



MMM-RS: A Multi-modal, Multi-GSD, Multi-scene Remote Sensing Dataset and Benchmark for Text-to-Image Generation

Jialin Luo, Yuanzhi Wang, Ziqi Gu, Yide Qiu, Shuaizhen Yao, Fuyun Wang, Chunyan Xu, Wenhua Zhang, Dan Wang, Zhen Cui

1. Nanjing University of Science and Technology, Nanjing, China
2. Beijing Institute of Spacecraft System Engineering, Beijing, China.

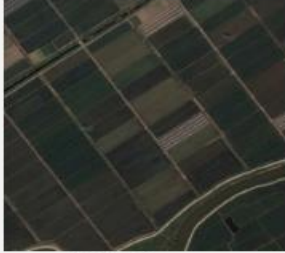
Reporter: **Jialin Luo**



Motivation

- Generating diverse remote sensing (RS) images that are tremendously different from general images in terms of scale and perspective remains a formidable challenge due to the lack of a comprehensive remote sensing image generation dataset with various modalities, ground sample distances (GSD), and scenes.

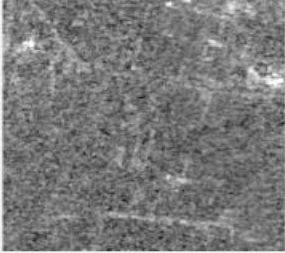

(a) A sample in RSICD dataset



Simple Text Prompt:
a river is next to many pieces of green farmlands.

RGB Image

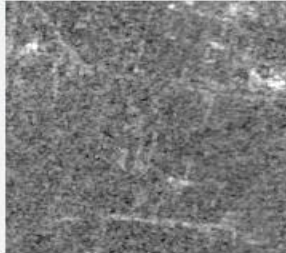

(b) A sample in SEN1-2 dataset



RGB Image

SAR Image

(c) A sample in our proposed MMM-RS dataset



Information-rich Text Prompt:
Ordinary precision resolution, fog, a satellite image shows a large area of farmland, Sentinel-2.

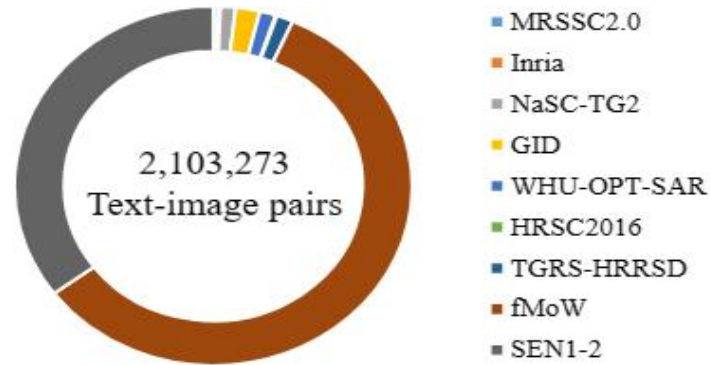
- Simple text prompt describing image content
- Ground Sample Distance (GSD) level
- Type of weather
- Type of satellite

