remote sensing

MDPI

Article

Cross-Visual Style Change Detection for Remote Sensing Images via Representation Consistency Deep Supervised Learning

Jinjiang Wei [1], Kaimin Sun [1,2,*], Wenzhuo Li [3], Wangbin Li [4], Song Gao [1], Shunxia Miao [1], Yingjiao Tan [1], Wei Cui [1] and Yu Duan [1]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China; weijinjiang@whu.edu.cn (J.W.); gaosong@whu.edu.cn (S.G.); shunxiamiao@whu.edu.cn (S.M.); yjtann@whu.edu.cn (Y.T.); cuiwei@whu.edu.cn (W.C.); 2023106190026@whu.edu.cn (Y.D.)
[2] State Key Laboratory of Geo-Information Engineering, Xi'an 710054, China
[3] School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430072, China; liwenzhuo@wust.edu.cn
[4] Qinhuangdao Branch Campus, Northeastern University, Qinhuangdao 066004, China; liwangbin@neuq.edu.cn
* Correspondence: sunkm@whu.edu.cn; Tel.: +86-135-5409-5825

Abstract: Change detection techniques, which extract different regions of interest from bi-temporal remote sensing images, play a crucial role in various fields such as environmental protection, damage assessment, and urban planning. However, visual style interferences stemming from varying acquisition times, such as radiation, weather, and phenology changes, often lead to false detections. Existing methods struggle to robustly measure background similarity in the presence of such discrepancies and lack quantitative validation for assessing their effectiveness. To address these limitations, we propose Representation Consistency Change Detection (RCCD), a novel deep learning framework that enforces global style and local spatial consistency of features across encoding and decoding stages for robust cross-visual style change detection. RCCD leverages large-kernel convolutional supervision for local spatial context awareness and global content-aware style transfer for feature harmonization, effectively suppressing interference from background variations. Extensive evaluations on S2Looking and LEVIR-CD+ datasets demonstrate RCCD's superior performance, achieving state-of-the-art F1-scores. Furthermore, on dedicated subsets with large visual style differences, RCCD exhibits more substantial improvements, highlighting its effectiveness in mitigating interference caused by visual style errors. The code has been open-sourced on GitHub.

Keywords: change detection; deep supervision; representation consistency; remote sensing; error validation

Academic Editors: Xiaobo Liu, Yaoming Cai and Pedram Ghamisi

Received: 15 January 2025
Revised: 17 February 2025
Accepted: 22 February 2025
Published: 25 February 2025

Citation: Wei, J.; Sun, K.; Li, W.; Li, W.; Gao, S.; Miao, S.; Tan, Y.; Cui, W.; Duan, Y. Cross-Visual Style Change Detection for Remote Sensing Images via Representation Consistency Deep Supervised Learning. *Remote Sens.* **2025**, *17*, 798. https://doi.org/10.3390/rs17050798

1. Introduction

Change detection (CD) techniques extract altered regions from bi-temporal remote sensing images, providing a crucial means to monitor Earth's surface changes over time [1]. CD plays a vital role in environmental monitoring [2], disaster assessment, and urban planning, enabling the tracking of deforestation, glacier retreat, damage assessment, and urban development. This information facilitates informed decision-making and effective resource management [3,4]. Consequently, CD has become an indispensable tool with significant societal and environmental benefits, and the development of accurate and efficient CD methods remains an active research area [5].

In the era of interconnected satellite constellations [6], the increasing demand for accurate and timely CD from bi-temporal images faces challenges from visual style interferences caused by varying acquisition conditions. These interferences, primarily stemming from radiation differences, atmospheric effects, and seasonal variations, introduce significant difficulties in accurately identifying genuine changes, as shown in Figure 1. These non-change-related pixel value variations manifest as spectral, textural, and semantic discrepancies between images, leading to false detections. For instance, illumination variations can cause shifts in spectral signatures, making similar objects appear different. Atmospheric conditions like haze can weaken textural features, potentially obscuring genuine changes or introducing spurious ones. Moreover, seasonal changes, such as vegetation growth cycles, can lead to substantial semantic changes, even if irrelevant to the user's objective. Consequently, these visual style discrepancies pose a significant obstacle to achieving high accuracy in CD, particularly in measuring the similarity between unchanged and changed areas across spectral, textural, and semantic domains. This necessitates the development of robust CD algorithms that can effectively mitigate the impact of such interferences.

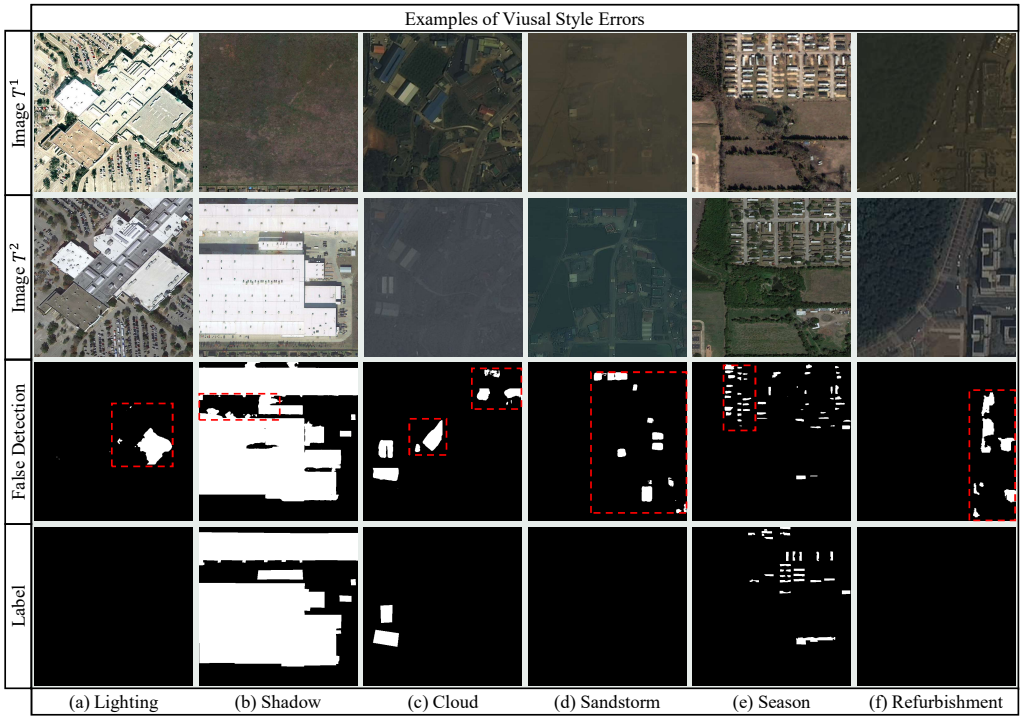

Figure 1. Six examples include the T^1 image, T^2 image, false detection map, and label. Red boxes highlight the significant areas of false detection.

Current CD research primarily focuses on mitigating visual style interferences through specialized datasets, sophisticated feature enhancements, and refined deep supervision. However, accurately quantifying background similarity amid complex visual style variations and establishing robust evaluation protocols remain challenging, as exemplified by the false detections arising from complex backgrounds in Figure 1. Traditional and machine learning CD methods struggle with varying visual styles [7], particularly radiation and non-linear weather/phenological changes. Deep learning, such as Siamese networks for CD, has emerged as a dominant approach due to its capacity for high-dimensional semantic understanding [8]. Dedicated visual CD datasets [9,10] facilitate method adaptation to specific interference patterns, yet they often lack the full spectrum of real-world variations. Methods explicitly addressing visual style interferences primarily leverage feature enhancement strategies, including difference feature enhancement [11,12], multi-

level feature fusion [13–15], and global semantic enhancement [16–18], to suppress false positives. Deep supervision methods, employing techniques like metric learning [9] and multi-scale supervision [19], also play a vital role. However, these approaches often lack effective feature constraints and measurements, hindering accurate similarity measurement in complex backgrounds by failing to establish consistent feature contrast during Siamese feature extraction. Furthermore, robustly quantifying visual style interference errors remains challenging, hindering comprehensive method assessment. Next, we provide a detailed introduction to the progress in related work for these two challenges.

Recent advances in deep learning [8] for change detection have achieved significant progress through encoder–decoder architectures that leverage weight-shared, pre-trained backbones to extract multi-temporal features [14,20–22]. Researchers enhance robustness via multi-task learning [23,24], training constraints [25–27], and deep supervision [19], with innovations including pyramid networks for salient feature refinement [28], multi-scale fusion frameworks like MSFF-CDNet [29], and temporal modeling with P2V-CD [18]. SEIFNet further integrates global-local cues to refine boundaries [30], while deep supervision strategies suppress noise through multi-level supervision [31], atrous spatial pyramid pooling [32], attention-based fusion [33], and GAN constraints [34]. However, despite these efforts, Siamese CD frameworks remain limited by unsystematic designs, sensitivity to interference, and insufficient utilization of advanced structures. Crucially, existing methods often neglect comprehensive constraints on bi-temporal backbone features, failing to ensure comparable and identically distributed representations during Siamese feature extraction [1,2,19], which undermines their robustness in complex scenarios.

A persistent challenge in optical remote sensing CD is visual style interference caused by radiometric, seasonal, and phenological variations [20,35]. While specialized datasets like LEVIR-CD+ [9] and S2Looking [10] improve adaptability, feature enhancement strategies dominate current solutions, including difference amplification [12], multi-level fusion [19], and global semantic integration [30]. Techniques such as multi-level difference enhancement [36] and similarity-aware attention networks [37] aim to suppress false positives, yet critical gaps persist: (1) inadequate constraints during Siamese feature extraction hinder consistent feature contrast in complex backgrounds, and (2) robust protocols for quantifying style interference errors remain underdeveloped. Traditional and deep learning methods struggle to disentangle style variations from genuine changes [7], while existing datasets lack real-world heterogeneity [11,13]. This limitation perpetuates false detections in scenarios with large visual style differences (Figure 1), underscoring the need for systematic feature constraints and standardized evaluation metrics [14–18].

