Temporal-Spectral-Spatial Unified Remote Sensing Dense Prediction

Sijie Zhao, Feng Liu, Xueliang Zhang*, Hao Chen*, Pengfeng Xiao, Lei Bai

***Abstract*—The proliferation of diverse remote sensing data has spurred advancements in dense prediction tasks, yet significant challenges remain in handling data heterogeneity. Remote sensing imagery exhibits substantial variability across temporal, spectral, and spatial (TSS) dimensions, complicating unified data processing. Current deep learning models for dense prediction tasks, such as semantic segmentation and change detection, are typically tailored to specific input-output configurations. Consequently, variations in data dimensionality or task requirements often lead to significant performance degradation or model incompatibility, necessitating costly retraining or fine-tuning efforts for different application scenarios. This paper introduces the Temporal-Spectral-Spatial Unified Network (TSSUN), a novel architecture designed for unified representation and modeling of remote sensing data across diverse TSS characteristics and task types. TSSUN employs a Temporal-Spectral-Spatial Unified Strategy that leverages meta-information to decouple and standardize input representations from varied temporal, spectral, and spatial configurations, and similarly unifies output structures for different dense prediction tasks and class numbers. Furthermore, a Local-Global Window Attention mechanism is proposed to efficiently capture both local contextual details and global dependencies, enhancing the model's adaptability and feature extraction capabilities. Extensive experiments on multiple datasets, encompassing tasks like building detection and land use/land cover classification with diverse TSS configurations, demonstrate that a single TSSUN model effectively adapts to heterogeneous inputs and unifies various dense prediction tasks. The proposed approach consistently achieves or surpasses state-of-the-art performance, highlighting its robustness and generalizability for complex remote sensing applications without requiring task-specific modifications.**

***Index Terms*—Remote Sensing, Dense Prediction, Unified Network, Change Detection, Semantic Segmentation, Deep learning**

I. INTRODUCTION

With the continuous advancement of remote sensing technologies and the increasing diversity of data acquisition methods [1], the field of remote sensing has entered a phase of rapid development [2]. Massive and multi-source remote sensing data have been widely applied to various dense prediction tasks, playing a crucial role in applications such as urban expansion monitoring [3], land cover classification [4], disaster damage assessment [5], crop classification and yield estimation [6], and environmental pollution monitoring [7]. Remote sensing imagery exhibits high heterogeneity across three key dimensions—temporal, spectral, and spatial—posing significant challenges for unified processing due to variations in temporal length, number of spectral channels, and spatial resolution in practical applications [8].

From the task perspective, dense prediction in remote sensing primarily involves three core categories: semantic segmentation, semantic change detection, and binary change detection. These tasks can be formally defined as follows: given a remote sensing image of shape (T_1, C_1, H_1, W_1), the deep learning model is expected to produce a prediction of shape (T_2, C_2, H_2, W_2), where T_1 and T_2 denote the temporal lengths of the input and output, C_1 is the number of input channels, C_2 is the number of output classes, and (H_1, W_1) and (H_2, W_2) denote the spatial dimensions of the input and output, respectively. Typically, $H_1 = H_2$ and $W_1 = W_2$. In semantic segmentation, the model performs multi-class land cover extraction from a single time-point image, corresponding to $T_2 = 1$ and $C_2 \geq 2$; in semantic change detection, the model extracts land cover information at each time step to analyze semantic differences between any two time points, corresponding to $T_2 = T_1$ and $C_2 \geq 2$; in binary change detection, the model identifies whether changes occur between adjacent time points, corresponding to $T_2 = T_1 - 1$ and $C_2 = 2$.

In recent years, deep learning methods, supported by the growing abundance of remote sensing data, have achieved notable success in dense prediction tasks, leading to the emergence of a wide range of high-performing models across all three task categories. However, existing models are typically designed for fixed input-output configurations, i.e., specific (T_1, C_1, H_1, W_1) and (T_2, C_2, H_2, W_2). Even slight variations in any of these dimensions can lead to significant performance degradation or complete incompatibility. In particular, changes in T_1, C_1, and (H_1, W_1) reflect the diversity of remote sensing data along the Temporal- Spectral-Spatial (TSS) dimensions; variations in T_2 correspond to different task types, while differences in C_2 relate to varying classification requirements. Given that remote sensing applications often differ across these dimensions, models trained in one scenario typically struggle to generalize to others, requiring additional training or fine-

This research is supported by the the Shanghai Municipal Science and Technology Major Project, *Corresponding author: Xueliang Zhang and Hao Chen*

Sijie Zhao, Xueliang Zhang, and Pengfeng Xiao are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: zsj@smail.nju.edu.cn; zxl@nju.edu.cn; xiaopf@nju.edu.cn).

Feng Liu is with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China (e-mail: liufeng2317@sjtu.edu.cn)

HaoChen and Lei Bai are with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China (email: justchenhao@buaa.edu.cn; bailei@pjlab.org.cn).

Codes are available at *https://github.com/walking-shadow/Official_TSSUN*

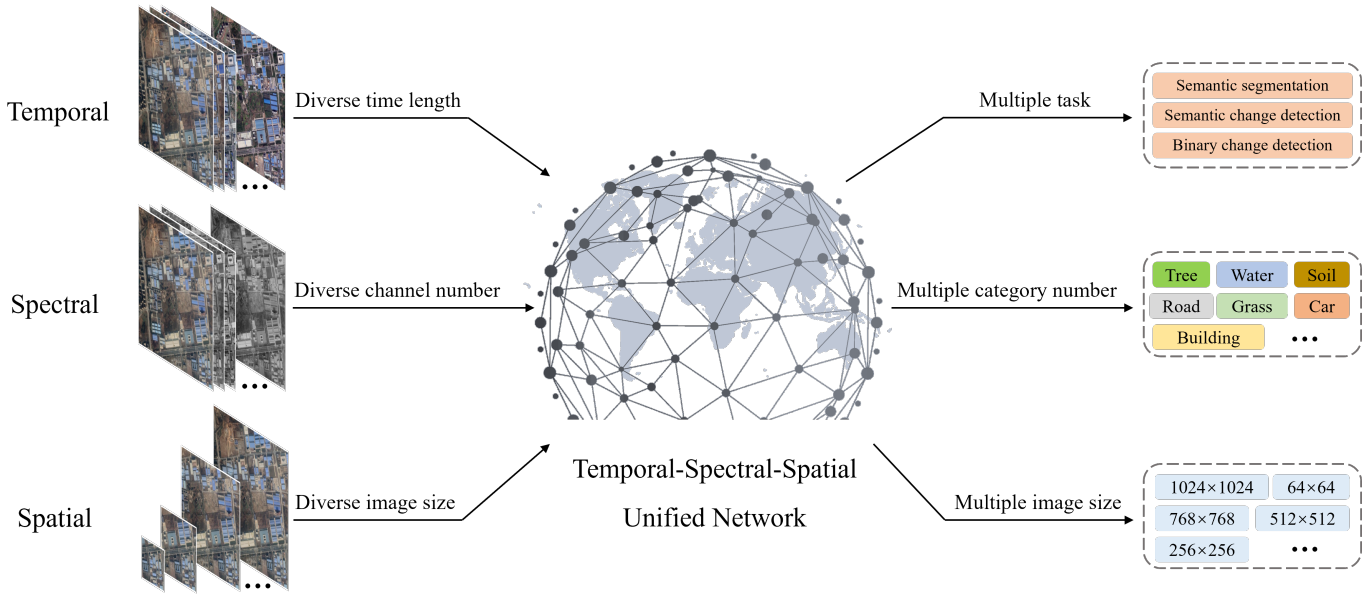

Fig. 1. Illustration of Temporal-Spectral-Spatial Unified Network. TSSUN is capable of handling input and output with arbitrary temporal, spectral, and spatial dimensions, supports dense prediction for any number of classes, and unifies semantic segmentation, binary change detection and semantic change detection tasks.

tuning, which introduces considerable computational and time costs.

To address these issues, this study propose a Temporal-Spectral-Spatial Unified Network (TSSUN) that enables unified representation and modeling of remote sensing data across the temporal, spectral, and spatial dimensions, as shown in Figure 1. TSSUN is highly flexible, capable of handling any combination of (T_1, C_1, H_1, W_1) inputs and (T_2, C_2, H_2, W_2) outputs. It supports inputs with varying temporal lengths, spectral bands, and spatial resolutions, and unifies all three dense prediction tasks. Moreover, it can produce outputs with an arbitrary number of semantic classes and maintains strong performance across diverse configurations.

