Stochastic Attention Refinement for Remote Sensing Change Detection: Learning Adaptive Modulation Patterns Through Contextual Pattern Embedding

Mostafa Etemadinia [a], Saeed Sharifian [a, *]

a. Department of Electrical Engineering, Amirkabir University of Technology, 15914, Tehran, Iran

* Corresponding author: sharifian_s@aut.ac.ir (Saeed Sharifian)

Keywords

Remote Sensing;
Change Detection;
Attention Mechanism;
Stochastic Refinement;
Multi-scale Fusion.

Abstract

Accurate change detection in remote sensing imagery requires sophisticated multi-scale temporal feature integration and attention mechanisms. Current methods suffer from suboptimal multi-scale information utilization due to uniform attention deployment and insufficient feature discriminability from deterministic processing. We propose a novel framework addressing these limitations through two key innovations. First, an Adaptive Scale-Context Attention module with resolution-aware orchestration strategically applies spatial attention at higher scales for precise boundary delineation and channel attention at lower scales for semantic feature selection. Second, and most importantly, we introduce a stochastic attention refinement mechanism that revolutionizes attention-based change detection by learning adaptive modulation patterns through contextual pattern embedding. This stochastic framework employs posterior and prior distributions to model context-dependent enhancement patterns, applying learned contextual representations to dynamically calibrate attention scores and significantly improve feature discriminability beyond deterministic approaches. Our method processes bi-temporal images through dual-stream encoders, applies Adaptive Scale-Context Attention modules with stochastic enhancement across multiple scales, and reconstructs change maps through semantically-aware upsampling. Extensive experiments on four benchmark datasets demonstrate superior performance: we achieve 93.82% F1-score on DSIFN-CD, 94.37% F1-score on WHU building dataset, 91.85% F1-score on LEVIR-CD, and 76.19% F1-score on MSRS-CD, while maintaining computational efficiency with only 6.8M parameters. Comprehensive ablation studies validate the effectiveness of both resolution-aware attention orchestration and stochastic enhancement, establishing a new paradigm for efficient and accurate remote sensing change detection.

1. Introduction

Change detection is a fundamental task in remote sensing that identifies and quantifies differences between bi-temporal images of the same geographical area. This capability enables critical applications, including monitoring deforestation, tracking urban expansion, assessing natural disaster damages, and supporting post-disaster reconstruction [1,2]. The unique requirement to simultaneously analyze temporally separated images distinguishes change detection from conventional computer vision tasks, necessitating specialized architectural designs and sophisticated fusion strategies [3].

The evolution of remote sensing change detection has shifted from traditional statistical approaches to sophisticated deep learning methodologies. Convolutional Neural Networks (CNNs) have become the dominant framework for automatic feature learning from raw imagery [4,5], offering superior capability to extract hierarchical features and capture complex spatial-temporal patterns. However, their reliance on local receptive fields limits handling of intricate spatial-temporal dependencies [6]. This has driven research toward attention mechanisms for selective feature focusing and long-range modeling, as well as stochastic methods for adaptive and robust representations [7,8].

Attention mechanisms have revolutionized change detection by enabling selective focus on semantically relevant regions while suppressing irrelevant information. Early approaches enhanced Siamese networks, such as hybrid Siamese architectures [9] that highlight change information, and multi-scale attention Siamese networks [10] for high-resolution detection. Attention-integrated Siamese U-shaped structures [11] incorporated attention into encoder-decoder frameworks for better multi-temporal comparison. Advanced fusion strategies include the Attention-Guided Multiscale Context Aggregation Network [12], which uses Fully Attentional Pyramid Modules with channel-wise cross-fusion transformers, and multi-scale multi-attention mechanisms [13] for discriminative representations. The Bibranch Fusion Network [14] combines CNN and transformer branches with axial cross-attention. Cross-level methods like Cross-Level Attentive Feature Aggregation (CLAFA) [15] reformulate merging via multiplicative channel and additive gated attention, while Hierarchical Attention Network (HANet) [16] integrates multiscale features with lightweight self-attention to address pixel imbalances. Recent applications include MDANet [17] for urban change detection with difference and attention fusion, dual-attention-guided networks [18] for spatial-channel capture, DAFNet [19] for attention-regularized differences, and channel self-attention with generative adversarial networks [20,21].

Probabilistic and stochastic methods leverage uncertainty quantification for robust temporal comparison [22,23]. Deep Probabilistic Change Models (DPCM) [24] provide unified probabilistic graphical modeling for various change scenarios. Uncertainty frameworks [25] use Support Vector Machine estimates for robust detection from very high-resolution imagery. Diffusion models like DDPM-CD [26] employ Denoising Diffusion Probabilistic Models as feature extractors, and GCD-DDPM [27] generates change maps with variational inference. Semi-supervised approaches such as DCENet [28] incorporate Kullback-Leibler divergence and probabilistic contrast loss for hyperspectral detection. Broader applications include EDiffSR [29] for image super-resolution using diffusion models.

Despite these advances, existing methods exhibit key limitations. Attention-based approaches often apply uniform strategies across scales, failing to leverage those higher resolutions benefit from spatial attention for boundary precision, while lower scales suit channel attention for semantic selection [12,16]. Deterministic mechanisms lack adaptive enhancement based on contextual patterns, limiting subtle change capture [22]. Probabilistic methods focus on global enhancement or reconstruction without integrating stochastic refinement into attention mechanisms or addressing multi-scale deployment. The direct application of stochastic processes to attention mask refinement via contextual embeddings remains unexplored.

This study addresses these gaps by introducing two synergistic innovations: (1) an Adaptive Scale-Context Attention (ASCA) module with resolution-aware orchestration, employing spatial attention at the three highest scales for precise boundary localization and channel attention at the two lowest for semantic discrimination; (2) a stochastic attention refinement framework that learns adaptive modulation patterns through contextual pattern embedding and probabilistic inference (posterior and prior distributions), dynamically calibrating attention scores for enhanced temporal comparison.

The main objectives are to develop and validate this framework for superior performance while maintaining efficiency. Contributions include:

1. An adaptive multi-scale context attention module with strategic deployment of spatial attention at the three highest scales and channel attention at the two lowest scales, optimized for different resolution levels to maximize information extraction and improve detection accuracy through resolution-aware orchestration.
2. A novel stochastic attention refinement framework that learns adaptive modulation patterns through contextual pattern embedding, significantly improving dynamic attention calibration and discriminability for enhanced temporal feature comparison through probabilistic attention mechanisms.
3. Comprehensive experimental validation demonstrating substantial improvements in change detection accuracy across four benchmark datasets, with detailed ablation studies confirming the effectiveness of each component while maintaining computational efficiency.

The remainder of this paper is organized as follows: Section 2 describes the materials and methods, including the ASCA module and stochastic refinement. Section 3 reports experimental results and comparisons. Section 4 provides ablation studies, and Section 5 concludes the work.

2. Materials and Methods

This section presents a comprehensive framework for remote sensing change detection that addresses suboptimal multi-scale information utilization and deterministic feature discriminability through two key innovations: an Adaptive Scale-Context Attention (ASCA) module with Resolution-Aware Orchestration, and a Stochastic Attention Refinement mechanism with Contextual Pattern Embedding. Figure 1 presents the overall architecture. The framework consists of four interconnected components: a dual-stream feature extraction network (Section 2.1), the novel Adaptive Scale-Context Attention module that implements Resolution-Aware Orchestration strategies, enhanced by Stochastic Attention Refinement through the ContextualModulationEngine. (Section 2.2), a progressive upsampling strategy that preserves both semantic coherence and spatial precision (Section 2.3), and a classification module that generates pixel-wise change predictions (Section 2.4).

2.1. Dual-Stream Feature Extraction Network

Our feature extraction framework builds upon the EfficientNetV2 [30] paradigm to the unique requirements of bi-temporal change detection. The choice of this architecture stems from its optimal balance between parameter efficiency and representational capacity, making it particularly suitable for processing high-resolution satellite imagery pairs. Its modules are initialized with pretrained weights from ImageNet to leverage learned low-level feature representations and accelerate convergence.

Our architectural design incorporates four hierarchical feature extraction stages (FE_1-FE_4), each optimized for specific representational tasks. As illustrated in Figure 2, the initial stages (FE_1-FE_3) employ Fused-MBConv blocks for efficient low-level feature extraction, while the deepest stage (FE_4) utilizes MBConv blocks to capture high-level semantic representations.

Given bi-temporal input images I^0 and I^1 with dimensions $3 \times H \times W$, the multi-scale feature extraction process operates as:

$$\begin{aligned} F_k^0 &= \mathrm{FE}_k(I^0;\theta_k), \quad k \in \{1,2,3,4\} \\ F_k^1 &= \mathrm{FE}_k(I^1;\theta_k), \quad k \in \{1,2,3,4\} \end{aligned}, \tag{1}$$

where F_k^0 and F_k^1 represent the extracted feature maps from reference and test images at scale k, respectively, FE_k denotes the feature extraction operation at layer k, and θ_k represents the shared learnable parameters at scale k.

2.2. Adaptive Scale-Context Attention

The Adaptive Scale-Context Attention (ASCA) module represents the core innovation of our approach, addressing the fundamental limitations of uniform attention deployment across resolution levels and deterministic feature processing. Figure 3 illustrates the complete ASCA architecture with dual-stream processing and contextual modulation components.

2.2.1. Stochastic Attention Refinement Framework

A fundamental challenge in change detection is achieving sufficient discriminative power in attention mechanisms to capture subtle temporal changes and complex spatial-temporal patterns. Traditional deterministic attention approaches fail to adaptively modulate attention scores based on contextual dependencies, limiting their effectiveness in challenging scenarios. To address this limitation, we introduce a novel Stochastic Attention Refinement framework implemented through the ContextualModulationEngine module that learns adaptive attention modulation patterns through Contextual Pattern Embedding.

2.2.2. RefinerNetwork with Enhanced Discriminability

Our stochastic framework integrates directly with the scale-context attention mechanism through dedicated RefinerNetwork branches that enhance input features from $C \times H \times W$ to $C' \times H \times W$, where $C' = C/2$ represents the output channels for each temporal stream. These RefinerNetwork components, as shown in Figure 3, provide the foundation for stochastic modulation by creating refined feature representations. The RefinerNetwork operations can be formulated as:

$$\begin{aligned} F_k^{\text{ref}} &= \text{RefinerNetwork}_{\text{ref}}(F_k^0), \quad k \in \{0,1,2,3,4\} \\ F_k^{\text{test}} &= \text{RefinerNetwork}_{\text{test}}(F_k^1), \quad k \in \{0,1,2,3,4\} \end{aligned}, \tag{2}$$

where F_k^0 and F_k^1 represent the input feature maps (at the highest scale $F_0^0 = I^0$, and $F_0^1 = I^1$), k is the associated scale, and F_k^{ref} and F_k^{test} are the refined features that serve as the basis for stochastic enhancement.

2.2.3. Contextual Pattern Embedding for Dynamic Attention Calibration

The core innovation of our approach lies in the ContextualModulationEngine module that learns adaptive modulation patterns through Contextual Pattern Embedding rather than relying on fixed attention mechanisms. This stochastic component models the attention enhancement patterns as contextual embeddings, enabling context-dependent calibration that improves discriminative power for temporal comparison. Figure 4 provides a detailed view of the ContextualModulationEngine architecture, showing the prior and posterior sampling mechanisms during training and inference.