To address the challenges in robustly measuring background similarity and quantitatively evaluating errors, we propose the Representation Consistency Change Detection (RCCD) framework. This framework is explicitly designed to handle remote sensing images with cross-visual style variations, aiming to achieve robust CD by enforcing representation consistency between bi-temporal global style and local spatial consistency of features throughout the encoding and decoding pipeline. RCCD comprises an encoder, a decoder, and a deep supervisor. The encoder utilizes a pre-trained network with shared weights for multi-level feature extraction. The Unified Decoder and Detector (UDD), equipped with skip connections and residual units, performs scale-wise decoding to generate the CD map. The Representation Consistency (RC) deep supervisor, active only during training, operates at all scales throughout the encoding and decoding process. It consists of local spatial contexts based on the large kernel convolution supervised (LS) module and the Global content-aware Style transfer (GS) module, constraining feature representation consistency includes local context and global style promoting effective CD. The LS module comprehensively assesses correlations between background and target regions in the bi-temporal

features. The GS module computes the similarity of channel-wise auto-correlation Gram matrices, implicitly performing style distribution transfer to harmonize global feature representations. Furthermore, we introduce a novel validation method for visual style errors by utilizing Otsu's method to segment CD datasets into subsets with varying degrees of visual style differences based on content and style similarity measurements of pre-trained backbone features. This enables quantitative assessment of the effectiveness of cross-visual style CD algorithms. We rigorously evaluate RCCD on the LEVIR-CD+ and S2Looking datasets, including their subsets with significant visual style differences. Our results demonstrate that RCCD substantially improves cross-visual style CD performance, achieving state-of-the-art (SOTA) results.

1. RCCD framework: A novel fully convolutional Siamese neural network training framework with an encoder, decoder, and deep supervisor module, enabling efficient and effective mitigation of cross-visual style errors without incurring extra computational costs during inference.
2. Representation consistency deep supervised learning: A method integrating target local spatial context and global content awareness, constructing a loss function that effectively suppresses image pairs with significant visual style differences, promoting robust CD.
3. Visual style errors validation: A novel method leveraging pre-trained networks to partition datasets based on visual style differences, enabling quantitative assessment of cross-visual style CD algorithms.
4. SOTA performance: RCCD achieves SOTA accuracy, substantially improving cross-visual style error suppression in challenging scenarios. The code is available on GitHub: https://github.com/wjj282439449/RCCD (accessed on 3 October 2024).

This paper proceeds as follows: Section 2 presents our methodology; Section 3 details our experimental setup and results; Section 4 discusses the implications and limitations; and Section 5 concludes the paper.

2. Methodology

We propose the Representation Consistency Change Detection (RCCD) framework (Figure 2), a cross-visual-style CD method tailored for bi-temporal images. Unlike conventional Siamese encoder–decoder networks, RCCD incorporates a novel representation consistency deep supervision approach during training. This approach applies additional supervision to multi-scale features at each encoding and decoding stage, additional assessing the similarity between bi-temporal pyramid features extracted by the backbone network, considering both local spatial and global stylistic characteristics. The encoder utilizes the pre-trained model for robust feature extraction. We introduce a unified basic feature compression unit (UDD) for both the decoder and detector, enabling efficient dimensionality reduction and feature fusion. Finally, the detector outputs a binary change mask aligned with the input image dimensions.

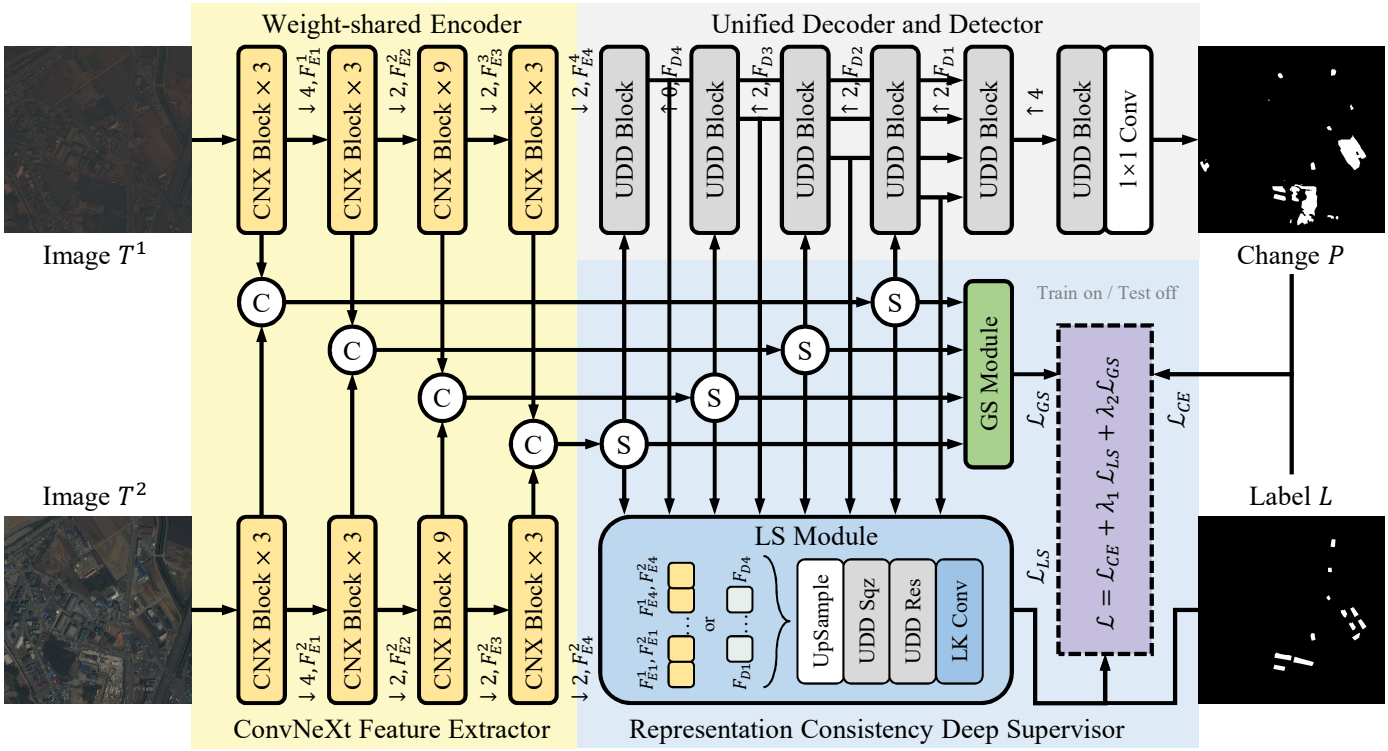

Figure 2. Illustration of the RCCD framework, including the representation consistency deep supervisor.

2.1. Overall Architecture

Given a pair of registered images $\{T^1, T^2\} \in \mathbb{R}^{C\times H\times W}$, RCCD predicts a change mask, denoted as *Mask*, using only the encoder, decoder, and detector components, excluding the supervisor module during inference. This process can be defined as follows:

$$Mask = \text{RCCD}(T^1, T^2) \tag{1}$$

The encoder generates a four-stage multi-scale pyramid of downsampled features, $\{F^t_{D1}, F^t_{D2}, F^t_{D3}, F^t_{D4}\}$, from each input image with a feature downsampling ratio of $\{4, 2, 2, 2\}$ between consecutive stages:

$$\{F^t_{E1}, F^t_{E2}, F^t_{E3}, F^t_{E4}\} = \text{Encoder}(T^t),\ t \in \{1, 2\} \tag{2}$$

The decoder upsamples input features via interpolation and a UDD module. Each level, except the lowest (F^1_{E4} and F^2_{E4}), concatenates its input with the previous level's UDD output before its own UDD processing, as follows:

$$F_{Dl} = \begin{cases} \text{Decoder}_{(l)}(F^1_{El}, F^2_{El}, F_{D(l+1)}) & l \in \{1, 2, 3\} \\ \text{Decoder}_{(l)}(F^1_{El}, F^2_{El}) & l = 4 \end{cases} \tag{3}$$

The detector utilizes two UDD modules as foundational building blocks, taking the outputs from each UDD level as input. It then employs a two-channel convolutional layer to generate the final CD prediction P and binary *Mask*, as follows:

$$Mask = Binary(P) = \text{Detector}(F_{D1}, F_{D2}, F_{D3}, F_{D4}) \tag{4}$$

During training, the supervisor employs three loss functions: Cross-Entropy Loss ($\mathcal{L}_{CE}$) to evaluate the similarity between the result (P) and the Label (L), and two L1 loss

functions to supervise local spatial (LS, Section 2.3) similarity ($\mathcal{L}_{LS}$) and global style (GS, Section 2.3.1) consistency ($\mathcal{L}_{GS}$), respectively:

$$\mathcal{L}_{CE} = -\frac{1}{HW}\sum_{n=1}^{HW}[L_n \log(P_n) + (1 - L_n)\log(1 - P_n)], \tag{5}$$

$$\mathcal{L}_{LS} = \frac{1}{8 \cdot C_l H_l W_l}\sum_{l=1}^{4}\sum_{n=1}^{C_l H_l W_l}(||LS(F_{El}^1, F_{El}^2)_n - L_n||_1 + ||LS(F_{Dl})_n - L_n||_1), \tag{6}$$

$$\mathcal{L}_{GS} = \frac{1}{4 \cdot C_l^1 C_l^2}\sum_{l=1}^{4}\sum_{n=1}^{C_l^1 C_l^2}(||GS(F_{El}^1)_n - GS(F_{El}^2)_n||_1), \tag{7}$$

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{LS} + \lambda_2 \mathcal{L}_{GS}, \tag{8}$$

where L_n represents each pixel between P and L. The El and Dl denote the l-th stage encoder and decoder. The LS and GS module are calculated as in Equations (18) and (20), and the LS module is capable of accepting two forms of input.

Therefore, the optimization process of RCCD for the parameters θ^* aims to achieve consistency with the ground truth L, as follows:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\text{RCCD}((T^1, T^2, \theta),\ L) \tag{9}$$

2.2. Siamese Encoder–Decoder Structure

CNX Encoder: Utilizing a hierarchical, pre-trained ConvNeXt (CNX) network [38], the encoder extracts multi-level features $\{F_{E1}, F_{E2}, F_{E3}, F_{E4}\}$ from input image pairs $\{T^1, T^2\} \in \mathbb{R}^{C \times H \times W}$, as defined in Equation (2). The resulting top-down encoding, within a Siamese structure, efficiently captures local features and streamlines the CD process.

In Figure 3a, each backbone feature level consists of a downsampling layer (with multipliers $n \in \{4, 2, 2, 2\}$) and multiple basic blocks; LN denotes Layer Normalization:

$$F'_{Dl} = \text{LN}(\text{Conv}^{(\text{n,n})})(F_{Dl}),\ l \in \{1, 2, 3, 4\} \tag{10}$$

where $F'_l \in \mathbb{R}^{C \times \frac{H}{n} \times \frac{W}{n}}$, Conv superscript "$^{(n,n)}$" indicates the kernel size and stride of the Convolution (Conv) operation, respectively. The CNX block computation is as follows:

$$\text{CNX}(F) = \text{Conv}(\text{GELU}(\text{Conv}(\text{LN}(\text{Conv}^{(7)}(F))))) \oplus F \tag{11}$$

$$F_{D(l+1)} = \text{CNX}_{(m)}(F'_{Dl}), m \in \{3, 3, 9, 3\} \tag{12}$$

where m is the per-stage CNX repetition count. Downsampling of F'_{Dl} followed by m CNX computations yields the input features for stage F_{l+1}. Superscript $^{(7)}$ denotes a 7×7 Conv, while the absence of a superscript indicates a 1×1 Conv. The two 1×1 Convs expand the channel dimension by a factor of 4 and then recover it.GELU [39] is an adaptive activation function. "$\oplus$" denotes element-wise addition.