Specifically, we introduce the Temporal-Spectral-Spatial Unified Strategy (TSSUS) to decouple and unify the input and output representations across the TSS dimensions. At the input stage, TSSUS leverages inherent spectral and spatial characteristics to achieve unified encoding of remote sensing imagery. Along the spectral dimension, it exploits the continuity across spectral bands and integrates spectral metadata to accomplish spectral harmonization. In parallel, along the spatial dimension, it harnesses the similarity among nearby regions, supplemented with spatial metadata, to produce a unified representation of spatial scales. By preserving the time dimension at the input, TSSUS is able to independently extract features from remote sensing images across distinct temporal instances. At the encoder–decoder junction, TSSUS facilitates feature-level fusion along the temporal axis, effectively mitigating interference from irrelevant information while adapting the output temporal length to suit the dense prediction task at hand [9], [10]. At the output stage, TSSUS further capitalizes on the structured nature of remote sensing dense prediction tasks in both spectral and spatial dimensions. In the spectral dimension, by modeling inter-class correlations, it enables prediction over an arbitrary number of classes. Meanwhile, in the spatial dimension, by exploiting the spatial structural consistency among neighboring regions, it significantly enhances the model's generalization capability when outputting data at varying spatial resolutions.

To further improve the model's ability to capture diverse TSS combinations, we design a Local-Global Window Attention (LGWA) mechanism. This module efficiently extracts local features using three overlapping window-based attention blocks with different shapes, followed by a global attention block that aggregates information at the global level. This design achieves a balance between computational efficiency and expressive power, enabling collaborative modeling of local and global features and significantly boosting the model's performance in complex remote sensing tasks.

In summary, the main contributions of this work are as follows:

1) A Temporal-Spectral-Spatial Unified Network is proposed to enable unified modeling across temporal, spectral, and spatial dimensions for dense prediction in remote sensing. TSSUN supports arbitrary TSS input configurations and accommodates various tasks, including semantic segmentation, semantic change detection, and binary change detection with flexible output class settings.
2) A Local-Global Window Attention mechanism is designed to efficiently capture both local and global contextual features, improving performance across a wide range of remote sensing prediction tasks.
3) Extensive experiments are conducted on multiple datasets with diverse TSS configurations, including building detection and land use/land cover (LULC) classification. The results show that a single TSSUN model can adapt to heterogeneous inputs, unify multiple

task types, and consistently achieve or exceed state-of-the-art (SOTA) performance.

II. Related Works

A. Task Unification in Remote Sensing Dense Prediction

Dense prediction tasks in remote sensing, particularly semantic segmentation, binary change detection, and semantic change detection, form the cornerstone of many Earth observation applications. While often treated as distinct problems, they share fundamental objectives and challenges.

Semantic segmentation (SS) in remote sensing aims to assign a thematic label to each pixel in an image, thereby partitioning it into meaningful regions corresponding to different land cover or land use classes [11]. Early approaches relied on traditional machine learning algorithms such as Support Vector Machines (SVMs) and Random Forests (RFs) operating on handcrafted features [12]. The advent of deep learning, particularly Fully Convolutional Networks (FCNs) [13], revolutionized the field, enabling end-to-end feature learning and significantly improving segmentation accuracy. Architectures like U-Net [14], with its encoder-decoder structure and skip connections, became foundational, inspiring numerous variants such as UNet++ [15] and the DeepLab series [15], [16], which introduced atrous convolutions and spatial pyramid pooling to handle scale variations.

Binary change detection (BCD) focuses on identifying areas of significant difference between two co-registered remote sensing images acquired at different times [17]. Traditional BCD methods include image differencing, principal component analysis (PCA)-based techniques [18], and change vector analysis (CVA) [19]. Deep learning has brought substantial advancements, with Siamese networks being a popular choice for learning discriminative features from image pairs [20], [21]. These networks typically process the bi-temporal images through shared-weight encoders and then compare the resulting feature maps to identify changes. U-Net-like architectures have also been adapted for BCD, often by modifying the input layers to accept concatenated bi-temporal images or difference images, and training the network to output a binary change map [22].

Semantic change detection (SCD), also known as "from-to" change detection, provides a more detailed analysis by not only identifying where changes have occurred but also classifying the type of land cover transition (e.g., forest to urban) [23]. This task is inherently more complex than BCD. A common traditional approach is post-classification comparison (PCC), where SS is performed independently on multi-temporal images, and the resulting maps are compared [24]. However, PCC suffers from error accumulation from the individual classification steps. More integrated methods aim to perform simultaneous segmentation and change identification. Deep learning models for SCD often employ dual-branch or multi-task learning frameworks, where separate decoders might be used for segmentation at each time step and for the final change map [25].

Critically analyzing these three tasks reveals profound interrelations and shared challenges. SCD inherently builds upon the principles of SS, as identifying "from-to" transitions requires accurate semantic understanding of the landscape at different time points. BCD can be viewed as a simplified abstraction of SCD, where all types of semantic changes are aggregated into a binary "change" or "no change" outcome. Thus, accurate SS is a prerequisite for high-quality SCD, and insights from BCD can potentially guide the focus of SCD algorithms. The complementary nature of these tasks is evident: SS provides the static land cover context, BCD highlights dynamic regions, and SCD offers a comprehensive narrative of landscape transformation. Information from one task can, in principle, regularize or enhance the performance of others; for instance, knowing the semantic class of a pixel can refine change boundaries, or detecting a change can prompt a more detailed semantic analysis. Despite these intrinsic relationships and the potential for synergistic modeling, there is currently a significant lack of models or frameworks that can effectively unify semantic segmentation, binary change detection, and semantic change detection within a single, cohesive architecture.

B. Data Dimensionality in Remote Sensing Dense Prediction

Remote sensing datasets utilized for dense prediction tasks are characterized by extensive variability across temporal, spectral, and spatial dimensions. This heterogeneity stems from the diverse array of satellite and airborne sensors, each with unique acquisition parameters, mission objectives, and coverage patterns, tailored for different application scenarios and geographical regions [26]. Such variability presents a formidable challenge for developing universally applicable and robust dense prediction models.

The temporal dimension in remote sensing data exhibits significant disparities. Revisit frequency, a critical factor for monitoring dynamic phenomena, ranges from multiple observations per day with sensors like MODIS [27], to several days (e.g., Sentinel-1 and Sentinel-2, with 5-12 day repeat cycles depending on latitude and constellation status [28], [29]), to 16 days for Landsat missions [30]. Consequently, the temporal length of image sequences available for analysis can vary from bi-temporal pairs, commonly used in BCD and SCD [[20], [23]], to dense time series comprising hundreds of observations, which are invaluable for applications like agricultural monitoring [31]. Models trained on data with a specific temporal frequency or sequence length may not generalize well to datasets with different temporal characteristics without substantial retraining or adaptation.

Spectral dimensionality is another source of major variation. The number of spectral bands can range from a single panchromatic band to a few multispectral bands (e.g., 4-bands in NAIP imagery, 13 bands in Sentinel-2 MSI [29]), to hundreds of narrow, contiguous bands in hyperspectral sensors like AVIRIS [32] or the upcoming EnMAP mission [33]. Each sensor captures information from different portions of the electromagnetic spectrum, with varying band central wavelengths and bandwidths. This spectral diversity allows for the discrimination of different materials and land cover types based on their unique spectral signatures. However, it

also means that models developed for one sensor may not be directly applicable to data from another sensor without strategies to handle the differing channel counts and spectral information content [34].

Spatial characteristics, primarily ground sampling distance (GSD) or resolution, also vary widely. Data can range from very high resolution (VHR), with GSDs of less than 1 meter, to high resolution (HR, 1-10 meters, e.g., Sentinel-2, PlanetScope), to lower resolution. VHR imagery is suitable for detailed urban mapping or object detection [35], while medium to low-resolution data is often used for regional or global land cover mapping and monitoring [36]. The spatial extent of individual scenes also differs, impacting the scale of analysis. Models trained on imagery of a particular GSD often struggle when applied to data with significantly different spatial resolutions, as object appearance and contextual cues change drastically.

The inherent heterogeneity in temporal frequency, spectral composition, and spatial resolution across remote sensing datasets thus poses a substantial hurdle for developing universally applicable dense prediction models. Consequently, there is a pressing research gap concerning the development of adaptive model architectures or unified data processing strategies that can effectively ingest and interpret such diverse time-spectrum-space data formats at the input level for comprehensive and robust dense prediction.

III. Methodology

A. Problem Formulation

Dense prediction tasks in remote sensing generally refer to performing structured semantic inference on input images at the pixel level. While these tasks vary in nature, their core objective remains consistent: to extract rich semantic information from remote sensing data that includes temporal, spectral, and spatial dimensions, and to output structured prediction results. This paper addresses three typical tasks—semantic segmentation, semantic change detection, and binary change detection—and formulates them as a unified tensor mapping problem. Let the input remote sensing image be represented as a four-dimensional tensor $X \in \mathbb{R}^{T_1 \times C_1 \times H_1 \times W_1}$, where T_1 denotes the length of the time sequence, C_1 represents the number of spectral channels, and H_1 and W_1 denote the height and width of the image, respectively. The model output is represented as a tensor $Y \in \mathbb{R}^{T_2 \times C_2 \times H_2 \times W_2}$, where T_2 represents the output time dimension, C_2 denotes the number of semantic categories, and H_2 and W_2 represent the spatial dimensions of the output image. In most cases, the input and output are spatially aligned, i.e., $H_1 = H_2$ and $W_1 = W_2$.

