ISPRS Journal of Photogrammetry and Remote Sensing 202 (2023) 599–609

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

An attention-based multiscale transformer network for remote sensing image change detection

Wei Liu [a,b], Yiyuan Lin [a], Weijia Liu [a], Yongtao Yu [c,*], Jonathan Li [d]

[a] School of Software, East China Jiaotong University, Nanchang 330013, China
[b] Thinvent Digital Technology Co., Ltd., Nanchang, 330096, China
[c] Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223003, China
[d] Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Keywords:
Change detection
Attention mechanism
Transformer
Multiscale

ABSTRACT

The bi-temporal change detection (CD) is still challenging for high-resolution optical remote sensing data analysis due to various factors such as complex textures, seasonal variations, climate changes, and new requirements. We propose an attention-based multiscale transformer network (AMTNet) that utilizes a CNN-transformer structure to address this issue. Our Siamese network based on the CNN-transformer architecture uses ConvNets as the backbone to extract multiscale features from the raw input image pair. We then employ attention and transformer modules to model contextual information in bi-temporal images effectively. Additionally, we use feature exchange to bridge the domain gap between different temporal image domains by partially exchanging features between the two Siamese branches of our AMTNet. Experimental results on four commonly used CD datasets – CLCD, HRSCD, WHU-CD, and LEVIR-CD – demonstrate the effectiveness and efficiency of our proposed AMTNet approach. The code for this work will be available on GitHub.[1]

1. Introduction

Bi-temporal change detection (CD) is a crucial task in remote sensing (RS), which involves comparing and identifying changes between co-registered RS images of the same area at different times. It has numerous applications, including disaster assessment, urban planning, agricultural surveys, resource management, and environmental monitoring. With the rapid development of Earth observation technology, vast numbers of high-resolution optical RS images have drawn significant attention to CD technology. However, this emergence also presents new challenges for CD technology due to various factors such as complex textures, seasonal variations, climate changes, and evolving requirements. Despite progress made with deep learning techniques on large-scale high-resolution RS images, bi-temporal CD remains one of the most challenging tasks for analyzing such data.

In recent decades, researchers have designed methods for optical RS image CD to address these challenges. Traditional CD methods with handcrafted features can achieve good results in simple scenes but typically perform poorly in complex scenes. Deep learning-based algorithms perform better than traditional counterparts because they can learn discriminant features from vast amounts of high-quality samples. Among these deep learning-based algorithms are those based on deep convolutional neural networks (ConvNets) or transformer networks, which perform better.

Deep ConvNets have been extensively used in CD to extract discriminative features recently. These deep feature extractors include classical ConvNets and their extended architectures, such as ResNet (He et al., 2016), UNet (Ronneberger et al., 2015), and HRNet (Sun et al., 2019). There are two commonly used feature extraction strategies: single-branch structures (Gao et al., 2020; Wang et al., 2018) and Siamese networks (Zhan et al., 2017; Wang et al., 2022). The single-branch CD networks adopt an early fusion strategy to fuse the input images before inputting them into the CD network. The Siamese network employs a late fusion strategy that typically fuses features extracted from two independent subnetworks. The Siamese structure has been more commonly used in recent years because it performs better than single-branch structures.

Deep ConvNets need to model context information in both spatial and temporal domains to capture changes in RS images effectively. Various methods have been developed to integrate feature aggregation or attention mechanisms into ConvNets, further improving the CD performance. The single-branch structure typically uses concatenation, difference, or summation operations for image-level feature fusion. On

* Corresponding author.
E-mail address: allennessy@hyit.edu.cn (Y. Yu).

[1] https://github.com/linyiyuan11/AMT_Net

https://doi.org/10.1016/j.isprsjprs.2023.07.001
Received 7 February 2023; Received in revised form 30 May 2023; Accepted 1 July 2023
Available online 15 July 2023

the other hand, the dual-branch Siamese structure usually employs single-scale (Zhan et al., 2017; Chen et al., 2019b, 2020; Mesquita et al., 2019; Liu et al., 2019; Xiang et al., 2021) or multiscale (Bao et al., 2020; Zhang and Shi, 2020; Chen et al., 2019a; Lin et al., 2021; Chen et al., 2022) feature fusion strategies.

Attention mechanisms can enhance feature representation by directing the network's focus toward information associated with changed areas. An attention-based CD network automatically highlights important information related to changed areas and suppresses features associated with unchanged areas in positions or channels. Consequently, recent works (Wang et al., 2022) have incorporated attention mechanisms into CD tasks. Typically, these attention mechanisms are implemented in three ways: spatial attention (Zhang et al., 2020; Peng et al., 2020), channel attention (Liu et al., 2020; Jiang et al., 2020), and self-attention (Chen and Shi, 2020; Zhou et al., 2022). Although previous studies have shown promising results using self-attention mechanisms for modeling long-range dependencies, they are computationally inefficient.

The transformer structure has gained attention in computer vision (CV) tasks such as image classification, semantic segmentation, and object detection. It efficiently models global contextual information through an encoder–decoder architecture compared to pure ConvNets. This success has motivated the development of a few CD algorithms based on transformers with impressive results (Chen et al., 2021; Bandara and Patel, 2022; Wang et al., 2022; Liu et al., 2022). However, while the transformer structure is widely used in NLP tasks, its application in CD needs improvement. Specifically, further research is needed to combine multiscale strategy and attention mechanism with the transformer structure for better performance.

Inspired by recent advancements in CD, we developed an attention-based multiscale transformer network (AMTNet) that combines the strengths of ConvNets, transformers, multiscale modules, and attention mechanisms. The AMTNet is a Siamese Network that utilizes the CNN-transformer structure with ResNet (He et al., 2016) as its backbone to extract multiscale features from input image pairs. We then employ attention and transformer modules to model contextual information in bi-temporal images. Additionally, we use feature exchange to bridge the domain gap between different temporal image domains by partially exchanging features between the two Siamese branches.

The following summarizes the main contributions of this paper.

1. To emphasize important information in channels or positions automatically, the proposed framework effectively integrates the attention mechanisms and the transformer, combining the advantages of ConvNets, transformers, multiscale, and attention mechanisms.
2. To bridge the domain gap between different temporal image domains, we apply feature exchange to our AMTNet and analyze the impact of different settings of the feature exchange module on the performance of the proposed framework.
3. We devise a channel attention module to make the multiscale CD network more focused on channels that significantly impact the change analysis.
4. We performed comprehensive experiments on four widely used CD datasets: CLCD (Liu et al., 2022), HRSCD (Daudt et al., 2019), WHU-CD (Ji et al., 2018), and LEVIR-CD (Chen and Shi, 2020). Our results demonstrate that our CD framework outperforms the most advanced CD frameworks currently available, and effectively improves the representation of changed features.

2. Related work

Recently, many CD works have achieved impressive results based on feature fusion or transformers. Most of these studies for CD tasks focus on feature aggregation, attention mechanisms, and transformers. In this section, we introduce the relevant work from three aspects: feature aggregation, attention mechanism, and transformer.

2.1. Feature aggregation

Modeling the context information in both spatial and temporal domains is crucial to capture changes in RS images. Many efforts have been made to model the context information to integrate feature aggregation into ConvNets. The single-branch structure typically performs feature fusion directly at the image level using operations such as difference, concatenation, or summation. There are two methods of feature fusion for the Siamese network structure (Jiang et al., 2022), including single-scale fusion (Chen et al., 2019b; Zhan et al., 2017; Chen et al., 2020; Mesquita et al., 2019; Liu et al., 2019; Xiang et al., 2021) and multiscale fusion (Bao et al., 2020; Xuan et al., 2021; Zhang and Shi, 2020; Chen et al., 2019a; Wang et al., 2022).

Single-scale fusion usually fuses features from the top level of the two Siamese branches. Multiscale Siamese networks typically fuse features at multiple levels in a low-to-high manner. In RS images, changed objects are usually irregular and multiscale. Compared with single-scale fusion, multiscale fusion combining the hierarchical feature maps can efficiently detect change regions at various scales and achieve good results. Compared with the shallow features of neural networks, the deep features in the neural networks contain richer semantic information. Because of the semantic difference between deep and shallow features, directly using logical operations to fuse features at different levels can easily confuse features. Moreover, most existing multiscale models cannot model contextual relationships, which are crucial for identifying changes of interest in RS images.

2.2. Attention mechanism

An attention-based CD network can automatically highlight important information related to the changed areas and suppress features associated with unchanged areas in positions or channels. Therefore, various CD works have introduced attention mechanisms into CD tasks recently. These attention mechanisms are typically implemented in three ways, including spatial attention (Zhang et al., 2020; Peng et al., 2020), channel attention (Liu et al., 2020; Jiang et al., 2020; Fang et al., 2021), and self-attention (Chen et al., 2020; Chen and Shi, 2020).

