Semantic Change Detection with Asymmetric Siamese Networks

Kunping Yang, Gui-Song Xia, *Senior Member, IEEE,* Zicheng Liu, Bo Du, *Senior Member, IEEE,*
Wen Yang, *Senior Member, IEEE,* Marcello Pelillo, *Fellow, IEEE,* Liangpei Zhang, *Fellow, IEEE*

***Abstract*—Given two multi-temporal aerial images, semantic change detection aims to locate the land-cover variations and identify their change types with pixel-wise boundaries. This problem is vital in many earth vision related tasks, such as precise urban planning and natural resource management. Existing state-of-the-art algorithms mainly identify the changed pixels by applying homogeneous operations on each input image and comparing the extracted features. However, in changed regions, totally different land-cover distributions often require heterogeneous features extraction procedures *w.r.t* each input. In this paper, we present an *asymmetric siamese network* (ASN) to locate and identify semantic changes through feature pairs obtained from modules of widely different structures, which involve areas of various sizes and apply different quantities of parameters to factor in the discrepancy across different land-cover distributions. To better train and evaluate our model, we create a large-scale well-annotated *SEmantic Change detectiON Dataset* (SECOND), while an *Adaptive Threshold Learning* (ATL) module and a *Separated Kappa* (SeK) coefficient are proposed to alleviate the influences of label imbalance in model training and evaluation. The experimental results demonstrate that the proposed model can stably outperform the state-of-the-art algorithms with different encoder backbones.**

***Index Terms*—Aerial images, semantic change detection, asymmetric siamese network, benchmark dataset, separated kappa.**

I. Introduction

CHANGE detection in multi-temporal aerial images [1], [2], [3], [4], [5], which aims to locate and analyze the regions of land-cover variations on the earth surface, is a crucial image interpretation task related to many applications such as, precise urban planning [6] and natural resource management [7], [8], [9]. Given a pair of multi-temporal images, most existing methods focus on detecting the locations of changed pixels between the input images, namely binary change detection (BCD) *e.g.* [1], [2], [10], [11]. However, since it overlooks pixels' categories, BCD often fails to depict the semantic change information that are highly demanded in subsequent applications. Hence, developing methods that can simultaneously extract changed regions and identify their land-cover classes in multi-temporal images, *i.e.* semantic change detection (SCD), has become an active research topic in recent years [3], [12].

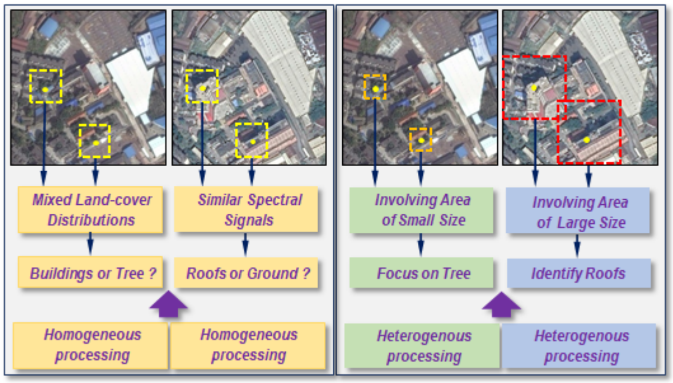

Fig. 1. In aerial images, land-cover objects appearing at different geometrical structures and mixed distributions across multi-temporal images, which we call asymmetric changes, make it difficult to locate and analyze land-cover variations through existing methods with homogeneous image processings *w.r.t* each input. In contrast with existing methods, we are motivated to design some heterogenous image processings, which we call locally asymmetric, to factor in the discrepancy across different land-cover distributions and provide extra information for SCD problem.

An intuitive solution to achieve SCD is to first partition the input images into semantic regions and then compare the segmentation results for identifying change types. However, this direct solution makes the underlying assumption that semantic categories in multi-temporal images are independent and is typically troubled by two problematic aspects: 1) the changed regions of the same semantic category cannot be distinguished; 2) the intrinsic correlation across categories will be overlooked.

In order to take advantage of categorical correlation in multi-temporal images, many existing SCD methods rely on the architecture of siamese networks [13] and have reported promising results [3], [12]. However, in contrast to the application scenarios in [14], [15], the involved siamese networks in SCD would face with difficulties in locating and identifying changed regions conforming to different mixed land-cover distributions in each multi-temporal image, which we call asymmetric changes. For instance, focusing on the marked positions (yellow points) in Fig.1, to better identify regions with mixed distributions of *tree* and *buildings*, concentrating on small area (within orange boxes) could provide delicate details. But to alleviate the categorical ambiguity between *impervious*

The study of this paper is funded by the National Natural Science Foundation of China (NSFC) under grant contracts No.61922065, No.61771350 and No.41820104006. It is also partially funded by the Major Projects of Technological Innovation in Hubei Province (2019AEA170).

K. Yang, G.-S. Xia, Z. Liu, B. Du and L. Zhang are with the State Key Lab. of LIESMARS and the School of Computer Science, Wuhan University, Wuhan, 430072, China. Email: {*kunpingyang, guisong.xia, zicheng.liu, dubo, zlp62*}@whu.edu.cn.

W. Yang is with the School of Electronic Information, Wuhan University, Wuhan, 430072, China. Email: yangwen@whu.edu.cn.

M. Pelillo is with DAIS, University of Venice, 30172, Italy. Email: pelillo@unive.it.

Corresponding author: Gui-Song Xia (guisong.xia@whu.edu.cn).

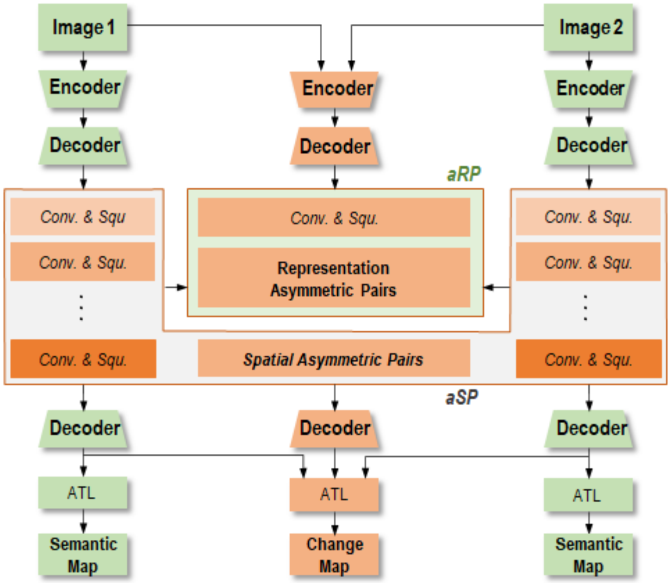

Fig. 2. *Asymmetric Siamese Network* (ASN) for SCD. ASN utilizes siamese encoders to map input multi-temporal images into feature space, while the siamese decoders are leveraged to obtain semantic maps. Similarly, encoder and decoders in change detection branch are designed to obtain change map. In contrast to traditional siamese network, ASN utilizes several convolutional sequences and squeeze gates in proposed aSP and aRP to obtain feature pairs deriving from widely different structures, which we call asymmetric feature pairs, to provide extra information. Furthermore, the designed ATL is exploited to adaptively revise the output deflections based on the combinations of raw model outputs through slight extra convolutional layers.

surface and the roofs of *buildings*, features involving more surroundings (within red boxes) are preferred. Furthermore, feature representation capabilities are often required to be adaptive. For example, pixels in one image belonging to objects with complex structures (*e.g. buildings* mixed with *tree*) often prefer modules with stronger representation capabilities, which however may make obvious flaws on those with simple structures (*e.g.* single *impervious surface*) in another image due to the over-fitting. In summary, mixed land-cover classes and various land-cover distributions often make object properties in each input image different, and cannot be modeled well by solely using symmetric architectures as in siamese networks. Thus, it is of great importance to design a deep model that can better depict the semantic but asymmetric changes in images.

In this article, we address the asymmetric properties of SCD problem by exploiting siamese networks. As illustrated in Fig. 2, we propose an *Asymmetric Siamese Network* (ASN) to extract changed pixels through two modules, *i.e., asymmetric Spatial Pyramid* (aSP) and *asymmetric Representation Pyramid* (aRP). Leveraging designed convolution sequences of different structures, aSP and aRP obtain features through several siamese feature pyramids deriving from input images. Specifically, in aSP and aRP, we design weighted dense connected topological architectures, where different feature pairs across the obtained siamese feature pyramids deriving from each input are linked with various edges. Although the whole architecture is symmetric, most of these feature pairs are obtained by widely different structures, which we call asymmetric feature pairs and locally asymmetric structures. Dynamic branch weights further adjust the importance of each edge according to each input. Containing designated receptive fields and representation capabilities, these asymmetric feature pairs are able to focus on areas of various sizes and implicate different representation capabilities. As we shall see, compared with traditional siamese network, ASN can better depict asymmetric changes between mixed targets and illegible area.

To better train and evaluate the proposed model, we create a well-annotated *SEmantic Change detectiON Dataset* (SECOND) to set up a new benchmark. Although existing SCD datasets contain abundant categorical information, they are often not big enough [12], which are inadequate to develop SCD algorithms with good generalization ability. Meanwhile, the annotations of some SCD datasets are unable to identify changes between the same land-cover class [3]. The proposed SECOND is with 4662 pairs of images in 30 change types. Especially, we annotate the semantic categories and changed pixels separately in SECOND dataset, which makes changed regions between the same land-cover class available.

Last but not least, model training and evaluation are always influenced by label imbalance due to the overwhelming categories, *e.g. non-change* pixels. Specifically, severe label imbalance would make models tend to collapse. Thus, we propose an *Adaptive Threshold Learning* (ATL) module to adaptively revise the deflections of semantic change outputs caused by label imbalance through learnable output adjustments, where the logical relationships between each semantic category can be explored. Moreover, existing metrics used to evaluate change detection algorithms, such as Overall Accuracy (OA) and Kappa coefficient (κ), are inherited from classification tasks, which would cause unreasonable scores due to the neglect of the dominant *non-change* pixels. Thus, we further present a *Separated Kappa* (SeK) coefficient as a modified evaluation measurement for semantic change detection task, which separates the *non-change* class from other change types to reduce the effects of label imbalance. Compared with OA and κ, SeK is more in line with human scoring in SCD problem.

Our main contributions in this paper are threefold.

- We propose an asymmetric siamese network, *i.e.* ASN, to factor in the discrepancy across different land-cover distributions in each multi-temporal image, which can alleviate the categorical ambiguity through extra information provided by heterogenous processings.
- We create a large-scale semantic change detection dataset, *i.e.* SECOND, to better train deep models and as a new benchmark for the SCD problem. This SECOND also enables us to distinguish changed regions between the same land-cover class.
- We design an *Adaptive Threshold Learning* module and a *Separated Kappa*, *i.e.* ATL and SeK, to alleviate influences of label imbalance, which can adaptively revise the output deflections and fix unreasonable scores computed with traditional metrics, *e.g.* OA and κ, respectively.

II. Related Work

A. Location of Changed Regions

In order to locate changed regions in multi-temporal images, early works formulated the task as a BCD problem and relied on certain probability statistical formulations [1], [2], [16], [17], [18], [19] or change vector analysis [20], [21], [22], [23], [24]. The statistical formulations and change vector analysis are able to explore the dissimilarity between features corresponding to changed and unchanged regions. Meanwhile, in order to embed pixel-wise correspondences, graphical model based algorithms are proposed [25], [26], which can model the spatial regularity during the optimization process. Further to expand the model capacities, deep learning networks [5], [27], [28], [29], [30] are designed with elaborate structures to depict more diverse scenes by searching optimal parameters, where huge parameter space ensures stronger model capacities. Moreover, as illustrated in Sec.I, considering in the discrepancy across land-cover distributions in input images could provide extra information when depicting land-cover distributions in the cases shown in Fig.1. Thus, it is necessary to explore some asymmetric architectures in conventional deep networks. In this paper, we propose an ASN to leverage locally asymmetric structures, which would be able to detect changed regions through change detection branch more precisely compared with traditional siamese networks.

