Remote Sensing Temporal Vision-Language Models: A Comprehensive Survey

Chenyang Liu, Jiafan Zhang, Keyan Chen, Man Wang, Zhengxia Zou, *Senior Member, IEEE*, and Zhenwei Shi*, *Senior Member, IEEE*

Beihang University

Abstract*—Temporal image analysis in remote sensing has traditionally centered on change detection, which identifies regions of change between images captured at different times. However, change detection remains limited by its focus on visual-level interpretation, often lacking contextual or descriptive information. The rise of Vision-Language Models (VLMs) has introduced a new dimension to remote sensing temporal image analysis by integrating visual information with natural language, creating an avenue for advanced interpretation of temporal image changes. Remote Sensing Temporal VLMs (RSTVLMs) allow for dynamic interactions, generating descriptive captions, answering questions, and providing a richer semantic understanding of temporal images. This temporal vision-language capability is particularly valuable for complex remote sensing applications, where higher-level insights are crucial. This paper comprehensively reviews the progress of RSTVLM research, with a focus on the latest VLM applications for temporal image analysis. We categorize and discuss core methodologies, datasets, and metrics, highlight recent advances in temporal vision-language tasks, and outline key challenges and future directions for research in this emerging field. This survey fills a critical gap in the literature by providing an integrated overview of RSTVLM, offering a foundation for further advancements in remote sensing temporal image understanding. We will keep tracing related works at *https://github.com/Chen-Yang-Liu/Awesome-RS-Temporal-VLM

***Index Terms*—Remote Sensing, Temporal Image Understanding, Vision-Language Model, and Large Language Model.**

I. Introduction

REMOTE sensing technology acquires the Earth's surface image information through various platforms such as satellites and drones [1]–[4]. It plays a crucial role in key areas including environmental monitoring, urban planning, disaster warning and assessment [5]–[8]. Earlier remote sensing image interpretation primarily focused on the analysis of single-temporal images, including tasks such as land cover classification [9], [10], object detection [11], [12], and semantic segmentation [13], [14]. However, a single-temporal image reflects only the surface conditions at a specific moment and fails to capture dynamic changes across time.

With the rapid advancements in remote sensing technology and equipment, the capability to acquire multi-temporal remote sensing images has been significantly improved [15]–[17]. Multi-temporal remote sensing images provide surface feature information at certain locations across different time points, offering new avenues for dynamic monitoring of surface changes [18], [19]. This temporal dimension is critical, as it allows researchers to analyze trends over time, leading to a more comprehensive understanding of environmental dynamics. Early research on temporal image understanding mainly focused on change detection technology, which locates the changed area by comparing images from different periods, such as vegetation cover changes or the emergence of new buildings [20], [21]. However, change detection often only detects the location of changes at the visual level, and lacks a comprehensive higher-level semantic understanding of the changes, such as the type of changing object, the changing state over time, and the relationship between objects [22]–[24].

Recently, vision-language models such as Llava [25] and GPT-4 [26] have achieved groundbreaking advancements, leading to a growing interest in Vision-Language Models (VLMs) within the multi-modal research domain [27]–[29]. VLMs integrate computer vision and natural language processing technologies, facilitating a comprehensive understanding of both visual and textual information. Different from visual models that focus on a single image modality, VLMs not only identify target objects within images but also comprehend the relationships among them, generating descriptive language or answering questions. This capability holds immense potential for applications in the remote sensing field [30]–[32]. In the remote sensing community, previous studies have explored various VLMs, such as image captioning [33]–[36], visual question answering (VQA) [37]–[40], visual question generation [41], [42], text-to-image retrieval [43]–[45], and visual grounding [46]–[48]. Some recent studies have explored remote sensing visual language models based on large language models (LLMs), such as RSGPT [49], GeoChat [50], H2RSVLM [51], LHRS-Bot [52] and EarthGPT [53]. However, these VLMs mainly focus on single-temporal remote sensing images and cannot achieve multi-temporal remote

The work was supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160401), the National Natural Science Foundation of China under Grant 62125102, the Beijing Natural Science Foundation under Grant JL23005, and the Fundamental Research Funds for the Central Universities. *(Corresponding author: Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn))*

Chenyang Liu, Jiafan Zhang, Keyan Chen, Man Wang, and Zhenwei Shi are with the Image Processing Center, School of Astronautics, with the Beijing Key Laboratory of Digital Media, and with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Chenyang Liu is also with Shen Yuan Honors College of Beihang University, Beijing 100191, China.

Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

最终版

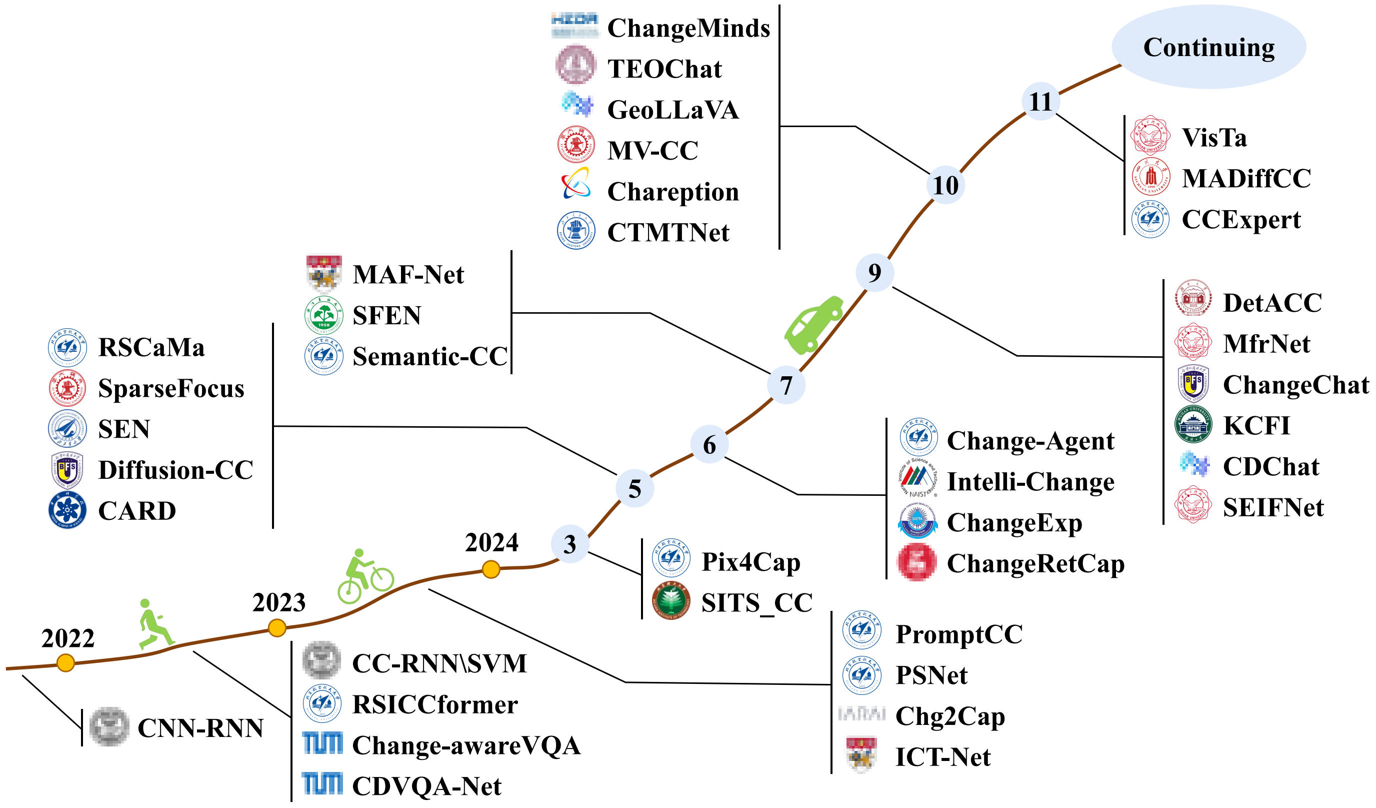


Fig. 1. Timeline of representative RS-TVLMs. This field is experiencing rapid growth. For additional resources and daily updates, visit our GitHub page at *https://github.com/Chen-Yang-Liu/Awesome-RS-Temporal-VLM*

sensing image understanding.

As VLMs evolve, research on multi-temporal remote sensing images has entered a new developmental stage. Researchers are increasingly exploring Remote Sensing Temporal VLMs (RS-TVLMs) specifically designed for temporal image understanding, addressing tasks such as change captioning [22], [23]and change visual question answering [54], [55]. The research on RS-TVLMs enriches the tools available for temporal image interpretation. Language, as a vehicle for human communication and knowledge [56], enhances the model's higher-level understanding when incorporated into the analysis of temporal images. By integrating temporal visual information with language, RS-TVLMs can recognize targets and changes, generate descriptions, answer relevant questions, and engage in multi-modal interactions, thus extending the temporal image interpretation beyond mere visual judgments.

Fig. 1 illustrates some representative RS-TVLMs along with their publication dates, indicating that research in this area can be traced back to 2021. Currently, the number of related studies is rapidly increasing. Despite the growing interest in RS-TVLMs, systematic reviews remain scarce. Many existing studies focus on isolated methods for specific tasks, making it challenging for researchers to gain a holistic view of the field's progress and future directions.

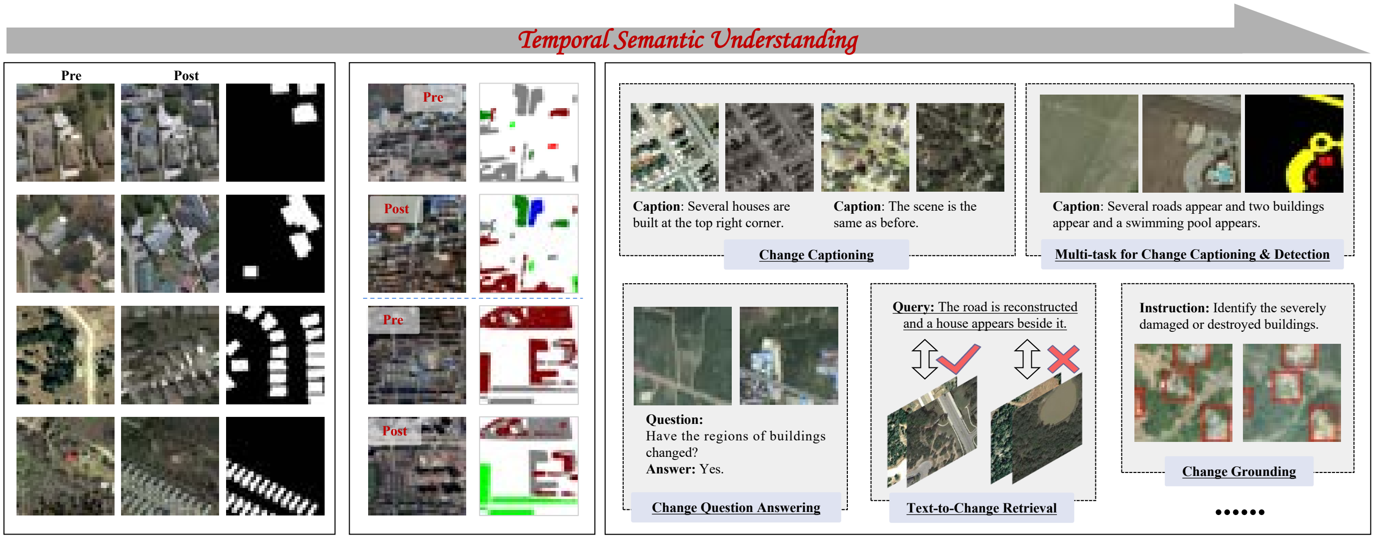
Contribution. In light of the rapid advancements and promising development of RS-TVLMs, we have composed this survey to acquaint researchers with the fundamental concepts, principal methodologies, datasets, evaluation metrics, and current progress in tasks such as change captioning and change visual question answering. To the best of our knowledge, we are the first survey on RS-TVLMs. By reviewing existing studies, we seek to outline clear pathways and future directions for research in this domain, addressing a current gap in related reviews and establishing a foundation for future research on RS-TVLMs for remote sensing temporal image understanding.

II. From Change Detection to Temporal Vision-Language Understanding

A. Remote Sensing Change Detection

Change detection (CD) is a basic task in the analysis of temporal remote sensing imagery, aimed at comparing remote sensing images taken at different times to identify pixel-level change area masks [57]–[59]. Based on the type of masks, change detection can be classified into two categories: binary change detection and semantic change detection. Early research focused primarily on binary change detection [60]–[62], which determines areas of change but lacks deeper semantic information. To address this limitation, researchers have developed semantic change detection methods that not only locate regions of change but also identify the types of land cover within those areas [63]–[65]. For example, in monitoring urban expansion, semantic change detection models can distinguish between buildings, roads, or green spaces that appear before and after the change, rather than merely identifying the locations where changes occurred.

Change detection meets visual question answering


Temporal Semantic Understanding
Pre
Post
Pre
Post
Pre
Post
Caption: Several houses are built at the top right corner.
Caption: The scene is the same as before.
Change Captioning
Caption: Several roads appear and two buildings appear and a swimming pool appears.
Multi-task for Change Captioning & Detection
Question: Have the regions of buildings changed?
Answer: Yes.
Change Question Answering
Query: The road is reconstructed and a house appears beside it.
Text-to-Change Retrieval
Instruction: Identify the severely damaged or destroyed buildings.
Change Grounding
••••••

(a) Binary Change Detection (b) Semantic Change Detection (c) Vision-Language Understanding

Fig. 2. Three concepts in temporal image understanding: (a) Binary change detection, locating areas of change without providing deeper contextual information. (b) Semantic change detection, offering insights into the types of changes. (c) Vision-Language Understanding, incorporating high-level language rather than being limited to visual-level interpretation.