RGB Image

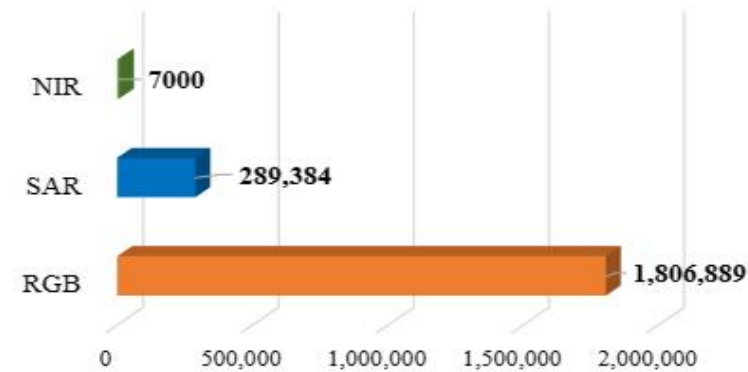
SAR Image

The construction of MMM-RS

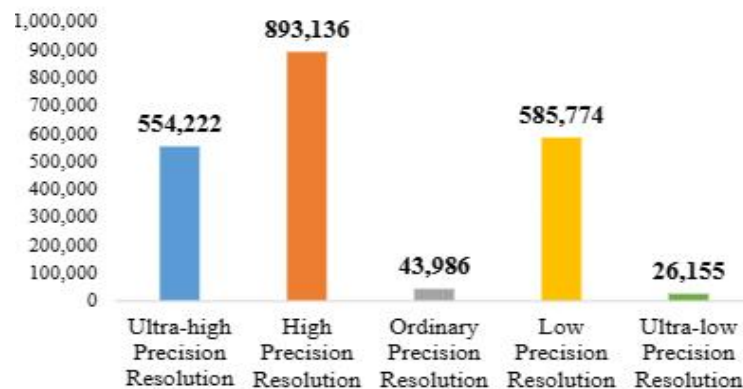
- In this paper, we propose a Multi-modal, Multi-GSD, Multi-scene Remote Sensing (MMM-RS) dataset and benchmark for text-to-image generation in diverse remote sensing scenarios.



