Feature Spectrum Learning for Remote Sensing Change Detection

Qi Zang[1*], Dong Zhao[1*], Shuang Wang[1✉], Dou Quan[1], Zhun Zhong[2]
[1] School of Artificial Intelligence, Xidian University, Shaanxi, China
[2] School of Computer Science and Information Engineering, Hefei University of Technology, China
shwang@mail.xidian.edu.cn

Abstract

Change detection (CD) holds significant implications for Earth observation, in which pseudo-changes between bitemporal images induced by imaging environmental factors are key challenges. Existing methods mainly regard pseudo-changes as a kind of style shift and alleviate it by transforming bitemporal images into the same style using generative adversarial networks (GANs). Nevertheless, their efforts are limited by the complexity of optimizing GANs and the absence of guidance from physical properties. This paper finds that the spectrum transformation (ST) has the potential to mitigate pseudo-changes by aligning in the frequency domain carrying the style. However, the benefit of ST is largely constrained by two drawbacks: 1) limited transformation space and 2) inefficient parameter search. To address these limitations, we propose a ***Fea****ture* ***Spect****rum learning (FeaSpect) that adaptively eliminate pseudo-changes in the latent space. For the drawback 1), FeaSpect directs the transformation towards style-aligned discriminative features via feature spectrum transformation (FST). For the drawback 2), FeaSpect allows FST to be trainable, efficiently discovering optimal parameters via extraction box with adaptive attention and extraction box with learnable strides. Extensive experiments on challenging datasets demonstrate that our method remarkably outperforms existing methods and achieves a commendable trade-off between accuracy and efficiency. Importantly, our method can be easily injected into other frameworks, achieving consistent improvements.*

1. Introduction

As one of the fundamental Earth observation tasks, change detection (CD) utilizes bitemporal remote sensing data to achieve dynamic understanding of the Earth's surface and provides pixel-wise ground object change maps for land use survey, urban expansion, and ecological environment monitoring [4, 17, 25–27, 42, 43, 49, 55–58]. Driven by deep learning (DL) and dominated by the siamese structure [3], change detection has gained tremendous progress in recent years [1, 31].

Nevertheless, challenges persist in handling pseudo-changes between bitemporal images induced by variations in the imaging environment (sun elevations or seasons) [8, 23, 48, 60], as shown in Fig. 1(a). This issue significantly hampers the network's ability to accurately recognize actually changed pixels. The fundamental cause of pseudo-changes is the disagreement between sensor-measured values and the object's spectral reflectance and radiance [13, 41], *i.e.*, spectral shift. Existing transformation-based methods regard this spectral shift as essentially analogous to a visual appearance or style shift. These methods alleviate this issue mainly by leveraging generative adversarial networks (GANs) to transform bitemporal image styles into the same one [14, 15, 32]. However, these efforts are limited by the complexity of optimizing GANs and the absence of guidance from physical properties, leading to transformed images susceptible to distortion from artifacts (See the orange rectangles in Fig. 1(b)). Inspired by the recently proven effective spectrum transformation (ST), a frequency-domain carrying style-oriented style alignment method [46], we explore its potential application for the first time in the CD task to mitigate pseudo-changes. We deploy the ST to the CD task by replacing the low-frequency (cropped with fixed extraction box) Fourier amplitude spectrum of an image with that of its paired image. Compared with GANs, ST attains better visual consistency with targeted low-level statistics (See Fig. 1(c)) and yields more accurate detection performance (See Table 3).

However, as shown in Fig. 1(d), there have two drawbacks: *1) Limited Transformation Space.* The transformation space of ST aims for visual consistency of images, while the optimization space for high-level perception tasks focuses on extracting discriminative features of images. This mismatched learning goal causes the feature to lose essential semantic information (See the pink rectangles in Fig. 1(a)-(c)), consequently diminishing discriminabil-

* Equal contribution. ✉ Co-corresponding author.

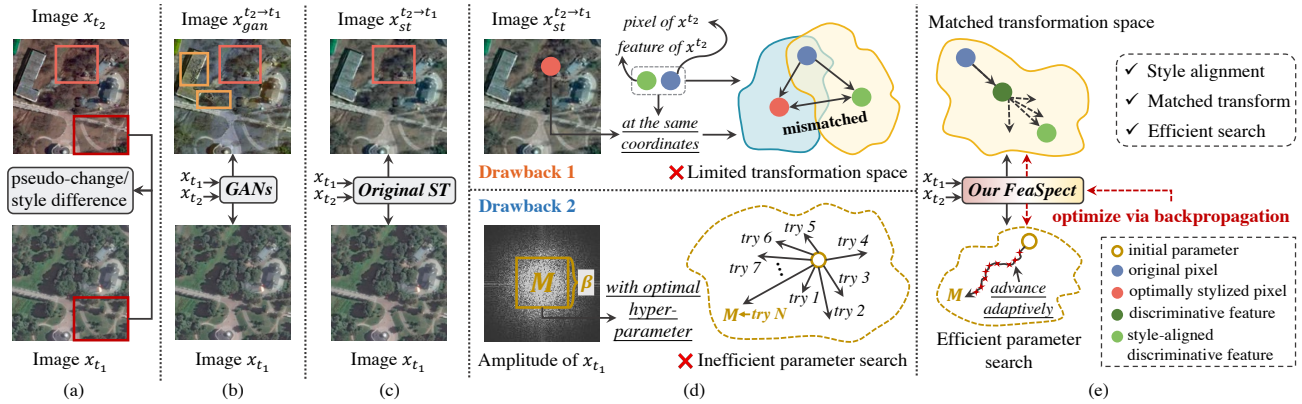

Figure 1. (a) Illustration of pseudo-change. (b) Style-aligned images obtained by *GANs*. (c) Style-aligned images obtained by *spectrum transformation (ST)*. (d) Illustration of two drawbacks of ST. (e) Illustration of our *FeaSpect* solving these two drawbacks.

ity. *2) Inefficient Parameter Search.* The hyper-parameter of the fixed extraction box in ST, *i.e.*, the non-differentiable stride, necessitates cross-validation to identify its optimal value. This empirical manual search in a gigantic search space hinders efficient discovery of optimal values, leading to imprecise style shielding.

This paper aims to tackle the above two drawbacks of the ST, unlocking its promising potential in the CD. To this end, we present a ***Fea**ture **Spect**rum learning (FeaSpect)* approach (See Fig. 1(e)), aiming to enable the model to adaptively eliminate the style differences in the latent space. Based on the identified challenges, *i.e.*, how to matched transform and how to efficient search, we proceed as follows: For the former, we formulate the ST into the feature space to align the styles between high-dimensional bitemporal features encoding essential semantics, called *feature spectrum transformation (FST)*, directing the transformation toward style-aligned discriminative features. For the latter, we devise two learning mechanisms for FST from two aspects, *extraction box with adaptive attention (EB-AA)* and *extraction box with learnable strides (EB-LS)*, to capture style characterization via gradient descent at a lower computational cost. The extraction box is not fixed by hyper-parameter but is trainable, allowing backpropagation to uncover optimal characterization regions in a virtually infinite search space. In the EB-AA, based on a specific stride, the amplitude values within the extraction box are adaptively focused to avoid extracting less informative amplitude values. In the EB-LS, we propose learning the strides of the extraction box along the horizontal and vertical directions as well as customize a way to make strides differentiable during learning.