Early change detection approaches were traditionally categorized into algebra-based, transformation-based, and classification-based methods. These methods relied on techniques such as change vector analysis (CVA) [66], [67], principal component analysis (PCA) [68], [69], multivariate alteration detection (MAD) [70], [71], and post-classification comparisons [72], [73]. Traditional methods have laid the foundation for modern change detection techniques [74]. With the emergence of deep learning, the field of remote sensing change detection has witnessed significant advancements [75]–[78]. Compared to traditional methods, deep learning significantly improves the performance of change detection with its superior multi-temporal feature learning capability. These methods utilize various architectures including convolutional neural networks (CNNs) [79]–[81], recurrent neural networks (RNNs) [82]–[86], auto-encoders [87]–[89], and Transformers [90]–[94].

To improve change detection accuracy, many recent works have been developed for improving the feature representation ability and the change discrimination ability of the model, such as proposing different feature fusion strategies [95]–[97], designing different attention modules [98], [99], and introducing the self-attention mechanism [90], [98]. For instance, [99] proposed a pyramid feature-based attention-guided Siamese network (PGA-SiamNet), in which a co-attention module captures the correlation between bi-temporal images and a context fusion strategy fuses low-level and high-level features.

B. Vision-Language Understanding for Temporal Images

Recent advancements in natural language processing (NLP) and multimodal learning have catalyzed the development of Temporal Vision-Language Models (RS-TVLMs), which have become a central research focus for interpreting temporal remote sensing imagery. As illustrated in Fig. 2, the progression from binary change detection to semantic change detection and then to vision-language understanding represents a shift from traditional visual analysis to a more comprehensive multimodal semantic understanding that combines both visual and textual information.

RS-TVLMs, often built upon advanced architectures like Transformers [100], integrate computer vision with natural language processing. This integration allows models to better capture patterns across multiple time points, enabling a more comprehensive understanding of spatiotemporal changes. Specifically, RS-TVLMs can perform a variety of tasks, such as change captioning [22], [23]and change visual question answering [54], [101], change grounding [55], and text-to-change retrieval [102].

RS-TVLMs have greatly enriched the tools available for temporal image interpretation and have significantly expanded the potential applications of remote sensing data, enabling more accurate and comprehensive interpretations [103]. For example, in disaster monitoring, RS-TVLMs can analyze satellite images of regions affected by events like earthquakes or floods and generate comprehensive textual reports. These reports not only describe the changes but also identify the affected regions and assess the severity of the damage, providing valuable insights for emergency response teams. As RS-TVLMs continue to evolve, they are expected to drive more automated and context-aware interpretive capabilities, ultimately revolutionizing remote sensing temporal image analysis.

This paper aims to provide a comprehensive review of the progress in RS-TVLMs. We will explore the evolution of these models, highlight key advancements in each area, discuss the available datasets, and identify potential directions for future research.

III. Preliminary Knowl

A. Basic Language Models

*1) **LSTM**:* The Long Short-Term M work [82] is a specialized type of Recu (RNN) designed to address the challen term dependencies in sequential data. Tr suffer from issues like vanishing and during the training process, particularly sequences. LSTM overcomes these limi a memory cell and a set of gating mec the flow of information, thus enabling long-range dependencies effectively. He components of the LSTM architecture.

The forget gate determines which i previous time step should be discarded f It computes the forget gate output f_t us input x_t and the previous hidden state activation function:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] +$$

where W_f is the weight matrix, b_f is th the sigmoid activation function.

The input gate controls the flow of the memory cell. It computes the input candidate memory cell state $\tilde{c}_t$:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

where W_i and W_c are the weight matrices, and b_i and b_c are the respective bias terms. The tanh function ensures that the candidate memory cell state is bounded between -1 and 1.

The memory cell state c_t is updated by combining the previous memory state c_{t-1} and the new candidate memory state $\tilde{c}_t$, weighted by the forget and input gates:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

where $\odot$ denotes element-wise multiplication. This step allows the network to retain relevant information over long sequences while discarding less useful data.

The output gate determines the current hidden state h_t, which serves as the output of the LSTM for the current time step. It computes the output gate value o_t and generates the hidden state by applying the output gate to the memory cell state:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \odot \tanh(c_t)$$

Through these mechanisms, LSTM effectively manages the storage and flow of information, enabling it to capture long-term dependencies in sequential data. This capability has made LSTM widely applicable across various fields such as natural language processing (NLP), speech recognition, and time series forecasting. In recent years, extensions of the standard LSTM model, such as the xLSTM [104], have been proposed to address specific limitations and further enhance the model's performance in more complex tasks.

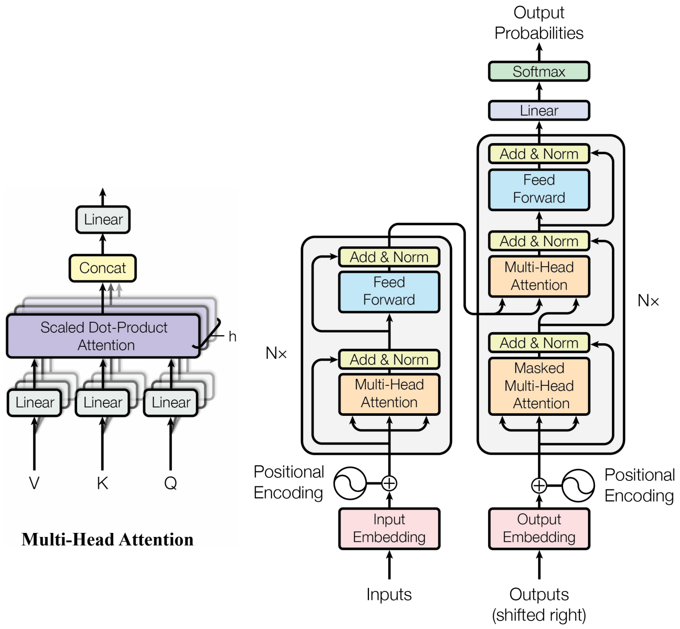


Fig. 3. The Transformer architecture [100].

*2) **Transformer**:* The Transformer model, introduced by Vaswani *et al.* [100], represents a fully attentional sequence model. Unlike traditional recurrent models such as LSTMs, which rely on sequential processing, the Transformer leverages a self-attention mechanism to process the entire input sequence in parallel, providing it with a global receptive field. This parallelization leads to improved efficiency and scalability. The Transformer has proven to be highly effective in training on vast amounts of data, making it the backbone of many state-of-the-art LLMs, such as GPT [105], T5 [106], and LLaMA [107].

The self-attention mechanism is the heart of the Transformer model. It allows each token in the input sequence to attend to every other token, assigning different attention scores based on their relevance. Formally, given a sequence of input embeddings $\mathbf{X_l} \in \mathbb{R}^{N \times C}$, the attention score is computed as follows:

$$Attention(Q, K, V) = Softmax(\frac{\mathbf{QK}^T}{\sqrt{d_m}})\mathbf{V}$$

$$\mathbf{Q} = \mathbf{X_l W^Q}$$

$$\mathbf{K} = \mathbf{X_l W^K}$$

$$\mathbf{V} = \mathbf{X_l W^V}$$

where $\mathbf{W^Q} \in \mathbb{R}^{C \times d_m}, \mathbf{W^K} \in \mathbb{R}^{C \times d_m}, \mathbf{W^V} \in \mathbb{R}^{C \times d_m}$, and $\mathbf{W^O} \in \mathbb{R}^{(h \times d_m) \times C}$ are learned projection matrices. The $Softmax$ function ensures that the attention scores are normalized, allowing the model to focus on the most relevant tokens. The result is a weighted sum of the values V based on the attention scores.

To capture a richer representation of the relationships between tokens, the Transformer utilizes multi-head attention, which performs several attention operations in parallel. Each attention head processes the sequence independently. Their outputs are concatenated and linearly transformed to form the final output. This allows the model to capture multiple types of relationships between tokens at different levels of abstraction.

As shown in Fig. 3, the Transformer model consists of two main components: the encoder and the decoder, each of which is composed of multiple stacked sublayers. Each sub-layer of the encoder contains a multi-head self-attention mechanism followed by a feed-forward network. For the decoder, each sub-layer contains an additional cross-attention layer that attends to the encoder's output. Another key feature of the decoder is the masked self-attention mechanism, which ensures that each token can only attend to previous tokens in the sequence (and not future ones). This masking is essential for autoregressive generation tasks, such as text generation or machine translation, where the model must predict the next token based on the preceding ones.

Since its introduction, the Transformer has been adapted and extended in numerous ways to tackle a wide range of tasks beyond traditional NLP, such as computer vision [108], [109], multimodality [27], [110], medicine [111], [112], and remote sensing [113], [114].

*3) **Mamba**:* Recently, State Space Models (SSMs) [115], have emerged as promising sequence models due to their global receptive field and linear computational complexity. Mamba [116], in particular, has demonstrated impressive performance across various language and vision tasks [117]–[122].

SSMs are based on the idea of continuous-time dynamic systems, which map a sequence $x(t) \in \mathbb{R}$ to $y(t) \in \mathbb{R}$ through an implicit hidden state $h(t) \in \mathbb{R}^N$. These models are inspired by linear time-invariant (LTI) systems, and their behavior is governed by the following set of equations:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t)$$

$$y(t) = \mathbf{C}h(t)$$

where $h'(t)$ represents the derivative of the hidden state $h(t)$, $A \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$ and $\mathbf{C} \in \mathbb{R}^{1\times N}$ define the system's evolution and output projections.

To adapt the continuous system for discrete sequence processing, SSMs employ a discretization technique known as zero-order hold (ZOH), which converts the continuous-time parameters, $\mathbf{A}$ and $\mathbf{B}$ into discrete-time equivalents, $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$. This transformation relies on a timescale parameter $\mathbf{\Delta} \in \mathbb{R}^D$, and is expressed as follows:

$$\bar{\mathbf{A}} = \mathbf{exp}(\mathbf{\Delta A})$$

$$\bar{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\mathbf{exp}(\mathbf{\Delta A}) - \mathbf{I}) \cdot \mathbf{\Delta B} \approx \mathbf{\Delta} B$$

The SSM model can then be represented as:

$$h_k = \bar{\mathbf{A}}h_{k-1} + \bar{\mathbf{B}}x_k$$

$$y_k = \mathbf{C}h_k$$

where $\{x_1, x_2, ..., x_K\}$ is the input sequence and $\{y_1, y_2, ..., y_K\}$ is the output sequence.

Different from traditional SSMs, Mamba [116] breaks the LTI limitation by introducing a selective state space mechanism. Specially, it allows the parameters $\bar{\mathbf{B}}$, $\mathbf{C}$, and $\mathbf{\Delta}$ dynamically vary based on the input sequence. This adaptability enables the model to selectively attend to different parts of the input, making it contextually aware and capable of processing sequences in a more flexible manner.

B. Large Language Models

LLMs have recently attracted significant attention due to their impressive performance across a wide range of natural language processing tasks [123]–[125]. Built on the advanced Transformer architecture with scaling law [126], LLMs leverage vast amounts of data and computational resources during pre-training to develop robust language understanding and generation capabilities. These models are capable of excelling in downstream tasks through fine-tuning or even in zero-shot or few-shot learning [127]–[129]. LLMs can be typically classified into three main categories based on their architectural design: Encoder-only models (e.g., BERT [130]), Encoder-Decoder models (e.g., T5 [106]), and Decoder-only models (e.g., GPT series, LLaMA [131], Gemini [132]).

The training process of an LLM typically consists of several key stages. The first stage is pre-training, during which the model learns a generalized language representation from large-scale, unlabeled text data through self-supervised learning. Common pre-training tasks include Masked Language Modeling (MLM) and Autoregressive Language Modeling (ALM) [133], both of which help the model learn contextual relationships within text. Following pre-training, the model undergoes a supervised Instruction fine-tuning phase, where it is further trained on labeled data specific to improve its performance on the specific task [134]. In some cases, an additional Reward Modeling is introduced, where external feedback signals—such as human ratings or user behavior data—are used to evaluate the quality of the model's outputs. Reinforcement Learning approaches leverage the reward signals to further optimize the model, ensuring that the outputs align more closely with human expectations [135].

This multi-phase training approach equips LLMs with both strong language comprehension and the ability to generate high-quality, diverse textual content. Recently, the success of models like ChatGPT [136] and GPT-4 [26] has firmly established the autoregressive, Decoder-only architecture as the dominant paradigm in current LLM research. The latest developments, such as GPT-o1, LLaMA-3 [137], and Qwen-2.5 [138], [139], demonstrate a growing trend toward multimodal capabilities, and the integration of AI agent functionalities.

IV. Remote Sensing Temporal Vision-Language Models

Current research on vision-language understanding in remote sensing temporal images focuses primarily on several key areas: change captioning, change visual question answering, change retrieval, and change grounding. These tasks aim to enhance the interpretation of remote sensing temporal images by leveraging multimodal modeling and language understanding. Additionally, with the development of LLMs, some recent research explores integrating LLMs to further improve vision-language understanding for remote sensing temporal images.

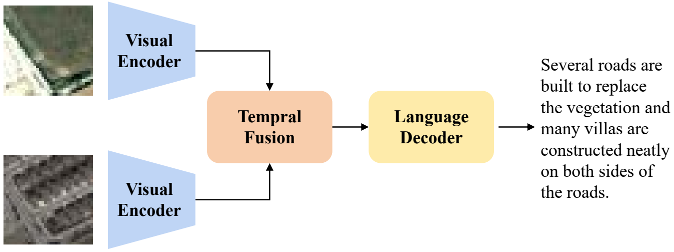
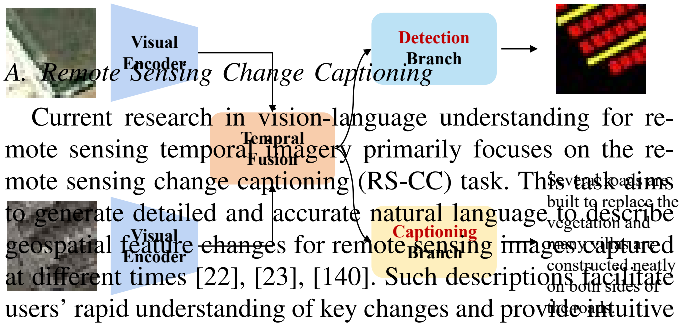

Visual Encoder
Visual Encoder
Tempral Fusion
Language Decoder
Several roads are built to replace the vegetation and many villas are constructed neatly on both sides of the roads.

