Feature-Guided Multitask Change Detection Network

Yupeng Deng, Jiansheng Chen, Anzhi Yue, Yu Meng, Jingbo Chen, and Yi Zhang

Abstract—Change detection is the discovery of changes in remote sensing images of the same region obtained at different times. Change detection algorithms based on deep neural networks have significant advantages over the traditional algorithms on high-resolution images. State-of-the-art (SOTA) change detection methods require sufficient labeled data to achieve good results, but semantic change detection requires not only binary change masks but also “from-to” change information, so large quantities of change labels are difficult to obtain. Achieving better semantic change detection accuracy with a limited number of labels remains an open problem in the remote sensing field. In this article, we propose a feature-guided multitask change detection network (MCDnet). Feature guidance is characterized by three steps: First, a multitask learning network that uses Siamese encoders to learn segmentation and change detection features simultaneously to realize mutual guidance between tasks is designed, second, a fine-grained feature fusion module to integrate and enhance change information under the guidance of symmetrical change features is constructed, and third, a contrastive loss function based on the a priori knowledge that the features of the changed regions are different while those of the unchanged regions are the same is proposed. The experimental results show that MCDnet achieves SOTA results on three public change detection datasets, including WHU-CD (F1: 94.46 IoU: 89.50), LEVIR (F1: 92.11 IoU: 85.37), and SECOND (mIoU: 73.1 IoU: 22.8). In addition, it is surprising that MCDnet is comparable to the SOTA models while using only 20% of the full training data.

Index Terms—Change detection (CD), contrastive loss, convolutional neural network (CNN), feature guided, multitask learning (MTL).

I. INTRODUCTION

Change detection (CD) captures the spatial changes between multitemporal satellite images due to man-made or natural phenomena [1]. It is one of the major problems in the earth observation field and has been extensively researched in recent decades [2]. Singh [3] defined CD as “the process of identifying differences in the state of an object or phenomenon by observing it at different times.” Some CD techniques have been proposed in the literature. However, the selection of the most suitable method or algorithm for CD is not easy in practice [4]. The CD methodologies can be divided into traditional CD algorithms and machine-learning-based CD algorithms [4].

The traditional CD algorithms can be summarized into the following three groups: direct comparison methods, transformation-based methods, and classification-based methods [5]. The direct comparison methods use image differencing [6], image ratioing [7], and regression analysis [8] to identify pixelwise changes. Transformation-based methods transform images to feature vectors and include change vector analysis [9] and principal component analysis [10]. Direct comparison and transformation-based methods require remotely sensed data to be acquired during the same phonological period [11]. Classification-based CD methods first determine the classification results for images and then compare them to identify the change areas [12], [13], [14], [58]. The CD accuracy of a classification-based method depends on the classification accuracy of each image used.

Along with the increase of image resolution, it is difficult for researchers to summarize the complex spatial information in total and form a general algorithm, so the data-driven algorithms dominate in high-resolution CD tasks. Machine-learning-based CD algorithms, including support vector machine [15], deep neural network [16], and decision tree [17] algorithms, attempt to estimate data properties based on data with labels. Machine-learning algorithms, especially deep neural networks, have shown great accuracy advantages over the traditional methods in the CD of very high-resolution (VHR) remote sensing images. Most deep-learning CD studies have focused on model design and data characterization.

In terms of model design, a semantic segmentation network can be directly used for CD. For example, Peng et al. [18] achieved good accuracy by performing CD based on UNet++. However, CD networks also have different characteristics from semantic segmentation networks. The input of semantic segmentation is a single image, while the input of CD is composed of two images of the same region acquired at different times. Therefore, in recent years, many CD researchers have proposed many models that differ from standard semantic segmentation networks. Siamese architectures [19] were first used for CD in [20]. Compared with previous CD deep neural networks [16], [21], the Siamese architecture used dual encoders to extract image feature pyramids separately. To further enhance the feature representation and generalization abilities of the Siamese
architecture, generative adversarial networks were combined with the Siamese CD network [22]. Unlike a Siamese network with dual encoders, a three-branch network was used for CD tasks [23], and its input includes bitemporal remote sensing images and their difference images. Recent research based on the combination of convolutional neural network (CNN) and transformer facilitates breakthroughs in CD tasks [55], [56], [57], [59], which rely on transformer’s powerful modeling capabilities to make feature fusion and extraction more robust.

In terms of data characterization, recent studies have gradually focused on exploiting the characteristics of CD data, which can be divided into two approaches. First, to alleviate the number of labels needed, Zheng et al. [24] proposed single-temporal supervised learning (STAR), which can exploit object changes in unpaired images, and Chen et al. [25] proposed a data-level solution that can generate bitemporal images containing changes by leveraging generative adversarial training. Second, some work has explored feature fusion methods so that a network can learn the change information easily, e.g., Daudt et al. [20] fused the pairs of feature vectors with concat/add/cut operations to explore their performance on different data.

While most of the pieces of work have focused on model design and data characterization for CD, there is a lack of work on CD scheme design from a feature-guided perspective. Specifically, network design can be divided into encoder feature guidance, feature fusion guidance, and contrastive loss guidance from the perspective of feature guidance.

