AllSpark: A Multimodal Spatiotemporal General Intelligence Model With Ten Modalities via Language as a Reference Framework

Run Shao, Cheng Yang, Qiujun Li, Linrui Xu, Xiang Yang, Xian Li, Mengyao Li, Qing Zhu, Yongjun Zhang, *Member, IEEE*, Yansheng Li, *Senior Member, IEEE*, Yu Liu, Yong Tang, Dapeng Liu, Shizhong Yang, and Haifeng Li, *Member, IEEE*

***Abstract*— RGB, multispectral, point, and other spatiotemporal modal data fundamentally represent different observational approaches for the same geographic object. Therefore, leveraging multimodal data is an inherent requirement for comprehending geographic objects. However, due to the high heterogeneity in structure and semantics among various spatiotemporal modalities, the joint interpretation of multimodal spatiotemporal data has long been an extremely challenging problem. The primary challenge resides in striking a trade-off between the cohesion and autonomy of diverse modalities. This trade-off becomes progressively nonlinear as the number of modalities expands. Inspired by the human cognitive system and linguistic philosophy, where perceptual signals from the five senses converge into language, we introduce the language as reference framework (LaRF), a fundamental principle for constructing a multimodal unified model. Building upon this, we propose AllSpark, a multimodal spatiotemporal general artificial intelligence model. Our model integrates ten different modalities into a unified framework, including 1-D (language, code, and table), 2-D (RGB, synthetic aperture radar (SAR), multispectral, hyperspectral, graph, and trajectory), and 3-D (point cloud) modalities. To achieve modal cohesion, AllSpark introduces a modal bridge and multimodal large language model (LLM) to map diverse modal features into the language feature space. To maintain modality autonomy, AllSpark uses modality-specific encoders to extract the tokens of various spatiotemporal modalities. Finally, observing a gap between the model's interpretability and downstream tasks, we designed modality-specific prompts and task heads, enhancing the model's generalization capability across specific tasks. Experiments indicate that the incorporation of language enables AllSpark to excel in few-shot classification tasks for RGB and point cloud modalities without additional training, surpassing baseline performance by up to 41.82%. Additionally, AllSpark, despite lacking expert knowledge in most spatiotemporal modalities and utilizing a unified structure, demonstrates strong adaptability across ten modalities. LaRF and AllSpark contribute to the shift in the research paradigm in spatiotemporal intelligence, transitioning from a modality-specific and task-specific paradigm to a general paradigm. The source code is available at https://github.com/GeoX-Lab/AllSpark.**

***Index Terms*— General intelligence model, large language model (LLM), multimodal machine learning, spatiotemporal data.**

Received 25 September 2024; revised 20 November 2024; accepted 5 January 2025. Date of publication 7 January 2025; date of current version 29 January 2025. This work was supported in part by the Major Program Project of Xiangjiang Laboratory under Grant 22XJ01010, in part by the National Natural Science Foundation of China under Grant 42271481, and in part by Using Computing Resources at the High-Performance Computing Platform of Central South University. *(Corresponding author: Haifeng Li.)*

Run Shao, Cheng Yang, Qiujun Li, Linrui Xu, Xiang Yang, Xian Li, Mengyao Li, and Haifeng Li are with the School of Geosciences and Info-Physics, Central South University, Changsha 410083, China, and also with the Xiangjiang Laboratory, Changsha 410205, China (e-mail: lihaifeng@csu.edu.cn).

Qing Zhu is with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China.

Yongjun Zhang and Yansheng Li are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China.

Yu Liu is with the School of Earth and Space Sciences, Peking University, Beijing 100871, China.

Yong Tang and Dapeng Liu are with Huawei Technologies Company Ltd., Shenzhen 518129, China.

Shizhong Yang is with BDS Micro Chip Inc., Changsha 410071, China.

Digital Object Identifier 10.1109/TGRS.2025.3526725

NOMENCLATURE

M	AllSpark.
f_i	Modal encoder for m_i.
Φ	Modal bridge.
T	Text tokenizer.
F	Multimodal LLM.
H_{task}	Task head.
L	Loss function.
Enc_n	n-layer transformer encoders.
Emb	Embedding layer.
$\text{Enc}_{\text{ResNet}}$	ResNet.
Grouper	Point grouper of the PointBERT [48].
Conv1d	1-D convolution layer.
FFN	Feedforward network.
σ	Softmax.
m_i	Modal i.
q	Query vectors of the bridge.
s_i	Tokens of m_i.
p_i	Text prompt for m_i.
θ	Model parameters.
y	Label.
t_i	Tokens of m_i.
W	Weights of the linear layer.
w_i	Word of the code or text.

I. INTRODUCTION

BENEFITING from the increasingly diverse observational methods available for spatiotemporal scenes, geographic objects can be described by various spatiotemporal modalities, such as RGB, synthetic aperture radar (SAR), multispectral, graph, point cloud, and trajectory data [7], [8], [9]. Each modality provides unique information about different aspects of the geographic object. Analogous to the human process of perceiving and understanding the world through multiple modalities, such as vision, hearing, and touch, joint interpretation of multimodal data is an inherent requirement for intelligent models to achieve cognition of geographic objects.

However, due to the inherent differences in the mechanisms of each modality, various modalities are often highly heterogeneous in both structure and semantics. For example, in terms of structure, a table is composed of rows and columns. Point clouds are represented by 3-D coordinates along with various feature values. A text is composed of sequences of words. In terms of semantics, RGB imagery reflects the electromagnetic characteristics of visible light bands emitted and reflected by geographic objects, whereas SAR imagery reflects the electromagnetic characteristics of microwaves emitted actively by radar for scattering from geographic objects.

For a long time, constrained by the high heterogeneity across various modalities mentioned above, researchers have often developed specific methods based on prior assumptions related to a particular modality or designed multimodal approaches for a few low-heterogeneity modalities. For instance, in single-modal research, Qi et al. [10] proposed PointNet for the point cloud modality, emphasizing the invariance of point cloud data ordering and the significance of global and local features. For language modality, Vaswani et al. [11] introduced the transformer, which focuses on the long-range dependencies within word sequences. For the graph modality, Kipf and Welling [12] and Li et al. [13] proposed the graph convolutional network (GCN) based on the adjacency relationships between nodes in a graph. In multimodal research, the fusion of optical and SAR imagery has been widely explored in both traditional and deep learning remote sensing [14]. Moreover, visual-language models have undergone rapid development in recent years [15], [16], [17], [18], [19]. The diverse prior assumptions associated with each modality have resulted in significant gaps between methods designed for different modalities, making it challenging to perceive and understand different modalities using a unified model.

We believe that the key challenge in addressing this issue lies in striking a trade-off between the cohesion and autonomy of diverse modalities. In our article, "cohesion" refers to the presence of mutually correlated shared information among modalities. For instance, both RGB and SAR images may describe the contour of the same object. "Autonomy," on the other hand, refers to the existence of unique information specific to each modality relative to others. For example, the RGB modality can describe an object's color and texture, while the SAR modality can capture the object's scattering properties in relation to radar waves. Cohesion forms the foundation for the interrelation between modalities, while autonomy highlights the value of multimodal joint interpretation—gaining a complete understanding of an object by integrating multiple modalities.

If we merely project data from different modalities into a shared representation space to emphasize intermodality cohesion, this approach risks losing modality-specific information, ultimately undermining the unique contributions of each modality and weakening the core value of multimodal collaboration. In contrast, if we excessively stress the autonomy between modalities, it may hinder the establishment of connections among them, limiting the model's ability to simultaneously perceive multiple modalities. Moreover, as the number of modalities increases, balancing cohesion and autonomy becomes progressively more challenging nonlinearly.

We observe that in the process of comprehending the world, humans integrate information from multiple modalities, such as hearing, touch, smell, and vision. The concepts formed through the parsing of these modalities ultimately converge in language. Humans engage in associating, reasoning, and expressive behaviors through language. In other words, language precisely encodes human perception and understanding of the world, providing clear definitions and meanings to abstract concepts from each modality. Inspired by this, we propose the language as reference framework (LaRF) as a fundamental principle for constructing multimodal models. It means that the abstract concepts derived from each modality should align with language, enabling joint interpretation in the unified representation space of language.

Building upon this, we propose a multimodal spatiotemporal general intelligence model [20], AllSpark, that integrates ten different modalities into a unified framework, including 1-D (language, code, and table), 2-D (RGB, SAR, multispectral, hyperspectral, graph, and trajectory), and 3-D (point cloud) modalities. As shown in Table I and Fig. 1, previous work either overlooked some important spatiotemporal modalities, such as hyperspectral and trajectory, or focused solely on natural images without considering remote sensing imagery. AllSpark covers a broader range of spatiotemporal modalities, such as multispectral, hyperspectral, graph, trajectory, and more, while demonstrating excellent few-shot learning capabilities and modality adaptability.

To achieve modal cohesion, AllSpark uniformly maps diverse modal features to the language feature space. To maintain the autonomy between modalities, AllSpark introduces specific modal encoders for each modality to extract independent tokens. Given the high heterogeneity among modality data and modality encoders, a significant dimensional gap exists between the tokens of each modality and the language modality. To address this issue, we introduce a modality bridge, a mechanism from perceiver [21], to accomplish dimensional mapping from each modality's tokens to the language tokens [22].

Finally, considering the existing gap between the interpretability of the multimodal large language model (LLM) and the specific downstream tasks, we design task heads and modality-specific text prompts for each downstream task to enhance the model's generalization capability. Given the powerful interpretability capabilities of the multimodal LLM, we adhere to a lightweight design principle in task heads.

Experiments demonstrate that AllSpark, despite lacking expert knowledge in most spatiotemporal modalities and utilizing a unified structure, achieves competitive accuracy in modalities such as RGB and spatiotemporal trajectories

TABLE I
AllSpark Integrates Ten Spatiotemporal Modalities

Model	1D			2D						3D
	Language	Code	Table	RGB	MSI	HSI	SAR	Traj	Graph	Point
EarthGPT [1]	✔			✔			✔			
SkySense [2]				✔	✔		✔			
PointLLM [3]	✔									✔
Table-LLaVA [4]	✔		✔							
OneLLM [5]	✔			✔						✔
Meta-transformer [6]	✔		✔	✔		✔			✔	✔
AllSpark	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