In summary, the contributions of this paper are threefold:

- We exploit spectrum transformation for the first time to alleviate the pseudo-changes in CD and indicate two drawbacks of deploying it. To tackle these two drawbacks, we propose a framework FeaSpect, adaptively eliminating the style differences in the feature space.
- We propose a feature spectrum transformation (FST) to maintain the semantics by aligning styles across bitemporal features. Further, we devise two learning mechanisms for FST, extraction box with adaptive attention/learnable strides, to efficiently find optimal configurations by gradient descent.
- Extensive experiments show that FeaSpect significantly surpasses existing methods and achieves a commendable trade-off between accuracy and efficiency. Moreover, our FeaSpect is a plug-and-play design, which can seamlessly integrates into any Siamese-dominated network, boosting performance without the need for additional parameters.

2. Related Work

Deep Learning-based Change Detection. Remote sensing change detection (CD) using deep convolutional neural networks [28] has garnered substantial attention due to its ability to automatically learn hierarchical features from bitemporal images. Various existing methods in this domain leverage the power of deep learning to alleviate the overdetection of pseudo-changes [1, 12, 18, 31, 36, 44]. Benefiting from the excellent performance of fully convolutional network (FCN) [35] on the segmentation task, many researchers apply it to the change detection [5, 10, 11, 30, 38, 52]. Daudt and Lei *et al.* [11, 30] first introduce the U-Net-based FCN architecture. Peng *et al.* [38] propose an end-to-end approach based on the improved UNet++ network and employ a multi-side output fusion strategy to combine change maps from different semantic levels. Daudt *et al.* [10] propose three fully convolutional siamese neural network structures with skip connections, *i.e.*, FC-EF, FC-Siam-conc, and FC-Siam-diff, obtaining multi-scale change maps by calculating the feature of the two weight-shared encoders at different sizes and decode them to obtain the de-

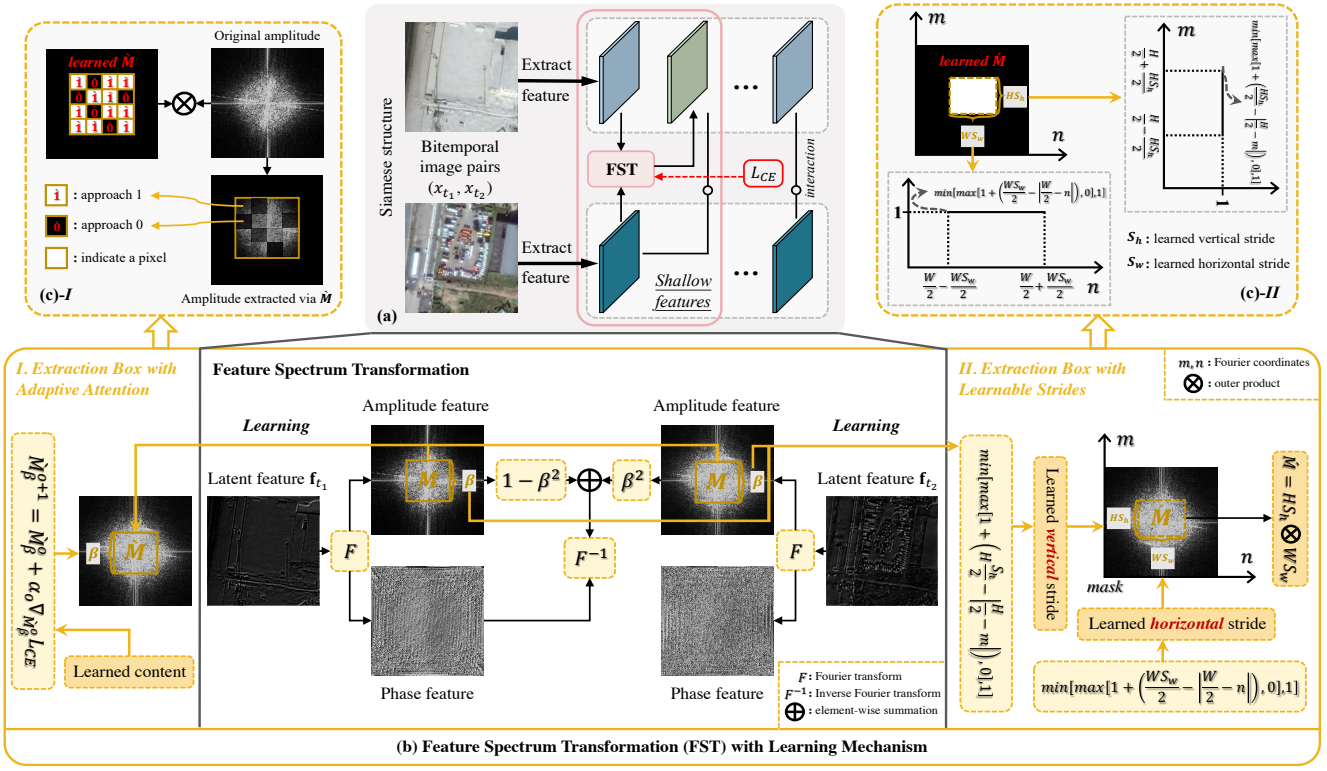

Figure 2. An overview of the proposed FeaSpect. (a) Illustration of the feature spectrum transformation (FST) embedded into the feature space. Only the first layer features are used as an example to demonstrate FST. FST is optimized during training via loss L_{CE}. (b) FST with two learning mechanisms aligns the styles of features of bitemporal images. (c) Illustration of two learning mechanisms learning parameters during training.

tection result. The advancements in existing works primarily revolve around designing complex network structures to enhance feature extraction capabilities [7, 16, 39, 54], optimizing loss functions to address sample imbalance [16, 50], incorporating various attention mechanisms to bolster feature discriminability [5, 7, 34]. Besides, there are several methods focusing on eliminating the style shift between bitemporal images, leveraging generative adversarial networks (GANs) to transform bitemporal image styles into the same one [14, 15, 32]. However, their efforts are limited as GANs is difficult to train and prone to cause the transformed image distortion. In this paper, we explore the potential of spectral transformation deployed in CD to mitigate pseudo-changes by aligning feature styles and address two of its drawbacks to align efficiently without compromising semantics.

Spectral Transformation for Cross-domain Task. Spectrum transformation has been identified as a viable approach to alleviate domain shift induced by imaging mechanisms. Yang *et al.* [46, 47] pioneer the application of this discovery to visual recognition tasks, aligning spectral distributions of source and target domains through Fourier amplitude spectrum transformation in the image space, yielding effective benefits in cross-domain visual recognition tasks. Building upon this, several works [29, 33, 45, 53] based on spectrum transformation have been proposed for mitigating style shifts caused by different imaging devices in cross-domain segmentation [20] and detection [21] tasks. Motivated by these developments, we first introduce spectrum transformation into the change detection task, aiming to alleviate the issue of pseudo-change induced by spectral (style) shifts between bitemporal data. In comparison to existing spectrum transformation methods, we innovate by presenting trainable feature spectrum learning, addressing the drawbacks of spectrum transformation technique in handling diverse temporal spectral shifts.