B. Identification of change types

Existing SCD algorithms mainly utilize two ways to identify the change type of each pixel. The first kind of models, *e.g.* [12], consider *non-change* as a special change type, which extract unchanged regions and other change types simultaneously. The second kind of methods, *e.g.* [3], [31], separately extract unchanged regions and identify change types. However, these models often fail to consider issues such as multi-scale objects or scenes of various complexities in the interpretation process. As discussed in recent semantic segmentation algorithms [32], [33], [34], parallel structures with diverse receptive fields significantly improve the model performances when dealing with multi-scale objects. In views of this insight, in SCD problem, we need to explore adaptive structures to obtain features suitable for different land-cover distributions corresponding to each input image. Furthermore, dominant *non-change* pixels in SCD problem would make model tend to collapse during the training process [35], which also leads us to investigate the revision process of raw model outputs. In proposed ASN, we design structures with various dilation rates and parameter quantities to adapt different land-cover distributions, while the proposed ATL learns to adaptively update semantic predictions in the light of raw model outputs.

C. Datasets and Evaluation Metrics

Benchmark datasets and evaluation metrics are two important aspects related to training the designed models and measuring the generalization capability of SCD algorithms.

Existing datasets, *e.g.* [36], [37], [38], [39], are mainly created for the BCD problem. The lack of land-cover categorical information still limits their usage in SCD, although [5] has made an impressive contribution to alleviate data scarcity through active learning. A few benckmark datasets have been built for SCD [3], [12], [31]. Among them, the SCD datasets used in [12], [31] are not big enough to sufficiently train and evaluate SCD algorithms. While, the dataset in [3] utilizes two independently annotated land-cover maps to represent the change types, which ignores changed regions between the same land-cover class. Thus, in our proposed SECOND dataset, we check multi-temporal images simultaneously to annotate land-cover classes and changed regions separately, which can distinguish changed regions between the same land-cover class.

On the other hand, commonly used evaluation metrics for change detection tasks are mainly inherited from those used in classification problems, such as OA and κ used in [3], [12], which ignore the fact that unchanged regions are often of the overwhelming majority in change maps. As a consequence, models with dominant predictions of *non-change* pixels would get unreasonable high scores in terms of OA and κ. Thus, it is demanded to take into account the label imbalance for better evaluations. In this paper, we utilize mean Intersection Over Union (mIOU) [40] and design a new metric, *i.e.* SeK, to measure the SCD results, which can alleviate the label imbalance effects.

III. Methods

A. Problem Definition

Let $I_1, I_2 : \Omega \to \mathbb{R}^d$ denote two multi-temporal images of $d \in \mathbb{N}_+$ channels, with Ω being the image grid $\{0, 1, \dots, H-1\} \times \{0, 1, \dots, W-1\}$. Given a set $L = \{y_1, \cdots, y_N\}$ of N semantic categories, the SCD problem aims to find a mapping function $f_{I_1,I_2} : \Omega \mapsto L^2$ such that

$$\forall \mathbf{p} \in \Omega,\ f_{I_1,I_2}(\mathbf{p}) = \begin{cases} (0,\, 0) & \text{if } \mathcal{C}_{I_1,I_2}(\mathbf{p}) < \tau, \\ (l_1,\, l_2), & \text{otherwise,} \end{cases} \tag{1}$$

where $\mathcal{C}_{I_1,I_2}(\mathbf{p})$ measures the change probability of each pixel $\mathbf{p} \in \Omega$, $l_1, l_2 \in L$, and $(0,\, 0)$ indicates *non-change* class. τ is a scalar thresholding on $\mathcal{C}_{I_1,I_2}$. Thus, f_{I_1,I_2} can locate changed regions and identify their categories simultaneously.

B. An Intuitive Solution

An intuitive solution to obtain such a f_{I_1,I_2} is to first partition the input images I_1 and I_2 into semantic regions, *e.g.* with semantic segmentation algorithms, and then compare them to locate and identify semantic changes. Specifically, for $t = 1, 2$, let $\mathcal{M}_t : \Omega \mapsto \mathbb{R}^N$ be the semantic probability map of I_t, *i.e.* the probability vector $\mathcal{M}_t(\mathbf{p}) \in \mathbb{R}^N$ indicates the possibility of pixel $\mathbf{p} \in \Omega$ belonging to each semantic category in L. Denoting $l_t = \arg\max_{l \in L} \mathcal{M}_t(\mathbf{p})$ as the semantic category of pixel $\mathbf{p}$ in I_t, we have

$$f_{I_1,I_2}(\mathbf{p}) = \begin{cases} (0,\, 0), & \text{if } l_1 = l_2, \\ (l_1,\, l_2), & \text{otherwise.} \end{cases} \tag{2}$$

As $\max \mathcal{M}_1(\mathbf{p}) \cdot \max \mathcal{M}_2(\mathbf{p}) = \max(\mathcal{M}_1^T(\mathbf{p}) \times \mathcal{M}_2(\mathbf{p}))$, for Eq. (2), we have

$$(l_1, l_2) = \arg \max_{(l_1,l_2) \in L^2} (\mathcal{M}_1^T(\mathbf{p}) \times \mathcal{M}_2(\mathbf{p})),$$

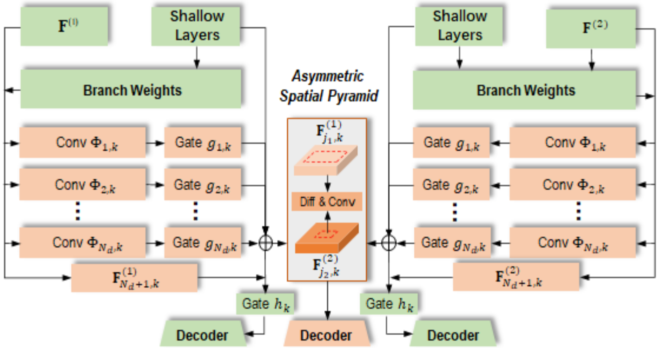

Fig. 3. The proposed aSP module with length of 1. Index k controls the channel numbers of layers in each convolution sequence, while j_1,j_2 indicate different receptive fields. Each squeeze gate consists of the concatenation operator, convolution layers and skip connections. aSP exploits asymmetric spatial feature pairs with diverse spatial information.

where the operation × represents matrix product. However, due to the non-relevance between $\mathcal{M}_1$ and $\mathcal{M}_2$, the intuitive solution actually implies that *the semantic categories in each image are independent for every pixel*. However, this underlying assumption is quite different from the reality. The intrinsic correlation between each category is important and the model performance would be limited by overlooking the categorical correlation. Moreover, according to Eq. (2), this intuitive solution also can not identify changed regions between same land-cover classes due to the same output form with *non-change* pixels.

C. SCD with Conventional Siamese Networks

To take categorical correlation into account, the state-of-the-art algorithm, *i.e.* HRSCD.str4 [3], utilizes siamese semantic segmentation branches with an extra change detection branch to address the SCD problem. During the parameter optimization process, the siamese branches would influence each other through the gradient flows and skip connections.

However, it is worth noticing that HRSCD.str4 relys on totally symmetric structures. More precisely, the response values in feature maps of change detection branch are related to the areas of same size and obtained by nearly the same mapping functions on multi-temporal images. Besides, the siamese decoders utilize single-line structure with skip connection, which are not fully aware of various object scale distributions *w.r.t.* each input. As we shall see in Sec.VI, this symmetric architecture makes it difficult to locate and identify semantic changes related to widely different land-cover distributions across multi-temporal images in some asymmetric changes.

D. Asymmetric Siamese Network for SCD

1) Overall Architecture: In order to depict the aforementioned asymmetric changes that beyond the descriptive capability of conventional siamese networks, we propose to integrate features deriving from heterogenous processes that are adaptive to each input. Specifically, as illustrated in Fig.2, the proposed ASN exploits siamese feature pairs *w.r.t.* the input images, denoted as I_1, I_2, in semantic segmentation branches, based on which designed aSP and aRP generate feature pairs with various receptive fields and representation capabilities through convolution sequences with different dilation rates and parameter quantities. These feature pairs are further integrated into the change detection branch and semantic segmentation branches to obtain predictions of *non-change* pixels and semantic categories, which finally compose the semantic change detection results.

2) Asymmetric Spatial Pyramid (aSP): In aSP, we design several parallel convolution sequences $\{\Phi_{j,k}\}_{1\leq j\leq N_d,\,1\leq k\leq N_r}$ of diverse structures, where each sequence is a convolutional operation set, denoted as $\Phi_{j,k} = \{\phi_{i,j,k}\}_{1\leq i\leq N_c}$. Specifically, N_d, N_c and N_r are positive integers, denoted as $\mathbb{N}_+$. For each $\phi_{i,j,k}$, output channel number is set as $c_i \cdot r_k$, where channel hyper-parameter $c_i \in \mathbb{N}_+$ and multiplication hyper-parameter $r_k \in \mathbb{N}_+$ are used to control the feature representation capabilities. Besides, spatial hyper-parameter $d_j \in \mathbb{N}_+$ is utilized to embed $\Phi_{j,k}$ with various spatial information, when the dilation rate of $\phi_{i,j,k}$ is set as d_j.

As illustrated in Fig.3, given features $\mathbf{F}^{(1)}, \mathbf{F}^{(2)}$ deriving from I_1, I_2 respectively in siamese semantic segmentation branches, aSP integrates features calculated by each $\phi_{i,j,k} \in \Phi_{j,k}$. Then, we utilize squeeze gate $g_{j,k}$ consisting of convolutions and skip connections to reduce the computational complexity. Concretely, for $t = 1, 2$, we have

$$\mathbf{F}^{(t)}_{i,j,k} = v^{(t)}_{i,j} \cdot \phi_{i,j,k}\big(\mathbf{F}^{(t)}\big), \tag{3}$$

$$\mathbf{F}^{(t)}_{j,k} = g_{j,k}\big(\cup_{i=1}^{N_c} \mathbf{F}^{(t)}_{i,j,k},\, \mathbf{F}^{(t)}\big), \tag{4}$$

where $\cup$ is the concatenation operator. $v^{(t)}_{i,j}$ is the element in normalized branch weights $v^{(t)} \in \mathbb{R}^{N_c\cdot N_d}$, which are obtained by global pooling and multi-layer perceptron based on a shallow layer in semantic segmentation branches. As a kind of attention weights, $v^{(t)}$ could be adjusted adaptively based on each input. The skip connection in squeeze gate $g_{j,k}$ connects $\mathbf{F}^{(t)}$ with the output to avoid gradient vanishing. The integrated feature map $\mathbf{F}^{(t)}_{j,k}$ embeds spatial information of various area sizes *w.r.t.* each index j. Given each index k, $\{\mathbf{F}^{(1)}_{j,k}\}_{1\leq j\leq N_d}$ and $\{\mathbf{F}^{(2)}_{j,k}\}_{1\leq j\leq N_d}$ compose a pair of siamese spatial feature pyramid.