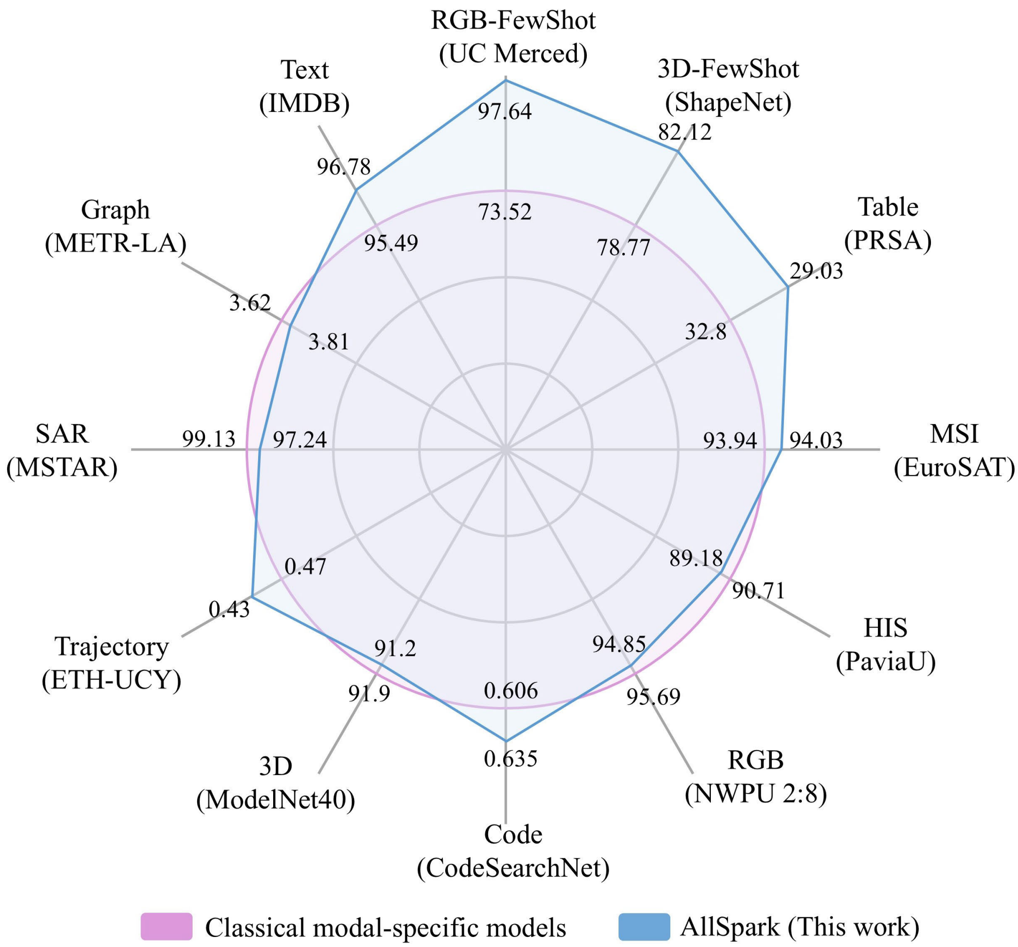


Fig. 1. AllSpark demonstrates excellent adaptability across up to ten heterogeneous modalities and shows outstanding few-shot learning capabilities in RGB and point cloud modalities.

compared to state-of-the-art models. Specifically, in the RGB modality, the accuracy of AllSpark is only 0.84 lower than that of the SOTA model, and in the trajectory modality, the average displacement error (ADE) metric differs by only 0.07 compared to that of the SOTA model. Additionally, AllSpark exhibits excellent adaptability in various other modalities, including point cloud, multispectral, hyperspectral, table, graph, and code. Theoretically, our proposed model has the potential for seamless extension to an arbitrary number of modalities.

In other words, our contributions can be summarized as follows.

1) We first propose a unified multimodal spatiotemporal general model, AllSpark, that successfully integrates ten spatiotemporal modalities into a single model.
2) Inspired by the human cognitive system and linguistic philosophy, we propose the LaRF, which offers a novel solution to balance cohesion and autonomy among multiple modalities.
3) Experiments indicate that AllSpark demonstrates strong few-shot learning capabilities and supports ten modalities. Theoretically, our proposed model has the potential for seamless extension to an arbitrary number of modalities.

II. Related Work

Leveraging multimodal data is an inherent requirement for achieving cognitive recognition of geospatial objects. An ideal multimodal model should possess the capability to integrate all the modalities for joint interpretation. Hence, a crucial trend in the research of intelligent methods in the spatiotemporal domain is the continual increase in the number of modalities available for joint interpretation.

Initially, early researchers often constructed single-modal expert models based on prior assumptions about a specific modality, achieving remarkable success within each respective modality. In recent years, with a deeper understanding of single-modal interpretation methods, numerous researchers have attempted to integrate several low-heterogeneity modalities to construct multimodal interpretation approaches. However, as the number of modalities increases, the challenge of balancing cohesion and autonomy among the modalities becomes increasingly difficult.

In the following, we recall the development of intelligent methods in the spatiotemporal domain from the perspective of the continually increasing number of modalities, and finally, we present the principles and approach of our proposed model, AllSpark.

A. Single-Modal Model

For 1-D modalities, we focus on code, language, and table. Given the excellent characteristics of code, such as strict syntax, unambiguous nature, and ability to interact with machines, code is treated as a separate modality.

Feng et al. [23] pretrained a model, CodeBERT, which facilitates the mutual transfer of information between code and natural language modalities. For the language modality, landmark contributions include a transformer [11], BERT [24], and the GPT series [25], [26], [27], [28], which have inspired subsequent series of works.

The table is one of the commonly used modalities for recording and expressing information. TabNet, proposed by Arik and Pfister [29], employs a sequence attention mechanism to achieve feature selection in the table modality, thereby enabling interpretable and more efficient learning.

We categorize RGB, multispectral, hyperspectral, SAR, graph, and trajectory as 2-D modalities.

Among the 2-D modalities, standard three-channel RGB images are among the most common. For the RGB modality, the ResNet proposed by He et al. [30], which is based on the importance of visual global and local information, and the vision transformer (ViT) introduced by Dosovitskiy et al. [31], which leverages a global attention mechanism, represent two landmark contributions.

An increase in the number of channels in images leads to multispectral and hyperspectral modalities. Huang et al. [32] proposed the STDCNN, leveraging the characteristic of a greater number of bands in multispectral images to simultaneously model the global spatial and spectral properties of multispectral images. In comparison to multispectral images, hyperspectral images have even more bands, often reaching hundreds. Based on this, Yang et al. [33] introduced the R-3D-CNN to further enhance the extraction of spectral features.

In the case of SAR images formed by active microwave radar, Chen et al. [34] introduced AConvNet, a widely used fully convolutional neural network for intelligent SAR image interpretation.

The trajectory modality reflects the temporal changes in the spatial positions of objects. Gupta et al. [35] proposed the social generative adversarial networks (GANs) based on the characteristic of trajectory multiplicity. This model combines historical trajectory information with social context information to predict multiple plausible future outcomes.

For the graph modality, Kipf and Welling [12] introduced the classic GCN, which is based on the adjacency relationships between nodes in the graph. Veličković et al. [36] proposed the GAT, which incorporates attention mechanisms into the graph modality.

Finally, we turn our attention to 3-D modalities: point cloud.

The point cloud modality captures information about the position, shape, color, texture, and other aspects of 3-D objects. Qi et al. [10] introduced the classic PointNet for the point cloud modality, emphasizing the importance of invariance through point data permutation and the significance of global and local features. Wu et al. [37] extended convolution operations to 3-D point clouds with the introduction of PointConv.

B. Multimodal Model

While the single-modal methods in Section II-A have demonstrated excellent performance within their respective modalities, they often face challenges in generalizing across multiple modalities due to their construction based on specific prior assumptions. Recognizing the intrinsic requirement for intelligent models to utilize multimodal data for geographic object cognition, numerous researchers have endeavored to balance the cohesion and autonomy among modalities to construct multimodal models.

Sadeghian et al. [38] extended the social GANs [35] and introduced RGB images to enhance scene data in Sophie, achieving better results in trajectory prediction tasks. Recognizing the high complementarity between RGB and SAR images, Hughes et al. [39] proposed a three-step deep neural network framework that utilizes a universal prediction of matching regions, generates heatmaps, and eliminates outliers to match RGB and SAR images. Li et al. [40] introduced the DTCDN, a model that employs a GAN network to migrate RGB and SAR images to the same feature space, facilitating target detection. Yang et al. [41] proposed a dual-stream convolutional network that uses high-resolution multispectral images to enhance the spatial resolution of hyperspectral images.

To achieve joint interpretation of multimodal data, traditional multimodal models typically design specific architecture based on the priors of certain modalities. For example, Hang et al. [42] proposed Coupled CNNs, which consist of two CNN networks to extract spectral–spatial features from hyperspectral data and elevation information from LiDAR data. They ultimately use both feature-level and decision-level fusion methods to integrate the heterogeneous features of the two modalities. Similarly, Zhang et al. [43] proposed SLA-Net, which designs specific network structures to extract and fuse spatial information and morphological characteristics from hyperspectral imagery. These methods are typically designed for specific purposes and modalities, making it difficult to extend them to more modalities.

Additionally, Gao et al. [44] proposed DFINet, which extracts self-correlation and cross correlation between multimodal data to deeply fuse hyperspectral and multispectral modality features. Likewise, Hong et al. [45] introduced S2FL, which extracts modality-specific subspaces for each modality and a shared subspace for all modalities, finally obtaining multimodal interpretation results through a unified projection. The issue with such methods is the lack of a unified alignment reference, limiting them to collaboration between a few modalities. As the number of modalities increases, the complexity of aligning multiple modalities will grow exponentially.

Notably, CLIP, proposed by Radford et al. [15], associates the RGB modality with the text modality using contrastive learning [46], [47]. With pretraining guided by weak supervisory signals from text, CLIP has demonstrated outstanding capabilities in both visual single-modal tasks and visual-language multimodal tasks, inspiring a series of subsequent works [16], [17], [18], [19]. Zhang et al. [6] introduced a meta-transformer, leveraging the contrastive learning paradigm from CLIP to pretrain a universal backbone network under the visual-language modality. It exhibits multimodal generalization abilities across various modalities, such as point cloud, infrared, and hyperspectral data. Han et al. [5] directly employed a multimodal LLM as a universal backbone network, proposing the One-LLM, which successfully unifies eight modalities, namely, images, audio, videos, and points. The success of these approaches implies the unique role of language modalities in multimodal models.

Building upon the aforementioned efforts, we systematically propose the fundamental principle of the LaRF. Guided by this principle, we balance cohesion and autonomy among diverse modalities and introduce a general intelligent model named AllSpark, which unifies ten spatiotemporal modalities and possesses the potential to extend to an arbitrary number of modalities.

III. Method

A. Language as Reference Framework

We observe that in the process of comprehending the world, humans integrate information from multiple modalities, such as hearing, touch, smell, and vision. The concepts formed through the parsing of these modalities ultimately converge in language. Humans engage in associating, reasoning, and expressive behaviors through language. In other words, language precisely encodes human perception and understanding of the world, providing clear definitions and meanings to abstract concepts from each modality.

Inspired by our observation, we introduce the fundamental principle of LaRF to balance cohesion and autonomy among multiple modalities.

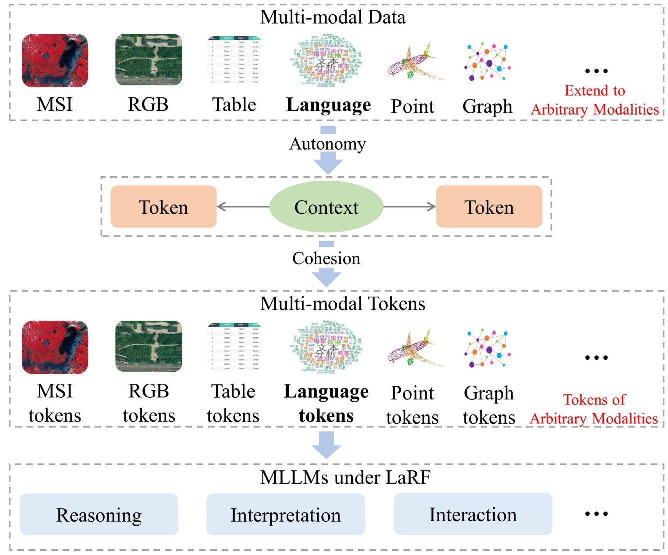


Fig. 2. Guided by the LaRF principle, multimodal data are transformed into a token-context structure akin to language, based on their respective prior assumptions. This approach preserves the autonomy of each modality while achieving cohesion between them, enabling the interpretation of multimodal data within a unified language representation space.