For each temporal stream (reference and test), we estimate contextual attention modulation patterns through dedicated Proj networks (Proj modules in Figure 4). The framework employs separate prior and posterior Proj networks to capture different aspects of the attention enhancement requirements. The prior projections process the original features:

$$\begin{aligned} \mathbf{h}_{\text{ref}}^{\text{prior}} &= \text{Proj}_{\text{ref}}^{\text{prior}}(F_k^0) \\ \mathbf{h}_{\text{test}}^{\text{prior}} &= \text{Proj}_{\text{test}}^{\text{prior}}(F_k^1) \end{aligned}, \tag{3}$$

During training, posterior Proj networks incorporate ground truth information for enhanced learning, as illustrated in the bottom section of Figure 4:

$$\begin{aligned} \mathbf{h}_{\text{ref}}^{\text{post}} &= \text{Proj}_{\text{ref}}^{\text{post}}(\text{Concat}(F_k^0, \mathbf{y})) \\ \mathbf{h}_{\text{test}}^{\text{post}} &= \text{Proj}_{\text{test}}^{\text{post}}(\text{Concat}(F_k^1, \mathbf{y})) \end{aligned}, \tag{4}$$

where $\mathbf{y}$ represents the ground truth change mask (available during training), spatially interpolated to match the feature resolution through the Resize operation shown in Figure 4.

The ContextualModulationEngine estimates both prior and posterior distributions using a single contextual embedding per stream, as shown in the Gaussian distribution modeling component of Figure 4:

$$\begin{aligned} \mathbf{z}_{k,\text{ref}}^{\text{prior}} &\sim q_\phi(\mathbf{z} \mid \mathbf{h}_{\text{ref}}^{\text{prior}}) = \mathcal{N}(\boldsymbol{\mu}_{\text{ref}}, \boldsymbol{\sigma}_{\text{ref}}^2) \\ \mathbf{z}_{k,\text{ref}}^{\text{post}} &\sim q_\phi(\mathbf{z} \mid \mathbf{h}_{\text{ref}}^{\text{post}}) = \mathcal{N}(\boldsymbol{\mu}_{\text{ref}'}, \boldsymbol{\sigma}_{\text{ref}'}{}^2) \end{aligned}, \tag{5}$$

where q_ϕ denotes the contextual embedding posterior parameterized by neural networks that compute mean and log-variance from GlobalAveragePooling features through LinearProj operations. Similar distributions are computed for the test stream.

2.2.4. Efficient Dynamic Attention Calibration

The estimated contextual attention patterns from the ContextualModulationEngine are applied directly to the attention scores through a novel calibration scheme that enables adaptive attention refinement based on learned contextual patterns while maintaining computational tractability. The attention scores for the ASCA module are first computed using the AttentionHeads applied to concatenated bi-temporal features:

$$\begin{aligned} A_k^{\text{ref}} &= \text{ASCA-AttentionHeads}(\text{Concat}(F_k^0, F_k^1)), \quad k \in \{0,1,2,3,4\} \\ A_k^{\text{test}} &= \text{ASCA-AttentionHeads}(\text{Concat}(F_k^0, F_k^1)), \quad k \in \{0,1,2,3,4\} \end{aligned}, \tag{6}$$

where A_k^{ref} and A_k^{test} represent the base attention scores at scale k generated by the ASCA module.

The Stochastic Attention Refinement is then applied through learned modulation parameters derived from the contextual embeddings. During training, contextual embeddings are sampled from the posterior distribution, while during inference, the prior mean is used:

$$\mathbf{z}_{k,\text{stream}} = \begin{cases} \text{sample from } q_\phi(\mathbf{z} \mid \mathbf{h}_{\text{stream}}^{\text{post}}), & \text{if training} \\ \boldsymbol{\mu}_{\text{stream}}^{\text{prior}}, & \text{if inference} \end{cases}, \tag{7}$$

The modulation parameters γ and β are computed through Linear transformations of the contextual embeddings, as illustrated in the right portion of Figure 4:

$$\begin{aligned} [\boldsymbol{\gamma}_{k,\text{ref}}, \boldsymbol{\beta}_{k,\text{ref}}] &= \text{Split}(\text{Linear}(\mathbf{z}_{k,\text{ref}})) \\ [\boldsymbol{\gamma}_{k,\text{test}}, \boldsymbol{\beta}_{k,\text{test}}] &= \text{Split}(\text{Linear}(\mathbf{z}_{k,\text{test}})) \end{aligned}, \tag{8}$$

where Linear denotes a linear transformation that maps the contextual embeddings to modulation parameters, and Split divides the output into scale and bias components. The enhanced attention scores are computed through element-wise modulation:

$$\begin{aligned} \tilde{A}_k^{\text{ref}} &= A_k^{\text{ref}} \odot \text{Softplus}(\boldsymbol{\gamma}_{k,\text{ref}}) + \boldsymbol{\beta}_{k,\text{ref}} \\ \tilde{A}_k^{\text{test}} &= A_k^{\text{test}} \odot \text{Softplus}(\boldsymbol{\gamma}_{k,\text{test}}) + \boldsymbol{\beta}_{k,\text{test}} \end{aligned}, \tag{9}$$

where $\odot$ denotes element-wise multiplication, $\text{Softplus}(\cdot)$ ensures positive scaling factors, and $\tilde{A}_k^{\text{ref}}$ and $\tilde{A}_k^{\text{test}}$ represent the stochastically enhanced attention scores.

2.2.5 Resolution-Aware Orchestration.

Standard attention mechanisms in change detection typically employ uniform attention strategies across different scales, failing to leverage the distinct characteristics and requirements of different resolution levels. To address this limitation, we introduce a Resolution-Aware Orchestration strategy through specialized ASCA-AttentionHeads.

At higher scales (scales 0, 1, 2), we deploy spatial attention ($1 \times H \times W$) to achieve precise boundary localization (ASCA-S), while at lower scales (scales 3, 4), we employ channel attention ($C \times 1 \times 1$) to enhance semantic feature discrimination (ASCA-C):

$$\begin{aligned} A_k^{\text{ref}}, A_k^{\text{test}} &= \text{ASCA-S}(\text{Concat}(F_k^0, F_k^1)), \quad k \in \{0,1,2\} \\ A_k^{\text{ref}}, A_k^{\text{test}} &= \text{ASCA-C}(\text{Concat}(F_k^0, F_k^1)), \quad k \in \{3,4\} \end{aligned}, \tag{10}$$

The stochastic refinement is then applied to these resolution-aware attention scores. This strategic deployment maximizes information extraction efficiency at each resolution level: spatial attention at higher scales enables precise boundary detection by focusing on spatially relevant regions, while channel attention at lower scales enhances semantic discrimination by selecting the most informative feature channels. The enhanced attention masks are then applied to the refined features through element-wise multiplication:

$$\begin{aligned} F_k^{\text{enhanced,ref}} &= F_k^{\text{ref}} \odot \tilde{A}_k^{\text{ref}}, \quad k \in \{0,1,2,3,4\} \\ F_k^{\text{enhanced,test}} &= F_k^{\text{test}} \odot \tilde{A}_k^{\text{test}}, \quad k \in \{0,1,2,3,4\} \end{aligned}, \tag{11}$$

where $F_k^{\text{enhanced, ref}}$ and $F_k^{\text{enhanced, test}}$ are the final enhanced feature maps produced by the ASCA module, as shown by the final outputs in Figure 3.

2.2.6 Training Objective with Stochastic Regularization.

To ensure proper learning of adaptive attention modulation patterns in the ContextualModulationEngine, we introduce a training objective that combines the primary change detection loss with stochastic regularization. During training, the model optimizes the primary change detection loss alongside a Kullback-Leibler (KL) divergence term that regularizes the contextual attention modulation distributions:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{change}} + \beta \mathcal{L}_{\text{KL}} + \lambda \mathcal{L}_{\text{mod}} \\ \mathcal{L}_{\text{KL}} &= \sum_k \Big[\text{KL}(q_\phi(\mathbf{z}_{k,\text{ref}} \mid \mathbf{h}_{\text{ref}}^{\text{post}}) \,\|\, q_\phi(\mathbf{z}_{k,\text{ref}} \mid \mathbf{h}_{\text{ref}}^{\text{prior}})) \\ &\quad + \text{KL}(q_\phi(\mathbf{z}_{k,\text{test}} \mid \mathbf{h}_{\text{test}}^{\text{post}}) \,\|\, q_\phi(\mathbf{z}_{k,\text{test}} \mid \mathbf{h}_{\text{test}}^{\text{prior}})) \Big] \\ \mathcal{L}_{\text{mod}} &= \sum_{k,s} \left(\| \boldsymbol{\gamma}_{k,s} \|_2^2 + \| \boldsymbol{\beta}_{k,s} \|_2^2 \right) \end{aligned}, \tag{12}$$

where $\mathcal{L}_{\text{change}}$ is the primary change detection loss (Binary Cross Entropy), β is the KL weighting factor that controls the strength of stochastic regularization, k indexes the ASCA modules, $s \in \{\text{ref,test}\}$ indexes the temporal streams, and λ is the modulation regularization weight.

The KL regularization ensures that the learned attention modulation distributions remain well-behaved and do not deviate excessively from the prior.

2.3. Progressive Upsampling with Multi-Scale Fusion

Our upsampling approach addresses the critical challenge of preserving both semantic coherence and spatial precision during feature map reconstruction. The strategy combines multi-scale information through carefully designed fusion operations, ensuring that the enhanced features from our ASCA modules are effectively propagated to the final resolution.

At the deepest scale, the module processes enhanced feature maps from the ContextualModulationEngine independently. As processing advances to higher scales, upsampled features from lower scales are strategically concatenated with same-scale enhanced features, creating increasingly rich representations. As illustrated in Figure 5, the progressive upsampling architecture operates through multiple stages with feature fusion at each level.

Formally, let $F_k' \in \mathbb{R}^{H_k \times W_k \times C_k}$ denote the enhanced feature map at scale k after ASCA processing. The progressive upsampling process operates as:

$$\begin{aligned} F_3^{\text{Up}} &= \text{Upsampler}(F_4'; \omega_4) \\ F_{k-1}^{\text{Up}} &= \text{Upsampler}(\text{Concat}(F_k^{\text{Up}}, F_{k-1}'); \omega_k), \quad k \in \{3,2,1\} \end{aligned}, \tag{13}$$

where $\text{Upsampler}(\cdot)$ represents the upsampling operation with learnable parameters ω_k, Concat performs channel-wise concatenation of features, and F_{k-1}' denotes the enhanced feature map at scale $k-1$ after stochastic refinement processing.

The Upsampler module, as detailed in the Figure 5, comprises three synergistic stages. The initial stage employs bilinear upsampling to increase the spatial resolution of input feature maps while preserving spatial relationships. This standard interpolation method provides a stable foundation for subsequent processing stages. The second stage implements an enhanced residual architecture that facilitates gradient flow while computing enriched feature representations. The residual block contains two parallel paths: the main path processes feature through Conv3×3 → BN+ReLU → Conv3×3 operations (shown as "Residual ×2" in the figure), while the skip connection uses Conv1×1 for dimension alignment:

$$\begin{aligned} F_{\text{residual}} &= \text{BN}(\sigma(\text{Conv}_{3\times3}(\text{BN}(\sigma(\text{Conv}_{3\times3}(F_{\text{upsampled}})))))) \\ &\quad + \text{BN}(\sigma(\text{Conv}_{1\times1}(F_{\text{upsampled}}))) \end{aligned}, \tag{14}$$

where σ denotes the PReLU activation function, and β_i represent batch normalization parameters.

The final stage utilizes the Channel-Wise Enhancement module (shown in the upper right of Figure 5) to extract additional discriminative information through adaptive channel attention. This module employs global average pooling followed by linear transformations to compute channel-wise attention weights:

$$s_c = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{\text{residual}}(i, j, c)$$
$$z_c = \sigma_s(\text{Linear}_2(\sigma_r(\text{Linear}_1(s_c; W_1); W_2))) \quad , \tag{15}$$

where s_c represents global average pooling for channel c, Linear_1 and Linear_2 are linear layers with weights W_1 and W_2, σ_r denotes ReLU activation, and σ_s represents sigmoid activation. The enhanced output is computed as:

$$F_{\text{enhanced}}(i, j, c) = z_c \cdot F_{\text{residual}}(i, j, c), \tag{16}$$

2.4. Classifier

After obtaining the final upsampled feature map, we combine the features extracted directly from bi-temporal images using the ASCA module F_0' of size $C_1 \times H \times W$ with the upsampled feature map F_0^{Up} of size $C_2 \times H \times W$. The fusion operation concatenates these feature maps along the channel dimension, resulting in the final feature map:

$$F_{\text{final}} = \text{Concatenate}(F_0^{'}, F_0^{Up}), \tag{17}$$

where $F_{\text{final}} \in \mathbb{R}^{C \times H \times W}$ is the fused feature map with $C = C_1 + C_2$. This fusion ensures that both high-level semantic information (from the upsampled feature map) and enhanced low-level spatial details (from the ASCA module) are preserved for subsequent classification.