3. Proposed Method

3.1. Background and Overview

Siamese Structure in Change Detection [3]. Given a bitemporal image pair (x^{t_1}, x^{t_2}), it is fed into a two-branch feature extractor with shared structure and weights. Then, the extracted features are classified to obtain a change map

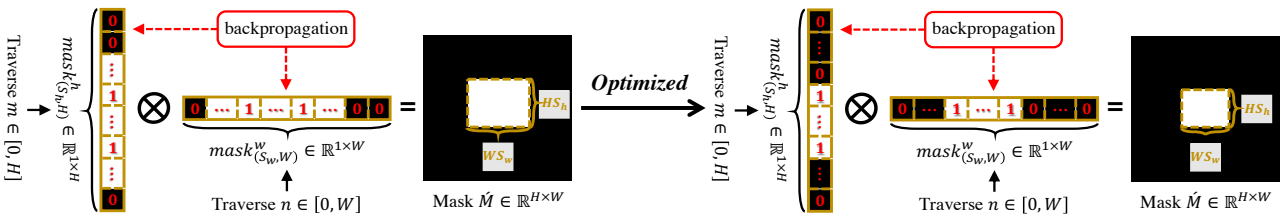

Figure 3. Illustration of optimizing strides (Eq. (7)) in the EB-LS module.

$\mathbf{P}_{out}$,

$$\mathbf{P}_{out} = h(\mathbf{f}_{t_1}(x^{t_1}) \circ \mathbf{f}_{t_2}(x^{t_2})),$$
$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{j=1}^{N}[y_j \log(\mathbf{P}^j_{out}) + (1-y_j)\log(1-\mathbf{P}^j_{out})], \tag{1}$$

where h is the classifier, $\mathbf{f}_{t_1}$ and $\mathbf{f}_{t_2}$ are the features, and $\circ$ is an interactive operation. $\mathcal{L}_{CE}$ is the binary cross-entropy loss function, N is the number of image pairs in a mini-batch and y_j is the ground truth for the j-th image pair. In this way, variations caused by environmental factors are also encoded in the features, so the model is prone to detect these as changes, *i.e.*, pseudo-changes. The pseudo-changes directly affect the calculation in the $\circ$ operation, leading to the misjudgment of the h. More analysis is in the Supplementary 2.

Spectrum Transformation in Image Space [46]. In domain adaptation, given images x_s and x_t from different domains, a fast Fourier transform $\mathcal{F}$ is performed on them to obtain a style-representable amplitude spectrum $\mathcal{F}_A$ and a content-representable phase spectrum $\mathcal{F}_P$. Then, aligning styles between two images can be formalized as,

$$x^{s\to t} = \mathcal{F}^{-1}((M_\beta \mathcal{F}_A(x_t) + (1-M_\beta)\mathcal{F}_A(x_s), \mathcal{F}_P(x_s)),$$
$$M_\beta = \mathbb{1}_{(h,w)\in[-\beta H:\beta H, -\beta W:\beta W]}, \tag{2}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform that maps spectral signals (phase and amplitude) back to image space, and M_β is a mask controlled by a hyper-parameter β to extract the low-frequency part of the amplitude. The modified amplitude of x_s, with its phase unaltered, is mapped back to the image $x^{s\to t}$, whose content is the same as x_s, but will resemble the style of the x_t.

Spectrum Transformation in Feature Space. As mentioned before, such a spectrum transformation in the image space has limited transformation space (Drawback 1). To solve this drawback, we introduce this spectrum transformation into the feature space based on the siamese structure to align the style between the $\mathbf{f}_{t_1}$ and $\mathbf{f}_{t_2}$, as shown in Fig. 2. This can ensure that the semantics are encoded by high-dimensional features while modifying the amplitude spectrum, *i.e.*, *feature spectrum transformation (FST)*. In addition, the styles mainly manifest in the shallow layers of the network while the contents mainly manifest in its deep layers [37]. Given this fact, the FST is only embedded into the shallow layers of the feature extractor, which further matches the transformation direction with the stated goal (style-aligned discriminative feature).

Depart from drawback 1, the FST is inefficient to obtain low-frequency amplitude spectrum by a fixed extraction box (Drawback 2). We develop two learning mechanisms for the FST, *extraction box with adaptive attention (EB-AA) and extraction box with learnable strides (EB-LS)*, to allow it to be learnable. The following Sections 3.2 and 3.3 describe these two mechanisms in detail.

3.2. Extraction Box with Adaptive Attention

The values in the region cropped from the amplitude features by the fixed extraction box do not necessarily all contribute to style representation. Adopting less informative values as replacement objects will weaken the effectiveness of style alignment. Therefore, we devise the extraction box with adaptive attention (EB-AA) for the FST to adaptively focus on those more informative values within the extraction boxes through network optimization.

In our implementation, we set the mask M_β controlled by the hyper-parameter β to be optimizable during training. We apply an open gradient operator to the M_β, allowing gradients to flow to this mask. Through the calculation of the loss function $\mathcal{L}_{CE}$ and the backpropagation of the errors, not only the network parameters θ are optimized, but the values in the M_β are also adjusted. This end-to-end optimization enables the learned content of M_β to be well-suited for the detection task in this study, which is formulated as,

$$\grave{M}^{o+1}_\beta = \grave{M}^{o}_\beta + \alpha_o \nabla_{\grave{M}^{o}_\beta} \mathcal{L}_{CE}(\mathbf{f}_{t_1}, \mathbf{f}_{t_2}, \grave{M}^{o}_\beta, \theta^o), \tag{3}$$

where α_o is the learning rate and $\grave{M}^{o+1}_\beta$ is the iteratively updated mask. With continuous optimization, the values of less informative regions in the mask $\grave{M}_\beta$ gradually approach 0 while the values of more informative regions approach 1. This compels the feature extractor to dynamically identify amplitude values with strong style representation, effectively filtering out redundant information and avoiding inconsequential replacements. More details can be found in the Supplementary 1.2.

3.3. Extraction Box with Learnable Strides

With the above optimized process, the extraction box of the amplitude is learnable for its internal values. However, this process is only differentiable with respect to these contents but not with respect to the strides (*i.e.*, β) of the control representation region. Thus, one still needs to provide β as hyper-parameters for the FST. To further improve efficiency, we propose the extraction box with learnable strides (EB-LS) for the FST, allowing to learn the size of the box via backpropagation instead of applying a box with a fixed size.

We design the learnable box $\acute{M}$ as the outer product between two strides $(S_h, S_w) \in [1, H) \times [1, W)$, each of which alongs with the horizontal and vertical axises respectively. S_h and S_w are defined by two differentiable 1D masking functions,

$$mask^{h}_{(S_h,H)}(m) = min\Big[max\Big[1+(\frac{S_h}{2}-|\frac{H}{2}-m|), 0\Big], 1\Big],$$
$$mask^{w}_{(S_w,W)}(n) = min\Big[max\Big[1+(\frac{S_w}{2}-|\frac{W}{2}-n|), 0\Big], 1\Big], \tag{4}$$

where $m \in [0, H]$ and $n \in [0, W]$ are the integers being traversed to determine the length of the mask vector. These two functions symmetrically set the stride around the center, with values of 1 in the middle interval and 0 on both sides (See Fig. 2*(c)-II*). As shown in Fig. 3, based on the two 1D functions, we build the 2D differentiable mask $\acute{M} \in \mathbb{R}^{H\times W}$,

$$\acute{M}(S_h, S_w, H, W) = mask^{h}_{(S_h,H)} \otimes mask^{w}_{(S_w,W)}, \tag{5}$$

which is parameterized by the shape of the input latent features and the strides. $\otimes$ is the outer product. Through backpropagation, the model is forced to learn the attention span for the amplitude features to find the optimal configuration,

$$\acute{M}^{o+1} = \acute{M}^{o} + \alpha_o \nabla_{S_h,S_w} \mathcal{L}_{CE}(\mathbf{f}_{t_1}, \mathbf{f}_{t_2}, \acute{M}^{o}, \theta^{o}). \tag{6}$$

During training, the amplitude feature is a multi-channel input, and all channels are required to learn different strides due to the diversity of high-dimensional representations. In spectrum transformation, strides are typically tied along the spatial axes ($S_h = S_w$), and EB-LS can also do by sharing one learnable stride on both dimensions. We discuss this in Table 6, but the results show that learning different strides along both dimensions is more beneficial for our task.