In terms of the cohesion of multimodalities, the high heterogeneity between multiple spatiotemporal modalities is a major challenge, while the LaRF principle defines the alignment anchor between multimodalities as language explicitly. As shown in Fig. 2, we observe that language is encoded by tokens and their contexts, and this structure can be extended to most spatiotemporal modalities. Therefore, we can align highly heterogeneous spatiotemporal modalities to language modalities in structure and semantics, enabling multimodal interpretation in the unified representation space of language. Additionally, the pivotal role of natural language prompts is a key factor in the LaRF principle's ability to achieve cohesion across modalities.

In contrast, we can independently encode multiple spatiotemporal modalities into token sequences under their respective prior assumptions, so the LaRF principle does not lead to the loss of modal autonomy. More importantly, LaRF is not dependent on specific modalities; therefore, theoretically, as long as token representations of modalities can be obtained, the multimodal model guided by LaRF can be extended to arbitrary modalities.

In summary, the significance of LaRF is as follows.

1) *Alignment Capability:* Language can accurately encode both cohesion and autonomy information across multiple modalities. Aligning each modality with the language modality enables a unified representation in the same feature space, addressing the challenge of high heterogeneity among modalities.
2) *Reasoning Capability:* Language, as a tool for human thought and expression, inherently possesses the ability to perform complex reasoning. Each modality, when represented in a unified space with LaRF, inherits the reasoning capability of language, unlocking the potential for multimodal joint reasoning.

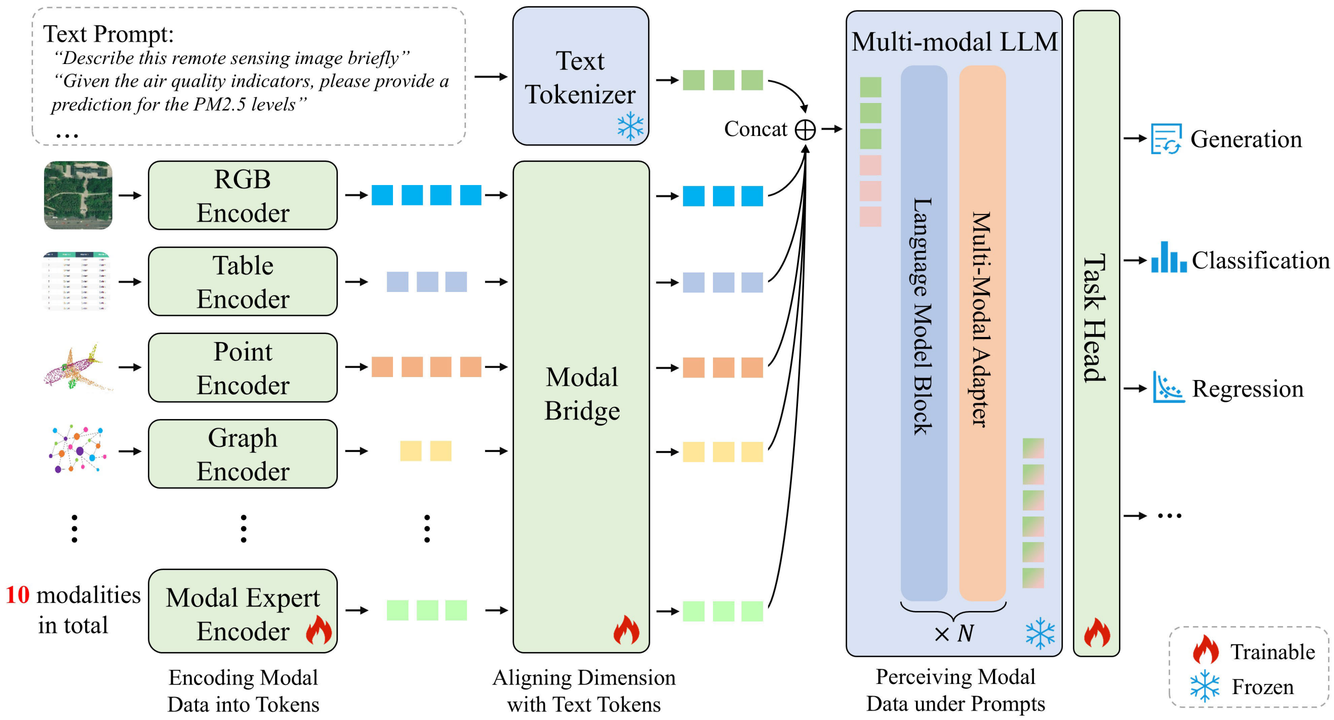


Fig. 3. AllSpark architecture. Multimodal data are extracted by their respective modal encoders into token sequences. Following dimension alignment with modality-specific text prompt tokens via a modal bridge, both the text prompt tokens and modality tokens are passed into a large language multimodal model for interpretation. The interpretation results are then aligned with downstream tasks through task-specific heads.

3) *Interpretability:* Deep learning methods have long been characterized as "black boxes." However, a multimodal intelligent system constructed based on the LaRF can directly leverage language as a tool. This facilitates the output of interpretable reasoning chains that humans can understand, thereby achieving true explainable artificial intelligence.
4) *Interactivity:* Language not only aids humans in understanding intelligent models but also facilitates intelligent models in understanding humans. In an intelligent system guided by the LaRF, humans can directly express their needs using natural language. This iterative correction of the model's output based on human interaction will become a new paradigm for the training and inference of intelligent models.
5) *Scalability:* The multimodal system guided by the LaRF is agnostic to specific modalities. New modalities need to establish a mapping to the language model only to participate in joint reasoning with other modalities. Therefore, theoretically, a multimodal model based on LaRF can be extended to an arbitrary number of modalities.

B. Overview

Guided by the principles mentioned above, AllSpark consists of five modules: the modal encoder, modal bridge, text tokenizer, multimodal LLM, and task head. The overall architecture is depicted in Fig. 3.

To maintain autonomy among modalities, we designed modality-specific encoders to encode highly heterogeneous data into modality-independent tokens (for details, see Section III-C). However, the dimensions of tokens outputted by modal encoders are still inconsistent. To parse in the unified representational space of language, we introduced the modal bridge from the Lynx [49]. The modal bridge aims to project tokens from each modality into the dimension of the multimodal LLM (Section III-D for details). The formalization of this process is as follows:

$$s_i = \Phi(f_i(m_i), q). \quad (1)$$

Here, $\{m_{\mathrm{RGB}}, m_{\mathrm{MSI}}, m_{\mathrm{HSI}}, \ldots, m_i, \ldots,\}$ represents the inputs from various modalities, where m_{RGB} represents the RGB modality, m_{MSI} represents the multispectral modality, and so on. f_i is defined as the modal encoder for the m_i modality, Φ represents the modal bridge, and $q \in R^{N*D}$ represents N learnable vectors of dimension D in the modality bridge, where D is set to 4096, representing the dimensionality of the multimodal LLM. The input data m_i of each modality are mapped to a token sequence $s_i \in R^{N*D}$ of the same dimension as the language model. Nomenclature section summarizes the main mathematical symbols and their meanings in the model.

To achieve cohesion among modalities, we employ a unified multimodal LLM to parse data from various modalities. The text tokenizer and multimodal LLM in AllSpark are based on the visual-language model Lynx [49]. To extend Lynx to ten spatiotemporal modalities, we designed specific text prompts for each modality to guide the model in correctly parsing information from each modality. Additionally, Lynx incorporates several lightweight multimodal adapter layers internally to accommodate multimodal inputs. We continue this design and do not freeze the parameters of the adapter layers during training to enhance the model's adaptability to other spatiotemporal modalities. Finally, we acknowledge that a gap exists between the parsing results of the model and those of the downstream task. Therefore, we design specific task heads for each task to enhance the model's generalization capability.

The entire model M can be formalized as follows:

$$M(m_i, p_i) = H_{\text{task}}(F(s_i \oplus T(p_i))) \tag{2}$$

where p_i represents the text prompt of m_i, T denotes the text tokenizer, $\oplus$ indicates the concatenation operation of text tokens and modality tokens in the sequence, F represents the multimodal LLM, and H_{task} signifies the task head.

All the tasks in our experiments are supervised tasks, with y denoting the labels, L representing the loss function, and θ representing the learnable parameters in our model. The optimization objective of the model can be formalized as follows:

$$\theta_i = \arg\min_{\theta} L(y, M(m_i, p_i); \theta_i). \tag{3}$$

C. Independent Encoder for Each Modality

The modal encoder aims to encode the raw data of each modality into a token sequence, formalized as $t_i = f_i(m_i)$, where $t_i \in R^{n*d}$. We designed different modal encoders for each modality to maintain autonomy among modalities. The following provides individual introductions for each modality:

1) 1-D Modal: Code & Language: Code is essentially a specialized form of language. But due to its distinct properties such as having a strict syntax, being unambiguous, and being capable of interacting with machines, we separate it as a distinct modality. Therefore, to avoid ambiguity, we will use "text" and "language" interchangeably to distinguish between the natural language and code modalities. Exploring intelligent methods for the code modality is crucial for reliable AI reasoning and achieving interaction between intelligent models and the real world. Since the Lynx model is a language model, we do not design an additional modal encoder for the code and text modalities. Instead, we directly utilize Lynx's text tokenizer, i.e., $f(m_{\text{Text/Code}}) = T(m_{\text{Text/Code}})$, where $m_{\text{Text/Code}} = \{w_1, w_2, w_3, \ldots,\}$ represents the sequence of words in the text or code.

Table: A table can be viewed as a sequence of rows, with each row containing several fields, or columns, i.e., $m_{\text{Table}} \in R^{\text{row}*\text{col}}$. The modal encoder for the table modality inherits the design from TabFormer [50]: first, based on the different degrees of discreteness for each column attribute, we use independent Embedding layers to encode discrete and continuous values separately. Subsequently, we employ a single-layer transformer encoder to further extract features. The entire process can be formalized as $f(m_{\text{Table}}) = \text{Enc}_1(\text{Emb}(m_{\text{Table}}))$.

2) 2-D Modal: RGB: RGB imagery represents the visible light spectrum and reflects the electromagnetic characteristics of objects that emit or reflect visible light waves. It is the most common modality in the field of computer vision. RGB imagery is a standard three-band image, i.e., $m_{\text{RGB}} \in R^{H*W*3}$. For the modal encoder of this modality, we adopted the visual encoder from the Lynx model: EVA [51]. The EVA is a large visual model composed of 40 stacked transformer blocks with a width of 1408. During the experiments, AllSpark loaded the official weights of the EVA model and froze them during training.

MSI: Multispectral imagery is a modality extensively studied in the remote sensing field. It incorporates multiple nonvisible light bands, such as near-infrared, shortwave infrared, coastal atmospheric aerosol, and cirrus bands. Therefore, the number of channels in MSIs is usually greater than the three bands in RGB imagery, i.e., $m_{\text{MSI}} \in R^{H*W*C}$, where $C > 3$. We extended the PatchEmbed of the standard ViT [31], modifying its channel count to match the number of bands in the input multispectral imagery. This adaptation allows it to serve as the feature encoder for the multispectral modality.

HSI: Hyperspectral imagery increases the number of bands compared to multispectral imagery, often reaching hundreds of bands, with each band containing rich information. Unlike RGB and multispectral imagery, where an image serves as a single sample, in hyperspectral imagery, all bands for each pixel are treated as a single sample, i.e., $m_{\text{HSI}} \in R^{1*1*C}$. In the modal encoder for hyperspectral imagery, we first use a linear projection layer to expand the feature dimensions for each pixel. This process is formalized as $W * m_{\text{HSI}}^T$, where $W \in R^{1*d}$ is the weight matrix of the linear projection layer. Subsequently, we use a 12-layer standard transformer encoder to extract its features. The entire process can be represented as $f(m_{\text{HSI}}) = \text{Enc}_{12}(W * m_{\text{HSI}}^T)$.

