Multimodal Foundation Model Based Siamese Network for Change Detection in Remote Sensing Imagery

Ruiqian Zhang, Yan Qin, Xiaogang Ning and Zhang Hanchao (China, PR)

Speaker: Ruiqian Zhang

Date: May 22, 2024

- **Name** ： Ruiqian Zhang
- **Institution:** Chinese Academy of Surveying and Mapping
- **Degree:** PhD in Engineering
- **Research Interests:** Image processing, computer vision, remote sensing, deep learning
- **E-mail:** zhangrq@casm.ac.cn; zhangruiqian@whu.edu.cn

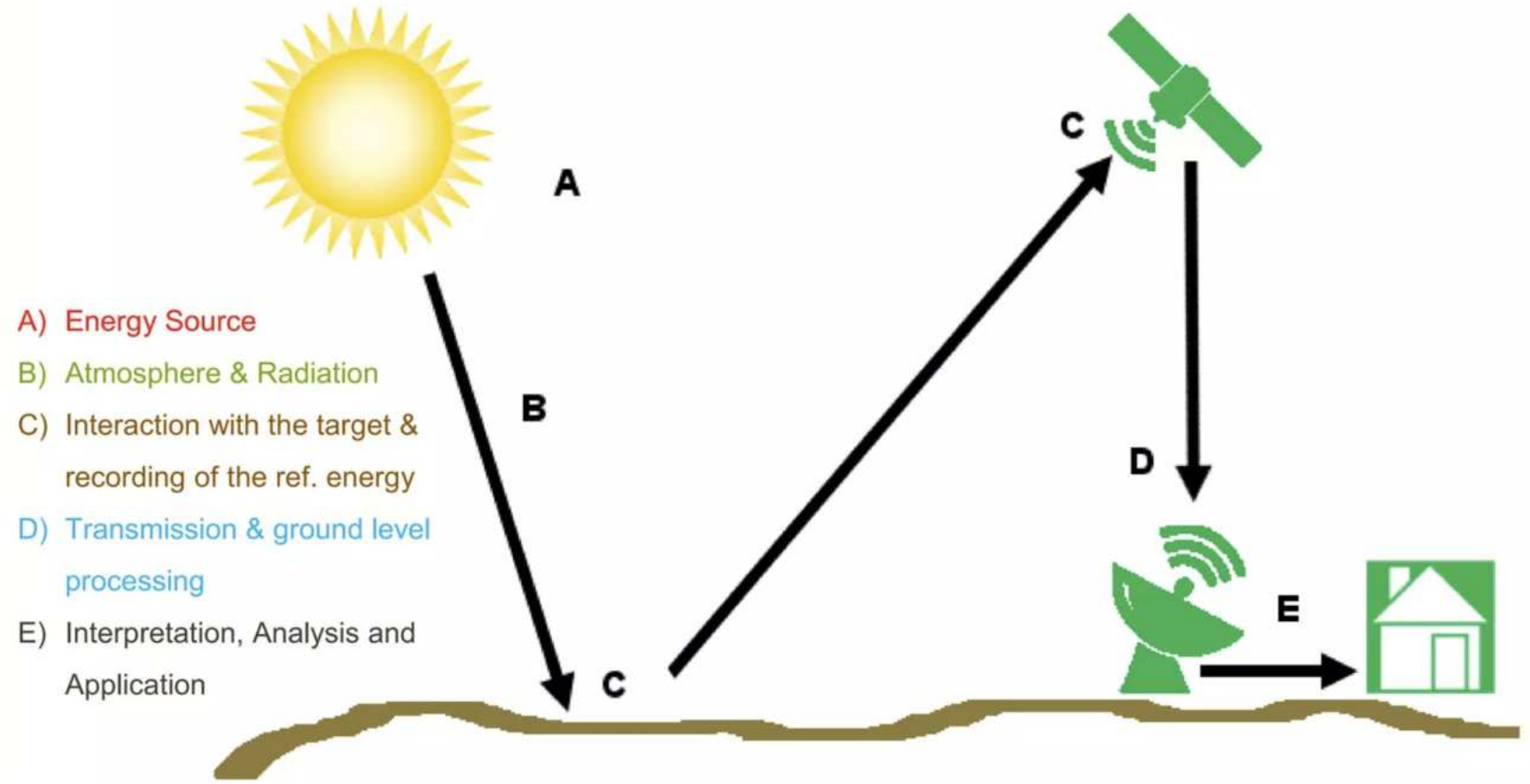
Homepage

Outline

- Introduction
- Our Proposed Approach
- Experimental Results
- Conclusions

Remote Sensing

A method of obtaining information about the properties of an object without coming into physical contact with it