Fig. 4. The general framework for remote sensing change captioning.


Visual Encoder
Visual Encoder
Temporal Fusion
Detection Branch
Captioning Branch
Several roads are built to replace the vegetation and many villas are constructed neatly on both sides of the roads.
Text

A. Remote Sensing Change Captioning

Current research in vision-language understanding for remote sensing temporal imagery primarily focuses on the remote sensing change captioning (RS-CC) task. This task aims to generate detailed and accurate natural language to describe geospatial feature changes for remote sensing images captured at different times [22], [23], [140]. Such descriptions facilitate users' rapid understanding of key changes and provide intuitive semantic support for decision-making and analysis of temporal remote sensing data. Change captioning requires models to accurately identify significant changes and translate them into natural, coherent language. This transformation process not only depends on precise visual change recognition but also demands high language generation capability to ensure both accuracy and fluency in the language.

Previous change captioning approaches are typically based on deep learning and follow a three-stage architecture as illustrated in Fig. 4: visual encoding, bi-temporal fusion, and language decoding. Each stage significantly impacts the overall model performance, so recent studies have concentrated on improving these three areas. Some representative methods are summarized in Table I.

Visual encoding aims to extract rich semantic features from bi-temporal images, commonly utilizing Siamese encoders that facilitate comparison between bi-temporal images. Encoders are typically based on Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs). CNNs excel at capturing spatial details. ViTs are capable of extracting broad geospatial information through global attention mechanisms. Many approaches leverage pre-trained image encoders, such as ResNet [141] or ViT [142]. For example, Chang *et al.* [22] use ResNet-101 as the encoder. Liu *et al.* [143] used ViT and compared the performance differences of ViT trained on ImageNet [144] and CLIP [145]. Additionally, some studies have explored training encoders specifically adapted for change extraction through self-supervised learning. For instance, Zhou *et al.* [24] propose a single-stream extractor pre-trained on a large-scale bi-temporal remote sensing image dataset, significantly enhancing the robustness of change feature extraction.

Most research focuses on improving the model's performance at the bi-temporal fusion stage, which is the core of change captioning. This stage aims to integrate bi-temporal features to capture latent temporal change patterns. During bi-temporal fusion, models should accurately identify significant differences between the two images while suppressing irrelevant pseudo-changes (e.g., variations due to lighting or weather). Previous research typically employs CNN or Transformer as basic modules and proposes some attention mechanisms to enhance the model's change perception ability. For example, Liu *et al.* [23] proposed a Transformer-based method named RSICCformer, which consists of multiple cross-encoding modules to use differential features, allowing attention to the changing areas of each image. Besides, they benchmarked some methods from the computer vision field [146], [147]. Chang *et al.* [22] proposed a hierarchical self-attention network consisting of multiple Transformer layers to dynamically focus on temporal change regions. Additionally, remote sensing images contain a variety of objects of different sizes. Some methods incorporate multi-scale strategies to further enhance the model's capability to identify diverse changes [148], [149].

The language decoder translates fused visual features into natural language descriptions. Early methods utilized Support Vector Machines (SVMs) or Long Short-Term Memory (LSTM) networks for language generation. Chouaf and Hoxha *et al.* [150], [151] compared the performance of RNN and support vector machine (SVM) as language decoders. Given the powerful generation ability of the Transformer decoder, RSICCformer [23] first introduced Transformers to the remote sensing change captioning task, employing cross-attention mechanisms to allow the model to focus on specific image regions during word generation. Although Transformers perform well, the computational complexity of the model grows quadratically as the sequence length increases. To address this challenge, recent research has introduced the Mamba model [116], which operates with linear complexity. Liu *et al.* [122] proposed Spatial Difference-aware SSM (SD-SSM) and Temporal-Traversing SSM (TT-SSM) to improve the ability of spatiotemporal joint modeling. Besides, they compare three different language decoders, including Mamba, generative pre-trained Transformer (GPT) style decoder, and Transformer decoder.

The above encoder-fusion-decoder framework process changed and unchanged image pairs in a coupled way. Unlike that, Liu *et al.* [143] proposed a decoupling paradigm that decouples change captioning into two questions: "whether a change has occurred" and "what change has occurred". They input the decoupling results into a pre-trained LLM for language generation through a multi-prompt learning strategy. The decoupling paradigm allows researchers to independently concentrate on improving the captioning of changed image pairs and that of unchanged image pairs.

B. Multi-task learning of Change Detection and Captioning

In remote sensing temporal image analysis, change detection and change captioning tasks focus on different levels of change information extraction [103], [166]. Change detection is primarily concerned with generating pixel-level change masks between bi-temporal images to identify and highlight changed areas. In contrast, change captioning aims to achieve a semantic-level understanding of these changes, including object attributes and contextual relationships. Given the intrinsic connection between these two tasks, recent research

TABLE I
SOME REPRESENTATIVE METHODS FOR REMOTE SENSING CHANGE CAPTIONING.

Model Name	Visual Encoder	Language Decoder	Contribution
CNN-RNN [150]	VGG-16	RNN	First exploration of the task and experiments on a private dataset.
CC-RNN/SVM [151]	VGG-16	RNN, SVM	Published two small datasets, explored two dual-temporal feature fusion strategies.
RSICCformer [23]	ResNet-101	Transformer Decoder	Proposed a large dataset LEVIR-CC and a dual-branch Transformer-based algorithm for change captioning.
PSNet [148]	ViT-B/32	Transformer Decoder	Proposed multiple progressive difference-aware layers for multi-scale feature extraction.
PromptCC [143]	ViT-B/32	GPT-2	Decoupled change caption into binary image change classification and fine-grained change perception. Proposed multiprompt learning strategy which can prompt the LLM to generate plausible captions.
Chg2Cap [22]	ResNet-101	Transformer Decoder	Proposed attentional encoder to accurately localize the variations between diachronic image features.
ICT-Net [149]	ResNet-101	Transformer Decoder	Using ResNet for multi-scale feature extraction, Interactive Change-Aware Encoder for change representation.
SITS-CC [152]	ResNet-101	Transformer Decoder	Explored a new task of image time series change caption.
RSCaMa [122]	ViT-B/32	Mamba, Transformer Decoder, GPT-2	Proposed a temporal-spatial model for spatial and temporal SSM to improve temporal modeling. It introduced Mamba to this task.
SparseFocus [153]	ResNet-101	Transformer Decoder	Proposed a Sparse Focus Transformer (SFT) to focus changing regions and reduce computational complexity.
SEN [154]	ResNet with 6-channel	Transformer Decoder	Proposed a single-stream extractor network (SEN) with contrastive pre-training to improve vision encoder. It has a lower computational cost than the dual-stream extractor.
Diffusion-RSCC [155]	ResNet-101	Diffusion	Introduced Diffusion model into this task that can learn the cross-modal data distribution between the image pairs and change captions.
CARD [156]	ResNet-101	Transformer Decoder	Decoupled common features and difference features to promote unified representation in multi-change scenarios.
ChangeRetCap [102]	ResNet-101	Transformer Decoder	Proposed a foundation model considering both bi-temporal captioning and bi-temporal text-image retrieval tasks.
Intelli-Change [157]	ResNet-101	Transformer Decoder	Created a model called IntelliChange-RSCC for change captioning where it primarily consists of two important components: a cross-attention based encoder and a spatial attention based conceptual tokenizer module.
ChangeExp [158]	LLaVA-1.5	LLaVA-1.5	Studied the ability of large vision-language models (LVLM) to explain temporal changes in satellite images; proposed three prompting methods
MAF-Net [159]	ResNet-101	Transformer Decoder	Resnet extracts multi-scale dual-temporal features, and each scale's dual-temporal features are processed and fused by Transformer.
SFEN [160]	WideResNet	Transformer Decoder	Proposed a scale feature enhancement network.
MfrNet [161]	ResNet-18	Transformer Decoder	Introduced Joint Attention and Dense Feature Fusion Module (JADF) for refining features and reducing noise
SEIFNet [162]	ResNet-101	Transformer Decoder	Proposed a network with cross-time interaction and symmetric difference learning to model differences from coarse to fine representations.
MV-CC [163]	InternVideo2	Transformer Decoder	Proposed a mask-enhanced video model for change captioning to reduce the complex manual design of features extraction and fusion.
Chareption [164]	CLIP ViT-L/14	LLaMA-7B	Adjusted VIT and LLM through adapters to generate accurate and contextually relevant descriptions of changes.
MADiffCC [165]	Diffusion	Transformer Decoder	Use a diffusion feature extractor to capture the multi-level and bi-temporal feature. Proposed a gated cross-attention guided captioning decoder
CCExpert [140]	ViT	Qwen-2	Utilized the MLLM and design a Difference-aware Integration Module to capture fine-grained difference and constructed a CC-Foundation Dataset, comprising 200,000 pairs of images and 1.2 million captions.

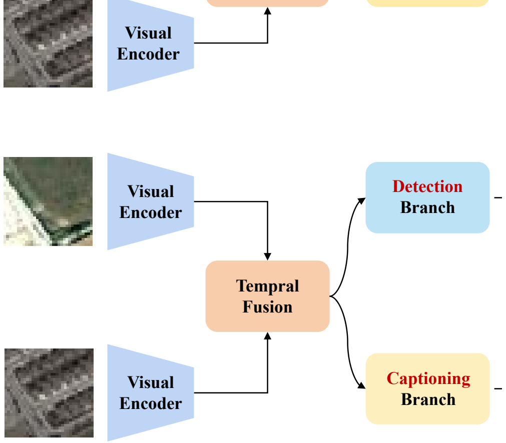


Fig. 5. The general framework for multi-task learning fr
detection and change captioning.

has integrated change detection and change c
unified multi-task learning framework to imp
efficiency and accuracy of change interpretati
sentative methods are summarized in Table II

Change-Agent [103] is one of the represer
this field, establishing a multi-task learning
has laid the foundation for subsequent studi
As shown in Fig. 5, this framework builds on
encoder and employs two task-specific branc
detection and change captioning, respectively.
sual encoding stage, the model extracts tempor
bi-temporal images, and these fused features s
sequent branches for each task. Notably, simil
change detection models, the change detection
utilizes multi-scale bi-temporal features extract
encoder to ensure precision and detail in th
Meanwhile, the change captioning branch us
only the deepest-level visual features to focus c
of the changes, with a design that closely resen
change captioning models.

Balancing the training of both tasks within the multi-task framework is a critical challenge. Current studies commonly apply weighted loss, combining the losses of change detection and change captioning with different weights. For instance, [166] adopt a metabalance strategy via adapting gradient magnitudes of auxiliary tasks proposed in [171], while [169] and [170] adopt a dynamic weight averaging strategy in [172].

Additionally, some recent studies have explored how change detection can specifically assist change captioning to improve descriptive accuracy [163], [173]. The core idea is that pixel-level change detection enhances the change captioning model's ability to recognize changes, particularly for small structures and under low-light conditions. For example, MV-CC [163] uses low-resolution change detection masks as explicit guidance to focus on the change region accurately.

C. Remote Sensing Change Visual Question Answering

The Remote Sensing Change Visual Question Answering (RS-CVQA) task aims to generate natural language responses based on temporal remote sensing images and user-specific questions. Unlike change detection and captioning tasks, RS-CVQA emphasizes interactive language engagement between users and temporal images, providing a more flexible and efficient means of acquiring information on changes within images. Fig. 6 illustrates a typical RS-CVQA model framework, which includes the following key stages: visual encoding, question encoding, multimodal interaction, and answer generation. Some representative methods are summarized in Table III.

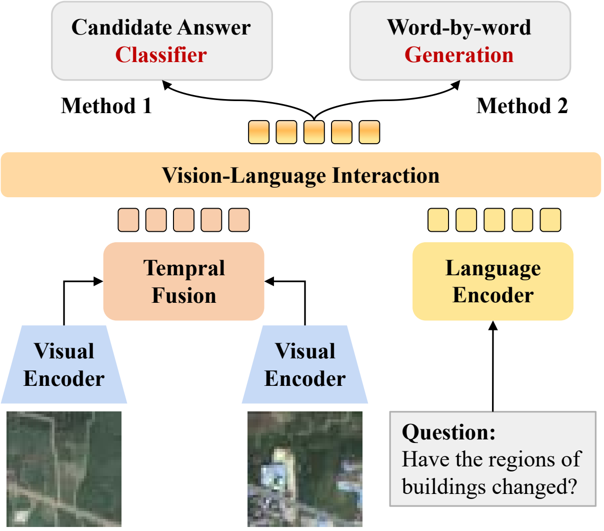


Fig. 6. The general framework for remote sensing change visual question answering.

In the visual encoding stage, the model generally employs a twin encoder to separately extract features from the bi-temporal remote sensing images, fusing these temporal features to capture change-related information within the images. In the question encoding stage, pre-trained language models such as BERT [130] or GPT [174] are commonly used to transform the user's complex question into a semantic embedding suitable for model understanding. In the multimodal interaction stage, attention mechanisms (e.g., self-attention and cross-attention) are widely applied to align and integrate visual changes and language features, allowing the model to focus on key change regions in the image based on semantic cues from the question. This multimodal interaction enhances the model's understanding of image changes and ensures that generated answers remain closely related to the visual content. Finally, the answer generation stage converts the fused multimodal features into a natural language response. Based on the answer generation method, RS-CVQA approaches are broadly divided into two categories: candidate-based RS-CVQA and word-by-word generative RS-CVQA.