The next stage employs a pixel-wise classifier network to generate the final change map predictions. This network consists of three sequential convolutional layers, each utilizing 1×1 kernels to perform channel-wise transformations. The architecture of the classifier can be expressed as:

$$CM = \sigma(\text{Conv}_3(\text{PReLU}(\text{Conv}_2(\text{PReLU}(\text{Conv}_1(F_{\text{final}})))))), \tag{18}$$

Where Conv_1, Conv_2, and Conv_3 represent 1×1 convolutional layers with learnable parameters. PReLU denotes the Parametric ReLU activation function. σ is the Sigmoid activation function.

2.5. Loss Function

Our framework employs a multi-component loss function that balances accurate change detection with proper stochastic attention refinement learning through the ContextualModulationEngine. The training objective consists of three primary components: binary cross-entropy for pixel-level supervision, KL regularization for attention modulation, and modulation parameter regularization.

The primary supervision is provided through binary cross-entropy (BCE) loss, which measures the pixel-wise agreement between predicted change probabilities and ground truth labels:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{19}$$

where N represents the total number of pixels, $y_i \in \{0,1\}$ denotes the ground truth change label for pixel i, and $\hat{y}_i \in [0,1]$ represents the predicted probability for pixel i being changed.

As detailed in the Stochastic Attention Refinement framework, the complete loss formulation incorporates both KL divergence regularization for the Contextual Pattern Embedding and modulation parameter regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \beta \mathcal{L}_{\text{KL}} + \lambda \mathcal{L}_{\text{mod}}$$
$$\mathcal{L}_{\text{KL}} = \sum_{k} \Big[\text{KL}(q_\phi(\mathbf{z}_{k,\text{ref}} \mid \mathbf{h}_{\text{ref}}^{\text{post}}) \| q_\phi(\mathbf{z}_{k,\text{ref}} \mid \mathbf{h}_{\text{ref}}^{\text{prior}})) + \text{KL}(q_\phi(\mathbf{z}_{k,\text{test}} \mid \mathbf{h}_{\text{test}}^{\text{post}}) \| q_\phi(\mathbf{z}_{k,\text{test}} \mid \mathbf{h}_{\text{test}}^{\text{prior}})) \Big] \quad , \tag{20}$$
$$\mathcal{L}_{\text{mod}} = \sum_{k,s} \left(\| \boldsymbol{\gamma}_{k,s} \|_2^2 + \| \boldsymbol{\beta}_{k,s} \|_2^2 \right)$$

where β is the KL weighting factor that controls the strength of Stochastic Attention Refinement regularization, λ is the modulation regularization weight, k indexes the attention modules across scales, $s \in \{\text{ref}, \text{test}\}$ indexes the temporal streams, $q_\phi(\mathbf{z}|\mathbf{h}^{\text{post}})$ represents the posterior distribution conditioned on features and ground truth labels through Contextual Pattern Embedding, $q_\phi(\mathbf{z}|\mathbf{h}^{\text{prior}})$ represents the prior distribution, and $\boldsymbol{\gamma}_{k,s}, \boldsymbol{\beta}_{k,s}$ represent the modulation parameters for each scale and stream generated by the ContextualModulationEngine module.

The KL regularization ensures that the learned attention modulation distributions through Dynamic Attention Calibration remain well-behaved and do not deviate excessively from the prior, which is crucial for maintaining meaningful adaptive attention enhancement while preventing overfitting in the Resolution-Aware Orchestration mechanism. The modulation parameter regularization term $\mathcal{L}_{\text{mod}}$ prevents excessive parameter magnitudes in the Stochastic Attention Refinement mechanism, ensuring stable training and improved generalization.

3. Results and Discussions

This section presents a comprehensive experimental evaluation of our proposed change detection framework, encompassing datasets, evaluation metrics, experimental settings, and detailed comparisons with state-of-the-art methods.

3.1. Datasets

We evaluate our method on four widely used public datasets that represent diverse change detection scenarios:

The **DSIFN-CD** dataset is a challenging binary change detection dataset that includes six pairs of high-resolution (2m) satellite images from six major cities in China, capturing a wide range of land-cover changes, such as those caused by natural disasters and urban development projects [31]. Following standard practice, 14400, 1360, and 192 cropped 256×256 image patches are used for training, validation, and testing, respectively.

The **WHU** building dataset consists of a single pair of high-resolution (0.075m) aerial images depicting a large urban area with complex building changes [32]. As the dataset does not provide predefined splits, the 256×256 cropped patches are randomly divided into 5947, 743, and 744 pairs for training, validation, and testing, respectively, ensuring unbiased evaluation of model generalization performance.

The **LEVIR-CD** dataset is a large-scale building change detection dataset containing 637 pairs of high-resolution (0.5m) satellite images that capture various building-related changes, including construction, demolition, and renovation [21]. The original dataset splits are used, and images are resized to 256×256 patches to maintain consistency across all experiments.

The **MSRS-CD** dataset is a multi-scale remote sensing change detection dataset consisting of 842 pairs of high-resolution (0.5m) satellite images sized 1024×1024, capturing changes such as new buildings, suburban expansion, vegetation changes, and road construction [33]. It is divided into training, validation, and test sets in a 7:1:2 ratio. To validate multi-scale capabilities, we use the original 1024×1024 images without cropping.

3.2. Evaluation Metrics

Five standard evaluation metrics are employed to comprehensively assess the performance of the proposed change detection model: Overall Accuracy (OA), Precision (Pr), Recall (Re), F1-Score, and Intersection over Union (IoU), defined as follows:

$$\begin{aligned} \mathrm{OA} &= \frac{TP+TN}{TP+FP+TN+FN} \\ \mathrm{Pr} &= \frac{TP}{TP+FP} \\ \mathrm{Re} &= \frac{TP}{TP+FN} \\ \text{F1-Score} &= 2\cdot\frac{\mathrm{Pr}\cdot\mathrm{Re}}{\mathrm{Pr}+\mathrm{Re}} \\ \mathrm{IoU} &= \frac{TP}{TP+FP+FN} \end{aligned}, \tag{21}$$

In these equations, TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) denote the number of correctly predicted changed pixels, correctly predicted unchanged pixels, incorrectly predicted changed pixels, and incorrectly predicted unchanged pixels, respectively.

Consistent with standard practice, a threshold of 0.5 is applied to the probabilistic output to generate binary change maps. These maps are then compared against the ground truth using the aforementioned metrics.

3.3. Experimental Setup

To ensure fair comparison with existing methods, all images are partitioned into non-overlapping 256×256 patches. Data augmentation techniques including random horizontal and vertical flipping, random rotation, and random perturbations to brightness and contrast are applied to enhance model generalizability while maintaining label consistency.

Model training is conducted using the AdamW optimizer [34] with an initial learning rate of 0.0001, following a cosine annealing scheduler [35] that decreases to $1e-9$. Each model is trained for 100 epochs with batch size 16. For the MSRS-CD dataset, due to GPU constraints with larger 1024×1024 images, we used batch size 8 with gradient accumulation to simulate an effective larger batch size. The KL regularization weight β for Stochastic Attention Refinement and the modulation parameter regularization weight λ for Dynamic Attention Calibration are set to 0.001 and 1e-5 for all experiments, respectively. Model selection is based on validation F1-Score, with the best-performing model evaluated on held-out test data.

To evaluate computational efficiency in terms of the inference time, we measured them for most baseline methods using their official implementations on an NVIDIA T4 GPU. The measurements were conducted with varying batch sizes on 256×256×3 input patches. For HSSENet [9] and RISNet [36], inference times are as reported in their original papers (per patch at batch size 8).

3.4. Comparison with State-of-the-Art Methods

We evaluate our approach against ten representative state-of-the-art methods spanning diverse architectural paradigms: DTCDSSCN [37], ChangeFormer [38], BIT-CD [6], IFNet [31], SNUNet [39], HSSENet [9], TinyCD [40], S^2CD [41], RISNet [36], and MSNet [33]. These methods encompass CNN-based, Transformer-based, and hybrid architectures with various attention mechanisms, providing comprehensive coverage of current change detection approaches.

Our method achieves superior performance across all four benchmark datasets, demonstrated in Tables 1, 2, 3, and 4. On DSIFN-CD, we achieve an F1-score of 93.82%, surpassing the previous best method (HSSENet) by 2.65%. The IoU improvement is even more substantial at 4.59%, demonstrating enhanced spatial accuracy. On the WHU data set, our approach achieves 94.37% F1-score, outperforming the second-best method (RISNet) by 2.46%. On LEVIR-CD, while the improvements are more modest, we still achieve consistent gains across all metrics, with F1-score reaching 91.85% compared to RISNet's 91.74%. On MSRS-CD, which specifically tests multi-scale feature handling, our method achieves 76.19% F1-score, outperforming MSNet (75.74%) and baselines like BIT-CD (74.11%) and ChangeFormer (71.65%), verifying strong multi-scale capabilities.

The superior performance can be attributed to our dual innovations. Our Resolution-Aware Orchestration strategy within the ASCA module optimally utilizes spatial attention at higher scales for precise boundary delineation and channel attention at lower scales for semantic discrimination. This targeted approach addresses the fundamental limitation of uniform attention deployment in existing methods, which fail to account for the different computational requirements at various feature scales. Meanwhile, our Stochastic Attention Refinement framework through the ContextualModulationEngine module adaptively modulates attention scores based on learned Contextual Pattern Embedding, providing significant improvements in discriminative power compared to the deterministic attention mechanisms employed in competing approaches.

3.5. Computational Efficiency Analysis

To validate the practical applicability of our approach, we analyze the trade-off between computational complexity and performance. Table 5 compares parameter efficiency, computational overhead, and inference time across state-of-the-art methods.

Our method demonstrates exceptional computational efficiency, requiring only 6.8M parameters and 11.04 GFLOPs, and 16.33 ms inference time while delivering state-of-the-art performance. This represents a remarkable efficiency gain compared to heavyweight methods like ChangeFormer (42.02M parameters, 202.87 GFLOPs, 74.46 ms) and IFNet (50.71M parameters, 41.18 GFLOPs, 36.44 ms), which achieve significantly lower accuracy. Even compared to HSSENet, which requires 10.95M parameters, 17.14 GFLOPs, and 13.03 ms inference time, our approach achieves superior performance with 40% fewer parameters, 36% lower computational cost, and comparable inference time.

Furthermore, while ultra-light models like RISNet (9.69 ms) and TinyCD (6.93 ms) offer faster processing, they lag in accuracy by 2-20% on key datasets. Our 16.33 ms inference time strikes an optimal balance, being 2-4x faster than heavier baselines like ChangeFormer and IFNet, and comparable to HSSENet, yet delivering the highest F1 scores (93.82% on DSIFN-CD and 94.37% on WHU).

The efficiency stems from our strategic architectural design. The EfficientNetV2-based backbone provides optimal parameter utilization, while our Resolution-Aware Orchestration within the ASCA module focuses computational resources where they yield maximum benefit. Our Stochastic Attention Refinement through the ContextualModulationEngine module, despite its sophistication, introduces minimal overhead due to its efficient Contextual Pattern Embedding formulation and direct Dynamic Attention Calibration approach. This combination enables optimal balance between accuracy and computational tractability, making the method highly suitable for practical deployment scenarios.