(a) Percentage of Different Datasets



(b) Number of Different Modalities



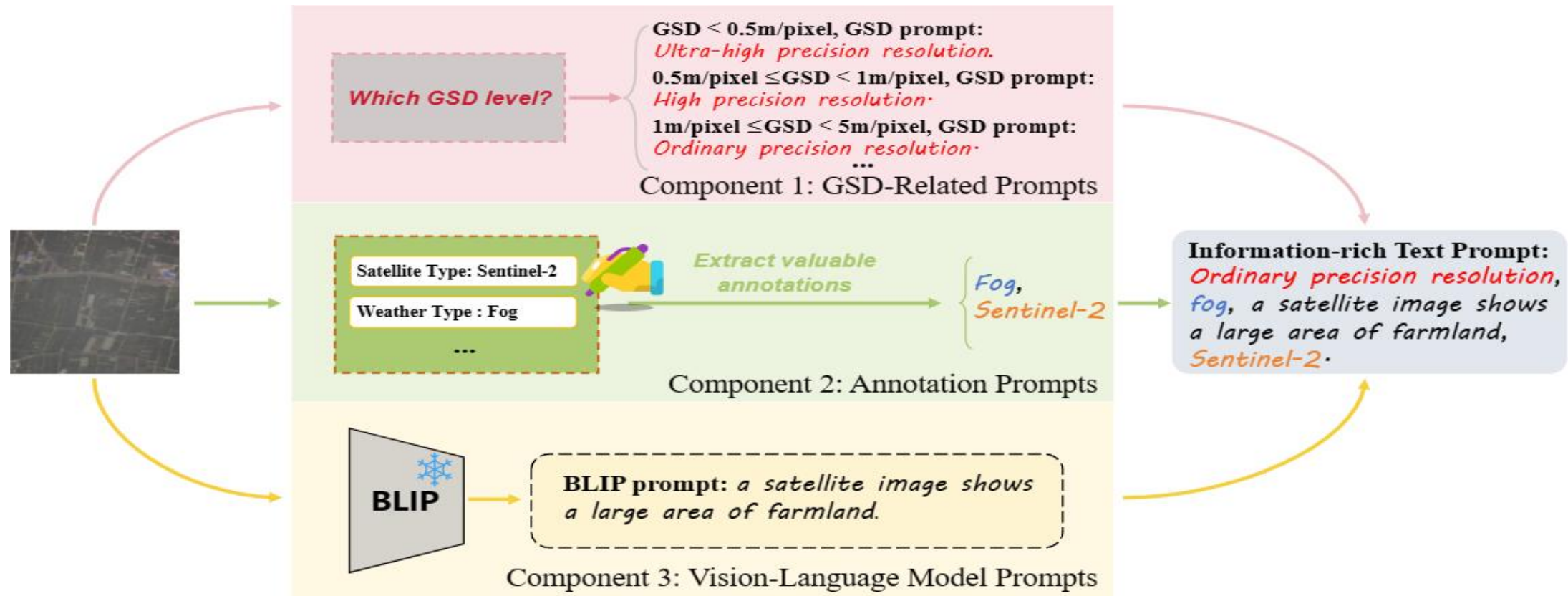
(c) Number of Different GSD Level



(d) Visualization of Category Distribution

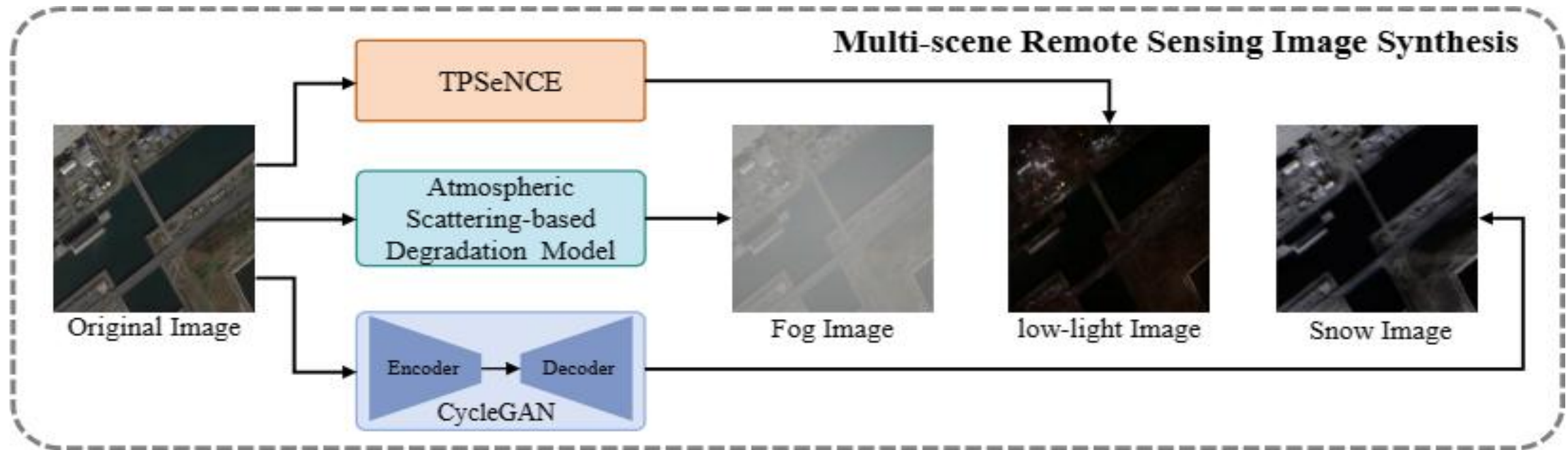
The construction of MMM-RS

(1) We first collect nine publicly available RS datasets and conduct standardization for all samples.



The construction of MMM-RS

(2) To bridge RS images to textual semantic information, we utilize a large-scale pretrained vision-language model to automatically output text prompts and perform hand-crafted rectification, resulting in information-rich text-image pairs.



The construction of MMM-RS

(3) We design some methods to obtain the images with different GSD and various environments (e.g., low-light, foggy) in a single sample.



Original Image
GSD: 0.3m/pixel



GSD: 2.4m/pixel



GSD: 1.2m/pixel



GSD: 0.6m/pixel



GSD: 0.3m/pixel

Experiments

Prompt

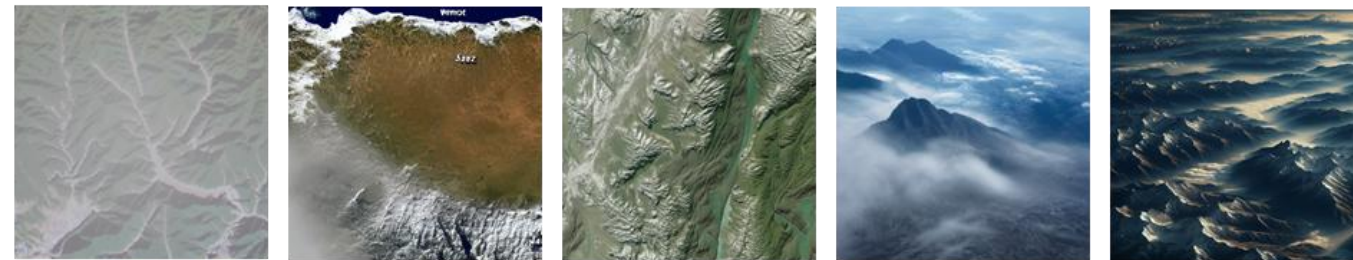
High precision resolution, a satellite image shows a small town in the middle of a field, GF2



High precision resolution, snow, a satellite image shows a park in the city, Google Earth



Ultra-low precision resolution, fog, a satellite image shows a large mountain range, T62



High precision resolution, night, a satellite image shows a street in a residential area, Google Earth