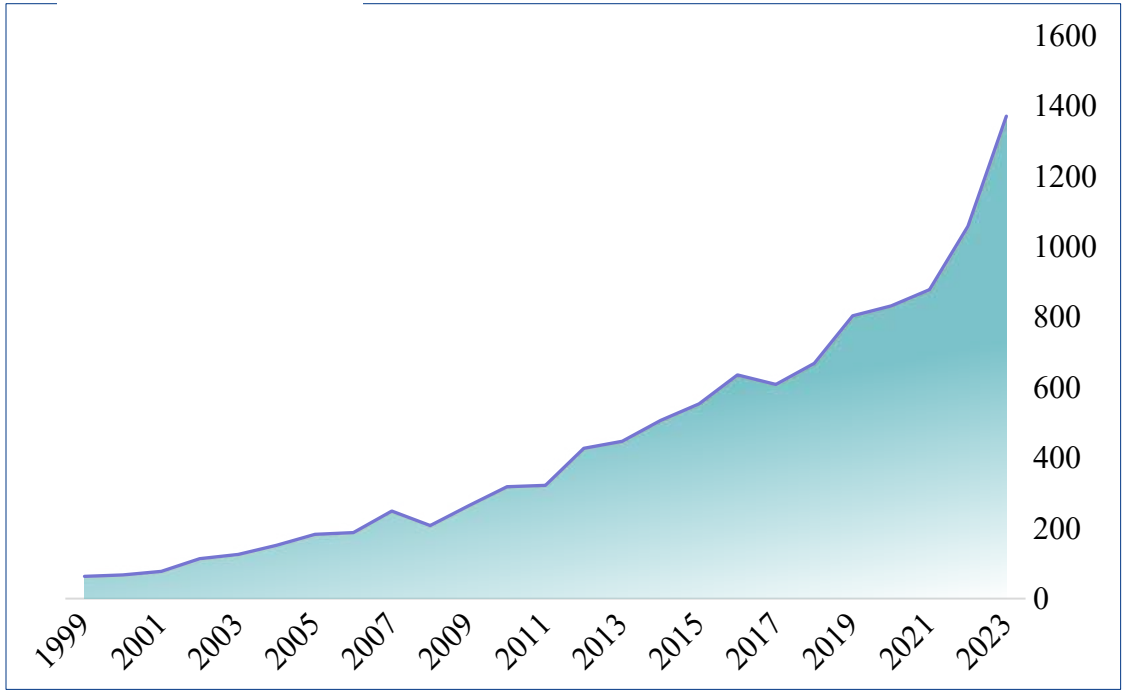


World

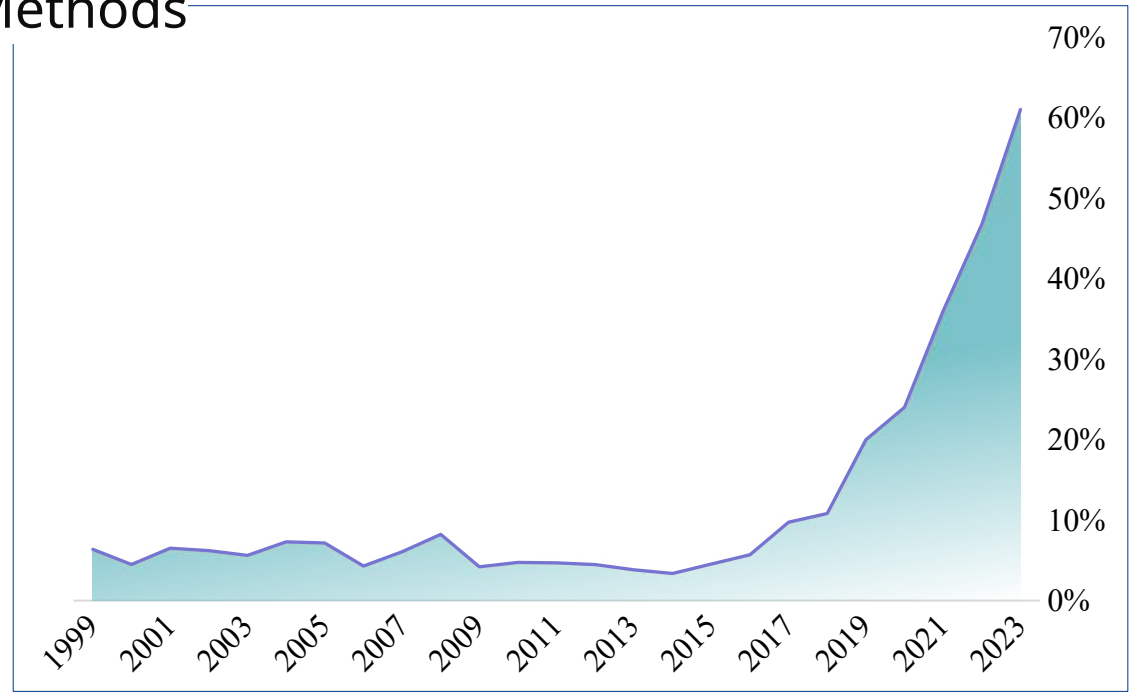
WorldVie

IKONOS-2, 0.82m

Number of Change Detection Publications



Proportion of Publications Using Deep Learning Methods



Sources : Web Of Science

However, change detection still faces many challenges. For instance, complex image backgrounds, seasonal variations, and difficulty detecting small changes

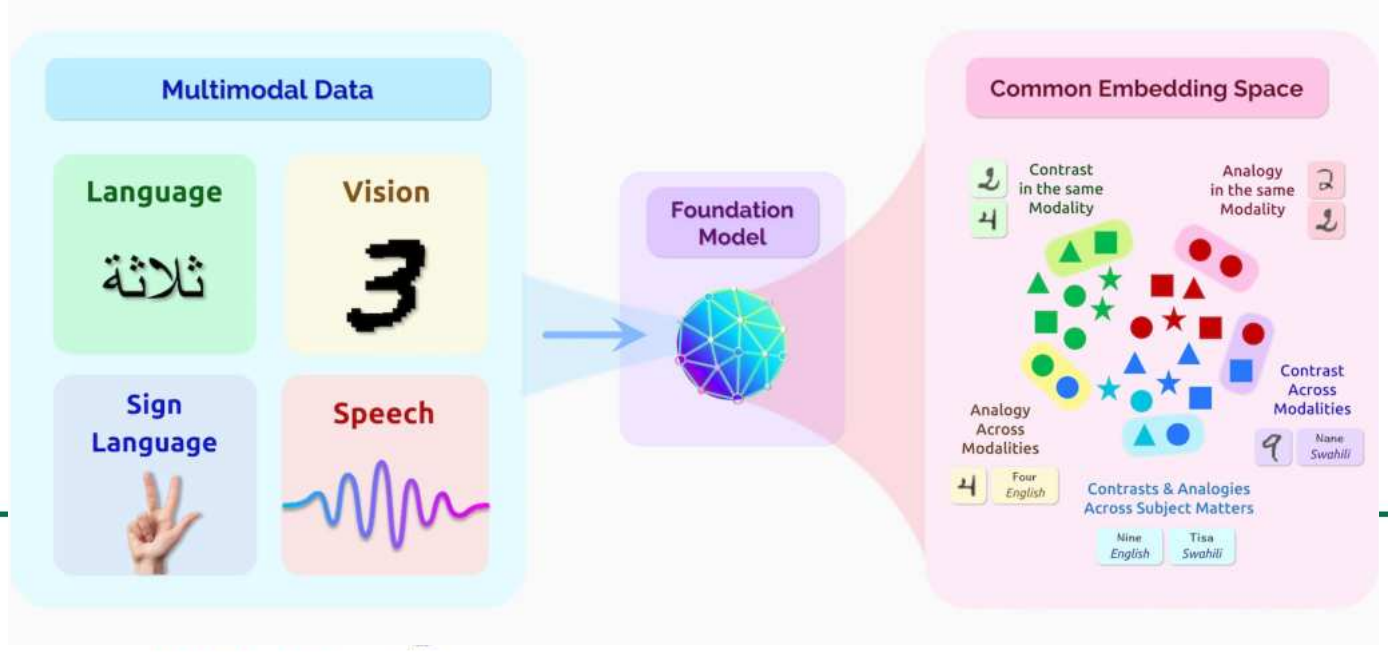
The gap between public sample datasets and real-world scenarios. This leads to poor feature generalization, making it hard to tackle these problems effectively.