3.6. Qualitative Analysis

The qualitative evaluation demonstrates the effectiveness of ASCA module across diverse change detection scenarios. Figures 6, 7, 8, and 9 present visualizations of our model's performance on DSIFN-CD, WHU building, LEVIR-CD, and MSRS-CD datasets, respectively.

Our model exhibits superior performance in detecting both subtle and significant structural modifications compared to state-of-the-art methods. The reference and test attention masks generated at the highest scale (rightmost columns) reveal the model's capability to identify general structural patterns and edge information through Dynamic Attention Calibration. Despite the limited parameter count of the ASCA layer at this scale, the attention masks demonstrate effective edge detection capabilities, consistently highlighting building boundaries and structural elements across all datasets.

The attention mechanism exhibits generalized responses to structural elements, which aligns with our framework where highest-scale attention performs broad edge detection and general change pattern identification. The visualizations show consistent activation along building boundaries in DSIFN-CD, structural edges in WHU building dataset and MSRS-CD, and residential perimeters in LEVIR-CD, confirming the discriminative power of our Stochastic Attention Refinement. This general edge detection approach through the ContextualModulationEngine module provides substantial advancement over previous methods, as evidenced by the superior change detection accuracy across diverse scenarios.

To complement the quantitative results, Figure 10 provides side-by-side visualizations of predicted change maps from our method and baseline models, allowing qualitative assessment of boundary precision, small object detection, and robustness to noise. For instance, in (f) and (g) featuring small building clusters and residential areas, our model excels in small object detection by capturing subtle changes as true positives with minimal false negatives. Similarly, (b) and (h) with linear structures and larger buildings highlight superior boundary precision through sharper, aligned edges via the ASCA module, while demonstrating robustness to environmental noise like shadows or terrain artifacts that plague methods like IFNet.

4. Ablation Study

We conduct comprehensive ablation studies to validate the effectiveness of our key innovations and design choices. The studies systematically analyze: (1) Resolution-Aware Orchestration deployment strategies within the ASCA module, (2) Stochastic Attention Refinement modulation schemes through the ContextualModulationEngine module, (3) the impact of KL divergence regularization weight β on Dynamic Attention Calibration performance.

4.1. Resolution-Aware Orchestration Deployment Analysis

To validate the Resolution-Aware Orchestration strategy within the Adaptive Scale-Context Attention module, we evaluate different attention configurations across the five scales of the proposed architecture. Our core hypothesis posits that spatial attention at higher scales optimizes boundary detection while channel attention at lower resolutions enhances semantic feature discrimination through the ContextualModulationEngine module. Since exhaustive evaluation of all 32 possible combinations would be computationally prohibitive, we focus on six strategic configurations that test this fundamental premise.

We evaluate: (1) spatial attention at all scales (Spatial-all), (2) channel attention at all scales (Channel-all), (3) spatial attention at 2 highest scales with channel attention at 3 lowest scales (Spatial2-Channel3), (4) spatial attention at 3 highest scales

with channel attention at 2 lowest scales (Spatial3-Channel2), (5) channel attention at 3 highest scales with spatial attention at 2 lowest scales (Channel3-Spatial2), and (6) channel attention at 2 highest scales with spatial attention at 3 lowest scales (Channel2-Spatial3).

The results in Table 6 provide strong empirical validation of our design hypothesis. Our proposed Spatial3-Channel2 configuration within the ASCA module achieves superior performance across both datasets: 93.82% F1 on DSIFN-CD and 94.37% F1 on WHU-CD, representing consistent improvements of 0.44-0.98 percentage points over alternativs. The performance differential between uniform attention strategies (Spatial-all: 93.13%, Channel-all: 93.38%) and our Resolution-Aware Orchestration approach demonstrates that neither pure spatial nor pure channel attention is optimal across all scales.

Key findings validate our theoretical framework: (1) the poor performance of Channel2-Spatial3 (92.84% F1 on DSIFN-CD) compared to our approach confirms that misaligned Resolution-Aware Orchestration deployment within the ASCA module significantly degrades performance; (2) the comparison between Spatial2-Channel3 (93.35%) and Spatial3-Channel2 (93.82%) reveals that the optimal transition from spatial to channel attention occurs at scale 3; (3) consistent results across both datasets demonstrate the generalizability of our Resolution-Aware Orchestration strategy, with IoU improvements (88.36% vs. 87.76% for second-best on DSIFN-CD) confirming enhanced boundary delineation capabilities through Dynamic Attention Calibration.

4.2. Stochastic Attention Refinement Modulation Analysis

To validate the Stochastic Attention Refinement framework within the ASCA, we compare different modulation strategies focusing on distribution choice (single vs. dual) and modulation target (features vs. attention scores). We evaluate five approaches: (1) no modulation baseline, (2) single distribution feature modulation, (3) dual distributions feature modulation, (4) dual distributions attention modulation, and (5) our proposed single distribution attention modulation.

For dual distribution feature modulation, separate scale and bias contextual embeddings are learned through Contextual Pattern Embedding:

$$\begin{aligned}\mathbf{u}_{k,\text{stream}} &\sim q_\phi(\mathbf{u} \mid \mathbf{h}^{\text{prior}}_{\text{stream}}) = \mathcal{N}(\boldsymbol{\mu}_u, \boldsymbol{\sigma}_u^2) \\ \mathbf{s}_{k,\text{stream}} &\sim q_\phi(\mathbf{s} \mid \mathbf{h}^{\text{prior}}_{\text{stream}}) = \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s^2)\end{aligned}, \tag{22}$$

The feature modulation is applied as:

$$\begin{aligned}\boldsymbol{\gamma}_{k,\text{stream}} &= \text{Linear}(\mathbf{s}_{k,\text{stream}}) \\ \boldsymbol{\beta}_{k,\text{stream}} &= \text{Linear}(\mathbf{u}_{k,\text{stream}}) \\ F_k^{\text{mod, stream}} &= \text{InstanceNorm}(F_k^{\text{stream}}) \odot \text{Softplus}(\boldsymbol{\gamma}_{k,\text{stream}}) + \boldsymbol{\beta}_{k,\text{stream}}\end{aligned}, \tag{23}$$

Where stream is either ref or test.

The same dual distribution modulation can be applied to attention scores within the ASCA module through the ContextualModulationEngine.

Our proposed single distribution attention modulation uses a unified contextual embedding per stream through Contextual Pattern Embedding:

$$\begin{aligned}\mathbf{z}_{k,\text{stream}} &\sim q_\phi(\mathbf{z} \mid \mathbf{h}^{\text{prior}}_{\text{stream}}) = \mathcal{N}(\boldsymbol{\mu}_{\text{stream}}, \boldsymbol{\sigma}^2_{\text{stream}}) \\ [\boldsymbol{\gamma}_{k,\text{stream}}, \boldsymbol{\beta}_{k,\text{stream}}] &= \text{Split}(\text{Linear}(\mathbf{z}_{k,\text{stream}}))\end{aligned}, \tag{24}$$

The results in Table 7 demonstrate that our proposed single distribution attention modulation through Dynamic Attention Calibration within the ASCA module achieves the best performance: 93.82% F1 on DSIFN-CD and 94.37% on WHU-CD. Key findings include: (1) any Stochastic Attention Refinement significantly outperforms the baseline, validating adaptive attention enhancement through the ContextualModulationEngine module integrated with the ASCA architecture; (2) attention score modulation consistently outperforms feature modulation, confirming that directly modulating attention mechanisms through Dynamic Attention Calibration within the ASCA module is more effective than modifying intermediate features; (3) single distribution approaches achieve superior performance compared to dual distribution methods, indicating that unified contextual embeddings through Contextual Pattern Embedding capture attention modulation patterns more effectively than separate modeling.

4.3. KL Divergence Regularization Weight Analysis

To determine the optimal KL divergence regularization weight β in the Stochastic Attention Refinement framework within the ASCA module, we conduct a systematic evaluation across different regularization strengths. The regularization weight controls the balance between change detection performance and the constraint on learned contextual embedding distributions through the ContextualModulationEngine module, ensuring they remain well-calibrated while providing sufficient flexibility for Dynamic Attention Calibration.

We evaluate four different β values: no regularization ($\beta = 0$), weak regularization ($\beta = 0.0001$), moderate regularization ($\beta = 0.001$), and strong regularization ($\beta = 0.01$). Each configuration is tested on both DSIFN-CD and WHU-CD datasets to assess the generalizability of our findings.

The results in Table 8 reveal that $\beta = 0.001$ achieves optimal performance with 93.82% F1 on DSIFN-CD and 94.37% on WHU-CD. Key observations include: (1) without regularization ($\beta = 0$), the model achieves reasonable performance but fails to fully exploit Stochastic Attention Refinement due to unconstrained contextual embedding distributions in the ContextualModulationEngine module; (2) weak regularization ($\beta = 0.0001$) provides modest improvement by introducing mild constraints on the learned distributions through Contextual Pattern Embedding; (3) moderate regularization ($\beta = 0.001$) achieves the optimal balance, allowing sufficient flexibility for Dynamic Attention Calibration within the ASCA module while maintaining well-calibrated distributions; (4) strong regularization ($\beta = 0.01$) degrades performance by overly constraining the contextual embedding distributions, limiting the model's capacity to learn effective attention enhancement patterns through the

ContextualModulationEngine module. The optimal $\beta = 0.001$ ensures that learned contextual embedding distributions remain close to the prior while providing adequate degrees of freedom for context-dependent Dynamic Attention Calibration within the ASCA architecture.

5. Conclusion

In In this paper, we proposed a novel framework for remote sensing change detection that overcomes two key limitations of existing methods: suboptimal multi-scale information utilization and deterministic feature discriminability. Our approach introduces two complementary innovations: an Adaptive Scale-Context Attention (ASCA) module with Resolution-Aware Orchestration and a Stochastic Attention Refinement mechanism via Contextual Pattern Embedding. The ASCA module strategically applies spatial attention at higher resolutions for precise boundary delineation and channel-wise attention at lower resolutions for robust semantic selection, enhanced by Dynamic Attention Calibration. The Stochastic Attention Refinement, implemented through the ContextualModulationEngine, adaptively modulates attention scores based on learned contextual dependencies, significantly outperforming deterministic alternatives.

Extensive experiments on four public datasets (DSIFN-CD, WHU building dataset, LEVIR-CD, and MSRS-CD) demonstrate that our method achieves state-of-the-art performance in F1-score, IoU, and overall accuracy, while remaining computationally efficient. Ablation studies confirm the individual contributions of Resolution-Aware Orchestration and Stochastic Attention Refinement to improved detection accuracy. These results underscore the value of adaptive, context-aware attention mechanisms in advancing remote sensing change detection.

Authors' Credit

First Author
Mostafa Etemadinia: Conceptualization, Methodology, Data Curation, Investigation, Software, Validation, Writing – original draft.
Second Author
Saeed Sharifian: Conceptualization, Methodology, Formal Analysis, Supervision, Validation, Writing – review & editing.