Unified Decoder and Detector (UDD): The UDD module, illustrated in Figure 3b, combines compressed multi-level feature representations with decoder outputs, using skip connections to the encoder for improved gradient flow. The UDD comprises a UDD Squeeze unit (UDDSqz) and a UDD Residual unit (UDDRes), both composed of several Basic Units (BU): Conv $\rightarrow$ BN $\rightarrow$ GELU:

$$\text{UDD}(\dots) = \text{UDDRes}(\text{UDDSqz}(\dots)) \tag{13}$$

$$F'_{Dl} = \text{UDDSqz}(\dots) = \text{BU}_{(l)}(\dots) \tag{14}$$

$$F_{Dl} = \text{UDDRes}(\dots) = \text{BU}_{(l)}(\text{BU}_{(l)}(F'_{Dl})) \oplus F'_{Dl} \tag{15}$$

where Equation (14) utilizes a 1×1 Conv. Equation (15) employs 3×3, 1×1, respectively.

Decoder and Detector: The decoder collects outputs from the encoder and utilizes a UDD module at each layer. The multi-layer outputs then serve as input for the detector, which receives $F_{D1}, F_{D2}, F_{D3}, F_{D4}$ from the decoder. As depicted in Figure 2, the features are interpolated to the same shape and Equation (13) is applied twice. Finally, a 1×1 Conv to generates the change *Mask*:

$$Mask = \text{Softmax}(\text{Conv}(\text{UDD}(\text{UDD}(F_{D1}, F_{D2}, F_{D3}, F_{D4})))) \tag{16}$$

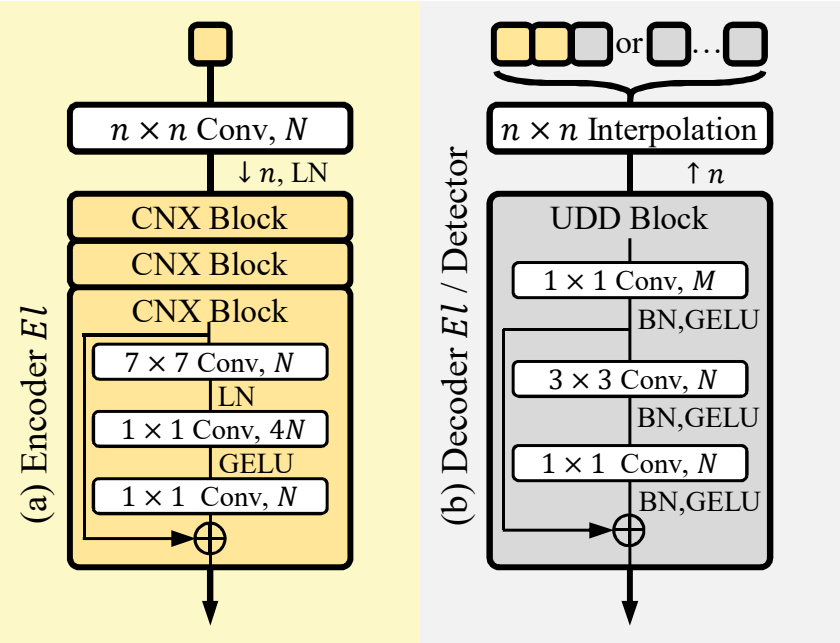

Figure 3. Details of CNX and UDD blocks in various stages of the encoder, decoder, and detector.

2.3. Representation Consistency Deep Supervised Learning

Unlike other Siamese CD methods, RCCD employs a representation consistency deep supervision training strategy. This framework comprises local spatial contexts, large-kernel supervised modules, and global content-aware style transfer modules, generating additional loss function constraint terms, $\mathcal{L}_{LS}$ and $\mathcal{L}_{GS}$, respectively. During training, features generated at each level, F^1_{El}, F^2_{El} and F_{Dl}, are subjected to deep supervision (detailed in Section 2.2), comprising the local spatial similarity large-kernel loss (LS loss, Equation (6)), and the Global content-aware Style transfer similarity loss (GS loss, Equation (7)). Both feature similarity losses are computed after Softmax, utilizing the L1 norm as the similarity metric.

2.3.1. Local Spatial Context Large-Kernel Supervised (LS) Module

The LS module assesses local spatial correlation using large-scale convolutional kernels to constrain the representational consistency of features within a defined local spatial extent. It comprises three key components: channel compression, residual module [40], and large-kernel linear discrimination. The channel compression module reduces feature map dimensionality, mitigating computational complexity. The residual module expands the receptive field and facilitates discriminative local spatial feature extraction:

$$\mathrm{UDD}(F_l) = \mathrm{UDDRes}_{(\mathrm{m})}(\mathrm{UDDSqz}(F_l)), F_l \in \{(F_{El}^1, F_{El}^2), F_{Dl}^D\} \tag{17}$$

The large-kernel linear discrimination module, with adjustable parameters, outputs a measure of local spatial similarity used by the loss function (Equation (6)) during the training, as follows:

$$\mathrm{LS}(F_l) = \mathrm{Softmax}(\mathrm{Conv}^{(\mathrm{k})}(\mathrm{UDD}(F_l))) \tag{18}$$

2.3.2. Global Content-Aware Style Transfer (GS) Module

The GS module implicitly performs style transfer through global content similarity assessment, constraining features from images with varying styles to possess representationally consistent global style descriptors.

Firstly, the GS module flattens spatial dimensions and then transposes the spatial and channel dimensions of each features.

$$F_{El}^t \in \mathbb{R}^{C_l \times H_l \times W_l} \rightarrow F_{El}^t \in \mathbb{R}^{C_l \times H_l W_l} \rightarrow F_{El}^{t'} \in \mathbb{R}^{H_l W_l \times C_l} \tag{19}$$

Next, the matrix multiplication between the original features and their transposed counterparts for global style descriptor (Gram matrix).

$$GS(F_{El}^t) = \frac{1}{C_l H_l W_l}(F_{El}^t \otimes F_{El}^{t'}) \in \mathbb{R}^{C_l \times C_l} \tag{20}$$

Finally, the loss function constrains (Equation (7)) the global style descriptors of features from different temporal phases, ensuring that content-aware, consistent features can be extracted even from images exhibiting diverse visual styles.

2.4. Visual Style Errors Validation

To demonstrate the effectiveness of our algorithm in mitigating visual style discrepancies, we propose a novel visual style error validation method (Figure 4), leveraging cosine similarity to measure both content and style similarity between pyramid features extracted using a pre-trained backbone network (without fine-tuning). We then employ Otsu's method [41] to segment these similarity values, yielding two subsets: the large visual style difference (Large-VS) subset (low content and style similarity) and the small visual style difference (Small-VS) subset (high content and style similarity). Improved algorithm performance on the Large-VS subset, compared to the Small-VS subset, demonstrates effective mitigation of visual style discrepancies and robust CD performance. The detailed procedure is as follows:

1. Extract multi-level features: utilize a pre-trained backbone network to extract pyramid features from each image pair:

$$\{F_{E1}^t, F_{E2}^t, F_{E3}^t, F_{E4}^t\} = \mathrm{Backbone}(T^t),\ t \in \{1, 2\} \tag{21}$$

2. Calculate mean content similarity ($\bar{C}$): average the cosine similarity between corresponding feature maps across all levels for both temporal phases:

$$\bar{C} = \mathrm{CosSim}(F_{El}^1, F_{El}^2) = \frac{1}{4}\sum_{l=1}^{4} \frac{F_{El}^1 \cdot F_{El}^2}{\|F_{El}^1\| \|F_{El}^2\|}, F_{El}^t \in \mathbb{R}^{C_l H_l W_l} \tag{22}$$

3. Calculate mean style similarity ($\bar{S}$): average the Gram matrix-based cosine similarity (Equation (20)) between corresponding feature maps across all levels:

$$\bar{S} = \frac{1}{4}\sum_{l=1}^{4}\frac{GS(F_{El}^1)\cdot GS(F_{El}^1)}{\|GS(F_{El}^1)\|\|GS(F_{El}^1)\|}, GS(F_{El})^t \in \mathbb{R}^{C_l C_l} \tag{23}$$

4. Apply Otsu's method: utilize Otsu's method to determine the optimal segmentation thresholds T_C and T_S for content and style similarity, respectively, based on the sets $\{\bar{C}_1, \bar{C}_2, \ldots, \bar{C}_n\}$ and $\{\bar{S}_1, \bar{S}_2, \ldots, \bar{S}_n\}$.
5. Construct visual style subsets, $\wedge$ denotes the intersection of sets:
Large-VS Subset (low content and low style similarity):

$$\bar{C} < T_C \wedge \bar{S} < T_S. \tag{24}$$

Small-VS Subset (high content and high style similarity):

$$\bar{C} \geq T_C \wedge \bar{S} \geq T_S. \tag{25}$$

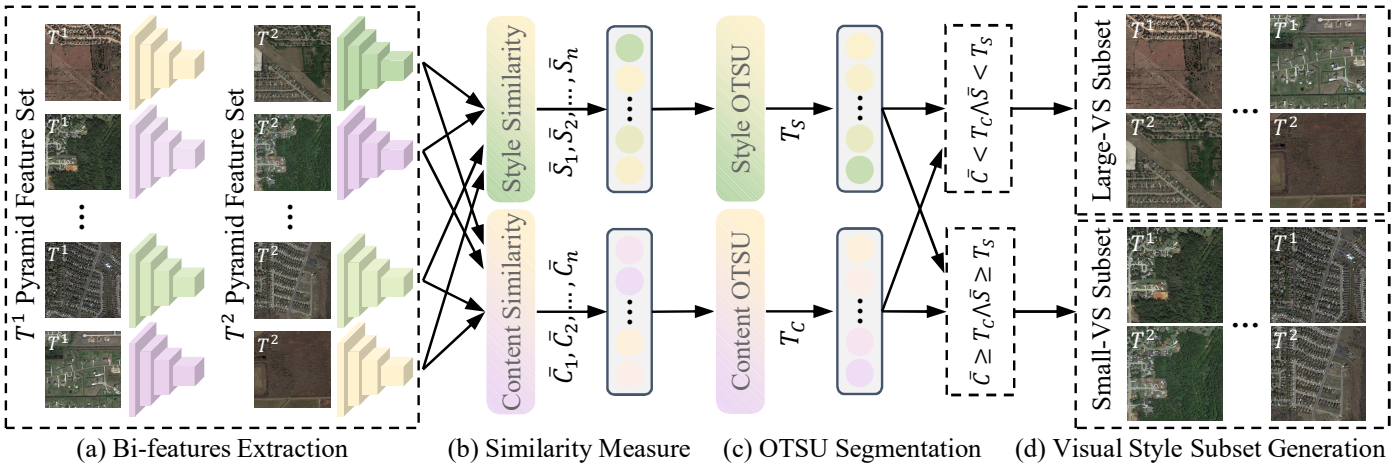

Figure 4. Workflow for generating Large- and Small-VS subsets based on visual style differences.

