IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 November 2023, accepted 14 November 2023, date of publication 16 November 2023,
date of current version 21 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3333360

RESEARCH ARTICLE

Local-Specificity and Wide-View Attention Network With Hard Sample Aware Contrastive Loss for Street Scene Change Detection

ENQIANG GUO AND XINSHA FU
School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510640, China
Corresponding author: Enqiang Guo (gmayday1997@163.com)

This work was supported by in part by the National Natural Science Foundation of China under Grant 51978283, and in part by the Changsha Natural Science Foundation under Grant kq2202212.

ABSTRACT Following the intuitive idea of detecting changes by directly measuring dissimilarities between pairs of features, change detection methods based on feature similarity learning have emerged as a crucial field. However, large variances in the scale and location of required contextual information and heavy imbalance between easy and hard samples remain challenging issues. To address the first issue, we propose the Local-Specificity and Wide-View Attention Network (LSWVANet), which features a series of attention modules named Local-Specificity and Wide-View Attention Modules (LSWVAMs). Each LSWVAM consists of two contextual feature units: the Local-Specificity Feature Pyramid unit, which extracts part-specific contexts at the fine-grained level to focus on subtle changes within local discriminative parts, and the Wide-View Feature Pyramid unit, which extracts wide-view contexts at the long-range level to highlight significant changes in large-scale regions. To tackle the second issue, we introduce a novel sample-specific loss function called Hard Sample-Aware Contrastive Loss (HSACL), which is designed to downweight easy samples from both changed and unchanged categories, thereby rapidly shifting the training focus towards the informative hard samples. We demonstrate the effectiveness of our method through experiments on three challenging datasets, VL-CMU-CD, PCD2015 and PSCD, and report the experimental results showing that our approach achieves state-of-the-art accuracy.

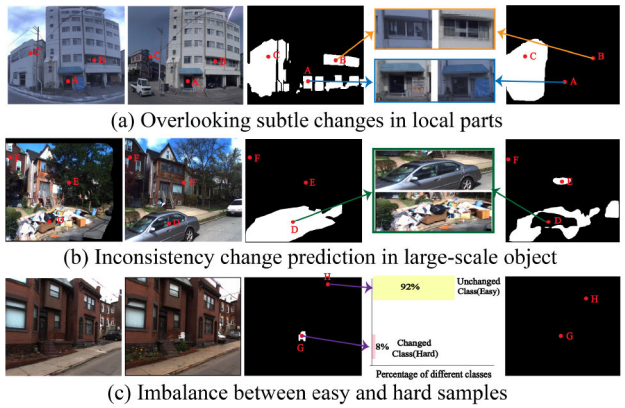
INDEX TERMS Change detection, hard sample, feature similarity learning, attention mechanism.

I. INTRODUCTION

Street-view scene change detection (SCD) is a crucial computer vision task with a wide range of applications, including urban planning [1], [2], [3], traffic surveillance [4], [5], abandoned object detection [6], [7], disaster evaluation [8], action recognition [9], [10] and self-driving [11], [12]. With the emergence of self-driving cars and robotic patrols, accurate navigation and planning based on map information have become increasingly important. Many researchers [1], [11], [12] use street-view change detection algorithms to update map information. Therefore, improving the accuracy of change detection model is a critical challenge in SCD.

With the powerful feature representation of convolutional neural networks [13], [14], [15] (CNNs), fully convolutional networks [14], [16] (FCNs) have been widely used in the field of change detection. FCN-based methods can be broadly classified into two categories. The first category is semantic segmentation-based methods [12], [16], [17], [18], which treat changed and unchanged samples as a binary classification problem. The second category is feature similarity learning-based methods (FSL) [8], [19], [20], which regard changes as feature dissimilarities and employ the Euclidean distance to measure the dissimilarity between sample pairs. In this paper, our focus is on FSL methods. To achieve more discriminative features, Zhan et al. [19] and Guo et al. [21] adopt the contrastive loss (CL) to ensure that unchanged regions yield lower distances while changed regions yield higher distances. Although these models have

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.



(a) Overlooking subtle changes in local parts

(b) Inconsistency change prediction in large-scale object

(c) Imbalance between easy and hard samples

FIGURE 1. Illustration of challenges in street-view change detection. (a) Overlooking subtle changes in local parts. (b) Inconsistency change detection in large-scale object. (c) Imbalance between easy and hard samples.

achieved high accuracy, the challenge of handling hard samples remains a crucial concern in change detection. This challenge can be delineated in two ways:

(1) There is significant variance in the scale and location of contextual information required for detecting changes. Firstly, the required contextual information of local discriminative parts needs to consider local specificity. In Fig.1(a), subtle changes such as windows (marked as A) and doors (marked as B) can be easily overlooked. Therefore, it is crucial to extract part-specific discriminative features to highlight these subtle changes. Secondly, the required contextual information of large-scale change regions also needs to consider long-range consistency. In Fig.1(b), local features attributed to the changed category vary dramatically due to the differences in textural information or lighting conditions, resulting in inconsistent predictions.

(2) During the training process, there exists a significant imbalance both within and between classes regarding easy and hard samples. Firstly, an extreme class imbalance between changed and unchanged samples (8% vs. 92%) causes neural networks to favor majority classes and ignore minority ones. Consequently, minority class samples (e.g., sample G) tend to become hard samples. Secondly, even within each class, hard samples are less abundant compared to easy ones. For example, within the changed class, hard samples like sample A are fewer than easy samples like sample C; similarly, within the unchanged class, hard samples like E are fewer than easy samples like sample F. This class imbalance hampers performance because the cumulative contribution of numerous easy samples can overwhelm the contribution of the fewer hard samples, making the training process less efficient.

To address the first issue, several strategies have been proposed in previous studies, including multiscale feature fusion [21], [22], multilevel feature fusion [23], adding auxiliary information [24], [25], [26], gate-based attention [27], [28], [29] and pairwise affinity attention [30], [31], [32]. Among these methods, gate-based attention methods that cooperate with gating functions (e.g., sigmoid function) often show considerable performance, because attention mechanisms theoretically enable the focus on the most relevant contextual information across various scale. These methods generally involve two steps: first, extracting contextual features, and then generating an attention mask based on these contextual features. In other words, gate-based attention methods heavily rely on contextual information. Specifically, Squeeze-and-Excitation (SE) variants [33], [34], Convolutional Block Attention Module (CBAM) variants [29], [35] and Split Attention variants [36], [37] employ global average pooling (GAP) to aggregate global channel or spatial contextual information. However, these methods often sacrifice local details, resulting in the loss of small-scale change regions. Approaches like U-shaped attention [38] and hourglass-based attention [39], [40] have been proposed to preserve rich feature map details and enhance the ability to detect small-scale change regions. Constrained by limited receptive field, these methods may struggle to ensure long-range consistency. Furthermore, scale-adaptive attention module [41] and local-global attention module [42] have been proposed to simultaneously capture local details and global semantic features for detecting multi-scale change regions. Despite the development of addressing scale variance, there is still significant room for improvement in detecting subtle changes within local distinctive parts. This limitation arises because local details are not fine-grained enough to distinguish subtle appearance differences. Therefore, the attention network for SCD requires further works on extracting fine-grained discriminative features and enhancing local specificity for the detection of subtle changes.

To address the aforementioned limitations, we propose Local-Specificity and Wide-View Attention Network (LSWVANet), which features a series of attention modules named (LSWVAM). Each LSWVAM consists of two contextual feature units: (1) a Local-Specificity Feature Pyramid unit (LSFP) models part-specific discriminative contexts at the fine-grained level, and (2) a Wide-View Feature Pyramid unit (WVFP) models wide-view contexts at the long-range level. The LSFP comprises four branches, each employing distinct partition methods. Within each branch, independent convolutional layers are used to extract part-specific discriminative features from distinctive spatial parts, thereby enhancing local specificity. Meanwhile, the WVFP unit consists of four branches, each integrating two successive dilation convolutional layers to capture long-range context features, thereby improving semantic consistency. Unlike other attention modules [29], [30], [33], [41], the advantage of LSWVANet is that it not only extracts long-range context features but also emphasizes local specificity within local parts, enabling the module to highlight subtle changes at the fine-grained level.

To tackle the class imbalance issue, several studies in the field of FSL have introduced reweighting loss techniques,

including the weighted contrastive loss [19], the weighted double-margin contrastive loss [30], and the combined loss [43]. These techniques aim to mitigate the imbalance issues by suppressing sample losses from the unchanged class while highlighting sample losses from the changed class. Despite the advances in addressing interclass imbalance, these methods often result in undesirable effects, such as suppressing informative hard samples from the unchanged class, while simultaneously upweighting easy samples from the changed class. This limitation arises from the fact that these class-specific loss functions struggle to effectively distinguish between easy and hard samples within each class at the fine-grained level, thus failing to upweighting hard samples from the unchanged class and downweighting easy samples from the changed class.

In our paper, we propose a sample-specific loss function called Hard Sample-Aware Contrastive Loss (HSACL). Inspired by focal loss [44], which quantifies hard samples using confidence scores, we adopt sample distance as a metric to measure hardness. Based on this definition, HSACL concentrates on optimizing samples with larger distance values from the changed class and samples with smaller distance values from the unchanged class, both of which are categorized as hard samples. In contrast to the class-specific losses [19], [30], the advantage of HSACL lies in downweighting the losses of easy samples from both the changed and unchanged classes based on their distance values, and rapidly shifting the training focus towards informative hard samples, such as subtle changes (e.g., sample A in Fig.1(a)) and background overactivation (e.g., sample E in Fig.1(b)).

We demonstrate the effectiveness of the proposed LSWVANet with HSACL through experiments on three challenging street-view datasets: the VL-CMU-CD [12], PCD2015 [8], and PSCD [24]. Our experimental results show that it achieves state-of-the-art accuracy. Furthermore, we validate the effectiveness of the proposed method through visualization of the distance distribution under polar coordinates and feature latent space distribution. Our main contributions are as follows:

(1) We propose a novel attention network called LSWVANet for street-view change detection, which features a series of attention modules named LSWVAMs. Each LSWVAM consists of two contextual feature units: (1) LSFP, which extracts local-specific contexts at the fine-grained level to focus on subtle changes within local discriminative parts, and (2) WVFP, which extracts wide-view contexts at the long-range level to highlight significant changes in large-scale regions.

(2) We propose a novel sample-specific loss function called HSACL, which is designed to identify hard samples from both changed and unchanged classes at the fine-grained level, subsequently downweighting the loss assigned to easy samples while rapidly shifting the training focus towards the hard samples.

(3) We conduct comprehensive comparative experiments on three challenging datasets: VL-CMU-CD, PCD2015 and PSCD. The results demonstrate that the proposed method achieves impressive performance and outperforms state-of-the-art methods by a considerable margin.

The rest of this work is structured as follows. The related works on change detection, attention mechanisms and reweighting loss are described in Section II. Section III describes our proposed method in detail. The experimental setup and results are presented in Section IV. The discussion and conclusions of this paper are presented in Section V and Section VI, respectively.

II. RELATED WORK

A. SCENE CHANGE DETECTION

In general, a change detection algorithm based on a CNN comprises two main components, a feature embedding and a detection head with a loss function.

1) FEATURE EMBEDDING

From the feature embedding perspective, CNN-based methods can be classified as single-stream, double-stream and hybrid structures. Single-stream structures, shown in Fig.2(a), fuse RGB image channels at $\{T_0, T_1\}$ using methods like channel concatenation [45], differential channel fusion [46], or nonlinear image fusion [47], [48]. Fig.2(b) and Fig.2(c) show that the double-stream frameworks detect changes based on high-level features. Hybrid structures are mostly building spatial-spectral and spatial-temporal information using temporal module, such as long short-term memory (LSTM) [49]. In the context of street-view change detection, double-stream structures with weight-sharing siamese networks are widely adopted. Specifically, well-established CNN models, such as DeconvNet [12], UNet [13], FCN8S [14], and Deeplab [21] have been introduced for change detection. However, these baseline methods still face challenges in dealing with significant variations in the scale and location of contextual information. To tackle this challenge, Varghese et al. [22] proposed a hierarchically dense connection to capture multi-scale features. Lei et al. [23] proposed a novel multi-level feature fusion network to hierarchically exploit channel information. Dense optical flow [25], [26] methods have also been proposed to model spatial correspondences between images at different times. Additionally, attention modules, like pairwise affinity attentions [30], [36] and gate-based attention methods [27], [38] have been proposed to locate the changed areas at various scales. Despite the remarkable performance achieved by previous methods, most methods focus on addressing scale variance, but they often overlook spatial variance, which is crucial for detecting subtle changes. Therefore, we propose a novel attention module called LSWVAM, which emphasizes local specificity within discriminative local parts. More details on related attention methods will be provided in subsection B.

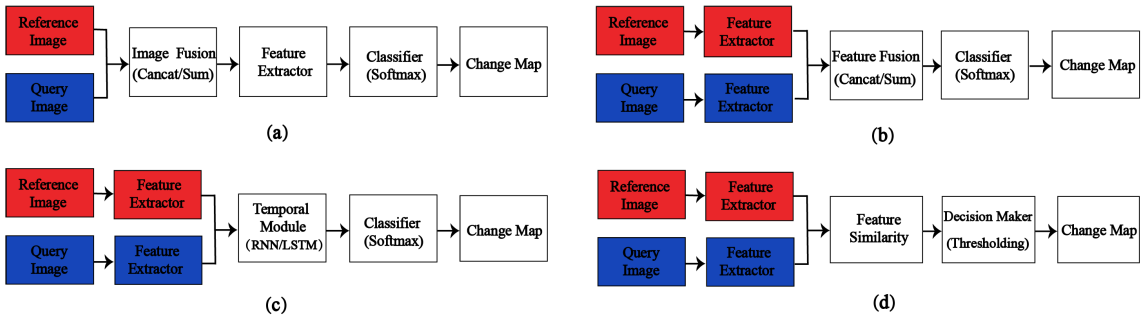


FIGURE 2. Schematic diagram of different architecture configurations based on CNNs including feature embedding structures and detection heads: (a) single-stream structure with classification head; (b) double-stream structure with classification head; (c) double-stream structure based on feature similarity learning with binary decision head; and (d) hybrid structure based on spatial-temporal feature learning with classification head.

2) DETECTION HEAD WITH LOSS FUNCTION

In general, the detection head can be divided into supervised and unsupervised algorithms. Many unsupervised methods rely on clustering techniques, including K-Nearest Neighbors [49], superpixels [50], K-means [51] and fuzzy c-means [52]. For CNN-based algorithms, the detection heads include semantic segmentation (SS) and feature similarity learning (FSL), illustrated in Fig.2(b) and Fig.2(d). SS treats changed and unchanged samples as a binary classification problem and is optimized using cross-entropy. In contrast, FSL detects changes by measuring the dissimilarity between sample pairs and is optimized by contrastive loss. However, a limitation of these loss functions is that they assign the same weights to all samples. To address this issue, numerous reweighting losses [19], [30] have been proposed to mitigate interclass imbalance. Nevertheless, these reweighting losses are class-specific. In this paper, we introduce a novel loss named HSACL, which can identify hard samples from both changed and unchanged classes. More details on related reweighting loss will be provided in subsection C.