Funding

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

Conflicts of interest

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

References

1. Hemati, M., Hasanlou, M., Mahdianpari, M., et al. "A systematic review of landsat data for change detection applications: 50 years of monitoring the earth", *Remote Sens. (Basel).*, **13**(15), p. 2869 (2021). https://doi.org/10.3390/rs13152869
2. Wang, Z., Peng, C., Zhang, Y., et al. "Fully convolutional siamese networks based change detection for optical aerial images with focal contrastive loss", *Neurocomputing*, **457**, pp. 155–167 (2021). https://doi.org/10.1016/j.neucom.2021.06.059
3. Bai, T., Wang, L., Yin, D., et al. "deep learning for change detection in remote sensing: a review", *Geo-spatial Information Science*, **26**(3), pp. 262–288 (2023). https://doi.org/10.1080/10095020.2022.2085633
4. Shafique, A., Cao, G., Khan, Z., et al. "Deep learning-based change detection in remote sensing images: a review", *Remote Sens. (Basel).*, **14**(4), p. 871 (2022). https://doi.org/10.3390/rs14040871
5. Khelifi, L. and Mignotte, M., "Deep learning for change detection in remote sensing images: comprehensive review and meta-analysis", *IEEE Access*, **8**, pp. 126385–126400 (2020). https://doi.org/10.1109/ACCESS.2020.3008036
6. Chen, H., Qi, Z., and Shi, Z., "Remote sensing image change detection with transformers", *IEEE Transactions on Geoscience and Remote Sensing*, **60**, pp. 1–14 (2022). https://doi.org/10.1109/TGRS.2021.3095166
7. Saleh, A., Sheaves, M., Jerry, D., et al. "Adaptive deep learning framework for robust unsupervised underwater image enhancement", *Expert Syst. Appl.*, **268**, p. 126314 (2025). https://doi.org/10.1016/j.eswa.2024.126314
8. Niu, Z., Zhong, G., and Yu, H., "A review on the attention mechanism of deep learning", *Neurocomputing*, **452**, pp. 48–62 (2021). https://doi.org/10.1016/j.neucom.2021.03.091
9. Yan, L. and Jiang, J., "A hybrid siamese network with spatiotemporal enhancement and two-level feature fusion for remote sensing image change detection", *IEEE Transactions on Geoscience and Remote Sensing*, **61**, pp. 1–17 (2023). https://doi.org/10.1109/TGRS.2023.3268294
10. Li, J., Zhu, S., Gao, Y., et al. "Change detection for high-resolution remote sensing images based on a multi-scale attention siamese network", *Remote Sens. (Basel).*, **14**(14), p. 3464 (2022). https://doi.org/10.3390/rs14143464
11. Song, L., Xia, M., Jin, J., et al. "SUACDNet: Attentional change detection network based on siamese u-shaped structure", *International Journal of Applied Earth Observation and Geoinformation*, **105**, p. 102597 (2021). https://doi.org/10.1016/j.jag.2021.102597

12. Xu, X., Yang, Z., and Li, J., "AMCA: Attention-guided multiscale context aggregation network for remote sensing image change detection", *IEEE Transactions on Geoscience and Remote Sensing*, **61**, pp. 1–19 (2023). https://doi.org/10.1109/TGRS.2023.3272006
13. Zhang, M., Liu, Z., Feng, J., et al. "Remote sensing image change detection based on deep multi-scale multi-attention siamese transformer network", *Remote Sens. (Basel).*, **15**(3), p. 842 (2023). https://doi.org/10.3390/rs15030842
14. Song, L., Xia, M., Weng, L., et al. "Axial cross attention meets cnn: bibranch fusion network for change detection", *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **16**, pp. 21–32 (2023). https://doi.org/10.1109/JSTARS.2022.3224081
15. Wang, G., Cheng, G., Zhou, P., et al. "Cross-level attentive feature aggregation for change detection", *IEEE Transactions on Circuits and Systems for Video Technology*, **34**(7), pp. 6051–6062 (2024). https://doi.org/10.1109/TCSVT.2023.3344092
16. Han, C., Wu, C., Guo, H., et al. "HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images", *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **16**, pp. 3867–3878 (2023). https://doi.org/10.1109/JSTARS.2023.3264802
17. Jiang, S., Lin, H., Ren, H., et al. "MDANet: A high-resolution city change detection network based on difference and attention mechanisms under multi-scale feature fusion", *Remote Sens. (Basel).*, **16**(8), p. 1387 (2024). https://doi.org/10.3390/rs16081387
18. Ren, H., Xia, M., Weng, L. et al. "Dual-attention-guided multiscale feature aggregation network for remote sensing image change detection", *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **17**, pp. 4899–4916 (2024). https://doi.org/10.1109/JSTARS.2024.3362370
19. Ma, C., Yin, H., Weng, L., et al. "DAFNet: A novel change-detection model for high-resolution remote-sensing imagery based on feature difference and attention mechanism", *Remote Sens. (Basel).*, **15**(15), p. 3896 (2023). https://doi.org/10.3390/rs15153896
20. Wang, Z., Zhang, Y., Luo, L., et al. "CSA-CDGAN: Channel self-attention-based generative adversarial network for change detection of remote sensing images", *Neural Comput. Appl.*, **34**(24), pp. 21999–22013 (2022). https://doi.org/10.1007/s00521-022-07637-z
21. Chen, H. and Shi, Z., "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection", *Remote Sens. (Basel).*, **12**(10) (2020). https://doi.org/10.3390/rs12101662
22. Shen, J., Zhang, C., Zhang, M., et al. "Learning remote sensing aleatoric uncertainty for semi-supervised change detection", *IEEE Transactions on Geoscience and Remote Sensing*, **62**, pp. 1–13 (2024). https://doi.org/10.1109/TGRS.2024.3437250
23. Ferchichi, A., Ferchichi, A., Hendaoui, F., et al. "Deep Learning-based Uncertainty Quantification for spatio-temporal environmental Remote Sensing: A systematic literature review", *Neurocomputing*, **639**, p. 130242 (2025). https://doi.org/10.1016/j.neucom.2025.130242
24. Zheng, Z., Zhong, Y., Zhao, J., et al. "Unifying remote sensing change detection via deep probabilistic change models: from principles, models to applications", *ISPRS Journal of Photogrammetry and Remote Sensing*, **215**, pp. 239–255 (2024). https://doi.org/10.1016/j.isprsjprs.2024.07.001
25. Salah, M., "Uncertainty management for robust probabilistic change detection from multi-temporal geoeye-1 imagery", *Applied Geomatics*, **13**(2), pp. 261–275 (2021). https://doi.org/10.1007/s12518-020-00346-z
26. Chaminda Bandara, W. G., Nair, N. G., and Patel, V. M., "DDPM-CD: Denoising diffusion probabilistic models as feature extractors for remote sensing change detection", *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp. 5250–5262 (2025). https://doi.org/10.1109/WACV61041.2025.00513
27. Wen, Y., Ma, X., Zhang, X., et al. "GCD-DDPM: A generative change detection model based on difference-feature-guided DDPM", *IEEE Transactions on Geoscience and Remote Sensing*, **62**, pp. 1–16 (2024). https://doi.org/10.1109/TGRS.2024.3381752
28. Luo, F., Zhou, T., Liu, J., et al. "DCENet: Diff-feature contrast enhancement network for semi-supervised hyperspectral change detection", *IEEE Transactions on Geoscience and Remote Sensing*, **62**, pp. 1–14 (2024). https://doi.org/10.1109/TGRS.2024.3374600
29. Xiao, Y., Yuan, Q., Jiang, K., et al. "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution", *IEEE Transactions on Geoscience and Remote Sensing*, **62**, pp. 1–14 (2024). https://doi.org/10.1109/TGRS.2023.3341437
30. Tan, M. and Le, Q. V, "EfficientNetV2: Smaller models and faster training", *arXiv preprint arXiv:2104.00298* (2021). https://doi.org/10.48550/arXiv.2104.00298
31. Zhang, C., Yue, P., Tapete, D., et al. "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images", *ISPRS Journal of Photogrammetry and Remote Sensing*, **166**, pp. 183–200 (2020). https://doi.org/10.1016/j.isprsjprs.2020.06.003
32. Ji, S., Wei, S., and Lu, M., "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set", *IEEE Transactions on Geoscience and Remote Sensing*, **57**(1), pp. 574–586 (2019). https://doi.org/10.1109/TGRS.2018.2858817
33. Liu, S., Zhao, D., Zhou, Y., et al. "Network and dataset for multiscale remote sensing image change detection", *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **18**, pp. 2851–2866 (2025). https://doi.org/10.1109/JSTARS.2024.3522135
34. Loshchilov, I. and Hutter, F., "Decoupled Weight Decay Regularization", *arXiv preprint arXiv:1711.05101* (2019). https://doi.org/10.48550/arXiv.1711.05101
35. Loshchilov, I. and Hutter, F., "SGDR: Stochastic gradient descent with warm restarts", *arXiv preprint arXiv:1608.03983* (2017). https://doi.org/10.48550/arXiv.1608.03983

36. Chong, Y., Ge, X., and Pan, S., "RISNet: Robust ill-posed solver for remote sensing image change detection", *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **18**, pp. 7625–7643 (2025). https://doi.org/10.1109/JSTARS.2025.3541260
37. Liu, Y., Pang, C., Zhan, Z., et al. "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model", *IEEE Geoscience and Remote Sensing Letters*, **18**(5), pp. 811–815 (2021). https://doi.org/10.1109/LGRS.2020.2988032
38. Bandara, W. G. C. and Patel, V. M., "A transformer-based siamese network for change detection", *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, pp. 207–210 (2022). https://doi.org/10.1109/IGARSS46834.2022.9883686
39. Fang, S., Li, K., Shao, J., et al. "SNUNet-CD: A densely connected siamese network for change detection of vhr images", *IEEE Geoscience and Remote Sensing Letters*, **19**, pp. 1–5 (2022). https://doi.org/10.1109/LGRS.2021.3056416
40. Codegoni, A., Lombardi, G., and Ferrari, A., "TINYCD: A (not so) deep learning model for change detection", *Neural Comput. Appl.*, **35**(11), pp. 8471–8486 (2023). https://doi.org/10.1007/s00521-022-08122-3
41. Wang, L., Fang, Y., Li, Z., et al. "Summator–subtractor network: modeling spatial and channel differences for change detection", *IEEE Transactions on Geoscience and Remote Sensing*, **62**, pp. 1–12 (2024). https://doi.org/10.1109/TGRS.2024.3349638

List of Captions

Figure. 1. Overall architecture of the proposed change detection framework. The dual-stream encoder processes bi-temporal images (I^0, I^1) through shared feature extraction blocks (FE_1-FE_4) at multiple scales (S_0-S_4). Adaptive Scale-Context Attention modules implement Resolution-Aware Orchestration (ASCA-S for spatial attention, ASCA-C for channel-wise attention) enhanced by Stochastic Attention Refinement through the ContextualModulationEngine, followed by progressive upsampling (UP) and final classification (CLS) to produce change maps (CM).
Figure. 2. Architecture of the dual-stream feature extraction network showing the four hierarchical stages FE_1-FE_4.
Figure. 3. Architecture of the ASCA-C module showing the dual-stream processing with RefinerNetwork components, ContextualModulationEngine, and AttentionHeads. The architecture processes reference and test features through parallel streams, applying contextual modulation and dynamic attention calibration to produce enhanced feature representations.
Figure. 4. Detailed architecture of the ContextualModulationEngine showing the dual-mode operation: (top) prior sampling during inference using feature projections, and (bottom) posterior sampling during training incorporating ground truth information. The module generates contextual embeddings through Gaussian distributions and produces modulation parameters γ and β for dynamic attention calibration.
Figure. 5. Progressive upsampling architecture with multi-scale fusion. The diagram shows the complete upsampling pipeline from the deepest feature maps ($C_4 \times H/16 \times W/16$) enhanced by the ContextualModulationEngine to the final output ($C_0 \times H \times W$).
Figure. 6. Visualization of our model performance on DSIFN-CD dataset. From left to right: reference image, test image, ground truth, prediction with classification results, reference attention mask, and test attention mask generated at highest scale.
Figure. 7. Visualization of our model performance on WHU building dataset. From left to right: reference image, test image, ground truth, prediction with classification results, reference attention mask, and test attention mask generated at highest scale.
Figure. 8. Visualization of our model performance on LEVIR-CD dataset. From left to right: reference image, test image, ground truth, prediction with classification results, reference attention mask, and test attention mask generated at highest scale.
Figure. 9. Visualization of our model performance on MSRSCD dataset. From left to right: reference image, test image, ground truth, prediction with classification results, reference attention mask, and test attention mask generated at highest scale.
Figure. 10. Visualization results of the baselines and our model tested on the LEVIR-CD dataset. Where white, green, red, and black, respectively, represent true positive, false negative, false positive, and true negative. Figures (a)–(h) display the detection results on the LEVIR-CD dataset.