Attention-based methods can effectively model context information compared with purely convolution-based approaches. DTCDSCN (Liu et al., 2020) designs a dual attention module to enhance the feature representation. DTCDSCN combines the advantages of channel and spatial attention well. To reconstruct the change map, DSIFN/IFN (Zhang et al., 2020) uses image difference features to fuse multiscale deep features of the input bi-temporal images with attention modules. SNUNet (Fang et al., 2021) models contextual information and improves the representation of intermediate features with ensemble channel attention. In order to provide adaptive weight for feature fusion, MFP-Net (Xu et al., 2021) utilizes a channel attention algorithm to generate a channel-wise weight vector. There are also some efforts (Chen et al., 2020; Wang et al., 2022) to combine spatial attention and channel attention to improve feature representation.

However, it is still difficult for most of these methods to model the long-range context information in space–time. The existing attention-based CNN methods either employ attention mechanisms to each temporal image/feature separately to enhance its feature representation (Liu et al., 2020) or directly utilize a channel or spatial attention module to reweight the fused features (Jiang et al., 2020). Some recent works design self-attention-based architectures (Chen et al., 2020; Zhou et al., 2022; Chen and Shi, 2020) to model the contextual-semantic relations between arbitrary pairs of pixels in space–time and obtain promising results. However, these algorithms are computationally inefficient. Their computational complexity increases exponentially as the number of pixels increases. Changed objects in RS images vary in size and type. It is often difficult to set the appropriate patch size for the best performance only using attention mechanisms.

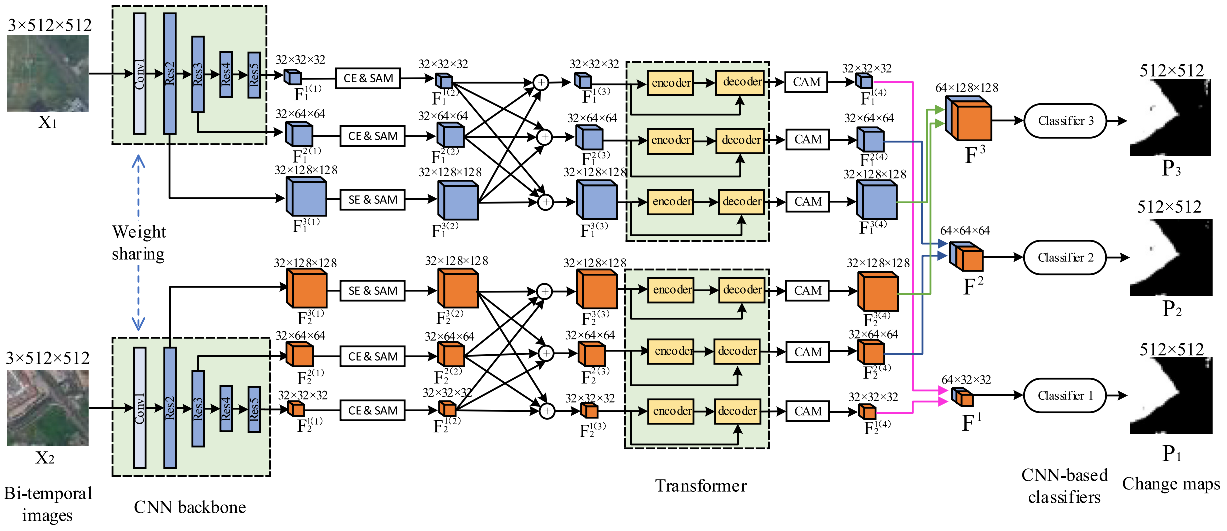

Fig. 1. The framework of the proposed scheme. The proposed framework efficiently combines attention mechanisms and transformers, leveraging the benefits of ConvNets, multiscale processing, and attention mechanisms. For more details, please view in color and zoom in.

2.3. Transformer-based network

Transformers were first introduced to solve problems for sequence-to-sequence learning and have been increasingly widely used in the field of natural language processing (NLP). Transformer modules have the strong ability to model long-range context dependencies with ease. Transformer-based CD methods have shown comparable or even better performance than CNN-based counterparts in a series of CV tasks in recent years, including image classification (Vaswani et al., 2017; Wu et al., 2020), object detection (Ding et al., 2019; Carion et al., 2020), and semantic segmentation (Strudel et al., 2021; Zheng et al., 2021).

Because of its remarkable performance on NLP and CV tasks, the transformer structure has received increasing attention in the RS community to improve performance on various RS data analysis tasks, such as image time-series classification (He et al., 2019; Yuan and Lin, 2020), scene classification (Bazi et al., 2021), hyperspectral image classification (Li et al., 2020), and CD (Liu et al., 2022; Wang et al., 2022). The transformer structure typically takes tokens or patches as inputs and makes the tokens or patches interact with each other and figure out where more attention is needed. The CD network BIT (Chen et al., 2021) models context relationships by incorporating a feature differencing-based network with a transformer module. It encodes the input image into several patches that contain rich contextual information. PSTNet (Song et al., 2022) gradually extracts changed parts in the image through an iterative sampling method by continuously extracting and optimizing feature information. MSCANet (Liu et al., 2022) first uses CNN backbones to extract hierarchical features from the bi-temporal pair and then uses a transformer module to further model and aggregate semantic information. MTCNet (Wang et al., 2022) designs a multiscale transformer for obtaining features at different scales in bi-temporal images. Changer (Fang et al., 2022) is a Siamese network, extracts multilayered features from the input images and then uses feature change operations to exchange features between the two branches of the CD network.

Although the transformer structure has achieved impressive results in remote sensing CD tasks, its application in CD still needs improvement. In this work, we combine the advantages of ConvNets, transformers, multiscale, and attention mechanisms to enhance the representation of the change features. Extensive experiments on four CD datasets, CLCD (Liu et al., 2022), HRSCD (Daudt et al., 2019), WHU-CD (Ji et al., 2018), and LEVIR-CD (Chen and Shi, 2020), demonstrate the effectiveness of our method.

3. The proposed method

3.1. Overview

As shown in Fig. 1, our AMTNet is a Siamese structure combining CNN, multiscale, transformer, and the attention mechanism. It uses ResNet (He et al., 2016) as the backbone to extract multiscale features from the raw input image pair. Then, it uses the attention and transformer modules to further model contextual information in bi-temporal images. In addition, we use feature exchange to bridge the domain gap between different temporal image domains by partially exchanging bi-temporal features between the two Siamese branches of the network.

Specifically, let I_1 and I_2 denote images of the same area taken at two different times, respectively. Let $X_i(i \in \{1,2\}) \in \mathbb{R}^{3\times H\times W}$ represent the raw feature map of image $I_i(i \in \{1,2\})$. The CNN backbones of two subnetworks of the CD network share the same weights. The following summarizes the process of the AMTNet:

1. First, for image $X_i(\forall i \in \{1,2\}) \in \mathbb{R}^{3\times H\times W}$, three feature maps $F_i^{1(1)}$, $F_i^{2(1)}$ and $F_i^{3(1)}$ with different scales are extracted from the ResNet backbone.
2. Next, the above feature map $F_i^{j(1)}(\forall j \in \{1,2,3\})$ is partially exchanged with the feature of the same scale from the other branch of the Siamese network, and then fed into a spatial attention module (SAM), and the feature map $F_i^{j(2)}$ is obtained.
3. Then, the feature map $F_i^{j(2)}$ is merged with the other two feature maps derived from the input image X_i by sampling and addition, and the fused feature map $F_i^{j(3)}$ is obtained.
4. Afterwards, $F_i^{j(3)}$ is input into the transformer and the channel attention module (CAM) successively, and the feature map $F_i^{j(4)}$ is obtained.
5. Finally, feature maps of the same scale from the two subnetworks are concatenated pairwise along the channel dimension and input into the corresponding CNN-based classifiers to get three predicted change maps P_1, P_2, and P_3. Only P_1 will be used as the predicted change map during the testing process.

Different modules of this algorithm can complement each other. Specifically, the attention mechanisms can enhance the feature representation by making the network focus on information related to the changed areas. The transformer module can handle long-range dependencies with ease. The multiscale structure can derive features of various scales, which helps detect changed objects of diverse scales. Because of the semantic difference between deep and shallow features,

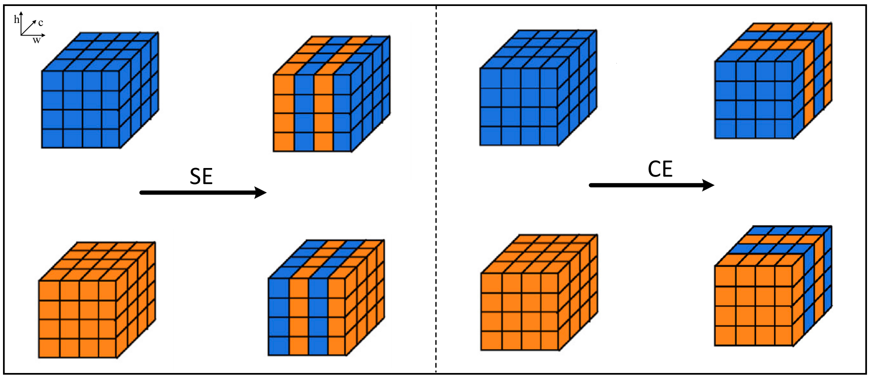

Fig. 2. Illustration of feature exchange. The left and right subgraphs represent spatial exchange and channel exchange, respectively. In both subgraphs, two features on the left are partially exchanged to obtain the corresponding two features on the right.

directly using logical operations to fuse features at different levels can easily cause feature confusion. The proposed SAM and the CAM can effectively emphasize the informative regions and channels, respectively. The following introduces the critical components of the AMTNet in detail.