Further to make receptive fields flexible and adaptive to each input during the changed region extraction, we design dense connected architectures to link $\mathbf{F}^{(1)}_{j,k}$ and $\mathbf{F}^{(2)}_{j,k}$ across siamese branches. Given index k, for $1 \leq j_1, j_2 \leq N_d$, we have

$$\mathbf{M}_{j_1,j_2,k} = w^{(1)}_k \cdot \mathbf{F}^{(1)}_{j_1,k} - w^{(2)}_k \cdot \mathbf{F}^{(2)}_{j_2,k}\,, \tag{5}$$

where $w^{(1)}_k, w^{(2)}_k \in \mathbb{R}^{N_r}$ are normalized branch weights calculated from shallow layers in semantic segmentation branches. Features maps $\mathbf{F}^{(1)}_{j_1,k}$ and $\mathbf{F}^{(2)}_{j_2,k}$ are mostly generated from locally asymmetric structures and involve different sizes of image areas *w.r.t* each input, which makes $\{\mathbf{M}_{j_1,j_2,k}\}_{1\leq k\leq N_r}$ *locally asymmetric in terms of spatial information*. After the convolution operations, $\{\mathbf{M}_{j_1,j_2,k}\}_{1\leq k\leq N_r}$ is passed through the decoder in change detection branch to calculate the change probability map $\mathcal{C}_{I_1,I_2}$, where N_r represents the length of aSP.

Besides, we fuse $\mathbf{F}^{(t)}_{j,k}$ along index j, saying that for $t = 1, 2$,

$$\mathbf{F}^{(t)}_k = h_k\big(w^{(t)}_k \cdot \cup_{j=1}^{N_d+1} \mathbf{F}^{(t)}_{j,k},\, \mathbf{F}^{(t)}_s\big), \tag{6}$$

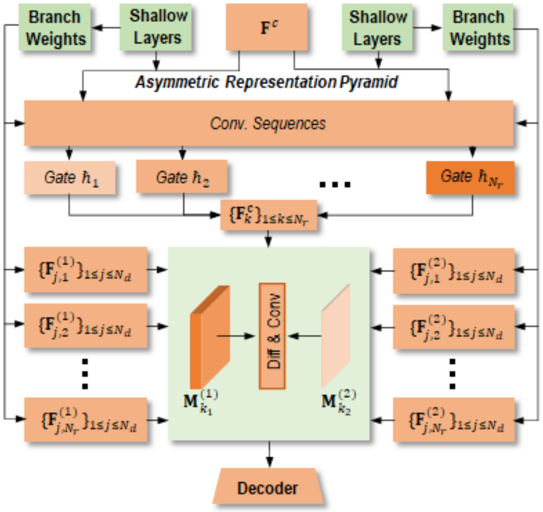
Branch Weights
Shallow Layers
$\mathbf{F}^c$
Shallow Layers
Branch Weights
Asymmetric Representation Pyramid
Conv. Sequences
Gate $\hbar_1$
Gate $\hbar_2$
...
Gate $\hbar_{N_r}$
$\{\mathbf{F}^c_k\}_{1\leq k\leq N_r}$
$\{\mathbf{F}^{(1)}_{j,1}\}_{1\leq j\leq N_d}$
$\{\mathbf{F}^{(1)}_{j,2}\}_{1\leq j\leq N_d}$
$\{\mathbf{F}^{(1)}_{j,N_r}\}_{1\leq j\leq N_d}$
Diff & Conv
$\mathbf{M}^{(1)}_{k_1}$
$\mathbf{M}^{(2)}_{k_2}$
$\{\mathbf{F}^{(2)}_{j,1}\}_{1\leq j\leq N_d}$
$\{\mathbf{F}^{(2)}_{j,2}\}_{1\leq j\leq N_d}$
$\{\mathbf{F}^{(2)}_{j,N_r}\}_{1\leq j\leq N_d}$
Decoder

Fig. 4. The proposed aRP. Index k_1, k_2 indicate different feature representation capabilities. Receiving spatial feature pyramids from aSP, aRP fuses asymmetric representation feature pairs with various representation capabilities.

where h_k represents the squeeze gate concatenating global features, denoted as $\mathbf{F}^{(t)}_{N_d+1,k}$, with $\{\mathbf{F}^{(t)}_{j,k}\}_{j<N_d+1}$. Also, h_k connects shallow features $\mathbf{F}^{(t)}_s$ deriving from I_t to avoid gradient vanishing. In this way, $\mathbf{F}^{(t)}_k$ with various receptive fields is passed through the siamese decoder in siamese semantic segmentation branches to obtain semantic probability maps $\mathcal{M}_1$, $\mathcal{M}_2$. Then, semantic prediction maps $\mathcal{L}_{\mathcal{M}1}$, $\mathcal{L}_{\mathcal{M}2}$ can be obtained by selecting the semantic category with the largest probability *w.r.t.* each position.

3) Asymmetric Representation Pyramid (aRP): As illustrated in Fig.4, given deep features $\mathbf{F}^c$ in change detection branch, we leverage convolution sequences and squeeze gates to obtain $\mathbf{F}^c_{j,k}$ in the same way as that obtaining $\mathbf{F}^{(t)}_{j,k}$. Moreover, We apply normalized branch weights $w^c = (w^c_1, \cdots, w^c_{N_r}) \in \mathbb{R}^{N_r}$ to integrate $\mathbf{F}^c_{j,k}$ along the index j. Concretely, we have

$$\mathbf{F}^c_k = \hbar_k\big(w^c_k \cdot \cup_{j=1}^{N_d+1}\mathbf{F}^c_{j,k},\, \mathbf{F}^c_s\big), \tag{7}$$

where $\hbar_k$ is also designed squeeze gate consisting of convolution layers and skip connections. Similarly, $\hbar_k$ concatenates global features, denoted as $\mathbf{F}^c_{N_d+1,k}$, with $\{\mathbf{F}^c_{j,k}\}_{j<N_d+1}$, while $\hbar_k$ connects shallow features $\mathbf{F}^c_s$ calculated from I_1, I_2 with the output.

As illustrated in Fig.4, we integrate $\mathbf{F}^c_k$ with $\{\mathbf{F}^{(1)}_{j,k}\}_{1\leq j\leq N_d}$ and $\{\mathbf{F}^{(2)}_{j,k}\}_{1\leq j\leq N_d}$ $w.r.t$ each index k by concatenation and convolution to obtain $\mathbf{M}^{(1)}_k$ and $\mathbf{M}^{(2)}_k$ respectively. All these $\{\mathbf{M}^{(1)}_k\}_{1\leq k\leq N_r}, \{\mathbf{M}^{(2)}_k\}_{1\leq k\leq N_r}$ compose a pair of siamese representation feature pyramid in aRP.

Further to make feature representation capabilities flexible and adaptive to each input, we also link $\mathbf{M}^{(1)}_k, \mathbf{M}^{(2)}_k$ in pairs. For $1 \leq k_1, k_2 \leq N_r$, we have

$$\mathbf{M}^c_{k_1,k_2} = \mathbf{M}^{(1)}_{k_1} - \mathbf{M}^{(2)}_{k_2}. \tag{8}$$

$\mathbf{M}^{(1)}_{k_1}, \mathbf{M}^{(2)}_{k_2}$ contain different representation capabilities *w.r.t.* the index k_1, k_2, which make $\mathbf{M}^c_{k_1,k_2}$ implicates different representation capabilities *w.r.t.* each input, namely *locally asymmetric in terms of representation capabilities*. Finally, after convolution operations, $\{\mathbf{M}^c_{k_1,k_2}\}_{1\leq k_1,k_2\leq N_r}$ is also passed through the decoder in change detection branch to calculate the change probability map $\mathcal{C}_{I_1,I_2}$.

Given the semantic prediction maps $\mathcal{L}_{\mathcal{M}1}, \mathcal{L}_{\mathcal{M}2}$ and change probability map $\mathcal{C}_{I_1,I_2}$, we can formulate our model as follows:

$$\forall \mathbf{p},\ f_{I_1,I_2}(\mathbf{p}) = \begin{cases} (0,0), & \mathcal{C}_{I_1,I_2}(\mathbf{p}) < \tau, \\ (\mathcal{L}_{\mathcal{M}1}(\mathbf{p}), \mathcal{L}_{\mathcal{M}2}(\mathbf{p})), & \text{otherwise.} \end{cases}$$

4) Loss Function: We utilize fully supervised learning to optimize parameters in the proposed ASN. Specifically, given the ground truth $\mathcal{G}(\mathbf{p}, I_1, I_2)$, which not only locates changed pixels but also indicates semantic categories, we can get ground truth for semantic probability maps ($\mathcal{L}_{\mathcal{G}1}$, $\mathcal{L}_{\mathcal{G}2}$) and change probability map $\mathcal{L}_{\mathcal{G}_c}$. Then, we have

$$\mathcal{L} = \alpha\mathcal{E}(\mathcal{M}_1, \mathcal{L}_{\mathcal{G}1}) + \beta\mathcal{E}(\mathcal{M}_2, \mathcal{L}_{\mathcal{G}2}) + \mathcal{E}(\mathcal{C}_{I_1,I_2}, \mathcal{L}_{\mathcal{G}_c}), \tag{9}$$

where $\mathcal{E}$ is the cross entropy function. α and β are loss weights to adjust respective loss terms. Stochastic gradient descent (SGD) is then used to reduce the total loss and obtain the optimal parameters, through which we train all branches simultaneously.

5) Adaptive Threshold Learning: After the traditional training process, we design an adaptive threshold learning module to search optimal thresholds based on different outputs. Given each probability map before softmax, denoted as $\mathcal{M}^{raw}_1$, $\mathcal{M}^{raw}_2$ and $\mathcal{C}^{raw}_{I_1,I_2}$, for $t = 1, 2$, we have

$$\hat{\mathcal{M}}^{raw}_t = \mathcal{M}^{raw}_t + \gamma\psi_1(\mathcal{M}^{raw}_t), \tag{10}$$

$$\hat{\mathcal{C}}_{I_1,I_2} = s_{max}(\mathcal{C}^{raw}_{I_1,I_2} + \gamma\psi_2(\cup_{t=1}^2 \hat{\mathcal{M}}^{raw}_t)), \tag{11}$$

where ψ_1, ψ_2 are both convolution layer series with length of 2 and s_{max} represents softmax function. We fix the parameters in the whole model except ψ_1, ψ_2 and re-train the model by utilizing the total loss in Sec. III-D4 calculated based on $s_{max}(\hat{\mathcal{M}}^{raw}_t)$ and $\hat{\mathcal{C}}_{I_1,I_2}$ with categorical weights. Thus, the model can be re-formulated as

$$\hat{f}_{I_1,I_2}(\mathbf{p}) = \begin{cases} (0,0), & \hat{\mathcal{C}}_{I_1,I_2}(\mathbf{p}) < \tau, \\ (\mathcal{L}_{\hat{\mathcal{M}}1}(\mathbf{p}), \mathcal{L}_{\hat{\mathcal{M}}2}(\mathbf{p})), & \text{otherwise.} \end{cases} \tag{12}$$

$\mathcal{L}_{\hat{\mathcal{M}}1}(\mathbf{p}), \mathcal{L}_{\hat{\mathcal{M}}2}(\mathbf{p})$ are the updated semantic prediction maps obtained based on $s_{max}(\hat{\mathcal{M}}^{raw}_1), s_{max}(\hat{\mathcal{M}}^{raw}_2)$ respectively. During the parameter optimization, ATL explores the logical relationships between semantic categories within considered neighbourhood, which can adaptively update the semantic prediction maps and revise the threshold deflections caused by label imbalance.

IV. The Second Dataset

Although several change detection datasets have been proposed [3], [36], [37], [38], [39], [41], only few of them contain land-cover categorical information, which is needed for SCD. A natural way to create an SCD dataset is comparing multi-temporal land-cover maps in the same geographic locations [3], which would neglect change types such as the demolition and reconstruction of *buildings*. Moreover, due to the required dense labors, a large-scale SCD dataset is hard to acquire, while the annotations of land-cover classes also require professional knowledge.

