column is marked in bold, and the second-best value is underlined. Fig. 4, 5, and 6 contain the qualitative results of CDNet, STANet, BIT, L-UNet, DSIFN, SNUNet, and P2V-CD, where we color the TP areas in yellow, the FP areas in red, the FN areas in blue, and the TN areas in black. The output change maps of FC-EF, FC-Siam-conc, and FC-Siam-diff are not presented, because these methods achieved less competitive results in our experiments.

1) Experimental Results on the SVCD Dataset: In spite of the advantage of occupying smaller memory footprints, FC-EF, FC-Siam-conc, FC-Siam-diff, and CDNet achieve the worst results in terms of all three metrics. The attention-based method BIT shows a comprehensive superiority over its counterpart, STANet. The RNN-based method LUNet outperforms the attention-based methods in

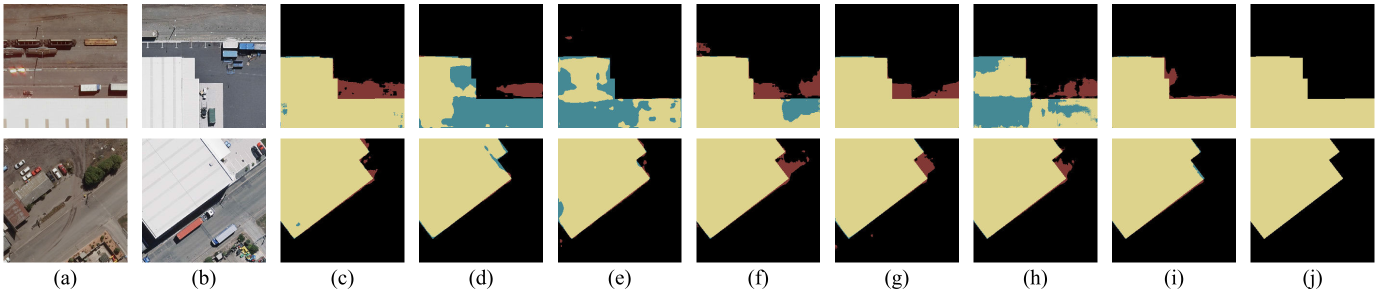


Fig. 6. Qualitative results of the different methods on the WHU dataset: (a) the RGB image of the first date, (b) the RGB image of the second date, (c) CDNet, (d) STANet, (e) BIT, (f) L-UNet, (g) DSIFN, (h) SNUNet, (i) P2V-CD, and (j) the grount-truth labels. We highlight the TP areas in yellow, the FP areas in red, and the FN areas in blue. The black color denotes the TN areas. Please zoom in for a better view.

terms of both the precision and recall on this dataset. SNUNet and DSIFN obtain the second- and third-highest F1 scores, respectively, demonstrating their advantages in the resistance to the pseudo changes caused by seasonal variation. Finally, our proposed P2V-CD method achieves the best overall performance, with a precision of 0.9857, a recall of 0.9826, and an F1 score of 0.9842. It is noteworthy that the P2V-CD model has a moderate number of parameters (5.42 M), which is significantly smaller than DSIFN (35.99 M in total, with 21.28 M trainable parameters) and SNUNet (12.03 M).

A perceptual comparison of the CD methods is given in Fig. 4. Apparently, misclassified changed pixels exist in large quantities in the results of all methods except P2V-CD. In particular, there is a noticeable part of the building missing in the results of CDNet, STANet, and DSIFN. Also, the foot trails are missing in the results of CDNet, STANet, and BIT. Generally, the change maps of P2V-CD have the most desirable visual effect, which appear closest to the ground-truth.

2) Experimental Results on the LEVIR-CD Dataset: From Table II, FC-EF, FC-Siam-conc, and FC-Siam-diff again achieve the least satisfying detection results. The F1 score of DSIFN is on par with that of CDNet (the difference is within 0.001), but DSIFN achieves relatively higher precision and lower recall. Additionally, only marginal performance gains are observed from the results of CDNet to those of STANet. This is because STANet does not keep a sensible balance between precision and recall. The precision of STANet is the highest among all methods, which is 0.9338, yet the recall is much lower, which is only 0.8658. L-UNet and BIT are two competitive methods on this dataset, with the F1 scores over 0.90. SNUNet also shows prominent performance by attaining the second best results in the recall and F1 score metrics. Our method P2V-CD maintains excellent performance on the LEVIR-CD dataset, with an outstanding recall rate of 0.9060 and the highest F1 score of 0.9194.