3. Experiments and Analysis

3.1. Experimental Settings

3.1.1. Datasets

Following Section 2.4, both test sets were partitioned into Large-VS and Small-VS subsets based on content and style similarity. The LEVIR-CD+ (348 pairs): 56 and 226 pairs; the S2Looking (1000 pairs): 157 and 559 pairs, respectively. The detailed proportions of image pairs with similar styles and content in the two datasets are presented in Table 1.

Table 1. Distribution of visual differences in the S2Looking and LEVIR-CD+ datasets

Category	S2Looking (Pairs)	LEVIR-CD+ (Pairs)
$\bar{C} \geq T_C \wedge \bar{S} \geq T_S$	559	226
$\bar{C} \geq T_C \wedge \bar{S} < T_S$	60	17
$\bar{C} < T_C \wedge \bar{S} \geq T_S$	224	49
$\bar{C} < T_C \wedge \bar{S} < T_S$	157	56
Total	1000	348

LEVIR-CD+: This CD datas contains over 985 pairs of bi-temporal building images (0.5 m/pixel) from 20 different regions in USA, spanning five years. Each image has dimen-

sions of 1024 × 1024 pixels. The dataset is split into training (65%), validation (25% of training), and test (35%) sets. It includes significant lighting, shadow, and phenology variances [9].

S2Looking: This dataset comprises 5000 pairs of satellite images (0.5–0.8 m/pixel) captured over three years, each with dimensions of 1024 × 1024 pixels. The dataset (3500 training, 500 validation, and 1000 test images) incorporates significant weather and radiation variability [10].

3.1.2. Evaluation Metrics

This evaluation includes comparisons with recent bi-temporal CD methods and ablation studies. We evaluate the CD quality using four metrics: Precision (Pre.), Recall (Rec.), F1-measure (F1), and Intersection over Union (IoU), computed from True Positives (TPs), False Positives (FPs), and False Negatives (FNs). Higher scores across all metrics indicate superior performance.

$$\text{Pre.} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{26}$$

$$\text{Rec.} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{27}$$

$$\text{F1-measure (F1)} = 2 \times \frac{\text{Pre.} \times \text{Rec.}}{\text{Pre.} + \text{Rec.}} \tag{28}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{29}$$

3.1.3. Implementation Details

We trained and tested RCCD and mainstream methods on LEVIR-CD+, S2Looking, and VS subsets using a single 4090 GPU. Dataset images were cropped to 256 × 256 pixel size, maintaining the original training and test set divisions. Models were selected from a valid set.The RCCD backbone was initialized with pre-trained parameters. Model optimization employed the AdamW [42] optimizer with a learning rate of 5×10^{-5}, a batch size of 32. Data augmentation included random photometric and geometric distortions. Our RCCD was implemented in PyTorch 2.1, while others used their official implementations.

3.2. Comparisons on Mainstream Methods

We compared our RCCD model with ten other CD methods on the S2Looking and LEVIR-CD+ datasets, including UNet [13], FC-Diff [11], FC-Conc [11], IFNet [31], SNUNet [14], L-UNet [15], BIT [16], ICIF-Net [17], P2V [18], CDNeXt [43], and the RCCD-B baseline model (lacking GS and LS modules). We evaluated the robustness of these methods against visual style errors using Large-VS and Small-VS subsets. The results of FC-Diff, L-UNet, BIT, ICIF-Net, P2V, CDNeXt, and our RCCD were visualized on both datasets.

3.2.1. Results on S2Looking Dataset

Figure 5 visually compares CD results on images with varying style differences. Unlike other methods, RCCD demonstrates robustness to variations in radiation, weather, phenology, and refurbishment, exhibiting broader applicability and effectively mitigating visual style errors. Under consistent radiation (e.g., rows 1–3), RCCD shows fewer missed detections and more accurate change boundaries. In the presence of irrelevant changes like clouds and vegetation phenology (e.g., rows 4–8), RCCD generates fewer false positives (e.g., rows 5, 7, and 8). In the fifth set of images, the changes between cultivated land and bare land are not of interest in this dataset; however, the significant semantic and color differences have led most methods to misclassify these changes. In the ninth set of images, vegetation is detected as change due to varied visual characteristics across different images caused by phenological variations, whereas buildings and bare land, presenting similar

visual appearances due to dust, result in omissions during detection. For uninteresting changes like road construction, our method demonstrates greater stability. These results confirm RCCD accurately extracts changes in interest under complex background interference, effectively suppressing visual style variations compared to mainstream methods.

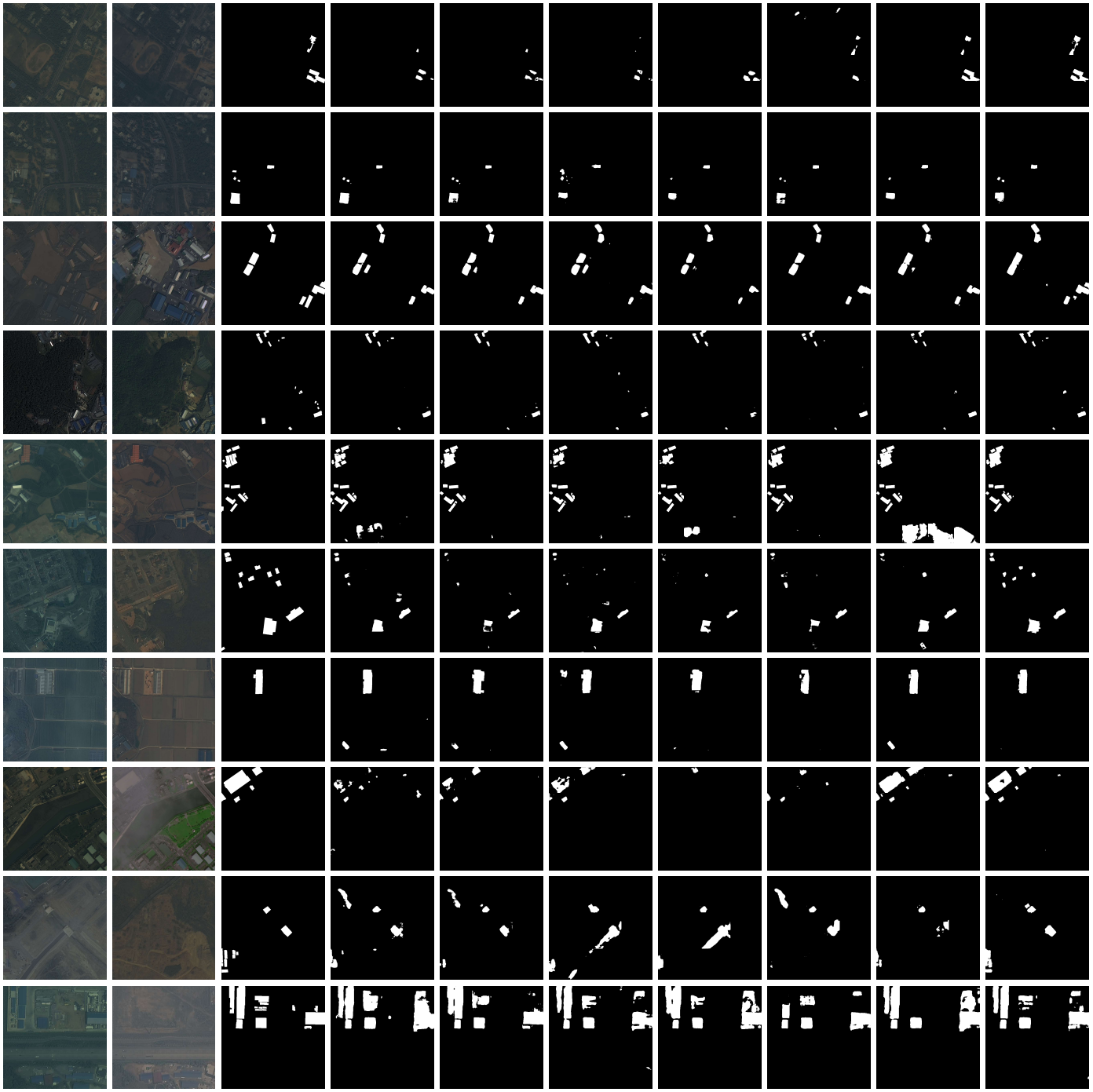

Figure 5. Visualizations of different methods on S2Looking.

Table 2 compares the performance of different methods on the S2Looking test set. Our RCCD method achieved the highest F1 and IoU scores across the full dataset and showed substantial improvement on the Large-VS subset. Compared to CDNeXt, RCCD achieved a 0.89% and 1.95% increase in F1-score on the full dataset and the Large-VS subset, respectively. Compared to the third-best method, these increases were 6.15% and 7.47%,

representing relative improvements of 119.1% and 21.5% on the Large-VS subset. Compared to the RCCD baseline, our method further improved F1-score by 1.81% (full dataset) and 2.57% (Large-VS subset), a relative increase of 41.9%. Consistent accuracy improvement was also observed on the Small-VS subset, indicating the LS and GS components do not negatively impact performance in other scenarios. This suggests our method effectively mitigates visual style discrepancies. Interestingly, feature differencing methods performed better on the Small-VS subset, while self- and cross-attention methods excelled on the Large-VS subset.

Table 2. Quantitative performance of RCCD and competing methods on the S2Looking. Color convention: best, 2nd-best, and **3rd-best**.