Table 1. Comparison of the proposed method with state-of-the-art methods on DSIFN-CD dataset.
Table 2. Comparison of the proposed method with state-of-the-art methods on WHU building dataset.
Table 3. Comparison of the proposed method with state-of-the-art methods on LEVIR-CD dataset.
Table 4. Comparison of the proposed method with state-of-the-art methods on MSRS-CD dataset.
Table 5. Parameter efficiency versus F1-Score comparison among state-of-the-art methods.
Table 6. Ablation study on Resolution-Aware Orchestration deployment configurations within the ASCA module.
Table 7. Ablation study on Stochastic Attention Refinement modulation schemes through the ContextualModulationEngine module within the ASCA architecture.
Table 8. Ablation study on KL divergence regularization weight β.

Figures

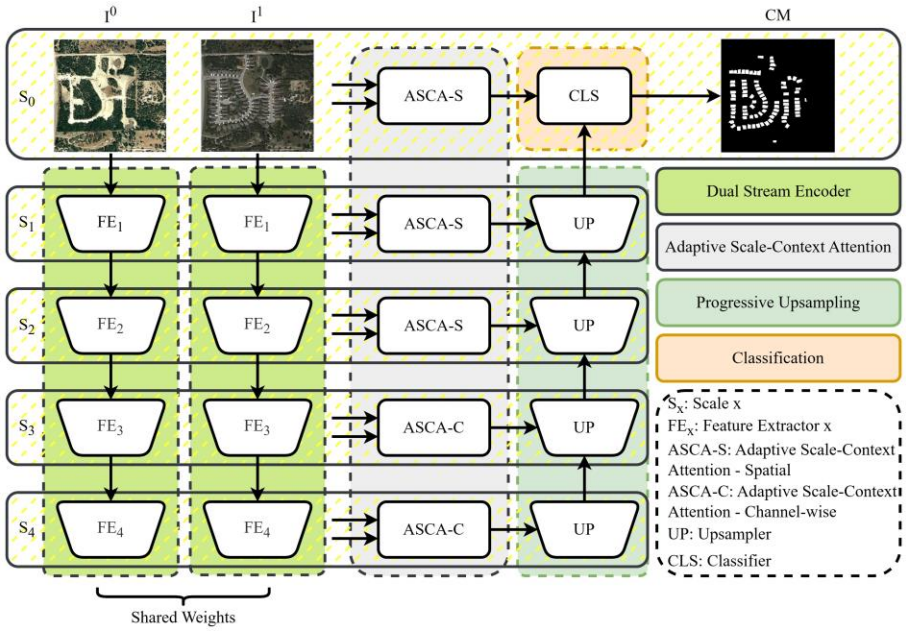

Figure. 1

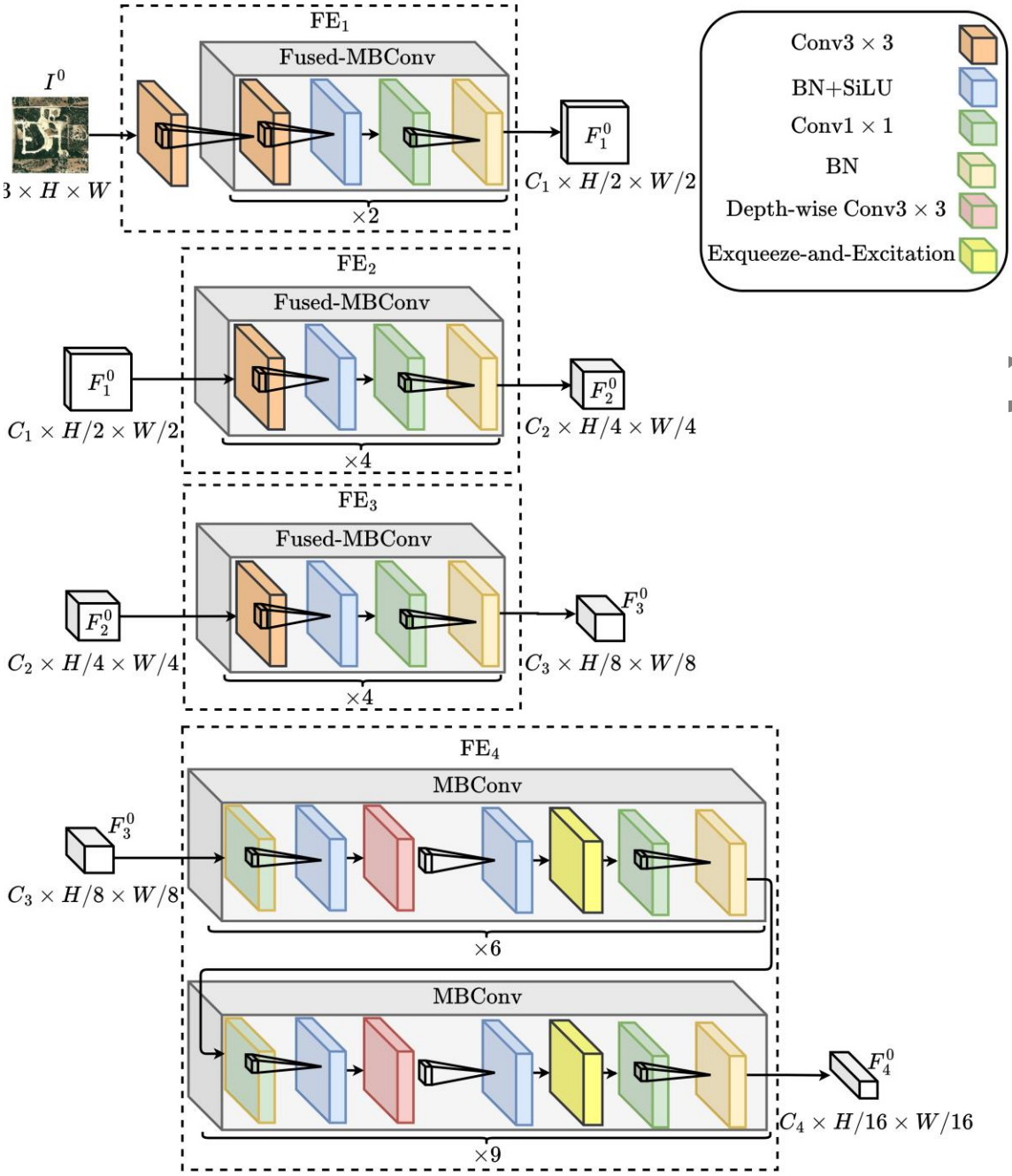

Figure. 2

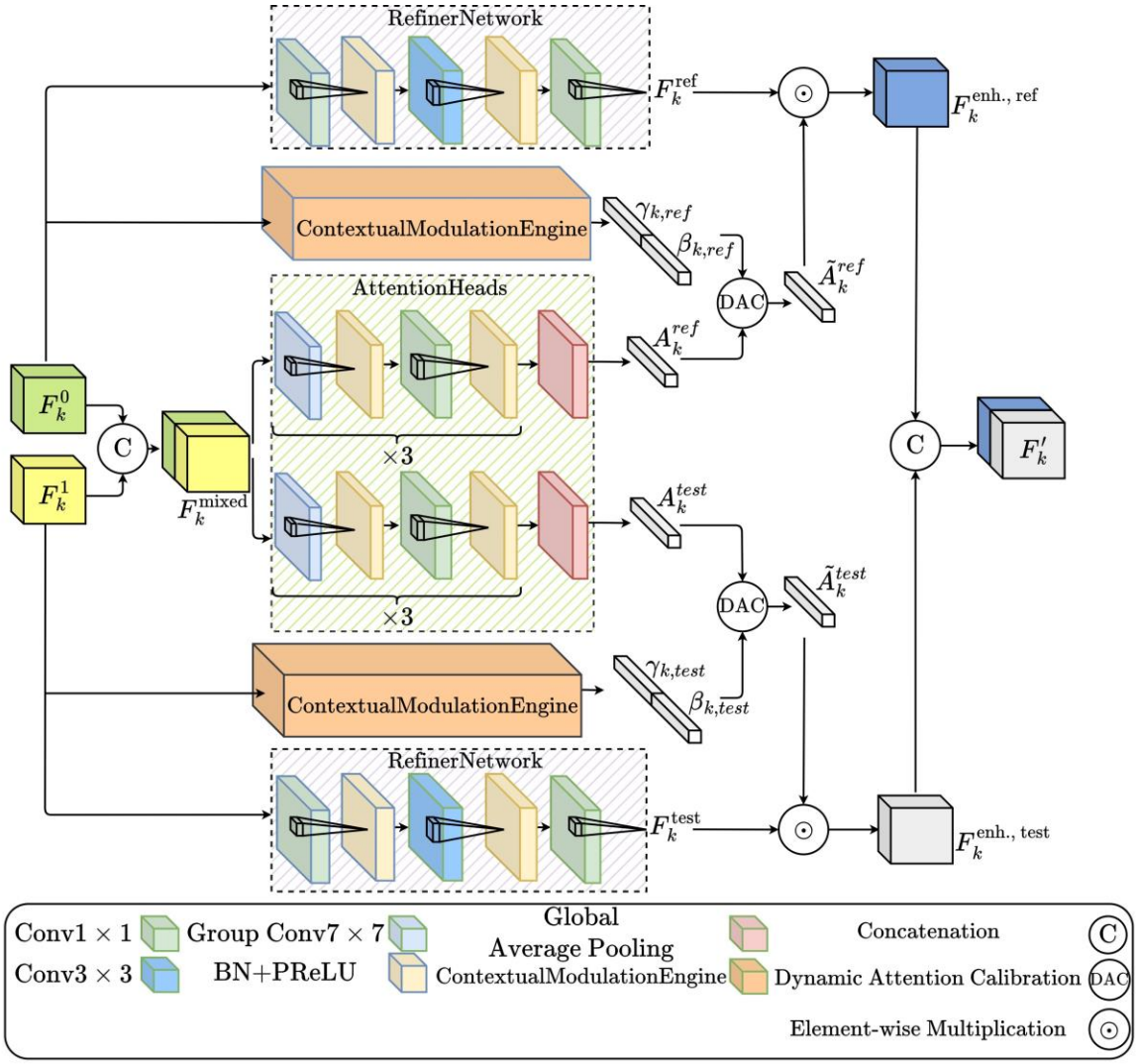

Figure. 3

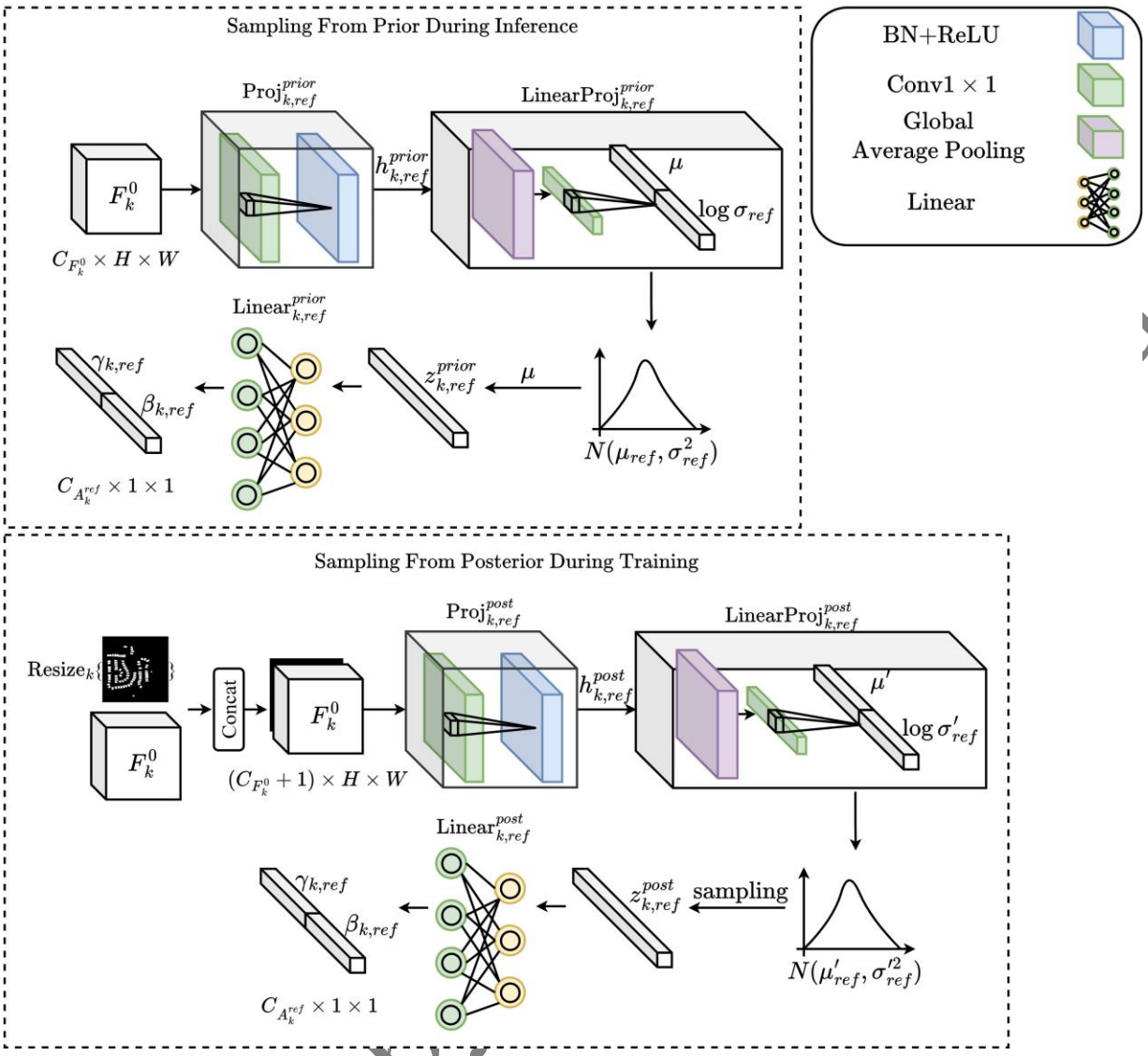

Figure. 4

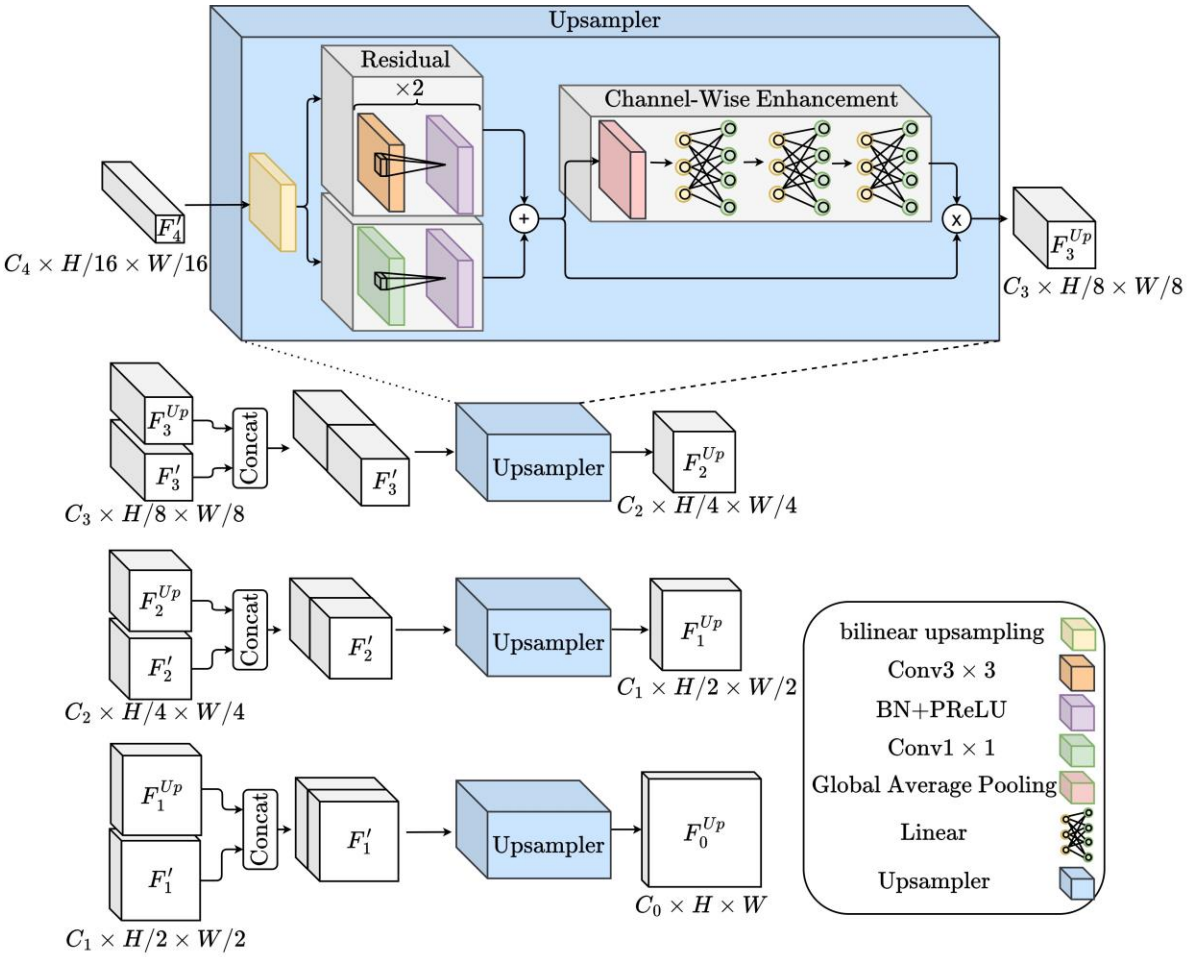

Figure. 5

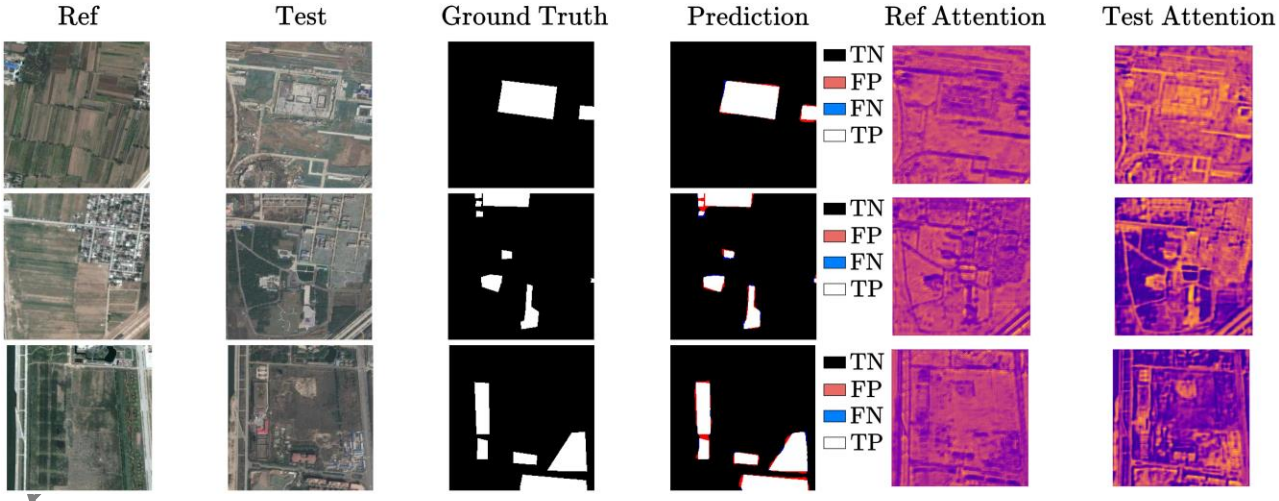

Figure. 6

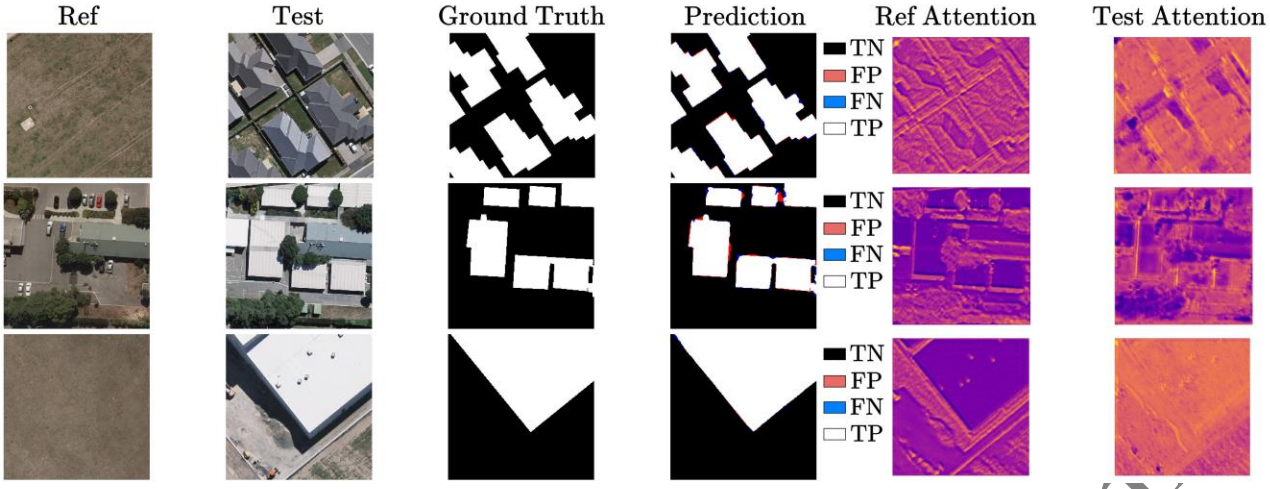

Figure. 7

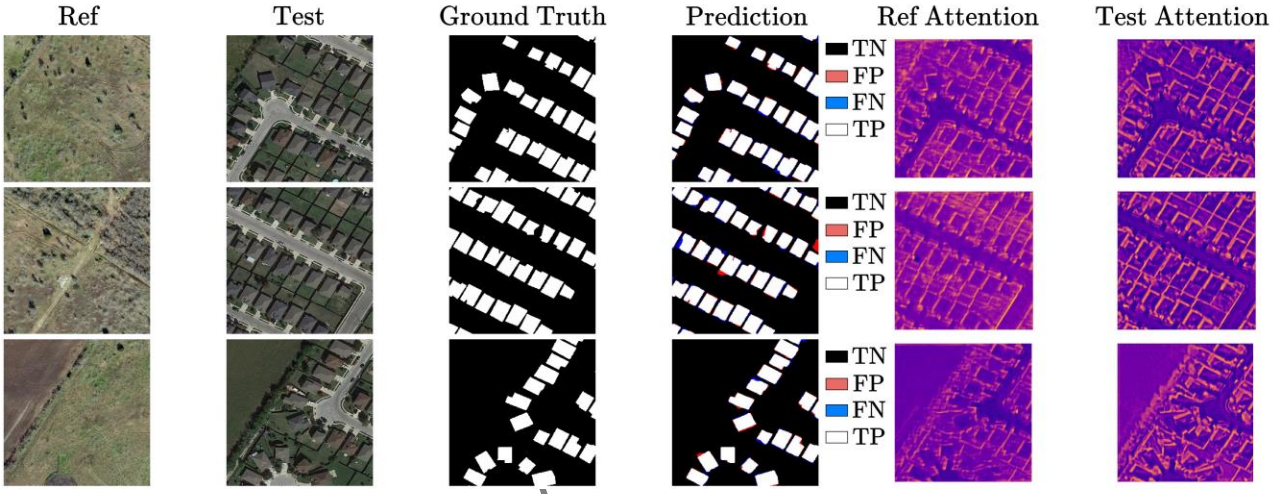

Figure. 8

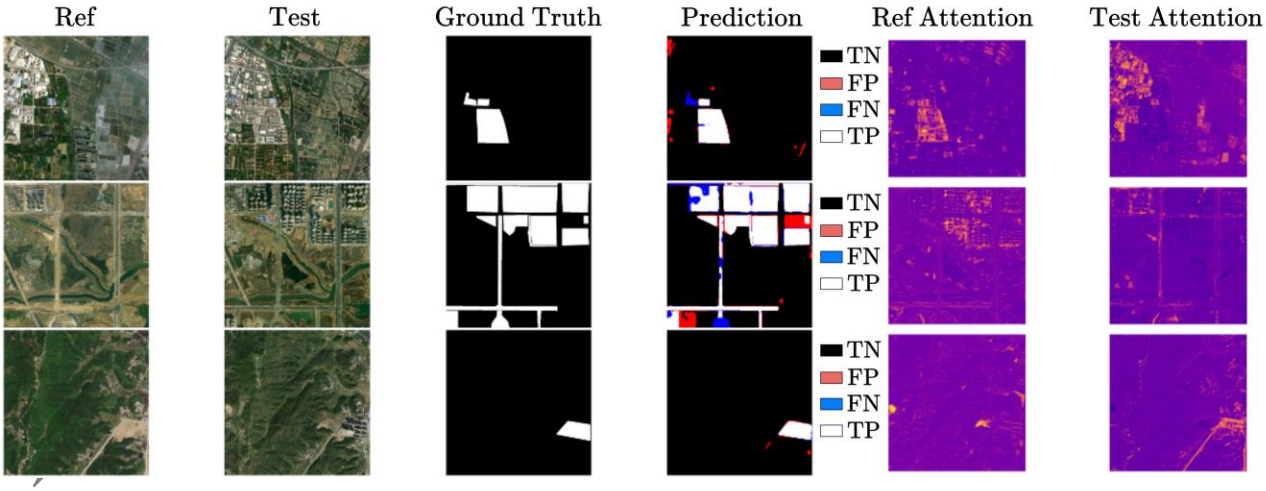

Figure. 9

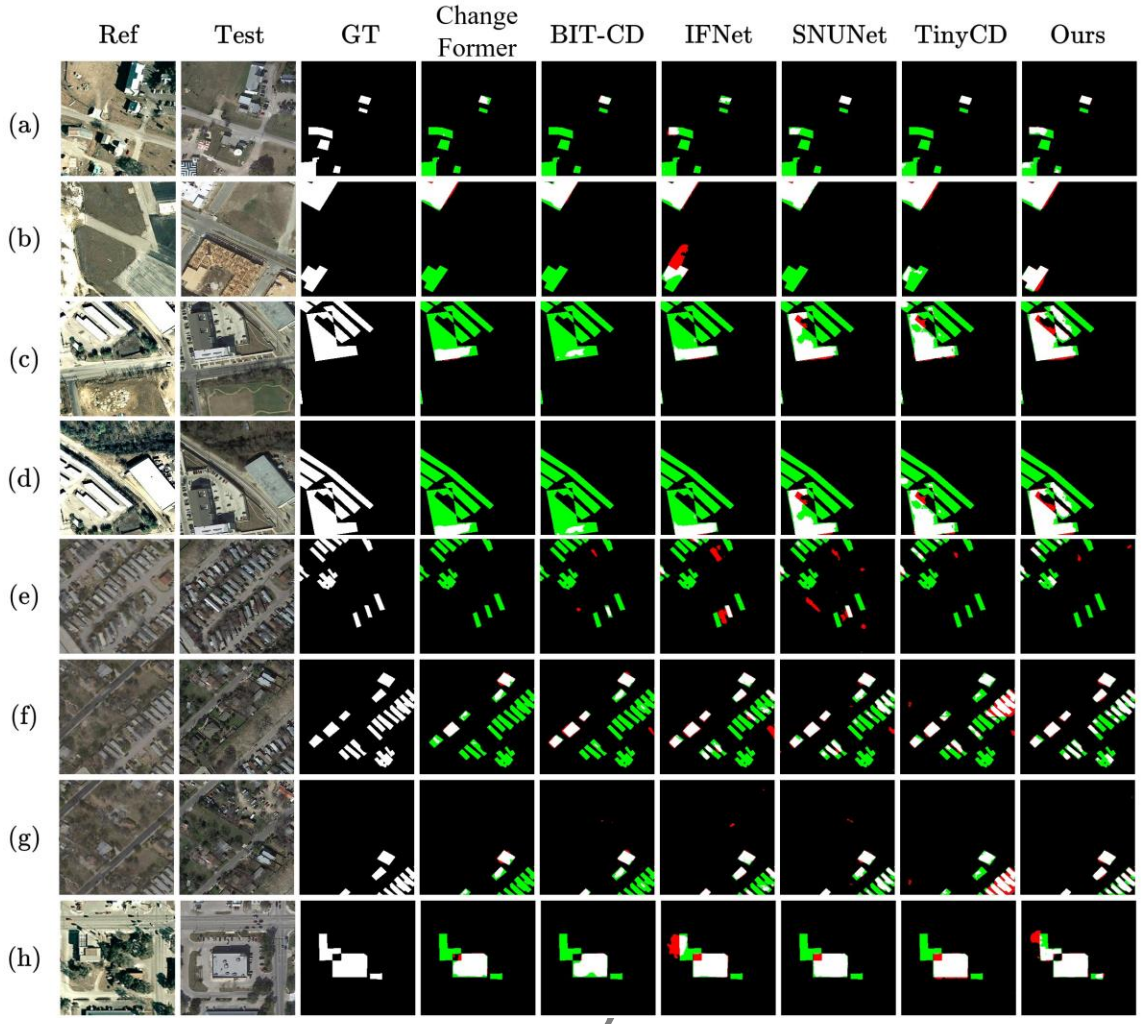

Figure. 10

Tables

Table 1

Method	Pr %	Re %	F1 %	IoU %	OA %
DTCDSSCN [37]	86.99	89.15	87.15	77.10	95.60
ChangeFormer [38]	89.90	86.01	87.91	78.43	95.98
BIT-CD [6]	81.44	74.38	77.75	63.59	92.77
IFNet [31]	91.27	80.99	80.99	75.17	95.45
SNUNet [39]	89.21	87.06	88.13	78.77	96.01
HSSENet [9]	92.52	89.86	91.17	83.77	97.04
TinyCD [40]	72.07	75.44	73.72	58.38	90.86
S2CD [41]	82.31	88.14	84.13	72.57	94.34
RISNet [36]	88.29	89.32	88.81	79.86	92.77
Ours	**92.77**	**94.90**	**93.82**	**88.36**	**97.88**