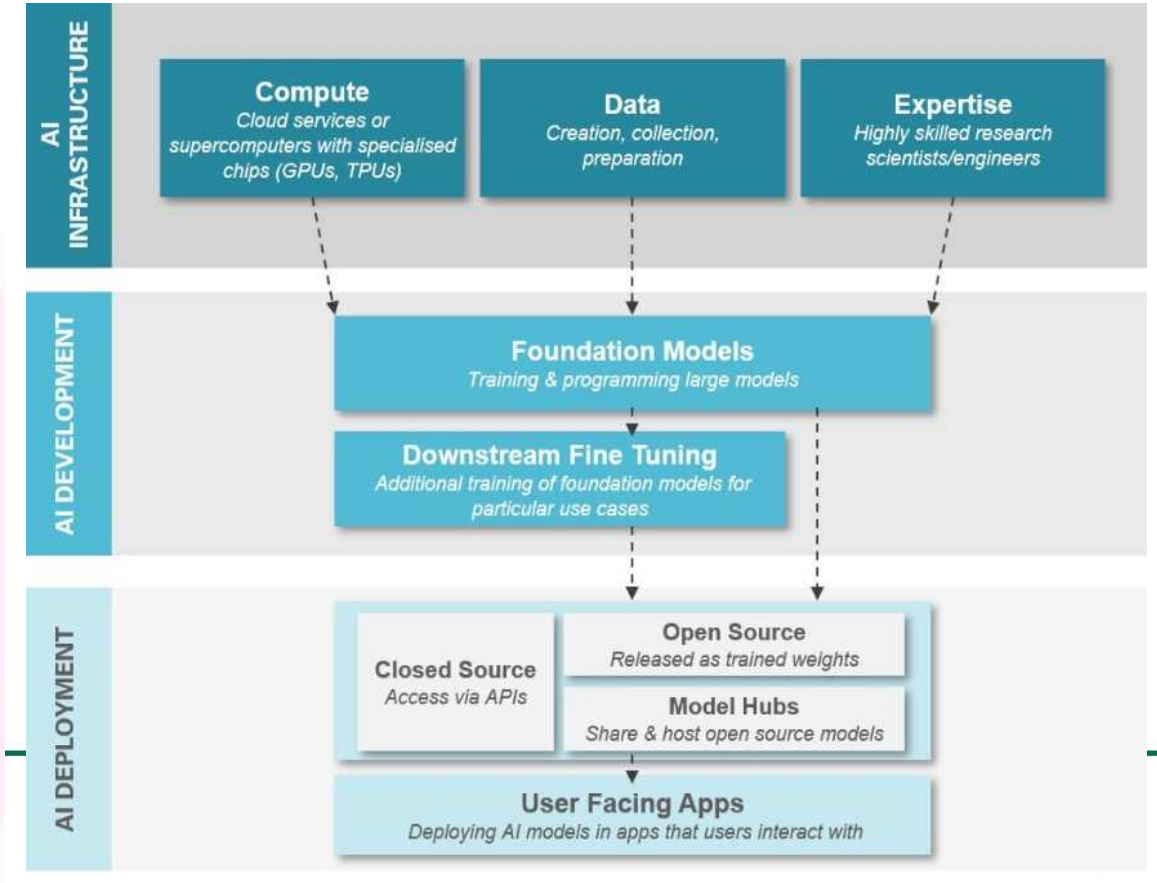
Outline

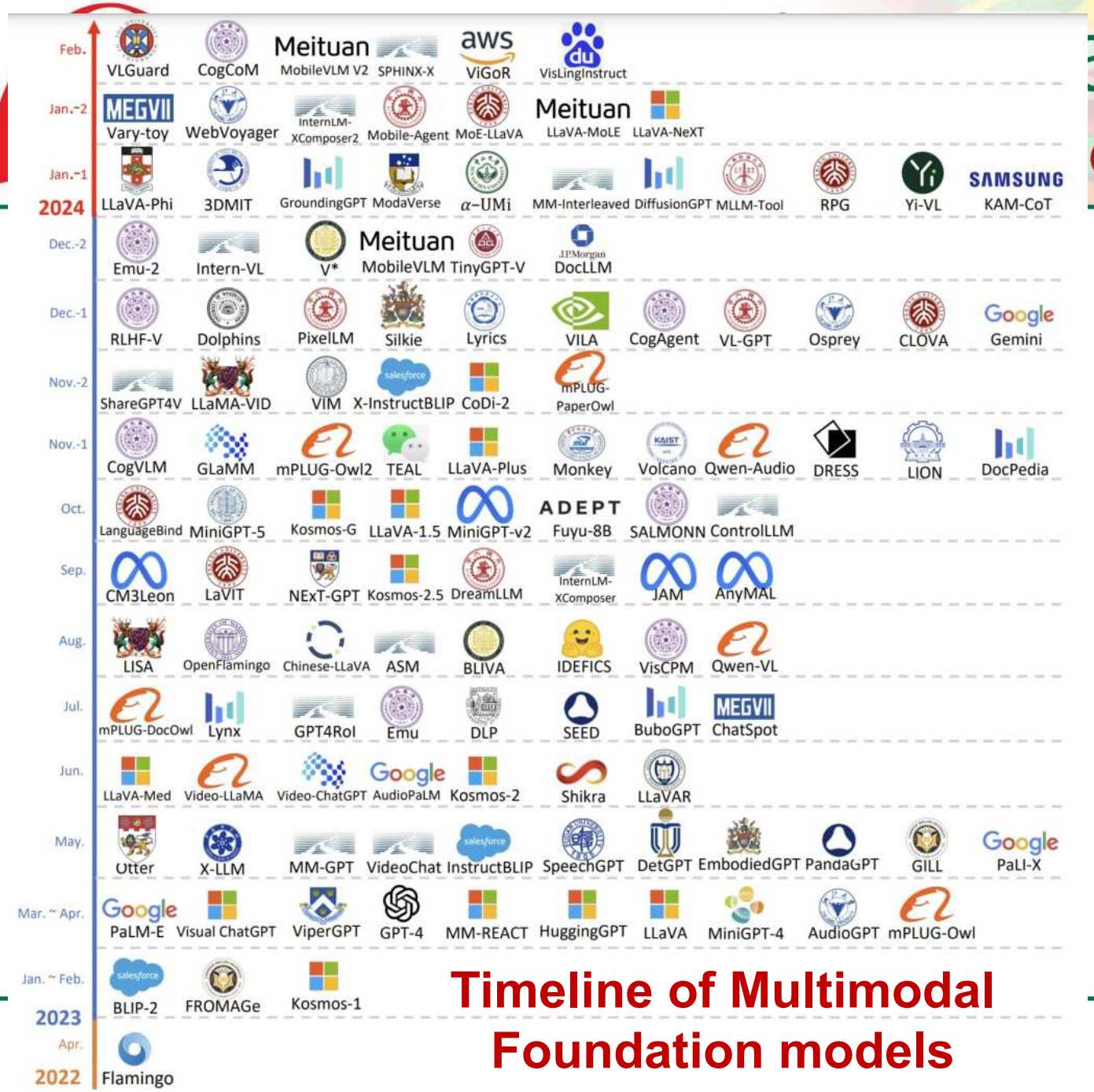
- ☐ Introduction
- ☐ Our Proposed Approach
- ☐ Experimental Results
- ☐ Conclusions