For encoder feature guidance, semantic segmentation tasks can assist CD tasks [26], and the dependent encoders can be used for CD and semantic segmentation. Similar to the traditional postclassification CD [12], [13], [14], we further consider whether the feature extraction process of a Siamese network encoder can be guided by semantic segmentation. Therefore, we design a multitask CD network (MCDnet) with CD and semantic segmentation and experimentally show that it not only improves the accuracy of CD but also greatly reduces the CD label needed.

For feature fusion guidance, after the Siamese encoder feed acquires the feature pyramid, the paired features must be fused, and then a decoder is used to generate the probability map. FC-Siam-conc [20] fused features by concatenation, and FC-Siam-diff [20] fused features by subtraction. Many recent studies still use concatenation for feature fusion [18], [22], [25], [27], [28]. We found that, for small datasets, such as that in [29], the accuracy of cut is better than that of concatenation, while for recent large datasets, such as LEVIR [30], the accuracy of concatenation is better than that of cut. Summarizing the previous work, we believe that a feature fusion module should have three characteristics, i.e., symmetry, information integrity, and feature enhancement. Therefore, we redesigned a feature fusion module to improve the convergence speed and accuracy of the model without additional computational effort.

For contrastive loss guidance, most CD-related studies [18], [25], [31], [32] adopted cross-entropy loss [33]. There are also some studies that design customized loss functions [34], and a suitable loss function can make the model achieve higher accuracy. Unlike recent CD models, this article proposes a multitask CD model that can output the results of semantic segmentation and CD simultaneously. It can be speculated that the semantic segmentation results are different in changed regions, while they are the same in unchanged regions. Thus, we design a contrastive loss function to maximize the feature distance in changed regions and minimize a similar feature distance in unchanged regions.

The contributions of this work can be summarized as follows.
1) We explore the feature complementarity between segmentation and CD, based on which MCDnet with better CD performance and fewer sample requirements is proposed. MCDnet achieves the state-of-the-art (SOTA) accuracy on the WHU-CD [30], LEVIR [37], and SECOND [38] datasets.
2) A change contrast loss (CCL) function is used to perform feature alignment in CD and is based on the property that the features of unchanged regions are similar while those of changed regions are different.
3) Through extensive research and experiments, we speculate that the feature fusion module should have characterized by symmetry, information integrity, and change region enhancement. A new feature fusion module is proposed according to the conjecture, and the effectiveness of this module is proved by experiment.

The rest of this article is organized as follows. Section II reviews the related work. Section III illustrates the proposed method. In Section IV, experiments are designed to evaluate the proposed method. Finally, Section V concludes this article.

II. RELATED WORK

A. Deep-Learning-Based CD Methods

Sakurada and Okatani [35] were the first to introduce fully convolutional networks for the task of scene CD in computer vision. In the field of remote sensing, FC-Siam-conc and FC-Siam-diff [20] first used the end-to-end Siamese model to find land cover changes.

With the widespread application of deep learning [23], [25], [36], [37], [38] in CD, there has been much research on models [20], [36], data augmentation [25], and loss functions [31], [38]. However, most CD research focuses on binary CD, while few studies focus on multiclass CD and multitask CD.

B. Multitask Learning

Multitask learning (MTL) aims to improve model generalization by leveraging domain-specific information contained in the training signals of related tasks. In the deep-learning era, MTL translates to designing networks capable of learning shared representations from multitask supervised signals [39]. MTL can offer several advantages relative to single-task learning, e.g., reduced calculations and improved performance. Vandenhende et al. [26] showed that MTL can be beneficial, as it allows for the acquisition of inductive bias through the inclusion of related additional tasks into a training pipeline.

A number of works [26], [29], [40] have exhibited models that are combinations of off-the-shelf backbone networks and several task-specific heads. This kind of model relies on an encoder (i.e., a backbone network) to learn a generic representation, which is
Fig. 1. Overview of the proposed framework (MCDnet). The encoder outputs multiscale pyramidal characteristics $P_1, P_2, P_3, P_4, P_5$; then, the decoder composed of CFFM and SB outputs the CD prediction, and the decoder composed of SB outputs the segmentation prediction. Arrows to represent the flow of features.

then used by the task-specific heads to obtain the predictions for each task. In [40], an MTL strategy showed robustness against adversarial attacks, while Zamir et al. [41] indicated that applying cross-task consistency in MTL not only improves generalization but also allows for domain-shift detection.

In the remote sensing field [42], a novel multiscale and multi-task deep-learning framework for automatic road extraction was proposed; the model was able to build a relationship between the learning tasks and simultaneously complete the road detection and centerline extraction tasks. Tan et al. [43] utilized road and junction segmentation cues to guide exploration and achieved better road alignment. Cipolla et al. [29] addressed the problem of preserving semantic segmentation boundaries in the high-resolution satellite imagery by introducing a novel multitask loss. The loss leverages multiple output representations of the segmentation mask and guides the network to focus more on pixels near boundaries.

### III. PROPOSED METHOD

In this section, we first introduce the proposed MCDnet in detail. Then, we illustrate CFFM and CCL proposed in this article.