The qualitative results shown in Fig. 5 provide a more intuitive comparison of the CD methods. We observe that most of the comparative CD methods suffer from spurious changes or missed detection in the heavily built-up areas. In contrast, the proposed P2V-CD method managed to precisely detect the construction of small buildings, even when trained with a highly imbalanced training set. In order to make a comprehensive analysis, we also show a case where P2V-CD does not achieve satisfying visual result in the last row of Fig. 5. In this case, the scene is more complex than the first two cases, and there are misclassified pixels in the results of all methods, including the proposed P2V-CD. This indicates that the ability of the proposed method to deal with more complex scenes needs to be further enhanced.

3) Experimental Results on the WHU Dataset: From Table III, FC-EF, FC-Siam-conc, and FC-Siam-diff are not observed to outperform the other methods. CDNet achieves a better precision and recall than the previous three methods, but its F1 score is still 0.0057 lower than that of STANet. We also observe that L-UNet and BIT, which achieve high evaluation scores on the other two datasets, fail to reproduce the good performance. It is reasonable to infer from the low precision that there are a considerable number of omissions in their detection results. We argue that these two methods are more sensitive to sample size, such that it is easier for them to suffer from overfitting when trained on the smaller WHU dataset. Albeit with the largest model size, DSIFN prevents overfitting with the help of the pretrained encoders. It achieves the highest precision of 0.9626 and the second-highest F1 score of 0.9127 on this dataset. In comparison, the SNUNet method, which also has abundant network parameters, is not able to compete with the much smaller CDNet model on this dataset. Our proposed P2V-CD method shows absolute dominance in the F1 score, leading by at least 0.0110 against the other methods.

False alarms are a major issue in the results of CDNet and DSIFN, as seen in Fig. 6. Besides, STANet and SNUNet incur both false and missed detections. For BIT and L-UNet, the observations on the change maps agree with the previous observations from Table III, as there are extensive FN areas in the results. The L-UNet method cannot effectively eliminate the falsely detected pixels in its change maps, either. The P2V-CD method achieves the best visual results among all approaches. Not only does P2V-CD delineate the changed areas with more complete boundaries, but it also significantly suppresses the pseudo changes in the detection result.

Collectively, the persistent good performance of P2V-CD suggests that it is a compelling approach for both the general-purpose CD and the building CD. Our method is also more robust to the massive existence of negative samples in the dataset, or a relatively small sample size. We believe that it is the explicit modeling of time, as well as the space-time decoupled design, that encourages the P2V-CD model to productively exploit valid temporal clues and localize changed areas with a higher accuracy.

TABLE IV

QUANTITATIVE RESULTS OF THE DIFFERENT VARIANTS OF P2V-CD ON THE SVCD DATASET

Model	Precision	Recall	F1 score	Params (M)
P2V-NoTE	0.9741	0.9643	0.9692	5.83
P2V-NoSE	0.9792	0.9742	0.9767	5.70
P2V-2DOnly	0.9826	0.9794	0.9810	5.52
P2V-2DOnlyFlat	0.9716	0.9669	0.9692	5.63
P2V-LateFusion	0.9829	0.9797	0.9813	5.46
P2V-LearnL	0.9849	0.9820	0.9835	5.42
P2V-LearnNL	0.9793	0.9758	0.9776	5.42
P2V-NoDS	0.9835	0.9808	0.9822	**5.42**
P2V-CD	**0.9857**	**0.9826**	**0.9842**	**5.42**

C. Ablation Study

In order to assess the contribution of each individual component of our proposed network independently, we ablate the P2V-CD model with various modifications on the architecture. There are eight variants of our network, in addition to the full model (denoted as P2V-CD). They are:

1) P2V-NoTE: We remove the temporal encoder, along with the lateral connections, to form a single-stream architecture.

2) P2V-NoSE: We remove the spatial encoder, add two additional T-Blocks in the temporal encoder, and connect the output of the TAMs directly with the corresponding decoder blocks to form a single-stream architecture.

3) P2V-2DOnly: We substitute all 3D convolutions in the temporal encoder with their 2D counterparts. The temporal dimension is preserved, and each 2D convolutional filter is shared by all pseudo video frames.

4) P2V-2DOnlyFlat: This variant is the same as P2V-2DOnly, except that the pseudo video is flattened, with all its frames concatenated in the channel dimension.