Trajectory: The trajectory modality reflects the changing information of an object over time and space and is composed of a series of 2-D coordinate points, i.e., $m_{\text{Trajectory}} \in R^{l*2}$, where l represents the sequence length of trajectory points. The encoder for the trajectory modality inherits the design from TUTR [52]: first, a linear layer is used to expand the dimensions of the 2-D trajectory features. This step can be formalized as $W * m_{\text{Trajectory}}$, where $W \in R^{d*(l*2)}$ is the weight matrix of the linear projection layer. Then, we use a 2-layer transformer encoder to extract its features. The entire process can be formalized as $f(m_{\text{Trajectory}}) = \text{Enc}_2(W * m_{\text{Trajectory}})$.

SAR: The SAR modality is a type of active remote sensing that reflects the electromagnetic characteristics of objects with respect to microwave backscatter. Due to differences in polarization modes, the final product of SAR imagery is typically a two-band or single-band image, i.e., $m_{\text{SAR}} \in R^{H*W*2}$. Therefore, we designed a simple three-layer convolutional network as the modal encoder.

Graph: A graph is constructed from a series of nodes and edges, where the attributes of the nodes and the adjacency attributes of the nodes reflect the majority of the features of the graph, i.e., $m_{\text{Graph}} \in R^{K*d}$, where K is the number of nodes and d is the node feature dimension. In AllSpark, the modality encoder of a graph is based on the STAEformer [53], [54], whose main design idea is to first use a linear layer to extend its feature dimension and then use several Embedding layers to separately encode features such as the node's characteristics, spatial characteristics, and temporal characteristics. The entire process can be formalized as $f(m_{\text{Graph}}) = \text{Emb}_{\text{node}}(W * m_{\text{Graph}}^{T}) \oplus \text{Emb}_{\text{spatial}}(W * m_{\text{Graph}}^{T}) \oplus \text{Emb}_{\text{time}}(W * m_{\text{Graph}}^{T})$, where $W \in R^{\text{hidden}*d}$ is the weight of the linear layer.

3) 3-D Modal: Point Cloud: A point cloud is typically composed of 3-D coordinates and feature values $m_{\text{PointCloud}} \in R^{K*(d+3)}$, where K represents the number of 3-D points and d represents the dimensionality of the point cloud features, reflecting information such as the spatial position, shape, color, and texture of objects. The encoder for the point cloud modality inherits the design from PointBERT [48]: first, point cloud data are grouped and encoded to unify the number of points simultaneously inputted. This step can be represented as $\text{PointGroup} = \text{Grouper}(m_{\text{PointCloud}}) \in R^{G*N*3}$, where G represents the number of groups and N represents the number of points in each group. Next, the grouped results are input into a 1-D convolutional layer to extract feature vectors for each group: $f_{\text{Group}} = \text{Conv1d}(\text{PointGroup}) \in R^{G*d}$. Finally, the feature vectors for each group are input into a standard 12-layer transformer encoder to extract their global features. The entire process can be formalized as $f(m_{\text{PointCloud}}) = \text{Enc}_{12}(\text{Conv1d}(\text{Grouper}(m_{\text{PointCloud}})))$.

D. Modal2Language Bridge

Although the modal encoders have transformed data from various heterogeneous modalities into a unified token sequence, there are still differences in dimensions between different modal tokens, making it difficult to perceive by a multimodal LLM. The modal bridge, based on the Perceiver [21], aims to perform dimensional projection from tokens of various modalities to tokens of the language modality. In its implementation, the modal bridge consists of stacked cross-attention layers and feedforward neural network layers.

In the cross-attention layer, we predefine a learnable query vector $Q \in R^{N*D}$, where D is the internal dimensionality of the language model and N serves as a hyperparameter that can be flexibly adjusted to accommodate inputs from different modalities. The keys and values in the cross-attention layer are the features outputted by the modal encoders.

The feedforward neural network inherits the classic design from the original transformer and consists of two linear layers with an inserted activation layer.

The entire process can be formalized as follows:

$$\Phi(Q, s_i) = \text{FFN}\left(\sigma\left(\frac{\text{QW}_q^T\left(s_i W_k^T\right)^T}{\sqrt{D}}\right) s_i W_v^T\right). \quad (4)$$

Here, $W_q \in R^{D*\text{hidden}}$, $W_k \in R^{d*\text{hidden}}$, and $W_v \in R^{d*\text{hidden}}$ are the linear projection layer weights defined inside the cross-attention layer for Q, K, and V, respectively. σ denotes the softmax operation.

E. Task-Guided Text Prompts and Task Heads

To extend the visual-language multimodal model to ten spatiotemporal modalities without intervention from modality expert knowledge, we designed specific text prompts and task heads for each modality and task. Text prompts are used to guide the multimodal language model in correctly interpreting each modality's data, while task heads are employed to match the model's parsing results with specific downstream tasks.

TABLE II
SUMMARY OF HYPERPARAMETERS

Dimension	Modal	Max Lr	Max Epochs	Warm-up Epochs
1D	Language	9.0×10^6	5	1
	Code	1.0×10^5	4	1
	Table	2.0×10^5	30	3
2D	RGB	5.0×10^5	50	5
	MSI	2.0×10^5	50	5
	HSI	1.0×10^4	30	3
	SAR	9.0×10^6	30	3
	Trajectory	1.0×10^5	30	5
	Graph	8.0×10^5	10	2
3D	Point Cloud	3.0×10^5	100	10

We manually designed one to four specific text prompts for each modality. During the training process, to enhance model performance, we employed a strategy of diversifying prompts, randomly selecting one prompt for each forward pass. However, during testing, for the sake of result stability and reproducibility, the prompt was fixed to be the first prompt among all prompts. Table XVI provides a list of all the text prompts.

To ensure the transferability of modalities across different tasks, the design principle for task heads is to be as simple and lightweight as possible.

For classification tasks or downstream tasks that can be formalized as classification tasks, we uniformly use a simple single-layer linear layer as the task head. For instance, we implemented standard classification tasks on the RGB, MSI, SAR, and point cloud modalities. Although the task involving the HSI modality is segmentation, it can be formalized as a per-pixel classification task. Therefore, a single-layer linear layer is used as the task head for the mentioned modalities.

For regression tasks on the table, trajectory, and graph modalities, we also use a linear layer to perform regression predictions. The only difference from classification tasks is the addition of an unscaled operation without learnable parameters. Since the Lynx itself is a language model, the code and text modalities directly use its native text decoder.

IV. Experiment

A. Setup

Our experiments aim to demonstrate the following: 1) the distinct advantages of AllSpark over traditional models and 2) AllSpark's ability to understand ten spatiotemporal modalities simultaneously.

For the former, we believe that since language is a reference framework, AllSpark's advantage over traditional models lies in the richer semantic meaning of its features. In LLMs [26], [27], [55], [56], a common way to measure this property is through few-shot learning. Therefore, we evaluated AllSpark's performance on few-shot classification tasks in the RGB and point cloud modalities without any extra training steps. The experimental results can be found in Section IV-C.

For the latter, we select a task for each modality to conduct the evaluation. Following the principles of simplicity and reproducibility, we choose the widely studied datasets in each modality's respective field and employ similar experimental settings across all modalities. Specifically, we use the AdamW optimizer with a learning rate schedule based on cosine annealing. The hyperparameters are adjusted slightly in terms of training epochs and learning rates for different experiments. The specific details on dataset selection can be found in Section IV-B. Table II summarizes the hyperparameter settings for the experiments on each modality. The experimental results can be found in Section IV-D.

B. Dataset

Below, we will provide detailed explanations of datasets in order.

Language: The IMDB [57] dataset is a binary sentiment analysis dataset consisting of 50 000 reviews from the Internet Movie Database (IMDb) labeled as positive or negative. Additionally, the dataset includes some unlabeled data. In our experiments, only the labeled data from the IMDB dataset were utilized for supervised sentiment classification tasks.

Code: CodeSearchNet [58] is a large-scale dataset of function code and its documentation from GitHub that covers six programming languages: Go, Java, JavaScript, PHP, Python, and Ruby. The task performed on the code modality is code document generation, and we tested it on the Ruby and JavaScript.

Table: The PRSA [65] dataset is a collection of air quality data from multiple stations in Beijing that contains hourly measurements of air pollutants. The data spans from March 1, 2013, to February 28, 2017, across 12 monitoring stations. In our experiments, we used various features, including time, station information, four air pollutant variables (SO_2, NO_2, CO, and O_3), and six meteorological variables (temperature, pressure, dew point temperature, amount of precipitation, wind speed, and wind direction). The task was to predict the concentration of PM2.5. We split the data into training (40%) and testing sets.

RGB: NWPU-RESISC45 [66] is a large-scale open dataset for visible light remote sensing image scene classification. The dataset included 45 land use categories, such as airplanes, baseball diamonds, beaches, and commercial areas. Each category included 700 remote sensing images, for a total of 31 500 images. The image size is 256*256 pixels, and we selected a version of the dataset split by the official release using 20% of the data for training.

MSI: The EuroSAT [67] dataset is a multispectral dataset for land use and land cover (LULC) classification. The samples are sourced from the Sentinel-2 optical satellite and include all 13 bands. The data are categorized into ten classes for a total of 27 000 images. We adopted a random 9:1 split for training and testing.

HSI: The Pavia University dataset is a high spatial resolution hyperspectral dataset acquired by the ROSIS sensor. It comprises 103 bands with a size of 610*340 pixels. The dataset includes nine land cover categories, such as asphalt, meadows, and gravel. We used a 4:6 split for training and testing.

Trajectory: The ETH-UCY [68], [69] dataset is a widely used benchmark for pedestrian trajectory prediction and is divided into five subsets: ETH, HOTEL, UNIV, ZARA1, and ZARA2. In our experiments, we utilized the ETH subset.

SAR: The MSTAR [70] dataset is an SAR dataset designed for military stationary target recognition that comprises ten categories of military targets. We employed the standard operating conditions (SOCs) dataset preprocessing method proposed by Chen et al. [34], ensuring that the serial numbers and target configurations are consistent between the test and training sets while the aspects and depression angles differ.

Graph: METR-LA is a traffic dataset collected from loop detectors on the Los Angeles highways spanning from March 1, 2012, to June 30, 2012. The task is traffic flow prediction.

Point Cloud: ModelNet40 [71] is a synthetic point cloud dataset consisting of 40 object categories and a total of 12 311 point cloud objects. We follow the official dataset split, with 9843 objects used for training and 2468 for testing.

C. Few-Shot Learning

To demonstrate AllSpark's unique advantages over traditional models, we tested its few-shot performance on the RGB and point cloud modalities. It is worth noting that traditional few-shot learning methods typically require additional training steps. For instance, common approaches like ProtoNet [61] and MatchingNet [59] involve randomly splitting support and query sets on the training set for supervised training, a step known as meta-learning. Thanks to the integration of natural language, AllSpark requires no extra training step and can directly evaluate few-shot classification accuracy on the test set, significantly outperforming baseline models.

1) RGB: We evaluated AllSpark's RGB image few-shot classification performance on the UC-Merced [72] and WHU-RS19 [73], [74] datasets, following the dataset splits from [62]. Specifically, the UC-Merced dataset uses six classes—Beach, Golf course, Mobile home park, River, Sparse residential, and Tennis court—as the test set, while the WHU-RS19 dataset uses five classes—Commercial, Meadow, Pond, River, and Viaduct—as the test set. AllSpark does not require meta-learning on the training set and is evaluated on 600 episodes directly on the test set, with 15 query samples per episode. As shown in Table III, we report the results for both 5-way 1-shot and 5-way 5-shot settings. The experiments demonstrate that AllSpark significantly outperforms baseline models without requiring any training, highlighting the advantage of LaRF.