B. ATTENTION MECHANISM ON CHANGE DETECTION

Attention mechanisms on change detection can be divided into two types: (1) pairwise affinity attention method and (2) gate-based attention method. The core idea of pairwise affinity attention method is to use dot product to build pixel-to-pixel relations. Specifically, Chen and Shi [27] utilizes the self-attention mechanism to establish long-range dependencies for capturing global-view features. Chen et al. [30] also propose a dual attention module that automatically focuses on the most relevant channel and spatial information. However, these methods are computationally intensive, particularly for high-resolution street-view images. In contrast, gate-based attention techniques offer a lightweight and computationally efficient alternative. The core idea of gate-based attention method is to serve as feature filters that highlight important features and suppress unnecessary features in different locations. Specifically, SENets [33], [34] employ GAP to leverage coarse channel wise attention relationships. Similarly, CBAM [29], [35] and Split Attention [36], [37] utilize GAP to aggregate global contextual information, emphasizing important features while suppressing unimportant ones. To preserve rich feature maps details, U-shaped attention variants [39] and hourglass-based attention [40], [41] are proposed to enhance the ability to detect small-scale change areas. Furthermore, scale-adaptive attention module [38] and local-global attention module [42] have been proposed to simultaneously capture local details and global semantic features for detecting multi-scale change regions. However, local details are not fine-grained enough to distinguish subtle changes. In this paper, we utilize part-specific feature strategy to extract fine-grained discriminative features and enhance local specificity for the detection of subtle changes.

C. REWEIGHTING LOSS ON CHANGE DETECTION

Considering that not all samples contribute equally to training a model, reweighting loss is a common practice in hard sample mining methods [44], [53], [54], [55]. The fundamental concept behind reweighting loss is to decrease the loss assigned to easy samples and increase the loss assigned to hard samples. In the field of change detection, Lei et al. [23], and Song and Jiang [56] propose the weighted cross-entropy loss function to address interclass imbalance issues. Similarly, Chen et al. [30] also proposes a weighted double-margin contrastive loss to mitigate the effects of unchanged regions by setting weight coefficients. Li et al. [43] combines weighted binary cross-entropy loss and dice coefficient loss to address the imbalance of positive and negative samples. These class-specific reweighting losses can improve the interclass imbalance between changed and unchanged samples but still face challenges in addressing the imbalance between easy and hard samples within each class. In this paper, we introduce a novel sample-specific loss called HSACL, which identifies hard samples from both changed and unchanged classes and assigns different weights to samples based on their distance values. The core idea of HSACL is inspired by focal loss [44]. However, a major difference exits: focal loss is a scaled cross entropy loss that

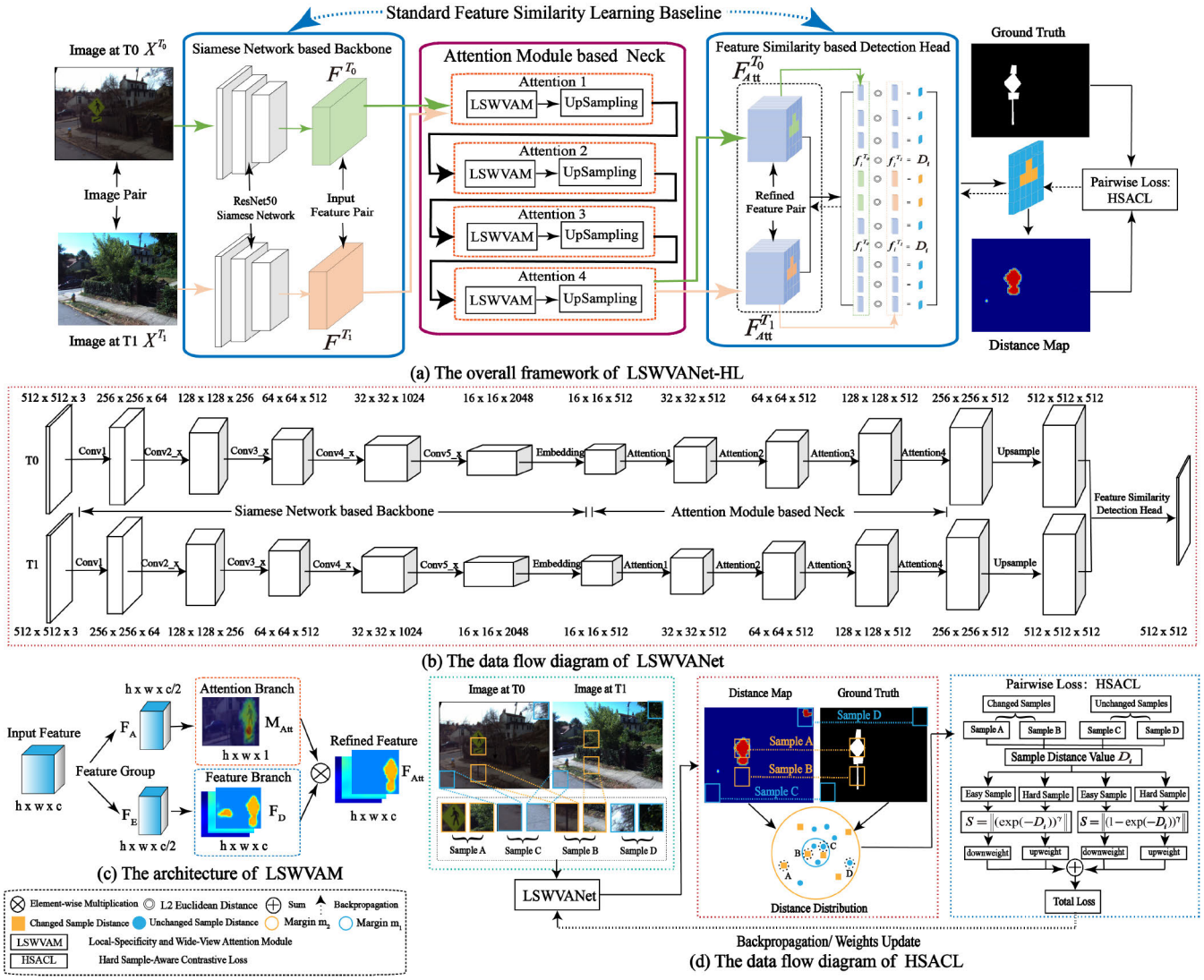


FIGURE 3. (a) The overall architecture of LSWVANet-HL, which combines LSWVANet with the optimization of HSACL. LSWVANet consists of two main components: a SFSL baseline and an attention module-based neck. The SFSL component, enclosed in the blue box, comprises two parts: (1) a siamese network-based backbone and (2) a feature similarity-based detection head. The attention module-based neck, enclosed in the red box, consists of four stacked LSWVAMs.(b) The data flow diagram of the proposed LSWVANet. (c) The architecture of LSWVAM. (d) The data flow diagram of HSACL.

employs prediction confidence as a measure of hardness, while HSACL is a scaled contrastive loss that utilizes distance values to measure hardness.

III. METHODOLOGY

In this section, we introduce our change detection framework called LSWVANet-HL, which combines the deep attention architecture of LSWVANet with the optimization of HSACL. As illustrated in Fig.3(a), LSWVANet is built upon a Standard Feature Similarity Learning (SFSL) baseline, whose components are shown in the blue box. SFSL aims to detect change regions by treating changes as feature dissimilarities and subsequently employs the Euclidean distance to measure dissimilarity between pairs of images captured at different times. To adapt to location and scale variations, LSWVANet incorporates a novel attention module-based neck enclosed in the red box, which consists of four sequentially stacked LSWVAMs. The schematic representation of LSWVANet's data flow is demonstrated in Fig.3(b). Delving into specifics, LSWVAM aims to generate a spatial attention mask that exhibits both local specificity and long-range consistency, enabling the model to highlight features within local discriminative parts and large-scale changed regions. This module is a lightweight component and can be easily integrated into the SFSL baseline. Furthermore, LSWVANet is optimized by HSACL. Fig.3(d) illustrates the data flow of HSACL, where it identifies hard samples from both the changed and unchanged classes based on the sample distance value. The HSACL dynamically adjusts the loss weights, downweighting easy samples and upweighting hard examples from both classes. Subsequently, we provide detailed information on the SFSL baseline in subsection A, describe the novelty of LSWVAM

in subsection B, and detail the design of our proposed HSACL in subsection C.

A. STANDARD FEATURE SIMILARITY LEARNING BASELINE

As depicted in Fig.3, SFSL consists of two parts: (1) a siamese network-based backbone and (2) a feature similarity-based detection head. Specifically, the image pairs $(X^{T_0}, X^{T_1}) \in \mathbb{R}^{C\times H\times W}$ pass through a Siamese network-based backbone to yield a feature pair $(F^{T_0}, F^{T_1}) \in \mathbb{R}^{c\times h\times w}$. Subsequently, for a feature sample $(f_i^{T_0}, f_i^{T_1}) \in \mathbb{R}^c$ at the i^{th} position within the $h \times w$ feature map, the feature similarity-based detection head evaluates the dissimilarity of the feature pair using the Euclidean distance D_i. Finally, change regions are identified by selecting the pixels whose distance values exceed the predefined threshold.

To enhance the distance value for changed samples and reduce it for unchanged samples, it is common to utilize contrastive loss (CL) in optimizing the SFSL-based change detection model. For clarity, during the training phase, we formulate a pair of feature samples s_i as $\{(f_i^{T_0}, f_i^{T_1}), y_i, D_i\}$, where y_i denotes the corresponding label of s_i, and D_i denotes the Euclidean distance between L2-normalized feature vectors $f_i^{T_0}$ and $f_i^{T_1}$. We define s_i as a changed sample or positive sample when $y_i = 1$, and as an unchanged sample or negative sample when $y_i = 0$. CL is formulated in Equation 1 as follows:

$$CL = \begin{cases} (D_i - m_1)^2 & y_i = 0 \\ (max(0, m_2 - D_i)))^2 & y_i = 1 \end{cases} \quad (1)$$

The parameters m_1 and m_2 represent the margins for positive and negative samples, respectively. In our case, we set m_1 to 0 and m_2 to 2.

Despite achieving impressive results in change detection, SFSL still faces challenges related to hard samples due to significant variance in the scale and location of contextual information and the heavy imbalance between easy and hard samples. We aim to address these challenges from two perspectives: (1) We introduce the LSWVAM as a neck component that considers not only the broader contextual information but also emphasizes part-specific contexts within local parts and (2) We utilize the HSACL to shift the training focus towards hard samples.

B. LOCAL-SPECIFICITY AND WIDE-VIEW ATTENTION MODULE

1) OVERVIEW

The structure of LSWVAM is illustrated in Fig.3(c). Each LSWVAM is composed of two branches: a feature branch and an attention branch. In the feature branch, feature encoding is performed using two 3×3 dilated convolutions, which project the input feature $F_E \in \mathbb{R}^{c/2\times h\times w}$ into decoded features $F_D \in \mathbb{R}^{c\times h\times w}$. The attention branch, on the other hand, is responsible for projecting the input features $F_A \in \mathbb{R}^{c/2\times h\times w}$ into an attention mask $M_{Att} \in \mathbb{R}^{h\times w}$, where the response of this mask reflects the most relevant contextual information. Our attention mechanism empowers the module to effectively highlight subtle changes in local parts, as well as significant changes in larger-scale regions. Finally, the aggregated features $F_{Att} \in \mathbb{R}^{c\times h\times w}$ are represented by the weighted combination of the attention map $M_{Att} \in \mathbb{R}^{h\times w}$ and the decoded feature map $F_D \in \mathbb{R}^{c\times h\times w}$. The refined feature map F_{Att} is computed as follows:

$$F_{Att} = F_D \otimes M_{Att} \quad (2)$$

where $\otimes$ denotes elementwise multiplication.

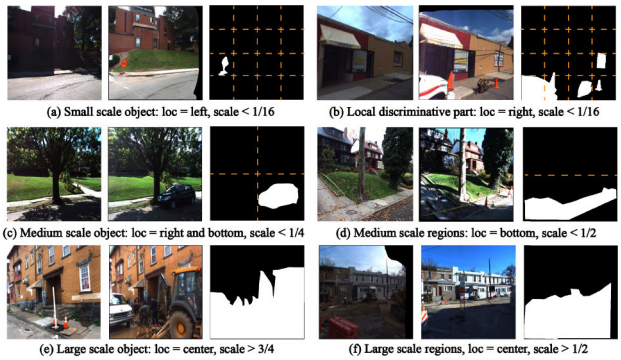


FIGURE 4. Change regions vary at spatial scales and locations.

2) ATTENTION BRANCH

The proposed attention branch generates a spatial attention map, denoted as $M_{Att} \in \mathbb{R}^{h\times w}$, to highlight or suppress features in different locations. Therefore, it is crucial to determine which contextual features should be focused on. As observed in Fig.4, contextual information related to change regions exhibits two characteristics: local specificity and long-range consistency. Firstly, contextual features must account for local specificity within discriminative local parts. For instance, small-scale objects like the road signals in Fig.4(a) are situated on the left side and occupy approximately 1/16 of the image, while subtle changes, as shown in local parts like the windows in Fig.4(b), are located on the right and also occupy about 1/16 of the image. Secondly, contextual features should also consider long-range consistency within large-scale objects. For example, the change regions depicted in Fig.4(d), Fig.4(e), and Fig.4(f) have large spatial scales and occupy almost the entire image. Based on these observations, we introduce a novel Local-Specificity Feature Pyramid (LSFP) to extract part-specific contextual information, emphasizing the local discriminative change parts and a Wide-View Feature Pyramid (WVFP) to capture long-range contextual information for large-scale change regions. Fig.5 illustrates the processes of LSFP and WVFP.