In candidate-based RS-CVQA, the answer generation module is designed as a multi-class classifier, selecting the optimal answer from a predefined set of candidate answers. Yuan *et al.* [175] first introduced this task, categorizing answers into several fixed classes and allowing the classifier to select an answer directly from them. This approach is computationally efficient and stable, making it suitable for tasks where the target is well-defined, and change types are fixed. However, due to its reliance on a limited predefined answer pool, this method is less flexible and may not be suitable for addressing

TABLE II
Some representative methods for multi-task learning of change detection and change captioning.

Model Name	Visual Encoder	Language Decoder	Contribution
Pix4Cap [167]	ViT-B/32	Transformer Decoder	Utilized a BIT model pre-trained on the LEVIR-CD dataset to establish CD pseudo-labels for the LEVIR-CC dataset, introducing a change detection branch in the change caption network to learn pseudo-labels.
Change-Agent [103]	ViT-B/32	Transformer Decoder; also includes a change detection branch	Built a change agent comprising an LLM and a multi-level change interpretation model. Proposed a dual-branch multi-task model and LEVIR-MCI dataset containing masks and descriptions.
Semantic-CC [168]	SAM	Vicuna	Semantic-CC leveraged the latent knowledge of the SAM foundational model to alleviate the high generalization algorithm's dependence on extensive annotations and generated more comprehensive and accurate change descriptions through pixel-level semantic guidance from change detection (CD).
DetACC [173]	ResNet-101	Transformer Decoder	Enhanced its change description capabilities by utilizing explicit visual change information.
KCFI [169]	ViT	Qwen	Proposes a novel multimodal framework for remote sensing change captioning, guided by key change features and instruction tuning (KCFI). Extracted key change features and leveraged LLM to improve the accuracy of the description. Used dynamical weights to optimize the losses for CC and CD.
ChangeMinds [166]	Swin Transformer	Transformer Decoder and CD decoder	Proposed ChangeLSTM module that utilizes the recent XlSTM to process features from both directions to obtain a universal change-aware representation. Proposed the Multi-task Predictor that uses the unified change decoder to fuse change-aware representations and employs multi-

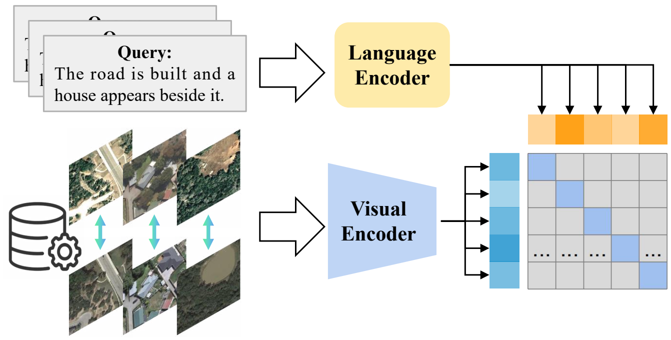
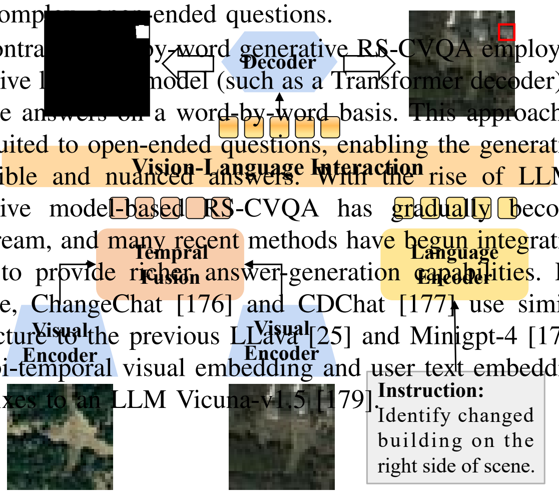
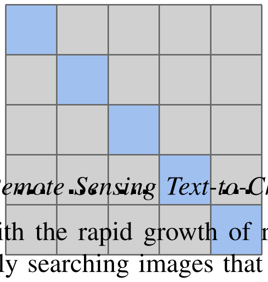


Fig. 7. The general framework for remote sensing text-to-change retrieval.

more complex open-ended questions.

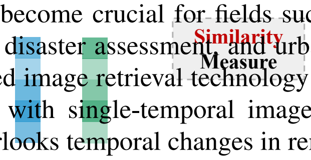

Decoder
Vision-Language Interaction
Tempral Fusion
Language Encoder
Visual Encoder
Visual Encoder
Instruction:
Identify changed building on the right side of scene.

In contrast, word-by-word generative RS-CVQA employs a generative language model (such as a Transformer decoder) to generate answers on a word-by-word basis. This approach is more suited to open-ended questions, enabling the generation of flexible and nuanced answers. With the rise of LLMs, generative model-based RS-CVQA has gradually become mainstream, and many recent methods have begun integrating LLMs to provide richer answer-generation capabilities. For example, ChangeChat [176] and CDChat [177] use similar architecture to the previous LLaVA [25] and Minigpt-4 [178], using bi-temporal visual embedding and user text embedding as prefixes to an LLM Vicuna-v1.5 [179].



D. Remote Sensing Text-to-Change Retrieval

With the rapid growth of remote sensing image data, efficiently searching images that meet specific user requirements has become crucial for fields such as environmental monitoring, disaster assessment, and urban planning. Traditional text-based image retrieval technology matches user-provided query text with single-temporal images. However, this technology overlooks temporal changes in remote-sensing images, making it difficult to meet users' retrieval needs for dynamic scenes.

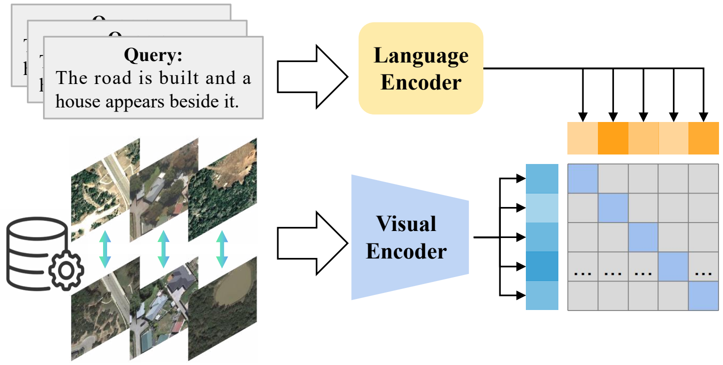


Remote Sensing Text-to-Change (RSI-TCR) has emerged to address this limitation. Its core objective is efficiently retrieving bi-temporal image pairs that meet user-input queries describing image changes. RSI-TCR significantly reduces the manual effort required for filtering large datasets, enhancing the usability of massive remote sensing data. This technology has shown great value in real-world scenarios. For instance, in disaster management, RSI-TCR can rapidly locate temporal images of affected areas based on query text (e.g., "flood inundation"), providing essential data for post-disaster emergency response.

Compared to traditional text-based image retrieval tasks, which involve binary matching between "text" and "image", RSI-TCR is more complex due to its tri-modal matching requirements—"pre-event image", "post-event image", and "text". This complexity requires models to process the complex relationship between spatiotemporal changes and textual information within a multi-modal semantic space.

TABLE III
Some representative methods for remote sensing change visual question answering.

Model Name	Visual Encoder	Language Decoder	Contribution
change-aware VQA [175]	CNN		
CDVQA-Net [54]	CNN		
ChangeChat [176]	CLIP-ViT		
CDChat [177]	CLIP ViT-L/14		
TEOChat [101]	CLIP ViT-L/14		
GeoLLaVA [180]	Video encoder		
VisTA [55]	Shared CLIP image Encoder		



Ferrod *et al.* [102] first studied the RSI-
LEVIR-CC dataset [23] and proposed a f
task, as illustrated in Fig. 7. In their approa
the Chg2Cap model [22] is used to extract
embedding from bi-temporal images. The u
text is encoded into a textual embedding us
decoder. They then align image change en
query text embedding through a contras
function, specifically InfoNCE [181].

One of the core challenges in RSI-TCR i
negatives. Specifically, an image pair labe
sample in a training batch may actually be
that matches the query text, which can disru
This problem is common in many tasks th
learning, and there are solutions to solve t
[183]. To address this problem, Ferrod *et a*
two common strategies to improve retrieval
plex change scenarios: 1) False Negative E
Potential false negatives are excluded from
to prevent interference. 2) False Negative Attraction: Potential false negatives are re-labeled as positive samples to better align with the true relationships in the data.

E. Remote Sensing Change Grounding

Remote Sensing Change Grounding (RS-CG) aims to identify and localize change regions referred to by the user-provided query text within bi-temporal remote sensing images. By incorporating natural language as a query modality, RS-CG significantly enhances user interaction flexibility compared to traditional change detection, which is limited to fixed category outputs.

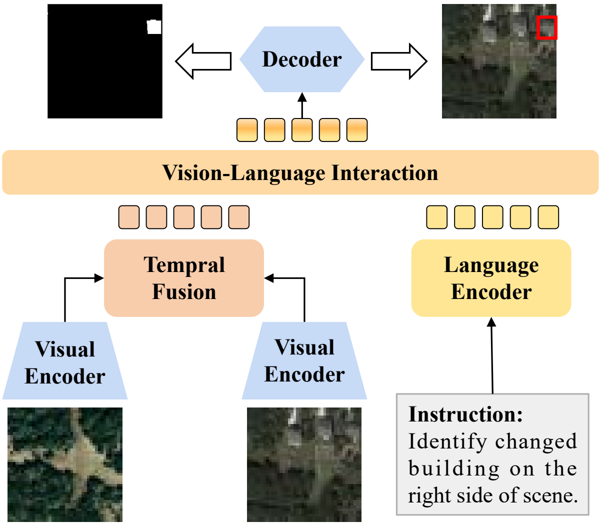

Decoder
Vision-Language Interaction
Tempral Fusion
Language Encoder
Visual Encoder
Visual Encoder
Instruction:
Identify changed building on the right side of scene.

Fig. 8. The general framework for remote sensing change grounding.

The outputs of RS-CG are typically presented in two forms: bounding boxes and pixel-level masks, as illustrated in Fig 8. Bounding boxes annotate change regions with rectangular outlines, providing an intuitive spatial location of target changes. Pixel-level masks, on the other hand, offer precise delineations of the shapes and boundaries of change regions, making them ideal for fine-grained analysis.

Irvin *et al.* [101] adopted a model architecture inspired by LLaVA-1.5 [25]. They used a temporally shared ViT-L/14 to encode the temporal images, with the embeddings mapped through an MLP before being fed into LLaMA-2 [107]. The

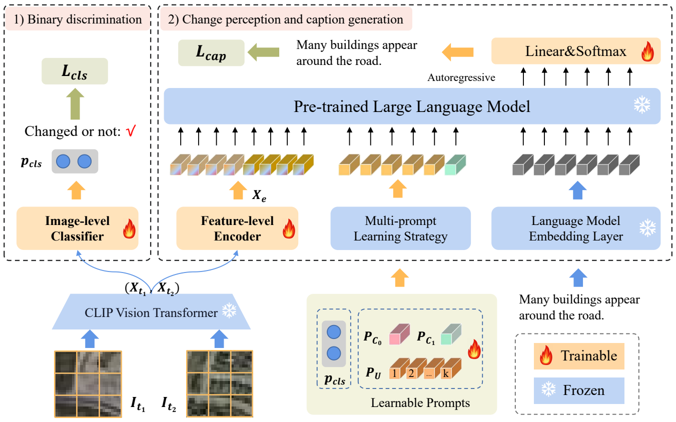


Fig. 9. Overview of a remote sensing change captioning method proposed in [143], using Large Language Models.

LLM outputs the coordinates of the bounding boxes in a textual format, effectively grounding the detected changes in the input query. Li *et al.* [55] proposed a novel multi-task model named VisTA, designed for Change Detection Question Answering and Grounding. VisTA is capable of answering user questions while simultaneously generating pixel-level change masks associated with the textual answers. The textual answers are generated through a two-layer MLP, while the mask decoder is composed of two attention blocks. This dual-output approach allows VisTA to provide both semantic and visual explanations, making it a versatile solution for RS-CG tasks.

V. Large Language Models Meets Temporal Images

With their remarkable natural language generation and understanding capabilities, LLMs have demonstrated significant value in multimodal tasks [27], [110]. In remote sensing temporal image understanding, LLMs offer novel perspectives by integrating visual features with language, enabling advanced analysis and reasoning for spatiotemporal semantics. This section discusses the recent advancements of LLMs in temporal image understanding, focusing on their applications in change captioning, change visual question answering, and intelligent agent construction.

A. LLM-based Change Captioning

Change Captioning, a key task in temporal image understanding, aims to translate changes between images into semantic natural language descriptions. PromptCC [143] is an early pioneering work introducing LLMs into this task. As shown in Fig. 9, PromptCC employs a shared visual encoder to extract features from bi-temporal images. A feature-level encoder fuses these features to incorporate rich change semantics, which are then passed to GPT-2 [174] as prefix tokens. GPT-2 translates these visual tokens into accurate language descriptions. To maximize the potential of LLMs in this task, PromptCC introduces a classifier-based multi-prompt learning strategy, significantly improving language generation accuracy without fine-tuning the GPT-2 model. This work established a solid foundation for leveraging LLMs in the change captioning task.