In order to set up a new benchmark for SCD problems with adequate quantities, sufficient categories and proper annotation methods, in this paper we present SECOND, a well-annotated semantic change detection dataset. Different from the datasets used in [12], [31], to ensure data diversity, we firstly collect 4662 pairs of aerial images from several platforms and sensors. These pairs of images are distributed over the cities such as Hangzhou, Chengdu, and Shanghai. Each image has size 512 × 512 and is annotated at the pixel level. The annotation of SECOND is carried out by an expert group of earth vision applications, which guarantees high label accuracy. Moreover, in contrast with [3], we utilize land-cover map pairs and *non-change* masks to represent the change types. The *non-change* masks equal the ground truth for BCD and are used to blacken the unchanged pixels in siamese semantic segmentation branches during the training process, which can avoid semantic categorical ambiguity. Through this annotation method, we can distinguish *non-change* pixels from changed pixels between the same land-cover class.

For the change type in the SECOND dataset, we focus on 6 main land-cover classes, *i.e.*, *non-vegetated ground surface, tree, low vegetation, water, buildings* and *playgrounds*, that are frequently involved in natural and man-made geographical changes [42], [43], [44], [45]. It is worth noticing that, in the new dataset, non-vegetated ground surface (*n.v.g. surface* for short) mainly corresponds to *impervious surface* and *bare land*. In summary, these 6 selected land-cover classes result in 30 common change types (including *non-change*). Through the random selection of image pairs, the SECOND reflects real distributions of land-cover classes when changes occur. Several samples of SECOND are displayed in Fig.5, where we can see the data diversity and label accuracy. The SECOND dataset is available at http://www.captain-whu.com/project/SCD.

V. Evaluation Metrics

A. Overall accuracy and Kappa coefficient

Existing works often utilize Overall Accuracy (OA) and Kappa coefficient (κ) that are commonly used for measuring classification performance to evaluate change detection algorithms [3], [12]. Given a confusion matrix $Q = \{q_{ij}\}$, where q_{ij} indicates the number of pixels that are identified as the i-th change type and actually belong to the j-th change type (*non-change* is set as the first change type), then OA is defined as

$$\mathrm{OA} \triangleq \rho = \sum_{i=1}^{C} q_{ii} / \sum_{i=1}^{C}\sum_{j=1}^{C} q_{ij}, \tag{13}$$

where C is the total number of change types. Due to the equivalence of each pixel in the calculation of OA, dominant *non-change* pixels would cause unreasonable scores. While, as a statistic calculated from confusion matrix (a kind of contingency table), κ measures the consistency between outputs and labels, which is less affected by the label imbalance [46]. More precisely, we have

$$\kappa = (\rho - \eta)/(1-\eta), \tag{14}$$

where $\eta = \sum_{j=1}^{C}(q_{j+}\cdot q_{+j})/(\sum_{i=1}^{C}\sum_{j=1}^{C} q_{ij})^2$ with q_{j+} and q_{+j} being as the row sum and column sum of the confusion matrix Q.

However, the dominant *non-change* pixels still mislead the scores obtained by κ. Given a change detection data sample, *i.e.* a pair of images and a sequence of change detection results, we collect visual scores between 0 and 1 *w.r.t.* each result from 11 remote sensing image interpretation experts. Meanwhile, we calculate evaluation scores of each result based on OA and κ. As illustrated in Fig. 6, the collapse model with constant *non-change* predictions would get unreasonable high scores in OA. Besides, although κ tends to zero when the model gradually collapses, the score of second to last result still gets 0.23 in κ, which is too high compared with human scoring.

B. Separated Kappa (SeK) coefficient

In order to alleviate the influence of label imbalance, we utilize mIOU to evaluate BCD results and propose a SeK coefficient to evaluate SCD results. Specifically, given a confusion matrix Q, we have

$$\mathrm{IOU}_1 = q_{11}/(\sum_{i=1}^{C} q_{i1} + \sum_{j=1}^{C} q_{1j} - q_{11}), \tag{15}$$

$$\mathrm{IOU}_2 = \sum_{i=2}^{C}\sum_{j=2}^{C} q_{ij}/(\sum_{i=1}^{C}\sum_{j=1}^{C} q_{ij} - q_{11}), \tag{16}$$

where IOU_1 measures the identification of *non-change* pixels and IOU_2 evaluates the extraction of changed regions. Then, we have

$$\mathrm{mIOU} = \frac{1}{2}(\mathrm{IOU}_1 + \mathrm{IOU}_2). \tag{17}$$

Involving in IOU_2, mIOU considers more about changed regions.

On the other hand, the true positive of *non-change* pixels q_{11} always dominates the calculation of κ. Thus, we separate q_{11} in the calculation of SeK. We also utilize IOU_2 to further emphasize changed pixels. Specifically, we define

$$\mathrm{SeK} = e^{(\mathrm{IOU}_2 - 1)} \cdot (\hat{\rho} - \hat{\eta})/(1-\hat{\eta}), \tag{18}$$

with

$$\hat{\rho} = \sum_{i=2}^{C} q_{ii}/(\sum_{i=1}^{C}\sum_{j=1}^{C} q_{ij} - q_{11}),$$

$$\hat{\eta} = \sum_{j=1}^{C}(\hat{q}_{j+}\cdot\hat{q}_{+j})/(\sum_{i=1}^{C}\sum_{j=1}^{C} q_{ij} - q_{11})^2,$$

where $\hat{q}_{j+}$ and $\hat{q}_{+j}$ represent the row sum and column sum of confusion matrix without q_{11}. The exponential form enlarges the discernibility compared with simple multiplication when evaluating models with better performance.

As illustrated in Fig.6, compared with κ and OA, models with apparently poor performances on small change types would get low scores in SeK no matter how good the performances on BCD are. Moreover, the Mean Square Error (MSE) between SeK and human scores is 0.003. While, MSE *w.r.t.* OA and κ are 0.212 and 0.028 respectively, which further validates the rationality of SeK.

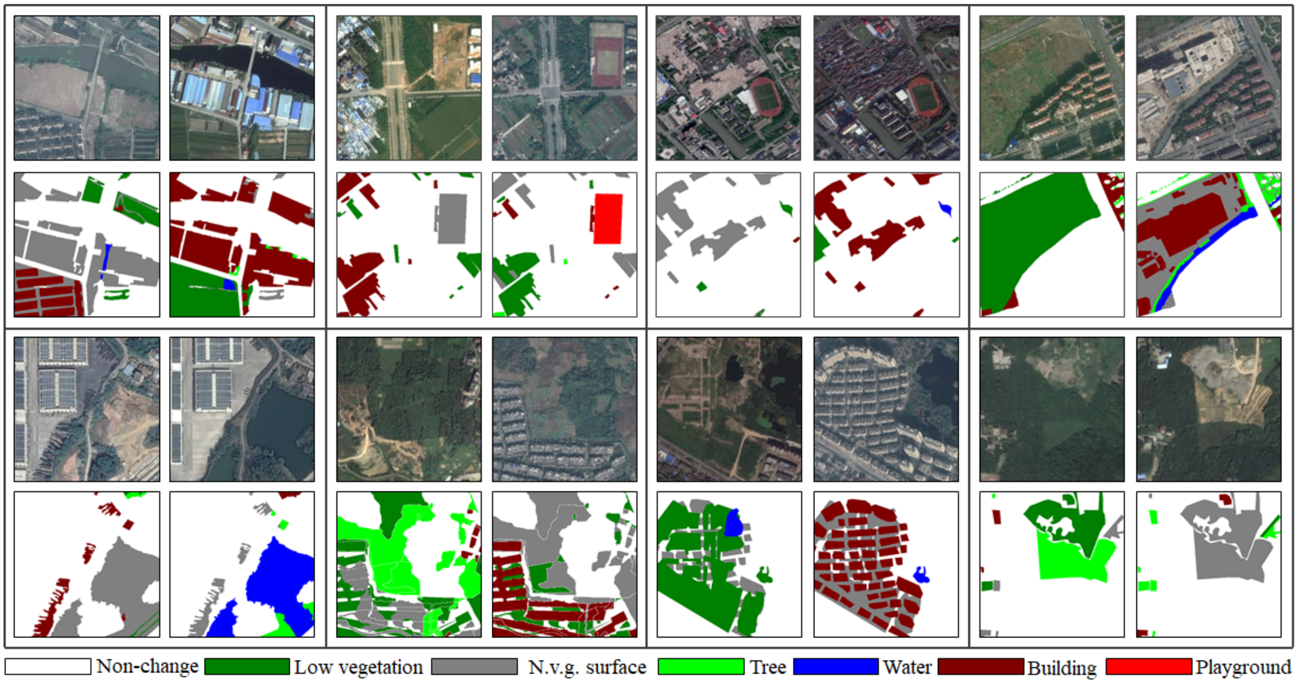

Fig. 5. Several samples of our proposed SECOND dataset. Color white indicates *non-change* regions, while other colors indicate different land-cover classes. Ground truth for SCD can be obtained by comparing the annotated land-cover classes.

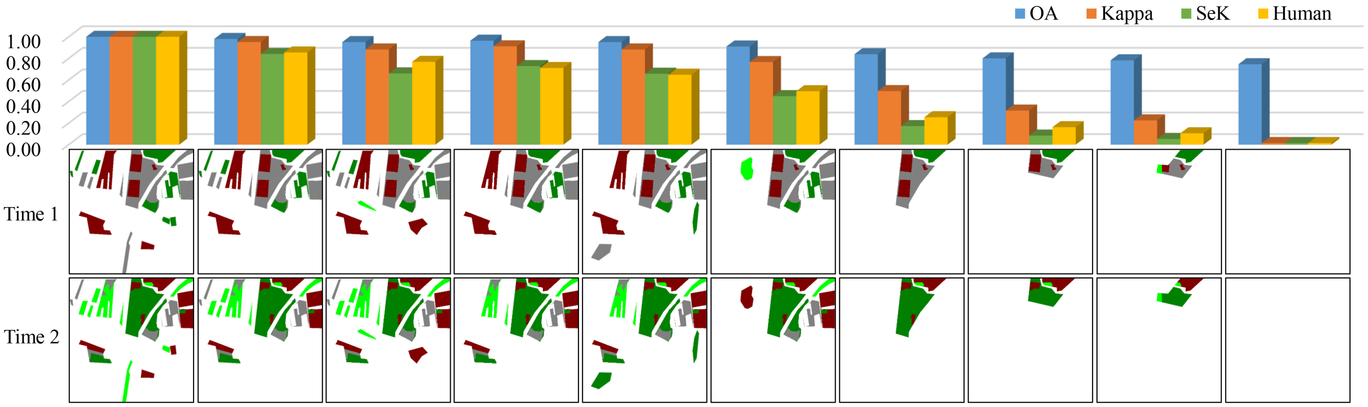

Fig. 6. Comparison between SeK and other metrics. SeK alleviates the influence of label imbalance caused by *non-change* pixels and shows most similarity with human scoring, while outputs with dominant predictions of *non-change* get unreasonable high scores in OA and κ.

VI. Experiments and Analysis

In this section, we evaluate the proposed ASN on the SECOND dataset. We first clarify the experiment settings in Sec. VI-A. Then, in Sec. VI-B and Sec. VI-C, we discuss several existing structures and compare ASN with several methods, including natural extensions of classical BCD algorithms and the state-of-the-art SCD algorithms [3], [27] using backbones with different basic blocks [47], [48], [49]. Further in Sec. VI-D, we demonstrate the effectiveness of our proposed modules by feature visualization. Finally, in Sec. VI-E, we remove the aSP, aRP and asymmetric feature pairs to make comparison with original ASN and verify the merits of each term in the proposed model.