Local-Specificity Feature Pyramid: To achieve a balance between hierarchical feature embedding and computational efficiency, we adopt the multi-groups structure through channel-wise splitting operations proposed in ShuffleNet [57]. Within each group, we employ different part-specific feature strategies to capture diverse local-specificity context features. As depicted in Fig.5, given

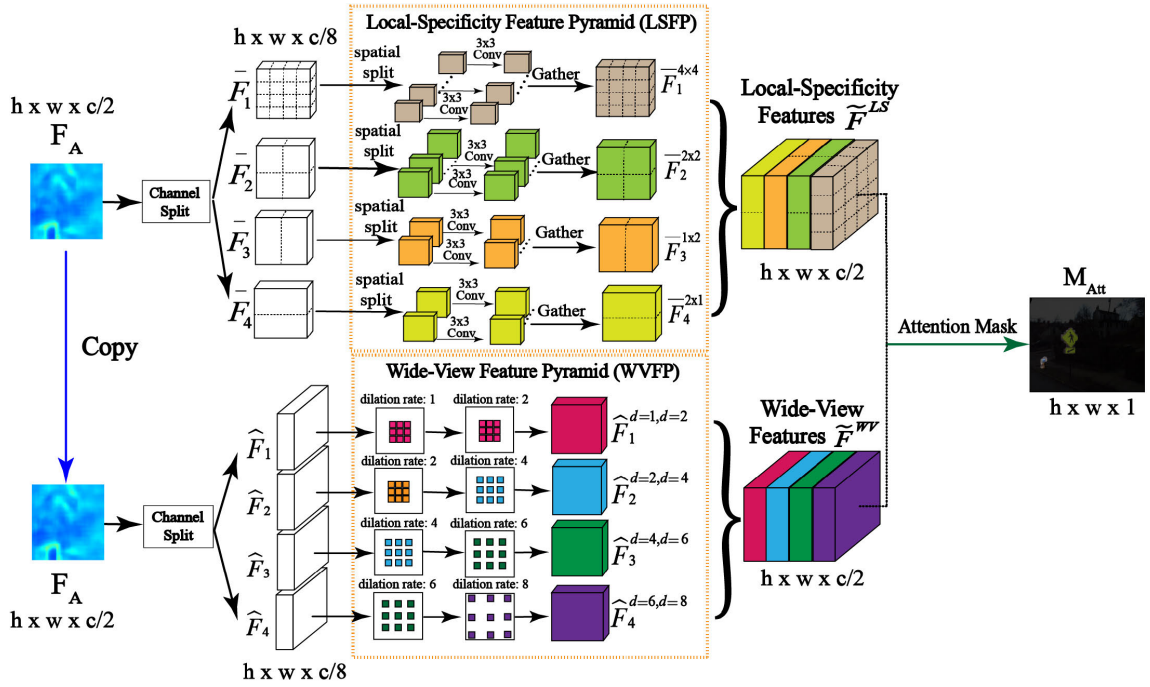


FIGURE 5. Illustration of the proposed attention branch, which utilizes a novel local-specificity feature pyramid and wide-view feature pyramid to extract part-specific and long-range context features.

the input feature map $F_A \in \mathbb{R}^{c\times h\times w}$, we first perform a channel split operation to obtain four groups of features $\{\bar{F}_1; \bar{F}_2; \bar{F}_3; \bar{F}_4\} \in \mathbb{R}^{(c/4)\times h\times w}$. For each feature group, we divide the feature map into several non-overlapping patches. Considering that change regions have different spatial scales and locations, we split four groups of feature maps into different feature parts at various spatial scales as follows. (1) The first feature map $\bar{F}_1 \in \mathbb{R}^{(c/4)\times h\times w}$ is divided into 4×4 parts, where each part $\bar{f}_1^{4\times4}$ has a spatial size of $(c/4) \times (h/4) \times (w/4)$; (2) The second feature map $\bar{F}_2 \in \mathbb{R}^{(c/4)\times h\times w}$ is divided into 2×2 parts, where each part $\bar{f}_2^{2\times2}$ has a spatial size of $(c/4) \times (h/2) \times (w/2)$; (3) The third feature map $\bar{F}_3 \in \mathbb{R}^{(c/4)\times h\times w}$ is divided into 1×2 parts along the width axis, where each part $\bar{f}_3^{1\times2}$ has a spatial size of $(c/4) \times h \times (w/2)$; and (4) The fourth feature map $\bar{F}_4 \in \mathbb{R}^{(c/4)\times h\times w}$ is divided into 2×1 parts along the height axis, where each part $\bar{f}_4^{2\times1}$ has a spatial size of $(c/4) \times (h/2) \times w$. After obtaining these feature parts, for each feature group, we apply independent convolutional layers with a 3×3 kernel size for each part to extract local-specific context features. These convolutional layers capture distinctive features from different spatial parts, facilitating the extraction of context features characterized by local specificity. Finally, we aggregate all feature parts across height and width axes and concatenate the four groups of features to obtain local-specificity context features $\widetilde{F}^{LS}$ along the channel axis. This process can be computed as follows:

$$\overline{F}_i^{N_h^i\times N_w^i} = Gather(g_{3\times3}(\overline{f}_i^{N_h^i\times N_w^i})) \quad \text{for} \quad i = 1, 2, 3, 4 \tag{3}$$

$$\widetilde{F}^{LS} = Concat\left\{\overline{F}_i^{N_h^i\times N_w^i}; \quad \text{for} \quad i = 1, 2, 3, 4\right\} \tag{4}$$

Here, $\overline{f}_i^{N_h^i\times N_w^i}$ represents the split of the i^{th} feature group into feature parts, with $N_h^i \times N_w^i$ denoting the number of feature parts. In our paper, the default settings for the number of feature parts $\{(N_h^i, N_w^i)|\text{for } i = 1, 2, 3, 4\}$ corresponding to the i^{th} feature group are determined according to the order of the set $\{(4, 4), (2, 2), (1, 2), (2, 1)\}$. The term $g_{3\times3}$ indicates convolution with a 3×3 kernel size. 'Gather' represents the operation that aggregates feature parts across height and width axes, while 'Concat' represents the operation that aggregates features across channel axes.

Wide-View Feature Pyramid: Similar to LSFP, we utilize the multi-groups structure in WVFP through channel-wise splitting operations to capture diverse long-range context features. As depicted in Fig.5, given the input feature map $F_{Att} \in \mathbb{R}^{C\times H\times W}$, we perform a channel split operation to obtain four groups of features $\left\{\hat{F}_1; \hat{F}_2; \hat{F}_3; \hat{F}_4\right\} \in \mathbb{R}^{c/4\times h\times w}$. Then, for each group feature, we apply two 3×3 dilated convolutions with different dilation rates to incorporate long-range context features. Finally, we concatenate all four groups of features to obtain wide-view context features $\widetilde{F}^{WV}$ along the channel dimension. The process described above can be computed as follows:

$$\hat{F}_i^{d_1^i=n,d_2^i=m} = g_{3\times3}^{d_2^i=m}(g_{3\times3}^{d_1^i=n}(\hat{F}_i)) \text{ for } i = 1, 2, 3, 4 \tag{5}$$

$$\widetilde{F}^{WV} = Concat\left\{\hat{F}_i^{d_1^i=n,d_2^i=m}; \quad \text{for} \quad i = 1, 2, 3, 4\right\} \tag{6}$$

TABLE 1. The detailed architecture of LSWVAM, k: kernel size, o: output channel, d: dilate rate, LSFP: Local-specificity feature pyramid, WVFP: Wide-view feature pyramid.

Branch Name		Parameters
Feature Branch		$\begin{bmatrix} k=3\times3, o=256, d=2 \\ k=3\times3, o=512, d=2 \end{bmatrix} \times 1$
Attention branch	LSFP	$\begin{bmatrix} k=3\times3, o=64 \end{bmatrix} \times 16$ $\begin{bmatrix} k=3\times3, o=64 \end{bmatrix} \times 4$ $\begin{bmatrix} k=3\times3, o=64 \end{bmatrix} \times 2$ $\begin{bmatrix} k=3\times3, o=64 \end{bmatrix} \times 2$
	WVFP	$\begin{bmatrix} k=3\times3, o=64, d=1 \\ k=3\times3, o=64, d=2 \end{bmatrix} \times 1$ $\begin{bmatrix} k=3\times3, o=64, d=2 \\ k=3\times3, o=64, d=4 \end{bmatrix} \times 1$ $\begin{bmatrix} k=3\times3, o=64, d=4 \\ k=3\times3, o=64, d=6 \end{bmatrix} \times 1$ $\begin{bmatrix} k=3\times3, o=64, d=6 \\ k=3\times3, o=64, d=8 \end{bmatrix} \times 1$
Attention Mask		$\begin{bmatrix} k=3\times3, o=1024 \\ k=1\times1, o=1 \end{bmatrix} \times 1$

Here, $g_{3\times3}^{d_1^i=n}$ denotes the first dilated convolution for the i^{th} feature group with a 3×3 kernel size and a dilation rate of n. In our paper, the default settings for the two successive dilation rates $\{(d_1^i, d_2^i)|\text{for } i = 1, 2, 3, 4\}$ corresponding to the i^{th} feature group are determined following the order of the set $\{(1, 2), (2, 4), (4, 6), (6, 8)\}$.

Spatial Attention Mask Generation: After obtaining the local-specificity context features $\widetilde{F}^{LS}$ and wide-view context features $\widetilde{F}^{WV}$, we first concatenate them along the channel dimension and apply a 1×1 convolution, batch normalization and sigmoid function to produce the spatial attention mask $M_{Att} \in \mathbb{R}^{h\times w}$.

$$M_{Att} = \sigma\left(BN\left(Conv_{1\times1}\left(Concat(\widetilde{F}^{LS}; \widetilde{F}^{WV})\right)\right)\right) \quad (7)$$

where σ denotes the sigmoid function, BN denotes batch normalization, concat denotes feature channel concatenation, and $Conv_{1\times1}$ denotes the convolutional filter with the 1×1 kernel. The detailed architecture of LSWVAM is shown in Table 1.

C. HARD SAMPLE-AWARE CONTRASTIVE LOSS

1) CL ANALYSIS

As mentioned above, the most common loss function for SFSL-based change detection methods is CL. To provide an intuitive understanding of how CL works (e.g., the impact of margins m_1 and m_2), we transform the Euclidean distance map to distance distribution under polar coordinates. Fig.6 shows an example of the distance distribution under polar coordinates. Specifically, the green dot represents the positive sample ($y_i = 1$), the blue star represents the negative sample ($y_i = 0$), the red circles represent the margins $m_1 = 0.0$ and $m_2 = 2.0$, and the Euclidean distance D_i represents the radial coordinate. For a more detailed explanation of the distance distribution under polar coordinates, we select some samples: sample $A(y_A, D_A)$, sample $B(y_B, D_B)$, sample $C(y_C, D_C)$ and sample $D(y_D, D_D)$. From the visualization of those selected samples, we can obtain three properties as follows:

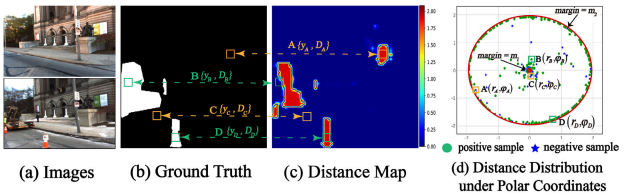


FIGURE 6. Conversion of the Euclidean distance map to a distance distribution under polar coordinates. (a) Input images, (b) ground truth, (c) Euclidean distance map, and (d) distance distribution under polar coordinates.

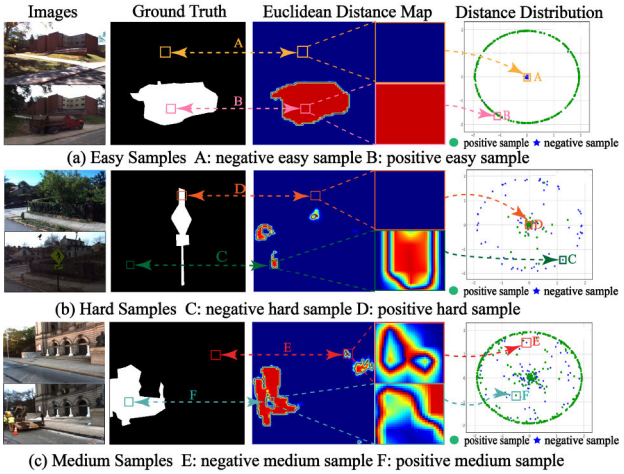


FIGURE 7. (a) Sample A is a negative easy sample and sample B is a positive easy sample, (b) Sample C is a negative hard sample and sample D is an positive hard sample, (c) Sample E is a negative medium sample and sample F is an positive medium sample.

(1) All the sample distances are in the range of $[m_1, m_2]$. In other words, margin m_2 defines the upper bound of the distance distribution, and margin m_1 defines the lower bound.

(2) After optimization by CL, the majority of positive sample distances (e.g., sample D) are constrained to the upper bound, and most of the negative sample distances (e.g., sample C) converge to the lower bound.

(3) Some positive samples exist with small distance values (e.g., sample B) are constrained to the lower bound; meanwhile, some negative samples with large distance values (e.g., sample A) converge to the upper bound.

In summary, although most samples have been well-optimized, there are always a small number of samples that are difficult to train. However, CL treats each sample equally and assigns the same weight to all samples. To address this limitation, we need to identify hard samples and adaptively assign different weights to samples based on their difficulty levels.

2) HARD SAMPLE DEFINITION

In this subsection, we focus on defining of the hard sample levels. As depicted in Fig. 7, we can observe three aspects regarding the sample distance distribution: (1) Positive

samples with large distance values (e.g., sample B) or negative samples with small distance values (e.g., sample A) tend to be easy samples; (2) Positive samples with small distance values (e.g., sample D) or negative samples with large distance values (e.g., sample C) are more likely to be hard samples; and (3) There are also positive medium samples (e.g., sample F) or negative medium samples (e.g., sample E) with distance values falling within the range between the lower and upper bounds. In that case, considering that samples have different difficulty levels, we partition all positive/negative samples into three levels using the sample distance value as an evaluation criterion. The definitions are as follows:

(1) We partition all positive samples into three levels, namely, 'positive easy sample', 'positive medium sample', and 'positive hard sample'. The positive hard sample set contains samples in which the distance D_i is in the interval of $[0, \tau_1]$, while the positive easy sample set contains samples in which the distance D_i is in the interval of $[\tau_2, 2.0]$ (the default margin value of contrastive loss is 2.0). The positive medium sample set contains samples in which the distance D_i is in the interval of $[\tau_1, \tau_2]$. In our work, we set τ_1 as 0.3 and τ_2 as 1.7.

(2) Similarly, we partition all negative samples into three sets, namely, 'negative easy sample', 'negative medium sample', and 'negative hard sample'. The negative easy sample set contains samples in which the distance D_i is in the interval of $[0, \tau_4]$, while the negative hard sample set contains samples in which the distance D_i is in the interval of $[\tau_3, 2.0]$. The negative medium sample set contains samples in which the distance D_i is in the interval of $[\tau_4, \tau_3]$. In our work, we set τ_4 as 0.3 and τ_3 as 1.7.

3) HSACL DEFINITION

Based on the hard sample definition, we propose HSACL to address the shortcomings of the CL. We design a modulating factor S^{pos} for positive samples as $\|(exp(-D_i))^{\gamma}\|$, while set the scaling factor S^{neg} for negative samples as $\|(1 - exp(-D_i))^{\gamma}\|$. Among them, γ is a hyperparameter that control the rate at which easy examples are down-weighted. More details are shown in Equation 8 and Equation 9:

$$HSACL = \begin{cases} S^{neg}(D_i - m_1)^2 & y_i = 0 \\ S^{pos}(max(0, m_2 - D_i)^2 & y_i = 1 \end{cases} \tag{8}$$

$$\begin{cases} S^{neg} = \|(1 - exp(-D_i))^{\gamma}\| \\ S^{pos} = \|(exp(-D_i))^{\gamma}\| \end{cases} \tag{9}$$

Similar to Equation 1, D_i denotes the Euclidean distance between the two L2-normalized feature vector of $(f_i^{T_0}, f_i^{T_1})$. m_1 and m_2 are the margins for positive and negative samples, respectively. In our work, we set m_1 to 0 and m_2 to 2. Fig.8(c)(d) describes the positive and negative sample weight distributions, and we analyse the distributions as follows:

(1) Assume that $s_i = \{(f_i^{T_0}, f_i^{T_1}), y_i, D_i\}$ is a positive sample. As shown in Fig.8(c), when D_i is close to 0, the sample s_i is a positive hard example, and the corresponding modulating factor $\|(exp(-D_i)^{\gamma}\|$ is close to 1, meaning that the loss is unaffected; when D_i is close to 2.0, the sample s_i is a positive easy sample, and the scaling factor is approximately equal to 0, which greatly reduces the loss contribution from easy examples.