3.2. Multiscale CNN backbone

As shown in Fig. 1, the AMTNet employs ResNet (He et al., 2016) with the initial fully connected layer removed as the backbone to extract multiscale features from the input images I_1 and I_2. The ResNet backbone comprises five main blocks, including a 7 × 7 convolutional layer and four residual blocks. For simplicity, these five building blocks will be referred to as Conv1, Res2, Res3, Res4, and Res5, respectively. Res3 and Res4 perform downsampling with a stride of 2. For the input bi-temporal image $X_i(\forall i \in \{1,2\}) \in \mathbb{R}^{H\times W\times C}$, three feature maps $F_i^{1(1)}$, $F_i^{2(1)}$ and $F_i^{3(1)}$ with different scales are extracted from Res2, Res3 and Res5, respectively.

3.3. Feature exchange

Differences in solar illumination, observational weather, sensor, or season often create domain gaps between different temporal image domains. To address this issue, we utilize parameter-free feature exchange (Fang et al., 2022) to partially exchange bi-temporal features with the same scale between the two Siamese branches in either the channel or spatial dimensions. This mixing of features results in a more similar distribution of features between the two branches and helps bridge the domain gap.

As illustrated in Fig. 2, we exchange the elements of the corresponding positions of two feature maps, which are the same size and come from different branches of the Siamese network (such as $F_1^{1(1)}$ and $F_2^{1(1)}$). Feature exchange between the feature maps $F_1^{j(1)}$ and $F_2^{j(1)}$ $(\forall j \in \{1,2,3\})$ in the channel or spatial dimensions can be formulated as:

$$F_i^{j(1)}(n,c,h,w) = \begin{cases} F_i^{j(1)}(n,c,h,w), & M(n,c,h,w) = 0, \\ F_{2^{2-i}}^{j(1)}(n,c,h,w), & M(n,c,h,w) = 1, \end{cases} \tag{1}$$

where n, c, h, and w correspond to the batch, channel, height, and width dimensions, respectively. M is an exchange mask composed of only 1 and 0, indicating whether to perform the feature exchange operation at the corresponding position. We implement feature exchange between features $F_1^{1(1)}$ and $F_2^{1(1)}$, $F_1^{2(1)}$ and $F_2^{2(1)}$, $F_1^{3(1)}$ and $F_2^{3(1)}$, respectively. Intuitively, feature maps with high spatial resolution are more suitable for spatial exchange (SE) in the spatial dimension. For the large-scale feature $F_i^{3(1)}$, we adopt the SE operation. For the small-scale features $F_i^{1(1)}$ and $F_i^{2(1)}$, we adopt the channel exchange (CE) operation in the channel dimension.

3.4. Spatial attention module

With the exchanged feature being mixed, the obtained feature map $F_i^{j(1)}$ passes through the subsequent spatial attention module (SAM) (Woo et al., 2018). The SAM automatically emphasizes the important information related to the feature map $F_i^{j(1)}$ in positions. As presented in Fig. 3, the SAM implements an element-wise multiplication operation on each channel of $F_i^{j(1)}$ with the 2D spatial attention $M_s(F_i^{j(1)}) \in \mathbb{R}^{H\times W}$. Meaningful features associated with the changes in positions are assigned with greater weights. In this way, the SAM effectively highlights the change regions and suppresses irrelevant regions' features in bi-temporal images.

To obtain the spatial attention $M_s(F_i^{j(1)})$, we implement average-pooling and max-pooling operations along the channel axis and then concatenate the results of pooling operations to generate $M_s(F_i^{j(1)})$. Let $MaxPool$ and $AvgPool$ denote max-pooling and average-pooling, respectively. The spatial attention process with respect to $F_i^{j(1)}$ can be formulated as:

$$M_s(F_i^{j(1)}) = \sigma(f([AvgPool(F_i^{j(1)}); MaxPool(F_i^{j(1)})])), \tag{2}$$

where σ represents the Sigmoid function, f represents the convolution kernel operation.

Then, the feature map $F_i^{j(2)}$ is obtained through the SAM as follows:

$$F_i^{j(2)} = F_i^{j(1)} \otimes M_s(F_i^{j(1)}), \tag{3}$$

where $\otimes$ indicates element-wise multiplication. For each channel of the feature map, we use the same weight matrix (Ms) to highlight the information, by broadcasting M_s along the channel dimension during the multiplication.

3.5. Channel attention module

As shown in Fig. 1, $F_i^{j(2)}$ is fused with the other two features of image X_i at different scales successively by sampling and addition to generate the fused feature $F_i^{j(3)}$. After that, the fused feature goes through the transformer module and the CAM to generate the feature map $F_i^{j(4)}$. Transformers utilize encoder and decoder blocks. Various transformer modules (Wang et al., 2022; Liu et al., 2022) are plug-and-play in the proposed CD network. This paper uses the SAM and the transformer to model spatial context information and global context information, respectively. The CAM models the channel context information by highlighting the channels related to the changes. We describe the CAM in detail below.

As illustrated in Fig. 4, the multiple features share the same channel attention M_c for image X_i. To compute the channel attention, first, we fuse feature maps of the same scale of the two Siamese branches by element-wise summation and then apply max-pooling along the

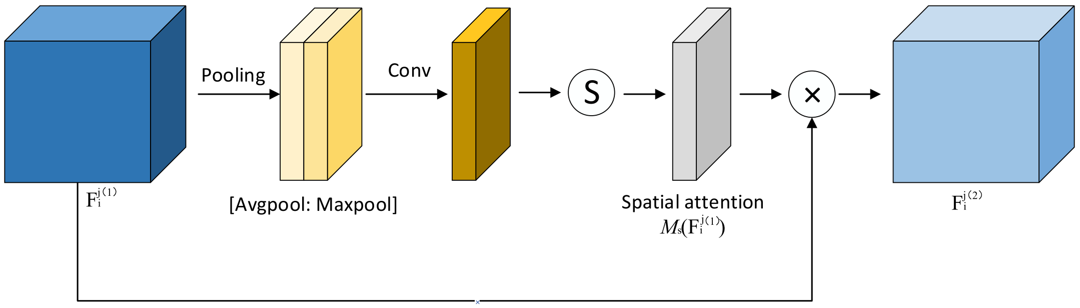

Fig. 3. Illustration of the spatial attention process. The role of the SAM is to automatically emphasize the important information related to the feature map $F_i^{j(1)}$ in positions.

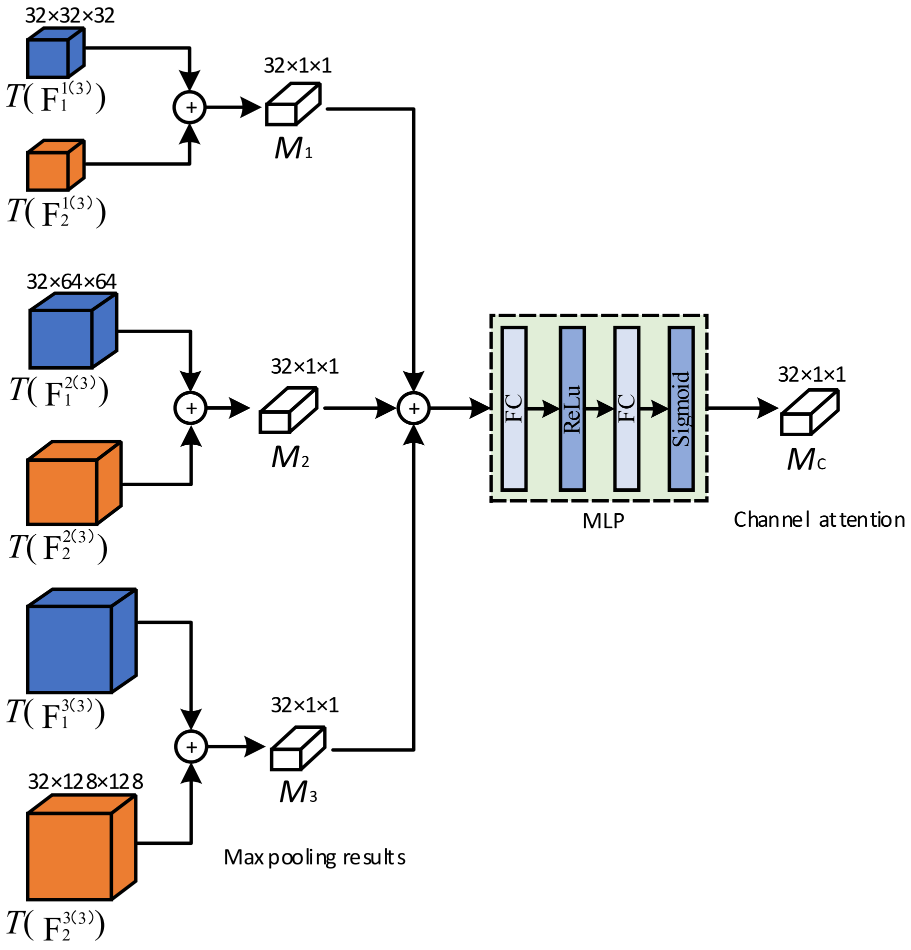

Fig. 4. Description of the channel attention module (CAM). The CAM models the contextual information by highlighting the channels related to the changes.