A. Experiment Settings

The change detection algorithms involved in the comparison experiments are as follows:

- FC-EF[27]: a BCD algorithm using single encoder-decoder structure.
- FC-conc[27]: a BCD algorithm using siamese encoders followed with single decoder branch and concatenation skip connections from the encoder to the decoder.
- FC-diff[27]: a BCD algorithm using siamese encoders followed with single decoder branch and difference skip connections from the encoder to the decoder.
- HRSCD.str1[3]: an algorithm corresponding to the intuitive solution to SCD problem discussed in Sec. III-B.
- HRSCD.str2[3]: a SCD algorithm using single encoder-decoder structure.
- HRSCD.str3[3]: a SCD algorithm using siamese semantic segmentation branches with change detection branch.
- HRSCD.str4[3]: a SCD algorithm using siamese semantic segmentation branches with change detection branch and difference skip connections from siamese encoders to the decoder of change detection branch.

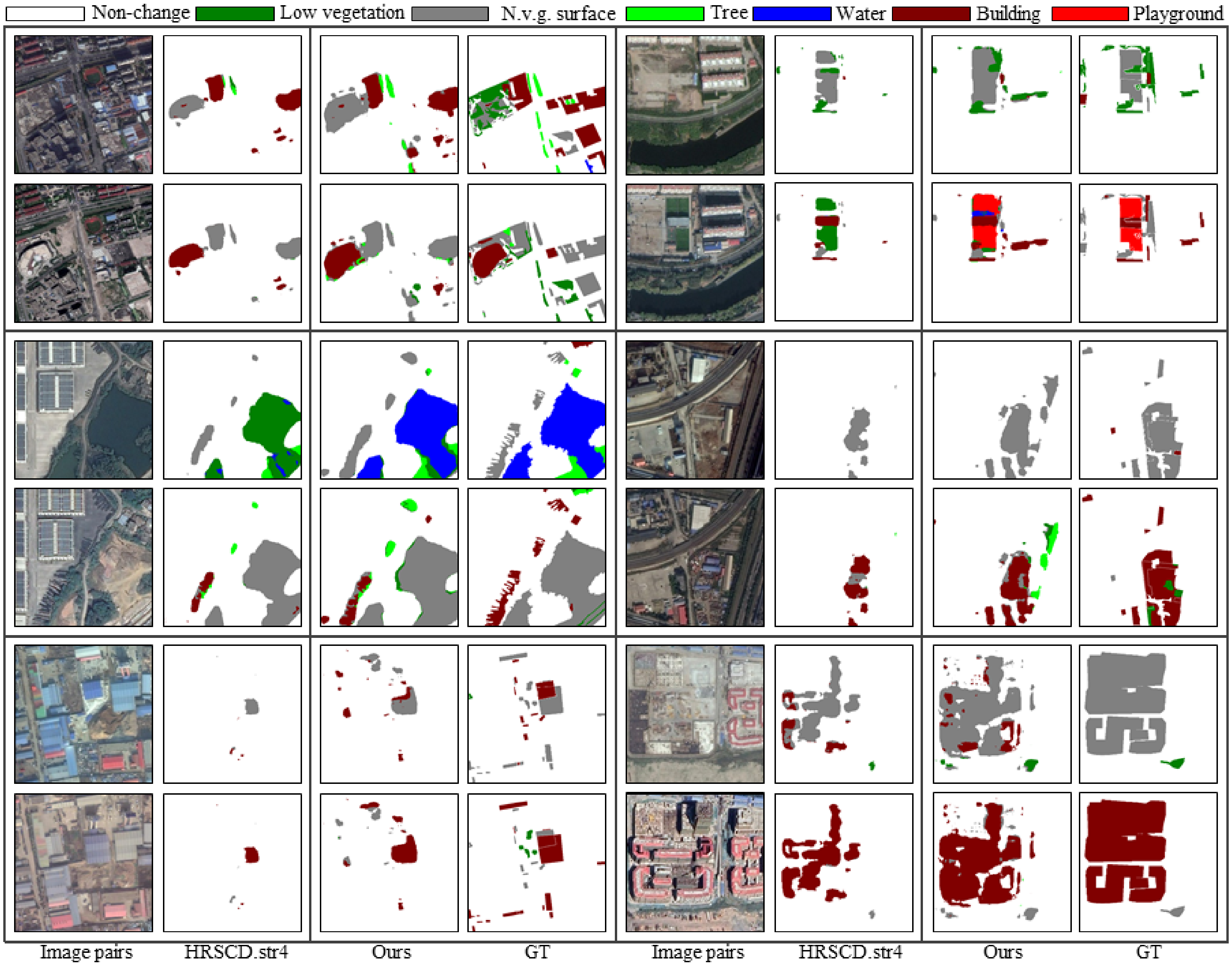

Fig. 7. Visual results of comparison with state-of-the-art method when the encoder is built on residual blocks. We mask the semantic prediction maps with change maps to represent the prediction of change type in each position, where our proposed ASN can better identify land-cover classes and alleviate false identifications of *non-change* pixels.

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS WHEN THE ENCODER IS BUILT ON RESIDUAL BLOCKS.

Methods	MS __ Flip __		MS ✓ Flip __		MS ✓ Flip ✓	
	mIOU	SeK	mIOU	SeK	mIOU	SeK
FC-EF [27]	59.3	5.7	59.3	5.7	59.0	5.9
FC-conc [27]	63.3	9.1	62.8	9.2	62.9	9.4
FC-diff [27]	61.9	8.8	60.9	8.6	61.0	8.7
HRSCD.str1 [3]	29.3	4.6	29.8	4.9	29.8	4.9
HRSCD.str2 [3]	59.7	6.3	59.4	6.5	59.4	6.6
HRSCD.str3 [3]	62.3	8.9	62.0	9.1	62.1	9.2
HRSCD.str4 [3]	67.5	13.7	67.8	14.4	67.9	14.5
ASN (Ours)	69.0	15.2	69.5	16.1	69.7	16.2
ASN-ATL (Ours)	**69.1**	**15.5**	**69.8**	**16.5**	**70.0**	**16.8**

ASN keeps the same basic architecture as HRSCD.str4, which contains 10 basic blocks in encoders and 13 basic blocks in decoders. We replace the difference skip connections with simple summations in ASN to reduce model sizes. The proposed aSP and aRP are embedded between the 6^{th} and 7^{th} basic blocks in the decoder branches. ASN-ATL applys adaptive threshold learning module on ASN. All the experiments are implemented on 4 Pascal V100 with memory of 16G based on Pytorch. All evaluated models are trained and tested under the same conditions without pre-trained models and other post-processings.

In the training process, SGD is utilized to search optimal parameters for 50 epochs, while extra parameters in ASN-ATL would be trained for 20 epochs with other parameters fixed. Random flip and random scale between 0.5 and 2 are utilized as the data augmentation. The initial learning rate is set as 0.005 and 'poly' policy is employed with the power of 0.9. Batch size is set as 4. Also, the momentum is set as 0.9 and weight decay is set as 0.0001. In the testing process, we apply flip strategy and multi-scale (MS) testing with 6 scales which are 0.5, 0.75, 1.0, 1.25, 1.5 and 1.75.

As for the hyper-parameters in ASN, considering in the image size in SECOND, we set spatial hyper-parameters $\{d_1, \cdots, d_{N_d}\}$ as $\{0, 6, 12\}$ to make feature receptive fields in aSP vary from local to the whole image. Besides, multiplication hyper-parameters $\{r_1, \cdots, r_{N_r}\}$ and channel hyper-parameters $\{c_1, \cdots, c_{N_c}\}$ are set as $\{16, 32, 64\}$ and $\{1, 2, 3, 4, 5\}$ to make features in aSP and aRP contain bal-

TABLE II

DETAIL CATEGORICAL RESULTS WITH THREE KINDS OF TESTING STRATEGIES WHEN THE ENCODER IS BUILT ON RESIDUAL BLOCKS. TOP SHEETS ARE WITHOUT TESTING STRATEGY. MIDDLE SHEETS ARE WITH MS TESTING STRATEGY. BOTTOM SHEETS ARE WITH MS AND FLIP TESTING STRATEGY. THE CATEGORICAL SEK IS LISTED IN THE MATRICES, WHILE THE CATEGORICAL INTERSECTION OVER UNION (IOU$_1$, IOU$_2$) OF BINARY CHANGE DETECTION IS LISTED BELOW EACH MATRICES.

Residual	HRSCD.str1(intuitive solution)						HRSCD.str4						ASN-ATL (Ours)					
	Low vegetation	N.v.g. surface	Building	Tree	Water	Playground	Low vegetation	N.v.g. surface	Building	Tree	Water	Playground	Low vegetation	N.v.g. surface	Building	Tree	Water	Playground
Low vegetation	–	30.0	22.7	19.4	21.7	3.3	–	41.7	30.8	22.2	19.9	8.4	–	**42.0**	**32.2**	**28.3**	**20.4**	**34.2**
N.v.g. surface	31.3	–	31.3	20.7	15.0	4.7	**44.1**	–	42.1	28.5	24.6	6.0	43.7	–	**44.7**	**30.4**	**29.5**	**41.9**
Building	21.9	31.3	21.4	12.6	19.2	0.0	**35.1**	47.2	26.1	18.1	**33.8**	0.0	33.1	**48.3**	**27.1**	**24.9**	31.1	**17.8**
Tree	16.9	19.1	16.1	–	1.1	0.0	12.0	26.0	**24.5**	–	**12.6**	0.0	**20.2**	**27.7**	23.7	–	5.9	0.0
Water	19.5	16.1	19.0	10.0	–	–	13.7	26.0	**29.2**	13.0	–	–	**25.9**	**29.8**	27.8	**21.7**	–	–
Playground	0.1	3.4	0.0	0.0	–	–	1.1	1.8	0.4	0.0	–	–	**17.1**	**29.7**	**16.5**	**1.3**	–	–
Non-change	IOU$_1$: 34.9			IOU$_2$: 23.8			IOU$_1$: **86.6**			IOU$_2$: 48.4			IOU$_1$:85.9			IOU$_2$:**52.4**		
Low vegetation	–	30.6	23.1	20.0	**22.0**	1.2	–	42.4	31.6	21.8	14.4	4.7	–	**43.4**	**32.5**	**28.5**	21.2	**37.6**
N.v.g. surface	31.5	–	31.9	21.4	14.3	1.7	44.7	–	42.8	29.2	19.8	4.8	**45.4**	–	**45.4**	**31.5**	**31.5**	**43.3**
Building	22.4	32.3	22.5	13.9	20.5	0.0	**36.1**	48.3	27.9	21.4	31.2	0.0	34.4	**49.4**	**28.7**	**25.9**	**36.0**	**20.1**
Tree	17.8	19.8	16.9	–	0.9	0.0	12.0	26.8	**25.6**	–	7.8	0.0	**21.0**	**28.7**	24.8	–	**10.3**	0.0
Water	16.4	15.9	17.8	11.0	–	–	7.5	22.0	26.3	13.0	–	–	**26.8**	**31.4**	**28.8**	**21.3**	–	–
Playground	0.0	1.9	0.0	0.0	–	–	0.1	1.7	0.3	0.0	–	–	**15.5**	**28.0**	**20.8**	**0.1**	–	–
Non-change	IOU$_1$: 35.6			IOU$_2$: 24.0			IOU$_1$: **86.9**			IOU$_2$: 48.8			IOU$_1$: 86.4			IOU$_2$: **53.3**		
Low vegetation	–	30.8	23.2	20.1	**21.9**	1.2	–	42.6	31.8	22.2	11.3	3.5	–	**43.6**	**32.7**	**28.3**	21.5	**37.7**
N.v.g. surface	31.6	–	31.9	21.4	14.1	2.3	45.0	–	42.9	29.4	18.9	4.4	**45.7**	–	**45.7**	**31.8**	**32.5**	**43.3**
Building	22.5	32.4	22.4	14.2	20.5	0.0	**36.3**	48.4	28.2	22.0	29.3	0.0	34.6	**49.7**	**29.1**	**26.4**	**35.5**	**20.1**
Tree	18.0	20.0	16.8	–	0.8	0.0	12.8	26.7	**26.2**	–	6.4	0.0	**21.3**	**28.6**	25.2	–	**13.4**	0.0
Water	16.1	15.4	18.1	9.6	–	–	8.7	22.2	25.7	15.4	–	–	**27.4**	**32.1**	**29.2**	**22.0**	–	–
Playground	0.0	2.4	0.0	0.0	–	–	0.0	1.6	0.3	0.0	–	–	**15.9**	**27.9**	**24.9**	**0.3**	–	–
Non-change	IOU$_1$: 35.6			IOU$_2$: 24.0			IOU$_1$: **87.0**			IOU$_2$: 48.8			IOU$_1$: 86.5			IOU$_2$: **53.5**		

TABLE III

COMPARISON WITH STATE-OF-THE-ART METHODS WHEN THE ENCODER IS BUILT ON XCEPTION BLOCKS.