2) Point: We evaluated AllSpark's point cloud few-shot classification performance on the ShapeNet [75] and ScanObjectNN [76] datasets, following the dataset settings from [77].

TABLE III
FEW-SHOT CLASSIFICATION RESULTS FOR THE RGB MODALITY

Model	Training-free	UC Merced		RS19	
		5way-1shot	5way-5shot	5way-1shot	5way-5shot
MatchingNet [59]	✗	48.18	67.39	67.68	85.01
RelationNet [60]	✗	50.07	65.22	65.01	79.75
ProtoNet [61]	✗	53.85	71.23	76.36	85.00
DLA-MatchNet [62]	✗	53.76	63.01	68.27	79.89
SPNet [63]	✗	57.64	73.52	81.06	88.04
AllSpark	✔	**95.58**	**97.64**	**97.16**	**98.94**

TABLE IV
FEW-SHOT CLASSIFICATION RESULTS FOR THE POINT MODALITY

Model	Training-free	ShapeNet		ScanObject	
		5way-1shot	5way-5shot	5way-1shot	5way-5shot
FSLGNN [64]	✗	64.98	76.14	29.91	32.77
RelationNet [60]	✗	65.88	76.25	45.32	55.43
ProtoNet [61]	✗	65.96	78.77	**44.75**	**59.81**
AllSpark	✔	**67.20**	**82.12**	35.70	53.33

TABLE V
RGB IMAGE CLASSIFICATION WITH ALLSPARK

Method	Publication	Acc(%)
CNN-CapsNet [78]	RS2019	89.03
DFAGCN [79]	TNNLS2021	89.29
D-CNN with GoogleNet [80]	TGRS2018	90.49
D-CNN with VGGNet [80]	TGRS2018	91.89
SCCov [81]	TNNLS2019	92.1
SeCo-ResNet-50 [82]	ICCV2021	92.91
MG-CAP [83]	TIP2020	92.95
LSENet [84]	TIP2021	93.34
MSANet [85]	JSTARS2021	93.52
IDCCP [86]	TGRS2021	93.76
MBLANet [87]	TIP2021	94.66
GRMANet-ResNet-50 [88]	TGRS2021	94.72
EMSNet [89]	TGRS2023	95.37
ViTAE-B + RVSA [90]	TGRS2022	**95.69**
AllSpark	ours	94.85

TABLE VI
MSI LAND COVER CLASSIFICATION WITH ALLSPARK

Expert Know.	Method	Publication	Acc(%)
✗	ResNet-18 [82]	ICCV2021	63.21
	ResNet-50 [91]	IGARSS2023	91.13
	InceptionNet [92]	ICIP2020	93.07
	EfficientNet [93]	IJSTARS2021	93.94
	AllSpark	**ours**	94.03
	ResNet-152 [91]	IGARSS2023	**96.63**
✔	MoCoV2 [82]	ICCV2021	89.51
	SeCo [82]	ICCV2021	93.14
	SEER [94]	Arxiv2022	97.6
	DINO-MC [95]	Arxiv2023	**98.78**

Similar to the RGB modality, AllSpark does not require meta-learning on the training set and is evaluated on 700 episodes directly on the test set, with 15 query samples per episode. As shown in Table IV, we report the results for both 5-way 1-shot and 5-way 5-shot settings. The experiments demonstrate that AllSpark surpasses most baseline models even on previously unseen point cloud datasets, showcasing its outstanding semantic richness and generalization capabilities.

D. Ability to Understand Ten Spatiotemporal Modalities

1) RGB: We evaluated the performance of AllSpark on the RGB image scene classification task using the NWPU-RESISC45 dataset, with the top-1 accuracy as the evaluation metric. AllSpark leverages expert knowledge from Lynx by loading its pretrained weights. Therefore, in Table V, we compare AllSpark with the state-of-the-art models. The results indicate that AllSpark outperforms most baseline models, with a margin of only 0.84 compared to that of the SOTA (95.69). This highlights AllSpark's exceptional perception and interpretation capabilities in the RGB modality.

2) MSI: We evaluated the performance of AllSpark on the MSI scene classification task using the EuroSAT dataset. In the experiment, all 13 spectral bands of the images were simultaneously input into the model. The model's objective was to correctly classify the images into one of the ten specified categories, and the evaluation metric chosen was the top-1 accuracy. AllSpark does not possess expert knowledge in the multispectral modality, so we categorize the baseline models into two groups, as shown in Table VI: those with expert knowledge intervention and those without. Expert knowledge intervention refers to baseline models pretraining on large datasets such as BigEarthNet and then fine-tuning on the EuroSAT dataset, while no expert knowledge indicates baseline models trained directly from scratch on the EuroSAT dataset. The results show that our model outperforms most models in the no expert knowledge group, with a margin of only 2.60 compared to the state-of-the-art model (ResNet-152). Furthermore, AllSpark lags behind the best result in the expert knowledge group by only 4.75, demonstrating its excellent adaptability to the multispectral modality.

TABLE VII
HSI PIXEL CLASSIFICATION WITH ALLSPARK

Method	OA	AA	Kappa
3D-CNN	75.24	80.26	68.34
CA-GAN	76.81	76.94	71.02
3D VS-CNN	81.63	83.86	76.46
RPNet	84.92	83.26	80.52
S-DMM	88.3	93.76	84.9
IFRF	88.38	85.99	84.97
AllSpark	89.18	86.65	85.32
DCFSL	90.71	90.2	87.73
TC-GAN	93.2	91.6	91
PRNet-RF	**95.6**	**94.96**	**94.27**

TABLE VIII
PM2.5 PREDICTION WITH ALLSPARK

Expert Arch.	Method	PM2.5		
		RMSE	MAE	R2
✗	GWO [97]	62.2	40.8	-
	TabBERT [50]	32.8	-	-
	AllSpark	**29.03**	**18.04**	**0.87**
✔	Stacked ResNet-LSTM [98]	40.68	23.75	0.8
	CBAM-CNN-Bi-LSTM [99]	**18.9**	**11.2**	**0.94**

3) HSI: We conducted a pixel classification task on the Pavia University dataset for the hyperspectral modality. The model treats all spectral bands of a single pixel as one sample and predicts the land cover category of that pixel. The reported metrics include overall accuracy (OA), average accuracy (AA), and kappa. Since AllSpark does not possess expert knowledge of the hyperspectral modality, we compared it with the semisupervised baselines summarized by Uchaev and Uchaev [96]. The results in Table VII demonstrate that AllSpark outperforms many hyperspectral image classification methods, such as IFRF and S-DMM, by a factor of 6.42 compared to the best result in terms of OA, highlighting AllSpark's superior adaptability to the hyperspectral modality.

4) Table: For the table modality, we evaluated AllSpark on the regression prediction task using the PRSA [65] dataset. The task involves predicting the concentration of PM2.5 in the air using features such as time, site, four air pollutants, and six meteorological variables (as detailed in Section IV-B). The performance metrics include the root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R^2).

Table VIII presents the comparative results between AllSpark and the baselines. It is worth noting that some works specifically focus on the prediction task on the PRSA dataset and design expert models with specific architectures based on dataset characteristics. For example, the CBAM-CNN-Bi-LSTM proposed by Li et al. [99] used a CNN to extract spatial dependencies between air monitoring stations and Bi-LSTM to capture the temporal dependencies of PM2.5 data. Similarly, the stacked ResNet-LSTM model proposed by Cheng et al. [98] employs a stacking LSTM strategy to enhance the extraction of temporal features in PM2.5 data. Therefore, we categorize baseline methods into two types: those with an expert architecture and those without an expert architecture. Among the models without expert architecture, our approach achieves the best performance among the baselines and is slightly inferior to the state-of-the-art method for models with expert architecture (CBAM-CNN-Bi-LSTM). This reflects the excellent adaptability of AllSpark to the table modality.

TABLE IX
CODE DOCUMENT GENERATION WITH ALLSPARK

Expert Know.	Method	MRR	
		Ruby	Javascript
✗	BIRNN[68]	0.084	0.153
	1D-CNN[68]	0.245	0.352
	selfAtt[68]	0.365	0.451
	NBoW[68]	0.429	0.461
	RoBERTa[67]	0.625	0.606
	AllSpark	**0.627**	**0.635**
✔	RoBERTa(Code)[67]	0.661	0.64
	CodeBERT[67]	0.693	0.706
	GraphCodeBERT[67]	**0.732**	**0.711**

5) Code: For the code modality, we evaluated the performance of AllSpark on the code document generation task using the CodeSearchNet dataset. This task involves generating corresponding documents based on the provided function code. We conducted tests for both the Ruby and JavaScript languages using the mean reciprocal rank (MRR) [100] as the evaluation metric.

Like in the previous modalities, since MSI-AGI does not possess expert knowledge of the code modality, we categorized the baselines into two groups: those with and without expert knowledge. As shown in Table IX, AllSpark achieved SOTA results in the group without expert knowledge, and the results were comparable to those of models trained with expert knowledge. This finding demonstrates the strong adaptability of AllSpark to the code modality.

6) Point Cloud: For the point cloud modality, we evaluated the performance of AllSpark on the ModelNet40 dataset for the classification task, with the top-1 accuracy as the metric. In the context of single-modal studies focused on point clouds, we observed that due to the unique 3-D structure of point cloud data, most works concentrate on designing specific structures to maintain properties such as permutation invariance and symmetry in 3-D point clouds. However, these structures designed for the unique priors of the modality are challenging to transfer across modalities. Additionally, some methods tend to pretrain on large point cloud datasets to acquire general modal expert knowledge before generalizing to specific downstream tasks to improve performance.

The AllSpark module is designed based on a general sequence-to-sequence architecture. As shown in Table X, in the absence of both modality expert architectural designs and modal expert knowledge, AllSpark still outperforms classical networks with point cloud-specific structures (PointNet and Kd-Net). This approach maintains comparability with the state-of-the-art PointGPT model, which has both a modal expert structure and modal expert knowledge. This finding suggested that AllSpark has significant potential for applications in the point cloud modality.

7) Trajectory: For the trajectory modality, we evaluated the performance of AllSpark on the ETH dataset in the trajectory

TABLE X
POINTCLOUD CLASSIFICATION WITH ALLSPARK

Method	Expert Architecture	Expert Knowledge	Acc(%)
PointNet [10]	✔	✗	89.2
Kd-net [101]	✔	✗	90.6
SPH3D-GCN [102]	✔	✗	91.4
PointNet++ [103]	✔	✗	91.9
SO-Net [104]	✔	✗	92.5
PointVGG [105]	✔	✗	93.6
PointBERT [48]	✗	✔	93.8
PointGPT [106]	✔	✔	**94.9**
AllSpark	✗	✗	91.2

TABLE XI
TRAJECTORY PREDICTION WITH ALLSPARK

Method	ADE/FDE
Social GAN	0.87/1.62
SoPhie	0.70/1.43
STAR	**0.36**/0.64
SGCN	0.63/1.03
CAGN	0.41/0.65
SIT	0.39/0.62
SocialVAE	0.47/0.76
PCENet	0.54/0.87
AgentFormer	0.45/0.75
MemoNet	0.40/0.61
SocialVAE+FPC	0.41/**0.58**
TUTR	0.40/0.61
AllSpark	0.43/0.69

prediction task. This task involves predicting possible 2-D trajectories based on a set of 2-D coordinate points within a certain time period. We report accuracy using the ADE and final displacement error (FDE) [52]. Given the future trajectory$\{x_t, y_t\}_{t=T_{\text{bos}}+1}^{T}$ (ground truth) and the predicted trajectory $\{\widehat{x_t}, \widehat{y_t}\}_{t=T_{\text{bos}}+1}^{T}$, the ADE and FDE are used to measure their $L2$ distances, calculated as follows:

$$\text{ADE} = \frac{1}{T_{\text{pred}}} \sum_{t=T_{\text{bos}}}^{T} \sqrt{\left(x_t - \widehat{x_t}\right)^2 + \left(y_t - \widehat{y_t}\right)^2} \quad (5)$$

$$\text{FDE} = \sqrt{\left(x_T - \widehat{x_T}\right)^2 + \left(y_T - \widehat{y_T}\right)^2}. \quad (6)$$

In Table XI, AllSpark is compared with the state-of-the-art trajectory prediction models. The results indicate that AllSpark, which uses a unified structure without trajectory modality expert knowledge, outperforms most expert models. It achieves a prediction accuracy close to that of the SOTA model (STAR), with a difference of only 0.07 in the ADE metric and 0.11 in the FDE metric. This finding suggested that AllSpark demonstrated excellent adaptability to the trajectory modality.