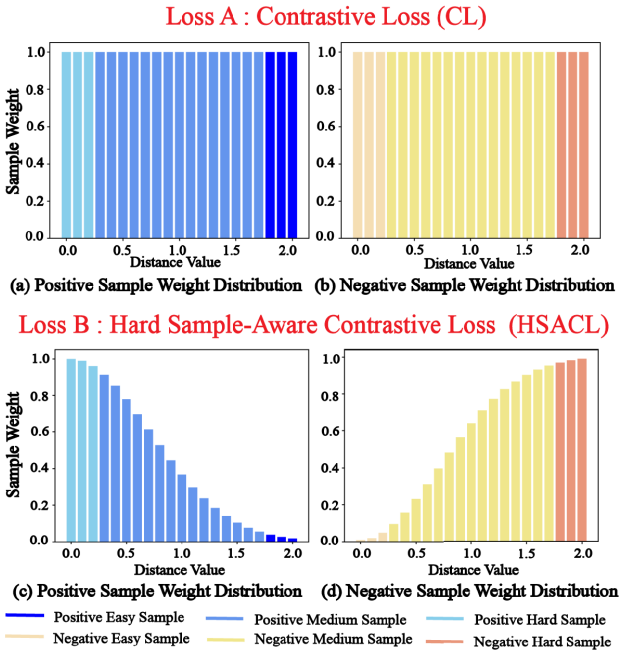


FIGURE 8. (a) Positive sample weight distribution of contrastive loss, (b) negative sample weight distribution of contrastive loss, (c) positive sample weight distribution of hard sample-aware contrastive loss, and (d) negative sample weight distribution of hard sample-aware contrastive loss.

(2) Similarly, assume that $s_i = \{(f_i^{T_0}, f_i^{T_1}), y_i, D_i\}$ is a negative sample. As shown in Fig.8(d), when D_i is close to 0, the sample s_i is a negative easy sample and the scaling factor $\|(1-exp(-D_i))^{\gamma}\|$ goes to 0; when D_i is close to 1, the sample is a negative hard sample, and the scaling factor is close to 1. It means that scaling factor can automatically suppress the negative easy examples and rapidly focus on hard examples.

(3) As mentioned above, the hyperparameter γ controls the rate at which samples are suppressed. When $\gamma = 1$, the weight assigned to easy sample is approximately 0.05, at which point the easy sample is not sufficiently suppressed. As γ increases, the corresponding factor heavily reduces the loss contribution from easy examples. When $\gamma = 5$, not only the loss weight of the easy sample goes to zero, but also the loss weight of medium example in the range [1.0,1.7] is overly reduced, which may affect the optimization of the medium example. Similarly, the loss weight of the negative easy sample decreases sharply as γ increases, while the loss weight assigned to negative hard sample equals 1, which means that optimization pay more attention to the hard samples. Notably, small γ values are not sufficiently suppressive for easy samples, while large γ values may impede the optimization of the medium samples. How to choose an appropriate γ will be discussed in next section.

IV. EXPERIMENT

In this section, we describe our experimental evaluation and conduct an ablation study of our proposed architecture. We apply our method to the task of street-view scene change detection and demonstrate competitive performance compared to the baseline on the VL-CMU-CD [12], PCD2015 [8], and PSCD [24] datasets.

A. IMPLEMENTATION DETAILS

In our experiment, we fine-tune the proposed method based on ResNet50 by removing its last classification layer. We initialize the learning rate as 1e-6 and train all models using stochastic gradient descent with a momentum of 0.95 and weight decay of 5e-5. During the training process, we conduct a series of ablation experiments on the validation set of the VL-CMU-CD dataset to determine the optimal model designs, such as patch sizes or dilation rates in LSWVAM, as well as the optimal hyperparameters, including the γ value in HSACL and the learning rate. The experiments are conducted on the PyTorch platform [58], and an Nvidia GTX TITAN X is used as the training hardware.

To improve performance during the training process, we employ three commonly used strategies from previous works [59], [60]: (1) Data augmentation techniques [61] enhance data diversity by introducing visual variability, including scale and color variations, addressing dataset imbalances [62], and preventing overfitting. In our work, we employ image transformations such as cropping, horizontal flipping, and the addition of random Gaussian noise. Specifically, for both the PCD 2015 and PSCD datasets, we use a sliding window of 28 pixels in width during cropping, generating 29 patches, each with a resolution of 224 $\times$ 224. Additionally, each cropped training sample is resized to 384 $\times$ 384 and flipped horizontally and vertically. For the VL-CMU-CD dataset, all training samples are resized to 512 $\times$ 512 and flipped horizontally and vertically. (2) Early stopping: To prevent overfitting and speed up computation, we implement early stopping to halt the training process if the model's F1-score metric stops improving on the validation dataset over a predefined number of consecutive epochs, known as patience. Specifically, we set a patience value of 10 for the VL-CMU-CD dataset and 7 for both the PCD2015 and PSCD datasets. (3) Learning rate decay: To speed up model convergence and improve the accuracy and stability of the trained model, we utilize the poly learning rate policy to gradually decrease the learning rate during the training process. Specifically, the learning rate is multiplied by $(1 - \frac{iter}{total\_iters})^{0.95}$, where $iter$ denotes the current number of iterations and $total\_iters$ denotes the total number of iterations. The total number of iterations depends on the number of training samples and the predefined number of training epochs. Specifically, we set the training epochs to 120 for the VL-CMU-CD dataset and 60 epochs for both the PCD2015 and PSCD datasets.

B. EVALUATION METRICS

To evaluate the performance of our method, we employ three evaluation metrics [63]: precision (P), recall (R), and F1-score ($F1$). These metrics are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$R = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{12}$$

where TP denotes the number of true positives, FP denotes the number of false positives and FN denotes the number of false negatives. P denotes the ratio of truly changed regions detected among all detected regions. R denotes the ratio of truly changed regions detected compared to the ground truth. The $F1$ is the harmonic mean of precision and recall, ranging from 0 to 1. From the definitions, we can observe that precision favours methods with a low false detection rate (low FP), while recall favours methods with a low missed detection rate (low FN). As a result, the $F1$ provides a balanced metric that demands lower values of both FP and FN and a higher TP value. A larger $F1$ indicates better performance, making it a more reliable metric for evaluating change detection.

C. DATASET

1) VL-CMU-CD DATASET

The VL-CMU-CD dataset [12] is a change detection dataset with challenging changes, including structural changes (e.g., building demolition and traffic signs) and noisy changes (e.g., lighting condition/weather/season changes, viewpoint changes, and dynamic changes). The dataset contains 152 sequences with 1362 image pairs. According to the data splits provided in [12], the training data consist of 97 sequences with 933 image pairs, and the testing set consists of 54 sequences with 429 image pairs.

2) PCD2015 DATASET

The PCD2015 dataset [8] consists of two subsets: Tsunami and GSV. Specifically, the Tsunami dataset describes the street scene changes after a tsunami disaster, including 200 image pairs, and the GSV dataset describes the street-view changes from Google Maps, including a total of 92 image pairs. To validate the model performance, we perform fivefold cross-validation as mentioned in [8].

3) PSCD DATASET

The PSCD dataset [24] is a panoramic semantic change detection dataset that contains a range of challenging factors, including dynamic illumination conditions and camera viewpoint differences. The PSCD dataset comprises 770 image pairs. To validate the model performance, we perform fivefold cross-validation as mentioned in [24].

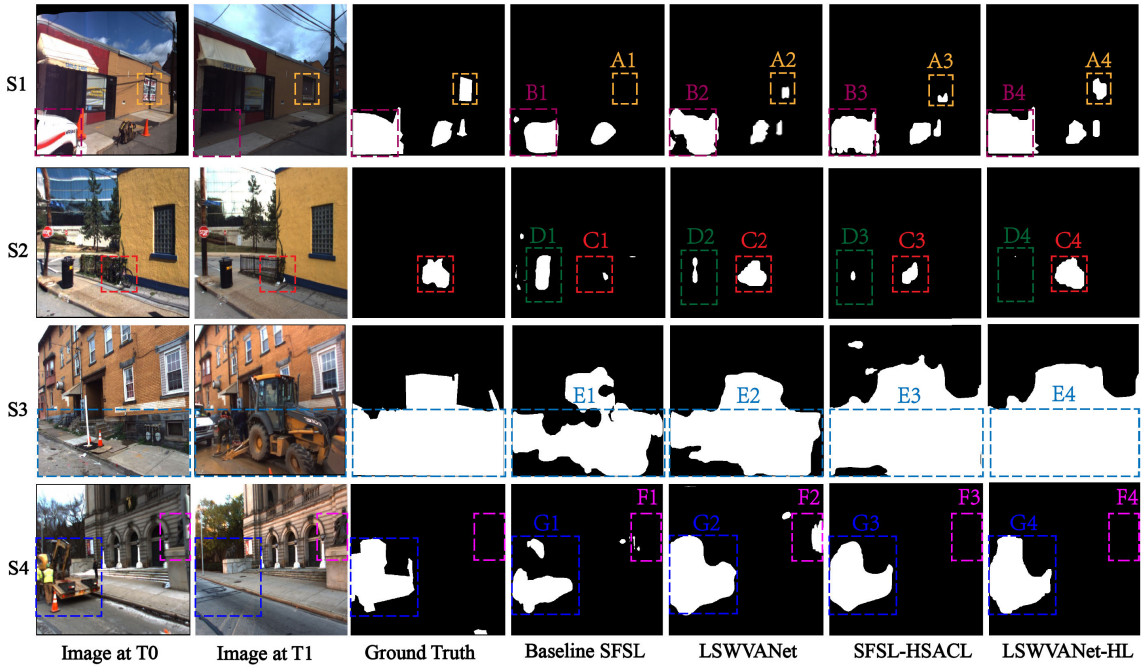


FIGURE 9. Visual quality comparison of different baseline methods. From left to right in each row: image at T0, image at time T1, ground truth, result generated by Baseline SFSL, result generated by LSWVANet, result generated by SFSL-HASCL, and result generated by LSWVANet-HL.

D. ABLATION STUDIES

1) BASELINE COMPARISON

We design comparison experiments on the validation set of the VL-CMU-CD dataset by progressively adding modules, including the baseline SFSL (optimized by CL), LSWVANet (integrated with LSWVAM and optimized by CL), SFSL-HSACL (optimized by HSACL), and LSWVANet-HL (integrated with LSWVAM and optimized by HSACL). Table 2 shows the experimental results. The baseline method SFSL achieves an $F1$ of 69.4% on the validation set of the VL-CMU-CD dataset. Compared to the baseline SFSL, LSWVANet and SFSL-HSACL yields improvements of 2.8% and 5.4%, respectively. Our LSWVANet-HL markedly outperforms the baseline methods and achieves an 8.8% improvement.

To validate the impact of LSWVAM, we conduct ablation studies comparing methods with and without LSWVAM under the same loss function. The comparison included baseline SFSL and LSWVANet, as well as SFSL-HSACL and LSWVANet-HL. The visual results in Fig. 9 confirm the effectiveness of LSWVAM in two key aspects: (1) Highlighting subtle changes in local parts. As depicted in Fig. 9, the baseline SFSL fails to detect subtle changes, such as the window areas within dashed box A1 in S1 and the bicycle within dashed box C1 in S2. In contrast to the spatial-invariant method (SFSL), LSFP in LSWVAM employs a part-specific learning strategy to model spatial variance, enabling the distinction of subtle visual differences among various local parts, such as the bicycle and fence in S2. (2) Enhancing consistency of large-scale objects. Fig. 9 illustrates that the baseline SFSL fails to ensure consistent detection results, such as the truck within dashed box E1 in S3 and the vehicle within dashed box G1 in S4. In contrast to single-scale context embedding (SFSL), WVFP in LSWVAM utilizes multi-groups dilated convolutions to model multi-scale contexts. This strategy establishes long-range relationships between different regions, improving prediction consistency, such as the truck within dashed box E2 in S3. A similar conclusion can also be drawn from the comparison between SFSL-HSACL and LSWVANet-HL. Compared with SFSL-HSACL, LSWVANet-HL, incorporating LSWVAM, exhibits a better ability to localize changes in local parts and large-scale objects.

TABLE 2. Performance comparison between different baseline methods on the VL-CMU-CD validation set. R: ResNet50, C: Contrastive Loss, A: Attention Module, H: Hard sample aware contrastive loss. The best results are in bold.

Methods	Adding Modules				P (%)	R (%)	F1 (%)
	R	C	A	H			
Baseline SFSL	√	√			68.3	70.5	69.4
LSWVANet	√	√	√		72.7	71.8	72.2
SFSL-HSACL	√			√	76.2	73.5	74.8
LSWVANet-HL	√		√	√	**79.1**	**77.3**	**78.2**

To validate the impact of HSACL, we conduct ablation studies comparing CL-based methods with HSACL-based methods using the same backbone architecture. As depicted in Fig. 9, it is evident that HSACL-based methods outperform CL-based methods. For instance, the result of baseline SFSL shows significant noise detection within dashed box D1 and

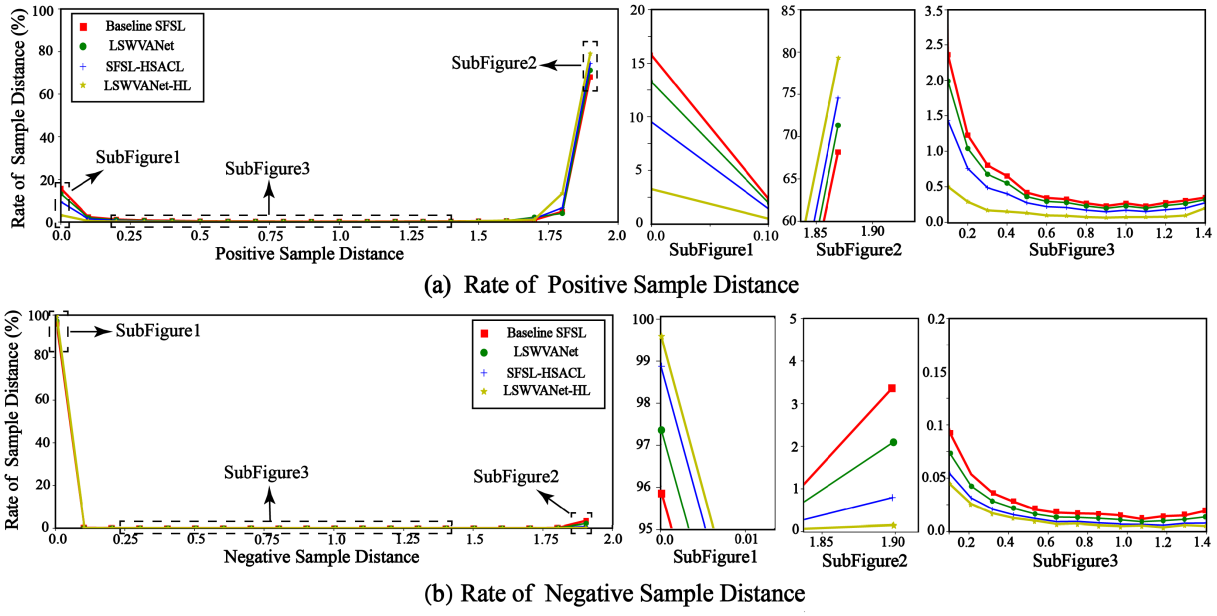


FIGURE 10. (a) Rate of positive sample distance, (b) Rate of negative sample distance.