Recently, the emergence of the ChatGPT series of pre-trained foundational models has sparked a new wave in the AI revolution.

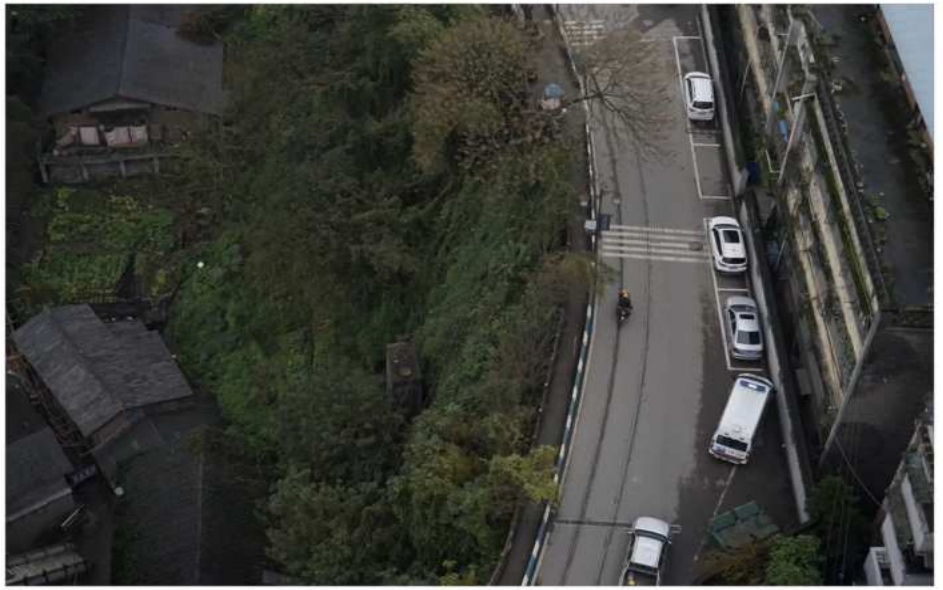
- Text
- Vision
- Multimodality





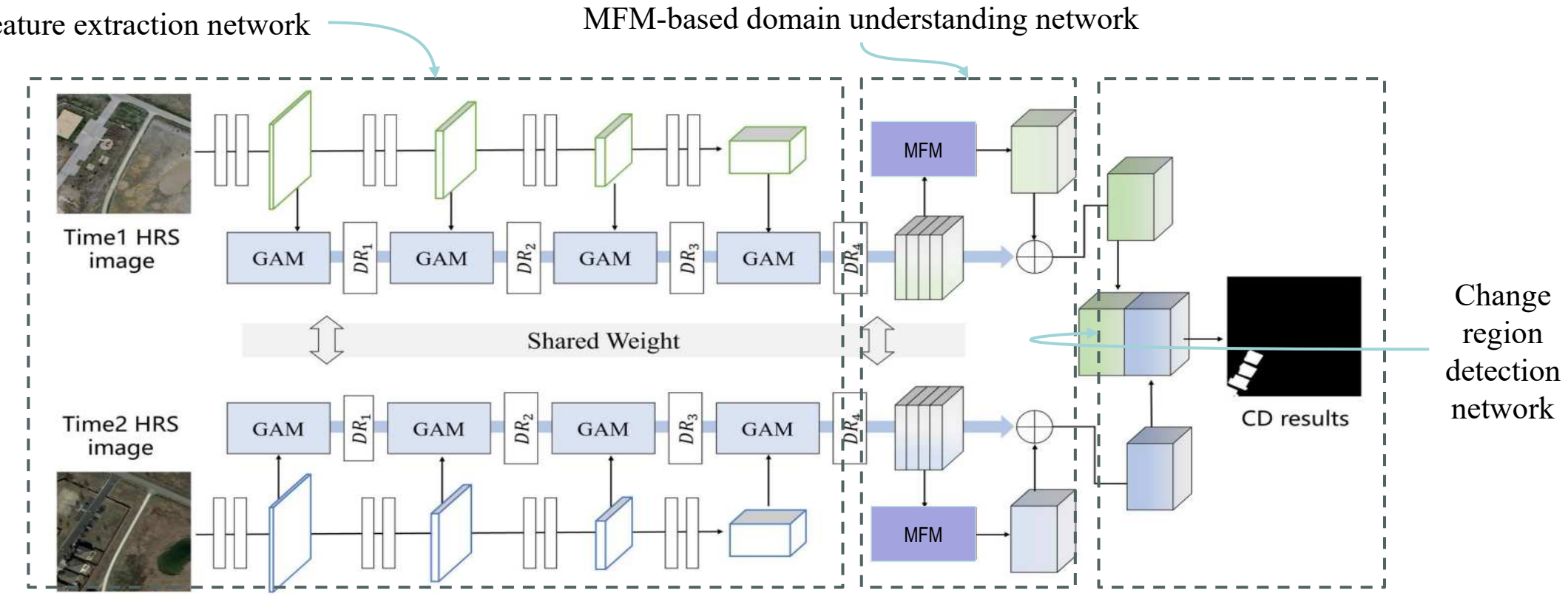