S2Looking	**Full Dataset**				**Large-VS Subset**				**Small-VS Subset**			
Method	**Pre.**	**Rec.**	**F1**	**IoU**	**Pre.**	**Rec.**	**F1**	**IoU**	**Pre.**	**Rec.**	**F1**	**IoU**
UNet	62.55	42.47	50.59	33.86	64.51	43.44	51.92	35.06	65.13	41.02	50.34	33.63
FC-Diff	61.65	**61.25**	61.45	44.35	58.01	**62.77**	60.29	43.15	64.49	**60.33**	62.34	45.28
FC-Conc	68.67	53.88	60.38	43.25	67.36	55.28	60.73	43.60	70.63	52.17	60.01	42.87
IFNet	66.18	48.56	56.02	38.90	61.18	50.77	55.49	38.40	68.65	45.51	54.74	37.68
SNUNet	69.54	53.82	60.68	43.55	67.76	55.75	61.17	44.06	69.99	52.92	60.27	43.13
L-UNet	**72.46**	53.26	61.39	44.29	70.51	53.75	61.00	43.88	**73.33**	51.95	60.82	43.70
BIT	67.74	55.95	61.28	44.18	67.05	56.44	61.29	44.18	69.31	54.63	61.10	43.99
ICIF-Net	**72.68**	50.77	59.78	42.63	68.94	52.97	59.91	42.77	**75.43**	47.58	58.36	41.20
P2V	68.74	51.61	58.96	41.80	67.04	52.75	59.04	41.89	70.46	50.42	58.77	41.62
CDNeXt-B	70.89	60.29	65.16	48.33	69.45	62.16	65.61	48.82	**73.81**	58.53	**65.29**	**48.47**
CDNeXt	70.78	**63.08**	**66.71**	**50.05**	**72.15**	**62.21**	**66.81**	**50.16**	70.08	**62.36**	**66.00**	**49.25**
RCCD-B	**72.96**	59.91	**65.79**	**49.02**	**74.11**	59.80	**66.19**	**49.47**	71.97	58.87	64.76	47.89
RCCD	71.10	**64.43**	**67.60**	**51.06**	**71.83**	**65.94**	**68.76**	**52.40**	71.36	**63.77**	**67.35**	**50.77**

3.2.2. Results on LEVIR-CD+ Dataset

Figure 6 visually compares CD results on images with significant variations in illumination, shadows, and vegetation phenology. RCCD demonstrates robustness to complex backgrounds, achieving higher detection rates (Small-VS) and lower false positive rates (Large-VS). Under consistent radiation (e.g., rows 1 and 2), other methods exhibit missed detections, as feature measurements at the decoding stage cannot effectively distinguish changed buildings from the ground. In scenarios with significant radiation differences (e.g., rows 3–6), other methods often misidentify rooftop illumination and building shadow variations as changes, while RCCD avoids these errors. In the third set of images, the strong contrasts and shadows in the later image due to lighting conditions cause most methods to interpret these dramatic visual differences as changes. Vegetation changes also pose challenges for other methods (e.g., rows 7–10), as dense foliage can obscure buildings, hindering accurate assessment. In the eighth set of images, occlusions from vegetation growth and scene style variations resulting from differences in lighting and shadows lead to a large number of false alarms. RCCD's large-scale convolutions, with their ability to induce bias in local spatial contexts, effectively determine the identity consistency between buildings and their surroundings.

Table 3 presents the performance of different methods on the LEVIR-CD+ test set, which exhibits significant variations in illumination, shadows, and phenology. Our RCCD achieved the highest accuracy across all evaluations, with substantial improvement on the Large-VS subset. Compared to CDNeXt, RCCD achieved a 0.54% higher F1-score (3.52% higher than IFNet). On the Large-VS subset, RCCD's F1-score improved by 1.12% (107.4% relative increase), indicating its effectiveness in mitigating interference from radiation and phenological changes. Compared to RCCD-B, our method achieved an additional

56.9% improvement on this subset. Comparable accuracy of RCCD, RCCD-B, and CDNeXt on the Small-VS subset suggests the overall improvement stems from the representation consistency learning, which mitigates visual style discrepancies. This highlights the crucial role of addressing such discrepancies for improved CD.

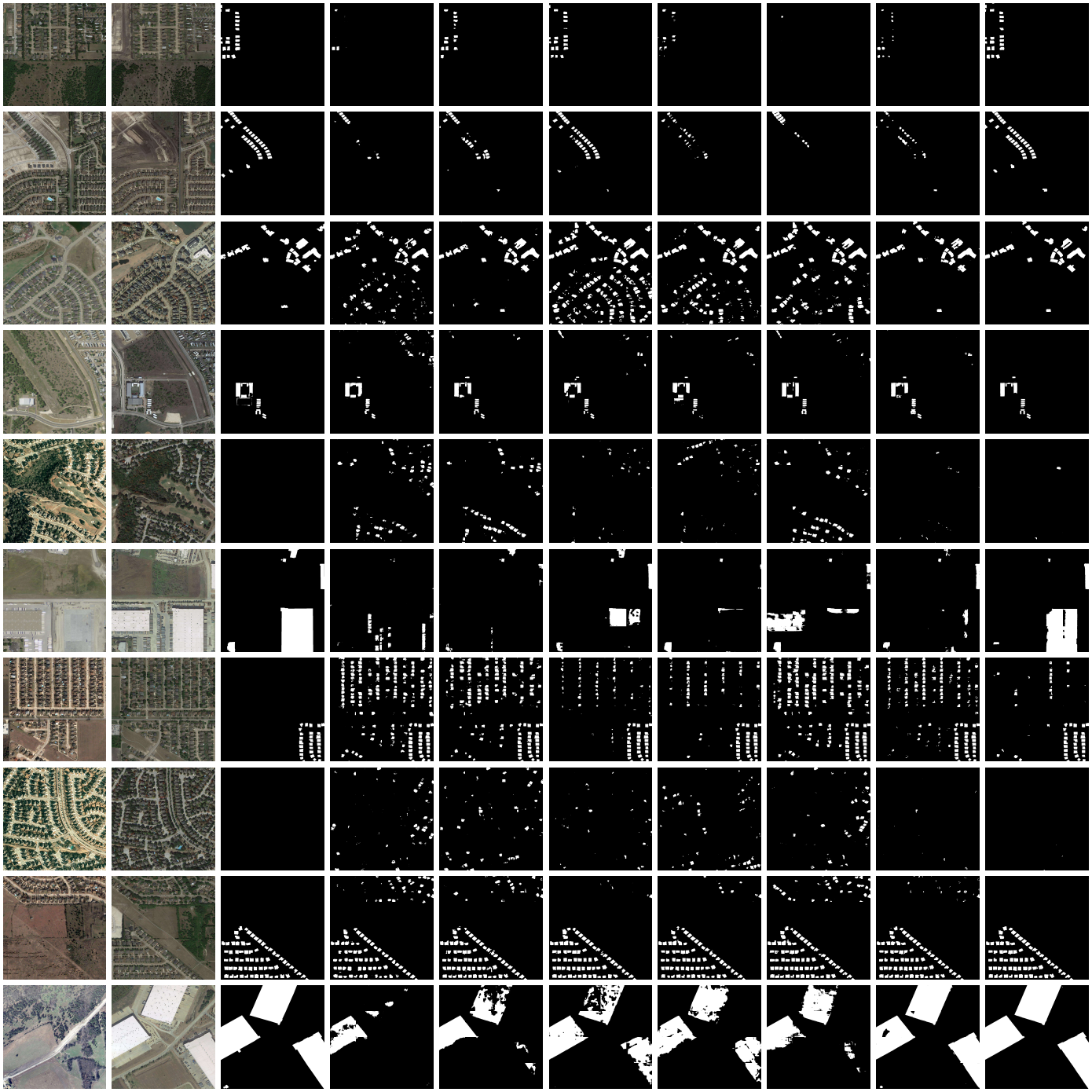

Figure 6. Visualizations of different methods on LEVIR-CD+.

Table 3. Quantitative performance of RCCD and competing methods on the LEVIR-CD+. Color convention: best, 2nd-best, and **3rd-best**.

LEVIR-CD+	Full Dataset				Large-VS Subset				Small-VS Subset			
Method	**Pre.**	**Rec.**	**F1**	**IoU**	**Pre.**	**Rec.**	**F1**	**IoU**	**Pre.**	**Rec.**	**F1**	**IoU**
UNet	73.23	75.25	74.22	59.01	81.43	73.20	77.09	62.73	66.41	72.38	69.27	52.98
FC-Diff	78.20	73.54	75.80	61.03	82.58	71.19	76.46	61.89	72.17	71.82	72.00	56.25
FC-Conc	71.05	77.66	74.21	58.99	80.96	75.19	77.97	63.89	60.14	77.04	67.55	51.00
IFNet	85.83	82.56	84.16	72.66	90.85	85.09	87.87	78.37	81.28	77.55	79.37	65.80
SNUNet	85.32	80.18	82.67	70.46	88.35	81.41	84.74	73.52	83.35	76.17	79.60	66.11
L-UNet	83.56	79.55	81.51	68.79	86.92	79.11	82.83	70.69	80.34	77.10	78.69	64.86
BIT	84.48	83.22	83.84	72.18	89.50	85.02	87.20	77.30	80.19	78.96	79.57	66.08
ICIF-Net	85.50	81.25	83.32	71.40	90.55	82.51	86.35	75.97	79.88	77.60	78.72	64.91
P2V	78.00	80.94	79.45	65.90	84.34	81.91	83.11	71.10	70.80	77.24	73.88	58.58
CDNeXt-B	86.64	83.99	85.29	74.35	89.32	85.57	87.40	77.63	84.95	80.64	82.74	70.56
CDNeXt	**89.68**	**84.73**	**87.14**	**77.21**	**92.73**	**86.36**	**89.43**	**80.88**	**87.01**	**81.47**	**84.15**	**72.63**
RCCD-B	**88.52**	**84.47**	**86.45**	**76.13**	**91.59**	**85.84**	**88.62**	**79.57**	**85.63**	**82.65**	**84.11**	**72.58**
RCCD	**89.11**	**86.30**	**87.68**	**78.06**	**91.92**	**89.21**	**90.55**	**82.73**	**86.02**	**82.95**	**84.46**	**73.10**

3.3. Ablation Study

The S2Looking and LEVIR-CD+ valid sets validated our proposed algorithm's effectiveness. These studies investigated various design choices: the number of UDDRes repetitions (m) for the LS module, the large-kernel Conv size (k), loss hyperparameters (λ_1, λ_2), and the representation consistency integration within encoder–decoder stages. These experiments ensured a rational and well-justified algorithm structure.

3.3.1. LS Module: UDDRes Depth Analysis

This ablation study (Table 4) investigated the impact of varying the number of UDDRes repetitions (m) within the LS module, while keeping all other hyperparameters fixed. We aimed to determine if the LS module benefits from similarity measurements at higher semantic levels. Results indicate that directly measuring the similarity between encoder and decoder features is optimal ($m = 1$), eliminating the need for extracting high-dimensional semantic representations. This is because the module's primary function is to capture local spatial feature similarity between target and background, effectively achieved through large convolutional kernels. While increasing UDDRes depth might lead to further performance gains, it would incur excessive computational costs during training, making it impractical for our current setup.