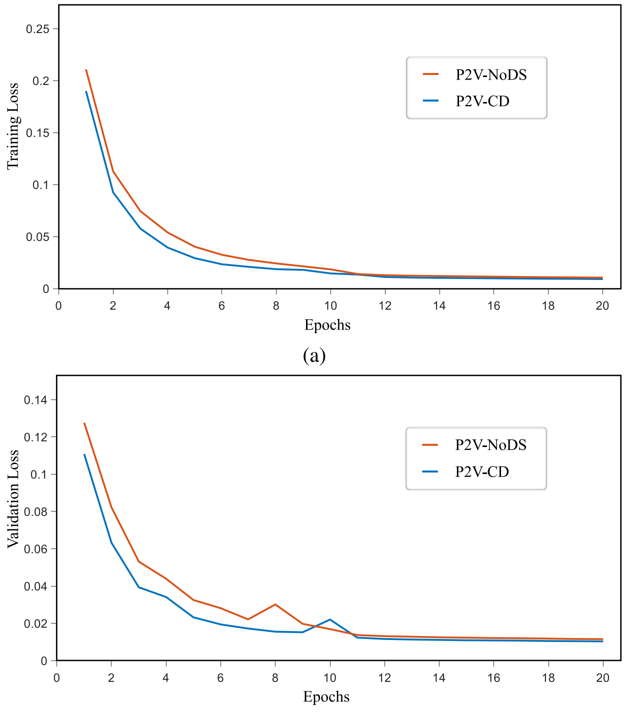
5) P2V-LateFusion: The lateral connections are removed. As an alternative, the output features of the two encoders are concatenated as the input to the decoder.

6) P2V-LearnL: A learning-based strategy is used to construct pseudo video frames from the bi-temporal image pair. More concretely, we use a convolution that operates on the temporal dimension, and the kernel size is 1.

7) P2V-LearnNL: This variant is the same as P2V-LearnL, except that two convolutional layers interleaved with non-linear activation is used instead of a single convolutional layer for the pair-to-video construction.

8) P2V-NoDS: We do not use the deep supervision technique for this variant.

To level the playing field, we kept all variants of P2V-CD comparable in model size by adjusting the number of convolutional filters and trained these networks using the same set of hyperparameters. The quantitative results from different ablation models on the SVCD dataset are presented in Table IV. Overall, it can be seen from the results that none of the ablation models outperforms the full P2V-CD model. Particularly, P2V-NoTE achieves the worst results, with the lowest recall and F1 score metrics. This signals that the temporal encoder is of prime importance in our proposed network. Removing the spatial encoder also causes a considerable decrease of detection accuracy, which is clearly seen by comparing the results of P2V-NoSE and P2V-CD. Besides, replacing the 3D convolutions with 2D convolutions while retaining the 3D data structures decreases the F1 score by 0.0032, which confirms the superiority of using 3D convolutions in the temporal encoder. A further reduction by 0.0118 in terms of the F1 score can be observed if the 3D features collapse to 2D structures. The performance gap between P2V-2DOnly and P2V-2DOnlyFlat indicates the importance of the TAM, as the capability of temporal aggregation is mainly provided by the TAMs when 2D convolutions are adopted in the temporal encoder. From the results of P2V-LearnL and P2V-LearnNL, we observe that using more advanced learning-based temporal interpolation strategies does not contribute to the performance of the CD model, as the model that adopts the simple linear strategy achieves better results in terms of all metrics. In contrast, we find that the most complex interpolation strategy we tried, the non-linear learning-based strategy, had a negative impact on the detection accuracy. We also observe that the absence of deep supervision leads to a significant performance drop in terms of all three metrics. For a more comprehensive comparison, we plot the loss curves of P2V-CD and P2V-NoDS during the training phase and the validation phase. The curves are shown in Fig. 7 (a) and (b), respectively. Based on Table IV and Fig. 7, we conclude that the deep supervision successfully accelerated the convergence of the network parameters and boosted the detection accuracy.



(a)

(b)

Fig. 7. Loss curves of models for each epoch: (a) the training loss curves of P2V-CD and P2V-NoDS and (b) the validation loss curves of P2V-CD and P2V-NoDS.

D. Effect of the Number of Frames

In this section, we conduct a sensitivity analysis to investigate how different values of N, the number of frames in

TABLE V

EFFECT OF THE NUMBER OF FRAMES ON THE SVCD DATASET

N	Precision	Recall	F1 score	MACs (G)
2	0.9797	0.9781	0.9789	**20.66**
4	0.9848	0.9816	0.9832	24.73
6	0.9853	0.9824	0.9838	28.79
8	0.9857	0.9826	0.9842	32.86
12	**0.9874**	0.9853	**0.9864**	40.99
16	0.9868	**0.9855**	0.9862	49.12

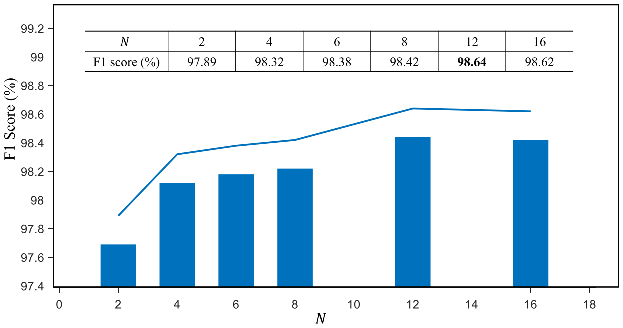


Fig. 8. Influence of the number of video frames N. The F1 score of the model gradually improves with the increase of N, yet a slight drop is observed from $N = 12$ to $N = 16$.