#### A. MCDnet

MCDnet consists of a Siamese encoder, change decoder, and segmentation decoder in the full mode, as shown in Fig. 1. According to the composition of different learning tasks, MCDnet has three variants: MCDnet-mtask, MCDnet-change, and MCDnet-seg. MCDnet-mtask is the complete variant that can obtain the segmentation and CD results simultaneously. MCDnet-change is a CD variant and is composed of a Siamese encoder and change decoder. MCDnet-seg is the segmentation variant and is composed of a Siamese encoder change decoder. Obviously, MCDnet-change and MCDnet-change can obtain only change or segmentation information.

The input of MCDnet is optical remote sensing images acquired before/after a change occurs, i.e., $X_{\text{post}}$, $X_{\text{pre}}$, and $X_{\text{post}}$. The Siamese encoder shares weight and extracts the depth pyramid features of both $X_{\text{post}}$ and $X_{\text{pre}}$. The change decoder uses the Siamese encoder features at the same time to obtain the change result $P_{\text{change}}$, while the segmentation decoder uses the Siamese encoder features separately to obtain the segmentation results $P_{\text{post}}$ and $P_{\text{pre}}$.

1) **The encoder** can be divided into two parts: a multiscale pyramid feature extractor and a switching path.

The multiscale pyramid feature extractor is shown in Fig. 2(a); it can extract features layer-by-layer with a descending scale...
Fig. 2. (a) Structure of multiscale pyramid feature extractor and BiFPN. (b)–(d) structures corresponding to the nodes of different colors.

using common backbone networks, such as ResNet [48] and EfficientNet [49]. We add a channel adaptation layer in the first layer with 3×3 convolution, the input channel is the image channel, and the output channel is 3 so that the remote sensing images with more than three channels can be pretrained on large-scale datasets, such as ImageNet [50]. Then, five scale features are drawn from the encoder. For example, when the input is 512×512×C (C is the image band) and the depth scale feature size is \{256×256×C₀, 128×128×C₁, 64×64×C₂, 32×32×C₃, 16×16×C₄\}, where \{C₀, C₁, C₂, C₃, and C₄\} are determined by the backbone network used, e.g., in the case of using EfficientNet-b1 [49], the number of channels is \{C₀ = 32, C₁ = 24, C₂ = 40, C₃ = 112, and C₄ = 320\}.

The switching path is designed based on a bidirectional feature pyramid network (BiFPN) [51] structure, which is shown in Fig. 2(a), and it is designed to enhance the features at different scales by constructing propagation pathways between different layers. Considering the differences in both the spatial resolution and feature size in remote sensing images, the proposed network is designed for images whose resolution ranges from 0.3 m to 2 m. For example, for building detection, the spatial features still exist after five downsampling steps at 0.3-m resolutions, while at 2-m resolutions, the spatial features exist for only three downsampling steps. Although different networks can be trained for different resolutions, the difference in features due to the change in resolution has an impact on the accuracy. Therefore, we add the structure of BiFPN into the classification network for feature extraction. By enhancing the exchange of information between features at different scales, the network can be adapted to detect targets with different sizes on different resolution images. With the above structure, when the input is a pair of 512×512×C images, the output of the encoder is \{256×256×C_{bifpn}, 128×128×C_{bifpn}, 64×64×C_{bifpn}, 32×32×C_{bifpn}, 16×16×C_{bifpn}\}, where \{C_{bifpn} ∈ \{32, 64, 128\}\}. In the proposed model \(C_{bifpn} = 32\), the pretemporal phase output features are \{\(P₁, P₂, P₃, P₄, \) and \(P₅\)\}, and the post-temporal phase pyramid features are \{\(P₁', P₂', P₃', P₄', \) and \(P₅'\)\}.

2) The segmentation decoder consists of four segmentation blocks (SBs), as shown in Fig. 3. The SB upsamples the input features and then concatenates the same-scale features, fuses the features through the conv-BN-ReLu block, weights the channels using SE attention, and then extracts the features using the conv-BN-ReLu block. Finally, the segmentation probability map is obtained by a segmentation head.

3) The change decoder consists of five change blocks denoted by CFFM and SB, as shown in Fig. 4, each of which is composed of CFFM, as shown in Fig. 4, and an SB. The SEB upsamples the input features, concatenates the same-scale features, fuses the features through the conv-BN-ReLu block, weights the channels using SE attention, and then extracts features using the conv-BN-ReLu block. Finally, the CD probability map is obtained by a segmentation head.
between each $x_{i,j}$ and $x'_{i,j}$. The number of channels is changed from $C$ to 1 by the MSE operator, so the sigmoid operator is used to narrow its value domain to between 0 and 1. Ultimately, the CFFM output can be expressed as follows:

$$I_n = \text{Concat}[I_{\text{add}*I_{\text{mse}} + I_{\text{add}}*I_{\text{mse}} + I_{\text{cut}}]}.$$  \hspace{1cm} (3)

C. CCL

The proposed MCDnet-mtask model learns both segmentation and change annotations by sharing encoders, which enables mutually beneficial accuracy promotion for both segmentation and CD. The reasons behind this are that the semantic segmentation task guides the encoder to focus more on the features of objects of interest, which makes the CD task make better use of pairwise features for change discovery; and the CD task is similar to contrast learning, by which features extracted by the Siamese encoders can be aligned and propagated to enhance the robustness of feature extraction.