Moreover, in Eq. (6), the optimization gradient required by $\acute{M}^{o}$ and the gradient returned by $\mathcal{L}_{CE}$ do not match in magnitude, as the batch normalization in the neural network restricts the gradient's magnitude. Thus, to improve the optimization rate of our mechanism, we extend the two masking functions (Eq. (4)) corresponding to S_h and S_w as follows,

$$mask^{h}_{(S_h,H)}(m) = min\Big[max\Big[1+(\frac{HS_h}{2}-|\frac{H}{2}-m|), 0\Big], 1\Big],$$
$$mask^{w}_{(S_w,W)}(n) = min\Big[max\Big[1+(\frac{WS_w}{2}-|\frac{W}{2}-n|), 0\Big], 1\Big]. \tag{7}$$

In Eq. (7), we introduce scaling factors H and W to expand the strides of the $\acute{M}$ to HS_h and WS_w. As a result, the search scope for S_h and S_w is greatly narrowed. Unless otherwise specified, all FeaSpect in the experiments refer to the extended optimization version. More details can be found in the Supplementary 1.1.

Method	Backbone	Prec.(%)	Rec.(%)	F1(%)	IoU(%)	Params.
• *ConvNet-based:*						
FC-EF [10]	U-Net*	77.48	88.13	82.46	71.15	1.35M
FC-Siam-conc [10]	U-Net*	72.88	89.21	80.22	67.98	1.55M
FC-Siam-diff-res [10]	U-Net*	82.91	89.37	86.02	75.48	1.35M
FCN-PP [30]	U-Net*	83.05	91.80	87.21	77.31	27.81M
W-Net [19]	U-Net*	85.25	92.90	88.91	80.04	40.49M
CDGAN [19]	U-Net*	89.05	89.47	89.26	80.60	115.12M
STANet [9]	ResNet-18*	89.24	93.51	91.33	84.04	16.93M
DSAMNet [40]	ResNet-18*	93.35	92.41	92.90	86.60	17.00M
ESCNet [52]	ResNet-18*	90.03	95.82	92.83	86.63	5.12M
SEIFNet [22]	ResNet-18*	92.15	93.07	92.61	86.23	27.91M
ChangeSTAR [59]	ResNet-101*	95.41	94.10	94.75	90.02	52.60M
FeaSpect (ours)	U-Net*	**95.75**	**95.87**	**95.81**	**91.96**	1.35M
FeaSpect (ours)	ResNet-18*	**96.23**	**95.94**	**96.08**	**92.46**	5.12M
FeaSpect (ours)	ResNet-101*	**96.62**	**96.01**	**96.31**	**92.89**	52.60M
◦ *Transformer-based:*						
BIT [6]	ViTAEv2-S*	94.85	93.91	94.38	89.35	3.55M
VcT [24]	ViTAEv2-S*	94.92	93.01	93.96	88.60	3.59M
SwinsUNet [51]	Swin-Trans*	95.12	93.84	94.48	89.53	41.25M
ChangeFormer [2]	MiT-B2*	95.19	93.61	94.39	89.38	41.01M
FeaSpect (ours)	ViTAEv2-S*	**96.80**	**96.13**	**96.46**	**93.17**	3.55M
FeaSpect (ours)	MiT-B1	**97.10**	**96.65**	**96.87**	**93.94**	13.70M
FeaSpect (ours)	MiT-B2*	**97.41**	**97.13**	**97.27**	**94.68**	41.01M
FeaSpect (ours)	Swin-Trans*	**97.49**	**97.18**	**97.33**	**94.81**	41.25M

Table 1. Comparison results on SVCD dataset. */* defines the backbone model modified in different/same ways.

4. Experiments

4.1. Datasets and Experimental Setup

The details of all experiments, including the used datasets, evaluation protocols, implementation details, and the used network are demonstrated in the Supplementary 3.

4.2. Comparative Experiments

Comparison on SVCD Dataset. We report the comparison results of FeaSpect and different methods on SVCD dataset in Table 1, including *ConvNet-based* and *Transformer-based* methods. For ConvNet-based methods, FeaSpect using small backbone U-Net achieves the better performance and surpasses the second-highest results by 0.65%, 0.31%, 2.15%, and 3.94% in terms of precision, recall, F1-score, and IoU, demonstrating its better generalization. Benefiting from the multi-scale long-range details of transformer, FeaSpect using different transformer backbones further achieves new records and surpasses existing transformer-based methods with smaller network parameters.

Method	Backbone	Prec.(%)	Rec.(%)	F1(%)	IoU(%)	Params.
• *ConvNet-based:*						
FC-EF [10]	U-Net*	52.64	68.54	59.55	42.40	1.35M
FC-Siam-conc [10]	U-Net*	64.27	54.43	58.94	41.79	1.55M
FC-Siam-diff-res [10]	U-Net*	57.47	67.23	61.97	44.90	1.35M
FCN-PP [30]	U-Net*	59.09	61.35	60.20	43.06	27.81M
W-Net [19]	U-Net*	63.01	59.72	61.32	44.22	40.49M
CDGAN [19]	U-Net*	60.45	62.50	61.46	44.36	115.12M
STANet [9]	ResNet-18*	66.89	63.06	64.92	48.06	16.93M
DSAMNet [40]	ResNet-18*	67.72	64.51	66.08	49.34	17.00M
ESCNet [52]	ResNet-18*	66.99	67.21	67.10	50.49	5.12M
SEIFNet [22]	ResNet-18*	67.68	68.43	68.05	51.58	27.91M
ChangeSTAR [59]	ResNet-101*	69.12	69.71	69.41	53.16	52.60M
FeaSpect (ours)	U-Net*	**69.47**	**70.27**	**69.87**	**53.69**	1.35M
FeaSpect (ours)	ResNet-18*	**70.35**	**71.43**	**70.89**	**54.90**	5.12M
FeaSpect (ours)	ResNet-101*	**72.20**	**72.01**	**72.10**	**56.38**	52.60M
○ *Transformer-based:*						
BIT [6]	ViTAEv2-S*	67.10	70.61	68.81	52.45	3.55M
VcT [24]	ViTAEv2-S*	65.31	73.98	69.38	53.11	3.59M
SwinsUNet [51]	Swin-Trans*	71.04	74.20	72.59	56.97	41.25M
ChangeFormer [2]	MiT-B2*	88.19	84.81	86.47	76.16	41.01M
FeaSpect (ours)	ViTAEv2-S*	**70.19**	**71.72**	**70.95**	**54.97**	3.55M
FeaSpect (ours)	MiT-B1	**70.83**	**72.01**	**71.42**	**55.54**	13.70M
FeaSpect (ours)	Swin-Trans*	**73.82**	**75.03**	**74.42**	**59.26**	41.25M
FeaSpect (ours)	MiT-B2*	**88.90**	**85.36**	**87.09**	**77.14**	41.01M

Table 2. Comparison results on DSIFN dataset. */* defines the backbone model modified in different/same ways.