the constructed pseudo video, affects the performance of P2V-CD. To this end, six experiments were carried out on the SVCD dataset, each with a different value of N in the setting. For $N = 12$ and $N = 16$, we trained the model for more epochs until convergence. The results are tabulated in Table V. For a more intuitive visualization, we also plot the trend of the F1 score with respect to the varying values of N in Fig. 8. Roughly speaking, the F1 score gradually improves with the increase of N. This is an expected result—with more frames in the video, the model gains more useful knowledge about change from a larger amount of data and at finer temporal resolution. From this perspective, the pair-to-video construction can be seen as a data augmentation strategy. However, the increase in the F1 score is not monotonous, as there is a small performance drop from $N = 12$ to $N = 16$. We attribute this drop to the saturation of data, which is likely explained by two underlying reasons. On the one hand, there is an upper limit on the amount of data that a CD model is able to effectively learn from, which is usually determined by the model capacity. On the other hand, excessively large N values will reduce the proportion of the real data, i.e. the original bi-temporal images, in the constructed pseudo frame sequence, making it easier for the CD model to overfit the specific pattern of how the video is constructed. It is also observable from Table V that the multiply-accumulate operations (MACs) surge when N is set to larger values. This sharply increasing computational overhead prevents the adoption of long pseudo frame sequences. Finally, we stress that although constructing videos with more frames significantly benefits the detection performance, it is not the only factor that contributes to the competitive results of P2V-CD. To demonstrate this, a basic fact is that even if with only two frames, P2V-CD is able to maintain its superior performance in comparison with P2V-NoTE (0.9789 vs. 0.9692 in terms of the F1 score).

E. Visualization on Feature Representation

Seeking a more convincing explainability of our approach, in this section, we analyze the behavior of some intermediate layers in P2V-CD through feature map visualization. First, we visualize the per-frame features from the second and the fourth T-Blocks in Fig. 9. We used the technique proposed in [56] to produce the visualization result. Two variants of P2V-CD, namely P2V-2DOnly and the full model, were selected for the comparison. From Fig. 9, a noticeable difference between the results of P2V-CD and P2V-2DOnly is that the features of P2V-2DOnly tend to be less informative. For P2V-2DOnly, we observe low spatial variability in the second to sixth frames of the output features from the second T-Block, and in the last three frames of the output features from the fourth T-Block. Following [14], we calculated the image entropy of the features for each case over the whole test set, to further quantify the information content of the features. We found that for P2V-CD, the output features from the second T-Block have an entropy value of 4.1445, and the output features from the fourth T-Block have an entropy value of 4.1005, while the corresponding entropy values for P2V-2DOnly are 3.7550 for the second T-Block, and 3.3291 for the fourth T-Block. This supports our idea that with the explicit consideration on the temporal interactions, the temporal encoder is able to effectively utilize a richer source of information, which goes beyond conducting simple compare-and-contrast operations.

In Fig. 10 we visualize the output features of the first and the third S-Blocks in the spatial encoder. The compared variants were P2V-CD and P2V-NoTE. It can be observed from Fig. 10 that for the shallower layer, i.e. the first S-Block, P2V-CD depicts finer spatial details than P2V-NoTE; for the deeper layer, i.e. the third S-Block, the feature map of P2V-CD better highlights important semantic objects that occur in the ground-truth labels. To make a quantitative analysis, we propose to measure the preservation of fine textures using the total variation (TV) metric and gauge the quantity of semantic information using the Pearson product-moment correlation coefficient (PCC) between the feature map and the ground-truth image. A higher TV value indicates the better preservation of the small spatial structures in the image, and a higher PCC represents the closer relationship between the features and the target change map. The TV and PCC values calculated on the test set are listed in Table VI, where we also present the relative alteration of TV and PCC from P2V-NoTE to the full model. The results in Table VI are consistent with those in Fig. 10: For the first S-Block, the features of P2V-CD have a 32.21% higher TV value, but only half the PCC of P2V-NoTE; for the third S-Block, the TV value of P2V-CD is 15.13% lower than that of P2V-NoTE, yet the PCC is 36.23% higher. This behavior can be explained as follows: The spatial encoder in P2V-CD is initially more concentrated on the spatial contexts, and thus the features extracted by the first S-Block are less concerned about the time-dependent changes and have