spatial dimension of the fused results. Next, we employ element-wise summation again to merge the multiscale results of the max-pooling operation and pass the fused result through a multi-layer perception (MLP) to obtain the channel attention M_c. The MLP consists of a full convolution layer followed by a ReLU activation function and a full convolution layer followed by a Sigmoid activation function. The following formulates the process in detail.

Let $T(F_i^{j(3)})$ denote the feature map obtained by passing $F_i^{j(3)}$ through the transformer module. The result of max-pooling of the fused feature of $T(F_1^{j(3)})$ and $T(F_2^{j(3)})$ can be represented as:

$$M_j = MaxPool(T(F_1^{j(3)}) \oplus T(F_2^{j(3)})), \tag{4}$$

where $\oplus$ denotes element-wise summation. Let r denote the reduction ratio of the channels. The channel attention map is:

$$\begin{aligned} M_c &= mlp(M_1 + M_2 + M_3) \\ &= \sigma(\mathbf{W}_2(ReLU(\mathbf{W}_1(M_1 + M_2 + M_3)))), \end{aligned} \tag{5}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C/r \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times C/r}$.

Finally, the feature map $F_i^{j(4)}$ is obtained through the CAM as follows:

$$F_i^{j(4)} = T(F_i^{j(3)}) \otimes M_c. \tag{6}$$

3.6. Overall loss function

As shown in Fig. 1, we concatenate the feature maps of the same scale of the two Siamese branches along the channel dimension. Three fused feature maps F^1, F^2, and F^3 obtained by pairwise concatenation. Then, these three fused feature maps are upsampled to the original image size and fed into their respective CNN-based classifiers, separately. The three classifiers have the same structure. Finally, three predicted maps P_1, P_2, and P_3 are obtained from the CNN-based classifiers. Let Y denote the ground truth, then the overall loss function for the CD task based on the cross-entropy (CE) loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{ce}(P_1, Y) + \mathcal{L}_{ce}(P_2, Y) + \mathcal{L}_{ce}(P_3, Y), \tag{7}$$

where $\mathcal{L}_{ce}(P_1, Y)$ is the CE loss between the predicted change map P_1 and the ground truth Y, likewise $\mathcal{L}_{ce}(P_2, Y)$ and $\mathcal{L}_{ce}(P_3, Y)$.

4. Experiments

4.1. Datasets and metrics

This paper conducts experiments on four popular CD datasets, including CLCD (Liu et al., 2022), HRSCD (Daudt et al., 2019), WHU-CD (Ji et al., 2018), and LEVIR-CD (Chen and Shi, 2020).

CLCD: The CLCD dataset is a public farmland dataset comprising 600 pairs of cropland change samples from Gaofen with size 512 × 512 pixels. These bi-temporal image pairs were taken in 2017 and 2019 in China's Guangdong Province. The spatial resolution of these images is in the range of 0.5 m to 2 m. There are two images and a binary label of cropland change for each sample group. All samples were divided into the training set, the validation set, and the testing set in a ratio of 6:2:2. Therefore, the sizes of the training set, the validation set, and the testing set are 360, 120, and 120, respectively.

HRSCD: The HRSCD dataset consists of 291 HR aerial image pairs of size 10,000 × 10,000 pixels. There is a binary ground truth change map for each bi-temporal pair. These samples were captured from rural and urban areas in Rennes and Caen, French. The original HR samples were cropped without overlapping, resulting in 4398 bi-temporal pairs with the size 512 × 512 pixels for cropland CD. Of the cropped bi-temporal pairs, 2639 are utilized for training, and 880 are utilized for testing.

WHU-CD: It is a public dataset for building CD, containing a pair of HR bi-temporal aerial images with a resolution of 0.2 m and a size of 32,507 × 15,354 pixels. It covers areas that have had earthquakes and have been rebuilt over the years, mainly building renovations. Following the typical setting on this dataset (Bandara and Patel, 2022), we crop the images into non-overlapping patches with a resolution of 256 × 256 pixels. The sizes of the training set, the validation set, and the test set are 5947, 744, and 744, respectively.

LEVIR-CD: The LEVIR-CD dataset is a public building CD dataset captured from the Google Earth. It comprises 637 HR bi-temporal pairs with a size of 1024 × 1024 pixels. These image pairs have a spatial resolution of 0.5 m, and span 5 to 14 years. The dataset includes complex changes in villa dwellings, small garages, high-rise apartments, and large warehouses. All bi-temporal image pairs are annotated utilizing binary labels. The images are cropped into non-overlapping patches with a resolution of 256 × 256 pixels. These patch pairs are randomly split into training/validation/testing sets with sizes of 7120/1024/2048.

To analyze the performance of our AMTNet and the comparison algorithms, we utilize the four most commonly used metrics for CD tasks, including precision, recall, F1-score, and Intersection over Union (IoU).

4.2. Baselines

In this paper, we compare our method with a series of most advanced bi-temporal CD networks, including pure CNN-based networks (Daudt et al., 2018; Xu et al., 2021; Fang et al., 2022), attention-based networks (Liu et al., 2020; Zhang et al., 2020; Fang et al., 2021), and transformer-based networks (Chen et al., 2021; Wang et al., 2022; Liu et al., 2022; Bandara and Patel, 2022).

1. FC-EF (Daudt et al., 2018) is a pure CNN network based on U-Net. It performs fusion at the image level. The CD network takes as input the concatenation of the bi-temporal image pair.
2. FC-Siam-conc (Daudt et al., 2018) is a variation on FC-EF. It is a Siamese Network sharing weights to extract multilevel features from bi-temporal images. Then, the multilayered features are fused using two fully connected layers.
3. FC-Sima-diff (Daudt et al., 2018) is also a variation on FC-EF. It performs fusion at the feature level by using a Siamese network to derive multilevel features for CD.
4. DTCDSCN (Liu et al., 2020) consists of two semantic segmentation sub-networks and a CD sub-network. It is a Siamese network based on a dual attention module using channel and spatial attention to improve the feature representation.
5. DSIFN/IFN (Zhang et al., 2020) merges image difference features with fused multiscale features of the input images through attention modules to reconstruct the change map.
6. SNUNet (Fang et al., 2021) models contextual information and improves the representation of intermediate features with ensemble channel attention. It is a densely connected Siamese network.
7. MFPNet (Xu et al., 2021) is an attention-based Siamese network that uses a multidirectional fusion pathway and an adaptive weighted fusion strategy to fuse features.
8. BiT (Chen et al., 2021) models context relationships by incorporating a feature differencing-based network with a transformer module. It encodes the input image into several patches that contain rich contextual information.
9. ChangeFormer (Bandara and Patel, 2022) uses a hierarchically structured transformer and an MLP as encoder and decoder. It consists of a hierarchical transformer encoder, four feature difference modules, and a light MLP decoder.
10. MTCNet (Wang et al., 2022) is a multiscale CNN transformer-based network that incorporates the attention mechanism. The convolutional block attention module consists of multiple spatial and channel modules.
11. MSCANet (Liu et al., 2022) is a multiscale CNN-transformer network in which a CNN is used to extract hierarchical features. Then, it uses a multiscale context aggregation based on the transformer structure to encode and decode the context information.

On each dataset, for the comparison algorithms that provided results, we cite results directly from the relevant papers; for the comparison algorithms that did not offer results, we use official or commonly used unofficial codes (if available) to reproduce the related algorithms as much as possible. To make the analysis more comprehensive, we also evaluate our method with different ResNet backbones, including ResNet-18 and ResNet-50.

4.3. Implementation details

We implement the proposed AMTNet and all the baselines in PyTorch, utilizing an NVIDIA GeForce RTX 3090 GPU. The AMTNet employs ResNets pre-trained on ImageNet (Krizhevsky et al., 2017) as the CNN backbone. The transformer module can be any commonly used transformer structure. Without losing generality, we use a transformer

Table 1
Comparison results on the CLCD dataset (%).