The encoder uses the CD label for feature alignment in this study, and we propose a new CD loss function inspired by the idea of contrast learning, namely, CCL. CCL aims to increase the feature variance in the changed region features and decrease the feature variance in the unchanged region.

Given the binary change ground truth $G_{\text{change}}$, whose size is scaled to that of $P_n$ and $P'_n$, CCL is calculated as follows:

$$P_{\text{post}_s} = P'_{n} * (1 - G_{\text{change}})$$  \hspace{1cm} (4)

$$P_{\text{pre}_s} = P_{n} * (1 - G_{\text{change}})$$  \hspace{1cm} (5)

$$C M_s = \sum_{i=1}^{h\times w\times c} \frac{(P_{\text{post}_s}[i] - P_{\text{pre}_s}[i])^2}{h\times w\times c}.$$  \hspace{1cm} (6)

$$P_{\text{post}_d} = P'_{n} * G_{\text{change}}$$  \hspace{1cm} (7)

$$P_{\text{pre}_d} = P_{n} * G_{\text{change}}$$  \hspace{1cm} (8)

$$P_{\text{pre}_d} = P_{n} * G_{\text{change}}$$  \hspace{1cm} (9)

$$C M_d = \sum_{i=1}^{h\times w\times c} \frac{(P_{\text{post}_d}[i] - P_{\text{pre}_d}[i])^2}{h\times w\times c}\quad(10)$$

$$\text{CCL} = \left( (1 - CM_s) + CM_d \right) / 2$$  \hspace{1cm} (11)

where $CM_s$ represents the MSE of the unchanged region and $CM_d$ represents the MSE of the changed region, so when $CM_{\text{Loss}} \to 0$, $CM_s \to 1$ and $CM_{n} \to 0$.}

IV. EXPERIMENTS

A. Dataset

To verify the effectiveness of the method proposed in this article, three datasets were selected, i.e., LEVIR, WHU-CD, and SECOND. LEVIR contains only the binary change labels of buildings, and WHU-CD contains both the change and segmentation labels of buildings. SECOND contains the change labels as well as the multicategorical segmentation labels for pre- and post-time phases within a changed region.
LEVIR: The LEVIR dataset consists of 637 VHR image patches collected from google earth as shown Fig. 5. The resolution of the images is 0.5 m, and the size is 1024×1024. It is a large-scale CD dataset and covers different kinds of buildings. We use the default dataset split, that is, 445/64/128 for training/validation/testing, respectively.

WHU-CD: The WHU building CD dataset consists of two-period aerial images with a resolution of 0.3 m as shown Fig. 6. These two-period images were obtained in 2012 and 2016. There are a variety of buildings with large-scale changes in the dataset. WHU-CD provides a standard train/validation/test split. We use the official test division.

SECOND: The SECOND dataset has 4662 pairs of aerial images obtained from several platforms and sensors as shown Fig. 7. These pairs of images are distributed over cities, such as Hangzhou, Chengdu, and Shanghai. Each image has a size of 512×512 and is annotated at the pixel level. The annotation of SECOND was carried out by expert earth vision applications, which guarantee high label accuracy. Moreover, the SECOND dataset utilizes land cover map pairs and nonchange masks to represent the change categories. We have only 2948 sets of data labels, so we use 2250 sets for training and 598 sets for validation. The accuracy of our results is evaluated by the official server.

B. Evaluation

LEVIR-CD and WHU-CD are binary category datasets. Therefore, we use precision, recall, intersection over union (IoU), and F1 score as the evaluation metrics. The values of IoU and F1 range from 0 to 1, and the higher the value is, the better the performance it indicates. The IoU metric tends to penalize single instances of bad classification more than the F1 score even when both can agree that this one instance is bad. Therefore, the fluctuations of these two metrics can reflect the problems detected by the model in two different dimensions. The evaluation metrics are calculated as follows:

$$\text{IoU} = \frac{TP}{TP + FP + FN}$$  \hspace{1cm} (12)
$$\text{F1} = \frac{2 \times TP}{2 \times TP + FP + FN}$$  \hspace{1cm} (13)
$$\text{precision} = \frac{TP}{TP + FP}$$  \hspace{1cm} (14)
$$\text{recall} = \frac{TP}{TP + FN}$$  \hspace{1cm} (15)

where TP denotes the true positives, FP denotes the false positives, and FN denotes the false negatives.