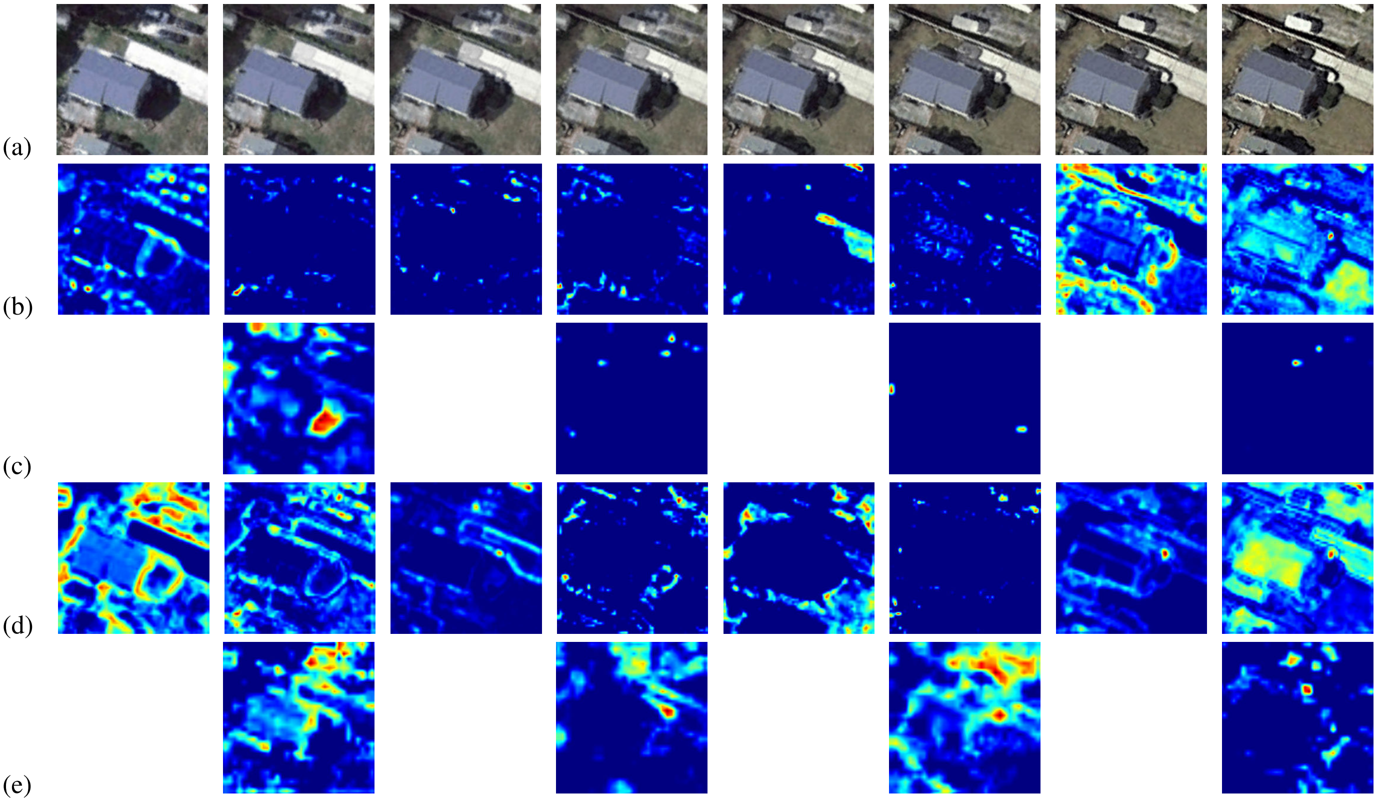


Fig. 9. Visualization results of the feature maps: (a) the pseudo video frames, (b) the output features of the second T-Block of P2V-2DOnly, (c) the output features of the fourth T-Block of P2V-2DOnly, (d) the output features of the second T-Block of P2V-CD, and (e) the output features of the fourth T-Block of P2V-CD. The red color indicates a high pixel value in the visualization results, while the blue color indicates a low pixel value.