Methods	MS __ Flip __		MS ✓ Flip __		MS ✓ Flip ✓	
	mIOU	SeK	mIOU	SeK	mIOU	SeK
FC-EF [27]	56.2	3.6	55.6	3.9	55.7	4.0
FC-conc [27]	61.1	7.3	60.4	7.4	60.5	7.5
FC-diff [27]	57.3	5.0	56.6	5.0	56.5	5.0
HRSCD.str1 [3]	29.3	4.7	30.1	4.9	30.2	5.0
HRSCD.str2 [3]	59.7	5.7	59.0	5.8	59.0	5.9
HRSCD.str3 [3]	62.1	8.4	61.9	8.7	61.9	8.8
HRSCD.str4 [3]	67.2	13.0	67.8	13.8	68.0	14.1
ASN (Ours)	68.4	14.3	68.8	15.2	69.0	15.5
ASN-ATL (Ours)	**69.0**	**14.9**	**69.6**	**15.9**	**69.9**	**16.3**

TABLE IV

COMPARISON WITH STATE-OF-THE-ART METHODS WHEN THE ENCODER IS BUILT ON SQUEEZE-AND-EXCITATION BLOCKS.

Methods	MS __ Flip __		MS ✓ Flip __		MS ✓ Flip ✓	
	mIOU	SeK	mIOU	SeK	mIOU	SeK
FC-EF [27]	60.5	6.9	60.1	6.9	60.1	6.9
FC-conc [27]	64.2	10.5	63.7	10.5	63.6	10.5
FC-diff [27]	63.1	9.7	62.3	9.6	62.3	9.7
HRSCD.str1 [3]	30.4	5.0	31.0	5.3	31.1	5.3
HRSCD.str2 [3]	62.3	8.3	61.9	8.3	61.9	8.3
HRSCD.str3 [3]	62.0	8.6	62.0	9.0	62.0	9.1
HRSCD.str4 [3]	67.9	14.6	68.1	15.3	68.2	15.4
ASN (Ours)	69.4	15.9	69.6	16.5	69.7	16.6
ASN-ATL (Ours)	**69.5**	**16.3**	**70.1**	**17.2**	**70.2**	**17.3**

anced representation capabilities with features in decoders. Then, the kernel size of convolution layers in aSP and aRP are all set as 3. We further set scalar threshold τ as 0.5. The hyper-parameter γ in ATL module is also set as 0.5, while we slightly adjust it to 0.4 for Xception backbone. For all the experiments, we split SECOND dataset into two subsets: 2968 sample pairs for training and 1694 sample pairs for testing. Finally, we set α and β in total loss as 1 in the training process.

B. Discussions on Basic Structures

As illustrated in Tab. I, HRSCD.str1 gets 29.8 in mIOU with full testing strategies, while FC-conc and FC-diff achieve 62.9 and 61.0 respectively. This result demonstrates that separate semantic segmentation application and false categorical independence assumption would limit the model performance, which could be addressed by the joint training of siamese network. Importantly, although FC-EF outperforms HRSCD.str1 dramatically in mIOU, they share similar results in SeK, which implicates that SeK would not simply increase merely with good BCD performance. The outperformances of FC-conc and FC-diff compared with FC-EF indicates that the siamese networks provide extra improvements compared with single encoder-decoder structure. Moreover, the outperformances of HRSCD.str3 and HRSCD.str4 show the superiority of separate

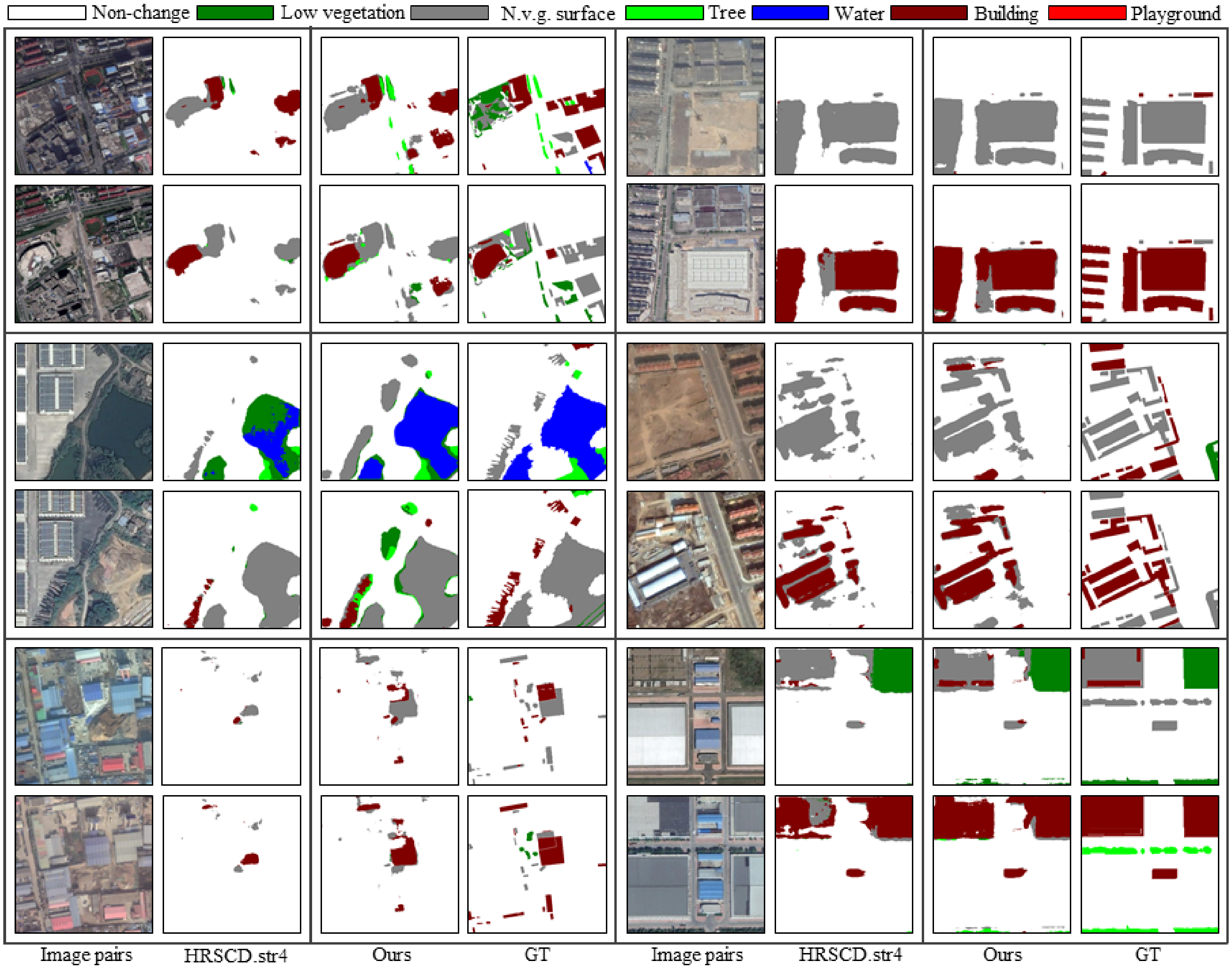

Fig. 8. Comparisons with state-of-the-art method when the encoder is built on Xception blocks. Left part of the figure shows the stable outperformance of ASN with different backbone, while the right part exhibits the generalization ability on different data samples.

changed region extraction and change type identification.

C. Comparison with State-of-the-art Algorithms

As shown in Tab.I, ASN-ATL achieves the best results, 69.1 in mIOU and 15.5 in SeK, without testing strategies. Then, with multi-scale strategy, ASN-ATL achieves 69.8 in mIOU and 16.5 in SeK, while the flip strategy provides extra 0.2 improvements in mIOU and 0.3 improvements in SeK. Our proposed model stably outperforms existing state-of-the-art algorithms, which achieves 2.1 improvements in mIOU and 2.3 improvements in SeK compared to HRSCD.str4 with full testing strategies. Also, as shown in Tab.I, ATL provides 0.3 improvement in mIOU and 0.6 improvement in SeK with full testing strategies. We can also see in Fig. 7, the proposed model can recover more details which leads to the better performance in SeK.

To further explore the detail comparison, we list the categorical SeK (calculated from 2 × 2 confusion matrix concentrating on each change type) and categorical intersection over union (IOU_1 and IOU_2) in Tab.II. Each row (except *non-change*) indicates land-cover class in the first image, while each column indicates land-cover class in the second image. Categorical SeK evaluates identification of all change types except *non-change*, while categorical intersection over union measures the results of *non-change* pixel extraction. We choose HRSCD.str4, HRSCD.str1 and ASN-ATL to be listed in Tab. II for detail comparsion. We can see that ASN-ATL achieves the best performance on an overwhelming majority of change types under all kinds of testing strategies.

TABLE V
ABLATION STUDY ON ASYMMETRIC FEATURE PAIRS WITH RESIDUAL BACKBONE. ASN-W/O-AF REPRESENTS RETAINING MULTI-SCALE STRUCTURES WITHOUT THE ASYMMETRIC FEATURE PAIR INTEGRATION.