SECOND is a multitask dataset that utilizes the common mIoU metric and a proposed coefficient named SeK to evaluate the results. Specifically, given a confusion matrix $Q$

$$\text{mIoU} = \frac{1}{2} (\text{IoU}_1 + \text{IoU}_2).$$  \hspace{1cm} (18)

Moreover, the true positive value of the unchanged pixels ($q_{11}$) always dominates the calculation of $\kappa$. Thus, $q_{11}$ is separated from the calculation of SeK. SECOND also utilizes IoU$_2$ to further emphasize the changed pixels. Specifically, it is defined as follows:

$$\text{SeK} = e^{\text{IoU}_2 - 1} \ast (\hat{\rho} - \hat{\eta}) / (1 - \hat{\eta})$$  \hspace{1cm} (19)

where

$$\hat{\rho} = \sum_{i=2}^{c} \sum_{j=2}^{c} q_{ij} / \left( \sum_{i=1}^{c} \sum_{j=1}^{c} q_{ij} - q_{11} \right)$$  \hspace{1cm} (20)
$$\hat{\eta} = \sum_{j=1}^{c} (q_{j+} : q_{+j}) / \left( \sum_{i=1}^{c} \sum_{j=1}^{c} q_{ij} - q_{11} \right)^2$$  \hspace{1cm} (21)

where $q_{j+}$ and $q_{+j}$ represent the row sum and column sum, respectively, of the confusion matrix without $q_{11}$. The exponential form enlarges the discernibility compared with simple multiplication.

C. Implementation Details

1) Data preprocessing: First, each image in the three datasets is normalized so that the mean is 0 and the variance is 1.

2) Data augmentation includes spatial, color, and noise transformations: Spatial transformations consist of random rotations, cropping, the exchange of pre/post data, and polygon shadows, and color transformations consist of HSV shifts and random brightness contrast operations. Noise transformation refers to the addition of Gaussian noise.

3) Training: We use Adam [52] as the optimizer, with an original learning rate of 0.001 and weight decay of 0.001. The learning rate scheduler is poly; when the epoch is 250, the learning rate is 0.

4) Inference: For both the LEVIR dataset and the SECOND dataset, the original data are input directly, normalized, and then inferred to obtain the results. For the WHU-CD dataset, overlapping sliding prediction with a sliding window of 1024×1024 and an overlapped area of 256 is performed.

The MCDnet method proposed in this article contains MCDnet-change, MCDnet-seg, and MCDnet-mtask, where MCDnet-change is evaluated on the LEVIR dataset and WHU-CD dataset, and MCDnet-mtask is evaluated on the SECOND dataset and WHU-CD dataset. MCDnet-seg is evaluated on the WHU-CD dataset to demonstrate that CD can yield gains for segmentation supervision.
DENG et al.: FEATURE-GUIDED MULTITASK CHANGE DETECTION NETWORK

Table I

| Methods            | Precision | Recall | F1  | IoU  |
|--------------------|-----------|--------|-----|------|
| FC-EF [20]         | 79.20     | 81.72  | 80.44| 67.27|
| FC-conc [20]       | 91.46     | 88.05  | 89.72| 81.36|
| FC-diff [20]       | 89.92     | 89.06  | 89.72| 80.92|
| BiDataNet [53]     | 92.50     | 87.14  | 89.74| 81.39|
| CDNet [35]         | 90.73     | 86.82  | 88.73| 79.75|
| UNet++_MSOF [36]   | 92.18     | 87.84  | 89.96| 81.75|
| DDCNN [18]         | 91.85     | 88.69  | 90.24| 82.21|
| CEECNet V1 [34]    | 93.73     | 89.93  | 91.79| 84.82|
| BIT [56]           | 89.24     | 89.37  | 89.31| 80.68|
| MCDnet-change      | 93.24     | 91.00  | 92.11| 85.37|

The bold values represent the highest number.

Table II

| Methods            | Precision | Recall | F1  | IoU  |
|--------------------|-----------|--------|-----|------|
| FC-EF [20]         | 85.27     | 75.11  | 79.87| 66.49|
| FC-conc [20]       | 92.67     | 55.59  | 69.49| 53.25|
| FC-diff [20]       | 88.76     | 66.40  | 75.97| 61.25|
| BiDataNet [53]     | 93.44     | 87.66  | 90.46| 82.58|
| CDNet [35]         | 87.64     | 81.43  | 84.42| 73.04|
| UNet++_MSOF [36]   | 93.03     | 86.29  | 89.53| 81.05|
| DDCNN [18]         | 96.11     | 89.73  | 92.81| 86.59|
| CEECNet V1 [34]    | 95.16     | 92.04  | 93.77| 88.23|
| BIT [56]           | 86.64     | 81.48  | 83.98| 72.39|
| MCDnet-change      | 95.34     | 92.82  | 94.06| 88.79|
| MCDnet-mlask       | 95.14     | 93.58  | 94.46| 89.50|

The bold values represent the highest number.

All the experimental codes are written using Python and PyTorch [52] for deep-learning training and inference, and the GPU is a 3090 graphics card with 24 GB of memory.

D. Comparative Method

For the LEVIR-CD and WHU-CD dataset, we compare our method with several SOTA CD methods.

1) **FC-EF, FC-conc, and FC-diff [20]**: FC-EF concatenates the bitemporal image pairs as the input, and FC-diff/FC-conc is based on FC-EF.

2) **BiDataNet [53]**: BiDataNet adds LSTM to U-Net.

3) **CDNet [35]**: CDNet was designed for street view CD.

4) **UNet++_MSOF [36]**: CD for high-resolution satellite images using improved UNet++.