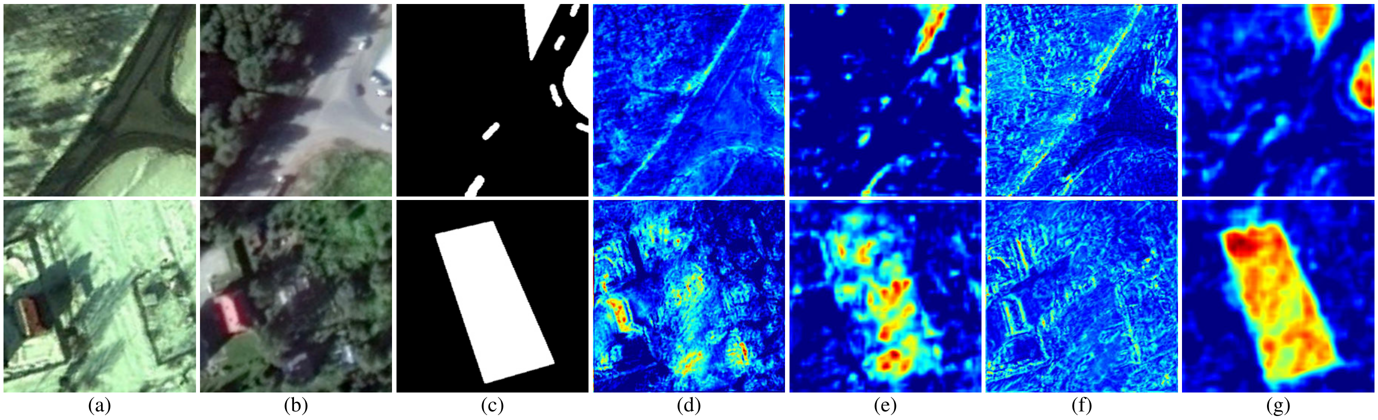


Fig. 10. Visualization results of the feature maps: (a) the RGB image of the first date, (b) the RGB image of the second date, (c) the ground-truth, (d) the output features of the first S-Block of P2V-No-TE, (e) the output features of the third S-Block of P2V-No-TE, (f) the output features of the first S-Block of P2V-CD, and (g) the output features of the third S-Block of P2V-CD. The red color indicates a high pixel value in the visualization results, while the blue color indicates a low pixel value.

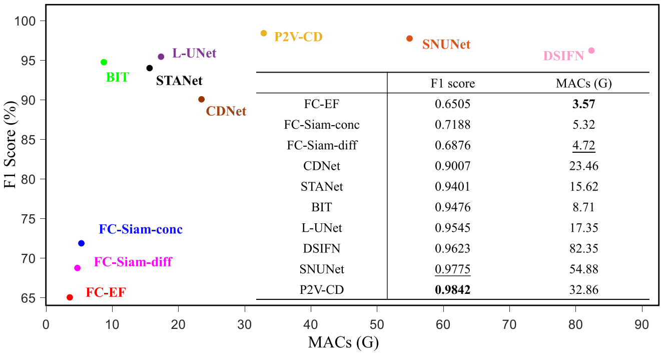
a lower PCC with the ground-truth. Later, with the integration of the temporal features from the lateral connections, the fused features gain more change information, and the third S-Block of the spatial encoder then yields more semantic features. For the P2V-NoTE model, however, the spatiotemporal coupling exists throughout the encoding process, such that the network has to simultaneously watch both space and time. As a result, the features extracted by the shallower layers fail to capture more spatial details, and the features extracted by the deeper layers are less relevant to the ground-truth change map. Based on the above analysis, we can conclude that the proposed P2V-CD functionally decouples the spatial and temporal aspects of the CD task, and we believe it is a key ingredient for the success of our approach on the three datasets.

F. Runtime Evaluation

In this section, we benchmark the proposed P2V-CD with nine SOTA DLCD methods by considering the trade-off between the detection accuracy and the computational cost. The MACs are reported from the inference on

TABLE VI
STATISTICS OF THE FEATURE MAP VISUALIZATION RESULTS ON THE SVCD DATASET

	P2V-NoTE		P2V-CD	
	S-Block 1	S-Block 3	S-Block 1	S-Block 3
TV	2885.50	266.03	3814.88 (+32.21%)	225.79 (−15.13%)
PCC	0.1305	0.1797	0.0623 (−47.74%)	0.2448 (+36.23%)



	F1 score	MACs (G)
FC-EF	0.6505	**3.57**
FC-Siam-conc	0.7188	5.32
FC-Siam-diff	0.6876	4.72
CDNet	0.9007	23.46
STANet	0.9401	15.62
BIT	0.9476	8.71
L-UNet	0.9545	17.35
DSIFN	0.9623	82.35
SNUNet	0.9775	54.88
P2V-CD	**0.9842**	32.86

Fig. 11. F1 score vs. MACs. We use different colors to denote different CD methods. The proposed P2V-CD method achieves the highest F1 score with a moderate number of MACs.

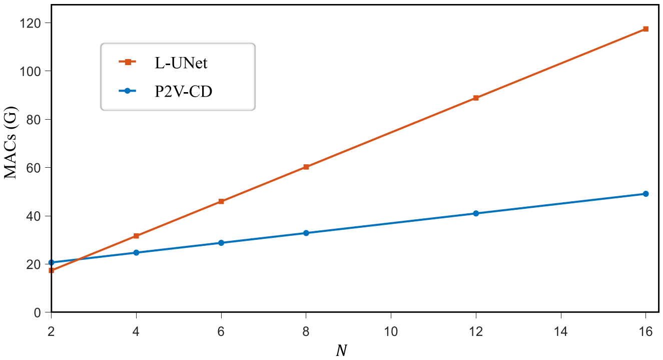


Fig. 12. MACs vs. the number of video frames N. The proposed P2V-CD method shows a more gradual upward trend in comparison to the L-UNet method.