Based on this representation, the three tasks considered in this paper can be uniformly described as follows:

- **Semantic Segmentation**: The model performs pixel-level multi-class classification on a remote sensing image at a given time point. This corresponds to the setup where $T_1 = 1$, $T_2 = 1$, and $C_2 \geq 2$.
- **Semantic Change Detection**: The model predicts semantic labels for each time point to analyze semantic changes between any two time points. This corresponds to $T_2 = T_1$, with $C_2 \geq 2$.
- **Binary Change Detection**: The model detects changes between consecutive time points without distinguishing specific categories of change. This corresponds to $T_2 = T_1 - 1$ and $C_2 = 2$.

It is important to note that the input data may differ significantly in terms of time length T_1, spectral channels C_1, and spatial scale (H_1, W_1), depending on the task type or dataset. To address this variability, we propose a unified modeling framework that aims to learn a mapping function f_θ, such that

$$f_\theta : \mathbb{R}^{T_1 \times C_1 \times H_1 \times W_1} \to \mathbb{R}^{T_2 \times C_2 \times H_2 \times W_2}, \tag{1}$$

which can adapt to any input configuration and output dense prediction results that meet the requirements of the target task. This abstract modeling provides the mathematical foundation for the modular design of the subsequent methods.

B. Overview of the Temporal-Spectral-Spatial Unified Network (TSSUN)

The proposed Temporal-Spectral-Spatial Unified Network (TSSUN) is designed as a unified modeling framework capable of addressing diverse dense prediction tasks across varying data modalities and spatial resolutions, as illustrated in Figure 2. TSSUN employs the Temporal-Spectral-Spatial Unified Strategy (TSSUS), which systematically integrates input alignment, feature extraction, and output decoding into a cohesive processing pipeline.

In the input stage, TSSUN introduces the Spectral-Spatial Unified Module (SSUM), which encodes heterogeneous spectral and spatial data into a unified representation. By leveraging spectral and spatial priors through a guided alignment mechanism, SSUM ensures data consistency across spectral and spatial dimensions, establishing a coherent input foundation for subsequent feature modeling. During the feature extraction stage, TSSUN employs the Local-Global Window Attention (LGWA) mechanism as the primary feature modeling unit, effectively capturing multi-scale features in TSS inputs. LGWA integrates three distinct window-based attention blocks, each tailored to capture local and global features at different scales. This structure balances computational efficiency and representational power, facilitating the extraction of spatial structures across multiple scales. At the encoder-decoder junction, TSSUN incorporates the Temporal Unification Module (TUM), which focuses on temporal feature fusion. TUM mitigates temporal redundancy and adjusts the temporal length of the output based on task-specific requirements, ensuring temporal coherence in the generated predictions. In the output stage, TSSUN reintroduces SSUM in a mirrored configuration to further reinforce spectral-spatial consistency during the decoding process. By progressively decoding spectral and spatial dimensions, SSUM enables unified modeling for both semantic segmentation with arbitrary class numbers and multi-resolution spatial reconstruction, thereby enhancing TSSUN's capability to handle multi-task, multi-modal data effectively.

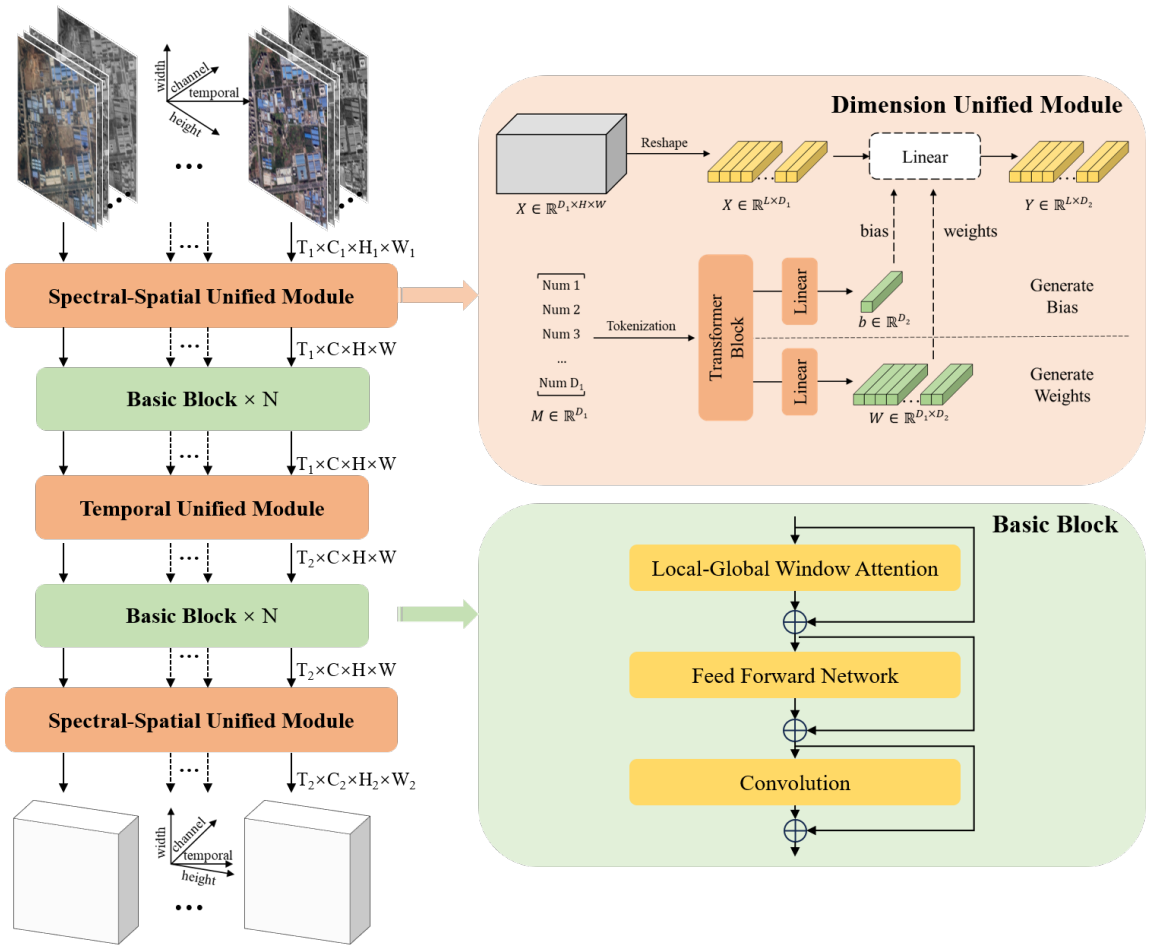

Fig. 2. Overview of the TSSUN architecture, comprising the Spectral-Spatial Unified Module (SSUM) for input alignment, Local-Global Window Attention (LGWA) blocks for feature extraction, and the Temporal Unified Module (TUM) for temporal fusion and output adaptation. The top right panel details the Dimension Unified Module (DUM), while the bottom right panel illustrates the LGWA block structure.

Overall, TSSUN leverages TSSUS to comprehensively integrate the input, feature extraction, and output stages, establishing a unified framework for heterogeneous data representation and dense prediction. The LGWA module further enhances the model's capacity to effectively capture both local and global contextual information, ensuring a balanced representation of multi-scale and multi-modal features. The following sections provide a detailed analysis of the two core components, TSSUS and LGWA, including their structural design and functional mechanisms.

C. Temporal-Spectral-Spatial Unified Strategy (TSSUS)

The Temporal-Spectral-Spatial Unified Strategy (TSSUS) is designed to decouple and unify the representation of inputs and outputs across temporal, spectral, and spatial dimensions. In the input stage, TSSUS leverages the inherent characteristics of spectral and spatial dimensions to achieve unified encoding of remote sensing imagery. Specifically, in the spectral dimension, TSSUS captures the continuity among spectral bands and incorporates spectral metadata to ensure spectral alignment. In the spatial dimension, TSSUS extracts spatial scale representations by modeling the similarity among neighboring regions and incorporating spatial metadata. By retaining the temporal dimension at the input stage, TSSUS enables independent feature extraction for remote sensing images across different time instances. At the encoder-decoder interface, TSSUS performs feature-level fusion along the temporal axis, effectively suppressing irrelevant information while adaptively adjusting the output temporal length to accommodate dense prediction tasks. In the output stage, TSSUS further leverages the structured characteristics of dense prediction tasks in remote sensing across spectral and spatial dimensions. In the spectral dimension, it models inter-class relationships to facilitate predictions with arbitrary class numbers. In the spatial dimension, it enhances the model's generalization capability for multi-resolution data outputs by leveraging the spatial structure consistency among neighboring regions.