Table 4. Impact of UDDRes repetition (m) on LS module performance. Color convention: best, 2nd-best, and **3rd-best**.

m-UDDRes	LEVIR-CD+		S2Looking	
	F1	**IoU**	**F1**	**IoU**
1	**67.28**	**50.69**	**87.36**	**77.55**
2	66.65	49.98	86.79	76.66
3	**66.81**	**50.16**	**87.11**	**77.17**
4	66.70	50.04	86.50	76.21
5	**66.87**	**50.23**	**87.16**	**77.25**

3.3.2. LS Module: Large-Kernel Size Effectiveness

The large kernel size (k) in our LS module plays a pivotal role in capturing local spatial context for building target detection, serving as a core parameter for measuring

the similarity of local spatial contexts (Figure 7a,b). As shown in Table 5, our experiments indicate that the optimal kernel sizes vary by dataset: $k = 15$ yields the highest F1-score of 67.13% for S2Looking, while $k = 11$ achieves a peak F1-score of 87.50% for LEVIR-CD+. These optimal values align closely with the datasets' ground resolutions—S2Looking varies between 0.5 and 0.8 m/pixel (averaging around 0.65 m/pixel), LEVIR-CD+ has a consistent resolution of 0.5 m/pixel. The ratio of ground resolutions (0.65/0.5 = 1.3) closely matches the ratio of optimal kernel sizes ($15/11 \approx 1.36$), suggesting that the ideal kernel size for measuring building target similarity in local spatial contexts is approximately 23 times the ground resolution (in meters). Note that if the targets of interest include other objects (e.g., vehicles, roads, or trees), the optimal kernel size may need re-evaluation. While larger kernels up to these optimal values improve performance, they also incur a modest increase in computational cost—parameters increase from 49.61 M to 50.58 M and FLOPs from 17.46 G to 17.96 G when k increases from 1 to 15. Beyond the optimal sizes, further increases in k lead to diminished performance improvements relative to the additional computational load.

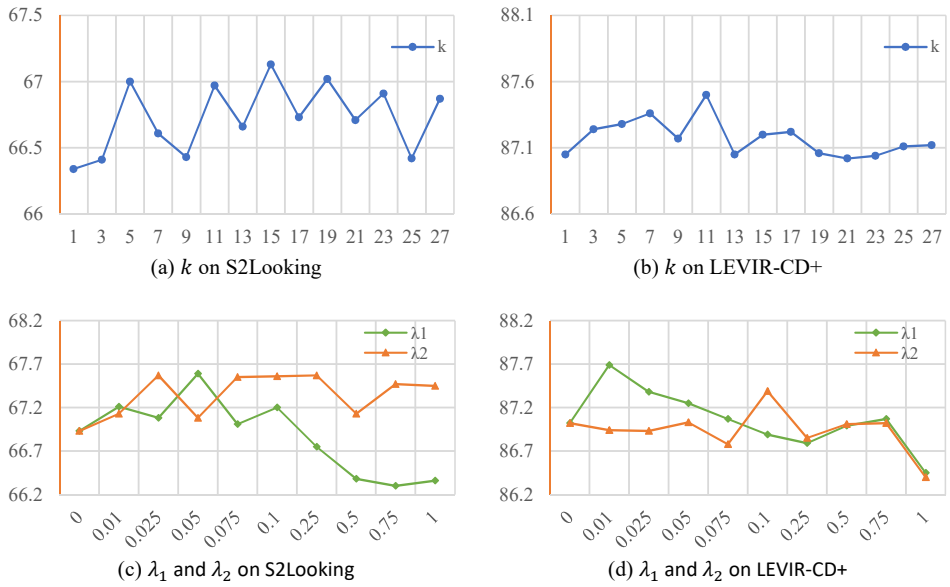

Figure 7. Impact of kernel size (k) and loss weights (LS λ_1 and GS λ_2) on F1.

Table 5. Impact of LS module kernel size (k) on S2Looking and LEVIR-CD+, encompassing Parameters (Params) count, FLoating-Point OPerations (FLOPs), and F1-score comparisons in RCCD training. Color convention: best, 2nd-best, and **3rd-best**.

k	S2Looking F1	LEVIR-CD+ F1	Params (M)	FLOPs (G)
1	66.34	87.05	49.61	17.46
3	66.41	87.24	49.65	17.48
5	**67.00**	**87.28**	49.72	17.52
7	66.61	87.36	49.82	17.57
9	66.43	87.17	49.96	17.64
11	66.97	87.50	50.13	17.73
13	66.66	87.05	50.34	17.84
15	67.13	87.20	50.58	17.96
17	66.73	87.22	50.86	18.10
19	67.02	87.06	51.17	18.26
21	66.71	87.02	51.52	18.44
23	66.91	87.04	51.90	18.63
25	66.42	87.11	52.31	18.84
27	66.87	87.12	52.76	19.07

3.3.3. Representation Consistency: Loss Weighting

Figure 7c,d analyze the effectiveness of loss function hyperparameters λ_1 and λ_2 in Equation (8). Optimal weights for the LS and GS modules need to be adjusted for different datasets. We sampled values for both weights (0 to 1), keeping other parameters fixed at 0 and adjusting either λ_1 or λ_2 individually. Results show incorporating the LS or GS module within a certain range improves accuracy. On S2Looking (more pronounced global style differences), the GS module generally outperforms the LS module. Conversely, on LEVIR-CD+ (greater local content variations), the LS module performs better. Since both types of variations are typically present, fine-tuning the loss function weighting is necessary. The principle is to select the minimum weights that still yield performance improvements, avoiding negative effects from excessively large weights.

Figure 7c,d and Table 6 reveal three key findings about loss weight selection: (1) Performance Stability: Both modules tolerate moderate weight variations while maintaining >95% of maximum gains. For S2Looking, GS achieves peak F1 (67.57%) at $\lambda_2 = 0.025 - 0.25$ (10× range), while LEVIR-CD+ sustains >87% F1 with $\lambda_1 = 0.01 - 0.075$ (7.5× range). (2) Dataset-Specific Optimization: S2Looking benefits more from GS (max $\triangle$F1 = +0.64% vs. baseline), as its wider viewpoint variations create stronger spectral discrepancies than structural changes. LEVIR-CD+ prefers LS (max $\triangle$F1 = +0.67%) due to its nadir perspective and 0.5 m resolution emphasizing structural details. (3) The selection principle of optimal weights follow: $\lambda_1 = 0.05/\lambda_2 = 0.025$ for S2Looking and $\lambda_1 = 0.01/\lambda_2 = 0.1$ for LEVIR-CD+.

Table 6. Impact of the loss weighting parameter for LS (λ_1) and GS (λ_2) modules on S2Looking and LEVIR-CD+ performance. Color convention: best, 2nd-best, and **3rd-best**.

Value	S2Looking F1		LEVIR-CD+ F1	
	λ_1	λ_2	λ_1	λ_2
0	66.93	66.93	87.02	87.00
0.01	**67.21**	67.13	**87.69**	86.94
0.025	67.08	**67.57**	**87.38**	86.93
0.05	**67.59**	67.08	**87.25**	**87.03**
0.075	67.01	**67.55**	87.07	86.78
0.1	**67.2**	**67.56**	86.89	**87.39**
0.25	66.75	**67.57**	86.79	86.85
0.5	66.38	67.13	86.99	87.01
0.75	66.30	67.47	87.07	**87.02**
1	66.36	67.45	86.45	86.40

3.3.4. Representation Consistency: Encoder and Decoder Stages

Analyzing Representation Consistency (RC) learning in encoding and decoding stages (Table 7) with fixed parameters allows exploration of reasonable framework structures. Firstly, the RC module acting independently in either stage shows varying responses across datasets, but generally outperforms the baseline (no RC). Secondly, for S2Looking (significant visual style differences), representation consistency between bi-temporal features is more crucial in encoding than decoding. For LEVIR-CD+ (complex background details), representation control in decoding yields better performance. In conclusion, full-stage representation consistency is optimal for addressing visual style issues, effectively handling both global style differences and local content discrepancies.

Table 7. Effectiveness of representation consistency learning across encoder–decoder stages. ✓: LS/GS module activated, ✗: LS/GS module deactivated. Color convention: best, 2nd-best, and **3rd-best**.

Encoder Stage				Decoder Stage				S2Looking		LEVIR-CD+	
E1	**E2**	**E3**	**E4**	**D4**	**D3**	**D2**	**D1**	**F1**	**IoU**	**F1**	**IoU**
✓	✗	✗	✗	✗	✗	✗	✗	66.47	49.77	87.17	77.17
✗	✓	✗	✗	✗	✗	✗	✗	66.74	50.09	87.07	77.10
✗	✗	✓	✗	✗	✗	✗	✗	**67.09**	**50.47**	86.86	76.77
✗	✗	✗	✓	✗	✗	✗	✗	66.51	49.82	87.17	77.25
✗	✗	✗	✗	✓	✗	✗	✗	66.67	50.01	87.04	77.05
✗	✗	✗	✗	✗	✓	✗	✗	66.31	49.60	87.00	76.99
✗	✗	✗	✗	✗	✗	✓	✗	66.36	49.65	86.97	76.95
✗	✗	✗	✗	✗	✗	✗	✓	66.47	49.78	86.54	76.27
✗	✗	✗	✗	✗	✗	✗	✗	65.79	49.02	86.45	76.13
✗	✗	✗	✗	✓	✓	✓	✓	**66.76**	**50.11**	**87.27**	**77.41**
✓	✓	✓	✓	✗	✗	✗	✗	66.29	49.58	**87.20**	**77.30**
✓	✓	✓	✓	✓	✓	✓	✓	**67.28**	**50.69**	**87.42**	**77.66**

3.4. Visual Style Subset Visualization

Figure 8 visualizes our visual style error validation method across content and style difference dimensions, using uniformly sampled examples. Based on the magnitude of these differences (large or small), we categorize them into four scenarios: small content/small style, small content/large style, large content/small style, and large content/large style. Visual comparison reveals that large content differences generally exhibit localized semantic variations (e.g., vegetation transforming into bare land, changes in building surroundings, uninteresting changes like parking lots). Conversely, large style differences tend to exhibit image-wide inconsistencies in radiation, weather, and phenology. Consequently, subsets exhibiting both large content and large style differences can be effectively identified. This method effectively partitions a CD dataset into two subsets for validating visual style errors, providing a quantifiable representation of this problem.