5) **DDCNN [18]**: Difference-enhanced dense-attention CNN.

6) **CEECNet V1 [34]**: CEECNet V1 defines a new loss function as well as a new spatial and channel convolution attention layer.

7) **BIT [56]**: A transformer-based CD model.

For the SECOND dataset, its task is multiclass semantic CD, and only official evaluations are currently available, so we compare MCDnet-mlask with ASN/ASN-ATL [38].

E. Results on the LEVIR Dataset

For the LEVIR dataset, since it contains only binary change labels, the proposed MCDnet-change is compared with SOTA CD models, as shown in Table I. MCDnet-change achieves higher F1 and IoU scores than the recently proposed CEECNetV1 model.

We show four sets of experimental results in Fig. 8. In the first set, the building color and ground color are similar, and MCDnet-change extracts buildings more completely and gives more accurate edge details; in the second and third sets, MCDnet-change performs better in extracting details of both small and large buildings, but there is a misidentification when classifying houses and containers. In the fourth set, MCDnet-change extracts more small buildings compared with CEECNetV1, while the edges of large buildings are accurate. The experimental results show that MCDnet-change can consider both large and small buildings and can extract better details of building changes.

F. Results on the WHU-CD Dataset

The WHU-CD dataset provides both change labels and building segmentation labels, so we use it to validate both MCDnet-change and MCDnet-mlask. It is important to note that our experiments and accuracy evaluation, as shown in Table II, are carried out on an entire remote sensing image scene. MCDnet-change achieves an F1 score of 94.06% and an IoU score of 88.79%, and MCDnet-mlask achieves an F1 score of 94.36% and an IoU score of 89.32%. It is obvious that the introduction of the segmentation task can promote the CD task.

We show four sets of experimental results, as shown in Fig. 9. The first row shows that MCD-mlask can effectively detect the missed change region of MCD-change and CEECNet, and MCD-mlask outperforms MCD-change in detecting the building edges due to the use of segmentation information. The second row shows that both MCDnet-mlask and MCDnet-change are significantly better than the recent SOTA CEECNet for scenarios where only building roof renovations have occurred and there is no significant change in the building structure. In the third row, the building extraction results are shown, and both CEECNet and MCD-mlask can extract the changed buildings completely. In the fourth row, MCD-mlask can detect more medium and large buildings than CEECNet, but there are still missed detections.
G. Results on the SECOND Dataset

The accuracy on the SECOND dataset is shown in Table III, where MS denotes the multiscale test and Flip denotes the flip enhancement test. We use IoU to measure the CD accuracy and Sek to measure the classification accuracy. Table III presents that the CD and the classification accuracy are promoted by 2.9% and 5.5%, respectively. The accuracy promotion fluctuates with the adoption of data augmentation, e.g., IoU and Sek are promoted by 3% and 6.1%, respectively, when using multiscale enhancement, and the IoU value reaches its peak when multiscale and flip enhancement are adopted. MCD-mtask is superior to the ASN-ATL model in terms of both the accuracy of CD and changed region classification, i.e., semantic CD.

Fig. 10 shows the experimental results on the two groups of image pairs. The group shown in the first two rows mainly covers changes from low vegetation to buildings, changes from buildings to bare ground, and changes from fragmented trees to low vegetation. The results indicate that the proposed method has fewer missed detections than ASN-ATL and HRSCD.str4, and the small change from fragmented trees to low vegetation is also correctly detected. The group shown in the second two rows mainly contains large change areas from low vegetation to tree changes that are overlaid with the shadows of buildings. The result obtained by MCD-mtask is similar to the ground truth, while the ASN-ATL and HRSCD.str4 models fail to distinguish between pseudochanges, such as shadows, and real changes.

H. Ablation Studies

To demonstrate the contributions of mtask, CFFM, and CCL, we carry out alignment studies on the WHU-CD dataset. As shown in Table IV, mtask, CCL, and CFFM bring approximately 1.15%, 0.3%, and 0.72% F1 score improvements, respectively.

1) mtask Analysis: MCDnet-change is a network trained with CD labels, while MCDnet-seg is a network trained with semantic segmentation labels. As the combination
TABLE III
RESULTS ON THE SECOND DATASET

| Methods         | MS | Flip | MS | Flip | MS | Flip |
|-----------------|----|------|----|------|----|------|
|                 | ☐  | ☐    | ☐  | ☐    | ☐  | ☐    |
|                 | mIoU | Sek  | mIoU | Sek  | mIoU | Sek  |
| FC-EF [20]      | 59.3 | 5.7  | 59.3 | 6.7  | 59.0 | 5.9  |
| FC-conc [20]    | 63.3 | 9.1  | 62.8 | 9.2  | 62.9 | 9.4  |
| FC-diff [20]    | 61.9 | 8.8  | 60.9 | 8.6  | 61.0 | 8.7  |
| HRSCD.str4 [54] |     |      |     |      |     |      |
| ASN [38]        | 69.0 | 15.2 | 69.5 | 16.1 | 69.7 | 16.2 |
| ASN-ATL [38]    | 69.1 | 15.5 | 69.8 | 16.5 | 70.0 | 16.8 |
| MCDnet-mtask    | 72.0 | 21.0 | 72.8 | 22.6 | 73.1 | 22.8 |

The table presents the results for different methods on the second dataset. The bold values represent the highest number.