Methods	MS __ Flip __		MS ✓ Flip __		MS ✓ Flip ✓	
	mIOU	SeK	mIOU	SeK	mIOU	SeK
ASN-w/o-AF	67.7	13.5	68.2	14.4	68.3	14.5
ASN	**69.0**	**15.2**	**69.5**	**16.1**	**69.7**	**16.2**

We then substitute all the basic residual blocks in the encoder with Xception blocks and Squeeze-and-Excitation blocks. As illustrated in Tab.III and Tab.IV, our proposed ASN still outperforms other models under all kinds of testing

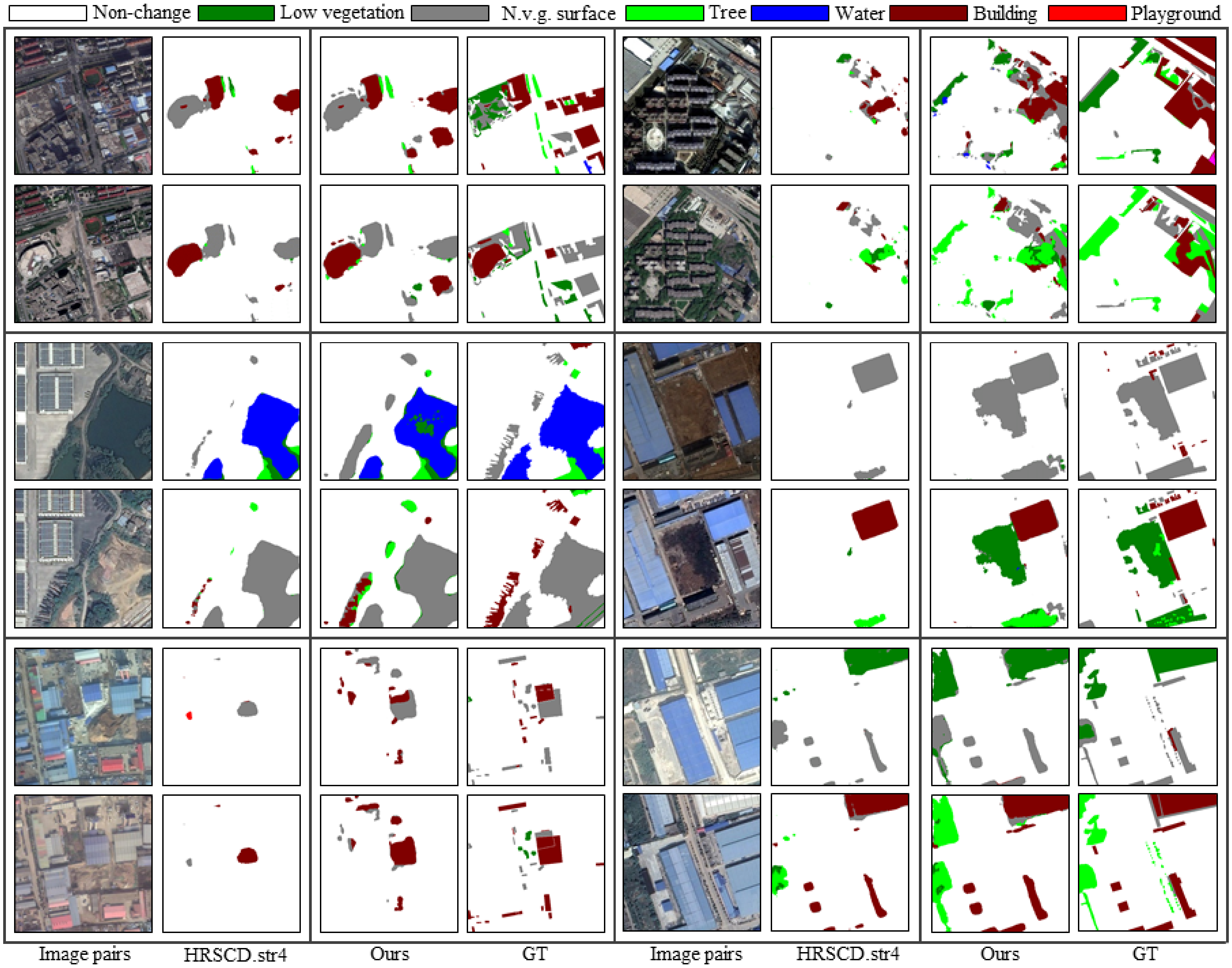

Fig. 9. Comparisons with state-of-the-art method when the encoder is built on Squeeze-and-Excitation blocks. ASN also shows stable outperformance and generalization ability.

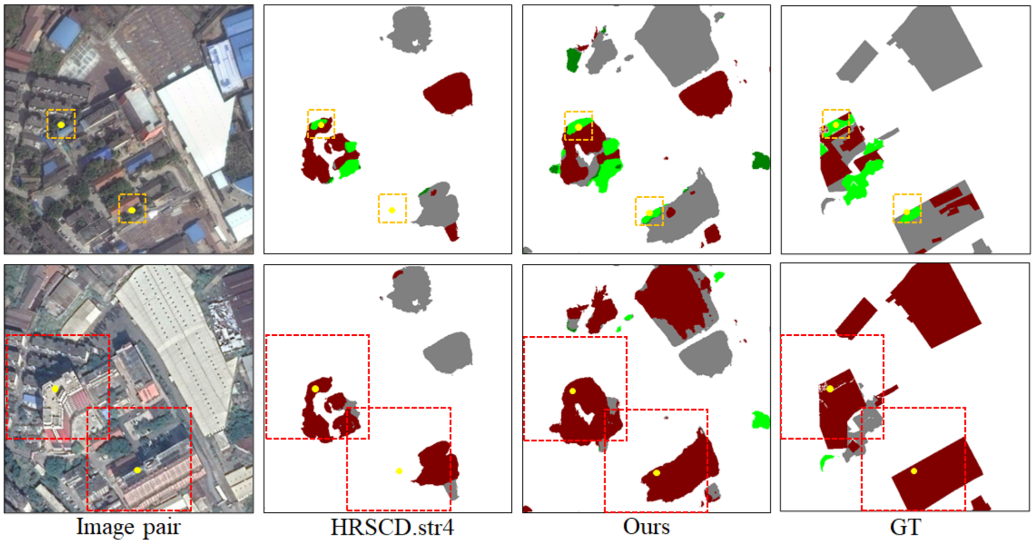

Fig. 10. Result visualizations of the case discussed in Sec. I when the encoder is built on residual blocks. ASN can extract changed regions and identify change types more precisely, while addressing these asymmetric changes.

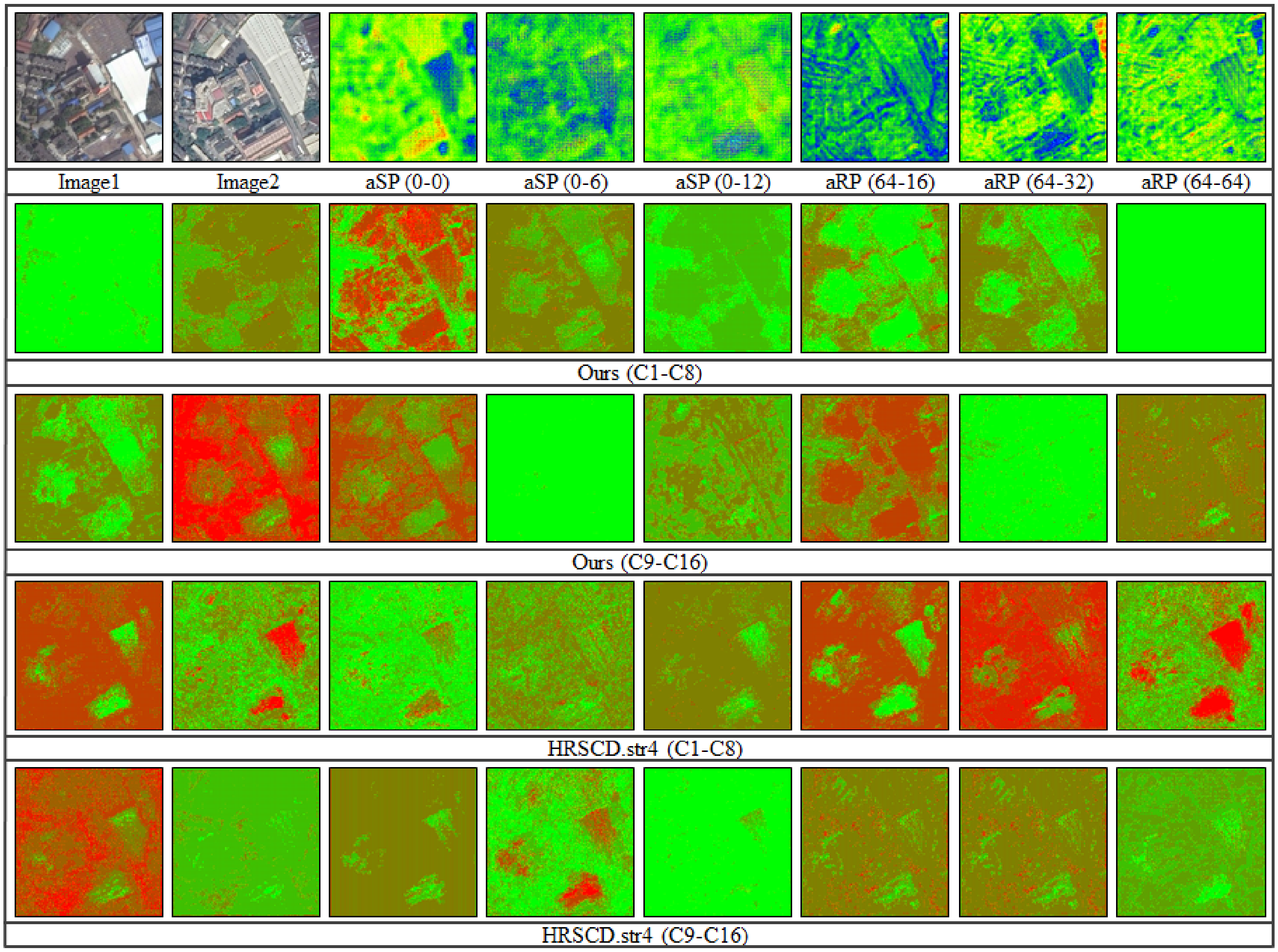

Fig. 11. Feature visualizing in ASN and HRSCD.str4 with residual blocks. The first row shows the mean value of normalized feature samples *w.r.t.* different integrations of asymmetric feature pairs. Features deriving from some asymmetric feature pairs, such as aRP(64-32) in the top right, contain more distinguishable values in regions of asymmetric changes. The rest rows show all 16 channels of corresponding features in the change detection branch of ASN and HRSCD.str4. C1-C8 means first 8 channels, while C9-C16 means last 8 channels. Similarly, features in ASN contain more distinguishable values in regions of asymmetric changes compared with HRSCD.str4.

TABLE VI
ABLATION STUDY OF ASN. W/O-R&S REPRESENTS WITHOUT ASP AND ARP, WHILE W/O-R REPRESENTS WITHOUT ARP.

Ablation	Residual			Xception			SENet		
study	w/o-R&S	w/o-R	ASN	w/o-R&S	w/o-R	ASN	w/o-R&S	w/o-R	ASN
mIOU	68.7	68.9	**69.7**	68.3	68.8	**69.0**	68.5	69.2	**69.7**
SeK	14.8	15.4	**16.2**	14.6	15.4	**15.5**	15.0	15.8	**16.6**

processes. Especially, under full testing strategies, ASN-ATL outperforms HRSCD.str4 by 1.9 in mIOU and 2.2 in SeK with the Xception blocks, while HRSCD.str4 is inferior to ASN-ATL by 2.0 in mIOU and 1.9 in SeK with Squeeze-and-Excitation blocks. Meanwhile, ATL provides 0.9 improvements in mIOU and 0.8 improvements in SeK with Xception blocks, while 0.5 improvements in mIOU and 0.7 improvements in SeK can be seen with Squeeze-and-Excitation blocks. We can conclude that ASN keeps the outperformances with all kinds of backbones and ATL does improve the model performances. Moreover, Fig.7-9 report the visual results of HRSCD.str4 and ASN with all three utilized basic blocks. We list the results of the same data samples in the left parts of these figures, which shows the stable effectiveness of ASN with different encoder backbones. Also, we list the results of different data samples in the right parts to exhibit the generalization abilities of the proposed modules.