Method	CLCD			
	P	R	F1	IoU
FC-EF* (Daudt et al., 2018)	71.7	47.6	57.22	40.07
FC-Siam-conc* (Daudt et al., 2018)	73.27	52.91	61.45	44.35
FC-Sima-diff* (Daudt et al., 2018)	64.26	52.33	57.69	40.54
DTCDSCN (Liu et al., 2020)*	54.49	66.23	59.79	42.64
DSIFN/IFN (Zhang et al., 2020)	**79.07**	63.79	70.61	54.58
SNUNet*(Fang et al., 2021)	70.82	62.37	66.32	49.62
MFPNet (Xu et al., 2021)	76.42	60.74	70.22	54.11
BiT (Chen et al., 2021)	61.42	62.75	62.08	45.01
ChangeFormer (Bandara and Patel, 2022)	69.11	51.75	59.18	42.03
MTCNet (Wang et al., 2022)	–	–	–	–
MSCANet (Liu et al., 2022)	75.36	**67.64**	**71.29**	**55.39**
AMTNet-18 (ours)	**75.93**	**71.92**	**73.87**	**58.57**
AMTNet-50 (ours)	**78.64**	**75.06**	**76.81**	**62.35**

The symbol "*" indicates unofficial re-implemented results. The symbol "–" indicates that the relevant algorithm does not report the result and has no official or commonly used unofficial code. Color convention: **best**, **2nd-best**, and **3rd-best**.

Table 2
Comparison results on the HRSCD dataset (%).

Method	HRSCD			
	P	R	F1	IoU
FC-EF*(Daudt et al., 2018)	72.75	50.3	59.48	42.33
FC-Siam-conc*(Daudt et al., 2018)	72.23	47.53	57.34	40.19
FC-Sima-diff*(Daudt et al., 2018)	74.19	46.19	55.59	38.49
DTCDSCN*(Liu et al., 2020)	**75.79**	48.83	59.39	42.24
DSIFN/IFN (Zhang et al., 2020)	**77.0**	54.27	63.66	46.7
SNUNet*(Fang et al., 2021)	70.53	53.63	60.93	43.81
MFPNet (Xu et al., 2021)	**76.42**	54.98	63.95	47.01
BiT (Chen et al., 2021)	71.3	52.23	60.30	43.16
ChangeFormer (Bandara and Patel, 2022)	73.39	52.39	61.16	44.03
MTCNet (Wang et al., 2022)	–	–	–	–
MSCANet (Liu et al., 2022)	70.17	**59.97**	**64.67**	**47.79**
AMTNet-18 (ours)	70.31	**62.07**	**65.93**	**49.18**
AMTNet-50 (ours)	69.31	**65.01**	**67.09**	**50.48**

The symbol "*" indicates unofficial re-implemented results. The symbol "–" indicates that the relevant algorithm does not report the result and has no official or commonly used unofficial code. Color convention: **best**, **2nd-best**, and **3rd-best**.

Table 3
Comparison results on the WHU-CD dataset (%).

Method	WHU-CD			
	P	R	F1	IoU
FC-EF*(Daudt et al., 2018)	80.87	75.43	78.05	64.01
FC-Siam-conc*(Daudt et al., 2018)	68.62	87.30	76.84	62.39
FC-Sima-diff*(Daudt et al., 2018)	70.45	77.62	73.86	58.56
DTCDSCN*(Liu et al., 2020)	82.72	88.44	85.49	74.65
DSIFN/IFN (Zhang et al., 2020)	91.47	81.57	86.36	75.99
SNUNet*(Fang et al., 2021)	83.25	**91.35**	87.11	77.17
MFPNet (Xu et al., 2021)	88.44	89.02	88.73	79.74
BiT (Chen et al., 2021)	83.05	88.80	85.83	75.18
ChangeFormer (Bandara and Patel, 2022)	**92.89**	85.60	88.82	79.89
MTCNet (Wang et al., 2022)	–	91.90	82.65	70.43
MSCANet (Liu et al., 2022)	91.10	89.86	**90.47**	**82.60**
AMTNet-18 (ours)	**91.99**	**89.96**	**90.96**	**83.42**
AMTNet-50 (ours)	**92.86**	**91.99**	**92.27**	**85.64**

The symbol "*" indicates unofficial re-implemented results. The symbol "–" indicates that the relevant algorithm does not report the result and has no official or commonly used unofficial code. Color convention: **best**, **2nd-best**, and **3rd-best**.

Table 4
Comparison results on the LEVIR-CD dataset (%).

Method	LEVIR-CD			
	P	R	F1	IoU
FC-EF*(Daudt et al., 2018)	86.91	80.17	83.40	71.53
FC-Siam-conc*(Daudt et al., 2018)	91.99	76.77	83.69	71.96
FC-Sima-diff*(Daudt et al., 2018)	89.53	83.31	86.31	75.92
DTCDSCN*(Liu et al., 2020)	88.53	86.83	87.67	78.05
DSIFN/IFN (Zhang et al., 2020)	**94.02**	82.93	88.13	78.77
SNUNet*(Fang et al., 2021)	89.18	87.17	88.16	78.83
MFPNet (Xu et al., 2021)	**93.16**	89.08	**91.08**	**83.62**
BiT (Chen et al., 2021)	89.24	**89.37**	89.31	80.68
ChangeFormer (Bandara and Patel, 2022)	**92.05**	88.80	**90.40**	**82.48**
MTCNet (Wang et al., 2022)	–	89.62	90.24	82.22
MSCANet (Liu et al., 2022)	91.30	88.56	89.91	81.66
AMTNet-18 (ours)	90.62	**89.00**	89.80	81.49
AMTNet-50 (ours)	91.82	**89.71**	**90.76**	**83.08**

The symbol "*" indicates unofficial re-implemented results. The symbol "–" indicates that the relevant algorithm does not report the result and has no official or commonly used unofficial code. Color convention: **best**, **2nd-best**, and **3rd-best**.

module similar to MSCANet in the experimental comparison. The main difference is that our method does not use context aggregation between different transformer branches. For the convolution kernel in the SAM, we use a 7 × 7 convolution kernel performed with $padding = 3$. We train the framework with the input size of 512 × 512 pixels or 256 × 256 pixels, using a batch size of 8 and an initial learning rate of 1e-4. We optimize the model's parameters utilizing the AdamW optimizer (Loshchilov and Hutter, 2018) and set the weight decay to 0.01. We perform general data augmentation on the input bi-temporal images. The data augmentation operations include random rotation, vertical flip, and horizontal flip. We set the reduction ratio r of the channel attention module to 1. We train the AMTNet for 100 epochs. Using ResNet-50 as the backbone, the optimization process takes around 1 h, 7 h, 3.5 h, and 4.5 h on CLCD, HRSCD, WHU-CD, and LEVIR-CD, respectively. During the testing phase, no data augmentation operation is applied. We only use P_1 as the predicted change map during testing.

4.4. Comparison results and discussion

4.4.1. Performance on CLCD

As shown in Table 1, our method AMTNet-50 significantly performs better than all the baselines on the CLCD dataset in terms of recall, F1, and IoU. Specifically, AMTNet-50 attains the best recall, F1, and IoU of 75.06%, 76.81%, and 62.35%, respectively. It is challenging to have both precision and recall high simultaneously. However, our AMTNet-50 obtains a very high precision of 78.64%. Our method AMTNet-18 also performs better than all the comparison algorithms according to recall, F1, and IoU. It attains the second-best recall, F1, and IoU of 71.92%, 73.87%, and 58.57%, respectively. Although its recall rate drops to 75.93%, it has the third highest recall rate among these algorithms.

4.4.2. Performance on HRSCD

Table 2 presents the quantitative evaluation results on the HRSCD dataset. As presented in Table 2, similar to the performance on the CLCD dataset, AMTNet-18 and AMTNet-50 show obvious superiority over all the comparison algorithms according to recall, F1, and IoU. Specifically, AMTNet-50 attains the best recall, F1, and IoU of 65.01%, 67.09%, and 50.48%, respectively. AMTNet-18 attains the second-best recall, F1, and IoU of 62.07%, 65.93%, and 49.18%, respectively. AMTNet-50 and AMTNet-18 achieve a precision of 69.31% and 70.31%, respectively.

4.4.3. Performance on WHU-CD

Table 3 presents the comparative results on the WHU-CD dataset. The attention-based method MFPNet and the transformer-based methods can achieve good results on this dataset. Our AMTNet-50 obtains the best recall, F1, and IoU of 91.99%, 92.27%, and 85.64%, respectively. It also gets the second-best precision of 92.86%. AMTNet-18 achieves the second-best F1 and IoU of 90.96% and 83.42%, respectively. It achieves the third-best precision and recall of 91.99% and 89.96%, respectively. The multiscale transformer-based method MSCANet obtains very high precision, recall, F1, and IoU of 91.10%, 89.86%, 90.47%, and 82.60%, respectively.