8) SAR: For the SAR modality, the adaptability of AllSpark was tested on the MSTAR dataset, where the model is required to identify SAR images of ten military targets, and the metric used is the top-1 accuracy. In the experiment, the preprocessing of the MSTAR dataset followed the SOC settings from AConvNets [34]. Table XII presents the comparison results between AllSpark and the state-of-the-art model under these settings. AllSpark achieves 97.24% top1-accuracy, only 1.89% lower than the SOTA model. The experiments demonstrate that AllSpark is capable of effectively understanding SAR imagery.

TABLE XII
SAR CLASSIFICATION WITH ALLSPARK

Method	Acc(%)
EMACH [107]	88
SVM [107]	90
AdaBoost [107]	92
MSRC [108]	93.6
IGT [107]	95
MSS [109]	96.6
Cond Gauss [110]	97
M-PMC [111]	98.8
AConvNets [34]	**99.13**
AllSpark	97.24

TABLE XIII
TRAFFIC PREDICTION WITH ALLSPARK

Method	RMSE	MAE	R2
HI	6.8	14.2	10.15
GWNet	3.51	7.28	9.96
DCRNN	3.54	7.47	10.32
AGCRN	3.59	7.45	10.47
STGCN	3.6	7.43	10.35
GTS	3.59	7.44	10.25
MTGNN	3.47	7.21	9.7
STNorm	3.57	7.51	10.24
GMAN	3.44	7.35	10.07
PDFormer	3.62	7.47	10.91
STID	3.55	7.55	10.95
STAEformer	**3.34**	**7.02**	**9.7**
AllSpark	3.81	7.52	11.24

TABLE XIV
TEXT UNDERSTANDING WITH ALLSPARK

Method	Acc(%)
RoBERTa [112]	95.3
ULMFiT [113]	95.4
BERT [114]	95.49
Mixed VAT [115], [116]	95.68
LongFormer [117]	95.7
XLNet [118]	96.8
ERNIE-Doc-Large [119]	**97.1**
AllSpark	96.78

9) Graph: For the graph modality, the performance of AllSpark was evaluated on the traffic flow prediction task using the METR-LA dataset. The evaluation metrics include the RMSE, MAE, and R^2. Table XIII compares AllSpark and the state-of-the-art methods on the METR-LA dataset, with the baseline derived from [53]. The results show that AllSpark, without the intervention of modal expert knowledge, is only 0.47 away from the best result in terms of the RMSE, demonstrating its excellent adaptability to the graph modality.

10) Language: We tested AllSpark's natural language processing capabilities on the IMDB dataset [57] with the task of binary sentiment classification (positive or negative). AllSpark

TABLE XV
TRAINING AND INFERENCE COSTS

Modality	Total params.(M)	Trainable params.(M)	Training time	Inference MACs(G)
RGB	8176.02	689.80	∼18 hours	1695.98
MSI	7455.14	978.87	∼30 hours	363.02
HSI	7450.58	974.31	∼21 hours	374.61
Table	7486.67	1010.40	∼44 hours	410.00
Code	6876.02	268.66	∼59 hours	3606.88
Point	7594.81	1118.54	∼49 hours	557.38
Trajectory	7381.00	904.73	∼39 hours	538.58
SAR	7452.58	976.31	∼87 hours	373.63
Graph	7223.91	747.64	∼53 hours	1846.28
Language	6876.02	268.66	∼34 hours	3606.88

loads weights from the Lynx; therefore, it can be considered to possess expert knowledge in the natural language modality. Compared with the SOTA models on the IMDB dataset, as shown in Table XIV, AllSpark outperformed most language models, trailing the SOTA result by only 0.32. This highlights AllSpark's powerful understanding and analysis capabilities in natural language.

E. Training and Inference Costs

As the scale of model parameters grows, the training and inference costs of LLMs have been rapidly increasing. In this section, we provide detailed training and inference costs of AllSpark for reference. All experiments are conducted on two NVIDIA A6000 48G GPUs, using a numerical precision of torch.float32. The hyperparameter settings are detailed in Section IV-A.

Due to the presence of modality-specific encoders, the activated parameters in AllSpark vary when processing different modalities. Table XV summarizes the total parameters, trainable parameters, training time, and inference computational cost (measured in MACs) for each modality.

V. DISCUSSION

A. Limitations

Certainly, our work has several limitations, which will guide our future research directions.

1) *Lack of Interaction Between Different Modalities:* AllSpark only facilitates interactions between the language and other modalities, without involving more interactions, such as RGB and point cloud, or hyperspectral and multispectral. This is primarily due to constraints such as the lack of multimodal paired data. However, collecting large-scale paired data for ten modalities is nearly impossible, so we attempt to use language as the alignment reference for each modality, achieving indirect alignment between modalities using unpaired data. AllSpark represents the initial effort in this approach and has demonstrated strong adaptability across various modalities. Also, the adversarial examples also affect the robustness of the proposed model [120]. In the future, we plan to explore and expand our efforts in these directions in the further.
2) *Initial Work:* Our current work is still in its initial exploratory phase, and we have not carefully refined AllSpark's adaptability and performance on each modality. As a result, the model exhibits suboptimal performance on certain modalities, such as oblique photography, and video. The experimental results can be found in Tables XVII and XVIII. In the future, we plan to conduct more refined and targeted adjustments for each modality to enhance overall performance.
3) *Expensive Cost:* Multimodal LLMs, due to prolonged pretraining, typically possess universal reasoning capabilities in certain modalities. We generalize their applicability to other modalities by utilizing modality bridges to project other modalities onto the language modality. As shown in Section IV-E, although we freeze most of the parameters, fine-tuning even once on 2 A6000 GPUs often requires more than a day. Given the increasing training costs for large models, exploring methods to generalize their universal reasoning abilities is one of our future research directions.
4) *Interesting Phenomenon:* During our exploratory experiments, we discovered several interesting properties of large models. For instance, spatial information tends to degrade in large models, leading to collapse when performing dense prediction tasks such as segmentation and detection. Additionally, these models struggle to optimize when transferred to a small quantity of downstream data. Large models often require smaller hyperparameters and are sensitive to them. These observations might partially reveal the working mechanisms of large models, and we plan to conduct additional in-depth investigations into these phenomena in the future.

B. Potential of the LaRF

Inspired by the human cognitive system and linguistic philosophy, we propose the "LaRF" as the first principle for constructing our unified multimodal model. Its foreseeable potential includes at least the following three points.

1) *Efficient Generalization of Large Models:* Currently, the computational power and data scale required for training large models are rapidly expanding, and even the cost

TABLE XVI
LIST OF PROMPTS

Dimension	Modal	Prompts
1D	Language	1.Please determine if this movie review is positive or negative?
	Code	-
	Table	1.Please utilize the provided air quality indicators to accurately predict the concentrations of PM2.5 in the atmosphere 2.Given the air quality indicators, please provide a prediction for the PM2.5 levels at this particular moment
2D	RGB	1.This remote sensing image belongs to which of the following categories: [Category list of the NWPU dataset] 2.Describe this remote sensing image briefly3.Find a word that is most relevant to this remote sensing image 3.Find a word that is most relevant to this remote sensing image 4.Describe the key elements in this remote sensing image
	MSI	1.Based on the multi-spectral imagery feature description, please classify this object 2.Given the following multi-spectral imagery characteristics, please output the most fitting scene label
	HSI	1.Given the spectral information, can you help determine which class this pixel belongs to? 2.Here is the spectral data for a pixel. Considering the typical characteristics of land cover classes, could you provide a detailed analysis and suggest the most likely class for this pixel? 3.The spectral information for a pixel is given, but the data is noisy. Given the potential variability, which land cover classes should be considered as possible candidates for this pixel?
	Trajectory	1.Based on their past positions and movements in a crowded environment, predict the future trajectory of a selected pedestrian 2.Using the pedestrian trajectory data, along with additional information about the surrounding environment, predict the future path of the pedestrian 3.Given the current and past positions of a pedestrian and their neighboring pedestrians, predict the main pedestrian's trajectory
	SAR	1.Based on the SAR imagery feature description, please classify this object 2.Given the following SAR imagery characteristics, please output the most fitting scene label
	Graph	1.Given the current traffic data including vehicle flow rate, average speed, and time of day from the METRLA dataset, predict the traffic flow 2.Analyze the historical data on vehicle speeds and flow rates from the METRLA dataset for the past week. Identify any patterns or trends and predict the traffic conditions
3D	Point Cloud	1.Classify the provided point cloud sample into the correct category 2.Look at the point cloud data characteristics and classify the object 3.Please analyze the given point cloud dataset and determine which category it belongs to. Focus on the shape and structure evident in the point cloud

of fine-tuning large models is becoming prohibitive. Therefore, in the future, training large models for every domain will be almost impossible. Language, however, holds the potential for achieving efficient generalization of large models, expanding them from their native domains to additional domains at minimal cost. With our proposed AllSpark, we designed simple text prompts and task heads for each modality, demonstrating significant potential for multimodal expansion. In theory, AllSpark, built on the LaRF principle, can be extended to arbitrary modalities. In the future, we will conduct more in-depth research on the impact of text prompts and lightweight parameter modules on the generalization of large models.

2) *Interpretable Reasoning:* Deep learning models have often been referred to as “black boxes,” indicating that the reasoning process of these models is invisible and challenging to interpret. Research on the interpretability of deep learning models often relies on complex mathematical models and numerous assumptions, greatly limiting the practical application of deep learning methods in fields such as clinical medicine, military, and national resources where low fault tolerance or high confidentiality is crucial. However, language, as a tool for human thought and communication, provides models based on LaRF with the potential to use natural language directly for outputting reasoning chains and justifications.
3) *Transition From an End-to-End to an Interactive Paradigm:* The end-to-end paradigm refers to the learning approach where the model takes input and directly outputs results. In recent years, the end-to-end paradigm has become increasingly popular due to its simple and clear architecture and excellent performance. However, this approach also has clear disadvantages, such as uncontrollable internal operations, the need to optimize the whole for certain problems, and difficulty in pinpointing the cause of issues. A LaRF-based architecture has the potential to achieve an interactive paradigm in which users input raw data and corresponding text prompts and the model automatically performs relevant operations based on the prompts. Users can even iteratively adjust the text prompts based on the results. Therefore, in terms of both performance and controllability, the interactive paradigm has advantages that are incomparable to those of the end-to-end paradigm.