TABLE IV
ABLATION RESULTS ON THE WHU-CD DATASET

| mtask | CCL | CFFM | Precision | Recall | F1 | IoU |
|-------|-----|------|-----------|--------|----|-----|
| -     | -   | -    | 95.12     | 90.60  | 92.80 | 86.58 |
| -     | -   | ✓    | 96.04     | 91.14  | 93.52 | 87.83 |
| -     | ✓   | -    | 95.47     | 90.84  | 93.10 | 87.08 |
| ✓    | -   | -    | 95.31     | 92.64  | 93.95 | 88.06 |
| ✓    | ✓   | -    | 95.34     | 92.82  | 94.06 | 88.79 |
| ✓    | ✓   | ✓    | 95.61     | 92.98  | 94.28 | 89.18 |
| ✓    | ✓   | ✓    | 95.14     | 93.59  | 94.46 | 89.50 |

The table presents the ablation results for different conditions on the WHU-CD dataset. The bold values represent the highest number.

TABLE V
ABLATION RESULTS ON THE WHU-CD DATASET

| Methods        | Change | Segmentation 2012 | Segmentation 2016 |
|----------------|--------|-------------------|-------------------|
|                | F1     | IoU   | F1    | IoU   | F1    | IoU   |
| MCDnet-change  | 92.80  | 86.58 | -     | -     | -     | -     |
| MCDnet-seg     | -      | -     | 95.98 | 92.28 | 96.80 | 93.80 |
| MCDnet-mtask   | 94.46  | 89.50 | 96.22 | 92.73 | 97.23 | 94.61 |

The table presents the ablation results for different conditions on the WHU-CD dataset. The bold values represent the highest number.
Fig. 10. SECOND dataset. (a) Preimage and postimage. (b) HRSCD.str4 result. (c) ASN-ATL result. (d) MCDnet-ntask result. (e) Ground truth.

### TABLE VI
**DIFFERENT UNION METHODS ON THE WHU-CD DATASET**

| Model       | Union method | F1   | IoU  |
|-------------|--------------|------|------|
| MCDnet-change | Add          | 91.99| 85.78|
| MCDnet-change | Cut          | 92.12| 86.03|
| MCDnet-change | Concat       | 92.80| 86.58|
| MCDnet-change | CFM[55]      | 93.11| 87.20|
| MCDnet-change | TST[57]      | 92.96| 86.77|
| MCDnet-change | CFFM         | 93.52| 87.83|

### TABLE VII
**EFFICIENCY COMPARISON ON THE WHU-CD DATASET**

| Methods                  | Params (Mb) | Inference Time (s) |
|--------------------------|-------------|--------------------|
| FC-EF [20]               | 5.2         | 21                 |
| FC-conc [20]             | 5.9         | 21                 |
| FC-diff [20]             | 5.2         | 21                 |
| CDNet [35]               | 7.8         | 26                 |
| UNet++_MSOF [36]         | 34.3        | 27                 |
| DDCNN [18]               | 178.2       | 44                 |
| CEECNet VI [34]          | 121.2       | 46                 |
| BFI [56]                 | 132.9       | 63                 |
| MCDnet-change            | 39.1        | 39                 |
| MCDnet-ntask             | 56.8        | 52                 |

2) **Analysis of CFFM**: The feature guidance of CFFM is characterized by symmetry, information integrity, and change region enhancement. As shown in Table VI, CFFM improves the F1 score by 1.53% and 0.72% compared...
with add and concat fusion, respectively. The results indicate that implicit feature learning using CNNs is superior to direct prior information introduction, especially when the introduction may lead to information loss (add/cut), while a reasonably designed prior information fusion method with feature guidance can better improve the accuracy.

3) Analysis of CCL: CCL is able to fully exploit the relationships between images by explicitly aligning each layer of feature pairs. As shown in Table IV, the addition of CCL achieves the accuracy improvement of approximately 0.3% for both MCDnet-change and MCDnet-mtask.

I. Number of Labels and Accuracy

There is a constraint relationship between the accuracy efficiency of the deep-learning model, so in this study, based on the WUH-CD dataset, efficiency comparison experiments are conducted using a single 3090 graphics card, and the results are shown in Table VII. The accuracy achieved by the network proposed in this article is obviously superior at the same scale.

J. Number of Labels and Accuracy

Sufficient labeled images for CD are usually more difficult to obtain than those for target detection and semantic segmentation. Deep learning requires a large amount of data with the purpose of learning feature representations as well as prior knowledge, which means that the introduction of suitable prior knowledge should reduce the need for large amounts of data. To verify this inference, we examine the relationship between the number of labels and the corresponding accuracy on the WHU-CD dataset when prior knowledge is introduced.

As shown in Table VIII, MCDnet is comparable to the SOTA model when only 20% of the full training data are used. The highest accuracy improvement comes from the addition of MTL, which means that the segmentation task can be used to assist CD, thus greatly reducing the amount of CD samples needed.