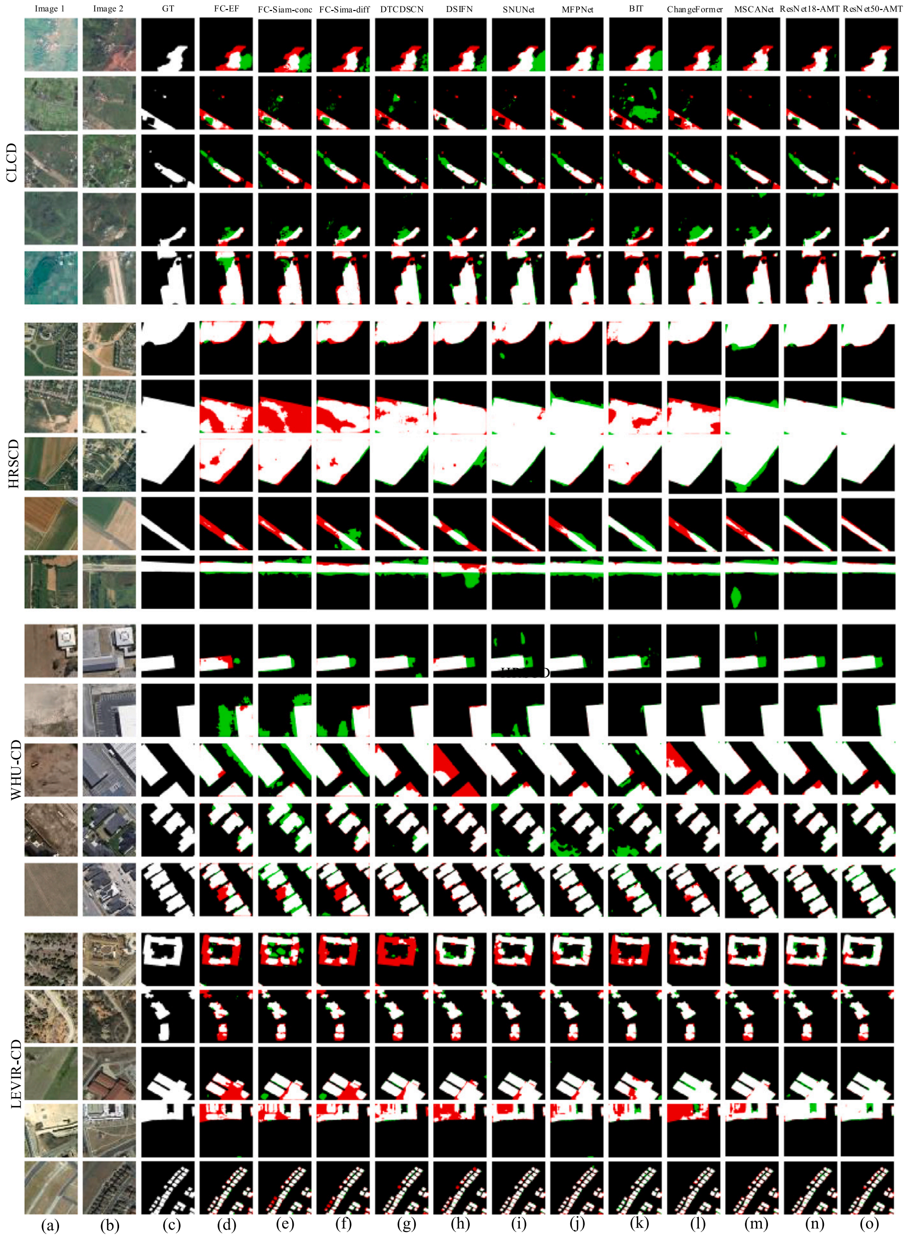

Fig. 5. Qualitative results for the CLCD, HRSCD, WHU-CD, and LEVIR-CD datasets are presented in the following images. True positives (TP) are represented by white pixels, while true negatives (TN) are black. False positives (FP) are green and false negatives (FN) are red. A better result is indicated by fewer red and green pixels. The images from left to right show: (a) Image 1, (b) Image 2, (c) ground truth, followed by the results of various models, including FC-EF, FC-Siam-conc, FC-Sima-diff, DTCDSCN, DSIFN, SNUNet, BIT, ChangeFormer, MFPNet, MSCANet, ResNet18-AMT, and ResNet50-AMT. For more details, please view in color and zoom in.

4.4.4. Performance on LEVIR-CD

Table 4 presents the comparative experimental results on the LEVIR-CD dataset. MFPNet and the transformer-based methods can achieve good results on this dataset. MFPNet performed very well, attaining precision, recall, F1, and IoU of 93.16%, 89.08%, 91.08%, and 83.62%, respectively. Its high performance is attributed to its complex multi-scale feature fusion structure and attention mechanism. Our AMTNet-ResNet obtained very high precision, recall, F1, and IoU of 91.82%, 89.71%, 90.76%, and 83.08%, respectively. Using ResNet as the CNN backbone, its precision, recall, F1, and IoU dropped slightly to 90.62%, 89.0%, 89.8%, and 81.49%, respectively. The transformer-based Siamese network ChangeFormer also performed well, achieving precision, recall, F1, and IoU of 92.05%, 88.8%, 90.40%, and 82.48%, respectively. Like MFPNet, ChangeFormer is also a heavy multiscale Siamese network.

4.4.5. Visualization comparison

To analyze and compare the effects of various CD algorithms, we visually compared different methods on four datasets: CLCD, HRSCD, WHU-CD, and LEVIR-CD. Fig. 5 shows that our AMTNet outperforms the comparison algorithms. Specifically, AMTNet-18 and AMTNet-50 produce significantly fewer red elements than most of the comparative algorithms on CLCD and HRSCD. Additionally, our algorithm can detect unlabeled changes in ground truth as demonstrated by the last bi-temporal image pairs of CLCD and HRSCD in Fig. 5. On building CD datasets WHU-CD and LEVIR-CD, the performance of AMTNet-18 and AMTNet-50 is comparable to or better than that of the best comparison algorithms. Most comparison algorithms struggle to accurately detect changes in irregular buildings with varying sizes; however, as shown in Fig. 5, both our AMTNet-18 and AMTNet-50 can still clearly identify building boundaries.

Table 5
Efficiency comparison analysis with different input sizes.

Method	512 × 512		256 × 256	
	FLOPs (G)	Params (M)	FLOPs (G)	Params (M)
FC-EF (Daudt et al., 2018)	14.29	1.35	3.58	1.35
FC-Siam-conc (Daudt et al., 2018)	21.29	1.55	5.33	1.55
FC-Sima-diff (Daudt et al., 2018)	18.91	1.35	4.73	1.35
DTCDSCN (Liu et al., 2020)	52.83	31.26	13.22	31.26
DSIFN/IFN (Zhang et al., 2020)	329.06	50.44	82.26	50.44
SNUNet (Fang et al., 2021)	219.33	12.03	54.83	12.03
MFPNet (Xu et al., 2021)	514.93	85.97	128.85	85.97
BiT (Chen et al., 2021)	42.37	3.49	8.75	3.49
ChangeFormer (Bandara and Patel, 2022)	230.25	32.03	202.79	41.03
MTCNet (Wang et al., 2022)	–	–	–	5.80
MSCANet (Liu et al., 2022)	59.08	16.42	14.80	16.42
AMTNet-18 (ours)	58.85	16.44	14.71	16.44
AMTNet-50 (ours)	86.23	24.67	21.56	24.67

The symbol "–" indicates that the relevant algorithm does not report the related information, and there is no official or commonly used unofficial code to re-implement the algorithm.

Table 6
Ablation results on the CD datasets (%). FE denotes feature exchange. The symbol "×" indicates that the corresponding module is removed. We use bold fonts to highlight the best results.