To facilitate the mapping from the original data space to the unified feature space across different variables (temporal, spectral, and spatial), TSSUS introduces the Dimension Unified Module (DUM), which serves as the foundational implementation for both the SSUM and the TUM. DUM employs metadata from TSS variable subsets to generate adaptive weights and biases for linear layers, thereby enabling adaptive feature mapping. As illustrated in the top-right subfigure of Figure 2, given an input tensor $X \in \mathbb{R}^{D_1 \times H \times W}$ and its

metadata $M \in \mathbb{R}^{D_1}$, where D_1 varies across different tasks and scenarios, DUM maps the input tensor to a unified feature $Y \in \mathbb{R}^{L \times D_2}$ through a hypernetwork generation process. The detailed procedure is as follows:

- **Metadata Embedding:** Metadata M is subjected to positional encoding and tokenization. A learnable class token [CLS] is prepended to the token sequence $T \in \mathbb{R}^{(D_1+1)\times d}$, where d denotes the embedding dimension.
- **Cross-Variable Relationship Modeling:** The token sequence is processed through multiple transformer blocks to capture the underlying relationships among metadata tokens.
- **Adaptive Parameter Generation:** The [CLS] token is linearly projected to generate bias parameters $b \in \mathbb{R}^{D_2}$, while the remaining tokens are projected to produce the weight matrix $W \in \mathbb{R}^{D_1 \times D_2}$. The resulting W and b form a linear layer that maps the input features with D_1 channels to output features with D_2 channels. The input X is first reshaped from (D_1, H, W) to (L, D_1), where $L = H \times W$, and then processed through the generated linear layer, resulting in an output of shape (L, D_2).

The DUM structure not only adaptively maps any TSS variable subset to a unified feature in terms of shape, but also effectively preserves the relationships between remote sensing data in original space, thereby enhancing the model's generalization capability across heterogeneous data.

D. Local-Global Window Attention (LGWA)

Transformer networks have become a prevalent structure in remote sensing tasks due to their exceptional capacity for modeling long-range dependencies. However, the computational complexity of traditional transformers increases quadratically with sequence length, posing significant challenges for the efficient processing of large-scale remote sensing data. To address this issue, this study proposes the Local-Global Window Attention (LGWA) module, which balances computational efficiency and modeling capability by capturing fine-grained features within local windows while simultaneously incorporating global contextual information.

LGWA adopts three selectable window attention shapes with varying sizes and configurations, as illustrated in Figure 3. These window shapes not only determine the feature sensitivity within each window but also directly impact computational efficiency. Therefore, selecting appropriate window configurations in different network blocks allows for an optimal trade-off between modeling effectiveness and computational resources. For a specific attention window, the input sequence $X \in \mathbb{R}^{L \times d_M}$ is projected into query, key, and value matrices Q, K, V through linear projections, formulated as:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V, \tag{2}$$

where $W^Q, W^K \in \mathbb{R}^{d_M \times d_k}$ and $W^V \in \mathbb{R}^{d_M \times d_v}$ are learnable weight matrices. The attention calculation within each window focuses on extracting fine-grained features, while different window configurations provide sensitivity to varying scales.

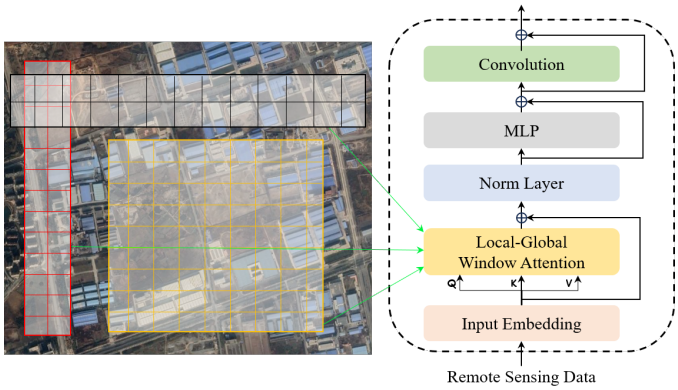

Fig. 3. Architecture of the proposed Local-Global Window Attention (LGWA)-based Transformer block, employing multiple window shapes to enhance feature extraction.

Next, attention scores are computed using the scaled dot-product attention mechanism:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3}$$

To further enhance the feature representation, the LGWA module employs the multi-head attention strategy, allowing each head to independently execute the above process and then concatenate the results:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(H_1, H_2, \ldots, H_h)W^O, \tag{4}$$

where $H_i = \text{Attn}(QW_i^Q, KW_i^K, VW_i^V)$, and $W^O \in \mathbb{R}^{hd_v \times d_M}$ projects the concatenated output back to the original dimension.

By adopting this modular design, the LGWA module effectively captures fine-grained features within local windows while integrating global contextual information. The flexible configuration of window shapes further optimizes computational efficiency, enabling the module to handle multi-scale remote sensing scenarios effectively.

IV. Experiments and Results

To validate the adaptability of TSSUN to arbitrary Temporal-Spectral-Spatial inputs and its capability to concurrently perform semantic segmentation, semantic change detection, and binary change detection tasks with support for a variable number of output classes, we conducted experiments on a total of six datasets across building and Land Cover/Land Classification (LULC) scenarios. For each scenario, a unified TSSUN model was trained using the combined training sets from all datasets within that scenario, and its performance was evaluated on their respective test sets.

In the building scenario, we selected the WHU, WHU-CD, LEVIR-CD, and TSCD datasets. While these datasets share the same number of channels in their input and label data, they exhibit variations in temporal and spatial dimensions, leading to differences in dense prediction task types. Specifically, the WHU dataset corresponds to a single-temporal building semantic segmentation task. The WHU-CD dataset is associated with a bi-temporal building semantic change detection task.

TABLE I
BRIEF INTRODUCTION OF THE EXPERIMENTAL DATASETS.

Name	Scene	Task	Input T	Output T	Input C	Output C	Image Size	Resolution	Images
WHU	Building	SS	1	1	3	1	512×512	0.3	8189
WHU-CD	Building	SCD	2	2	3	1	32207×15354	0.075	1
LEVIR-CD	Building	BCD	2	1	3	1	1024×1024	0.5	445
TSCD	Building	BCD	4	3	3	1	256×256	0.5	2700
LoveDA Urban	LULC	SS	1	1	3	7	1024×1024	0.3	5987
DynamicEarthnet	LULC	SCD&SS&BCD	24	24	4	6	1024×1024	3	54750

The LEVIR-CD dataset is used for a bi-temporal building binary change detection task. The TSCD dataset pertains to a multi-temporal building binary change detection task, as summarized in Table I.

In the LULC scenario, we chose the LoveDA Urban and Dynamic EarthNet datasets. These datasets differ in their temporal, spectral dimensions for both input and label data. The LoveDA Urban dataset corresponds to a single-temporal semantic segmentation task, while the Dynamic EarthNet dataset is used for multi-temporal semantic change detection, semantic segmentation and binary change detection tasks, as detailed in Table I.

The experiments across multiple datasets in the building scenario were primarily designed to verify TSSUN's adaptability to inputs and outputs with arbitrary temporal-spatial dimensions within this context. Furthermore, these experiments aimed to confirm its ability to simultaneously handle semantic segmentation, semantic change detection, and binary change detection tasks. The datasets in the LULC scenario were mainly used to validate TSSUN's adaptability to inputs and outputs with arbitrary temporal-spectral dimensions. Additionally, these experiments were intended to demonstrate its capacity to concurrently manage semantic segmentation, binary change detection and semantic change detection tasks while supporting a flexible number of output classes.

A. Datasets

We offer a brief description of the experimental building and LULC scene datasets in Table I.

1) Building scene datasets: The WHU Building dataset [37] is divided into two main components: one containing satellite imagery and another composed of aerial photos. In our study, we utilize the aerial photo subset, which consists of 8,189 images. These images are split into 4,736 for training, 1,036 for validation, and 2,416 for testing, each with a spatial resolution of 0.3 meters. In total, this subset represents over 22,000 buildings covering an area in excess of 450 square kilometers. Our experiments were conducted using the original partitioning scheme and image dimensions (512×512) as specified by the WHU dataset.

The WHU-CD dataset [37] includes bitemporal very high-resolution (VHR) aerial images taken in 2012 and 2016, which clearly highlight major changes in building structures. The dataset is partitioned into non-overlapping patches of 1024×1024 pixels. These patches are further allocated into training, validation, and test sets following a 7:1:2 ratio.

The LEVIR-CD dataset [25] is an extensive resource for change detection, comprising VHR Google Earth images with a resolution of 0.5 m/pixel. These images capture a variety of building transformations over periods ranging from 5 to 14 years, with a particular emphasis on construction and demolition events. The bitemporal images have been expertly annotated using binary masks, where a label of 1 denotes a change and 0 signifies no change. In total, there are 31,333 labeled instances of building modifications. Our experimental setup used the dataset's original image dimensions of 1024×1024 and adhered to the provided data partitioning scheme.

The TSCD dataset [38] is constructed from WorldView-2 satellite imagery with a spatial resolution of approximately 0.5 m/pixel, acquired in 2016, 2018, 2020, and 2022. To mitigate external influences, the images underwent co-registration using manually selected control points and resampling to ensure a consistent coordinate framework. Building footprints were densely labeled for each temporal phase. Subsequently, three sets of change labels (2016–2018, 2018–2020, 2020–2022) were generated by performing differential operations on adjacent building distribution maps. The final TSCD dataset was created through uniform cropping and partitioning of these original images and derived labels.