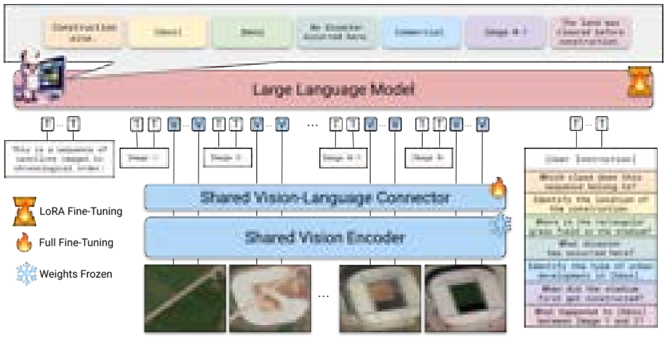


Fig. 10. Overview of a remote sensing temporal image understanding method proposed in [101], using Large Language Models.

Subsequent studies have adopted similar frameworks to enhance LLM performance in change captioning. Most methods focus on extracting discriminative change semantics for the LLM decoder. For instance, KCFI [169] introduces a Key Change Perception module to filter irrelevant regions, highlighting salient change features. Chareption [164] uses a cosine similarity-based change-aware module for acquiring representative change features. To effectively utilize LLMs' knowledge, Semantic-CC [168], combining SAM model [184] and Vicuna model [179], employs Low-Rank Adaptation (LoRA) [185] to reduce training costs. Chareption [164] designed a change adapter module into the LLM's attention layers, improving the adaptability of the LLM to the changing captioning. Some representative studies are shown in Table IV.

B. LLM-based Change Visual Question Answering

Researchers have recently begun exploring LLM applications in change-related visual question answering for temporal images. ChangeChat [176], an early research in this area, adopts an LLava-like architecture [25], bridging bi-temporal image features and LLMs via a simple MLP to enable change-related multiturn dialogue. Follow-up works such as CD-Chat [177] and GeoLLava [180] employ similar architecture.

More recently, TeoChat [101] expanded LLMs to address a broader range of temporal image understanding tasks. As shown in Fig. 10, TeoChat supports analyzing multi-temporal images of arbitrary temporal lengths and employs joint training mechanisms to simultaneously optimize multiple tasks, including Temporal Scene Classification, Change Detection, Change Referring Expressions, and Change Question Answering. This design enhances the model's flexibility and strengthens its capacity for comprehensive spatiotemporal reasoning. TeoChat encodes all task responses in natural language formats recognizable by LLMs, enabling unified task execution based on user instructions. Compared to earlier methods, TeoChat demonstrates the immense potential for temporal image change understanding and represents a shift from single-task solutions to a multi-task framework.

C. LLM-based Agents

LLMs, trained on large-scale corpora, possess extensive knowledge and powerful capabilities in instruction comprehension, planning, and reasoning. These strengths have spurred researchers' interest in LLM-based agents [186]–[189]. LLM-based agents typically use LLMs as core controllers, adopt modular designs, and integrate various tools and models to dynamically adapt to users' requirements, offering high flexibility and autonomy. For example, agents like PaLM-E [190] and EmbodiedGPT [191] seamlessly combine visual perception, language generation, and user interaction to deliver end-to-end intelligent services.

In the realm of temporal image understanding, LLM-based agents show emerging potential. Liu *et al.* [103] proposed a change interpretation agent named Change-Agent, employing LLMs as a central brain and integrating change interpretation models (e.g., change detection and captioning modules) as visual processing units, supplemented by external tools for extended functionalities. Change-Agent can understand complex user instructions and implement step-by-step execution through task planning. For instance, when tasked with counting changed objects, Change-Agent invokes a change detection model to generate change masks and autonomously writes and executes Python scripts to count changed objects, thereby avoiding the "hallucination" issues prevalent in traditional VQA models. Additionally, Change-Agent supports highly customized outputs, such as detecting specific types of changes (e.g., road changes), offering flexible solutions for temporal image analysis.

In summary, LLM-based agents exhibit significant advantages, overcoming the limitations of traditional models constrained to single tasks. However, research in this area remains nascent. Future advancements may focus on optimizing agent scheduling mechanisms, incorporating remote sensing domain-specific knowledge, and broadening the scope of interpretation tasks. These efforts will lay a robust foundation for deploying intelligent agents in practical temporal image understanding applications.

VI. Evaluation Metrics

This section outlines the evaluation metrics commonly employed to assess the above temporal vision-language tasks. Based on the differences in model outputs, we will introduce them into three groups, including language generation metrics, retrieval metrics, and localization metrics, each tailored to evaluate specific aspects of the tasks.

A. Language Generation Metrics

Language generation tasks, such as describing changes or answering questions based on remote sensing imagery, require metrics that evaluate the quality of the generated text. Commonly used metrics in previous studies are as follows:

1) ***BLEU:*** The BLEU (Bilingual Evaluation Understudy) [192] metric measures the overlap of n-grams (n=1, 2, ...) between generated text and ground truth. BLEU employs a brevity penalty to discourage overly short outputs. BLEU-4 is most commonly reported. Higher BLEU scores indicate better alignment.

TABLE IV
Some representative studies based on LLMs for remote sensing temporal image understanding. "CC" denotes "change captioning" and "CG" denotes "change grounding"

Method	Release Time	LLM	Fine-tuning	Task
PromptCC [143]	2023.06	GPT-2	Prompt Learning	CC
Change-Agent [168]	2024.07	Chatgpt	–	CC, CD
Semantic-CC [168]	2024.07	Vicuna	LoRA	CC
ChangeChat [176]	2024.09	Vicuna-v1.5	LoRA	CVQA, CG
KCFI [169]	2024.09	Qwen	Prompt	CC
CDChat [177]	2024.09	Vicuna-v1.5	LoRA	CVQA
TEOChat [101]	2024.10	LLaMA-2	LoRA	CVQA, CG
GeoLLaVA [180]	2024.10	LLaVA-NeXT	LoRA	CVQA
Chareption [164]	2024.10	LLaMA-7B	Adapter	CC
CCExpert [140]	2024.11	Qwen-2	LoRA	CC

2) ***ROUGE:*** Unlike BLEU, which focuses on precision, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [193] focuses on recall, measuring how much of the reference text is covered by the generated text. ROUGE$_N$ measures the overlap of n-grams between the generated and reference texts. The most commonly used is ROUGE$_L$, which measures the overlap of the longest common subsequence (LCS), taking into account structural similarity at the sentence level.

3) ***METEOR:*** Unlike BLEU, which primarily considers exact n-gram matches, METEOR (Metric for Evaluation of Translation with Explicit ORdering) [194] aligns words between the generated text and reference text using a more sophisticated approach. It incorporates exact word matches, stemming (matching words with the same root form, e.g., "run" and "running"), and synonyms (words with similar meanings), providing a more semantically aware evaluation. METEOR calculates precision and recall for the aligned words and combines them using a harmonic mean. Additionally, it includes a penalty term to account for word order mismatches.

4) ***CIDEr:*** CIDEr (Consensus-based Image Description Evaluation) [195] calculates the Term Frequency Inverse Document Frequency (TF-IDF) vectors of the n-gram of the generated text and the reference text, and then uses cosine similarity to measure semantic consistency. TF-IDF distinguishes the importance of different n-grams, with frequent phrases having lower weights and uncommon phrases having higher weights. CIDEr-D is a commonly used extension of the original CIDEr metric, incorporating additional processing to improve robustness to noisy or irrelevant variations in the generated text.

5) $\boldsymbol{S_m^*}$*:* The S_m^* metric provides a comprehensive approach to evaluating text generation by combining four widely used metrics, including BLEU, METEOR, ROUGE, and

CIDEr—each of which captures a distinct aspect of generation quality. By consolidating these metrics, S_m^* balances precision, recall, and semantic alignment, providing a holistic evaluation of text generation. The S_m^* metric is calculated as follows:

$$S_m^* = \frac{(BLEU_4 + METEOR + ROUGE_L + CIDEr)}{4}$$

*6) **BERTScore:*** BERTScore [196] is a modern evaluation metric for text generation tasks that leverages contextualized word embeddings from pre-trained language models, BERT [130]. Unlike traditional n-gram-based metrics (e.g., BLEU, METEOR), BERTScore focuses on semantic similarity, providing a more nuanced evaluation of generated text quality. It calculates precision and recall by measuring the cosine similarity between each token in the generated text and every token in the reference text. Finally, the F1 score is derived based on the calculated precision and recall.

B. Text-Image Retrieve Metrics

*1) **Recall@K:*** The Recall@K (R@K) [**?**] metric measures the coverage of all relevant items within the top K retrieved results, reflecting the comprehensiveness of the retrieval. K is typically set to 1, 5, or 10. Higher R@K values indicate that more relevant items are recalled within the top K results, particularly in scenarios with many potential matches. Additionally, the mean Recall (mR) is the average of R@1, R@5, and R@10, providing a comprehensive evaluation of the ability of the model to balance precision and recall across both narrow and broader retrieval contexts. The R@K metric is calculated as follows:

$$\mathrm{R@K} = \frac{\mathrm{TP@K}}{\mathrm{TP@K} + \mathrm{FN@K}}$$

where TP@k represents the number of correctly identified relevant items in the top-k results, and FN@k represents the number of relevant items that were not retrieved in the top-k results.

*2) **Precision@K:*** The Precision@K (Pr@K) metric focuses on evaluating the proportion of relevant items in the top K retrieved results, measuring the quality of the retrieval. A higher Pr@K value indicates that more relevant items are returned within the top K results, reflecting the ability of the model to capture highly relevant information accurately. When the ranking quality of retrieval results is crucial, the Mean Average Precision (MAP) is an effective metric. It combines precision and recall at different K values, assigning higher scores to better-ranked results, thereby providing a comprehensive evaluation of the overall performance of the image-text retrieval model. The Pr@K metric is calculated as follows:

$$\mathrm{Pr@K} = \frac{\mathrm{TP@K}}{\mathrm{TP@K} + \mathrm{FP@K}}$$

where FP@k represents the number of items incorrectly marked as relevant in the top-k results.

C. Localization Metrics

*1) **MIoU:*** MIoU (Mean Intersection over Union) metric [197] evaluates the overlap between the predicted region and the ground truth region, focusing not only on prediction accuracy but also on how well the model covers the target location, especially in complex backgrounds and occlusion scenarios. A high MIoU value indicates that the model can accurately identify and locate the object while minimizing incorrect predicted regions. The mathematical formula is as follows:

$$\mathrm{IoU}_i = \frac{|\,\mathrm{A}_i \cap \mathrm{B}_i\,|}{|\,\mathrm{A}_i \cup \mathrm{B}_i\,|}$$

$$\mathrm{MIoU} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \mathrm{IoU}_i$$

Where N is the number of samples, A_i and B_i are the predicted region and ground truth region of the i-th sample, respectively.

*2) **CIoU:*** Unlike MIoU, CIoU (Cumulative Intersection over Union) [197] metric evaluates the intersection over union ratio between the predicted regions and the ground truth regions across multiple IoU thresholds, and accumulates the overall performance of the model based on these thresholds. CIoU provides a more comprehensive performance evaluation, especially in complex scenarios with large prediction deviations or overlapping objects. The mathematical formula is as follows:

$$\mathrm{CIoU} = \frac{1}{\mathrm{T}} \sum_{i=1}^{\mathrm{T}} \mathrm{IoU}_i$$

Where T is the number of selected IoU thresholds, A_i and B_i represent the predicted region and the ground truth region at the i-th IoU threshold, respectively.

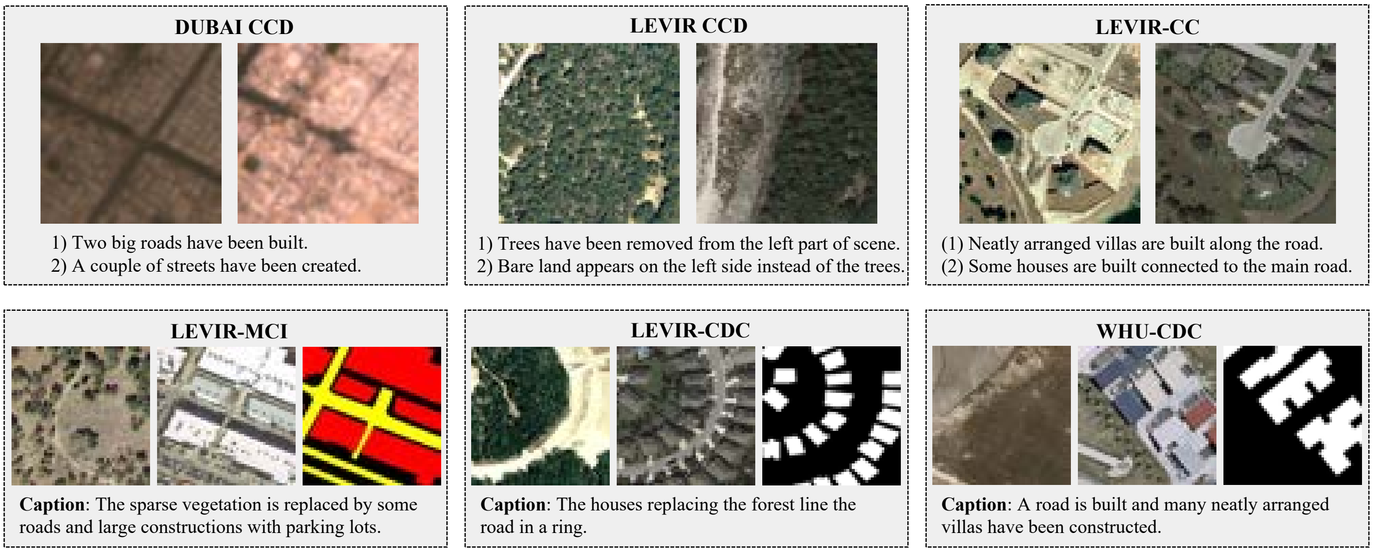
*3) **Precision@k:*** The Precision@k (Pr@k) metric also can be used to evaluate the model's capacity to accurately localize specific regions, particularly in identifying the correct region among multiple candidate predictions. It is computed by comparing the top-k predictions with the ground truth; if the correct region appears within the top-k results, it is considered correct. A higher Pr@k value signifies better performance in accurately locating the target region among the top-k predictions. However, Pr@k is typically evaluated alongside other performance metrics, such as IoU and MIoU, to offer a more comprehensive assessment of the model's overall performance.