a 256 × 256 image, and the F1 scores are calculated on the test set of the SVCD dataset. The scatter plot of F1 score vs. MACs is drawn in Fig. 11. As can be observed from Fig. 11, DSIFN and SNUNet obtain high F1 scores, but they require a huge amount of computation. L-UNet, STANet, BIT, and CDNet have moderate computational overhead, yet they fail to achieve the top performance. FC-Siam-conc, FC-Siam-diff, and FC-EF use the least computational resources, but their detection accuracy lags far behind the other methods. In contrast to these methods, our approach, P2V-CD, surpasses all its counterparts in terms of the F1 score, without being encumbered by an enormous computational complexity.

To further verify the superiority of using 3D CNNs over RNNs in terms of efficiency, we compare the MACs of P2V-CD and L-UNet provided different lengths of input sequences. The number of video frames is denoted as N. The resulting curves are displayed in Fig. 8. Both curves show a linear relationship of the MACs and N, but with different slopes. As the number of input images increases, a steep rise in terms of MACs can be observed from the L-UNet curve. On the contrary, the growth of MACs for P2V-CD is more modest, although from a higher starting point at $N = 2$. In the case of $N = 16$, P2V-CD have 49.12 GMACs, which is less than a half of L-UNet's 117.47 GMACs. An additional bonus of using 3D CNNs in favor of RNNs is that 3D convolutions process data in a parallel manner, which greatly speeds up the training and inference of the CD model.

V. CONCLUSION

In this paper, the detection of change is framed as a dense classification task for a transition video. Based on this idea, we have established a novel CD framework named P2V-CD, which automatically constructs pseudo video frames from the bi-temporal image pair and effectively learns the change information in a space-time decoupled manner. First, a new CD paradigm was developed, which provides an innovative perspective on the CD problem and facilitates the full utilization of temporal information. Second, the spatiotemporal coupling issue with respect to the CD encoders was discussed and addressed for the first time. Third, the TAM was adopted as a powerful feature adapter to bridge the gap between the spatial and temporal features. We have corroborated the effectiveness of our solution through comprehensive experiments. Our network achieved SOTA performance on three public datasets and succeeds in narrowing the gulf between video understanding and CD. However, we also note that the proposed P2V-CD performs less efficiently than pure 2D convolution-based models. In the future, an important direction for improvement is to make up for the shortcoming of P2V-CD in terms of computational complexity. Additionally, it will be interesting and promising to investigate the possibility of adapting latest advances in video understanding to the CD task, thereby propelling these two fields toward an in-depth integration.

ACKNOWLEDGMENT

The authors would like to thank the providers of the public datasets for their effort in change detection research. They would also like to thank the editor and the anonymous reviewers for their comments and suggestions.

REFERENCES

[1] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 609–630, Mar. 2013.

[2] A. A. Abuelgasim, W. D. Ross, S. Gopal, and C. E. Woodcock, "Change detection using adaptive fuzzy neural networks," *Remote Sens. Environ.*, vol. 70, no. 2, pp. 208–223, Nov. 1999.

[3] A. Frick and S. Tervooren, "A framework for the long-term monitoring of urban green volume based on multi-temporal and multi-sensoral remote sensing data," *J. Geovisualization Spatial Anal.*, vol. 3, no. 1, pp. 1–11, Jun. 2019.

[4] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin, "Review article digital change detection methods in ecosystem monitoring: A review," *Int. J. Remote Sens.*, vol. 25, no. 9, pp. 1565–1596, 2004.

[5] O. A. El-Kawy, J. Rød, H. Ismail, and A. Suliman, "Land use and land cover change detection in the western Nile delta of Egypt using remote sensing data," *Appl. Geography*, vol. 31, no. 2, pp. 483–494, Apr. 2011.

[6] D.-H. Kim et al., "Global, landsat-based forest-cover change from 1990 to 2000," *Remote Sens. Environ.*, vol. 155, pp. 178–193, Dec. 2014.

[7] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[8] P. Qin, Y. Cai, J. Liu, P. Fan, and M. Sun, "Multilayer feature extraction network for military ship detection from high-resolution optical remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11058–11069, 2021.

[9] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[10] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[11] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7 nos. 3–4, pp. 197–387, 2013.

[12] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.*, vol. 12, no. 10, p. 1688, Jan. 2020.

[13] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 982–986, Jun. 2019.

[14] F. I. Diakogiannis, F. Waldner, and P. Caccetta, "Looking for change? Roll the dice and demand attention," *Remote Sens.*, vol. 13, no. 18, p. 3707, Sep. 2021.