Model	Transformer	Multiscale	FE	SAM	CAM	CLCD		HRSCD		WHU		LEVIR	
						F1	IOU	F1	IoU	F1	IOU	F1	IoU
AMTNet	×	✓	✓	✓	✓	76.04	61.35	66.06	49.32	91.27	83.93	90.46	82.57
AMTNet	✓	×	✓	✓	✓	45.27	29.26	49.20	32.62	82.45	70.15	86.90	76.84
AMTNet	✓	✓	×	✓	✓	76.29	61.66	67.00	50.37	91.74	84.73	90.66	82.91
AMTNet	✓	✓	✓	×	✓	75.68	60.87	66.55	49.87	91.94	85.08	90.57	82.77
AMTNet	✓	✓	✓	✓	×	76.50	61.94	66.44	49.74	91.68	84.63	90.47	82.60
AMTNet	✓	✓	✓	✓	✓	**76.81**	**62.35**	**67.09**	**50.48**	**92.27**	**85.64**	**90.76**	**83.08**

4.4.6. *Model efficiency*

To compare model efficiency, we further analyze all comparison algorithms regarding the number of floating point operations (FLOPs) and parameters (Params). We use FLOPs and Params to measure the model's computational and space complexities, respectively. Table 5 presents the two metrics given the image input sizes 512 × 512 pixels and 256 × 256 pixels. As seen from Table 5, FC-EF, FC-Siam-conc, and FC-Sima-diff have the lowest FLOPs and Params. These three pure CNN-based models have very small computational and spatial complexities due to their simple structures. MFPNet, which performs very well on the HSRCD and LEVIR-CD datasets, has the highest numbers of FLOPs and Params because of the complicated multiscale strategy. ChangeFormer has the highest numbers of FLOPs and Params among the transfer-based networks due to the hierarchical transformer encoder. With the input size of 512 × 512 pixels, our AMTNet-18 has 58.85G FLOPs and 16.44M Params. AMTNet-50 has 86.23G FLOPs and 24.67M Params. With the input size of 256 × 256 pixels, our AMTNet-18 has 14.71G FLOPs and 16.44M Params. AMTNet-50 has 21.56G FLOPs and 24.67M Params. Table 5 further indicates that the proposed CD scheme can obtain state-of-the-art performance and have very low FLOPs and Params at the same time.

4.4.7. *Discussion*

Based on the experimental results mentioned above, we have made three observations: (1) Transformer-based CD methods outperform pure CNN-based and attention-based counterparts across all four CD datasets. (2) Our transformer-based method is more effective than BiT, ChangeFormer, MTCNet, and MSCANet as it achieves better F1 and IoU scores. (3) Our algorithm delivers good performance while maintaining high efficiency.

4.5. *Ablation experiments*

We performed ablation studies to confirm the effectiveness of each key component in the proposed method using ResNet-50 as the CNN backbone. To assess the significance of each component, we removed the related module from the CD network. Table 6 shows the experimental results on the four datasets. In addition, we also discussed the impact of pre-training of the ResNet-50 backbone on the ImageNet.

4.5.1. *Transformer*

To assess the effectiveness of the transformer module, we conducted an ablation study by removing it from our network. The results in Table 6 indicate a slight decrease in F1 scores on CLCD, HRSCD, WHU-CD, and LEVIR-CD datasets to 76.04%, 66.06%, 91.27%, and 90.46% respectively with a drop of only 0.77%, 1.03%, 1.00%, and 0.30%. Similarly, there is a reduction in IoU on all four datasets after removing the transformer module. Compared to attention mechanisms, transformers are better equipped to handle long-range dependencies. The comparison results highlight the importance of this module in our network architecture.

4.5.2. *Multiscale*

Remote sensing images often contain objects of varying sizes that have undergone changes. The multiscale structure can extract features of different scales, which is advantageous for detecting these changed objects. To demonstrate the effectiveness of this mechanism, we removed the branches corresponding to the CNN building blocks res2 and res3 from one branch in each of the two subnetworks in the CD network. As a result, its F1 score dropped significantly to 45.27%, 49.20%, 82.45%, and 86.90%, respectively, without multiscale mechanism compared to with it, where it is only reduced by 31.54%, 17.89%, 9.82%, and 3.86%. Similarly, its IoU also decreases by a significant margin when the multiscale mechanism is not used (29.26%,32 62%, 70.15%, and 76.84%) as opposed to when it is used (62.35%, 50.48%, 85.64%, and 83.08%). These experimental results clearly indicate that the multiscale structure has a significant impact on CD network performance and explain why recent successful CD algorithms typically adopt multiscale structures.

Table 7
The effect of different feature exchange settings (%).

Model	Setting			CLCD		HRSCD		WHU		LEVIR	
	Level_1	Level_2	Level_3	F1	IoU	F1	IoU	F1	IoU	F1	IoU
AMTNet	SE	CE	SE	74.13	58.89	65.44	48.63	91.82	84.88	90.24	82.21
AMTNet	CE	SE	SE	76.18	61.53	66.51	49.83	91.85	84.92	90.52	82.68
AMTNet	CE	CE	CE	76.50	61.94	66.04	49.30	91.78	84.81	90.70	82.98
AMTNet	SE	SE	SE	69.24	52.96	65.60	48.81	91.72	84.70	90.40	82.48
AMTNet	CE	CE	SE	**76.81**	**62.35**	**67.09**	**50.48**	**92.27**	**85.64**	**90.76**	**83.08**

Table 8
The effect of ImageNet pre-training for the CNN backbone (%).

Dataset	Pre-training	P	R	F1	IoU
CLCD	w/o	55.54	70.93	62.30	45.24
	w	**78.64**	**75.06**	**76.81**	**62.35**
HRSCD	w/o	**70.57**	58.06	63.71	46.74
	w	69.31	**65.01**	**67.09**	**50.48**
WHU	w/o	81.97	90.41	85.98	75.41
	w	**92.86**	**91.99**	**92.27**	**85.64**
LEVIR	w/o	90.10	87.87	88.97	80.14
	w	**91.82**	**89.71**	**90.76**	**83.08**

4.5.3. Feature exchange

Feature exchange can make the feature distributions of the two branches similar and achieve domain adaptation between the two branches to some extent. The feature exchange module partially exchanges feature in the channel or spatial dimensions between the two Siamese branches of the proposed AMTNet. We conducted ablation experiments by removing the feature exchange operations from our CD network. As seen from Table 6, feature exchange can improve F1 by 0.52%, 0.09%, 0.53%, and 0.10%, respectively. It can improve IoU by 0.69%, 0.11%, 0.91%, and 0.17%, respectively. Table 6 indicates that feature exchange can improve the feature representation of our AMTNet.

In addition, we also evaluated the effect of using different feature exchange settings. As presented in Table 7, Level_1, level_1, and level_3 correspond to the feature maps of the three scales from small to large, respectively. Using the settings of CE, CE, and SE for Level_1, level_2, and level_3, respectively, our AMTNet achieves the best performance on all four datasets. Table 7 indicates that features with high spatial resolution are more suitable to use SE; and vice versa. Based on this observation, we adopt SE to the feature maps $F_1^{3(1)}$. For the other two pairs of feature maps, we adopt the CE operation.

4.5.4. SAM

The SAM automatically emphasizes the vital information related to the feature maps in positions. We experiment with ablation by deleting the SAM from our network. Table 6 demonstrates that the proposed method has a small decrease in IoU and F1 on the four datasets without the SAM. Specifically, its F1 drops by 1.13%, 0.54%, 0.33%, and 0.19% on the four datasets, respectively. Its IoU drops by 1.48%, 0.61%, 0.56%, and 0.31%, respectively. The experimental results show that the SAM is a key component in the CD network.

In addition, we also analyzed the impact of using kernels of different sizes in the SAM. As presented in Table 9, we evaluated three settings: $f^{3\times3}$, $f^{5\times5}$, and $f^{7\times7}$, performed with $padding = 1$, $padding = 2$, and $padding = 3$, respectively. Table 9 indicates that the setting $f^{7\times7}$ with $padding = 3$ achieves the best performance on all four datasets.

4.5.5. CAM

The CAM makes the network focus on channels that greatly impact the change analysis. To validate the importance of CAM in the network, we experimented with ablation by deleting the CAM. As is shown in Table 6, in the absence of CAM, the performance of the AMTNet declines slightly across the four datasets according to both IoU and F1. The CAM can improve F1 by 0.31%, 0.65%, 0.59%, and 0.29% on the four datasets, respectively. It can improve IoU by 0.41%, 0.74%, 1.01%, and 0.48% on the four datasets, respectively. Table 6 shows that the CAM plays an important role in the proposed CD network.

4.5.6. Pre-training on the ImageNet dataset

As presented in Table 8, using the pretrained ResNet-50 backbone can improve the model's overall performance. Compared with the results on the LEVIR-CD dataset, the performance improvement on the CLCD dataset using the pre-trained model is obvious. Specifically, regarding precision, recall, F1, and IoU, the pre-training can achieve 23.10%, 4.13%, 14.51%, and 17.11% improvement on the CLCD dataset, respectively. It can obtain 1.72%, 1.84%, 1.89%, and 2.94% improvement on the LEVIR-CD dataset, respectively. The significant difference in performance improvement between the two datasets may stem from the significant difference in the size of the two training sets of LEVIR-CD and CLCD. The training set sizes for CLCD and LEVIR-CD are 360 and 7120, respectively. With sufficient training data, our AMTNet can achieve high performance on LEVIR-CD without using the pretrained backbone. Therefore, pre-training has a more significant effect on the CLCD dataset.