VII. Temporal Vision-Language Dataset

Temporal vision-language datasets are pivotal for developing models that understand and integrate temporal changes with language. These datasets capture changes over time in visual data, paired with textual descriptions or queries, enabling various tasks like change captioning, multi-task vision-language modeling, and change visual question answering. In this section, we categorize and introduce these datasets based on their annotation forms.

A. Dataset Matching Temporal Images and Text

Such datasets match temporal images with text. The text describes the visual changes in the image. These datasets support tasks like Change Captioning and Retrieval. The current datasets are compared in Fig. 11 and Table V.



DUBAI CCD

1) Two big roads have been built.
2) A couple of streets have been created.

LEVIR CCD

1) Trees have been removed from the left part of scene.
2) Bare land appears on the left side instead of the trees.

LEVIR-CC

(1) Neatly arranged villas are built along the road.
(2) Some houses are built connected to the main road.

LEVIR-MCI

Caption: The sparse vegetation is replaced by some roads and large constructions with parking lots.

LEVIR-CDC

Caption: The houses replacing the forest line the road in a ring.

WHU-CDC

Caption: A road is built and many neatly arranged villas have been constructed.

Fig. 11. The example of the datasets. For DUBAI CCD, LEVIR CCD, and LEVIR-CC, they match temporal images and text. For LEVIR-MCI, LEVIR-CDC, and WHU-CDC, they match temporal images, text, and masks

*1) **DUBAI CCD**:* The Dubai CCD dataset [151] focuses on urbanization changes in Dubai between 2000 and 2010. It is based on multispectral images from the Landsat 7 ETM+ sensor, which provides imagery in several bands with varying spatial resolutions. The dataset includes bitemporal images acquired in 2000 and 2010, with a spatial resolution of 30 meters. A total of 500 image tiles, each with a size of 50x50 pixels, were extracted for annotation. The changes include urban development such as roads, residential areas, buildings, and green spaces. Annotators provided 2500 change descriptions that range from simple to more complex captions, with an average sentence length of 7.35 words.

*2) **LEVIR CCD**:* The image size in the LEVIR CCD dataset [151] is 256x256 pixels, which is cropped from the 1024x1024 image in the LEVIR-CD dataset [98]. The image resolution is 0.5 m/pixel. The dataset contains a total of 500 pairs of bi-temporal images, each of which is annotated with 5 textual descriptions that detail the changes occurring between the two acquisitions. The dataset contains 2500 change descriptions in total, with an average sentence length of 15.12 words.

*3) **LEVIR-CC**:* Compared to previous small datasets (i.e., DUBAI CCD and LEVIR CCD), LEVIR-CC [23] is a large-scale change captioning benchmark dataset widely used in current research. It includes 10,077 image pairs with a spatial resolution of 0.5 m/pixel, each containing 256×256 pixels. The images are sourced from the LEVIR-CD dataset [98] and cover 20 regions across Texas, USA, with a time span of 5 to 14 years. The dataset captures various types of ground object changes, including buildings, roads, and vegetation. For each image pair, five human annotators provided descriptive sentences outlining the changes observed, resulting in a total of 50,385 change descriptions. The dataset was curated to exclude trivial changes, such as lighting variations, and emphasize substantial changes like the appearance or disappearance of objects. It offers a rich variety of descriptions with an average sentence length of 11 words.

B. Dataset Matching Temporal Images, Text, and Masks

These datasets extend the annotations to change detection masks. Combining change images, corresponding textual descriptions, and pixel-level masks provides a richer context for models to interpret spatial and semantic changes. These datasets support tasks like multi-task learning of Change Detection and Change Captioning. The current datasets are compared in Fig. 11 and Table VI.

*1) **LEVIR-MCI**:* The LEVIR-MCI dataset [103] is a multi-task change interpretation dataset, containing pixel-level change masks and semantic-level descriptive annotations. Derived from the previous LEVIR-CC dataset [23], LEVIR-MCI further annotates the bi-temporal images with change detection masks. It contains 10,077 bitemporal image pairs (256×256 pixels, 0.5 m/pixel resolution), each annotated with change detection masks for roads and buildings, alongside five descriptive sentences. This dual annotation approach supports precise spatial change detection and high-level semantic understanding. With over 40,000 annotated road and building change masks, LEVIR-MCI captures diverse object scales and deformations. The LEVIR-MCI dataset bridges the gap between fine-grained change detection and high-level semantic understanding. It provides a large-scale benchmark dataset for multi-task learning of change detection and change captioning, which has been widely used in current research.

*2) **LEVIR-CDC**:* The LEVIR-CDC dataset [170] is also an extension of the LEVIR-CC dataset [23], further providing binary building change detection masks. This dataset is similar to the LEVIR-MCI dataset [103] in that both datasets are annotated with change detection masks. However, the LEVIR-CDC dataset only provides masks for building change detection while the LEVIR-MCI dataset provides masks for multiple classes of change detection (i.e., buildings and roads). Besides, building change detection masks of both datasets are almost

TABLE V
Comparison of datasets matching temporal images and text.

Dataset	Image Size/Resolution	Image pairs	Captions	Annotation	Download Link
DUBAI CCD [151]	50×50 (30m)	500	2,500	Manual	https://disi.unitn.it/~melgani/datasets.html
LEVIR CCD [151]	256×256 (0.5m)	500	2,500	Manual	https://disi.unitn.it/~melgani/datasets.html
LEVIR-CC [23]	256×256 (0.5m)	10,077	50,385	Manual	https://github.com/Chen-Yang-Liu/LEVIR-CC-Dataset

TABLE VI
Comparison of datasets matching temporal images, text and masks.

Dataset	Image Size/Resolution	Image pairs	Captions	Pixel-level Masks	Annotation	Download Link
LEVIR-MCI [103]	256×256 (0.5m)	10,077	50,385	44,380 (building, road)	Manual	https://huggingface.co/datasets/lcybuaa/LEVIR-MCI
LEVIR-CDC [170]	256×256 (0.5m)	10,077	50,385	– (building)	Manual	https://huggingface.co/datasets/hygge10111/RS-CDC
WHU-CDC [170]	256×256 (0.075m)	7,434	37,170	– (building)	Manual	https://huggingface.co/datasets/hygge10111/RS-CDC

TABLE VII
Comparison of datasets matching temporal images and question-answer instructions.

Dataset	Temporal Images	Image Resolution	Instruction Samples	Change-related Task	Annotation	Download Link
CDVQA [54]	2,968 pairs (bi-temporal)	0.5m~3m	122,000	CVQA	Manual	https://github.com/YZHJessica/CDVQA
ChangeChat-87k [176]	10,077 pairs (bi-temporal)	0.5m	87,195	CVQA, Grounding	Automated	https://github.com/hanlinwu/ChangeChat
GeoLLaVA [180]	100,000 pairs (bi-temporal)	–	100,000	CVQA	Automated	https://github.com/HosamGen/GeoLLaVA
TEOChatlas [101]	– (variable temporal length)	–	554,071	Classification, CVQA, Grounding	Automated	https://github.com/ermongroup/TEOChat
QVG-360K [55]	6,810 pairs (bi-temporal)	0.1m~3m	360,000	CVQA, Grounding	Automated	https://github.com/like413/VisTA

the same because most of the masks are collected from the LEVIR-CD dataset [98].

*3) **WHU-CDC**:* The WHU-CDC dataset [170] focuses on the Christchurch region of New Zealand, documenting building construction after the 2011 earthquake. This dataset is derived from the WHU-CD change detection dataset [198], where the binary change detection masks reveal the changed building areas. The WHU-CDC dataset further provides change description annotations. It comprises 7,434 high-resolution bitemporal image pairs (256 × 256 pixels at 0.075 m resolution), annotated with five descriptive sentences per pair, resulting in a total of 37,170 captions, with a vocabulary size of 327 unique words. The dataset captures changes across five categories: buildings, parking lots, roads, vegetation, and water.

C. Dataset Matching Temporal Images and Question-Answer Instructions

These datasets pair temporal images with question-answering tasks. The annotations are designed to encourage deeper temporal reasoning, guiding models to analyze sequences of images over time and respond to natural language instructions or queries. These datasets support tasks like temporal visual question answering and grounding. A comparison of current datasets is shown in Table VII.

*1) **CDVQA**:* The CDVQA dataset [54] focuses on the change visual question-answering task. Built upon the SECOND semantic change detection dataset [199], the CDVQA dataset adopts a rule-based automated method to generate questions and answers, using semantic change mask information from the SECOND dataset. It contains 2,968 pairs of bi-temporal images and more than 122,000 question-answer pairs. These images cover multiple cities in China, including Shanghai, Hangzhou, and Chengdu, with spatial resolutions ranging from 0.5 m to 3 m. These images annotate non-change regions and six land-cover classes that capture the nature of changes: non-vegetated ground surfaces, buildings, playgrounds, water, low vegetation, and trees. The question types include whether changes occurred, types of changes, increase or decrease in changes, maximum/minimum changes, and change proportions.

*2) **ChangeChat-87k**:* The ChangeChat-87k dataset [176] is a large-scale change instruction dataset comprising 87,195 instructions tailored for change analysis. Building upon the LEVIR-MCI dataset [103], the ChangeChat-87k dataset is developed using an automated pipeline that combines rule-based methods and ChatGPT-assisted techniques to produce diverse instruction-response pairs. It supports six distinct instruction types: change captioning, binary change detection, category-specific change quantification, change localization, GPT-assisted change instruction, and multi-turn conversations.

*3) **GeoLLaVA**:* The GeoLLaVA dataset [180] is built upon the fMoW dataset [200], a high-resolution satellite imagery dataset featuring 62 categories and a global timespan from 2002 to 2017. To construct temporal image pairs, images were sorted by location and timestamp, ensuring a minimum 12-month interval between paired images. This process yielded 100,000 training pairs and 6,042 test pairs. Text annotations were generated using OpenAI's GPT-4o mini model with prompts that elicited independent descriptions of each image and explicit summaries of the changes between them. All data is structured in a conversational format compatible with VLM fine-tuning. However, using GPT-4o mini to describe image changes may not be reliable enough.

*4) **TEOChatlas**:* TEOChatlas [101] is a temporal Earth observation vision-language instruction-following dataset, containing 554,071 examples. It includes a diverse range of tasks, spanning both single-image and temporal instruction-following scenarios, to support spatial reasoning and complex temporal analysis. Temporal tasks cover categories like temporal scene classification, change detection, temporal referring expressions, and temporal question answering. The dataset incorporates variable-length temporal sequences sourced from diverse EO sensors, including bitemporal (xBD [201] and S2Looking [202]), pentatemporal (QFabric [203]), and multi-temporal (fMoW [200]) data. The dataset spans imagery from eight sensors, including commonly used platforms like Sentinel-2 and WorldView-2, enhancing its versatility and broad applicability across real-world scenarios.

*5) **QAG-360K**:* The QAG-360K dataset [55]represents a pioneering effort in advancing Change Detection Question Answering and Grounding (CDQAG) task. QAG-360K integrates a curated selection of high-quality remote sensing image pairs sourced from notable datasets, including Hi-UCD [204], SECOND [199], and LEVIR-CD [98]. These images, spanning 24 geographically diverse regions across Estonia, China, and the United States, offer resolutions ranging from 0.1 to 3.0 meters. The QAG-360K dataset comprises 6,810 image pairs annotated with semantic masks for 10 land-cover categories. The LLM is used to generate questions. Then, rule-based methods leverage the mask annotations of existing change detection datasets to generate text answers and corresponding pixel-level masks automatically. The dataset over 360K question-answer-mask triples automatically. Each pair of images contains an average of 53 triples. This dataset encompasses diverse inquiries such as identifying the presence, type, magnitude, and ratio of changes, providing a rich context for exploring complex temporal dynamics.

VIII. Future Prospects and Discussion

With the continuous development of remote sensing multi-modal learning, the research on remote sensing temporal image understanding has made significant progress, particularly in tasks such as change captioning and change visual question answering. However, challenges remain in this area, and future research could focus on the following aspects.

*1) **Large-scale Benchmark Datasets**:* At present, research in temporal vision-language understanding primarily relies on a few standard datasets, which are limited in terms of scale and diversity, thus failing to meet the growing demands of the field. Although some studies have attempted to extend dataset size through rule-based or LLM-based automated annotation methods, these approaches still face challenges in terms of annotation quality and diversity. Future research should focus on developing more comprehensive remote sensing temporal vision-language datasets that cover a wider range of scenarios and time points, in order to support more complex temporal image understanding tasks.

*2) **Temporal Vision-Language Foundation Models**:* Previous research in temporal vision-language understanding has primarily concentrated on single tasks such as change captioning and change visual question answering. Given the inherent relationships between various temporal vision-language understanding tasks, future research could explore the development of a unified temporal vision-language foundation model, which would enhance the flexibility and efficiency of temporal image analysis. By integrating the powerful reasoning capabilities of LLMs, the foundation model could address multiple tasks simultaneously and promote synergy between different tasks, thus improving overall model performance.

*3) **Variable Temporal Vision-Language Understanding**:* Current methods primarily focus on change analysis using dual-temporal images. With advancements in remote sensing technology and the increasing number of satellites with varying revisit cycles, multi-temporal remote sensing imagery is becoming more prevalent. The temporal coverage of image sequences is wide, ranging from hourly to yearly intervals, with varying sequence lengths. Future research should focus on effectively processing image sequences of arbitrary temporal lengths to capture richer spatiotemporal information. By enhancing the understanding of multi-temporal images, models will better grasp spatiotemporal trends and spatial distributions, thereby improving their ability to reason about complex change patterns. However, this research poses greater challenges in model design.