D. Visualization of Learned Features

Recalling the discussion in Sec.I, we can see in Fig.10 that ASN alleviates the categorical ambiguity caused by asymmetric changes. Moreover, as illustrated in Fig.11, samples of feature maps in ASN are more likely to contain distinguishable response values in changed regions compared with HRSCD.str4. Specifically, the first row shows some mean value maps of normalized features integrated by different feature pairs. Among these samples, we can see that some

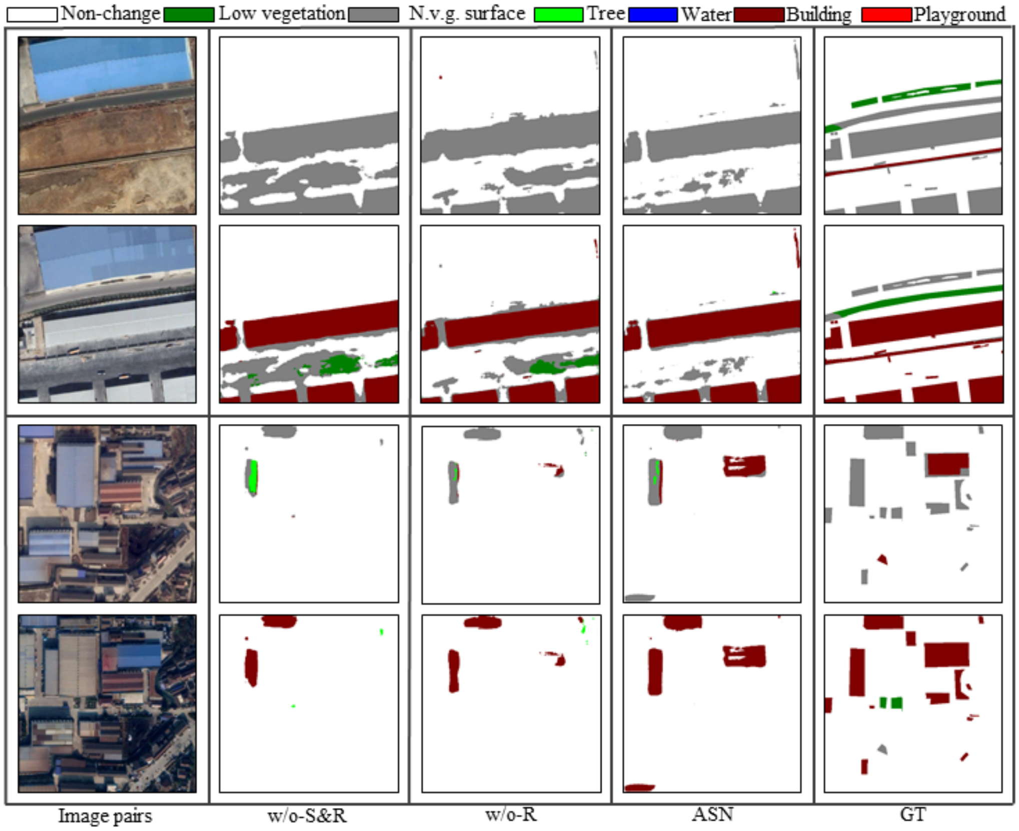

Fig. 12. Results of ablation study with residual blocks. With our proposed modules, the model can better extract illegible change types and alleviate the false identifications of changed pixels.

feature maps in aSP integrated by asymmetric feature pairs (0-6 and 0-12) contain more distinguishable response values in asymmetric changes (framed out in Fig.10). Similarly, feature map calculated from asymmetric feature pairs in aRP (64-32) also contains more distinguishable values in changed regions. Moreover, we can conclude from the rest rows of Fig.11 that feature maps integrated by asymmetric feature pairs in change detection branch are also more likely to contain different response values in asymmetric changes, while corresponding features in HRSCD.str4 contain similar response values in asymmetric changes with other unchanged regions.

E. Ablation Study

In order to further explore the validity of the proposed modules, we remove aRP and aSP successively to set up the ablation study and check corresponding influences on model performance. As illustrated in Tab.VI, aSP leads to 0.2 improvements in mIOU and 0.6 improvements in SeK with the residual blocks. Then, aRP further improves the model performance by 0.8 in mIOU and SeK. Besides, with the Xception block, aSP improves mIOU by 0.5 and SeK by 0.8, while aRP further improves mIOU by 0.2 and SeK by 0.1. Similarly, with the Squeeze-and-Excitation block, aSP brings about 0.7 improvements in mIOU and 0.8 improvements in SeK, while aRP raises 0.5 improvements in mIOU and 0.8 improvements in SeK. In summary, aSP and aRP improve model performances on BCD and SCD simultaneously. The effectiveness of our proposed modules can be verified.

Moreover, Fig.12 shows the visual results of the ablation study, where we can see that our proposed modules can better extract illegible changed regions, such as the changes between *buildings* across multi-temporal images. In the meantime, the false positive identification of changed regions could be reduced with the proposed modules.

Further to explore the effects of asymmetric feature pairs, we remove feature pair integrations and retain the multi-scale structures in aSP to see the influence on results with residual backbones. As we can see in Tab.V, the incomplete model suffers a decline in the performance, which verifies merits of the integration of asymmetric feature pairs.

VII. Conclusion

In this paper, we propose an asymmetric siamese network for semantic change detection to alleviate categorical ambiguity caused by asymmetric changes through locally asymmetric architecture. To better train deep models, we create a large scale well-annotated SECOND as a new benchmark, which includes the changed regions between the same land-cover class. Further, to alleviate the influence of label imbalance during model training and evaluation, we design an adaptive threshold learning module and an SeK to adaptively revise the output deflections and fix unreasonable scores computed with traditional metrics respectively. The experimental results show that the proposed model stably achieves the best results with different encoder backbones compared with state-of-the-art algorithms.

REFERENCES

[1] A. Robin, L. Moisan, and S. L. Hégarat-Mascle, "An a-contrario approach for subpixel change detection in satellite imagery," *IEEE TPAMI*, vol. 32, no. 11, pp. 1977–1993, 2010.

[2] A. Lanza and L. di Stefano, "Statistical change detection by the pool adjacent violators algorithm," *IEEE TPAMI*, vol. 33, no. 9, pp. 1894–1910, 2011.

[3] R. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *CVIU*, vol. 187, 2019.

[4] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE TIP*, vol. 14, no. 3, pp. 294–307, 2005.

[5] V. Ruzicka, S. D'Aronco, J. D. Wegner, and K. Schindler, "Deep active learning in remote sensing for data efficient change detection," in *ECML/PKDD Workshop on Machine Learning for Earth Observation*, 2020.

[6] K. Pollock, "Policy urban physics," *Nature*, vol. 531, pp. S64–S66, 2016.

[7] A. Arneth, "Climate science: Uncertain future for vegetation cover," *Nature*, vol. 524, pp. 44–45, 2015.

[8] J. R. Townshend, "Global land change from 1982 to 2016," *Nature*, vol. 560, pp. 639–643, 2018.

[9] A. S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, pp. 418–422, 2016.

[10] A. Huertas and R. Nevatia, "Detecting changes in aerial views of man-made structures," in *ICCV*, 1998.

[11] V. Rengarajan, A. N. Rajagopalan, R. Aravind, and G. Seetharaman, "Image registration and change detection under rolling shutter motion blur," *IEEE TPAMI*, vol. 39, no. 10, pp. 1959–1972, 2017.

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

[13] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using A "siamese" time delay neural network," *IJPRAI*, vol. 7, no. 4, pp. 669–688, 1993.

[14] K. Sakurada and T. Okatani, "Change detection from a street image pair using CNN features and superpixel segmentation," in *BMCV*, 2015, pp. 61.1–61.12.

[15] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "Changenet: A deep learning architecture for visual change detection," in *ECCV Workshops*, 2018, pp. 129–145.

[16] A. J. Lingg, E. G. Zelnio, F. Garber, and B. D. Rigling, "A sequential framework for image change detection," *IEEE TIP*, vol. 23, no. 5, pp. 2405–2413, 2014.

[17] J. Prendes, M. Chabert, F. Pascal, A. Giros, and J. Tourneret, "A new multivariate statistical model for change detection in images acquired by homogeneous and heterogeneous sensors," *IEEE TIP*, vol. 24, no. 3, pp. 799–812, 2015.

[18] S. Liu, C. Fu, and S. Chang, "Statistical change detection with moments under time-varying illumination," *IEEE TIP*, vol. 7, no. 9, pp. 1258–1268, 1998.

[19] F. Chatelain, J. Tourneret, J. Inglada, and A. Ferrari, "Bivariate gamma distributions for image registration and change detection," *IEEE TIP*, vol. 16, no. 7, pp. 1796–1806, 2007.

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

[21] L. Bruzzone and D. Fernández-Prieto, "An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images," *IEEE TIP*, vol. 11, no. 4, pp. 452–466, 2002.

[22] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE TGRS*, vol. 45, no. 1, pp. 218–236, 2007.

[23] L. Bruzzone and D. Fernández-Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE TGRS*, vol. 38, no. 3, pp. 1171–1182, 2000.

[24] M. J. Carlotto, "Detection and analysis of change in remotely sensed imagery with application to wide area surveillance," *IEEE TIP*, vol. 6, no. 1, pp. 189–202, 1997.

[25] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Simultaneous registration and change detection in multitemporal, very high resolution remote sensing data," in *CVPR Workshops*, 2015, pp. 61–69.

[26] A. Ghosh, B. N. Subudhi, and L. Bruzzone, "Integration of gibbs markov random field and hopfield-type neural networks for unsupervised change detection in remotely sensed multitemporal images," *IEEE TIP*, vol. 22, no. 8, pp. 3087–3096, 2013.

[27] R. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *ICIP*, 2018, pp. 4063–4067.

[28] Y. Chen, X. Ouyang, and G. Agam, "MFCNET: end-to-end approach for change detection in images," in *ICIP*, 2018, pp. 4008–4012.

[29] J. Liu, M. Gong, A. K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 3, pp. 545–559, 2018.

[30] M. Kolos, A. Marin, A. Artemov, and E. Burnaev, "Procedural synthesis of remote sensing images for robust change detection with neural networks," in *ISNN*, 2019, pp. 371–387.

[31] F. Pacifici, F. D. Frate, C. Solimini, and W. J. Emery, "An innovative neural-net method to detect temporal changes in high-resolution optical satellite imagery," *IEEE TGRS*, vol. 45, no. 9, pp. 2940–2952, 2007.

[32] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.

[33] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018, pp. 833–851.

[34] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019, pp. 5693–5703.

[35] Y. Li, T. Wang, B. Kang, S. Tang, C. Wang, J. Li, and J. Feng, "Overcoming classifier imbalance for long-tail object detection with balanced group softmax," in *CVPR*, 2020, pp. 10 988–10 997.

[36] J. Lopez-Fandino, A. S. Garea, D. B. Heras, and F. Argüello, "Stacked autoencoders for multiclass change detection in hyperspectral images," in *IGARSS*, 2018, pp. 1906–1909.

[37] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised multiple-change detection in VHR optical images using deep features," in *IGARSS*, 2018, pp. 1902–1905.

[38] R. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IGARSS*, 2018, pp. 2115–2118.

[39] C. Benedek and T. Szirányi, "Change detection in optical aerial images by a multilayer conditional mixed markov model," *IEEE TGRS*, vol. 47, no. 10, pp. 3416–3430, 2009.

[40] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

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

[42] G. B. Allison and C. J. Barnes, "Estimation of evaporation from non-vegetated surfaces using natural deuterium," *Nature*, vol. 301, no. 5896, pp. 143–145, 1983.

[43] X. Y. Tong, G. S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[44] Y. Long, G. S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "Dirs: On creating benchmark datasets for remote sensing image interpretation," *CoRR*, vol. abs/2006.12485, 2020.

[45] G. S. Xia, X. Bai, J. Ding, Z. Zhu, S. J. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *CVPR*, 2018, pp. 3974–3983.

[46] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[47] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017, pp. 5987–5995.

[48] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *CVPR*, 2017, pp. 1800–1807.

[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.