Table 2

Method	Pr %	Re %	F1 %	IoU %	OA %
DTCDSSCN [37]	93.35	86.82	89.97	81.76	99.23
ChangeFormer [38]	91.74	84.66	88.06	78.66	99.09
BIT-CD [6]	88.99	88.99	86.81	76.69	98.93
IFNet [31]	90.62	88.75	89.67	81.27	99.19
SNUNet [39]	90.04	86.42	88.20	78.89	99.08
HSSENet [9]	94.56	88.73	91.55	84.42	99.35
TinyCD [40]	91.72	91.76	91.74	84.74	99.34
S2CD [41]	92.45	88.74	90.19	82.13	99.27
RISNet [36]	92.33	91.49	91.91	85.03	99.47
Ours	**95.67**	**93.11**	**94.37**	**89.35**	**99.56**

Table 3

Method	Pr %	Re %	F1 %	IoU %	OA %
DTCDSSCN [37]	91.58	88.82	90.18	82.12	99.02

Method	Pr %	Re %	F1 %	IoU %	OA %
ChangeFormer [38]	90.80	88.56	89.67	81.27	98.96
BIT-CD [6]	92.34	87.71	89.96	81.76	99.00
IFNet [31]	91.78	89.75	90.61	82.84	99.06
SNUNet [39]	91.92	89.98	90.94	83.38	99.09
HSSENet [9]	92.63	90.35	91.48	84.29	99.14
TinyCD [40]	92.68	89.47	91.05	83.57	99.10
S2CD [41]	92.12	89.08	90.58	82.78	99.06
RISNet [36]	92.66	90.83	91.74	84.73	99.25
Ours	**92.72**	**90.99**	**91.85**	**84.92**	**99.18**

Table 4

Method	Pr %	Re %	F1 %	IoU %	OA %
ChangeFormer [38]	70.91	72.01	71.65	55.83	91.66
BIT-CD [6]	75.50	72.60	74.11	58.93	91.83
MSNet [33]	**76.79**	74.73	75.74	59.80	93.03
Ours	76.65	**75.73**	**76.19**	**61.54**	**93.05**

Table 5

Method	Param (M)	GFLOPs	Inference Time (ms)	F1 (%) (DSIFN)	F1 (%) (WHU)
DTCDSSCN [37]	41.07	7.21	12.58	87.15	89.97
ChangeFormer [38]	42.02	202.87	74.46	87.91	88.06
BIT-CD [6]	3.55	4.35	8.83	77.75	86.81
IFNet [31]	50.71	41.18	36.44	80.99	89.67
SNUNet [39]	12.03	27.44	32.76	88.13	88.20
HSSENet [9]	10.95	17.14	13.03	91.17	91.55
TinyCD [40]	0.28	1.45	6.93	73.72	91.74
S2CD [41]	3.22	9.26	6.33	84.13	90.19
RISNet [36]	11.7	26.35	9.69	88.81	91.91
Ours	6.8	11.04	16.33	**93.82**	**94.37**

Table 6

Configuration	DSIFN-CD F1 / IoU / OA (%)	WHU F1 / IoU / OA (%)