Method	Modeling Space	Modeling Way	Prec.(%)	Rec.(%)	F1(%)	IoU(%)
Baseline [10]	-	-	82.91	89.37	86.02	75.48
+ DLSF† [14]	Image	GT	88.70	89.90	89.30	80.66
+ FDA† [46]	Image	ST	89.23	91.49	90.35	82.39
+ IBN-Net† [37]	Feature	FN	91.98	92.75	92.36	85.81
+ FeaSpect (ours)	Feature	FSL	**95.75**	**95.87**	**95.81**	**91.96**

Table 3. Results on SVCD dataset. "GT" is generated transformation. "ST" is spectrum transformation. "FN" is feature normalization. "FSL" is feature spectrum learning. † defines the reproduced method.

Comparison on DSIFN Dataset. Table 2 reports the comparison results on the DSIFN dataset. Within ConvNet-based methods, we can obtain consistent observations that our FeaSpect improves detection performance by learning more robust feature representations. Among transformer-based methods, a noteworthy trend is observed wherein models with larger backbones significantly outperform others. Furthermore, in comparison to ConvNet-based, FeaSpect with transformer backbones obtains a substantial performance leap. This can be attributed to the capability of larger model to capture intricate feature relationships, ensuring a robust discriminative representation. This advantage is particularly evident in this complex scene due to the required generalizability, where training and testing data are from different locations or datasets (See the Supplementary 6 for more details on generalizability analysis).

Comparison on More Datasets and Tasks. To further verify the effectiveness and scalability of our method, we provide comparative results on other datasets and other tasks in the Supplementary 7 and 8.

Comparison with Domain Shift-Aware Methods. Table 3 presents the comparison between our method and other related methods focusing on domain shifts in change detection, domain adaptation or domain generalization. For change detection, we compare with DLSF, which uses generative adversarial networks (GANs) to transform image styles. For domain adaptation, we compare with FDA [46], *i.e.*, which aligns image styles via the spectrum transformation. For domain generalization, we compare with classic IBN-Net [37], which focuses on removing style by normalizing features. For fair comparison, all methods use FC-Siam-diff-res [10] as the baseline. Compared with the baseline model, the consistent improvement in detection performance across the four methods underscores the necessity of eliminating the style differences between bitemporal images. DLSF obtains unsatisfactory performance due to difficulty in GANs training and possible image distortion. FDA achieves limited performance gains, primarily because it inevitably disrupts essential semantics and may have suboptimal parameters. IBN-Net obtains more performance boosts than FDA. Notably, our method outperforms IBN-Net in accuracy, facilitated by the interaction between bitemporal features and the decoupling of style information, resulting in improved style alignment and semantic preservation. This suggests that a careful exploration of modeling approaches to eliminate style differences can further improve representation discriminability.

Method	Prec.(%)	Rec.(%)	F1(%)	IoU(%)	Δ FLOPs
• *ConvNet-based:*					
FC-EF [10]	77.48	88.13	82.46	71.15	–
+ FeaSpect	**95.53** (+18.05)	**95.98** (+7.85)	**95.75** (+13.29)	**91.85** (+20.70)	0.90G
FC-Siam-conc [10]	72.88	89.21	80.22	67.98	–
+ FeaSpect	**95.25** (+22.37)	**95.80** (+6.59)	**95.52** (+15.30)	**91.43** (+23.45)	0.90G
FCN-PP [30]	83.05	91.80	87.21	77.31	–
+ FeaSpect	**96.00** (+12.95)	**96.12** (+4.32)	**96.06** (+8.85)	**92.42** (+15.11)	3.62G
STANet [9]	89.24	93.51	91.33	84.04	–
+ FeaSpect	**96.85** (+7.61)	**96.88** (+3.37)	**96.86** (+5.53)	**93.92** (+9.88)	3.69G
DSAMNet [40]	93.35	92.41	92.90	86.60	–
+ FeaSpect	**97.42** (+4.07)	**96.30** (+3.89)	**96.86** (+3.96)	**93.91** (+7.31)	3.76G
ESCNet [52]	90.03	95.82	92.83	86.63	–
+ FeaSpect	**97.21** (+7.18)	**97.30** (+1.48)	**97.25** (+4.42)	**94.66** (+8.03)	1.78G
ChangeSTAR [59]	95.41	94.10	94.75	90.02	–
+ FeaSpect	**98.11** (+2.70)	**97.90** (+3.80)	**98.00** (+3.25)	**96.09** (+6.07)	14.26G
○ *Transformer-based:*					
BIT [6]	94.85	93.91	94.38	89.35	–
+ FeaSpect	**97.59** (+2.74)	**97.02** (+3.11)	**97.30** (+2.92)	**94.75** (+5.40)	3.81G
SwinsUNet [6]	95.12	93.84	94.48	89.53	–
+ FeaSpect	**98.02** (+2.90)	**97.71** (+3.87)	**97.86** (+3.38)	**95.82** (+6.29)	2.32G
ChangeFormer [2]	95.19	93.61	94.39	89.38	–
+ FeaSpect	**97.99** (+2.80)	**97.86** (+4.25)	**97.92** (+3.53)	**95.93** (+6.55)	2.88G

Table 4. Results of existing methods incorporating our method on SVCD dataset.

4.3. Compatibility of Our Method

Our method can serve as a complementary enhancement to existing methods in change detection, seamlessly integrating to boost their performance. To validate this, we incorporate the proposed feature spectrum transformation (FST) and its two learning mechanisms extraction box with adaptive attention (EB-AA) and extraction box with learnable strides (EB-LS) into the existing methods. The quantitative results on the SVCD dataset are listed in Table 4. Table 4 reveals that the incorporation of our method consistently improves the performance of these methods. The positive effect of our method can be attributed to its ability to assist other methods in adaptively mitigating model interference caused by style differences within image pairs. Except

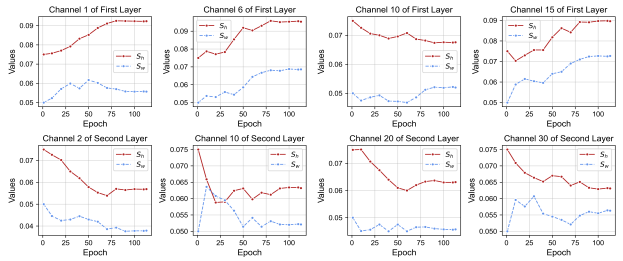

Figure 4. Trajectories of S_h and S_w for different channels in different layers during learning. Configuration (0.05, 0.075) is used as initial value.

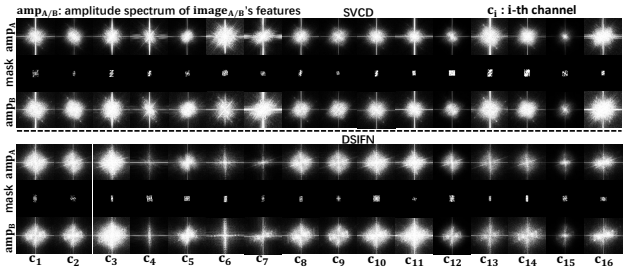

Figure 5. Visualization of different channels of the learned content on different datasets.

for FC-EF and FC-Siam-conc, the methods enhanced with our FeaSpect outperform ours. We attribute this to the fact that these methods fortify the foundational feature representation from other perspectives with their proposed strategies. By integrating our strategies, the generalization of features is further improved, boosting detection performance. Furthermore, it is noteworthy that after incorporating our method, the training process will introduce no additional parameters, and the added computation is negligible.