Experiments

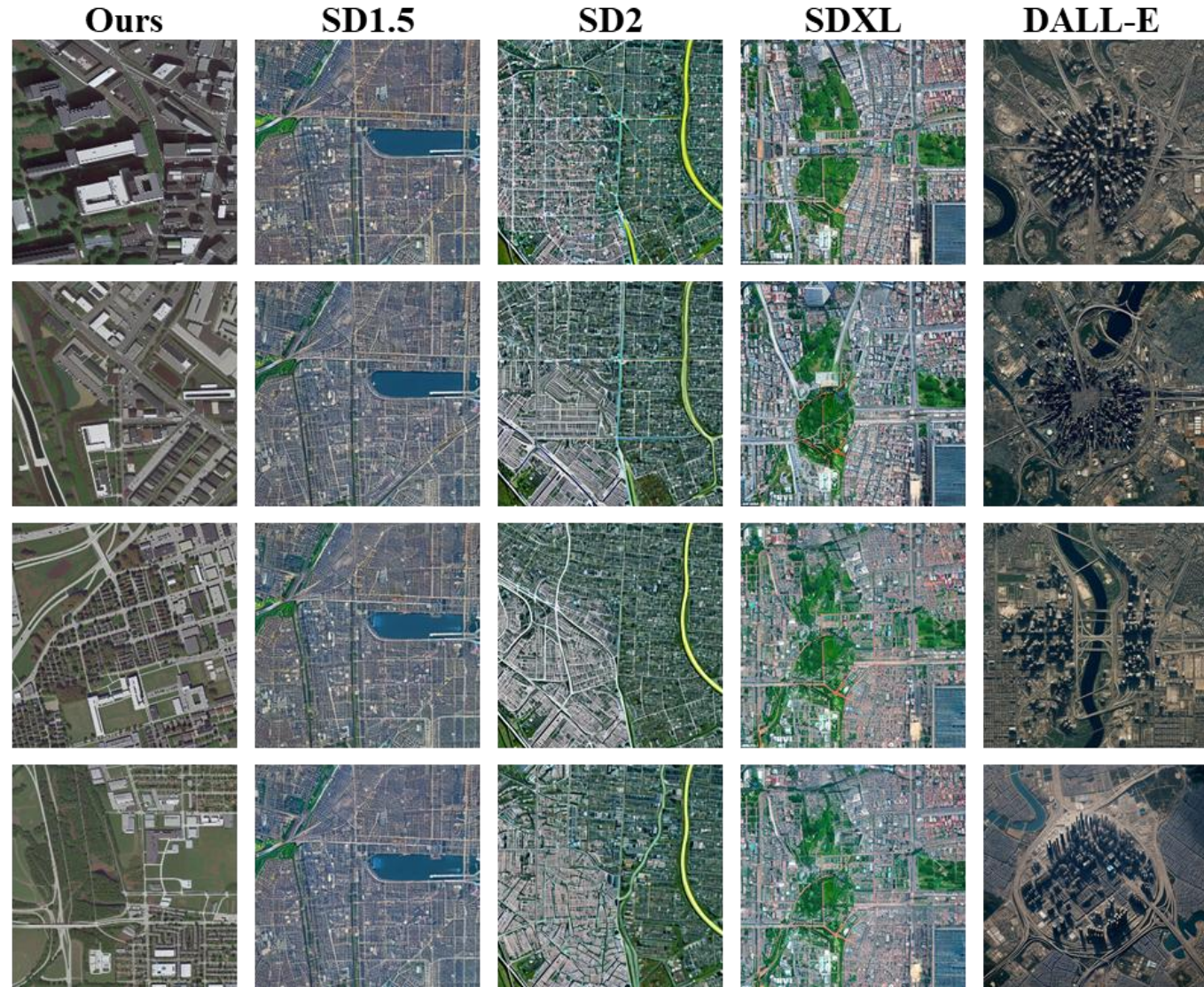
Prompt

Ultra-high precision resolution, a satellite image shows the area around a large city, Google Earth

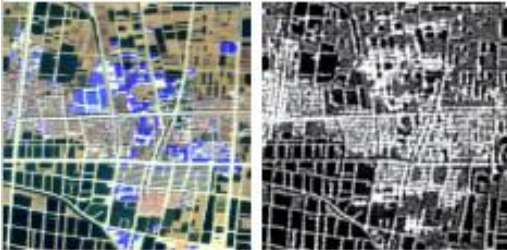
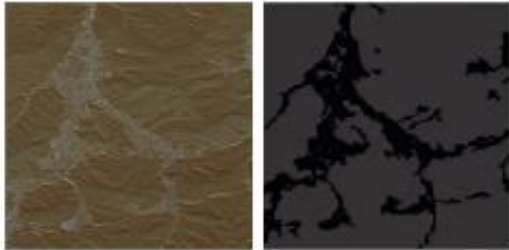
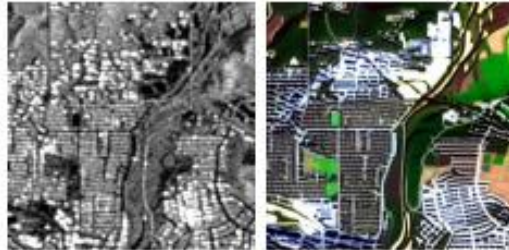