VI. Conclusion

Leveraging multimodal data is an inherent requirement for intelligent models to achieve geographic object cognition. Inspired by human cognitive systems and linguistic philosophy, we propose that the construction of multimodal models follows the fundamental principle of LaRF. Guided by this principle, we use language to balance the cohesion and autonomy of modalities, presenting a unified intelligent model, AllSpark, encompassing ten spatiotemporal modalities. The experimental results demonstrated that AllSpark exhibited excellent adaptability and application potential across various spatiotemporal modalities, highlighting the feasibility and potential of constructing multimodal models with LaRF. AllSpark remains an initial exploratory work, and in the future, we aim to delve deeper into the mechanisms guided by natural language, the efficient generalization of large models, and the transition to an interactive paradigm.

Appendix

A. List of Prompts

The text prompts we used are listed in Table XVI.

TABLE XVII
Video Classification With AllSpark

Method	Acc(%)
OPN	59.6
VCOP	72.4
SpeedNet	81.1
VTHCL	82.1
CVRL	94.4
VideoMAE v1	96.1
VideoMAE v2	99.6
AllSpark	27.5

TABLE XVIII
3-D Reconstruction With AllSpark

Method	PAG_6	PAG_{10}
MVSNet	81.15	91.44
CasMVSNet	95.45	98.02
Ada-MVS	96.14	98.1
UCSNet	96.25	98.45
AllSpark	6.4	10.4

B. Video

For the video modality, we evaluated AllSpark’s performance on action recognition tasks using the UCF101 [121] dataset. The UCF101 dataset is a human action recognition dataset comprising 101 action classes with a total of 13 320 video clips. The videos have a combined duration of 27 h and a resolution of 320*240 pixels and were sourced from YouTube.

The model is tasked with understanding videos and accurately classifying them into one of the 101 classes, with the evaluation metric being the top-1 accuracy. In Table XVII, we compare AllSpark with the current state-of-the-art models.

Currently, AllSpark’s adaptation to the video modality is not optimal, as it shows a significant difference from the baseline model results. We attribute this to two main reasons.

1) The high redundancy in video information increases the training cost for AllSpark. We trained it for only 3 epochs on the dataset.
2) AllSpark’s model architecture lacks flexibility for 3-D data, making it less effective at capturing temporal information.

C. Oblique Photography

For the oblique photography modality, we tested AllSpark’s performance on the 3-D reconstruction task using the WHU-OMVS [122] dataset. WHU-OMVS is an oblique photography dataset designed for 3-D reconstruction tasks. The dataset provides imagery from five different viewpoints, along with camera parameters and other relevant information. It consists of six areas, and in our experiments, area 1 is used as the training set, and area 2 is used as the test set.

The model takes five-view images as input, and the goal is to output depth maps for reconstructing 3-D models. The evaluation metric used was the percentage of accurate grids in total (PAG) [122], calculated by the following formula:

$$\mathrm{PAG}_a = \left(\frac{m_a}{m}\right) * 100. \quad (7)$$

The suffixes in the PAG represent different accuracy standards, where PAG_6 signifies an error within 0.6 m and PAG_{10} indicates an error within 1 m. Table XVIII compares AllSpark with popular multiview 3-D reconstruction models on the WHU-OMVS dataset.

AllSpark falls short in terms of accuracy compared to modality-specific expert models. We speculate two possible reasons: 1) dense spatial information gradually diminishes in the deep structure of large models, a point verified in our exploratory experiments involving segmentation, detection, and so on and 2) the model architecture lacks flexibility and cannot connect gradual features like expert 3-D reconstruction models such as Ada-MVS. Additionally, these methods lack the ability to design specific model structures for processing, severely restricting their performance. Adapting multimodal large models to dense prediction tasks and optimizing the architecture are future research directions.

REFERENCES

[1] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao, "EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5917820.

[2] X. Guo et al., "Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 27672–27683.

[3] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, "PointLLM: Empowering large language models to understand point clouds," 2023, *arXiv:2308.16911*.

[4] L. Chen, C. Huang, X. Zheng, J. Lin, and X. Huang, "TableVLM: Multi-modal pre-training for table structure recognition," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics Long Papers*, Toronto, ON, Canada, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Jul. 2023, pp. 2437–2449. [Online]. Available: https://aclanthology.org/2023.acl-long.137

[5] J. Han et al., "OneLLM: One framework to align all modalities with language," 2023, *arXiv:2312.03700*.

[6] Y. Zhang et al., "Meta-transformer: A unified framework for multimodal learning," 2023, *arXiv:2307.10802*.

[7] S. He et al., "STGC-GNNs: A GNN-based traffic prediction framework with a spatial–temporal Granger causality graph," *Phys. A, Stat. Mech. Appl.*, vol. 623, May 2023, Art. no. 128913.

[8] Q. Luo, S. He, X. Han, Y. Wang, and H. Li, "LSTTN: A long-short term transformer-based spatiotemporal neural network for traffic flow forecasting," *Knowl.-Based Syst.*, vol. 293, Jun. 2024, Art. no. 111637.

[9] H. Li, J. Cao, J. Zhu, Q. Luo, S. He, and X. Wang, "Augmentation-free graph contrastive learning of invariant-discriminative representations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp. 11157–11167, Aug. 2024.

[10] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[11] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 1–11.

[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[13] H. Li, J. Cao, J. Zhu, Y. Liu, Q. Zhu, and G. Wu, "Curvature graph neural network," *Inf. Sci.*, vol. 592, pp. 50–66, May 2022.

[14] S. C. Kulkarni and P. P. Rege, "Pixel level fusion techniques for SAR and optical images: A review," *Inf. Fusion*, vol. 59, pp. 13–29, Jul. 2020.

[15] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[16] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[17] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5583–5594.

[18] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 23716–23736.

[19] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.

[20] C. Tao, J. Qi, G. Zhang, Q. Zhu, W. Lu, and H. Li, "TOV: The original vision model for optical remote sensing image understanding via self-supervised learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4916–4930, 2023.

[21] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4651–4664.

[22] R. Shao, Z. Zhang, C. Tao, Y. Zhang, C. Peng, and H. Li, "Homogeneous tokenizer matters: Homogeneous visual tokenizer for remote sensing image understanding," *ISPRS J. Photogramm. Remote Sens.*, vol. 18, pp. 294–310, Dec. 2024.

[23] Z. Feng et al., "CodeBERT: A pre-trained model for programming and natural languages," 2020, *arXiv:2002.08155*.

[24] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, vol. 1, 2019, p. 2.

[25] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[27] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[28] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[29] S. O. Arík and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 6679–6687.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[31] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sens. Environ.*, vol. 214, pp. 73–86, Sep. 2018.

[33] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[34] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, Aug. 2016.

[35] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.

[36] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[37] W. Wu, Z. Qi, and L. Fuxin, "PointConv: Deep convolutional networks on 3D point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9621–9630.

[38] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1349–1358.

[39] L. H. Hughes, D. Marcos, S. Lobry, D. Tuia, and M. Schmitt, "A deep learning framework for matching of SAR and optical imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 166–179, Nov. 2020.

[40] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 14–34, Sep. 2021.

[41] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network," *Remote Sens.*, vol. 10, no. 5, p. 800, 2018.

[42] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.

[43] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, "Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501212.

[44] Y. Gao et al., "Hyperspectral and multispectral classification for coastal wetland using depthwise feature interaction network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5512615.

[45] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS J. Photogramm. Remote Sens.*, vol. 178, pp. 68–80, Aug. 2021.

[46] Z. Zhang, Z. Ren, C. Tao, Y. Zhang, C. Peng, and H. Li, "GraSS: Contrastive learning with gradient-guided sampling strategy for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5626814.

[47] J. Peng, D. Ye, B. Tang, Y. Lei, Y. Liu, and H. Li, "Lifelong learning with cycle memory networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 1–14, Jul. 2023.

[48] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-BERT: Pre-training 3D point cloud transformers with masked point modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19313–19322.

[49] Y. Zeng et al., "What matters in training a GPT4-style language model with multimodal inputs?" 2023, *arXiv:2307.02469*.

[50] I. Padhi et al., "Tabular transformers for modeling multivariate time series," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 3565–3569.

[51] Y. Fang et al., "EVA: Exploring the limits of masked visual representation learning at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19358–19369.

[52] L. Shi, L. Wang, S. Zhou, and G. Hua, "Trajectory unified transformer for pedestrian trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 9675–9684.

[53] H. Liu et al., "Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage.*, 2023, pp. 4125–4129.

[54] J. Zhu et al., "KST-GCN: A knowledge-driven spatial–temporal graph convolutional network for traffic forecasting," *IEEE Trans. Intell. Transport. Syst.*, vol. 23, no. 9, pp. 15055–15065, Jan. 2022.

[55] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[56] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[57] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 142–150.

[58] H. Husain, H.-H. Wu, T. Gazit, M. Allamanis, and M. Brockschmidt, "CodeSearchNet challenge: Evaluating the state of semantic code search," 2019, *arXiv:1909.09436*.

[59] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[60] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[61] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[62] L. Li, J. Han, X. Yao, G. Cheng, and L. Guo, "DLA-MatchNet for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7844–7853, Sep. 2020.

[63] G. Cheng et al., "SPNet: Siamese-prototype network for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5608011.

[64] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," 2017, *arXiv:1711.04043*.

[65] S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen, "Cautionary tales on air-quality improvement in Beijing," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 473, no. 2205, Sep. 2017, Art. no. 20170457.

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

[67] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[68] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 261–268.

[69] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Comput. Graph. Forum*, vol. 26, no. 3, pp. 655–664, Sep. 2007.

[70] E. Keydel, S. Lee, and J. Moore, "MSTAR extended operating conditions: A tutorial," *Proc. SPIE*, vol. 2757, pp. 228–242, Jun. 1996.

[71] Z. Wu et al., "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912–1920.

[72] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2010, pp. 270–279.

[73] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 173–176, Jan. 2011.

[74] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. MaÎtre, "Structural high-resolution satellite image indexing," in *Proc. 7th ISPRS TC Symp.*, Vienna, Austria, 2010, pp. 298–303.

[75] A. X. Chang et al., "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.

[76] M. A. Uy, Q.-H. Pham, B.-S. Hua, D. T. Nguyen, and S.-K. Yeung, "Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1588–1597.

[77] C. Ye, H. Zhu, B. Zhang, and T. Chen, "A closer look at few-shot 3D point cloud classification," *Int. J. Comput. Vis.*, vol. 131, no. 3, pp. 772–795, Mar. 2023.

[78] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using CNN-capsNet," *Remote Sens.*, vol. 11, no. 5, p. 494, 2019.

[79] K. Xu, H. Huang, P. Deng, and Y. Li, "Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5751–5765, Oct. 2021.

[80] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[81] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1461–1474, May 2019.

[82] O. Manas, A. Lacoste, X. Giro-i-Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9414–9423.

[83] S. Wang, Y. Guan, and L. Shao, "Multi-granularity canonical appearance pooling for remote sensing scene classification," *IEEE Trans. Image Process.*, vol. 29, pp. 5396–5407, 2020.

[84] Q. Bi, K. Qin, H. Zhang, and G.-S. Xia, "Local semantic enhanced ConvNet for aerial scene recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 6498–6511, 2021.

[85] G. Zhang et al., "A multiscale attention network for remote sensing scene images classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9530–9545, 2021.

[86] S. Wang, Y. Ren, G. Parr, Y. Guan, and L. Shao, "Invariant deep compressible covariance pooling for aerial scene categorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6549–6561, Aug. 2020.

[87] S.-B. Chen, Q.-S. Wei, W.-Z. Wang, J. Tang, B. Luo, and Z.-Y. Wang, "Remote sensing scene classification via multi-branch local attention network," *IEEE Trans. Image Process.*, vol. 31, pp. 99–109, 2021.