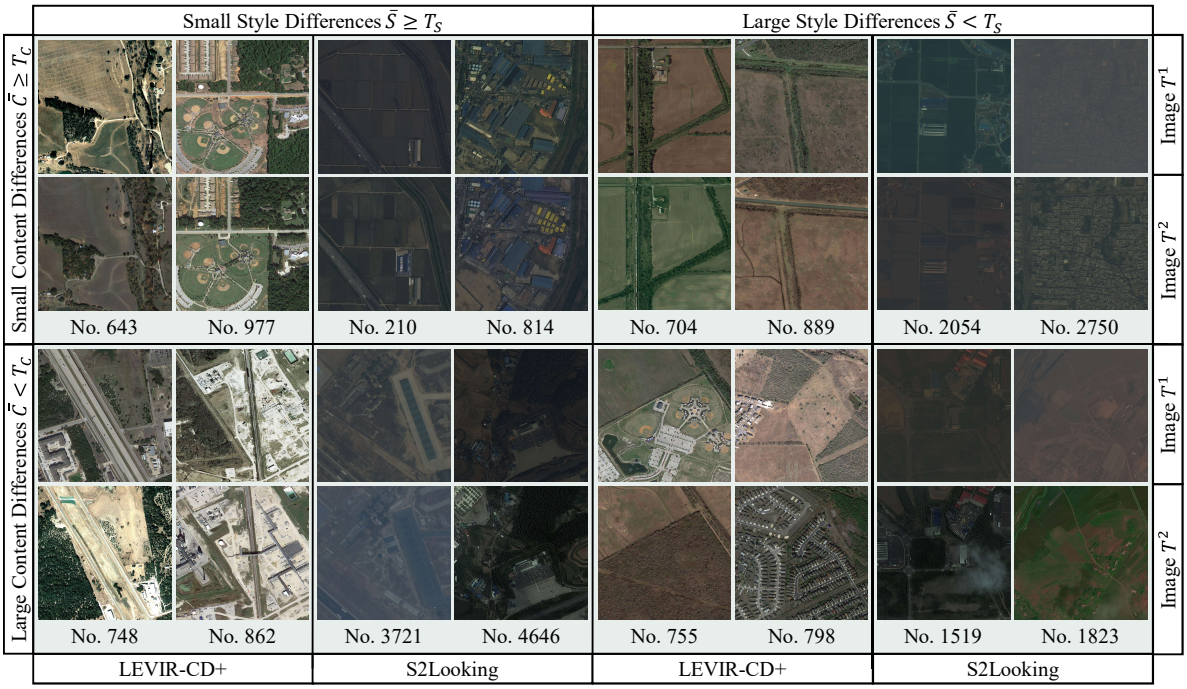

Figure 8. Visual style error validation via content/style differences.

3.5. Feature Interpretability of Representation Consistency

As illustrated in Figure 9, we employ a feature visualization technique—Gradient weighted Class Activation Mapping (GradCAM++) [44]—to perform an interpretability analysis of the features during the change detection process in RCCD. We compare the feature visualizations under three conditions: with the representation consistency learning module (w/LS and GS), with only the GS module (w/GS), and without any representation consistency learning modules (w/o LS and GS). The model with the representation consistency deep supervision module demonstrates stable feature responses throughout both the encoding and decoding stages—the heatmaps exhibit more uniform intensities and do not fluctuate dramatically despite variations in the visual styles of the input images. A similar global stabilization effect can be observed in the second row: with the addition of the GS module, the perceived activation distribution in the heatmap becomes even more uniform. In the first row, we observe that, compared to the second row (which lacks the LS module), there is a more pronounced suppression of pseudo-changes in the target objects and adjacent background during the decoding process. This underscores the necessity of local spatial context measurement. In contrast, the third row—corresponding to the configuration without any representation consistency learning module—exhibits the highest number of false alarms and is more susceptible to interference from vegetation and shadows, resulting in aberrant heatmap activations. Overall, the feature visualizations obtained during the RCCD prediction process confirm that the LS module effectively measures the similarity between the target and its local spatial context to suppress false alarms, while the GS module plays a crucial role in globally stabilizing images with varied visual styles, thereby ensuring the construction of comparable features.

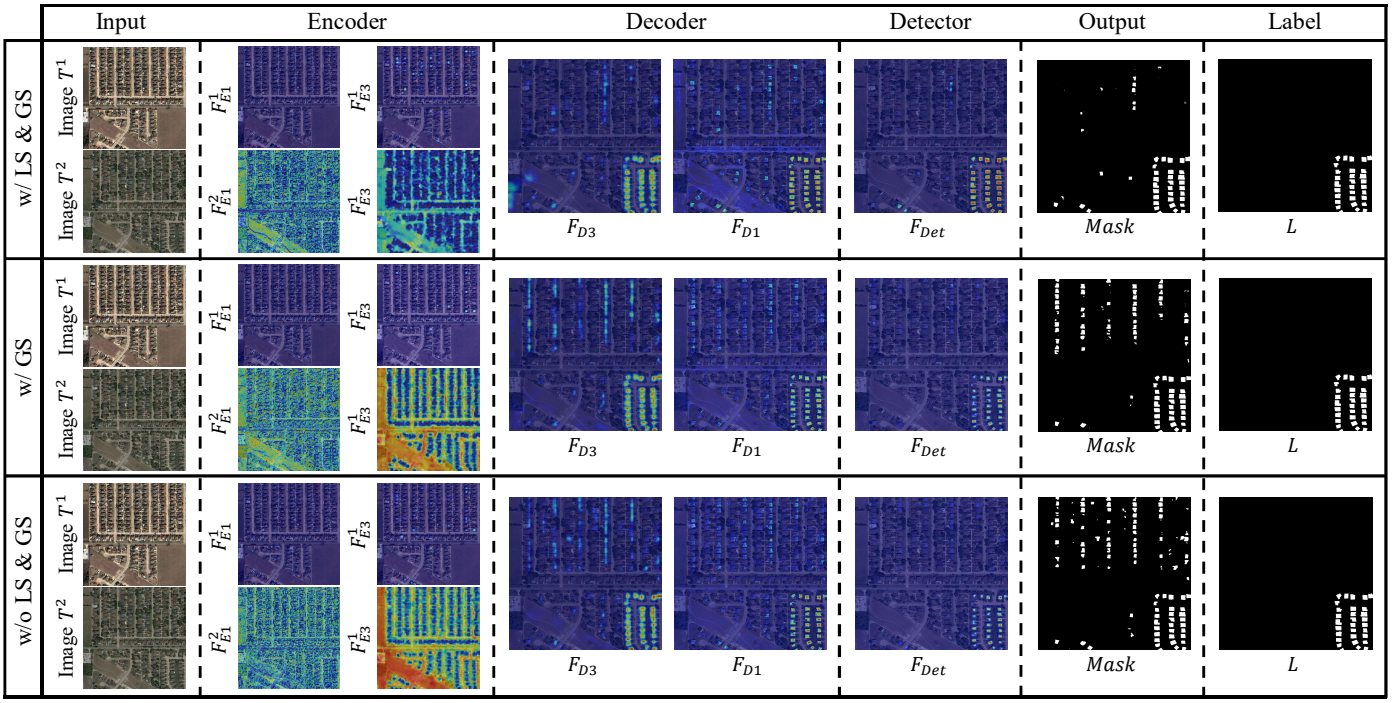

Figure 9. Comparison of feature visualizations for the representational consistency deep supervision method. "w/" indicates that the module is included, while "w/o" denotes its exclusion. The symbol definitions correspond to the formulas presented in Section 2.

3.6. Computational Complexity Analysis

The proposed RCCD framework is designed not only to achieve state-of-the-art (SOTA) accuracy in CD but also to maintain a favorable efficiency-accuracy trade-off, as shown

in Table 8. Specifically, the design of RCCD incorporates a representation consistency deep supervision strategy during training to enforce robust bi-temporal feature constraints. However, once trained, the representation consistency supervisor (comprising the local spatial and global style modules) is not required during inference. This design choice directly results in a substantial reduction in computational burden during the testing phase. Concretely, when the supervisor module is removed, experimental evaluations show that the effective parameter count decreases by approximately 24.5% and the FLOPs are reduced by 14.3%. This ensures that RCCD remains highly efficient in practical applications without sacrificing its strong predictive performance. Moreover, RCCD achieves the highest detection accuracy under these optimized conditions. On the S2Looking and LEVIR-CD+, RCCD outperforms the CDNeXt by improving 0.89% and 0.54% F1, respectively. In addition, compared with other methods such as UNet, FC-Diff, IFNet, SNUNet, and L-UNet, RCCD not only maintains competitive (or even superior) segmentation performance but does so with a significantly lower computational cost. For example, although IFNet exhibits strong detection performance, its large model size (35.73M parameters) and high FLOPs (82.26G) undermine its practical applicability. In contrast, RCCD achieves a well-balanced trade-off by leveraging the shared encoder–decoder structure and excluding additional inference overhead from the representation consistency modules.

Table 8. Computational complexity analysis across different methodologies, encompassing Parameters (Params) count, FLoating-Point OPerations (FLOPs), and F1-score comparisons on S2Looking and LEVIR-CD+ datasets. Lower values of Params and FLOPs indicate superior efficiency. RCCD-training denotes computational complexity during inference in the training phase, while RCCD-valid represents RCCD complexity during validation procedures. Color convention: best, 2nd-best, and **3rd-best**.

Method	S2Looking F1	LEVIR-CD+ F1	Params (M)	FLOPs (G)
UNet	50.59	74.22	1.35	3.58
FC-Diff	**61.45**	75.80	1.35	4.73
FC-Conc	60.38	74.21	1.54	5.33
IFNet	56.02	**84.16**	35.73	82.26
SNUNet	60.68	82.67	10.20	44.38
L-UNet	61.39	81.51	8.45	17.33
BIT	61.28	83.84	3.04	8.75
ICIF-Net	59.78	83.32	23.84	25.41
P2V	58.96	79.45	5.42	32.96
CDNeXt	**66.71**	**87.14**	39.42	15.76
RCCD-training	-	-	50.13	17.72
RCCD-valid	**67.60**	**87.68**	37.84	15.18

4. Discussion

The proposed RCCD framework demonstrates significant advancements in handling cross-visual style challenges for remote sensing CD. Our experimental results reveal three critical insights: First, enforcing representation consistency at both encoding and decoding stages proves more effective than single-stage constraints (Table 7), particularly for datasets combining global style variations and local content discrepancies. This hierarchical consistency learning aligns with recent findings in domain adaptation, suggesting that multi-level feature alignment better captures cross-domain relationships. Second, the dataset-specific optimal kernel sizes for LS modules (15 for S2Looking vs. 11 for LEVIR-CD+) correlate with spatial resolution differences (0.65 m vs. 0.5 m), establishing a practical guideline for kernel size selection as 23× ground resolution. Third, our visual style validation method

provides a novel quantitative framework for CD algorithm evaluation, addressing a critical gap in current benchmarking practices.