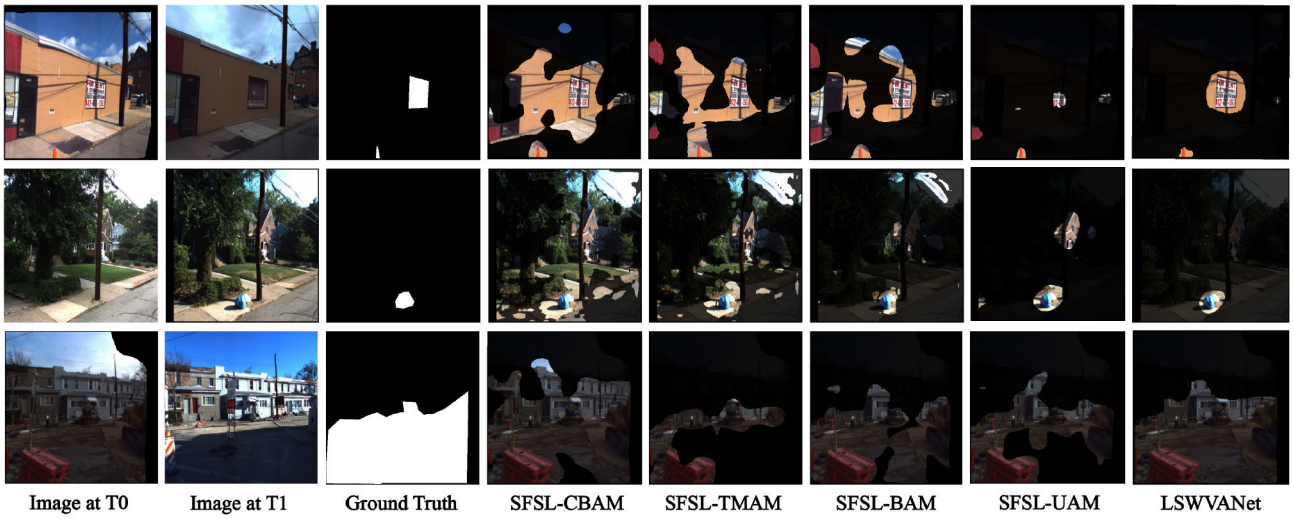


FIGURE 11. Visual quality comparison of attention masks based on different gate-based attention modules.

missed detection within dashed box C1 in S2. Similarly, the result of LSWVANet exhibits background overactivation within dashed boxes D2 in S2 and F2 in S3. In contrast to CL-based methods, HSACL-based methods present a cleaner background and smoother foreground. To further explore the superiority of HSACL, we also analyze the contributions of LSWVAM and HSACL to the accuracy improvement of the overall architecture. As shown in Table 2, SFSL-HSACL and LSWVANet-HL outperform LSWVANet by 2.6% and 6.0%, respectively, indicating that the main improvement is gained through HSACL. Although LSWVAM effectively detects subtle changes through a part-specific strategy, it may be sensitive to visual differences caused by image misalignment, leading to noise detection within dashed box F2 in S4. From the perspective of hard sample mining, missed detection correspond to positive hard samples, while noise detection relate to negative hard samples. Owing to the hard sample mining strategy, HSACL-based methods can effectively focus on these hard samples, ensuring that hard examples are 'well-optimized'.

To further verify the effectiveness of LSWVANet-HL in handling hard sample issues, we statistically analyse the positive/negative sample distance distribution at different distance intervals. Fig.10 shows the comparison of positive/negative sample distance distributions between LSWVANet-HL and other baseline methods. As shown in Fig.10, compared to the

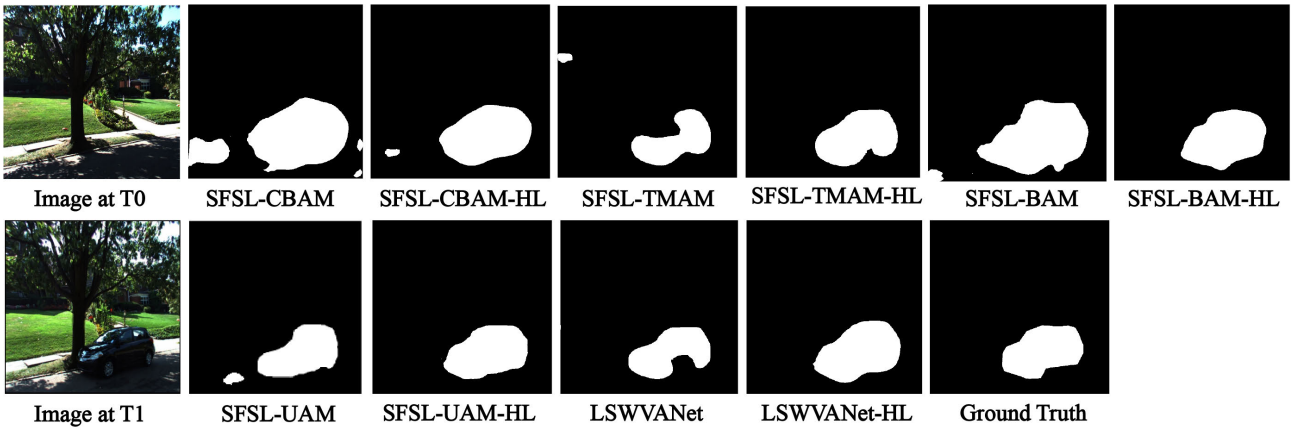


FIGURE 12. Visual quality comparison of different gate-based attention methods with and without HSACL.

TABLE 3. Comparison with other gate-based attention methods on with/without HSACL on the VL-CMU-CD validation set. CL: Contrastive loss, HSACL: Hard sample-aware contrastive loss. The best results are in bold.

Methods	Loss	P (%)	R (%)	F1 (%)
Baseline SFSL	CL	68.3	70.5	69.4
SFSL-CBAM [35]	CL	69.2	71.9	70.6
SFSL-TMAM [40]	CL	70.5	71.3	70.9
SFSL-BAM [15]	CL	71.4	71.6	71.5
SFSL-UAM [38]	CL	71.9	71.4	71.7
LSWVANet	CL	72.7	71.8	72.2
SFSL-CBAM [35]	HSACL	74.6	75.7	75.1
SFSL-TMAM-HL [40]	HSACL	74.9	75.5	75.2
SFSL-BAM-HL [15]	HSACL	77.0	75.8	76.4
SFSL-UAM-HL [38]	HSACL	77.1	76.7	76.9
LSWVANet-HL	HSACL	**79.1**	**77.3**	**78.2**

baseline SFSL, our method achieves a significant reduction in positive hard sample rates, with a decrease of 16.9% (22.3% vs. 5.4%), and a reduction in negative hard sample rates by 0.08% (0.09% vs. 0.01%).

2) COMPARISON WITH OTHER GATE-BASED ATTENTION METHODS

To explore the effectiveness of LSWVANet, we conduct an ablation experiment in comparison to other gate-based attention methods, including SFSL-CBAM [35], SFSL-BAM [15], SFSL-TMAM [40], and SFSL-UAM [38]. All models are built on the SFSL baseline with four stacked attention modules. The results in Table 3 show that our method significantly outperforms the baseline methods. Compared to the other gate-based attention methods, LSWVANet achieves an 1.6% improvement over SFSL-CBAM, 1.3% over SFSL-TMAM, 0.7% over SFSL-BAM, and 0.5% over SFSL-UAM.

Fig.11 illustrates a visual quality comparison of attention masks generated by different gate-based attention methods. The visualization results confirm the effectiveness of LSWVAM in two key aspects: (1) In comparison to CBAM and BAM, LSWVAM leverages the LSFP to localize part-specific discriminative change regions, significantly improving the detection of subtle changes in local parts (e.g., the window in the first row) and small-scale objects at various locations (e.g., the litter in the second row). (2) Compared to SFSL-TMAM and SFSL-UAM, LSWVANet employs the WVFP to extract long-range contextual features and enhance semantic consistency, thereby greatly improving the prediction of large-scale change regions (e.g., the construction regions in the third row).

TABLE 4. Comparison with other attention methods on computational cost on the VL-CMU-CD validation set. The best results are in bold.

Model Architecture	Params (M)	FLOPs (G)	Time (ms)	F1 (%)
Baseline SFSL	**44.74**	**78.04**	**18.3**	69.4
SFSL-CBAM [35]	46.51	125.12	25.8	70.6
SFSL-TMAM [40]	47.75	141.26	32.2	70.9
SFSL-BAM [15]	46.84	155.48	35.1	71.5
SFSL-UAM [38]	54.93	192.16	48.5	71.7
SFSL-SA [27]	45.58	391.83	74.6	72.0
SFSL-DA [30]	48.96	458.24	86.5	**72.2**
LSWVANet	47.43	174.37	42.8	**72.2**

We also evaluate the impact of HSACL on different gate-based attention methods. As shown in Table 3, HSACL achieves an absolute gain of 4.5%, 4.3%, 4.9%, and 5.2% in F1-score over SFSL-CBAM, SFSL-TMAM, SFSL-BAM and SFSL-UAM, respectively. Fig.12 provides a visual comparison of gate-based attention methods with and without HSACL. It is evident that HSACL focuses on mining negative hard samples such as noise detection and positive hard samples such as noise detection, resulting in a cleaner background and smoother foreground. Our proposed HSACL demonstrates robustness across different gate-based attention methods.

3) COMPUTATIONAL COST ANALYSIS

Street-view change detection algorithms are commonly employed in traffic surveillance and self-driving scenarios, making it crucial to balance real-time requirements with high accuracy. To validate the effectiveness of LSWVANet in terms of real-time performance, we conduct a computational

TABLE 5. Ablation study on patch sizes in LSFP module. The best results are in bold.

Modules	Patch Sizes $N_h \times N_w$							F1 (%)
	1x1	1x2	2x1	2x2	4x4	8x8	16x16	
Baseline SFSL								69.4
SFSL-CA	✓							69.9
SFSL-LSFP-SP				✓	✓	✓	✓	70.3
SFSL-LSFP-LP	✓	✓	✓	✓				70.5
SFSL-LSFP		✓	✓	✓	✓	✓		**70.8**

TABLE 6. Ablation study on dilation rates in WVFP module. The best results are in bold.

Modules	Dilation Rates (d_1, d_2)							F1 (%)
	(1,1)	(1,2)	(2,4)	(4,6)	(6,8)	(8,12)	(12,16)	
Baseline SFSL								69.4
SFSL-CA	✓							69.9
SFSL-WVFP-LR				✓	✓	✓	✓	70.5
SFSL-WVFP		✓	✓	✓	✓			**70.7**

TABLE 7. Varying γ for HSACL on the VL-CMU-CD validation set. The best results are in bold.

γ	Model	P (%)	R (%)	F1 (%)
0	SFSL	68.3	70.5	69.4
1	SFSL	73.9	72.4	73.2
2	SFSL	76.2	73.5	74.8
3	SFSL	72.1	71.2	71.7
4	SFSL	63.2	67.7	65.4
5	SFSL	62.7	67.5	65.1
0	LSWVANet	72.7	71.8	72.2
1	LSWVANet	75.9	76.8	76.4
2	LSWVANet	79.1	77.3	**78.2**
3	LSWVANet	76.2	75.1	75.5

TABLE 8. Comparison with other hard sample mining methods on the VL-CMU-CD validation set. OHEM: Online hard example mining, LC: Layer cascade, WCL: Weighted contrastive loss, HSACL: Hard sample-aware contrastive loss. The best results are in bold.

Methods	P (%)	R (%)	F1 (%)
Baseline SFSL	68.3	70.5	69.4
SFSL-WCL [19]	70.4	71.3	70.9
SFSL-OHEM [55]	71.8	72.4	72.1
SFSL-LC [54]	74.1	73.0	73.6
SFSL-HSACL	76.2	73.5	**74.8**

cost evaluation by comparing LSWVANet with several baseline methods using three metrics: the number of parameters (Params), the number of floating-point operations (FLOPs), and the inference time (Time (ms)). We evaluated both gate-based attention methods (GBA) and pairwise affinity attention methods (PAA), such as self-attention (SA [27]) and dual attention (DA [30]). All reported results are based on an input size of 512 × 512. Theoretically, for input features $F \in \mathbb{R}^{C\times H\times W}$, PAA computes pixel-to-pixel relations with a complexity of $\mathcal{O}(CH^2W^2)$, while GBA methods calculate attention masks with a complexity of $\mathcal{O}(CK^2HW)$, where K represents the convolution kernel size. Since the spatial size is larger than the kernel size, the computational costs of SFSL-SA and SFSL-DA are higher than those of other methods. From Table 4, it can be observed that PAA methods achieve higher accuracy but slower inference speeds. Specifically, LSWVANet achieves the same accuracy in F1 as DA but exhibits 2x faster inference speed and frames per second (FPS). Consequently, LSWVANet better meets real-time requirements than PAA methods under the same hardware conditions.

Among the GBA methods, SFSL-CBAM employs shallow convolutional layers, resulting in relatively low computational complexity. SFSL-BAM and SFSL-TMAM also employ channel reduction rates and stacked downsampling operations to reduce computational costs. SFSL-UAM uses a multibranch U-shape attention module to extract multi-scale features, effectively improving change detection performance, but also increasing model complexity. In contrast, LSWVANet adopts a feature grouping strategy to significantly reduce computational costs, enabling real-time inference at 23 FPS. Furthermore, LSWVANet leverages two well-designed feature pyramid units to extract relevant contextual features across local to global scales, achieving the highest accuracy in F1. The quantitative results in Table 4 demonstrate that LSWVANet achieves a favorable trade-off between real-time requirement and change detection accuracy when compared to other attention methods.

4) EFFECTIVENESS OF LSFP AND WVFP

To demonstrate the benefits of the LSFP and WVFP, we conduct a series of experiments using ablated variant modules. To address concerns about accuracy improvements due to extra parameters, we introduce a conventional attention module (CA) without patch-specific or dilated convolutional operations.