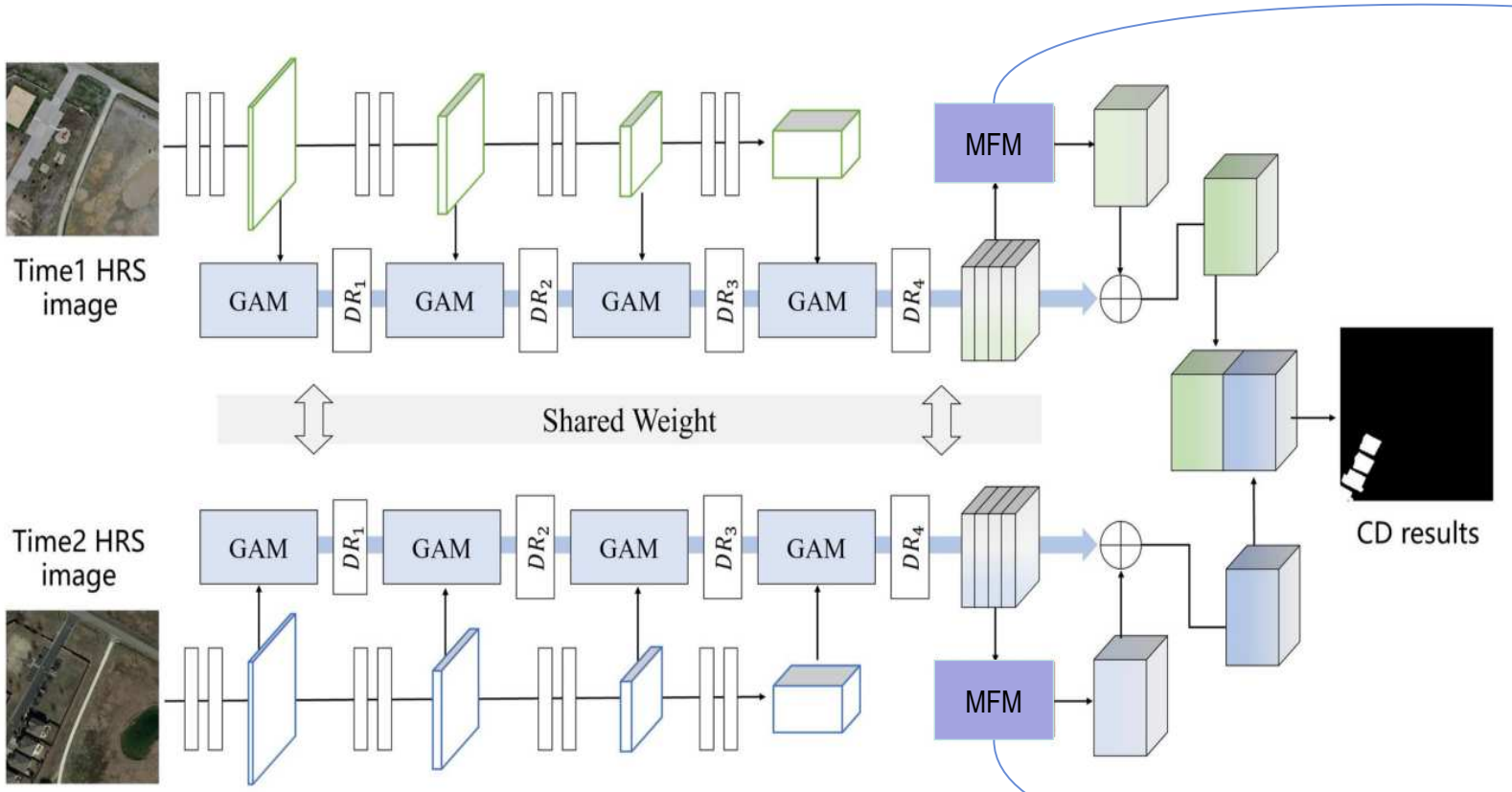
Timeline of Multimodal Foundation models

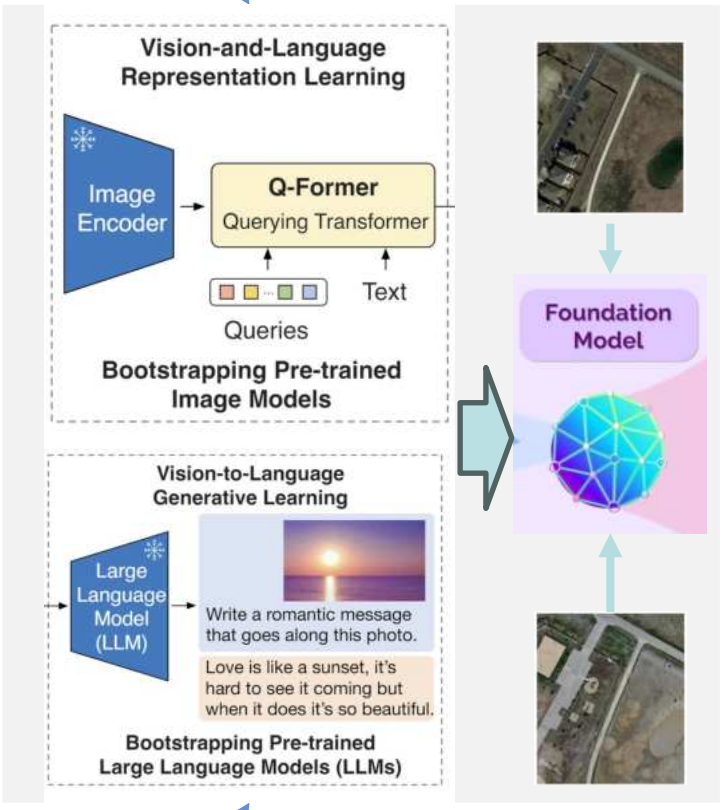


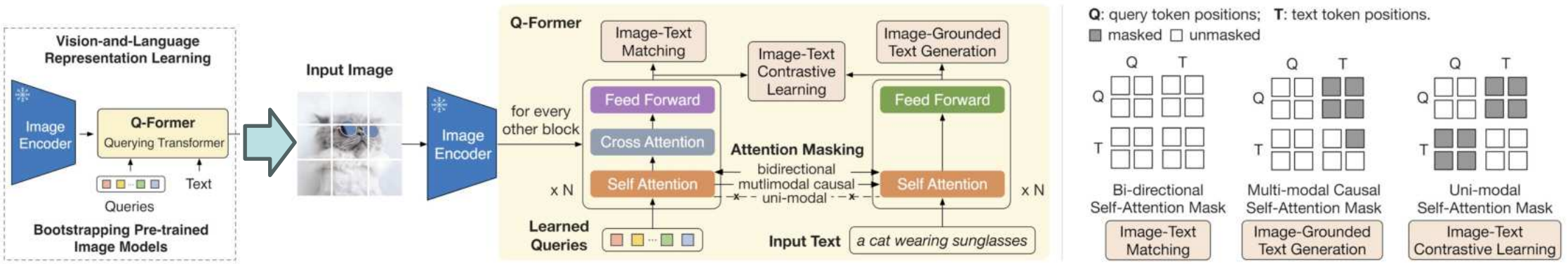
Panoramic Instance Segmentation of Drone Imagery by Segment Anything