The effectiveness of RCCD stems from its dual mechanism: (1) The GS module acts as a global style normalizer through Gram matrix alignment, reducing domain shifts caused by seasonal/illumination variations. (2) The LS module functions as a local context stabilizer, mitigating false alarms from transient environmental changes through large receptive field analysis. This dual approach outperforms existing single-strategy solutions like metric learning or attention mechanisms, particularly in complex scenarios combining both global and local variations. However, two limitations warrant discussion: (1) The current LS module focuses on rectangular context regions, potentially missing irregular-shaped contexts. Future work could explore deformable convolutions for adaptive context shaping. (2) While our style validation method effectively identifies challenging samples, it relies on pre-trained features that may not optimally represent CD-specific characteristics. Developing task-oriented style metrics could further enhance validation accuracy.

Practically, RCCD's robustness enables reliable CD in applications requiring cross-seasonal analysis, such as post-disaster assessment and illegal construction monitoring. The framework's modular design permits integration with various backbone architectures, suggesting broad applicability beyond building CD. Future research directions include extending RCCD to multi-temporal CD scenarios and investigating self-supervised adaptation for unseen style variations.

5. Conclusions

This paper introduced Representation Consistency Change Detection (RCCD), a transformative framework that fundamentally elevates CD in remote sensing by pioneering a novel paradigm of representation consistency learning. This framework fundamentally advances CD in optical remote sensing by systematically addressing visual style discrepancies. Through innovative representation consistency learning integrating global style transfer and local spatial context constraints, RCCD achieves outperformance on the benchmark (87.68% F1 on LEVIR-CD+, 67.60% on S2Looking), demonstrating particular efficacy in challenging scenarios with large visual style differences (90.55% and 68.76% F1 on respective Large-VS subsets). The proposed visual style validation method establishes a novel paradigm for quantitative error analysis in CD research, enabling targeted algorithm improvement. In the future, we will delve into the beneficial impact of temporal visual style inconsistencies in remote sensing imagery on foundational models and CD tasks, aiming to advance the technological frontiers and foster innovative applications within this domain.

Author Contributions: Conceptualization, J.W. and K.S.; methodology, J.W.; writing—original draft preparation, J.W.; writing—review and editing, K.S. and W.L. (Wangbin Li); validation, S.G.; visualization, S.M.; supervision, K.S.; project administration, W.L. (Wenzhuo Li); software, Y.T.; investigation, W.C.; data curation, Y.D.; funding acquisition, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (No. 2022YFB3902900), and the National Natural Science Foundation of China Projects (No. 42192583). Funded by State Key Laboratory of Geo-Information Engineering, NO. SKLGIE2023-M-3-2.

Data Availability Statement: No new data were created in this manuscript.

Acknowledgments: The authors would like to thank the editor and reviewers for their contributions to this paper.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Zhang, H.; Ma, G.; Fan, H.; Gong, H.; Wang, D.; Zhang, Y. SDCINet: A Novel Cross-Task Integration Network for Segmentation and Detection of Damaged/Changed Building Targets With Optical Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 422–446. [CrossRef]
2. Lv, Z.; Liu, T.; Benediktsson, J.A.; Falco, N. Land Cover Change Detection Techniques: Very-High-Resolution Optical Images: A Review. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 44–63. [CrossRef]
3. Li, W.; Sun, K.; Li, W.; Wei, J.; Miao, S.; Gao, S.; Zhou, Q. Aligning semantic distribution in fusing optical and SAR images for land use classification. *ISPRS J. Photogramm. Remote Sens.* **2023**, *199*, 272–288. [CrossRef]
4. Li, W.; Sun, K.; Li, W.; Huang, X.; Wei, J.; Chen, Y.; Cui, W.; Chen, X.; Lv, X. Assisted learning for land use classification: The important role of semantic correlation between heterogeneous images. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 158–175. [CrossRef]
5. Gao, S.; Sun, K.; Li, W.; Li, D.; Tan, Y.; Wei, J.; Li, W. A Building Change Detection Framework With Patch-Pairing Single-Temporal Supervised Learning and Metric Guided Attention Mechanism. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103785. [CrossRef]
6. Li, D.; Wang, M.; Guo, H.; Jin, W. On China's Earth Observation System: Mission, Vision and Application. *Geo-Spat. Inf. Sci.* **2024**, 1–19.
7. Bai, T.; Wang, L.; Yin, D.; Sun, K.; Chen, Y.; Li, W.; Li, D. Deep learning for change detection in remote sensing: A review. *Geo-Spat. Inf. Sci.* **2023**, *26*, 262–288. [CrossRef]
8. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
9. Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
10. Shen, L.; Lu, Y.; Chen, H.; Wei, H.; Xie, D.; Yue, J.; Chen, R.; Lv, S.; Jiang, B. S2Looking: A Satellite Side-Looking Dataset for Building Change Detection. *Remote Sens.* **2021**, *13*, 5094. [CrossRef]
11. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully convolutional siamese networks for change detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067. [CrossRef]
12. Peng, X.; Zhong, R.; Li, Z.; Li, Q. Optical remote sensing image change detection based on attention mechanism and image difference. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7296–7307. [CrossRef]
13. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
14. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
15. Papadomanolaki, M.; Vakalopoulou, M.; Karantzalos, K. A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7651–7668. [CrossRef]
16. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
17. Feng, Y.; Xu, H.; Jiang, J.; Liu, H.; Zheng, J. ICIF-Net: Intra-Scale Cross-Interaction and Inter-Scale Feature Fusion Network for Bitemporal Remote Sensing Images Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
18. Lin, M.; Yang, G.; Zhang, H. Transition Is a Process: Pair-to-Video Change Detection Networks for Very High Resolution Remote Sensing Images. *IEEE Trans. Image Process.* **2023**, *32*, 57–71. [CrossRef] [PubMed]
19. Pan, J.; Cui, W.; An, X.; Huang, X.; Zhang, H.; Zhang, S.; Zhang, R.; Li, X.; Cheng, W.; Hu, Y. MapsNet: Multi-level feature constraint and fusion network for change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *108*, 102676. [CrossRef]
20. Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 811–815. [CrossRef]
21. Eftekhari, A.; Samadzadegan, F.; Dadrass Javan, F. Building change detection using the parallel spatial-channel attention block and edge-guided deep network. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103180. [CrossRef]
22. Basavaraju, K.S.; Sravya, N.; Lal, S.; Nalini, J.; Reddy, C.S.; Dell'Acqua, F. UCDNet: A Deep Learning Model for Urban Change Detection From Bi-Temporal Multispectral Sentinel-2 Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–10. [CrossRef]
23. Gao, S.; Li, W.; Sun, K.; Wei, J.; Chen, Y.; Wang, X. Built-Up Area Change Detection Using Multi-Task Network with Object-Level Refinement. *Remote Sens.* **2022**, *14*, 957. [CrossRef]
24. Pang, C.; Wu, J.; Ding, J.; Song, C.; Xia, G.S. Detecting building changes with off-nadir aerial images. *Sci. China Inf. Sci.* **2023**, *66*, 140306. [CrossRef]
25. Shi, Q.; Liu, M.; Li, S.; Liu, X.; Wang, F.; Zhang, L. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]

26. Shen, Q.; Huang, J.; Wang, M.; Tao, S.; Yang, R.; Zhang, X. Semantic feature-constrained multitask siamese network for building change detection in high-spatial-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *189*, 78–94. [CrossRef]
27. Lei, J.; Gu, Y.; Xie, W.; Li, Y.; Du, Q. Boundary Extraction Constrained Siamese Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
28. Liu, T.; Gong, M.; Lu, D.; Zhang, Q.; Zheng, H.; Jiang, F.; Zhang, M. Building Change Detection for VHR Remote Sensing Images via Local–Global Pyramid Network and Cross-Task Transfer Learning Strategy. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
29. Wang, L.; Li, Y.; Zhang, M.; Shen, X.; Peng, W.; Shi, W. MSFF-CDNet: A Multiscale Feature Fusion Change Detection Network for Bi-Temporal High-Resolution Remote Sensing Image. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
30. Huang, Y.; Li, X.; Du, Z.; Shen, H. Spatiotemporal Enhancement and Interlevel Fusion Network for Remote Sensing Images Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
31. Zhang, C.; Yue, P.; Tapete, D.; Jiang, L.; Shangguan, B.; Huang, L.; Liu, G. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 183–200. [CrossRef]
32. Ding, Q.; Shao, Z.; Huang, X.; Altan, O. DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102591. [CrossRef]
33. Wang, D.; Chen, X.; Jiang, M.; Du, S.; Xu, B.; Wang, J. ADS-Net: An Attention-Based deeply supervised network for remote sensing image change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *101*, 102348. [CrossRef]
34. Lebedev, M.A.; Vizilter, Y.V.; Vygolov, O.V.; Knyaz, V.A.; Rubis, A.Y. Change detection in remote sensing images using conditional adversarial networks. *ISPRS-Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2018**, *XLII-2*, 565–571. [CrossRef]
35. Deepanshi; Barkur, R.; Suresh, D.; Lal, S.; Reddy, C.S.; Diwakar, P.G. RSCDNet: A Robust Deep Learning Architecture for Change Detection From Bi-Temporal High Resolution Remote Sensing Images. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *7*, 537–551. [CrossRef]
36. Chen, Z.; Zhou, Y.; Wang, B.; Xu, X.; He, N.; Jin, S.; Jin, S. EGDE-Net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement. *ISPRS J. Photogramm. Remote Sens.* **2022**, *191*, 203–222. [CrossRef]
37. Guo, H.; Su, X.; Wu, C.; Du, B.; Zhang, L. SAAN: Similarity-Aware Attention Flow Network for Change Detection With VHR Remote Sensing Images. *IEEE Trans. Image Process.* **2024**, *33*, 2599–2613. [CrossRef] [PubMed]
38. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11966–11976. [CrossRef]
39. Hendrycks, D.; Gimpel, K. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units. In Proceedings of the International Conference Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
41. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man. Cybern. B Cybern.* **1979**, *9*, 62–66. [CrossRef]
42. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
43. Wei, J.; Sun, K.; Li, W.; Li, W.; Gao, S.; Miao, S.; Zhou, Q.; Liu, J. Robust change detection for remote sensing images based on temporospatial interactive attention module. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103767. [CrossRef]
44. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]