5. Conclusion

The aim of this paper is to devise an attention-based multiscale CNN-transformer framework that combines the benefits of ConvNets, transformers, multiscale and attention mechanisms. The different modules in this algorithm complement each other. Spatial attention and channel attention can enhance feature representation by focusing on information related to changed areas, while the transformer module handles long-range dependencies with ease. The multiscale structure derives features of various scales. Additionally, we use feature exchange to bridge the domain gap between different temporal image domains by partially exchanging features in either the channel or spatial dimension between the two Siamese branches.

Our method outperforms advanced CD methods based on a series of quantitative and qualitative comparisons on four popular CD datasets. In future research, we plan to apply AMTNet in weakly supervised learning for CD tasks, particularly domain-adaptive ones. We hope our work inspires researchers to explore combining convolutional neural networks, attention mechanisms, and transformers or utilize our model for CD applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62076107, 42177156) and the Natural Science Foundation of Jiangxi Province (20232BAB203057, 20224ACB204021).

Table 9
Effect of using convolution kernels of different sizes in the SAM (%).

Model	Setting		CLCD		HRSCD		WHU		LEVIR	
	kernel	padding	F1	IOU	F1	IOU	F1	IOU	F1	IOU
AMT	3 × 3	1	76.1	61.42	66.21	49.49	91.48	84.3	90.36	82.41
AMT	5 × 5	2	76.33	61.72	66.13	49.4	91.38	84.12	90.56	82.74
AMT	7 × 7	3	**76.81**	**62.35**	**67.09**	**50.48**	**92.27**	**85.64**	**90.76**	**83.08**

References

Bandara, W.G.C., Patel, V.M., 2022. A transformer-based siamese network for change detection. In: IGARSS. pp. 207–210.

Bao, T., Fu, C., Fang, T., Huo, H., 2020. PPCNET: A combined patch-level and pixel-level end-to-end deep network for high-resolution remote sensing image change detection. IEEE Geosci. Remote Sens. Lett. 17 (10), 1797–1801.

Bazi, Y., Bashmal, L., Rahhal, M.M.A., Dayil, R.A., Ajlan, N.A., 2021. Vision transformers for remote sensing image classification. Remote Sens. 13 (3), 516–534.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: ECCV. pp. 213–229.

Chen, H., Qi, Z., Shi, Z., 2021. Remote sensing image change detection with transformers. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Chen, H., Shi, Z., 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sens. 12 (10), 1662–1684.

Chen, H., Wu, C., Du, B., Zhang, L., 2019a. Deep siamese multi-scale convolutional network for change detection in multi-temporal VHR images. In: MultiTemp. pp. 1–4.

Chen, H., Wu, C., Du, B., Zhang, L., Wang, L., 2019b. Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network. IEEE Trans. Geosci. Remote Sens. 58 (4), 2848–2864.

Chen, J., Yuan, Z., Peng, J., Chen, L., Huang, H., Zhu, J., Liu, Y., Li, H., 2020. DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 1194–1206.

Chen, P., Zhang, B., Hong, D., Chen, Z., Yang, X., Li, B., 2022. FCCDN: Feature constraint network for VHR image change detection. ISPRS J. Photogramm. Remote Sens. 187, 101–119.

Daudt, R.C., Le Saux, B., Boulch, A., Gousseau, Y., 2018. Urban change detection for multispectral earth observation using convolutional neural networks. In: IGARSS. pp. 2115–2118.

Daudt, R.C., Le Saux, B., Boulch, A., Gousseau, Y., 2019. Multitask learning for large-scale semantic change detection. Comput. Vis. Image Understand. 187, 102783–102792.

Ding, J., Xue, N., Long, Y., Xia, G.-S., Lu, Q., 2019. Learning roi transformer for oriented object detection in aerial images. In: CVPR. pp. 2849–2858.

Fang, S., Li, K., Li, Z., 2022. Changer: Feature interaction is what you need for change detection. pp. 1–11, arXiv preprint arXiv:2209.08290.

Fang, S., Li, K., Shao, J., Li, Z., 2021. SNUNet-CD: A densely connected siamese network for change detection of VHR images. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Gao, Y., Gao, F., Dong, J., Li, H.-C., 2020. SAR image change detection based on multiscale capsule network. IIEEE Geosci. Remote Sens. Lett. 18 (3), 484–488.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: CVPR. pp. 770–778.

He, J., Zhao, L., Yang, H., Zhang, M., Li, W., 2019. HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers. IEEE Trans. Geosci. Remote Sens. 58 (1), 165–178.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57 (1), 574–586.

Jiang, H., Hu, X., Li, K., Zhang, J., Gong, J., Zhang, M., 2020. PGA-SiamNet: Pyramid feature-based attention-guided siamese network for Remote Sens. orthoimagery building change detection. Remote Sens. 12 (3), 484–504.

Jiang, H., Peng, M., Zhong, Y., Xie, H., Hao, Z., Lin, J., Ma, X., Hu, X., 2022. A survey on deep learning-based change detection from high-resolution remote sensing images. Remote Sens. 14 (7), 1552–1582.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM. 60 (6), 84–90.

Li, Z., Chen, G., Zhang, T., 2020. A CNN-transformer hybrid approach for crop classification using multitemporal multisensor images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 847–858.

Lin, Z.A., Xha, B., Mi, Z.A., Zhen, S.A., Hao, Z.A., 2021. Object-level change detection with a dual correlation attention-guided detector. ISPRS J. Photogramm. Remote Sens. 177, 147–160.

Liu, M., Chai, Z., Deng, H., Liu, R., 2022. A CNN-transformer network with multi-scale context aggregation for fine-grained cropland change detection. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 4297–4301.

Liu, J., Chen, K., Xu, G., Sun, X., Yan, M., Diao, W., Han, H., 2019. Convolutional neural network-based transfer learning for optical aerial images change detection. IEEE Geosci. Remote Sens. Lett. 17 (1), 127–131.

Liu, Y., Pang, C., Zhan, Z., Zhang, X., Yang, X., 2020. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. IEEE Geosci. Remote Sens. Lett. 18 (5), 811–815.

Loshchilov, I., Hutter, F., 2018. Decoupled weight decay regularization. In: ICLR.

Mesquita, D.B., dos Santos, R.F., Macharet, D.G., Campos, M.F., Nascimento, E.R., 2019. Fully convolutional siamese autoencoder for change detection in UAV aerial images. IEEE Geosci. Remote Sens. Lett. 17 (8), 1455–1459.

Peng, X., Zhong, R., Li, Z., Li, Q., 2020. Optical remote sensing image change detection based on attention mechanism and image difference. IEEE Trans. Geosci. Remote Sens. 59 (9), 7296–7307.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241.

Song, X., Hua, Z., Li, J., 2022. PSTNet: Progressive sampling transformer network for remote sensing image change detection. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 8442–8455.

Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation. In: ICCV. pp. 7262–7272.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. NeurIPS 30, 1–15.

Wang, W., Tan, X., Zhang, P., Wang, X., 2022. A CBAM based multiscale transformer fusion approach for remote sensing image change detection. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 6817–6825.

Wang, Q., Zhang, X., Chen, G., Dai, F., Gong, Y., Zhu, K., 2018. Change detection based on faster R-CNN for high-resolution remote sensing images. Remote Sens. Letters. 9 (10–12), 923–932.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: Convolutional block attention module. In: ECCV. pp. 3–19.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P., 2020. Visual transformers: Token-based image representation and processing for computer vision. pp. 1–12, arXiv preprint arXiv:2006.03677.

Xiang, S., Wang, M., Jiang, X., Xie, G., Zhang, Z., Tang, P., 2021. Dual-task semantic change detection for remote sensing images using the generative change field module. Remote Sens. 13 (16), 3336–3350.

Xu, J., Luo, C., Chen, X., Wei, S., Luo, Y., 2021. Remote sens. change detection based on multidirectional adaptive feature fusion and perceptual similarity. Remote Sens. 13 (15), 3053–3056.

Xuan, H.A., Yb, B., Ying, L.A., Cs, B., Qiang, S.B., 2021. High-resolution triplet network with dynamic multiscale feature for change detection on satellite images. ISPRS J. Photogramm. Remote Sens. 177, 103–115.

Yuan, Y., Lin, L., 2020. Self-supervised pretraining of transformers for satellite image time series classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 474–487.

Zhan, Y., Fu, K., Yan, M., Sun, X., Wang, H., Qiu, X., 2017. Change detection based on deep siamese convolutional network for optical aerial images. IEEE Geosci. Remote Sens. Lett. 14 (10), 1845–1849.

Zhang, M., Shi, W., 2020. A feature difference convolutional neural network-based change detection method. IEEE Trans. Geosci. Remote Sens. 58, 7232–7246.

Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., Liu, G., 2020. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS J. Photogramm. Remote Sens. 166, 183–200.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR. pp. 6881–6890.

Zhou, Y., Wang, F., Zhao, J., Yao, R., Chen, S., Ma, H., 2022. Spatial-temporal based multihead self-attention for remote sensing image change detection. IEEE Trans. Circuits Syst. Video Technol. 32, 6615–6626.