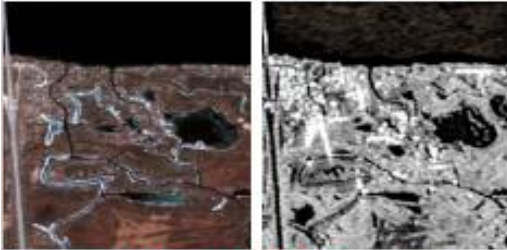
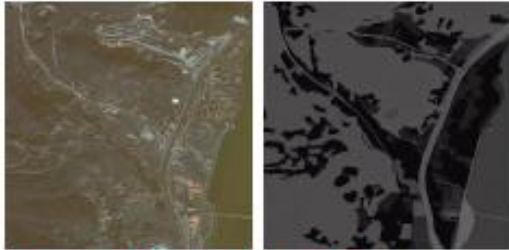
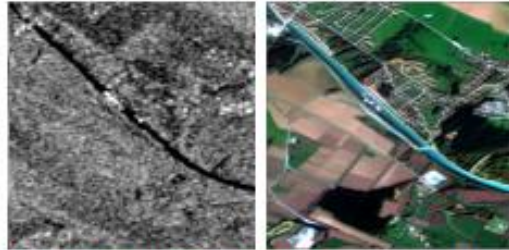
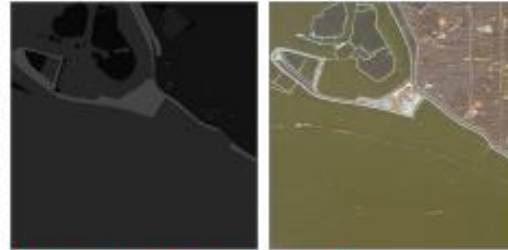
High precision resolution, a satellite image shows the area around a large city, Google Earth

Ordinary precision resolution, a satellite image shows the area around a large city, Google Earth

Low precision resolution, a satellite image shows the area around a large city, Google Earth



Experiments

RGB → SAR	RGB → NIR	SAR → RGB	NIR → RGB
			
<p><i>Low precision resolution, a satellite image shows some roads in farmland, Sentinel-2</i></p>	<p><i>Low precision resolution, a satellite image of a mountain with a river running through it, GF1</i></p>	<p><i>Low precision resolution, a satellite image shows a city, Sentinel-1</i></p>	<p><i>Low precision resolution, a satellite image shows a river flowing through the mountains, GF1</i></p>
			
<p><i>Low precision resolution, a satellite image shows large areas of lakes and rivers, Sentinel-2</i></p>	<p><i>Low precision resolution, a satellite image of a city with a river, GF1</i></p>	<p><i>Low precision resolution, a satellite image of a village with a river running through it, Sentinel-1</i></p>	<p><i>Low precision resolution, a satellite image of a city with a river, GF1</i></p>

Conclusion

- We construct a large-scale Multi-modal, Multi-GSD, and Multi-scene Remote Sensing (MMM-RS) dataset and benchmark for text-to-image generation in diverse RS scenarios, which standardizes 9 publicly available RS datasets with uniform and information-rich text prompts.
- To provide the various GSD samples, we design a GSD sample extraction strategy that extracts different GSD levels images for each sample and define the GSD-related text prompts describing different GSD levels. Furthermore, due to the lack of real-world multi-scene samples, we select some RGB samples and utilize existing techniques to synthesize samples with different scenes including fog, snow, and low-light environments.
- We use our proposed MMM-RS dataset to fine-tune the advanced Stable Diffusion, and perform extensive quantitative and qualitative comparisons to prove the effectiveness of our MMM-RS dataset. In particular, we use the aligned multi-modal samples (including RGB, SAR, and infrared modalities) in the MMM-RS dataset to train the cross-modal generation models based on ControlNet, and the visualization results demonstrates impressive cross-modal generation capabilities.