2) LULC scene datasets: The LoveDA dataset [39] consists of 5,987 high-resolution optical remote sensing images (with a ground sampling distance of 0.3 m) each sized at 1024×1024 pixels. It covers seven land cover classes: building, road, water, barren, forest, agriculture, and background. The dataset is divided into 2,522 training images, 1,669 images for validation, and 1,796 images for testing, all drawn from two distinct scenes—urban and rural—from three Chinese cities: Nanjing, Changzhou, and Wuhan. The dataset poses considerable challenges due to the presence of multiscale objects, complex backgrounds, and uneven class distribution.

The DynamicEarthnet dataset [40] comprises 55 daily Sentinel-2 Image Time Series (SITS) collected globally between January 1, 2018, and December 31, 2019. For each month, data from the first day is annotated, which results in 24 ground truth segmentation maps per Area of Interest (AoI). Each image is 1024×1024 pixels and multi-spectral, containing four channels (RGB plus near-infrared). The annotations cover general land-use and land-cover categories: impervious surface, agriculture, forest, wetlands, soil, and water. The 'snow' class appears in only a few AoIs and has been excluded from this study.

B. Baseline

To evaluate the effectiveness of the proposed TSSUN, we conducted comparative experiments with various benchmark

methods on the building and LULC scene datasets. The benchmark methods tested on the same dataset are based on the same splitting of the dataset and use the same data.

On the four building scene datasets, the compared CNN-based models include FCN [13], SegNet [41], U-Net [14], PSPNet [42], HRNet [43], MA-FCN [44], Deeplabv3+ [16], ResUNet [45], MAPNet [46], D-LinkNet [47], SIINet [48], FC-EF [20], FC-Siam-Diff [20], FC-Siam-Conc [20], STANet [25], DTCDSCN [49], SNUNet [50], CDNet [51], DDCNN [52], DASNet [53], DSIFN [54], HANet [55], USSFCNet [56], and SEIFNet [57], the compared transformer-based models include Segformer [58], ChangeFormer [59] and A2Net [60], and the CNN-transformer hybrid models include BDTNet [61], TransUNet [62], CMTFNet [63], BIT [64], MTCNet [65], MSCANet [66], AMTNet-50 [67], Contrast-COUD [38] and TS-COUD [38].

On the two LULC scene datasets, the compared CNN-based models include U-Net [14], Deepabv3+ [16], DANet [68], ResUNet-a [45], DASSN [69], HCANet [70], RAANet [71], A2-FPN [72] and CAC [73], the compared transformer-based models include LANet [74], SCAttNet [75], and TSViT [76], the CNN-transformer hybrid models include SAPNet [77], SSCBNet [78], UTAE [79], A2Net [80], SCanNet [81] and TSSCD [82].

C. Implementation Details

1) Data Augmentation: To validate the proposed methods, we adopted a minimalistic yet effective data augmentation strategy, deliberately refraining from complex augmentation schemes. Specifically, the employed transformations were limited to horizontal/vertical flipping (probability = 0.5) and transposition (probability = 0.5).

2) Training and Inference: The TSSUN model was implemented using PyTorch [83] and executed on a single RTX A100 GPU (80G). Due to the heterogeneous image resolutions across the datasets, the batch size was set to 16 for the four building scene datasets and 4 for the two LULC scene datasets. Our optimization strategy combined binary cross-entropy loss with Dice coefficient loss, facilitating a balanced performance optimization. The AdamW optimizer [84] was initialized with a learning rate of 0.0001 and a weight decay of 0.001. A learning rate scheduler was employed to reduce the learning rate by a factor of 0.1 if no increase in the mean F1-score was observed on the aggregate validation set for 5 consecutive epochs. The training process spanned 100 epochs, ensuring robust convergence, and the best performing checkpoints—corresponding to the maximum mean F1-scores achieved—were retained for the testing phase. Furthermore, in order to ensure comparability with existing methodologies, all models were initialized using the default PyTorch settings across all datasets.

3) Evaluation Metrics: The performance of the proposed models was quantitatively assessed using five principal metrics: overall accuracy (OA), precision (P), recall (R), F1-score, and intersection over union (IoU). For multi-temporal tasks and multi-category tasks, the average F1-score (AF) and mean IoU (mIoU) will be used. In addition, following the settings of the DynamicEarthnet dataset [40], we use the semantic change segmentation (SCS) metric, classagnostic binary change score (BC) and semantic segmentation score among changed pixels (SC) metrics to evaluate the model's performance on this dataset. Precision provides a measure of the incidence of false positives, while recall reflects the occurrence of false negatives. Due to the inherently inverse relationship between these two metrics, simultaneously securing high scores for both represents a significant challenge. The F1-score, calculated as the harmonic mean of precision and recall, offers a balanced assessment. The IoU metric evaluates spatial accuracy by quantifying the overlap between the predicted and ground truth changed pixels relative to their union.

D. Ablation Study

To ascertain the efficacy of the proposed Temporal-Spectral-Spatial Unified Strategy (TSSUS) and Local-Global Window Attention (LGWA), ablation studies were performed on the TSCD dataset. As comparative baselines, TSSUS was replaced with a strategy involving the direct unification of temporal, spectral, and spatial dimensions at both the input and output stages (referred to as "direct unification"). Concurrently, LGWA was substituted with standard global attention. This experimental design serves to highlight the distinct advantages of TSSUS in its independent extraction of temporal features and the capability of LGWA in the simultaneous capture of both local and global contextual information.

TABLE II
ACCURACY COMPARISON ON THE WHU DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Strategy	Attention	F1	IoU	OA
TSSUS	LGWA	66.48	49.79	98.05
Direct Unification	LGWA	63.22	46.22	94.72
TSSUS	Global Attention	65.91	49.15	97.21
Direct Unification	Global Attention	62.87	45.85	94.38

Table 2 presents the results of these ablation studies, indicating that both TSSUS and LGWA yield superior performance compared to their respective baseline alternatives. Specifically, TSSUS distinguishes itself from the direct unification approach. It begins by unifying the spectral and spatial dimensions at the input stage. Then, within the encoder, it independently extracts features for each temporal instance (i.e., each remote sensing image). Subsequently, these temporally distinct features are integrated via feature-level fusion at the interface between the encoder and decoder. This methodology—processing temporal information independently before fusion—effectively minimizes interference from irrelevant pixel information within the remote sensing imagery. Furthermore, TSSUS dynamically adjusts the temporal length of feature fusion to align with the specific requirements of the target dense prediction task. Following this, it optimizes the output for each time step independently within the decoder, thereby significantly enhancing the overall model efficacy.

In a similar vein, LGWA offers advantages over standard global self-attention. By employing a combination of variously shaped local windows alongside global self-attention mechanisms, LGWA is capable of concurrently extracting a rich

tapestry of local features and comprehensive global context. This simultaneous extraction of multi-level features is crucial for enabling the model to effectively perform dense prediction tasks across various scales.

E. Overall Comparison

1) Building: The efficacy of the proposed Temporal-Spectral-Spatial Unified Network was rigorously evaluated through extensive experiments on four benchmark remote sensing datasets: the WHU dataset for single-temporal semantic segmentation, the WHU-CD dataset for bi-temporal semantic change detection, the LEVIR-CD dataset for bi-temporal binary change detection, and the TSCD dataset for multi-temporal binary change detection. Our method was compared against several state-of-the-art approaches, with quantitative results summarized in Table III, IV, V and VI.

On the WHU dataset, characterized by its complex building footprints and significant variations in object scale, our proposed TSSUN achieves state-of-the-art performance. As detailed in Table III, TSSUN obtains the highest IoU of 91.00% and an F1-score of 95.29%. This superior performance can be attributed to TSSUN's Temporal-Spectral-Spatial Unified Strategy, particularly its spatial unification component, which effectively processes varying spatial resolutions, and the Local-Global Window Attention mechanism. The LGWA module, with its ability to capture both fine-grained local details and broader contextual information, is particularly adept at delineating intricate building boundaries and accurately segmenting buildings of diverse sizes.

TABLE III
ACCURACY COMPARISON ON THE WHU DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Methods	P (%)	R (%)	F1 (%)	IoU (%)
FCN [13]	92.29	92.84	92.56	86.16
SegNet [41]	93.42	91.71	92.56	86.15
U-Net [14]	94.50	90.88	92.65	86.31
PSPNet [42]	93.19	94.21	93.70	88.14
HRNet [43]	91.69	92.85	92.27	85.64
MA-FCN [44]	94.75	94.92	94.83	90.18
Deeplabv3+ [16]	94.31	94.53	94.42	89.43
ResUNet [45]	94.49	94.71	94.60	89.75
MAP-Net [46]	93.99	94.82	94.40	89.40
Segformer [58]	94.72	94.42	94.57	89.70
TransUNet [62]	94.05	93.07	93.56	87.89
CMTFNet [63]	90.12	**95.21**	92.59	86.21
TSSUN	**95.71**	94.87	**95.29**	**91.00**

For the bi-temporal semantic change detection task on the WHU-CD dataset, TSSUN demonstrates leading performance against other SOTA methods, as shown in Table IV. The WHU-CD dataset demands accurate identification of 'from-to' semantic transitions between two time points. TSSUN excels here due to its inherent design for unified temporal and spectral modeling. The TSSUS component allows the network to effectively learn temporal evolutionary patterns and relationships between different land cover classes, while the LGWA mechanism enhances the discrimination of subtle yet significant semantic changes from unchanged areas, leading to more precise change maps.