Spatial-all	93.13 / 87.15 / 97.59	94.20 / 89.04 / 99.55
Channel-all	93.38 / 87.59 / 97.72	94.23 / 89.10 / 99.55
Spatial2-Channel3	93.35 / 87.52 / 97.68	93.66 / 88.08 / 99.50
Spatial3-Channel2	93.82 / 88.36 / 97.88	94.37 / 89.35 / 99.56
Channel3-Spatial2	93.48 / 87.76 / 97.74	94.20 / 89.03 / 99.55
Channel2-Spatial3	92.84 / 86.63 / 97.48	94.21 / 89.05 / 99.55

Table 7

Modulation Strategy	DSIFN-CD F1 (%)	WHU-CD F1 (%)
No Modulation	93.37	93.89
Single Dist. + Feat. Mod.	92.83	94.05
Dual Dist. + Feat. Mod.	92.76	94.12
Dual Dist. + Attn. Mod.	93.39	94.20
Single Dist. + Attn. Mod. (Ours)	**93.82**	**94.37**

Table 8

β Value	DSIFN-CD F1 (%)	WHU-CD F1 (%)
0.0	80.26	93.99
0.0001	92.90	94.15
0.001	**93.82**	**94.37**
0.01	93.46	93.95

Biographies

Mostafa Etemadinia received the B.S. and M.S. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran. He is currently pursuing the Ph.D. degree with the Amirkabir University of Technology, Tehran, Iran. His research interests include computer vision, and machine learning, especially real-time machine learning and its applications in remote sensing.

Dr. Saeed Sharifian received his B.Sc. degree in electrical engineering from KNT University of Technology, Tehran, Iran, in 2000, and M.Sc. and Ph.D. degrees in digital electronic engineering from Amirkabir University of Technology, Tehran, Iran, in 2002 and 2008, respectively. He was vice chancellor for research and development of Iranian High-Performance Computing Research Center (HPCRC) from 2009 up to 2015. He is currently an Associate professor in the department of electrical

engineering, Amirkabir University of Technology, Tehran, Iran. His research interests include Artificial Intelligence, Deep learning, IoT, Cloud and Edge Computing, Big data analytics and Soft Computing.