*4) **Multi-modal Temporal Images**:* Existing research has largely focused on the temporal visual language understanding of optical imagery. However, with the diversification of satellite sensor types, future research could shift towards multi-modal temporal image understanding. Multi-modal imagery includes different types of data, such as optical, SAR, and infrared images, which provide rich and complementary information, enabling a more comprehensive understanding of spatiotemporal changes. Images from different modalities have distinct perceptual characteristics, each offering a unique perspective on the same temporal changes. For example, optical imagery is well-suited for capturing ground details, while SAR imagery can provide stable observations under adverse weather conditions.

*5) **Temporal Agents**:* Agents based on LLMs have demonstrated strong potential in multi-task execution and automated reasoning [186]. In the future, agents for remote sensing temporal image understanding could be designed to dynamically adjust task execution strategies based on user needs, autonomously performing complex tasks such as change captioning and changed object counting. Furthermore, integrating intelligent agents with external knowledge bases would further enhance their deep reasoning capabilities in complex temporal contexts, providing more flexible and efficient solutions for temporal remote sensing image understanding.

IX. Conclusion

By integrating computer vision with natural language processing, Remote Sensing Temporal Vision-Language Models (RS-TVLMs) have greatly enhanced the ability to analyze temporal remote sensing data, with applications in disaster monitoring, environmental analysis, and urban planning. This review has examined the advancements in RS-TVLMs, including the fundamental concepts, principal methodologies, datasets, and evaluation metrics. By reviewing existing studies, we seek to outline clear pathways and future directions for research in this domain. Besides, several challenges remain, such as collecting large-scale datasets, designing foundation models, and processing multi-temporal image sequences.

References

[1] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 22–36, 2016.

[2] Z. Zhang and L. Zhu, "A review on unmanned aerial vehicle remote sensing: Platforms, sensors, data processing methods, and applications," *drones*, vol. 7, no. 6, p. 398, 2023.

[3] L. Zhu, J. Suomalainen, J. Liu, J. Hyyppä, H. Kaartinen, H. Haggren *et al.*, "A review: Remote sensing sensors," *Multi-purposeful application of geospatial data*, vol. 19, 2018.

[4] P. Roy, M. Behera, and S. Srivastav, "Satellite remote sensing: sensors, applications and techniques," *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*, vol. 87, pp. 465–472, 2017.

[5] R. R. Navalgund, V. Jayaraman, and P. Roy, "Remote sensing applications: An overview," *current science*, pp. 1747–1766, 2007.

[6] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: A survey," *Earth Science Informatics*, vol. 12, pp. 143–160, 2019.

[7] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2207–2219, 2016.

[8] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," *Remote sensing of environment*, vol. 236, p. 111402, 2020.

[9] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[10] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

[11] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol. 159, pp. 296–307, 2020.

[12] Z. Li, Y. Wang, N. Zhang, Y. Zhang, Z. Zhao, D. Xu, G. Ben, and Y. Gao, "Deep learning-based object detection techniques for remote sensing images: A survey," *Remote Sensing*, vol. 14, no. 10, p. 2385, 2022.

[13] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, p. 114417, 2021.

[14] I. Kotaridis and M. Lazaridou, "Remote sensing image segmentation advances: A meta-analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 309–322, 2021.

[15] D. Tuia, K. Schindler, B. Demir, X. X. Zhu, M. Kochupillai, S. Džeroski, J. N. van Rijn, H. H. Hoos, F. Del Frate, M. Datcu, V. Markl, B. Le Saux, R. Schneider, and G. Camps-Valls, "Artificial intelligence to advance earth observation: A review of models, recent trends, and pathways forward," *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–25, 2024.

[16] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 270–294, 2022.

[17] 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.

[18] M. Noman, M. Fiaz, H. Cholakkal, S. Narayan, R. Muhammad Anwer, S. Khan, and F. Shahbaz Khan, "Remote sensing change detection with transformers trained from scratch," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[19] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 228–239, 2022.

[20] Z. Lv, H. Huang, X. Li, M. Zhao, J. A. Benediktsson, W. Sun, and N. Falco, "Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective," *Proceedings of the IEEE*, vol. 110, no. 12, pp. 1976–1991, 2022.

[21] L. Wang, M. Zhang, X. Gao, and W. Shi, "Advances and challenges in deep learning-based change detection for remote sensing images: A review through various learning paradigms," *Remote Sensing*, vol. 16, no. 5, p. 804, 2024.

[22] S. Chang and P. Ghamisi, "Changes to captions: An attentive network for remote sensing change captioning," *IEEE Transactions on Image Processing*, 2023.

[23] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[24] Q. Zhou, J. Gao, Y. Yuan, and Q. Wang, "Single-stream extractor network with contrastive pre-training for remote-sensing change captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[25] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[26] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[27] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[28] J. Wu, W. Gan, Z. Chen, S. Wan, and S. Y. Philip, "Multimodal large language models: A survey," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 2247–2256.

[29] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, p. nwae403, 2024.

[30] X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geoscience and Remote Sensing Magazine*, 2024.

[31] L. Bashmal, Y. Bazi, F. Melgani, M. M. Al Rahhal, and M. A. Al Zuair, "Language integration in remote sensing: Tasks, datasets, and future directions," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 4, pp. 63–93, 2023.

[32] Y. Zhou, L. Feng, Y. Ke, X. Jiang, J. Yan, X. Yang, and W. Zhang, "Towards vision-language geo-foundation model: A survey," *arXiv preprint arXiv:2406.09385*, 2024.

[33] S. Zhuang, P. Wang, G. Wang, D. Wang, J. Chen, and F. Gao, "Improving remote sensing image captioning by combining grid features and

transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[34] H. Kandala, S. Saha, B. Banerjee, and X. X. Zhu, "Exploring transformer and multi-label classification for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–1, 2022.

[35] C. Liu, R. Zhao, and Z. Shi, "Remote sensing image captioning based on multi-layer aggregated transformer," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–1, 2022.

[36] Z. Chen, J. Wang, A. Ma, and Y. Zhong, "Typeformer: Multiscale transformer with type controller for remote sensing image caption," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[37] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "Rsvqa: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020.

[38] C. Chappuis, V. Zermatten, S. Lobry, B. Le Saux, and D. Tuia, "Prompt-rsvqa: Prompting visual context to a language model for remote sensing visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1372–1381.

[39] Z. Zhang, L. Jiao, L. Li, X. Liu, P. Chen, F. Liu, Y. Li, and Z. Guo, "A spatial hierarchical reasoning network for remote sensing visual question answering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[40] ——, "A spatial hierarchical reasoning network for remote sensing visual question answering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[41] L. Bashmal, Y. Bazi, F. Melgani, R. Ricci, M. M. Al Rahhal, and M. Zuair, "Visual question generation from remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 3279–3293, 2023.

[42] S. Li, L. Mi, J. Castillo-Navarro, and D. Tuia, "Knowledge-aware visual question generation for remote sensing images," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 498–502.

[43] S. Zhang, Y. Li, and S. Mei, "Exploring uni-modal feature learning on entities and relations for remote sensing cross-modal text-image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[44] G. Hoxha, F. Melgani, and B. Demir, "Toward remote sensing image retrieval under a deep image captioning perspective," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4462–4475, 2020.

[45] H. Yu, F. Yao, W. Lu, N. Liu, P. Li, H. You, and X. Sun, "Text-image matching for cross-modal remote sensing image retrieval via graph neural network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 812–824, 2022.

[46] Y. Zhan, Z. Xiong, and Y. Yuan, "Rsvg: Exploring data and models for visual grounding on remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[47] Y. Sun, S. Feng, X. Li, Y. Ye, J. Kang, and X. Huang, "Visual grounding in remote sensing images," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 404–412.

[48] K. Li, D. Wang, H. Xu, H. Zhong, and C. Wang, "Language-guided progressive attention for visual grounding in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[49] Y. Hu, J. Yuan, C. Wen, X. Lu, and X. Li, "Rsgpt: A remote sensing vision language model and benchmark," *arXiv preprint arXiv:2307.15266*, 2023.

[50] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "Geochat: Grounded large vision-language model for remote sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 831–27 840.

[51] C. Pang, J. Wu, J. Li, Y. Liu, J. Sun, W. Li, X. Weng, S. Wang, L. Feng, G.-S. Xia *et al.*, "H2rsvlm: Towards helpful and honest remote sensing large vision language model," *arXiv preprint arXiv:2403.20213*, 2024.

[52] D. Muhtar, Z. Li, F. Gu, X. Zhang, and P. Xiao, "Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model," *arXiv preprint arXiv:2402.02544*, 2024.

[53] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao, "Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[54] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, "Change detection meets visual question answering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[55] K. Li, F. Dong, D. Wang, S. Li, Q. Wang, X. Gao, and T.-S. Chua, "Show me what and where has changed? question answering and grounding for remote sensing change detection," *arXiv preprint arXiv:2410.23828*, 2024.

[56] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[57] G. Cheng, Y. Huang, X. Li, S. Lyu, Z. Xu, H. Zhao, Q. Zhao, and S. Xiang, "Change detection methods for remote sensing in the last decade: A comprehensive review," *Remote Sensing*, vol. 16, no. 13, p. 2355, 2024.

[58] T. Bai, L. Wang, D. Yin, K. Sun, Y. Chen, W. Li, and D. Li, "Deep learning for change detection in remote sensing: a review," *Geo-spatial Information Science*, vol. 26, no. 3, pp. 262–288, 2023.

[59] D. Wen, X. Huang, F. Bovolo, J. Li, X. Ke, A. Zhang, and J. A. Benediktsson, "Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 68–101, 2021.

[60] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[61] Z. Lv, P. Zhong, W. Wang, Z. You, and N. Falco, "Multiscale attention network guided with change gradient image for land cover change detection using remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[62] J. Wang, Y. Zhong, and L. Zhang, "Change detection based on supervised contrastive learning for high-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[63] S. Xiang, M. Wang, X. Jiang, G. Xie, Z. Zhang, and P. Tang, "Dual-task semantic change detection for remote sensing images using the generative change field module," *Remote Sensing*, vol. 13, no. 16, p. 3336, 2021.

[64] Q. Zhu, X. Guo, W. Deng, S. Shi, Q. Guan, Y. Zhong, L. Zhang, and D. Li, "Land-use/land-cover change detection based on a siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 63–78, 2022.

[65] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sensing*, vol. 12, no. 10, p. 1688, 2020.

[66] W. A. Malila, "Change vector analysis: An approach for detecting forest changes with landsat," in *LARS symposia*, 1980, p. 385.

[67] O. A. Carvalho Júnior, R. F. Guimarães, A. R. Gillespie, N. C. Silva, and R. A. Gomes, "A new approach to change vector analysis using distance and similarity measures," *Remote Sensing*, vol. 3, no. 11, pp. 2473–2493, 2011.

[68] J. Deng, K. Wang, Y. Deng, and G. Qi, "Pca-based land-use change detection and analysis using multitemporal and multisensor satellite data," *International Journal of Remote Sensing*, vol. 29, no. 16, pp. 4823–4838, 2008.

[69] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and *k*-means clustering," *IEEE geoscience and remote sensing letters*, vol. 6, no. 4, pp. 772–776, 2009.

[70] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (mad) and maf postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sensing of Environment*, vol. 64, no. 1, pp. 1–19, 1998.

[71] 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.

[72] O. Abd El-Kawy, J. Rød, H. Ismail, and A. Suliman, "Land use and land cover change detection in the western nile delta of egypt using remote sensing data," *Applied geography*, vol. 31, no. 2, pp. 483–494, 2011.

[73] T. Chou, T. Lei, S. Wan, and L. Yang, "Spatial knowledge databases as applied to the detection of changes in urban land use," *International Journal of Remote Sensing*, vol. 26, no. 14, pp. 3047–3068, 2005.

[74] A. P. Tewkesbury, A. J. Comber, N. J. Tate, A. Lamb, and P. F. Fisher, "A critical synthesis of remotely sensed optical image change detection techniques," *Remote Sensing of Environment*, vol. 160, pp. 1–14, 2015.

[75] Z. Li, C. Yan, Y. Sun, and Q. Xin, "A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[76] C. Wu, B. Du, and L. Zhang, "Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[77] H. Zhang, H. Chen, C. Zhou, K. Chen, C. Liu, Z. Zou, and Z. Shi, "Bifa: Remote sensing image change detection with bitemporal feature alignment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[78] K. Chen, C. Liu, W. Li, Z. Liu, H. Chen, H. Zhang, Z. Zou, and Z. Shi, "Time travelling pixels: Bitemporal features integration with foundation model for remote sensing image change detection," *arXiv preprint arXiv:2312.16202*, 2023.

[79] 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.

[80] D. Zheng, Z. Wu, J. Liu, Y. Xu, C.-C. Hung, and Z. Wei, "Explicit change-relation learning for change detection in vhr remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[81] Z. Lv, P. Zhong, W. Wang, Z. You, and N. Falco, "Multi-scale attention network guided with change gradient image for land cover change detection using remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[82] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[83] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[84] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[85] B. Varma, N. Naik, K. Chandrasekaran, M. Venkatesan, and J. Rajan, "Forecasting land-use and land-cover change using hybrid cnn–lstm model," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[86] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource vhr images via deep siamese convolutional multiple-layers recurrent neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 4, pp. 2848–2864, 2019.

[87] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[88] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[89] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[90] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[91] 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.

[92] Q. Li, R. Zhong, X. Du, and Y. Du, "Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[93] H. Lin, R. Hang, S. Wang, and Q. Liu, "Diformer: A difference transformer network for remote sensing change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[94] 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.

[95] F. Rahman, B. Vasu, J. Van Cor, J. Kerekes, and A. Savakis, "Siamese network with multi-level features for patch-based change detection in satellite imagery," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2018, pp. 958–962.