First, we explore the impact of the LSFP. In Table 5, we present the results obtained under several settings: (1) the baseline model SFSL built on ResNet50, (2) SFSL-CA, with four branches and no patch split ($\{N_h, N_w\} = \{1, 1\}, \{1, 1\}, \{1, 1\}, \{1, 1\}$), (3) SFSL-LSFP-SP with four branches and smaller patch sizes ($\{N_h, N_w\} = \{2, 2\}, \{4, 4\}, \{8, 8\}, \{16, 16\}$), (4) SFSL-LSFP-LP with four branches and larger patch sizes ($\{N_h, N_w\} = \{1, 1\}, \{1, 2\}, \{2, 1\}, \{2, 2\}$) and (5) SFSL-LSFP with the default settings ($\{N_h, N_w\} = \{1, 2\}, \{2, 1\}, \{2, 2\}, \{4, 4\}$). As shown in Table 5, the proposed LSFP obtains the best performance, achieving an 1.4% improvement over the baseline SFSL, 0.9% over SFSL-CA, 0.5% over SFSL-LSFP-SP and 0.3% over the baseline SFSL-LSFP-LP. It is worth noting that the performance of LSFP-LP decreases when using smaller part sizes than those employed by LSFP. One possible explanation for this is that extremely localized parts may fail to offer sufficient context for describing semantic information, which can adversely affect change detection accuracy.

Next, we conduct comparative experiments for the WVFP. In Table 6, we present results from several settings: (1) the baseline model SFSL built on ResNet50, (2) SFSL-CA, with four branches and no dilation rates ($\{d_1, d_2\} = \{1, 1\}, \{1, 1\}, \{1, 1\}, \{1, 1\}$), (3) SFSL-WVFP-LR with four branches

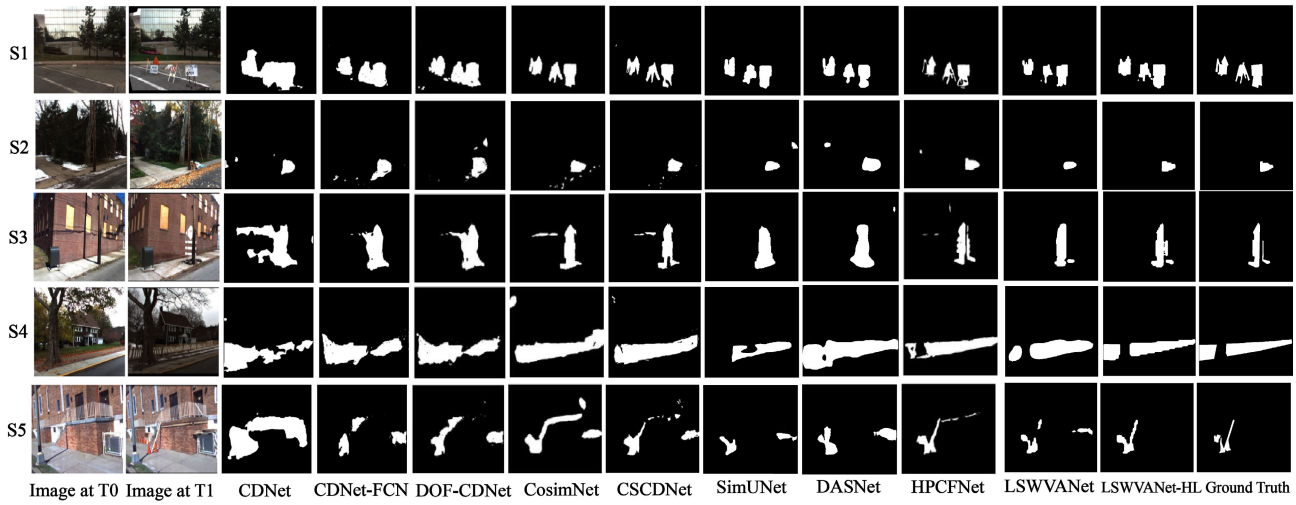


FIGURE 13. Visual quality comparison of different methods on a variety of challenging sequences of VL-CMU-CD testing set (S1-S5). From left to right in each row: Image at T0, Image at T1, CDNet [12], CDNet-FCN [12], DOF-CDNet [26], CosimNet [21], CSCDNet [64], SimUNet [38], DASNet [30], HPCFNet [23], LSWVANet, LSWVANet-HL and Ground Truth.

and larger dilation rates ($\{d_1, d_2\} = \{4, 6\}, \{6, 8\}, \{8, 12\}, \{12, 16\}$) and (4) SFSL-WVFP with the default settings ($\{d_1, d_2\} = \{1, 2\}, \{2, 4\}, \{4, 6\}, \{6, 8\}$). As shown in Table 6, the proposed SFSL-WVFP obtains the best performance, achieving an 1.3% improvement over the baseline SFSL, 0.8% over SFSL-CA, and 0.2% over SFSL-WVFP-LR. It's worth noting that the performance of WVFP-LR decreases when using larger dilation rates compared to WVFP, indicating that an overly large dilation rate introduces excessive irrelevant contextual information, leading to semantic confusion.

5) IMPORTANCE OF HYPERPARAMETER SELECTION FOR HSACL

As formulated in Equation 9, we employ the hyperparameter γ to downweight the easy samples. To determine the optimal value for γ, we conduct a series of experiments on the validation set of the VL-CMU-CD dataset, using both the baseline SFSL and LSWVANet models. We explore various values of γ within the range $\{0, 1, 2, 3, 4, 5\}$. The comparison results for different γ values are presented in Table 7. Through these comparative experiments, we discover that HSACL with $\gamma = 2$ significantly outperformed other hyperparameters. As a result, all subsequent experiments with HSACL are conducted using this optimal parameter setting as the default.

6) COMPARISON WITH OTHER HARD SAMPLE MINING METHODS

To further demonstrate the effectiveness of HSACL, we conduct an experiment comparing it with existing hard example mining methods, including Layer Cascade (LC) [54], Online Hard Example Mining (OHEM) [55] and Weighted Contrastive Loss (WCL) [19]. To ensure a fair comparison, we integrate all these methods into the same baseline (SFSL).

TABLE 9. Comparison of performance with the baseline method over the testing set of the VL-CMU-CD dataset. The best results are in bold.

Method	Publication	F1 (%)
CNN-Feat [8]	BMVC 2015	40.3
CDNet [12]	AR 2018	58.2
CDNet-FCN [12]	AR 2018	68.5
DOF-CDNet [26]	Arxiv 2017	68.8
CosimNet [21]	Arxiv 2018	70.6
CSCDNet [64]	ICRA 2020	71.0
SimUNet [38]	ARC 2021	71.4
DASNet [30]	RS 2020	72.1
HPCFNet [23]	TIP 2020	75.2
SimSac [25]	CVPR 2022	75.6
LSWVANet		72.0
LSWVANet-HL		**77.4**

As shown in Table 8, our method significantly outperforms other hard example mining techniques. Compared to WCL, OHEM, and LC, HSACL demonstrates substantial performance improvements of 3.9%, 2.7% and 1.2%, respectively. These results clearly indicate that HSACL enhances change detection.

E. COMPARISON WITH STATE-OF-THE-ART ALGORITHMS

1) RESULTS ON VL-CMU-CD DATASET

We evaluate existing advanced algorithms on the testing set of the VL-CMU-CD dataset. Specifically, we adopt ResNet50 as the backbone. The results are shown in Table 9. Compared with the baseline method based on semantic segmentation, LSWVANet improves accuracy by 3.5% over CDNet-FCN [12], 1.0% over CSCDNet [64]. In comparison to the baseline method based on feature similarity learning, LSWVANet enhances accuracy by 3.2% over DOF-CDNet [26], 1.4% over CosimNet [21], and 0.6% over SimUNet [38]. With the optimization of HSACL, our proposed LSWVANet-HL outperforms all existing advanced methods. Specifically, LSWVANet-HL achieves a 5.3%

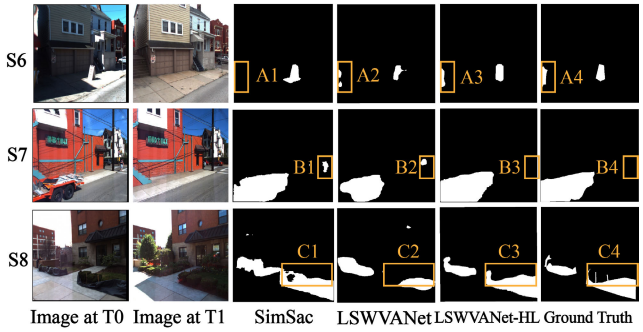


FIGURE 14. Visual quality comparison of SimSac and our proposed methods on a variety of challenging sequences of VL-CMU-CD testing set (S6-S8). From left to right in each row: Image at T0, Image at T1, SimSac [25], LSWVANet, LSWVANet-HL and Ground Truth.

improvement over DASNet [30], 2.2% over HPCFNet [23], and 1.8% over SimSac [25]. It's worth noting that compared with LSWVANet, LSWVANet-HL achieves a 5.4% improvement, indicating that the primary increase in accuracy is attributed to HSACL.

Furthermore, we select challenging sequences from the testing set, including scenes with lighting condition changes and seasonal variations, and conduct visual quality comparisons of different methods on various challenging scenes from the VL-CMU-CD testing set. As depicted in Fig.13, when compared with other baseline methods such as CDNet-FCN, DOF-CDNet, and SimUNet, LSWVANet refines the boundaries of large-scale objects, such as the fence in S4. Additionally, in comparison to SimSac in Fig.14, LSWVANet excels at identifying subtle changes, such as the road sign within box A2 in S6. However, LSWVANet still faces challenges with hard samples, resulting in missed detection within the dashed box C2 in S8 and noise detection within the dashed box B2 in S7. With the optimization of HSACL, LSWVANet-HL outperforms other existing advanced methods. We will discuss the performance improvement facilitated by HSACL in the following two aspects:

(1) Enhancing real change detection through positive hard sample learning: Distinguishing semantic changes is a key aspect of change detection tasks. When comparing the results with and without HSACL in Fig. 13 and Fig. 14, the missed detections produced by LSWVANet are refined by concentrating on training positive hard samples. For example, LSWVANet-HL improves semantic consistency in S4 and S8 and accurately locates small-scale objects like rubbish in S2. Compared with previous advanced methods, LSWVANet-HL also exhibits smoother foreground detection. For instance, in comparison to HPCFNet and DASNet, LSWVANet-HL enhances the performance of localizing foreground objects in S3 and S5. Compared to SimSac, LSWVANet-HL improves the ability to identify subtle changes in local parts within dashed box A3 in S6. Positive hard sample learning effectively enhances the identification of real changes and refines missed detections.

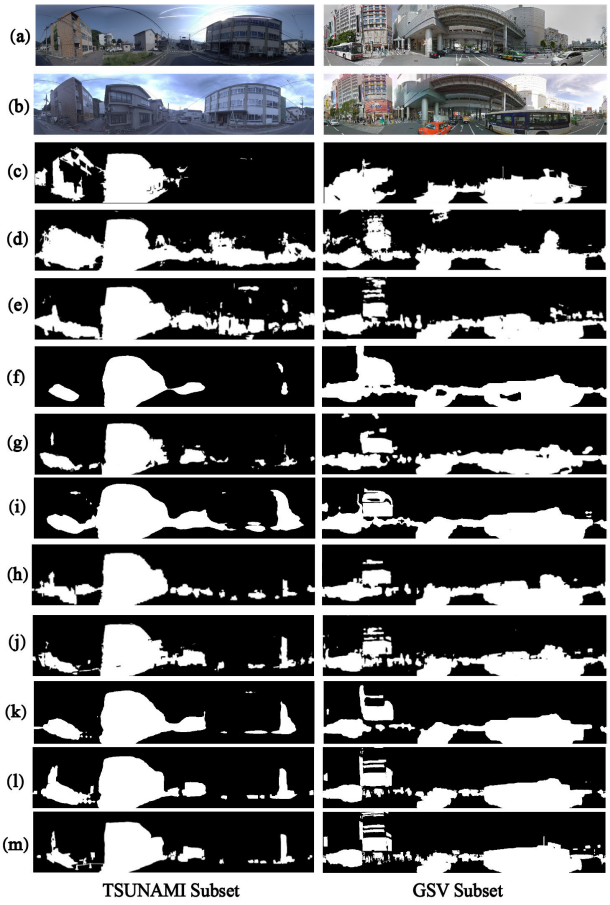


FIGURE 15. Visual quality comparison of different methods on the PCD2015 dataset. (a) Image at T0, (b) Image at T1, (c) CNN-Feat [8], (d) CDNet [12], (e) CosimNet [21], (f) SimUNet [38], (g) DOF-CDNet [26], (h) DASNet [30], (i) CSCDNet [64], (j) HPCFNet [23], (k) LSWVANet, (l) LSWVANet-HL, (m) Ground Truth.

(2) Rejecting noise detection through negative hard sample learning: Excluding noisy changes is another crucial aspect of the change detection task. As shown in Fig.13 and Fig.14, all the results predicted by baseline methods fail to eliminate fake changes, indicating that these models struggle to learn discriminative features and distinguish real changes from noisy changes. For instance, seasonal condition changes in S2 are misclassified as real changes by CDNet, DOF-CDNet, and CSCDNet. Lighting condition changes in S5 are also misclassified as structural changes by CSCDNet and HPCFNet. Similarly, SimSac fails to accurately locate structural changes, misclassifying the background as changes within box B1 in S7 and box C1 in S8. In contrast, LSWVANet-HL treats the noise detections as negative hard samples and places more emphasis on optimizing them until these hard samples are 'well-optimized'.

2) RESULTS ON PCD2015 DATASET

We evaluate the performance of LSWVANet-HL on PCD2015 dataset. From the comparison results described in Table 10, our method achieves state-of-the-art accuracy. Specifically, LSWVANet-HL has an 1.9% improvement

TABLE 10. Comparison of performance with other popular methods over the PCD2015 dataset. The best results are in bold.

Method	Publication	F1 (%)	
		Tsunami	GSV
CNN-Feat [8]	BMVC 2015	72.4	63.9
CDNet [12]	AR 2018	77.4	61.4
CosimNet [21]	Arxiv 2018	80.6	69.2
SimUNet [38]	ARC 2021	82.9	68.1
DOF-CDNet [26]	Arxiv 2017	83.8	70.3
DASNet [30]	RS 2020	84.1	74.5
CSCDNet [64]	ICRA 2020	85.9	73.8
HPCFNet [23]	TIP 2020	86.8	77.6
SimSac [25]	CVPR 2022	86.5	78.2
LSWVANet		84.6	74.4
LSWVANet-HL		**88.7**	**79.1**

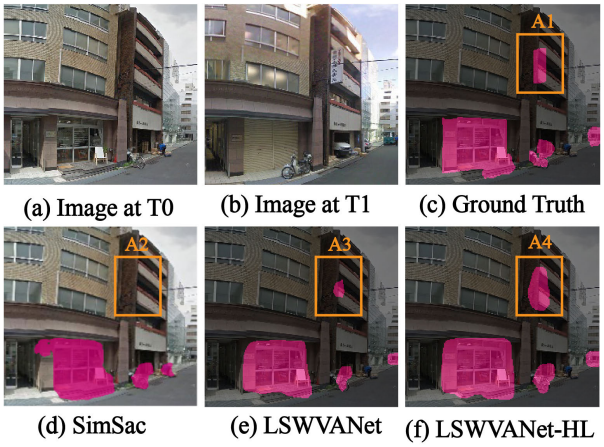


FIGURE 16. Visual quality comparison between SimSac, LSWVANet and LSWVANet-HL on the GSV dataset.