V. Conclusion

In this article, we investigate how to more fully exploit the potential of CNN-based CD models with feature guidance. First, we design an MTL network, MCDnet-mtask, with the goal of mutually beneficial feature learning introduced by multiple segmentation tasks. Then, we analyze the possibility of improving the accuracy under the guidance of contrast learning and design a loss function, CCL, which can maximize the feature distance in changed regions and minimize a similar feature distance in unchanged regions. Finally, a novel feature fusion module, CFFM, is proposed to integrate the learned features and enhance the change information under the guidance of symmetrical change features.

The algorithm proposed in this article is tested on three datasets, LEVIR, WHU-CD, and SECOND, and superior accuracy is achieved compared with the SOTA CD models. However, the model proposed in this article also has shortcomings, our model is for symmetric change scenarios, and the accuracy improvement for asymmetric change scenarios will not be so significant. In the future, we will continue to design reasonable feature-guided schemes based on the network proposed in this article. We will verify the effectiveness of the scheme proposed in this article on actual homeland data.

REFERENCES

[1] A. Asokan and J. Anitha, “Change detection techniques for remote sensing applications: A survey,” Earth Sci. Inform., vol. 12, no. 2, pp. 143–160, Mar. 2019, doi: 10.1007/s12145-015-00380-5.
[2] 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, May 2020, Art. no. 1688, doi: 10.3390/rs12101688.
[3] A. Singh, “Review article digital change detection techniques using remotely-sensed data,” Int. J. Remote Sens., vol. 10, no. 6, pp. 989–1003, Jun. 1989, doi: 10.1080/01431168908903939.
[4] D. Lu, P. Mausel, E. Brondízio, and E. Moran, “Change detection techniques,” Int. J. Remote Sens., vol. 25, no. 12, pp. 2365–2401, Jun. 2004, doi: 10.1080/0143116031000139863.
[5] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, “Change detection from remotely sensed images: From pixel-based to object-based approaches,” ISPRS J. Photogramm. Remote Sens., vol. 80, pp. 91–106, Jun. 2013, doi: 10.1016/j.isprsjprs.2013.03.006.
[6] P. R. Copin and M. E. Bauer, “Digital change detection in forest ecosystems with remote sensing imagery,” Remote Sens. Rev., vol. 13, no. 3/4, pp. 207–234, Apr. 1996, doi: 10.1080/02757259609532305.
[7] P. J. Howarth and G. M. Wickware, “Procedures for change detection using landsat digital data,” Int. J. Remote Sens., vol. 2, no. 3, pp. 277–291, Jul. 1981, doi: 10.1080/01431168108980362.
[8] A. K. Ludeke, R. C. Maggio, and L. M. Reid, “An analysis of anthropogenic deforestation using logistic regression and GIS,” J. Environ. Manage., vol. 31, no. 3, pp. 247–259, Oct. 1990, doi: 10.1007/s10301-008-0038-6.
[9] W. A. Malila, “Change vector analysis: An approach for detecting forest changes with landsat,” in Proc. Mach. Process. Remotely Sensed Data Symp., 1980, pp. 326–335.
R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, “Multitask learning for large-scale semantic change detection,” *Comput. Vis. Image Understanding*, vol. 187, 2019, Art. no. 102783.

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

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

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, 2022.

F. Samadi, G. Akbarizadeh, and H. Kaabi, “Change detection in SAR images using deep belief network: A new training approach based on morphological images,” *IET Image Process.*, vol. 13, no. 12, pp. 2255–2264, 2019.

S. Fang, K. Li, and Z. Li, “Changer: Feature interaction is what you need for change detection,” 2022, arXiv:2209.08290.

**Yupeng Deng** received the B.S. degree in communication engineering from the Hunan University of Science and Technology, Xiangtan, China, in 2017. He is currently working toward the Ph.D. degree in signal and information processing with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include landcover segmentation, change detection, and deep learning.

**Jiansheng Chen** received the Ph.D. degree in signal and information processing from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently an Assistant Professor with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include object detection and remote sensing image processing.

**Shiming Yi** received the M.S. degree in pattern recognition and intelligent system from the National University of Defense Technology, Changsha, China, in 2007. He is currently an Engineer with the Engineering Quality Supervision Center of Logistics Support Department of the Military Commission, Beijing, China. His research interests include the land survey and planning.

**Anzhi Yue** received the Ph.D. degree in agricultural information technology from China Agricultural University, Beijing, China, in 2012. He is currently an Associate Professor with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include the fields of land use and land-cover classification, with a focus on remote sensing technologies and their application in land survey and monitoring.

**Yu Meng** received the Ph.D. degree in signal and information processing from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2008. She is currently a Professor with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her research interests include change detection, knowledge graph, and image processing.

**Jingbo Chen** received the Ph.D. degree in signal and information processing from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2011. He is currently an Associate Professor with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include deep learning in optical remote sensing imagery.

**Yi Zhang** received the bachelor’s degree in automation from the Beijing University of Chemical Technology, Beijing, China, in 2016, and the doctorate degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2022. Her research interests include deep learning and remote sensing image target extraction.