[15] Y. Zhang, L. Fu, Y. Li, and Y. Zhang, "HDFNet: Hierarchical dynamic fusion network for change detection in optical aerial images," *Remote Sens.*, vol. 13, no. 8, p. 1440, Apr. 2021.

[16] 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.

[17] 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.

[18] 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.

[19] H. Chen, C. Wu, B. Du, and L. Zhang, "Deep Siamese multi-scale convolutional network for change detection in multi-temporal VHR images," in *Proc. 10th Int. Workshop Anal. Multitemporal Remote Sens. Images (MultiTemp)*, Aug. 2019, pp. 1–4.

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

[21] B. Fang, G. Chen, G. Ouyang, J. Chen, R. Kou, and L. Wang, "Content-invariant dual learning for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

[22] T. Zhang, F. Gao, J. Dong, and Q. Du, "Remote sensing image translation via style-based recalibration module and improved style discriminator," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[23] M. Zanetti, F. Bovolo, and L. Bruzzone, "Rayleigh-rice mixture parameter estimation via EM algorithm for change detection in multispectral images," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5004–5016, Dec. 2015.

[24] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based Markov random field model," *IEEE Trans. Image Process.*, vol. 29, pp. 757–767, 2020.

[25] 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)*, Piscataway, NJ, USA, Oct. 2018, pp. 4063–4067.

[26] W. Wu, D. He, T. Lin, F. Li, C. Gan, and E. Ding, "MVFNet: Multi-view fusion network for efficient video recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2943–2951.

[27] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15193–15202.

[28] L. R. Medsker and L. C. Jain, "Recurrent neural networks," *Design Appl.*, vol. 5, pp. 64–67, Dec. 2001.

[29] 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, Jul. 2020.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7651–7668, Sep. 2021.

[32] V. Sainte Fare Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12325–12334.

[33] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[34] X. Shu, L. Zhang, G.-J. Qi, W. Liu, and J. Tang, "Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3300–3315, Jun. 2022.

[35] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5281–5292, Aug. 2022.

[36] 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, May 2020.

[37] 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.

[38] V. Fontana, G. Singh, S. Akrigg, M. Di Maio, S. Saha, and F. Cuzzolin, "Action detection from a robot-car perspective," 2018, *arXiv:1807.11332*.

[39] C.-Y. Wu and P. Krahenbuhl, "Towards long-form video understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1884–1894.

[40] C. Wu, H. Chen, B. Du, and L. Zhang, "Unsupervised change detection in multitemporal VHR images based on deep kernel PCA convolutional mapping network," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12084–12098, Nov. 2022.

[41] Z. Wang, Z. Gao, L. Wang, Z. Li, and G. Wu, "Boundary-aware cascade networks for temporal action segmentation," in *Computer Vision—ECCV 2020* (Lecture Notes in Computer Science), vol. 12370, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 34–51.

[42] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[43] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 1. Cambridge, MA, USA: MIT Press, 2014, pp. 568–576.

[44] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6202–6211.

[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[47] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[48] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[49] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, vol. 38, G. Lebanon and S. V. N. Vishwanathan, Eds. San Diego, CA, USA: PMLR, 2015, pp. 562–570.

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

[51] 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.

[52] 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," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42-2, pp. 565–571, May 2018.

[53] 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.

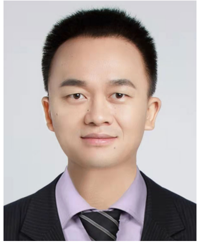
[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[55] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Auto. Robots*, vol. 42, no. 7, pp. 1301–1322, 2018.

[56] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Manhui Lin received the B.S. degree in space physics from Wuhan University, China, in 2019, where he is currently pursuing the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing.

His research interests include remote sensing change detection, image processing, and deep learning.



Guangyi Yang received the Ph.D. degree in electronic engineering from Wuhan University, China, in 2018.

He is currently a Senior Engineer with the School of Electronic Information, Wuhan University. His research interests include machine vision and image processing.

Hongyan Zhang (Senior Member, IEEE) received the B.S. degree in geographic information systems and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, China, in 2005 and 2010, respectively.

Since 2016, he has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. He is a Young Chang-Jiang Scholar appointed by the Ministry of Education of China. He scored 1st in the Track Challenge of the 2019 and 2021 Data Fusion Contest organized by the IEEE Image Analysis and Data Fusion Technical Committee. He has authored/coauthored more than 130 research articles. His research interests include image reconstruction for quality improvement, hyperspectral information processing, and agricultural remote sensing.

Dr. Zhang was the session chair of several international conferences. He serves as an Associate Editor for *Photogrammetric Engineering and Remote Sensing* and *Computers and Geosciences*. He is a Reviewer for more than 30 international academic journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, and *ISPRS Journal of Photogrammetry and Remote Sensing*.