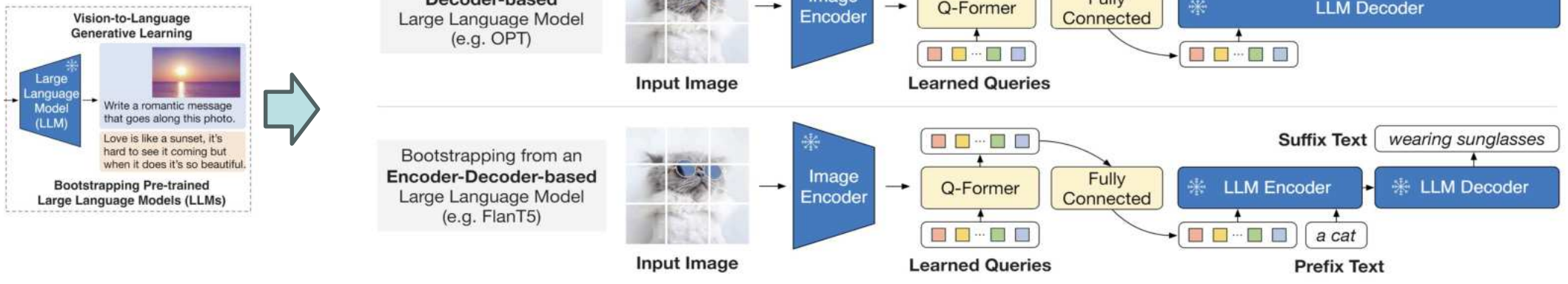
The proposed MFM Based Siamese Change Detection Network




Time1 HRS image
Time2 HRS image
GAM
DR_1
DR_2
DR_3
DR_4
Shared Weight
MFM
MFM
CD results


Vision-and-Language Representation Learning
Image Encoder
Q-Former
Querying Transformer
Text
Queries
Bootstrapping Pre-trained Image Models
Vision-to-Language Generative Learning
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained Large Language Models (LLMs)
Foundation Model


Vision-and-Language Representation Learning
Image Encoder
Q-Former
Querying Transformer
Text
Queries
Bootstrapping Pre-trained Image Models
Input Image
Image Encoder
Q-Former
Image-Text Matching
Image-Text Contrastive Learning
Image-Grounded Text Generation
for every other block
Feed Forward
Cross Attention
Self Attention
x N
Attention Masking
bidirectional
mutlimodal causal
uni-modal
Learned Queries
Input Text
a cat wearing sunglasses
Q: query token positions; T: text token positions.
masked unmasked
Bi-directional Self-Attention Mask
Image-Text Matching
Multi-modal Causal Self-Attention Mask
Image-Grounded Text Generation
Uni-modal Self-Attention Mask
Image-Text Contrastive Learning

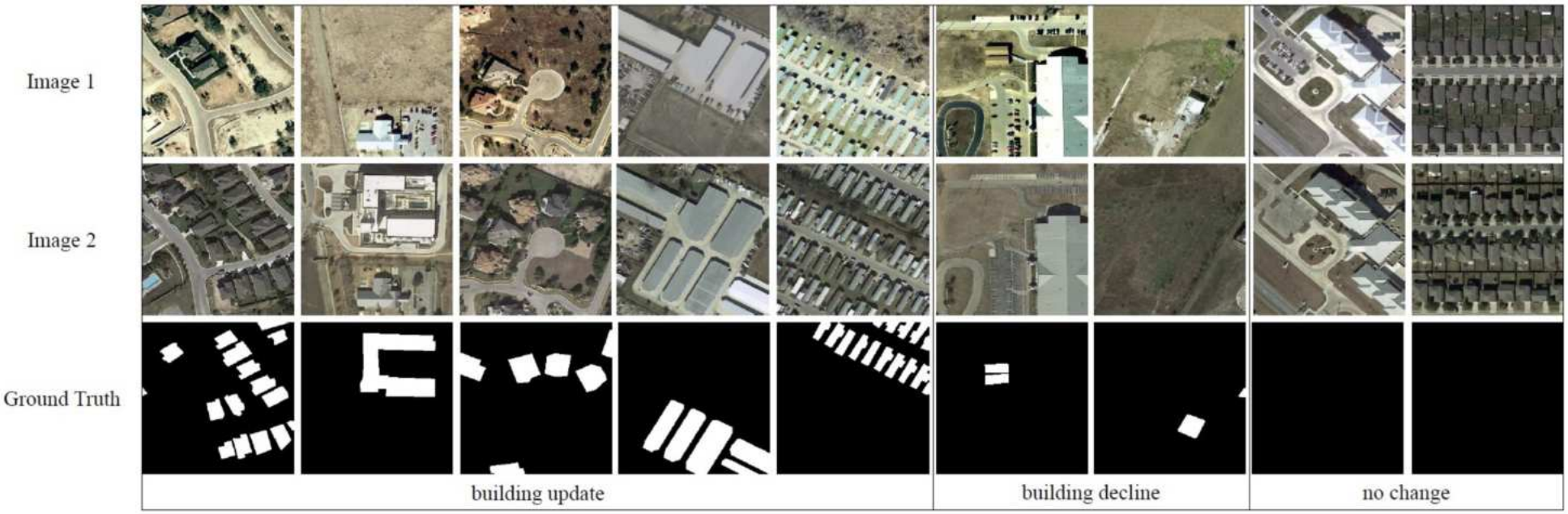

Vision-to-Language Generative Learning
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained Large Language Models (LLMs)
Bootstrapping from a Decoder-based Large Language Model (e.g. OPT)
Input Image
Image Encoder
Q-Former
Learned Queries
Fully Connected
Output Text
a cat wearing sunglasses
LLM Decoder
Bootstrapping from an Encoder-Decoder-based Large Language Model (e.g. FlanT5)
Input Image
Image Encoder
Q-Former
Learned Queries
Fully Connected
Suffix Text
wearing sunglasses
LLM Encoder
LLM Decoder
a cat
Prefix Text

Outline

- Introduction
- Our Proposed Approach
- Experimental Results
- Conclusions

LEVIR-CD, a large-scale remote sensing binary change detection dataset.