4.4. Learning Dynamics of Stride and Content

Stride. Fig. 4 illustrates the learning dynamics of the strides of the extraction box in the first two layers of the model. Four randomly selected channels from each layer are presented. During training, all strides deviate from the initial stride value while rapidly converging to the optimal value. Additionally, the learned S_h and S_w values exhibit different trends for various layers (first row↔second row). S_h and S_w learned in shallow layer tend to be larger, while the opposite is observed in deep layer.

Content. In the Fig. 5 (please zoom in), we visualize the learned content of the extraction boxes for each channel on SVCD and DSIFN datasets. Our EB-AA adaptively searches for valuable values in low-frequency amplitudes with the guidance of the learned strides by each channel. Besides, EB-AA preserves varying amounts of information for different datasets due to the distinct style shifts in each.

Model	FST	EB-AA	EB-LS	Prec.(%)	Rec.(%)	F1(%)	IoU(%)
Baseline [10]				82.91	89.37	86.02	75.48
FeaSpect-fixed	✓			92.20	93.86	93.02	86.96
FeaSpect-AA	✓	✓		93.90	95.59	94.74	90.00
FeaSpect-LS	✓		✓	95.01	95.65	95.33	91.07
FeaSpect (ours)	✓	✓	✓	**95.75**	**95.87**	**95.81**	**91.96**

Table 5. Ablation studies for each component of our method on SVCD dataset.

Initial Strides	Original ST [46]				FeaSpect (ours)			
	Prec.(%)	Rec.(%)	F1(%)	IoU(%)	Prec.(%)	Rec.(%)	F1(%)	IoU(%)
(0.025, 0.025)	87.57	95.33	91.28	83.97	95.74	95.61	95.67	91.71
⌜(0.025, 0.025)⌟	–	–	–	–	94.00 (↓1.74)	95.13 (↓0.48)	94.56 (↓1.11)	89.68 (↓2.03)
(0.025, 0.05)	**89.20**	**94.50**	**91.76**	**84.78**	95.72	95.86	95.79	91.92
(0.05, 0.05)	88.99	94.35	91.59	84.49	95.66	95.89	95.77	91.89
⌜(0.05, 0.05)⌟	–	–	–	–	93.94 (↓1.72)	95.95 (↑0.06)	94.93 (↓0.84)	90.36 (↓1.53)
(0.05, 0.075)	86.78	94.03	90.26	82.25	**95.75**	95.87	**95.81**	**91.96**
(0.075, 0.075)	86.29	94.57	90.24	82.22	95.71	95.79	95.75	91.85
⌜(0.075, 0.075)⌟	–	–	–	–	94.59 (↓1.12)	94.96 (↓0.83)	94.77 (↓0.98)	90.07 (↓1.78)
(0.05, 0.075)	87.41	94.42	90.78	83.12	95.70	95.84	95.77	91.88
(0.10, 0.10)	85.99	94.20	89.91	81.67	95.69	**95.90**	95.79	91.93
⌜(0.10, 0.10)⌟	–	–	–	–	94.78 (↓0.91)	95.11 (↓0.79)	94.94 (↓0.85)	90.38 (↓1.55)
Mean Accuracy	87.46	94.49	90.83	83.21	95.71	95.82	95.76	91.88

Table 6. Results on SVCD dataset with different initial strides. (·, ·) is the initial S_h and S_w values. ⌜⌟ indicates that vertical and horizontal values are shared during learning. The mean accuracy is calculated from the results of not sharing strides.

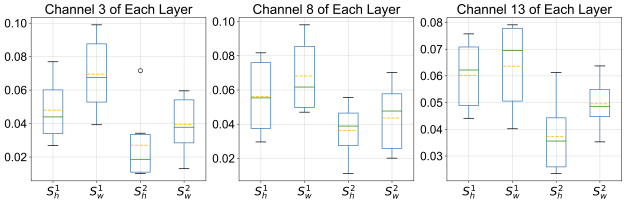

Figure 6. Distribution of learned S_h and S_w in different layers for different initialization configurations (same as Table 6). S_h^i/S_w^i indicates the values in the i-th layer.

4.5. Ablation Studies

Performance Impact of Each Component. Given the feature spectrum transformation (FST), extraction box with adaptive attention (EB-AA), and extraction box with learnable strides (EB-LS), we report the contribution of each component to the performance of our method in Table 5. On the four evaluation metrics, the base model (Baseline) achieves 82.91%, 89.37%, 86.02%, and 75.48%, respectively. Adding the FST with fixed content and stride to the base model (FeaSpect-fixed) improves performance on four evaluation metrics by 9.29%, 4.49%, 7.00%, and 11.48%. By adding the EB-AA based on the FST, FeaSpect-AA further boosts the performance of the model, which improved by 1.70%, 1.73%, 1.72%, and 3.04%. Adding the EB-LS submodule (FeaSpect-LS) improves the performance to a higher extent than adding the EB-AA, *i.e.*, 2.81%, 1.79%, 2.31%, and 4.11%. Finally, combining the FST with the EB-AA and EB-LS (FeaSpect) achieves the best performance on the evaluation metric, *i.e.*, 95.75%, 95.87%, 95.81%, and 91.96%. Thus, we can conclude that the three components, FST, EB-AA, and EB-LS, are complementary and each component contributes to these improvements.

Effect of Initial Stride on Performance. Table 6 reports

L_1	L_2	L_3	L_4	SVCD Dataset				DSIFN Dataset			
				F1(%)	IoU(%)	F1(%)	IoU(%)	F1(%)	IoU(%)	F1(%)	IoU(%)
				Baseline		Baseline		Baseline		Baseline	
				86.0	75.5	86.0	75.5	62.0	44.9	62.0	44.9
				FST (FeaSpect-fixed)		FeaSpect		FST (FeaSpect-fixed)		FeaSpect	
✓				91.7(+5.7)	84.8(+9.3)	94.6(+8.6)	89.7(+14.2)	64.2(+2.2)	47.3(+2.4)	67.7(+5.7)	51.2(+6.3)
✓	✓			**93.0**(+7.0)	**87.0**(+11.5)	**95.8**(+9.8)	**92.0**(+16.5)	**66.6**(+4.6)	**49.9**(+5.0)	**69.9**(+7.9)	**53.7**(+8.8)
		✓	✓	86.7(+0.7)	76.5(+1.0)	87.1(+1.1)	77.2(+1.7)	62.5(+0.5)	45.4(+0.5)	63.0(+1.0)	46.0(+1.1)
✓	✓	✓		92.2(+6.2)	85.2(+9.7)	95.7(+9.7)	91.8(+16.3)	64.6(+2.6)	47.8(+2.9)	69.7(+7.7)	53.5(+8.6)
✓	✓	✓	✓	91.9(+5.9)	85.1(+9.6)	95.7(+9.7)	91.7(+16.2)	64.4(+2.4)	47.4(+2.5)	69.6(+7.6)	53.3(+8.4)

Table 7. Results of applying FST to different layers. FeaSpect is FST with learnable content and strides.

the performance of spectrum transformation (ST) [46] and our FeaSpect with different initial strides on the SVCD dataset. We first note the significance of the stride parameter for ST performance, with a reduction from 89.20% to 85.99% across the best and worst configurations. Despite this, the performance fluctuations of ST for unfavorable initial values remain relatively moderate, indicating its potential in addressing pseudo-changes in change detection. In contrast, the FeaSpect shows greater resilience to poor initialization, consistently converging to high accuracy on four metrics. This highlights the efficacy of backpropagation in enabling FeaSpect to discover optimal configurations during training, avoiding cross-validation.