[96] B. Hou, Q. Liu, H. Wang, and Y. Wang, "From w-net to cdgan: Bitemporal change detection via deep learning techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 1790–1802, 2019.

[97] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.

[98] 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.

[99] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "Pga-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection," *Remote Sensing*, vol. 12, no. 3, p. 484, 2020.

[100] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[101] J. A. Irvin, E. R. Liu, J. C. Chen, I. Dormoy, J. Kim, S. Khanna, Z. Zheng, and S. Ermon, "Teochat: A large vision-language assistant for temporal earth observation data," *arXiv preprint arXiv:2410.06234*, 2024.

[102] R. Ferrod, L. Di Caro, and D. Ienco, "Towards a multimodal framework for remote sensing image change retrieval and captioning," *arXiv preprint arXiv:2406.13424*, 2024.

[103] C. Liu, K. Chen, H. Zhang, Z. Qi, Z. Zou, and Z. Shi, "Change-agent: Toward interactive comprehensive remote sensing change interpretation and analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[104] M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter, "xlstm: Extended long short-term memory," *arXiv preprint arXiv:2405.04517*, 2024.

[105] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[106] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[107] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[108] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[109] S. Jamil, M. Jalil Piran, and O.-J. Kwon, "A comprehensive survey of transformers for computer vision," *Drones*, vol. 7, no. 5, p. 287, 2023.

[110] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," *arXiv preprint arXiv:2202.10936*, 2022.

[111] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023.

[112] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—a contemplative retrospection," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106126, 2023.

[113] A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia *et al.*, "Transformers in remote sensing: A survey," *arXiv preprint arXiv:2209.01206*, 2022.

[114] R. Wang, L. Ma, G. He, B. A. Johnson, Z. Yan, M. Chang, and Y. Liang, "Transformers for remote sensing: A systematic review and analysis," *Sensors*, vol. 24, no. 11, p. 3495, 2024.

[115] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[116] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[117] X. Wang, S. Wang, Y. Ding, Y. Li, W. Wu, Y. Rong, W. Kong, J. Huang, S. Li, H. Yang, Z. Wang, B. Jiang, C. Li, Y. Wang, Y. Tian, and J. Tang, "State space model for new-generation network alternative to transformers: A survey," 2024.

[118] R. Xu, S. Yang, Y. Wang, B. Du, and H. Chen, "A survey on vision mamba: Models, applications and challenges," *arXiv preprint arXiv:2404.18861*, 2024.

[119] B. N. Patro and V. S. Agneeswaran, "Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges," *arXiv preprint arXiv:2404.16112*, 2024.

[120] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi, “Rsmamba: Remote sensing image classification with state space model,” *arXiv preprint arXiv:2403.19654*, 2024.

[121] S. Zhao, H. Chen, X. Zhang, P. Xiao, L. Bai, and W. Ouyang, “Rs-mamba for large remote sensing image dense prediction,” *arXiv preprint arXiv:2404.02668*, 2024.

[122] C. Liu, K. Chen, B. Chen, H. Zhang, Z. Zou, and Z. Shi, “Rscama: Remote sensing image change captioning with state space model,” *IEEE Geoscience and Remote Sensing Letters*, 2024.

[123] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang *et al.*, “Pre-trained models: Past, present and future,” *AI Open*, vol. 2, pp. 225–250, 2021.

[124] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, “A comprehensive overview of large language models,” *arXiv preprint arXiv:2307.06435*, 2023.

[125] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, “A survey on evaluation of large language models,” *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[126] T. Kumar, Z. Ankner, B. F. Spector, B. Bordelon, N. Muennighoff, M. Paul, C. Pehlevan, C. Ré, and A. Raghunathan, “Scaling laws for precision,” *arXiv preprint arXiv:2411.04330*, 2024.

[127] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, “Parameter-efficient fine-tuning for large models: A comprehensive survey,” *arXiv preprint arXiv:2403.14608*, 2024.

[128] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[129] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, “An empirical study of gpt-3 for few-shot knowledge-based vqa,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3081–3089.

[130] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[131] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.

[132] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, “Gemini: a family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.

[133] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, “Recent advances in natural language processing via large pre-trained language models: A survey,” *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2023.

[134] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu *et al.*, “Instruction tuning for large language models: A survey,” *arXiv preprint arXiv:2308.10792*, 2023.

[135] Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, and Q. Liu, “Aligning large language models with human: A survey,” *arXiv preprint arXiv:2307.12966*, 2023.

[136] T. OpenAI, “Chatgpt: Optimizing language models for dialogue,” *OpenAI*, 2022.

[137] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.

[138] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Dang *et al.*, “Qwen2. 5-coder technical report,” *arXiv preprint arXiv:2409.12186*, 2024.

[139] A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin *et al.*, “Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement,” *arXiv preprint arXiv:2409.12122*, 2024.

[140] Z. Wang, M. Wang, S. Xu, Y. Li, and B. Zhang, “Ccexpert: Advancing mllm capability in remote sensing change captioning with difference-aware integration and a foundational dataset,” *arXiv preprint arXiv:2411.11360*, 2024.

[141] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[142] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[143] C. Liu, R. Zhao, J. Chen, Z. Qi, Z. Zou, and Z. Shi, “A decoupling paradigm with prompt learning for remote sensing image change captioning,” *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[144] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[145] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[146] D. H. Park, T. Darrell, and A. Rohrbach, “Robust change captioning,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4623–4632.

[147] Y. Qiu, S. Yamamoto, K. Nakashima, R. Suzuki, K. Iwata, H. Kataoka, and Y. Satoh, “Describing and localizing multiple changes with transformers,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1951–1960.

[148] C. Liu, J. Yang, Z. Qi, Z. Zou, and Z. Shi, “Progressive scale-aware network for remote sensing image change captioning,” in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 6668–6671.

[149] C. Cai, Y. Wang, and K.-H. Yap, “Interactive change-aware transformer network for remote sensing image change captioning,” *Remote Sensing*, vol. 15, no. 23, p. 5611, 2023.

[150] S. Chouaf, G. Hoxha, Y. Smara, and F. Melgani, “Captioning changes in bi-temporal remote sensing images,” in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 2891–2894.

[151] G. Hoxha, S. Chouaf, F. Melgani, and Y. Smara, “Change captioning: A new paradigm for multitemporal remote sensing image analysis,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2022.

[152] W. Peng, P. Jian, Z. Mao, and Y. Zhao, “Change captioning for satellite images time series,” *IEEE Geoscience and Remote Sensing Letters*, 2024.

[153] D. Sun, Y. Bao, J. Liu, and X. Cao, “A lightweight sparse focus transformer for remote sensing image change captioning,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[154] Q. Zhou, J. Gao, Y. Yuan, and Q. Wang, “Single-stream extractor network with contrastive pre-training for remote sensing change captioning,” *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[155] X. Yu, Y. Li, and J. Ma, “Diffusion-rscc: Diffusion probabilistic model for change captioning in remote sensing images,” *arXiv preprint arXiv:2405.12875*, 2024.

[156] Y. Tu, L. Li, L. Su, Z.-J. Zha, C. Yan, and Q. Huang, “Context-aware difference distilling for multi-change captioning,” *arXiv preprint arXiv:2405.20810*, 2024.

[157] D. Vyshnav, L. Gutha, A. P. V. Manindra, and B. Karthikeyan, “Intelli-change remote sensing-a novel transformer approach,” in *2024 Second International Conference on Data Science and Information System (ICDSIS)*. IEEE, 2024, pp. 1–7.

[158] R. Tsujimoto, H. Ouchi, H. Kamigaito, and T. Watanabe, “Towards temporal change explanations from bi-temporal satellite images,” *arXiv preprint arXiv:2407.09548*, 2024.

[159] C. Chen, Y. Wang, and K.-H. Yap, “Multi-scale attentive fusion network for remote sensing image change captioning,” in *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2024, pp. 1–5.

[160] F. Zhang, W. Zhang, K. Xia, and H. Feng, “Scale-wised feature enhancement network for change captioning of remote sensing images,” *International Journal of Remote Sensing*, vol. 45, no. 17, pp. 5845–5869, 2024.

[161] K. Xu, Y. Han, R. Yang, X. Ye, Y. Guo, H. Xing, and S. Wang, “Mfrnet: A new multi-scale feature refining method for remote sensing image change captioning,” in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 7119–7123.

[162] Y. Li, X. Zhang, X. Cheng, P. Chen, and L. Jiao, “Inter-temporal interaction and symmetric difference learning for remote sensing image change captioning,” *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[163] R. Liu, K. Li, J. Song, D. Sun, and X. Cao, “Mv-cc: Mask enhanced video model for remote sensing change caption,” *arXiv preprint arXiv:2410.23946*, 2024.

[164] C. Wang, N. He, and B. Wang, "Chareption: Change-aware adaption empowers large language model for effective remote sensing image change captioning," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2024, pp. 342–355.

[165] Y. Yang, T. Liu, Y. Pu, L. Liu, Q. Zhao, and Q. Wan, "Remote sensing image change captioning using multi-attentive network with diffusion model," *Remote Sensing*, vol. 16, no. 21, p. 4083, 2024.

[166] Y. Wang, W. Yu, M. Kopp, and P. Ghamisi, "Changeminds: Multi-task framework for detecting and describing changes in remote sensing," *arXiv preprint arXiv:2410.10047*, 2024.

[167] C. Liu, K. Chen, Z. Qi, Z. Liu, H. Zhang, Z. Zou, and Z. Shi, "Pixel-level change detection pseudo-label learning for remote sensing change captioning," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 8405–8408.

[168] Y. Zhu, L. Li, K. Chen, C. Liu, F. Zhou, and Z. Shi, "Semantic-cc: Boosting remote sensing image change captioning via foundational knowledge and semantic guidance," *arXiv preprint arXiv:2407.14032*, 2024.

[169] C. Yang, Z. Li, H. Jiao, Z. Gao, and L. Zhang, "Enhancing perception of key changes in remote sensing image change captioning," *arXiv preprint arXiv:2409.12612*, 2024.

[170] J. Shi, M. Zhang, Y. Hou, R. Zhi, and J. Liu, "A multi-task network and two large scale datasets for change detection and captioning in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[171] Y. He, X. Feng, C. Cheng, G. Ji, Y. Guo, and J. Caverlee, "Metabalance: improving multi-task recommendations via adapting gradient magnitudes of auxiliary tasks," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2205–2215.

[172] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1871–1880.

[173] X. Li, B. Sun, and S. Li, "Detection assisted change captioning for remote sensing image," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 10 454–10 458.

[174] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[175] Z. Yuan, L. Mou, and X. X. Zhu, "Change-aware visual question answering," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 227–230.

[176] P. Deng, W. Zhou, and H. Wu, "Changechat: An interactive model for remote sensing change analysis via multimodal instruction tuning," *arXiv preprint arXiv:2409.08582*, 2024.

[177] M. Noman, N. Ahsan, M. Naseer, H. Cholakkal, R. M. Anwer, S. Khan, and F. S. Khan, "Cdchat: A large multimodal model for remote sensing change description," *arXiv preprint arXiv:2409.16261*, 2024.

[178] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[179] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.

[180] H. Elgendy, A. Sharshar, A. Aboeitta, Y. Ashraf, and M. Guizani, "Geollava: Efficient fine-tuned vision-language models for temporal change detection in remote sensing," *arXiv preprint arXiv:2410.19552*, 2024.

[181] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[182] T. Huynh, S. Kornblith, M. R. Walter, M. Maire, and M. Khademi, "Boosting contrastive self-supervised learning with false negative cancellation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2785–2795.

[183] L. Xu, H. Xie, F. L. Wang, X. Tao, W. Wang, and Q. Li, "Contrastive sentence representation learning with adaptive false negative cancellation," *Information Fusion*, vol. 102, p. 102065, 2024.

[184] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[185] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[186] S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou *et al.*, "Metagpt: Meta programming for multi-agent collaborative framework," *arXiv preprint arXiv:2308.00352*, 2023.

[187] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[188] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou *et al.*, "The rise and potential of large language model based agents: A survey," *arXiv preprint arXiv:2309.07864*, 2023.

[189] J. Xie, Z. Chen, R. Zhang, X. Wan, and G. Li, "Large multimodal agents: A survey," *arXiv preprint arXiv:2402.15116*, 2024.

[190] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[191] Y. Mu, Q. Zhang, M. Hu, W. Wang, M. Ding, J. Jin, B. Wang, J. Dai, Y. Qiao, and P. Luo, "Embodiedgpt: Vision-language pre-training via embodied chain of thought," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[192] Papineni, Kishore, Roukos, Salim, Ward, Todd, Zhu, and Wei-Jing, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[193] Lin and C. Yew, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004.

[194] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, ser. StatMT '07. USA: Association for Computational Linguistics, 2007, p. 228–231.

[195] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[196] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[197] C. Wu, Z. Lin, S. Cohen, T. Bui, and S. Maji, "Phrasecut: Language-based image segmentation in the wild," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 10 213–10 222.

[198] 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.

[199] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, M. Pelillo, and L. Zhang, "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[200] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6172–6180.

[201] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas, H. Choset, and M. Gaston, "Creating xbd: A dataset for assessing building damage from satellite imagery," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 10–17.

[202] L. Shen, Y. Lu, H. Chen, H. Wei, D. Xie, J. Yue, R. Chen, S. Lv, and B. Jiang, "S2looking: A satellite side-looking dataset for building change detection," *Remote Sensing*, vol. 13, no. 24, p. 5094, 2021.

[203] S. Verma, A. Panigrahi, and S. Gupta, "Qfabric: Multi-task change detection dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1052–1061.

[204] S. Tian, A. Ma, Z. Zheng, and Y. Zhong, "Hi-ucd: A large-scale dataset for urban semantic change detection in remote sensing imagery," *arXiv preprint arXiv:2011.03247*, 2020.