TABLE IV
ACCURACY COMPARISON ON THE WHU-CD DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Methods	P (%)	R (%)	F1 (%)	IoU (%)
FCN [13]	79.35	77.82	78.58	64.71
SegNet [41]	85.20	86.21	85.70	74.98
Deeplabv3+ [16]	89.24	90.91	90.07	81.93
U-Net [14]	83.19	84.02	83.60	71.83
PSPNet [42]	84.85	82.09	83.45	71.60
HRNet [43]	86.77	85.92	86.34	75.97
MA-FCN [44]	86.10	89.92	87.97	78.52
Segformer [58]	90.45	88.93	89.68	81.30
TransUNet [62]	**93.82**	89.33	91.52	84.37
RSM-CD	93.07	**91.20**	**92.13**	**85.40**

In the context of bi-temporal binary change detection on the LEVIR-CD dataset, which features a large number of buildings of various sizes and styles undergoing changes, TSSUN surpasses existing methods. Table V shows that our method achieves the highest F1-score of 91.59% and an IoU of 84.49%. The strength of TSSUN on this dataset lies in its robust temporal modeling capabilities, facilitated by TSSUS, which allows for consistent feature representation across different time points. Furthermore, the LGWA mechanism's proficiency in extracting salient local changes while considering global context ensures high accuracy in detecting both small and large-scale building changes, minimizing missed detections and false alarms often encountered with heterogeneous scene elements.

TABLE V
ACCURACY COMPARISON ON THE LEVIR-CD DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Methods	P (%)	R (%)	F1 (%)	IoU (%)
FC-EF [20]	86.91	80.17	83.40	71.53
FC-Siam-Diff [20]	89.53	83.31	86.31	75.91
FC-Siam-Conc [20]	91.99	76.77	83.69	71.96
DTCDSCN [49]	88.53	86.83	87.67	78.05
DSIFN [54]	**94.02**	82.93	88.13	78.77
STANet [25]	83.81	**91.00**	87.26	77.39
SNUNet [50]	89.18	87.17	88.16	78.83
HANet [55]	91.21	89.36	90.28	82.27
CDNet [51]	91.60	86.50	88.98	80.14
DDCNN [52]	91.85	88.69	90.24	82.22
BIT [64]	89.24	89.37	89.30	80.68
ChangeFormer [59]	92.05	88.80	90.40	82.47
MTCNet [65]	90.87	89.62	90.24	82.22
MSCANet [66]	91.30	88.56	89.91	81.66
AMTNet-50 [67]	91.82	89.71	90.76	83.08
TSSUN	93.17	90.07	**91.59**	**84.49**

Finally, on the TSCD dataset, which presents a challenging multi-temporal binary change detection scenario with longer image sequences, TSSUN achieves the best results, as indicated in Table VI, with an F1-score of 66.48% and an IoU of 49.79%. The TSCD dataset requires robust modeling of temporal dependencies across multiple observations. TSSUS is specifically designed to handle inputs with varying temporal lengths and to model the continuous evolution of land cover. This, combined with the LGWA's capacity to aggregate contextual information effectively over extended

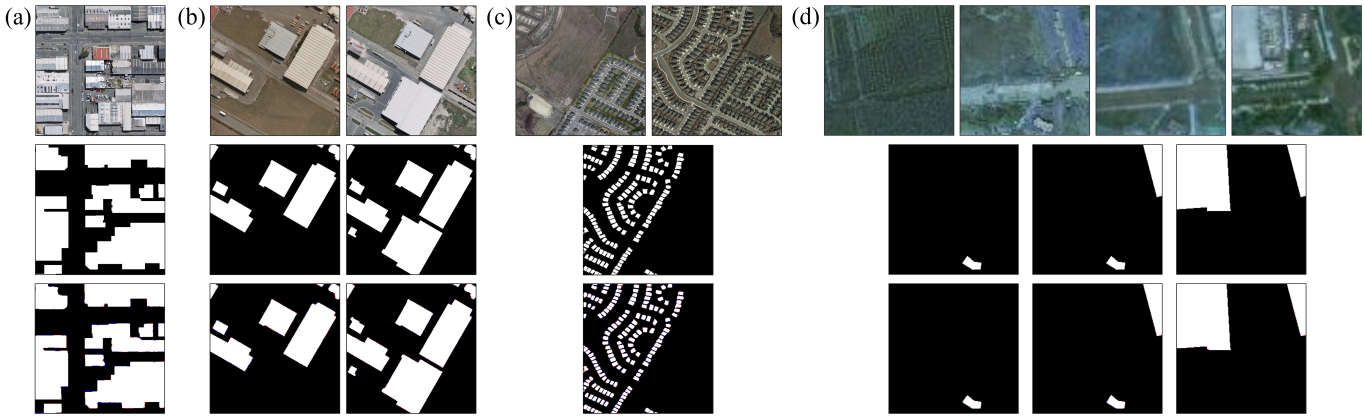

Fig. 4. Sample inference results on for building scene datasets. The input images, ground truths and predictions are shown in the first, second and third rows, respectively. Red areas denote false positives and blue areas denote false negatives. (a) WHU dataset sample. (b) WHU-CD dataset sample. (c) LEIVR-CD dataset sample. (d) TSCD dataset sample.

temporal sequences, enables TSSUN to accurately identify changes in complex, evolving landscapes where other methods might falter due to the increased temporal dimensionality.

TABLE VI
ACCURACY COMPARISON ON THE TSCD DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Methods	F1 (%)	IoU (%)	OA (%)
FC-EF [20]	53.39	37.86	97.03
FC-Siam-Conc [20]	41.51	28.05	96.60
FC-Siam-Diff [20]	39.65	26.83	96.48
SNUNet-CD [50]	63.22	47.22	97.61
USSFCNet [56]	55.68	39.80	97.30
A2Net [60]	53.16	37.19	97.14
SEIFNet [57]	60.37	44.01	97.41
Contrast-COUD [38]	64.35	48.45	97.72
TS-COUD [38]	65.24	49.33	97.90
TSSUN	**66.48**	**49.79**	**98.05**

Figure 4 presents a qualitative comparison of our method against other SOTA approaches on representative samples from all four datasets. Visually, TSSUN consistently produces more accurate and complete segmentation and change maps. The results exhibit fewer false positives and false negatives, particularly in challenging areas such as those with small objects, intricate boundaries, or subtle temporal variations. For instance, in building segmentation on the WHU dataset (Figure 4(a)), TSSUN generates sharper edges and more complete building shapes. Similarly, for change detection tasks on WHU-CD (Figure 4(b)), LEVIR-CD (Figure 4(c)), and TSCD (Figure 4(d)), our method demonstrates superior capability in precisely localizing changed regions while maintaining the integrity of unchanged areas. This visual superiority can be attributed to the model's enhanced capability, derived from the synergistic operation of TSSUS and LGWA, to learn highly discriminative features and effectively model contextual relationships across the temporal, spectral, and spatial dimensions, resulting in outputs that are more coherent and closely aligned with ground truth.

2) LULC: The efficacy of the proposed Temporal-Spectral-Spatial Unified Network is rigorously evaluated against SOTA methods on two challenging benchmark datasets: LoveDA for single-temporal semantic segmentation and DynamicEarthNet for multi-temporal land cover classification.

On the LoveDA dataset, TSSUN demonstrates superior performance, achieving the highest OA of 71.82% and mIoU of 65.73% as shown in Table VII. This leading performance can be attributed to TSSUN's architecture, particularly the Local-Global Window Attention mechanism, which effectively captures both fine-grained local details and broader contextual information. This capability is crucial for accurately segmenting multiscale objects and navigating the complex scenes and backgrounds characteristic of the high-resolution LoveDA imagery. Furthermore, the Temporal-Spectral-Spatial Unified Strategy enables robust feature representation by adeptly unifying spatial and spectral information, which is vital for handling the heterogeneity present in such diverse urban and rural landscapes with uneven class distributions.

For the multi-temporal DynamicEarthNet dataset, TSSUN again surpasses existing methods, yielding the top scores across all reported metrics: Semantic Change Segmentation (SCS) score of 29.9, class-agnostic Binary Change score (BC) of 38.9, and mIoU of 54.7 as shown in Table VIII. The success of TSSUN on this dataset underscores the effectiveness of the TSSUS in explicitly modeling the temporal dimension. By leveraging temporal meta-information to ensure consistency across sequences of varying lengths, TSSUN adeptly manages long-sequence image time series and discerns subtle land cover changes despite spectral variations across different time points—a key challenge in DynamicEarthNet. This robust temporal modeling, distinct from its spatial feature extraction strengths highlighted in the LoveDA analysis, allows for consistent and accurate classification and change analysis across the 24 annotated time steps.