TABLE 11. Comparison of performance with other popular methods over the PSCD dataset. The best results are in bold.

Method	Publication	F1 (%)
CosimNet [21]	Arxiv 2018	67.6
SiameseCDResNet [64]	ICRA 2020	69.7
CSCDNet [64]	ICRA 2020	69.8
LSWVANet-HL		**71.3**

in the Tsunami dataset and an 1.5% improvement in the GSV dataset compared with HPCFNet [23]. Compared with SimSac [25], LSWVANet-HL achieves an 2.2% improvement for the Tsunami dataset and an 0.9% improvement for the GSV dataset. The visualization results of the proposed method are shown in Fig.15 and Fig.16. We observe that LSWVANet-HL achieves a smoother change map (e.g., large buildings and small pedestrians) and fewer noise detection, which proves that handling hard examples can solve the semantic inconsistence problem. In particular, when compared with SimSac in Fig. 16, our proposed methods exhibit superior capability in detecting subtle changes, such as the signboard within boxes A3 and A4.

3) RESULTS ON PSCD DATASET

Table 11 presents the quantitative comparison results for the PSCD dataset. As shown in Table 11, the proposed LSWVANet-HL outperforms existing methods. Specifically, compared to the segmentation-based baseline method, LSWVANet-HL achieves an improvement of 1.6% over SiameseCDResNet [64] and 1.5% over CSCDNet [64] in F1. The visualization results of the proposed method are depicted in Fig.17. From the representation of dashed rectangles, LSWVANet-HL can effectively locate and delineate real changes. In contrast, the baseline method CSCDNet misses many subtle changes due to heavy occlusion (e.g., the building located at the yellow dashed rectangle, the traffic pole located at the blue dashed rectangle, the advertisement board located at the red dashed rectangle).

V. DISCUSSION

A. THE ROLE OF HSACL

To further validate the effectiveness of the HSACL during the optimization phase, we analyze the gradient distributions of the converged model. For a sample $s_i = \{y_i, D_i\}$, we treat the distance value D_i as the independent variable and compute the gradients by calculating the derivatives of the CL and HSACL with respect to the distance value D_i. For a positive sample, the gradient is computed as follows:

For CL:

$$\frac{\partial CL}{\partial D} = -2(m_2 - D) = G^{Pos} * \frac{\partial CL}{\partial D}$$
$$G^{Pos} = 1 \quad (\gamma = 0) \tag{13}$$

For HSACL:

$$\frac{\partial HSACL}{\partial D} = G^{Pos} * (-2(m_2 - D)) = G^{Pos} * \frac{\partial CL}{\partial D}$$
$$G^{Pos} = (e^{-D})^{\gamma}(\gamma(m_2 - D) + 2)/2 \quad (\gamma \geq 1) \tag{14}$$

Meanwhile, for a negative example, the gradient is computed as follows:

For CL:

$$\frac{\partial CL}{\partial D} = 2D = G^{Neg} * \frac{\partial CL}{\partial D}$$
$$G^{Neg} = 1 \quad (\gamma = 0) \tag{15}$$

For HSACL:

$$\frac{\partial HSACL}{\partial D} = G^{Neg} * (2D) = G^{Neg} * \frac{\partial CL}{\partial D}$$
$$G^{Neg} = (1 - e^{-D})^{\gamma - 1}(2 + \gamma D e^{-D})/2 \quad (\gamma \geq 1) \tag{16}$$

From Equation 14 and 16, it can be observed that the HSACL gradient calculation for positive and negative samples can be regarded as introducing scaling factors G^{Pos} and G^{Neg} to the CL gradient calculation. The curve of the gradient scaling factors G^{Pos} and G^{Neg} with respect to the distance value is illustrated in Fig.18 (a). For positive samples, when $\gamma = 0$, all gradient weights are equal to 1. As γ increases, the gradient weight assigned to easy samples approaches 0, while more weight concentrates on the hard samples. Similarly, for negative samples, the gradient weight assigned to the easy samples tends to 0, while the gradient weight assigned to the hard sample increases to 1. Consequently, HSACL prevents the numerous easy

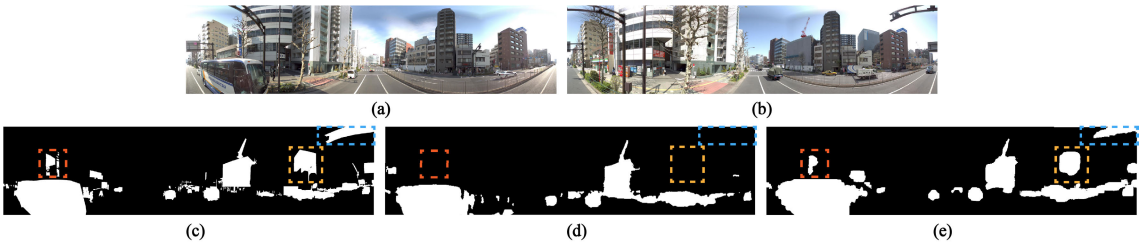
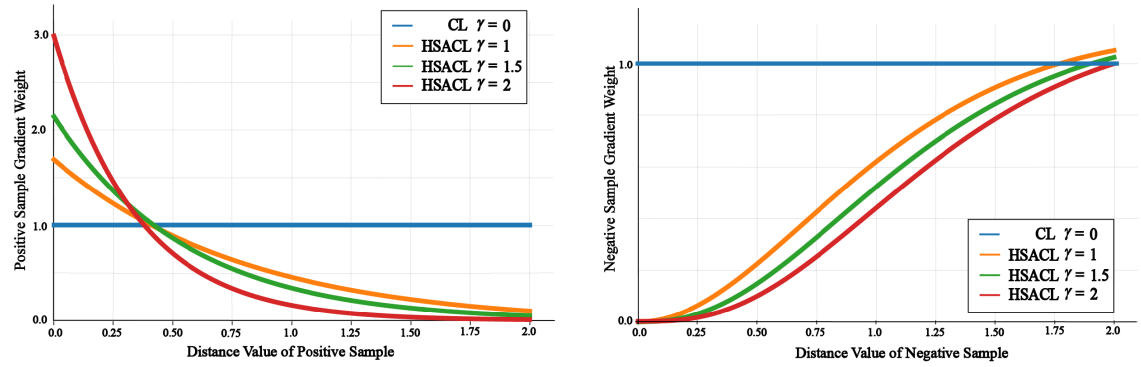
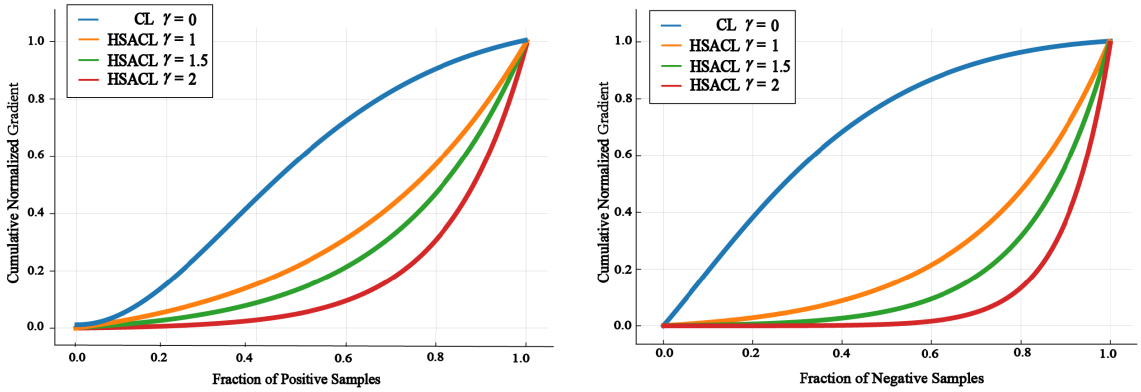


FIGURE 17. Visual quality comparison of different methods on the PSCD dataset. (a) Image at T0, (b) Image at T1, (c) Ground Truth, (d) CSCDNet [64], (e) LSWVANet-HL. Subtle changes within dashed rectangles present that our method has better detection results compared with CSCDNet.



(a) Sample gradient weight distribution for positive and negative samples



(b) Cumulative distribution functions of normalized gradient for positive and negative samples

FIGURE 18. (a) Sample gradient weight distribution for positive and negative samples for different values of γ. (b) Cumulative distribution functions for normalized gradient for positive and negative samples for different values of γ.

samples from dominating the gradients during the training process and forces the model updates to concentrate on the informative hard samples. To gain further insight into the HSACL, we analyze how HSACL addresses the intraclass and interclass imbalances between easy and hard samples.

1) INTRACLASS IMBALANCE ANALYSIS

We conduct an analysis of the intraclass imbalance using the cumulative gradient distribution function. For both positive and negative samples, we calculate the gradients and normalize them to ensure that their sum equals one. Then, we sort the normalized gradient values from low to high. The cumulative distribution of the normalized gradients for the positive and negative samples, with varying settings for γ, is illustrated in Fig.18 (b). When $\gamma = 0$, the gradients are dominated by numerous easy samples during the training process. Specifically, for positive samples, approximately 20% of the hardest positive samples contribute only 10% of the total gradients. With increasing γ, the majority of the gradient originates from a small fraction of hard samples, which become concentrated in the top 20% of the hardest samples. In detail, with $\gamma = 2$, approximately 20% of the hardest positive samples account for approximately 70% of the total gradients. Furthermore, the impact of γ on the negative samples is even more significant. With $\gamma = 0$, approximately 20% of the hardest negative samples account for only 5% of the total gradients of the negative samples. However, with $\gamma = 2$, approximately 20% of the hardest

TABLE 12. Ablation study on interclass imbalance with different values of γ.

Interclass Imbalance	NPR	PR	NR
CL ($\gamma = 0$)	3.27	0.223	0.766
HSACL ($\gamma = 1$)	1.71	0.367	0.632
HSACL ($\gamma = 2$)	**1.22**	**0.451**	**0.549**

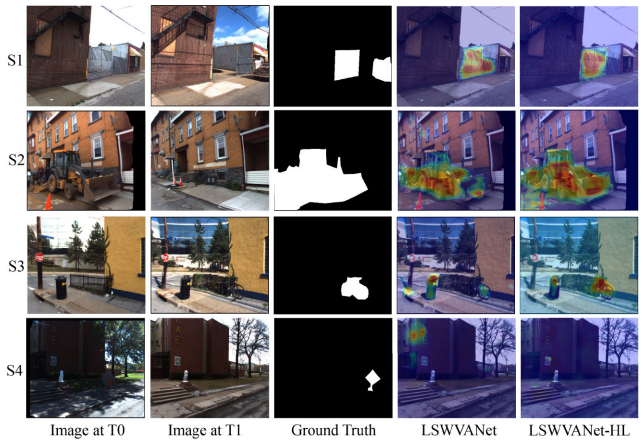


FIGURE 19. Visual quality comparison of the attention masks generated by the last LSWVAM between LSWVANet and LSWVANet-HL.

positive samples contribute to approximately 85% of the total gradients. Clearly, the HSACL can effectively downweight easy samples from both changed and unchanged classes, which addresses the intraclass imbalance issues.

2) INTERCLASS IMBALANCE ANALYSIS

We analyze the impact of HSACL on the interclass imbalance between positive and negative samples. We first separately compute the positive gradients (PG) and negative gradients (NG). Based on this, we calculate the ratio of negative gradients to positive gradients (NPR), the positive gradient rate (PR), and the negative gradient rate (NR). The calculations are as follows:

$$\begin{aligned} NPR &= \frac{NG}{PG} \\ PR &= \frac{PG}{PG + NG} \\ NR &= \frac{NG}{PG + NG} \end{aligned} \tag{17}$$

Due to the dynamic change in the number of positive and negative samples during HSACL optimization, we do not specifically set parameters to balance the importance of positive and negative samples. However, as shown in Table 12, due to the significant suppression of numerous negative easy samples, HSACL also mitigates the interclass imbalance problem.

3) THE IMPACT OF HSACL ON LSWVAM

To clearly demonstrate the superiority of the proposed HSACL, we compare the learned attention masks of the last attention module between LSWVANet and LSWVANet-HL. As depicted in Fig.19, the attention masks generated by LSWVANet provide insights into where the network should focus. However, there are still instances of failure, such as false attention applied to other objects (e.g., the door in S1 and the bin in S3) and background overactivation (e.g., the lighting condition noises in S4). Moreover, in Fig.19, it is evident that the attention mask with the optimization of HSACL covers the change regions more effectively than that of LSWVANet. For example, the local specificity of subtle changes in S1 and the long-range consistency of large-scale objects in S2 are further improved. Meanwhile, unnecessary noisy changes (e.g., the lighting condition changes in S3 and S4) are significantly suppressed. It appears that LSWVANet-HL has a better ability to concentrate on the most relevant subtle changes and ignore noisy changes with the optimization of HSACL.

B. HARD SAMPLE DISTANCE DISTRIBUTION

The greatest problem with hard samples is that positive samples with small distance values or negative samples with large distance values can increase the intraclass distance variance. Meanwhile, positive samples with large small values can entangle with negative samples, which may reduce interclass separability. To demonstrate the effectiveness of the proposed method, we visualize changes in the hard sample distance distribution with respect to the training epoch. The details of the hard sample distance distribution are illustrated in Fig.20.

From the visualization of the distance distribution, we can observe the following: (1) At the beginning of training processing (e.g., epoch 0), it is impossible to distinguish semantic changes from noisy changes because the positive and negative sample distance distributions are all mixed. (2) During the optimization (e.g., epoch 8 to epoch 90), the distance value of positive samples grows as the training process progresses, while the distance value of negative samples decreases, indicating that an increasing number of hard samples become easy samples by enlarging the distance of positive samples and reducing the distance of negative samples. However, there are still many hard samples, for example, negative samples with large distances, which give rise to incorrect activation in background regions, and positive samples with small distances, making the prediction fragmented. (3) As the training processing is finished, the majority of positive sample distances are constrained to the upper bound (margin $m_2 = 2$), and most of the negative sample distances tend to the lower bound (margin $m_1 = 0$), significantly reducing the intraclass distance variance and interclass distance separability. These observations prove the effectiveness of our proposed method.