Effect of Initial Stride on Stride Learning. Considering the diverse initial configurations outlined in Table 6, we depict the distribution of learned strides per layer for various initializations in Fig. 6. Three randomly selected channels from each layer are showcased. An intriguing observation emerges: despite converging to the same accuracy, different initializations yield significantly distinct stride configurations at convergence — whether between different layers in the same channel or between different channels in the same layer. Strides are not recklessly optimized but actively explores the optimal direction at each moment. This emphasizes that loss optimization involves a trade-off between preserving discriminative information, aligning styles, and efficiently progressing toward extreme points.

Effect of Sharing Stride Horizontally and Vertically. We control that strides are shared in both horizontal and vertical dimensions during learning to further discuss the effectiveness of stride learning. The results are detailed in Table 6. With the same initial stride, sharing the learnable stride slightly diminishes the model's detection performance compared to the one without sharing (95.74→94.00). This implies that regions representing style information display irregularities and that imposing equality between the two dimensions would compromise essential semantics.

Effect of Applying FST on Different Layers. We report the results of applying FST to different output layers (L_1-L_4) of 4 blocks in the Table 7. As shown in Table 7, we observe that: **I.** Adding FST to either shallow layers (L_1, L_2) or deeper layers (L_3, L_4) can consistently improve the performance (red numbers) of baseline, indicating that *our FST can be applied to all output layers.* **II:** Adding FST to shallow layers (L_1, L_2) achieves the best improvements. **III:** Based on this, further adding FST to deeper layers (L_3, L_4) fails to achieve additional improvement. Thus, applying FST into first two shallow layers is more reasonable, which aligns with the understanding that style information predominantly manifests in shallow layers.

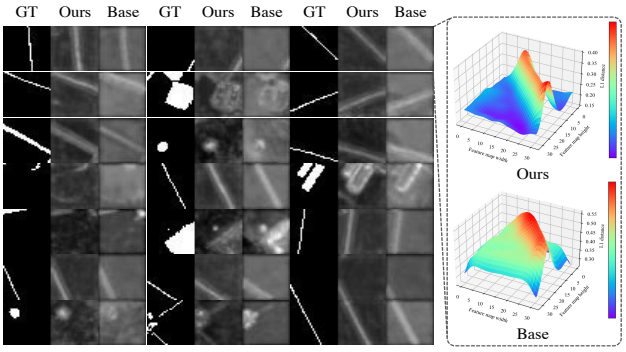

Figure 7. Visual comparison of feature distances between our method and baseline.

4.6. Analysis and Discussion of Features

To verify whether features are style-aligned while preserving semantics, we visualize the point-to-point L1 distance between deep features of bitemporal images in Fig. 7, comparing our method with the baseline (FC-Siam-diff-res [10]). A 3D view of one image pair is provided for clarity. For style alignment, our method shows significantly lower distances in unchanged areas, indicating better alignment of same semantic features. For semantic preservation, it maintains large distances in changed regions, preserving differences between distinct semantics. The consistently low distances in unchanged areas further confirm that our method encodes identical semantics similarly. Such style-aligned yet semantically discriminative features help highlight true changes and suppress pseudo-changes.

5. Conclusion

This paper addresses pseudo-changes in change detection (CD) by exploring spectrum transformation (ST) as a mitigation strategy, while identifying its limitations in transformation space and parameter efficiency. To overcome these issues, we propose feature spectrum learning (FeaSpect), which adaptively eliminates style differences in the feature space. FeaSpect aligns the optimization space of discriminative features and achieves optimal parameters with lower computational cost. Extensive experiments on two public datasets demonstrate its superior accuracy and generalization. Moreover, FeaSpect can be easily integrated into existing siamese-based models to enhance performance. We hope this work inspires new ST-based solutions for CD.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62271377, the National Key Research and Development Program of China under Grant Nos. 2021ZD0110400, 2021ZD0110404, the Key Research and Development Program of Shannxi (Program Nos. 2023YBGY244, 2023QCYLL28, 2024GX-ZDCYL-02-08, 2024GX-ZDCYL-02-17), the Key Scientific Technological Innovation Research Project by Ministry of Education, the Joint Funds of the National Natural Science Foundation of China (U22B2054).

References

[1] Mohammed El Amin Arabi, Moussa Sofiane Karoui, and Khelifa Djerriri. Optical remote sensing change detection through deep siamese network. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 5041–5044. IEEE, 2018. 1, 2

[2] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022. 5, 6

[3] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993. 1, 3

[4] Chi-Farn Chen, Nguyen-Thanh Son, Ni-Bin Chang, Cheng-Ru Chen, Li-Yu Chang, Miguel Valdez, Gustavo Centeno, Carlos Alberto Thompson, and Jorge Luis Aceituno. Multi-decadal mangrove forest change detection and prediction in honduras, central america, with landsat imagery and a markov chain model. *Remote Sensing*, 5(12):6408–6426, 2013. 1

[5] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020. 2, 3

[6] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 5, 6

[7] Jie Chen, Ziyang Yuan, Jian Peng, Li Chen, Haozhe Huang, Jiawei Zhu, Yu Liu, and Haifeng Li. Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1194–1206, 2020. 3

[8] Pan Chen, Bing Zhang, Danfeng Hong, Zhengchao Chen, Xuan Yang, and Baipeng Li. Fccdn: Feature constraint network for vhr image change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187:101–119, 2022. 1

[9] Gong Cheng, Ceyuan Yang, Xiwen Yao, Lei Guo, and Junwei Han. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. *IEEE transactions on geoscience and remote sensing*, 56(5):2811–2821, 2018. 5, 6

[10] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067. IEEE, 2018. 2, 5, 6, 7, 8

[11] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. High resolution semantic change detection. *arXiv preprint arXiv:1810.08452*, 2018. 2

[12] Huihui Dong, Wenping Ma, Yue Wu, Maoguo Gong, and Licheng Jiao. Local descriptor learning for change detection in synthetic aperture radar images via convolutional neural networks. *IEEE access*, 7:15389–15403, 2018. 2

[13] Yong Du, Philippe M Teillet, and Josef Cihlar. Radiometric normalization of multitemporal high-resolution satellite images with quality control for land cover change detection. *Remote sensing of Environment*, 82(1):123–134, 2002. 1

[14] Bo Fang, Li Pan, and Rong Kou. Dual learning-based siamese framework for change detection using bi-temporal vhr optical remote sensing images. *Remote Sensing*, 11(11): 1292, 2019. 1, 3, 6

[15] Bo Fang, Gang Chen, Guichong Ouyang, Jifa Chen, Rong Kou, and Lizhe Wang. Content-invariant dual learning for change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2021. 1, 3