V. DISCUSSION

The proliferation of diverse remote sensing data sources necessitates models capable of unified processing across varied data characteristics and task requirements. This research confronts the critical challenge of existing deep learning

TABLE VII
ACCURACY COMPARISON ON THE LOVEDA DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Methods	F1-score per category (%)							AF	OA	mIoU
	Background	Building	Road	Water	Barren	Forest	Agriculture			
U-Net [14]	50.21	54.74	56.38	77.12	18.09	48.93	66.05	53.07	51.81	47.84
DeepLabV3+ [16]	52.29	54.99	57.16	77.96	16.11	48.18	67.79	53.50	52.30	47.62
DANet [68]	54.47	61.02	63.37	79.17	26.63	52.28	70.02	58.14	54.64	50.18
ResUNet-a [45]	59.16	64.08	66.73	81.01	32.23	55.81	75.79	62.12	59.65	54.16
DASSN [69]	57.95	66.90	68.63	76.64	44.35	54.96	70.49	62.85	60.35	55.42
HCANet [70]	66.39	70.76	75.11	88.29	51.14	63.92	81.07	70.95	69.47	62.77
RAANet [71]	55.02	62.19	65.58	81.03	29.25	54.11	74.07	60.18	58.95	53.93
SCAttNet [75]	65.95	71.88	77.04	86.61	50.79	61.19	82.00	70.78	67.31	61.09
A2FPN [72]	65.17	73.32	75.19	88.01	48.82	59.96	79.71	70.03	66.89	61.14
LANet [74]	67.04	74.19	77.54	87.54	52.23	64.78	80.80	72.02	69.11	62.16
MSAFNet [85]	65.51	73.71	75.59	88.47	49.08	60.28	80.13	70.40	67.17	60.76
CLCFormer [86]	67.17	74.34	77.69	87.71	52.34	64.91	80.96	72.16	69.37	63.85
SAPNet [77]	67.50	75.06	78.12	88.35	53.10	65.50	81.30	73.04	70.12	63.45
SSCBNet [78]	68.25	75.90	**79.00**	89.10	54.00	66.20	82.15	74.05	70.95	64.58
TSSUN	**68.72**	**76.13**	78.83	**90.21**	**54.73**	**67.03**	**82.67**	**74.81**	**71.82**	**65.73**

TABLE VIII
ACCURACY COMPARISON ON THE DYNAMICEARTHNET DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Methods	SCS↑	BC↑	SC↑	mIoU↑
CAC [73]	17.7	10.7	**24.7**	37.9
U-Net [14]	17.3	10.1	24.4	37.6
TSViT [76]	23.0	34.1	11.8	50.5
UTAE [79]	25.9	38.0	13.8	53.7
A2Net [80]	22.2	32.9	11.5	47.2
SCanNet [81]	24.8	35.8	13.9	53.0
TSSCD [82]	12.0	19.4	4.7	33.9
TSSUN	**29.9**	**38.9**	21.8	**54.7**

models in remote sensing dense prediction, which are typically constrained to fixed input-output configurations and struggle to generalize across the inherent heterogeneity in temporal length, spectral channels, and spatial resolution of imagery, as well as diverse task specifications such as semantic segmentation, semantic change detection, and binary change detection. This rigidity imposes significant burdens in terms of retraining and fine-tuning for each specific scenario, limiting practical applicability in dynamic operational environments.

To surmount this limitation, we proposed the Temporal-Spectral-Spatial Unified Network, a novel architecture engineered for comprehensive unification. The cornerstone of TSSUN is the Temporal-Spectral-Spatial Unified Strategy, which systematically decouples and standardizes data representations. On the input side, TSSUS leverages dimension-specific meta-information to harmonize temporal sequences of varying lengths, align spectral inputs with differing channel counts, and reconcile features from disparate spatial resolutions. This is achieved by modeling the continuity in temporal evolution, the inherent relationships between spectral bands, and the local similarity in spatial regions. Concurrently, on the output side, TSSUS exploits shared structural elements among the three primary dense prediction tasks. It supports variable temporal output lengths by identifying common temporal patterns, accommodates an arbitrary number of output classes by capturing inter-class relationships, and ensures spatial consistency across resolutions. Furthermore, the integration of a Local-Global Window Attention mechanism enhances TSSUN's capacity. LGWA efficiently processes local contextual information through overlapping windowed attention blocks and aggregates global context via a subsequent global attention layer, striking a balance between computational load and feature representation power essential for complex remote sensing scenes.

The efficacy of the TSSUN framework in addressing the identified research problem is substantiated by extensive experimental validation. Results across multiple datasets, encompassing diverse Temporal-Spectral-Spatial configurations and tasks like building detection and land use/land cover classification, demonstrate that a single, pre-trained TSSUN model can successfully adapt to heterogeneous inputs and unify these distinct dense prediction tasks, consistently achieving performance comparable to or exceeding state-of-the-art specialized models.

Despite these advancements, the current TSSUN framework possesses certain limitations. Firstly, the TSSUS component relies on the availability and accuracy of explicit meta-information (e.g., acquisition dates, sensor band specifications, ground sampling distance) to guide the unification across temporal, spectral, and spatial dimensions. The performance may be suboptimal if such meta-information is imprecise, incomplete, or unavailable, potentially hindering the model's adaptability in data-scarce or poorly documented scenarios. Secondly, while the LGWA mechanism is designed for efficiency, the computational demand of its global attention component, although optimized, might still present a scalability challenge when processing extremely large-scale images or exceptionally long and dense temporal sequences, potentially requiring substantial memory and processing resources.

Future research will focus on mitigating these limitations and expanding the capabilities of TSSUN. One promising direction involves developing methods for the implicit learning or unsupervised inference of the requisite meta-information directly from the remote sensing data, thereby reducing dependency on external inputs and enhancing robustness. Another avenue for exploration is the refinement of the LGWA module, possibly by incorporating more sophisticated hierarchical attention structures or adaptive computational pruning techniques to further improve its efficiency and scalability for massive datasets. Additionally, future work could extend the

unification paradigm of TSSUN to encompass a broader array of remote sensing tasks and potentially integrate other data modalities, such as SAR or LiDAR, within a unified analytical framework.

VI. CONCLUSION

This paper addressed the critical challenge of processing heterogeneous remote sensing data for diverse dense prediction tasks, a common hurdle due to variations in temporal, spectral, and spatial characteristics. We introduced the Temporal-Spectral-Spatial Unified Network, a novel deep learning architecture. TSSUN demonstrates remarkable flexibility by handling arbitrary input configurations across these three dimensions and unifying key dense prediction tasks: semantic segmentation, semantic change detection, and binary change detection, with adaptable output class settings. The proposed Temporal-Spectral-Spatial Unified Strategy enables this unification by effectively decoupling and modeling data representations guided by meta-information. Furthermore, the designed Local-Global Window Attention mechanism significantly improves feature extraction by capturing both local and global contextual information efficiently. Comprehensive experimental validation on multiple datasets confirmed that a single TSSUN model consistently adapts to varied data inputs and task requirements, achieving or exceeding state-of-the-art performance. The TSSUN framework represents a substantial advancement towards universal remote sensing models, mitigating the need for task-specific designs and extensive retraining, thus paving the way for more scalable and cost-effective analysis of complex, multi-source geospatial data and fostering broader application of remote sensing technologies.

REFERENCES

[1] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote sensing of Environment*, vol. 202, pp. 18–27, 2017.

[2] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE geoscience and remote sensing magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[3] H. Taubenböck, T. Esch, A. Felbier, M. Wiesner, A. Roth, and S. Dech, "Monitoring urbanization in mega cities from space," *Remote sensing of Environment*, vol. 117, pp. 162–176, 2012.

[4] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland *et al.*, "High-resolution global maps of 21st-century forest cover change," *science*, vol. 342, no. 6160, pp. 850–853, 2013.

[5] S. Voigt, F. Giulio-Tonolo, J. Lyons, J. Kučera, B. Jones, T. Schneiderhan, G. Platzeck, K. Kaku, M. K. Hazarika, L. Czaran *et al.*, "Global trends in satellite-based emergency mapping," *Science*, vol. 353, no. 6296, pp. 247–252, 2016.

[6] J. Dong, X. Xiao, M. A. Menarguez, G. Zhang, Y. Qin, D. Thau, C. Biradar, and B. Moore III, "Mapping paddy rice planting area in northeastern asia with landsat 8 images, phenology-based algorithm and google earth engine," *Remote sensing of environment*, vol. 185, pp. 142–154, 2016.

[7] A. Van Donkelaar, R. V. Martin, M. Brauer, R. Kahn, R. Levy, C. Verduzco, and P. J. Villeneuve, "Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application," *Environmental health perspectives*, vol. 118, no. 6, pp. 847–855, 2010.

[8] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and F. Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[9] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "Fccdn: Feature constraint network for vhr image change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 187, pp. 101–119, 2022.