C. FEATURE LATENT SPACE DISTRIBUTION

The fundamental goal for addressing hard sample issues is learning a discriminative feature representation. To further explore how the proposed model achieves the goal, we use the t-SNE [65], [66] algorithm to visualize the feature latent

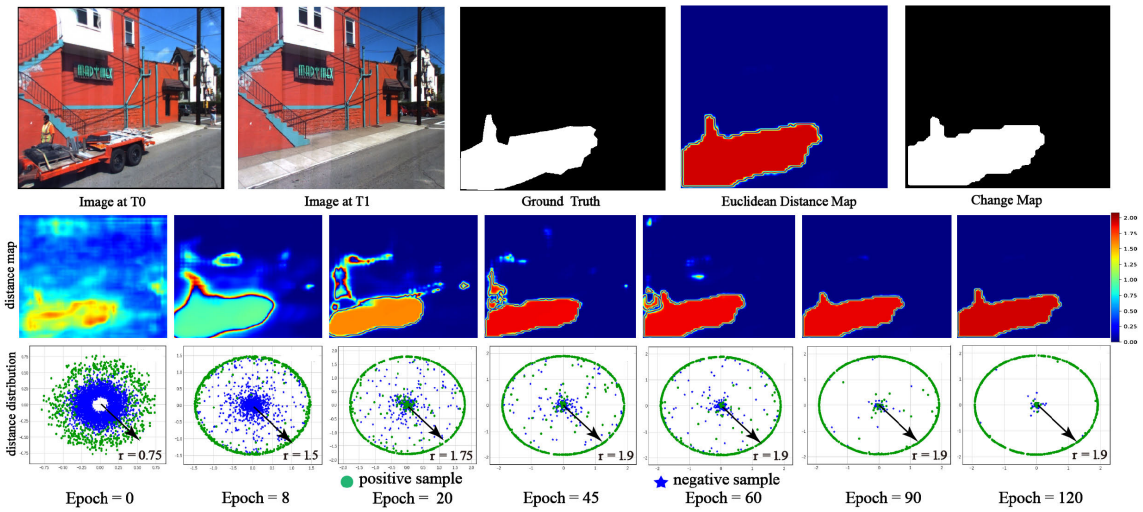


FIGURE 20. Visualization of changes in the hard sample distance distribution under polar coordinates during the training process.

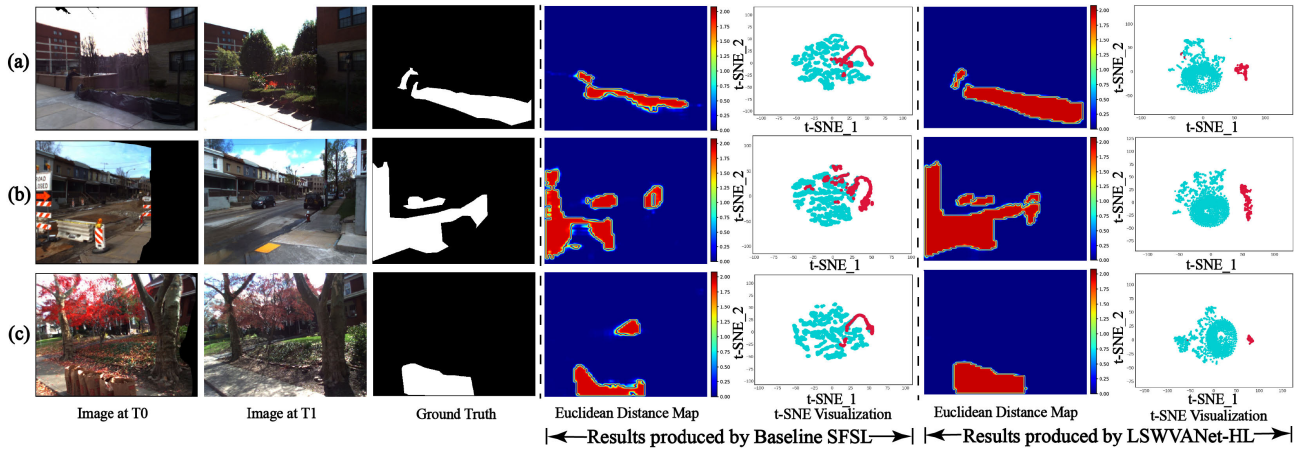


FIGURE 21. Comparison visualization result of feature latent space distribution between Baseline SFSL and LSWVANet-HL.

space distribution of the last feature layer and provide a visual comparison between Baseline SFSL and LSWVANet-HL. The results of the two-dimensional feature sample are illustrated in Fig.21, where red dots denote positive/changed samples and cyan dots represent negative/unchanged samples. From the comparison results described in Fig.21, the advantage of the proposed LSWVANet-HL is that it can keep the boundary smooth and make change detection consistent. Moreover, we observe that learned features extracted from LSWVANet-HL have larger interclass separability and smaller intraclass variations, making features scatter and gather more distinctly. It seems that handling hard sample issues forces intraclass compactness and interclass separability, contributing to learning more disentangled features and leading to better performance.

VI. CONCLUSION

In this paper, we proposed the Local-Specificity and Wide-View Attention Network to adapt to location and scale variations of change regions. Our attention network could not only take into account long-range contextual information but also emphasize the local specificity within discriminative local parts, enhancing the detection accuracy of subtle changes in local parts as well as significant changes in large-scale regions. To tackle the issue of heavy imbalance between easy and hard samples, we introduced a novel sample-specific loss function called Hard Sample-Aware Contrastive Loss, which downweights easy samples from both changed and unchanged categories, putting more focus on training informative hard samples. Experiments conducted on three datasets (i.e., VL-CMU-CD, PCD2015 and PSCD) clearly demonstrate the effectiveness of our approach.

In the future, we will focus on addressing the limitations in two aspects. Firstly, while LSWVAM captures both local-specificity and long-range contexts, the generation of the attention mask relies solely on spatial-wise features. Our future research will aim to incorporate more channel-wise and spatial-temporal information. Secondly, the supervised

learning approach requires a large number of annotations; however, acquiring annotations is time-consuming. Therefore, the next research will also focus on detecting changes using unsupervised methods or prompt learning methods.

REFERENCES

[1] K. Sakurada, T. Okatani, and K. Deguchi, "Detecting changes in 3D structure of a scene from multi-view images captured by a vehicle-mounted camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 137–144.

[2] Z. J. Yew and G. H. Lee, "City-scale scene change detection using point clouds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13362–13369.

[3] R. Sachdeva and A. Zisserman, "The change you want to see," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2023, pp. 3982–3991.

[4] S. A. Ahmed, D. P. Dogra, S. Kar, R. Patnaik, S.-C. Lee, H. Choi, G. P. Nam, and I.-J. Kim, "Query-based video synopsis for intelligent traffic monitoring applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3457–3468, Aug. 2020.

[5] S. Wan, X. Xu, T. Wang, and Z. Gu, "An intelligent video analysis method for abnormal event detection in intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4487–4495, Jul. 2021.

[6] H. Kim, J. Park, K. Min, and K. Huh, "Anomaly monitoring framework in lane detection with a generative adversarial network," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1603–1615, Mar. 2021.

[7] D. Roy, T. Ishizaka, C. K. Mohan, and A. Fukuda, "Detection of collision-prone vehicle behavior at intersections using Siamese interaction LSTM," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3137–3147, Apr. 2022.

[8] K. Sakurada and T. Okatani, "Change detection from a street image pair using CNN features and superpixel segmentation," in *Proc. Mach. Vis. Conf.*, 2015, pp. 1–12.

[9] T. P. Nguyen, C. C. Pham, S. V. Ha, and J. W. Jeon, "Change detection by training a triplet network for motion feature extraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 433–446, Feb. 2019.

[10] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with deeply transferred motion vector CNNs," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2326–2339, May 2018.

[11] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 900–918, Mar. 2020.

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

[13] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 1055–1059.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[15] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "A simple and light-weight attention module for convolutional neural networks," *Int. J. Comput. Vis.*, vol. 128, no. 4, pp. 783–798, Apr. 2020.

[16] S. Bu, Q. Li, P. Han, P. Leng, and K. Li, "Mask-CDNet: A mask based pixel change detection network," *Neurocomputing*, vol. 378, pp. 166–178, Feb. 2020.

[17] C. Xu, Z. Ye, L. Mei, W. Yang, Y. Hou, S. Shen, W. Ouyang, and Z. Ye, "Progressive context-aware aggregation network combining multi-scale and multi-level dense reconstruction for building change detection," *Remote Sens.*, vol. 15, no. 8, p. 1958, Apr. 2023.

[18] W. Wiratama, J. Lee, S.-E. Park, and D. Sim, "Dual-dense convolution network for change detection of high-resolution panchromatic imagery," *Appl. Sci.*, vol. 8, no. 10, p. 1785, Oct. 2018.

[19] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep Siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.

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

[21] E. Guo, X. Fu, J. Zhu, M. Deng, Y. Liu, Q. Zhu, and H. Li, "Learning to measure change: Fully convolutional Siamese metric networks for scene change detection," 2018, *arXiv:1810.09111*.

[22] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "ChangeNet: A deep learning architecture for visual change detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Munich, Germany, Sep. 2018, pp. 129–145.

[23] Y. Lei, D. Peng, P. Zhang, Q. Ke, and H. Li, "Hierarchical paired channel fusion network for street scene change detection," *IEEE Trans. Image Process.*, vol. 30, pp. 55–67, 2021.

[24] K. Doi, R. Hamaguchi, Y. Iwasawa, M. Onishi, Y. Matsuo, and K. Sakurada, "Detecting object-level scene changes in images with viewpoint differences using graph matching," *Remote Sens.*, vol. 14, no. 17, p. 4225, Aug. 2022.

[25] J.-M. Park, U.-H. Kim, S.-H. Lee, and J.-H. Kim, "Dual task learning by leveraging both dense correspondence and Mis-correspondence for robust change detection with imperfect matches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 13739–13749.

[26] K. Sakurada, W. Wang, N. Kawaguchi, and R. Nakamura, "Dense optical flow based change detection network robust to difference of camera viewpoints," 2017, *arXiv:1712.02941*.

[27] H. Chen and Z. Shi, "A spatial–temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, May 2020.

[28] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.

[29] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.

[30] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2021.

[31] Y. Zhou, F. Wang, J. Zhao, R. Yao, S. Chen, and H. Ma, "Spatial-temporal based multihead self-attention for remote sensing image change detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6615–6626, Oct. 2022.

[32] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "UniFormer: Unifying convolution and self-attention for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12581–12600, Oct. 2023.

[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[34] H. Zhang, M. Wang, F. Wang, G. Yang, Y. Zhang, J. Jia, and S. Wang, "A novel squeeze-and-excitation W-Net for 2D and 3D building change detection with multi-source and multi-feature remote sensing data," *Remote Sens.*, vol. 13, no. 3, p. 440, Jan. 2021.

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

[36] Q. Ke and P. Zhang, "CS-HSNet: A cross-Siamese change detection network based on hierarchical-split attention," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9987–10002, 2021.

[37] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2735–2745.

[38] Z. Jiezhong, C. Yong, K. Fuyang, and Z. Guorong, "Building change detection from high resolution remote sensing imagery based on Siam-UNet++," *Appl. Res. Comput.*, vol. 38, no. 11, pp. 3460–3465, 2021.

[39] L. Wang, L. Wang, Q. Wang, and P. M. Atkinson, "SSA-SiamNet: Spectral–spatial-wise attention-based Siamese network for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5510018.

[40] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458.

[41] T. Lei, D. Xue, H. Ning, S. Yang, Z. Lv, and A. K. Nandi, "Local and global feature learning with kernel scale-adaptive attention network for VHR remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7308–7322, 2022.
[42] Y. Huang, L. Zhang, W. Qi, C. Huang, and R. Song, "Contrastive self-supervised two-domain residual attention network with random augmentation pool for hyperspectral change detection," *Remote Sens.*, vol. 15, no. 15, p. 3739, Jul. 2023.
[43] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2999–3007.
[45] H. Dong, W. Ma, Y. Wu, M. Gong, and L. Jiao, "Local descriptor learning for change detection in synthetic aperture radar images via convolutional neural networks," *IEEE Access*, vol. 7, pp. 15389–15403, 2019.
[46] J. Geng, H. Wang, J. Fan, and X. Ma, "Change detection of SAR images based on supervised contractive autoencoders and fuzzy clustering," in *Proc. Int. Workshop Remote Sens. With Intell. Process. (RSIP)*, Shanghai, China, May 2017, pp. 1–3.
[47] T. Zhan, M. Gong, X. Jiang, and S. Li, "Log-based transformation feature learning for change detection in heterogeneous images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1352–1356, Sep. 2018.
[48] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.
[49] R. Ghosh, S. Phadikar, N. Deb, N. Sinha, P. Das, and E. Ghaderpour, "Automatic eyeblink and muscular artifact detection and removal from EEG signals using k-nearest neighbor classifier and long short-term memory networks," *IEEE Sensors J.*, vol. 23, no. 5, pp. 5422–5436, Mar. 2023.
[50] M. Gong, T. Zhan, P. Zhang, and Q. Miao, "Superpixel-based difference representation learning for change detection in multispectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2658–2673, May 2017.
[51] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019.
[52] Y. Li, C. Peng, Y. Chen, L. Jiao, L. Zhou, and R. Shang, "A deep learning method for change detection in synthetic aperture radar images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5751–5763, Aug. 2019.
[53] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8577–8584.
[54] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang, "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6459–6468.
[55] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 761–769.
[56] K. Song and J. Jiang, "AGCDetNet: An attention-guided network for building change detection in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4816–4831, 2021.
[57] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet v2: Practical guidelines for efficient CNN architecture design," in *Computer Vision—ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 122–138.
[58] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2019, pp. 8024–8035.
[59] R. Naushad, T. Kaur, and E. Ghaderpour, "Deep transfer learning for land use and land cover classification: A comparative study," *Sensors*, vol. 21, no. 23, p. 8083, Dec. 2021.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[61] X. Dai, X. Zhao, F. Cen, and F. Zhu, "Data augmentation using mixup and random erasing," in *Proc. IEEE Int. Conf. Netw., Sens. Control (ICNSC)*, Dec. 2022, pp. 1–6.
[62] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, and A. Almuhaimeed, "Stop oversampling for class imbalance learning: A review," *IEEE Access*, vol. 10, pp. 47643–47660, 2022.
[63] S. Yang, F. Song, G. Jeon, and R. Sun, "Scene changes understanding framework based on graph convolutional networks and Swin transformer blocks for monitoring LCLU using high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 15, p. 3709, Aug. 2022.
[64] K. Sakurada, M. Shibuya, and W. Wang, "Weakly supervised silhouette-based semantic scene change detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 6861–6867.
[65] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.
[66] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Sep. 2023.

ENQIANG GUO is currently pursuing the Ph.D. degree with the School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China. His research interests include intelligent transportation systems, semantic segmentation, deep metric learning, and attention mechanisms in deep learning models.

XINSHA FU gained the State Council Special Allowance, in 1993, and was exceptionally promoted as a Professor, in 1998. He works on highway planning and design, computer-aided engineering and design of highways, transportation infrastructure management systems, intelligent transportation systems, 3S technology, and teaching and research of traffic information. He has published over 60 articles and five monographs. He was awarded two second prizes and seven third prizes of provincial and ministry-level awards.