[88] B. Li, Y. Guo, J. Yang, L. Wang, Y. Wang, and W. An, "Gated recurrent multiattention network for VHR remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5606113.

[89] Y. Zhao, J. Liu, J. Yang, and Z. Wu, "EMSCNet: Efficient multisample contrastive network for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605814.

[90] D. Wang et al., "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2022, Art. no. 5607315.

[91] T. Nampally, J. Wu, and S. Dev, "Performance comparison of multi-spectral channels for land use classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2023, pp. 6178–6181.

[92] K. Zhang and H. Yang, "Semi-supervised multi-spectral land cover classification with multi-attention and adaptive kernel," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1881–1885.

[93] P. Gómez and G. Meoni, "MSMatch: Semisupervised multispectral scene classification with few labels," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11643–11654, 2021.

[94] P. Goyal et al., "Vision models are more robust and fair when pretrained on uncurated images without supervision," 2022, *arXiv:2202.08360*.

[95] X. Wanyan, S. Seneviratne, S. Shen, and M. Kirley, "Dino-mc: Self-supervised contrastive learning for remote sensing imagery with multi-sized local crops," 2023, *arXiv:2303.06670*.

[96] D. Uchaev and D. Uchaev, "Small sample hyperspectral image classification based on the random patches network and recursive filtering," *Sensors*, vol. 23, no. 5, p. 2499, Feb. 2023.

[97] H. Xie, L. Zhang, and C. P. Lim, "Evolving CNN-LSTM models for time series prediction using enhanced grey wolf optimizer," *IEEE Access*, vol. 8, pp. 161519–161541, 2020.

[98] X. Cheng, W. Zhang, A. Wenzel, and J. Chen, "Stacked ResNet-LSTM and CORAL model for multi-site air quality prediction," *Neural Comput. Appl.*, vol. 34, no. 16, pp. 13849–13866, Aug. 2022.

[99] D. Li, J. Liu, and Y. Zhao, "Prediction of multi-site $PM_{2.5}$ concentrations in Beijing using CNN-bi LSTM with CBAM," *Atmosphere*, vol. 13, no. 10, p. 1719, Oct. 2022.

[100] D. Guo et al., "GraphCodeBERT: Pre-training code representations with data flow," 2020, *arXiv:2009.08366*.

[101] R. Klokov and V. Lempitsky, "Escape from cells: Deep Kd-networks for the recognition of 3D point cloud models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 863–872.

[102] H. Lei, N. Akhtar, and A. Mian, "Spherical kernel for efficient graph convolution on 3D point clouds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3664–3680, Oct. 2020.

[103] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[104] J. Li, B. M. Chen, and G. H. Lee, "SO-Net: Self-organizing network for point cloud analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9397–9406.

[105] R. Li et al., "PointVGG: Graph convolutional network with progressive aggregating features on point clouds," *Neurocomputing*, vol. 429, pp. 187–198, Mar. 2021.

[106] G. Chen, M. Wang, Y. Yang, K. Yu, L. Yuan, and Y. Yue, "PointGPT: Auto-regressively generative pre-training from point clouds," 2023, *arXiv:2305.11487*.

[107] U. Srinivas, V. Monga, and R. G. Raj, "SAR automatic target recognition using discriminative graphical models," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 1, pp. 591–606, Jan. 2014.

[108] G. Dong, N. Wang, and G. Kuang, "Sparse representation of monogenic signal: With application to target recognition in SAR images," *IEEE Signal Process. Lett.*, vol. 21, no. 8, pp. 952–956, Aug. 2014.

[109] G. Dong and G. Kuang, "Classification on the monogenic scale space: Application to target recognition in SAR image," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2527–2539, Aug. 2015.

[110] J. A. O'Sullivan, M. D. DeVore, V. Kedia, and M. I. Miller, "SAR ATR performance using a conditionally Gaussian model," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 37, no. 1, pp. 91–108, Jan. 2001.

[111] J.-I. Park and K.-T. Kim, "Modified polar mapping classifier for SAR automatic target recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 2, pp. 1092–1107, Apr. 2014.

[112] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[113] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*.

[114] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6256–6268.

[115] D. S. Sachan, M. Zaheer, and R. Salakhutdinov, "Revisiting LSTM networks for semi-supervised text classification via mixed objective function," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6940–6948.

[116] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," 2016, *arXiv:1605.07725*.

[117] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[118] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 5753–5763.

[119] S. Ding et al., "ERNIE-doc: A retrospective long-document modeling transformer," 2020, *arXiv:2012.15688*.

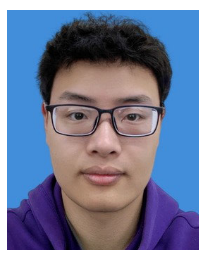
[120] J. Cui, W. Guo, H. Huang, X. Lv, H. Cao, and H. Li, "Adversarial examples for vehicle detection with projection transformation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5632418.

[121] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[122] J. Liu, J. Gao, S. Ji, C. Zeng, S. Zhang, and J. Gong, "Deep learning based multi-view stereo matching and 3D scene reconstruction from oblique aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 204, pp. 42–60, Oct. 2023.

Run Shao received the B.S. degree in geographic information science from Central South University, Changsha, China, in 2023, where he is currently pursuing the Ph.D. degree with the GeoXLab.

His research interests include computer vision, multimodal large language model, foundation model, and remote sensing image understanding.



Cheng Yang received the B.S. degree from Central South University, Changsha, China, in 2023, where he is currently pursuing the M.S. degree.

His research interests lie in visual language models and the application of compositional reasoning.

Qiujun Li is currently pursuing the master's degree with Central South University, Changsha, China.

Her research interests include computer vision and multimodal model interaction.

Linrui Xu is currently pursuing the master's degree with Central South University, Changsha, China.

His research interests include multimodal learning and deep learning.

Xiang Yang received the B.S. degree in remote sensing from Central South University, Changsha, China, in 2024, where he is currently pursuing the Ph.D. degree majoring in computer vision and deep learning.

Xian Li received the master's degree in mechanical engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Geosciences and Info-Physics, Central South University, Changsha, China.

From 2020 to 2023, he worked as an Image Processing and Computer Vision Engineer at China Railway Construction Heavy Industry Group Company Ltd., Changsha. His research interests include machine/deep learning, image processing, computer vision, and multimodal spatiotemporal general intelligence.

Mengyao Li received the B.S. degree in geographic information science from Central South University, Changsha, China, in 2023, where she is currently pursuing the master's degree.

Her research interests include causal inference and machine learning.

Qing Zhu received the Ph.D. degree in railway engineering from Northern Jiaotong University, Beijing, China, in 1995.

He was working as a Postdoctoral Fellow with Wuhan University, Wuhan, China, during 1995 to 1997. He was a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, during 1998 to 2013. He is currently a Professor with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China. His research interests include digital elevation model, 3D GIS, virtual geographic environment, and smart city.

Yongjun Zhang (Member, IEEE) received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively.

From 2014 to 2015, he was a Senior Visiting Fellow with the Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada. From 2015 to 2018, he was a Senior Scientist at the Environmental Systems Research Institute, Inc. (Esri), Redlands, CA, USA. He is currently the Dean of the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 150 research articles and one book. He holds 23 Chinese patents and 26 copyright-registered computer software. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource datasets, artificial intelligence-driven remote sensing image interpretation, integration of light detection and ranging (LiDAR) point clouds and images, and 3-D city reconstruction.

Dr. Zhang was a Key Member of ISPRS Workgroup II/I from 2016 to 2020. He is the PI Winner of the Second-Class National Science and Technology Progress Award in 2017 and the PI Winner of the Outstanding-Class Science and Technology Progress Award in Surveying and Mapping (Chinese Society of Surveying, Mapping and Geoinformation, China) in 2015. In recent years, he has also served as the session chair for above 20 international workshops or conferences. He has been frequently serving as a referee for over 20 international journals. He is the Coeditor-in-Chief of *The Photogrammetric Record*.

Yansheng Li (Senior Member, IEEE) received the B.S. degree in information and computing science from Shandong University, Weihai, China, in 2010, and the Ph.D. degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology, Wuhan, China, in 2015.

From 2017 to 2018, he was a Visiting Assistant Professor with the Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. He is currently a Full Professor with the School of Remote Sensing and Information Engineering, Wuhan University (WHU), Wuhan. He has authored more than 100 peer-reviewed journal articles and conference papers. His research interests include knowledge graph, deep learning, and their applications in remote sensing big data mining.

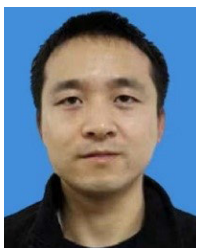
Dr. Li received the Young Surveying and Mapping Science and Technology Innovation Talent Award of the Chinese Society for Geodesy, Photogrammetry and Cartography in 2022. He received the recognition of the Best Reviewers of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) in 2021 and the Best Reviewers of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2022. He is an Associate Editor of IEEE TGRS, a Junior Editorial Member of *The Photogrammetric Record*, and a Lead Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING and *Remote Sensing*.

Yu Liu received the B.S., M.S., and Ph.D. degrees from Peking University, Beijing, China, in 1994, 1997, and 2003, respectively.

He is currently a Professor with the Institute of Remote Sensing and Geographic Information System, Peking University. His research interests mainly concentrate in humanities and social science based on big geo-data.

Yong Tang received the B.S. degree in communication engineering from Central South University, Changsha, China, in 2008.

He is currently the Coo of Huawei Hunan Kunpeng/Ascend Eco-Innovation Center, Changsha, and is responsible for AI technology enabling and computing power operation, joint development of AI application solutions, and the work of migrating, adapting, and tuning AI algorithms based on Ascend.

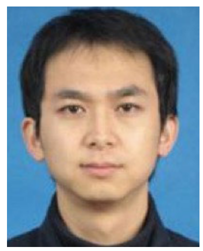


Dapeng Liu received the dual B.S. degree in computer science and technology as well as civil engineering surveying and mapping from Hunan University, Changsha, China, in 2012.

He is currently the Coo of Huawei Hunan Kunpeng/Ascend Eco-Innovation Center, Changsha, and is responsible for AI technology enabling and computing power operation, joint development of AI application solutions, and the work of migrating, adapting, and tuning AI algorithms based on Ascend. His research interests encompass artificial intelligence image recognition and software application development.

Shizhong Yang received the master's and Ph.D. degrees in prospecting engineering from China University of Geosciences, Wuhan, China, in 1994 and 1997, respectively.

He has been working on high-precision location services technology for nearly 30 years. He is currently the chairperson of Hunan BDS Micro-chipset Industrial Development Co.,Ltd., Vice Chairperson of China Geographic Information Industry Association, Vice chairperson of China Satellite Navigation Association, Vice chairperson of Hunan Satellite Application Association, and part-time vice chairperson of the 13th Committee of Changsha Association for Science and Technology. He presided over the compilation of China Aerospace Science and Technology Corporation civil navigation data format standards; Participate in the development and marketization of China's first domestic geographic information system MAPGIS, the first three-dimensional visual information system IMAGIS, the first full digital aerial photogrammetry system VIRTUOZO. He is an Associate professor.



Haifeng Li (Member, IEEE) received the master's degree in transportation engineering from South China University of Technology, Guangzhou, China, in 2005, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009.

He was a Research Associate with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, in 2011, and a Visiting Scholar with the University of Illinois at Urbana-Champaign, Urbana, IL, USA, from 2013 to 2014. He is currently a Professor at the School of Geosciences and Info-Physics, Central South University, Changsha, China. He has authored over 30 journal papers. He is a reviewer for many journals. His research interests include geo/remote sensing big data, machine/deep learning, and artificial/brain-inspired intelligence.