[16] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 3

[17] Jan Feranec, Gerard Hazeu, Susan Christensen, and Gabriel Jaffrain. Corine land cover change detection in europe (case studies of the netherlands and slovakia). *Land use policy*, 24 (1):234–247, 2007. 1

[18] Maoguo Gong, Tao Zhan, Puzhao Zhang, and Qiguang Miao. Superpixel-based difference representation learning for change detection in multispectral remote sensing images. *IEEE Transactions on Geoscience and Remote sensing*, 55 (5):2658–2673, 2017. 2

[19] Bin Hou, Qingjie Liu, Heng Wang, and Yunhong Wang. From w-net to cdgan: Bitemporal change detection via deep learning techniques. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3):1790–1802, 2019. 5, 6

[20] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6891–6902, 2021. 3

[21] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8988–8999, 2021. 3

[22] Yanyuan Huang, Xinghua Li, Zhengshun Du, and Huanfeng Shen. Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 5, 6

[23] Masroor Hussain, Dongmei Chen, Angela Cheng, Hui Wei, and David Stanley. Change detection from remotely sensed

images: From pixel-based to object-based approaches. *ISPRS Journal of photogrammetry and remote sensing*, 80:91–106, 2013. 1

[24] Bo Jiang, Zitian Wang, Xixi Wang, Ziyan Zhang, Lan Chen, Xiao Wang, and Bin Luo. Vct: Visual change transformer for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 5, 6

[25] Nada Kadhim, Monjur Mourshed, and Michaela Bray. Advances in remote sensing applications for urban sustainability. *Euro-Mediterranean Journal for Environmental Integration*, 1(1):1–22, 2016. 1

[26] AH Kerekes and M Alexe. Evaluating urban sprawl and landuse change using remote sensing, gis techniques and historical maps. case study: The city of dej, romania. *Analele Universităţii din Oradea, Seria Geografie*, 29(2):52–63, 2019.

[27] A Koltunov and SL Ustin. Early fire detection using nonlinear multitemporal prediction of thermal imagery. *Remote Sensing of Environment*, 110(1):18–28, 2007. 1

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[29] Jogendra Nath Kundu, Akshay R Kulkarni, Suvaansh Bhambri, Varun Jampani, and Venkatesh Babu Radhakrishnan. Amplitude spectrum transformation for open compound domain adaptive semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1220–1227, 2022. 3

[30] Tao Lei, Yuxiao Zhang, Zhiyong Lv, Shuying Li, Shigang Liu, and Asoke K Nandi. Landslide inventory mapping from bitemporal images using deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 16(6): 982–986, 2019. 2, 5, 6

[31] Yu Lei, Xiaodong Liu, Jiao Shi, Chao Lei, and Jing Wang. Multiscale superpixel segmentation with deep features for change detection. *Ieee Access*, 7:36600–36616, 2019. 1, 2

[32] Jia Liu, Wenjie Xuan, Yuhang Gan, Yibing Zhan, Juhua Liu, and Bo Du. An end-to-end supervised domain adaptation framework for cross-domain change detection. *Pattern Recognition*, 132:108960, 2022. 1, 3

[33] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021. 3

[34] Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang. 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):811–815, 2020. 3

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[36] Lichao Mou, Lorenzo Bruzzone, and Xiao Xiang Zhu. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):924–935, 2018. 2

[37] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 4, 6

[38] Daifeng Peng, Yongjun Zhang, and Haiyan Guan. End-to-end change detection for high resolution satellite images using improved unet++. *Remote Sensing*, 11(11):1382, 2019. 2

[39] Xueli Peng, Ruofei Zhong, Zhen Li, and Qingyang Li. Optical remote sensing image change detection based on attention mechanism and image difference. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7296–7307, 2020. 3

[40] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 5, 6

[41] PM Teillet, JL Barker, BL Markham, RR Irish, G Fedosejevs, and JC Storey. Radiometric cross-calibration of the landsat-7 etm+ and landsat-5 tm sensors based on tandem data sets. *Remote sensing of Environment*, 78(1-2):39–54, 2001. 1

[42] Shuang Wang, Dong Zhao, Chi Zhang, Yuwei Guo, Qi Zang, Yu Gu, Yi Li, and Licheng Jiao. Cluster alignment with target knowledge mining for unsupervised domain adaptation semantic segmentation. *IEEE Transactions on Image Processing*, 31:7403–7418, 2022. 1

[43] Shuang Wang, Qi Zang, Dong Zhao, Chaowei Fang, Dou Quan, Yutong Wan, Yanhe Guo, and Licheng Jiao. Select, purify, and exchange: A multisource unsupervised domain adaptation method for building extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1

[44] Wahyu Wiratama, Jongseok Lee, Sang-Eun Park, and Donggyu Sim. Dual-dense convolution network for change detection of high-resolution panchromatic imagery. *Applied Sciences*, 8(10):1785, 2018. 2

[45] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 3

[46] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, pages 4085–4095, 2020. 1, 3, 4, 6, 7, 8

[47] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[48] Qi Zang, Shuang Wang, Dong Zhao, Yang Hu, Dou Quan, Jinlong Li, Nicu Sebe, and Zhun Zhong. Generalized source-free domain-adaptive segmentation via reliable knowledge propagation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5967–5976, 2024. 1

[49] Qi Zang, Shuang Wang, Dong Zhao, Zhun Zhong, Biao Hou, and Licheng Jiao. Joint style and layout synthesizing: Toward generalizable remote sensing semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1

[50] Yang Zhan, Kun Fu, Menglong Yan, Xian Sun, Hongqi Wang, and Xiaosong Qiu. Change detection based on deep siamese convolutional network for optical aerial images. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1845–1849, 2017. 3

[51] Cui Zhang, Liejun Wang, Shuli Cheng, and Yongming Li. Swinsunet: Pure transformer network for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 5, 6

[52] Hongyan Zhang, Manhui Lin, Guangyi Yang, and Liangpei Zhang. Escnet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2, 5, 6

[53] Jingyi Zhang, Jiaxing Huang, Zichen Tian, and Shijian Lu. Spectral unsupervised domain adaptation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9829–9840, 2022. 3

[54] Mengya Zhang, Guangluan Xu, Keming Chen, Menglong Yan, and Xian Sun. Triplet-based semantic relation learning for aerial remote sensing image change detection. *IEEE Geoscience and Remote Sensing Letters*, 16(2):266–270, 2018. 3

[55] Dong Zhao, Shuang Wang, Qi Zang, Dou Quan, Xiutiao Ye, and Licheng Jiao. Towards better stability and adaptability: Improve online self-training for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11733–11743, 2023. 1

[56] Dong Zhao, Shuang Wang, Qi Zang, Dou Quan, Xiutiao Ye, Rui Yang, and Licheng Jiao. Learning pseudo-relations for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19191–19203, 2023.

[57] Dong Zhao, Shuang Wang, Qi Zang, Licheng Jiao, Nicu Sebe, and Zhun Zhong. Stable neighbor denoising for source-free domain adaptive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23416–23427, 2024.

[58] Dong Zhao, Qi Zang, Shuang Wang, Nicu Sebe, and Zhun Zhong. Connectivity-driven pseudo-labeling makes stronger cross-domain segmenters. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[59] Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15193–15202, 2021. 5, 6

[60] Zhi Zheng, Yi Wan, Yongjun Zhang, Sizhe Xiang, Daifeng Peng, and Bin Zhang. Clnet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:247–267, 2021. 1