[10] S. Zhao, X. Zhang, P. Xiao, and G. He, "Exchanging dual-encoder–decoder: A new strategy for change detection with semantic guidance and spatial localization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[11] J. Lv, Q. Shen, M. Lv, Y. Li, L. Shi, and P. Zhang, "Deep learning-based semantic segmentation of remote sensing images: a review," *Frontiers in Ecology and Evolution*, vol. 11, p. 1201125, 2023.

[12] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS journal of photogrammetry and remote sensing*, vol. 66, no. 3, pp. 247–259, 2011.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[16] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[17] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *International journal of remote sensing*, vol. 10, no. 6, pp. 989–1003, 1989.

[18] A. A. Nielsen, "The regularized iteratively reweighted mad method for change detection in multi-and hyperspectral data," *IEEE Transactions on Image processing*, vol. 16, no. 2, pp. 463–478, 2007.

[19] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 218–236, 2006.

[20] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[21] S. Zhao, H. Chen, X. Zhang, P. Xiao, L. Bai, and W. Ouyang, "Rs-mamba for large remote sensing image dense prediction," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[22] T. Chen, Z. Lu, Y. Yang, Y. Zhang, B. Du, and A. Plaza, "A siamese network based u-net for change detection in high resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2357–2369, 2022.

[23] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Transactions on Geoscience and Remote sensing*, vol. 38, no. 3, pp. 1171–1182, 2000.

[24] J.-F. Mas, "Monitoring land-cover changes: a comparison of change detection techniques," *International journal of remote sensing*, vol. 20, no. 1, pp. 139–152, 1999.

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

[26] F. E. Fassnacht, J. C. White, M. A. Wulder, and E. Næsset, "Remote sensing in forestry: current challenges, considerations and directions," *Forestry: An International Journal of Forest Research*, vol. 97, no. 1, pp. 11–37, 2024.

[27] C. O. Justice, E. Vermote, J. R. Townshend, R. Defries, D. P. Roy, D. K. Hall, V. V. Salomonson, J. L. Privette, G. Riggs, A. Strahler *et al.*, "The moderate resolution imaging spectroradiometer (modis): Land remote sensing for global change research," *IEEE transactions on geoscience and remote sensing*, vol. 36, no. 4, pp. 1228–1249, 1998.

[28] R. Torres, P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Rommen, N. Floury, M. Brown *et al.*, "Gmes sentinel-1 mission," *Remote sensing of environment*, vol. 120, pp. 9–24, 2012.

[29] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort *et al.*, "Sentinel-2: Esa's optical high-resolution mission for gmes operational services," *Remote sensing of Environment*, vol. 120, pp. 25–36, 2012.

[30] S. N. Goward, J. G. Masek, D. L. Williams, J. R. Irons, and R. Thompson, "The landsat 7 mission: Terrestrial research and applications for the 21st century," *Remote Sensing of Environment*, vol. 78, no. 1-2, pp. 3–12, 2001.

[31] M. Belgiu and O. Csillik, "Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis," *Remote sensing of environment*, vol. 204, pp. 509–523, 2018.

[32] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis *et al.*, "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (aviris)," *Remote sensing of environment*, vol. 65, no. 3, pp. 227–248, 1998.

[33] L. Guanter, H. Kaufmann, K. Segl, S. Foerster, C. Rogass, S. Chabrillat, T. Kuester, A. Hollstein, G. Rossner, C. Chlebek *et al.*, "The enmap spaceborne imaging spectroscopy mission for earth observation," *Remote Sensing*, vol. 7, no. 7, pp. 8830–8857, 2015.

[34] M. Yang, L. Jiao, B. Hou, F. Liu, and S. Yang, "Selective adversarial adaptation-based cross-scene change detection framework in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2188–2203, 2020.

[35] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 60–77, 2018.

[36] P. Gong, H. Liu, M. Zhang, C. Li, J. Wang, H. Huang, N. Clinton, L. Ji, W. Li, Y. Bai *et al.*, "Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017," *Sci. Bull*, vol. 64, no. 6, pp. 370–373, 2019.

[37] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on geoscience and remote sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[38] Y. Zhao, H.-C. Li, S. Lei, N. Liu, J. Pan, and T. Celik, "Coud: Continual urbanization detector for time series building change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[39] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," *arXiv preprint arXiv:2110.08733*, 2021.

[40] A. Toker, L. Kondmann, M. Weber, M. Eisenberger, A. Camero, J. Hu, A. P. Hoderlein, Ç. Şenaras, T. Davis, D. Cremers *et al.*, "Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 158–21 167.

[41] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[43] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[44] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using cnn and regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 2178–2189, 2019.

[45] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[46] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "Map-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 6169–6181, 2020.

[47] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 182–186.

[48] C. Tao, J. Qi, Y. Li, H. Wang, and H. Li, "Spatial information inference net: Road extraction using road-specific contextual information," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 155–166, 2019.

[49] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2020.

[50] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[51] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[52] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7296–7307, 2020.

[53] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2020.

[54] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.

[55] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, "Hanet: A hierarchical attention network for change detection with bi-temporal very-high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[56] T. Lei, X. Geng, H. Ning, Z. Lv, M. Gong, Y. Jin, and A. K. Nandi, "Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[57] Y. Huang, X. Li, Z. Du, and H. Shen, "Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[58] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[59] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 207–210.

[60] Z. Li, C. Tang, X. Liu, W. Zhang, J. Dou, L. Wang, and A. Y. Zomaya, "Lightweight remote sensing change detection with progressive feature aggregation and supervised attention," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[61] L. Luo, J.-X. Wang, S.-B. Chen, J. Tang, and B. Luo, "Bdtnet: Road extraction by bi-direction transformer from remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[62] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[63] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "Cmtfnet: Cnn and multiscale transformer fusion network for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[64] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[65] W. Wang, X. Tan, P. Zhang, and X. Wang, "A cbam based multiscale transformer fusion approach for remote sensing image change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6817–6825, 2022.

[66] M. Liu, Z. Chai, H. Deng, and R. Liu, "A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4297–4306, 2022.

[67] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, "An attention-based multiscale transformer network for remote sensing image change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 599–609, 2023.

[68] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[69] X. Li, F. Xu, X. Lyu, H. Gao, Y. Tong, S. Cai, S. Li, and D. Liu, "Dual attention deep fusion semantic segmentation networks of large-

scale satellite remote-sensing images," *International Journal of Remote Sensing*, vol. 42, no. 9, pp. 3583–3610, 2021.
[70] X. Li, F. Xu, R. Xia, X. Lyu, H. Gao, and Y. Tong, "Hybridizing cross-level contextual and attentive representations for remote sensing imagery semantic segmentation," *Remote Sensing*, vol. 13, no. 15, p. 2986, 2021.
[71] R. Liu, F. Tao, X. Liu, J. Na, H. Leng, J. Wu, and T. Zhou, "Raanet: A residual aspp with attention framework for semantic segmentation of high-resolution remote sensing images," *Remote Sensing*, vol. 14, no. 13, p. 3109, 2022.
[72] R. Li, L. Wang, C. Zhang, C. Duan, and S. Zheng, "A2-fpn for semantic segmentation of fine-resolution remotely sensed images," *International journal of remote sensing*, vol. 43, no. 3, pp. 1131–1155, 2022.
[73] X. Lai, Z. Tian, L. Jiang, S. Liu, H. Zhao, L. Wang, and J. Jia, "Semi-supervised semantic segmentation with directional context-aware consistency," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1205–1214.
[74] L. Ding, H. Tang, and L. Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 426–435, 2020.
[75] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 905–909, 2020.
[76] M. Tarasiou, E. Chavez, and S. Zafeiriou, "Vits for sits: Vision transformers for satellite image time series," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 418–10 428.
[77] X. Li, F. Xu, F. Liu, X. Lyu, Y. Tong, Z. Xu, and J. Zhou, "A synergistical attention model for semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.
[78] X. Li, X. Yong, T. Li, Y. Tong, H. Gao, X. Wang, Z. Xu, Y. Fang, Q. You, and X. Lyu, "A spectral–spatial context-boosted network for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 16, no. 7, p. 1214, 2024.
[79] V. S. F. Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4872–4881.
[80] Z. Li, C. Tang, X. Liu, W. Zhang, J. Dou, L. Wang, and A. Y. Zomaya, "Lightweight remote sensing change detection with progressive feature aggregation and supervised attention," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.
[81] L. Ding, J. Zhang, H. Guo, K. Zhang, B. Liu, and L. Bruzzone, "Joint spatio-temporal modeling for semantic change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
[82] H. He, J. Yan, D. Liang, Z. Sun, J. Li, and L. Wang, "Time-series land cover change detection using deep learning-based temporal semantic segmentation," *Remote Sensing of Environment*, vol. 305, p. 114101, 2024.
[83] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.
[84] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[85] X. Lyu, W. Jiang, X. Li, Y. Fang, Z. Xu, and X. Wang, "Msafnet: Multiscale successive attention fusion network for water body extraction of remote sensing images," *Remote Sensing*, vol. 15, no. 12, p. 3121, 2023.
[86] J. Long, M. Li, and X. Wang, "Integrating spatial details with long-range contexts for semantic segmentation of very high-resolution remote-sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.