Cropped samples size of 256 × 256



Consists of 637 VHR Google Earth image patch pairs with a size of 1024 × 1024 pixels

LIM-CD, A LARGE-SCALE REMOTE SENSING CHANGE DETECTION DATASET FOR INCREMENTAL MONITORING.

LIM-CD: A LARGE-SCALE REMOTE SENSING CHANGE DETECTION DATASET FOR INCREMENTAL MONITORING

Hanchao Zhang [1], Ruiqian Zhang [1,*], Xiaogang Ning [1], Xiao Huang [2], You He [1], Yixin Chen [1], Mingzhu Li [3], Wei Cui [4], Jiaming Wang [5]

[1] Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping, Beijing, China
[2] Department of Environmental Sciences, Emory University, Atlanta, GA, USA
[3] School of Surveying, Mapping and Geographic Information, Liaoning Technical University, Fuxin, Liaoning, China
[4] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, Hubei, China
[5] Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan, Hubei, China
E-mail address: zhangrq@casm.ac.cn

The LIM-CD dataset comprises an extensive collection of image data from 15 different sensors. And the dataset contains a total of 9259 images, including 6547 images in the training set, 1776 images in the validation set, and 936 images in the test set. All the images in the dataset have a resolution ranging from 0.5-2 meters with 512x512 pixels.

Hanchao Zhang, Ruiqian Zhang*, Xiaogang Ning, et al. LIM-CD: A Large-scale Remote Sensing Change Detection Dataset for Incremental Monitoring[C]// ISPRS Annals, GSW X-1/W1-2023, 903-910.

The selection of images in the LIM-CD dataset was conducted with care to ensure that it covers a wide range of imaging variations, such as different sources of images, acquisition years, backgrounds, and terrains.

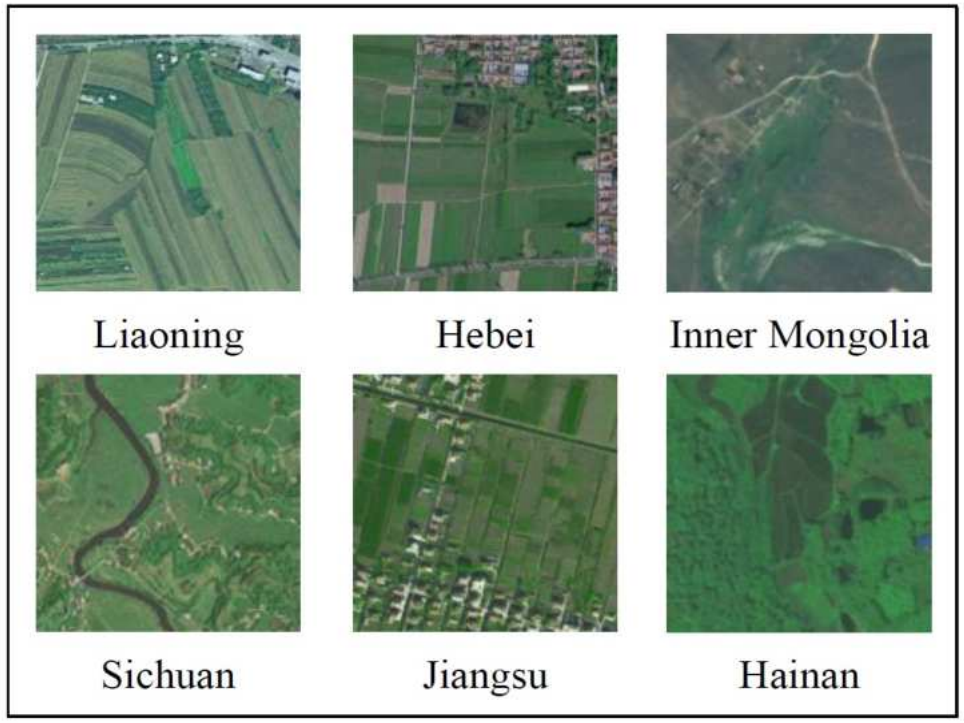


Figure 1. Different region image examples in the dataset

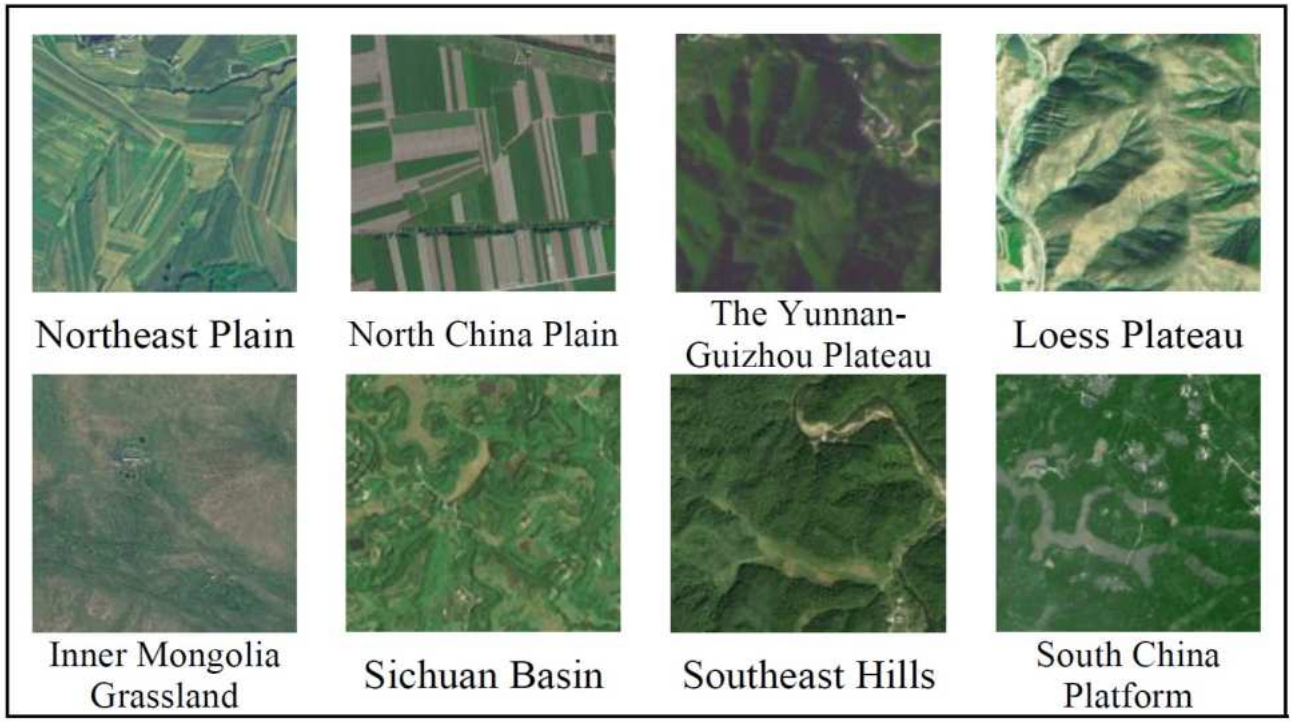


Figure 2. Different terrain image examples in the dataset

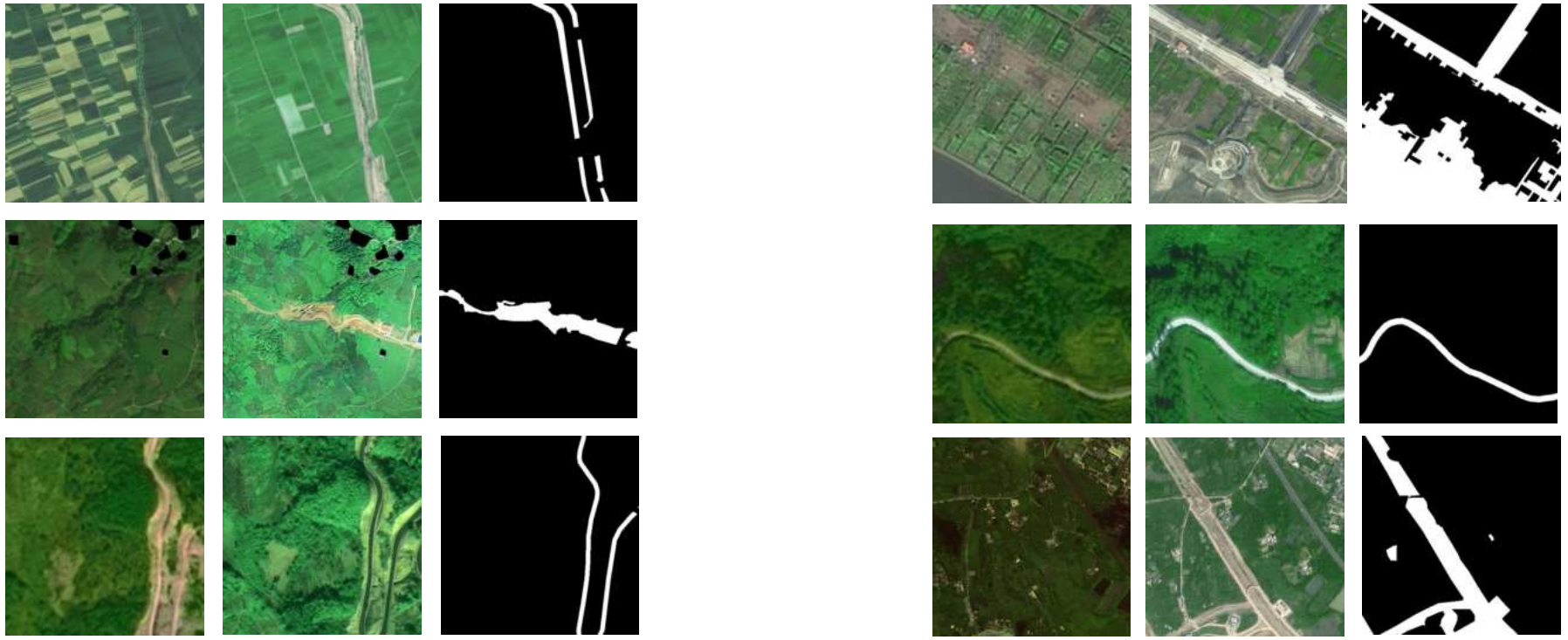
Comparisons of detection performance on Levir-CD

Method	OA	Precision	Recall	F1
FCEF	93.83	48.90	85.90	62.32
FC-Siam-diff	95.95	54.21	73.18	63.09
FC-Siam-conc	96.30	60.40	76.63	68.21
IFN	98.20	79.55	87.99	83.57
STANet	98.33	83.81	91.04	87.34
HRTNet	98.79	85.43	91.77	88.48
Our Method	99.08	90.08	92.08	91.07

Comparisons of detection performance on LIM-CD

Method	Precision	Recall	IOU	F1
FCEF	64.87	54.47	42.06	59.22
FC-Siam-diff	66.29	52.41	41.38	58.54
FC-Siam-conc	64.54	46.92	37.30	54.34
ISNet	66.41	54.63	42.80	59.95
ChangeFormer	70.84	45.36	38.22	55.31
BIT-CD	74.34	51.05	43.40	60.53
Our Method	75.44	51.31	43.97	61.08

Experimental results on the LEVIR-CD dataset
(From left to right: pre-phase image, post-phase image, ground truth and prediction result)



Experimental results on the LIM-CD dataset
(From left to right: pre-phase image, post-phase image, ground truth and prediction result)

Outline

- ☐ Introduction
- ☐ Our Proposed Approach
- ☐ Experimental Results
- ☐ Conclusions

Multimodal Foundation Model Based Siamese Network for Change Detection in Remote Sensing Imagery

- We developed a novel Multimodal Foundation Model (MFM) Based Siamese Change Detection Network, which comprises a feature extraction network, an MFM-based domain understanding network, and a change region detection network.
- Our framework has undergone extensive experimental analysis and benchmarking against state-of-the-art methods on two publicly available datasets.
- The results demonstrate the exceptional efficacy and logical coherence of our proposed framework and its individual components, validating their applicability in change detection scenarios.

SUSTAINABLE DEVELOPMENT GOALS | International Federation of Surveyors supports the Sustainable Development Goals

Commission

Commission's name

Serving Society for the Benefit of People and Planet

Homepage