

Is synthetic data from generative models ready for image recognition?

GISLab Short-Term Course 2025 Summer

Zhenyuan Chen

School of Earth Science, Zhejiang University

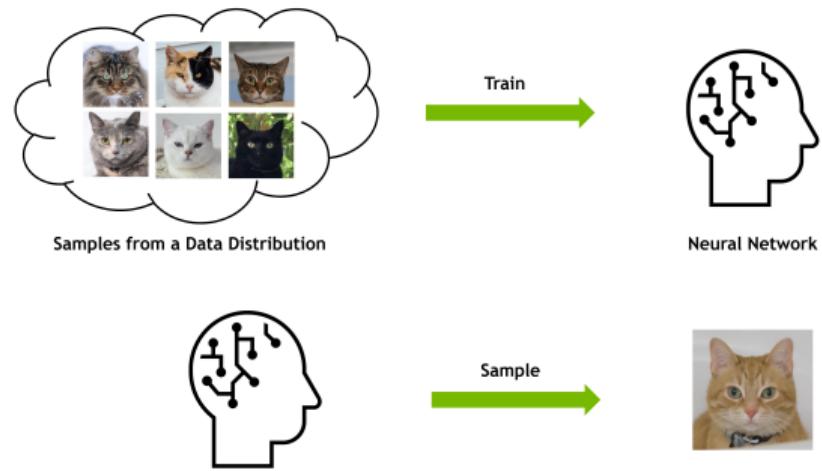
2025
bili_sakura@zju.edu.cn

Outline

- ▶ Generative Models and Diffusion Methods
- ▶ Applications in Remote Sensing and Synthetic Data
- ▶ Project Assignment

Generative Modeling

Deep Generative Learning Learning to generate data



2

Figure: Illustration of generative modeling (Vahdat Arash, Song, and Meng, 2023).

Timeline of Generative Models

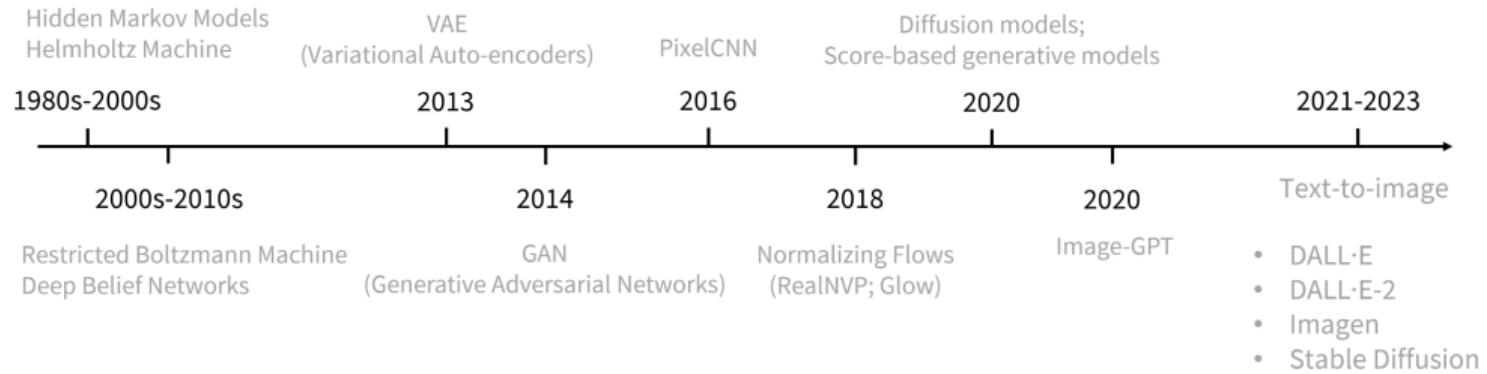


Figure: Timeline of key developments in generative models (Deng, 2024).

Background: Diffusion Models

Denoising diffusion models consist of two processes:

- ▶ A forward diffusion process that gradually adds noise to the input.
- ▶ A reverse denoising process that learns to generate data by denoising.

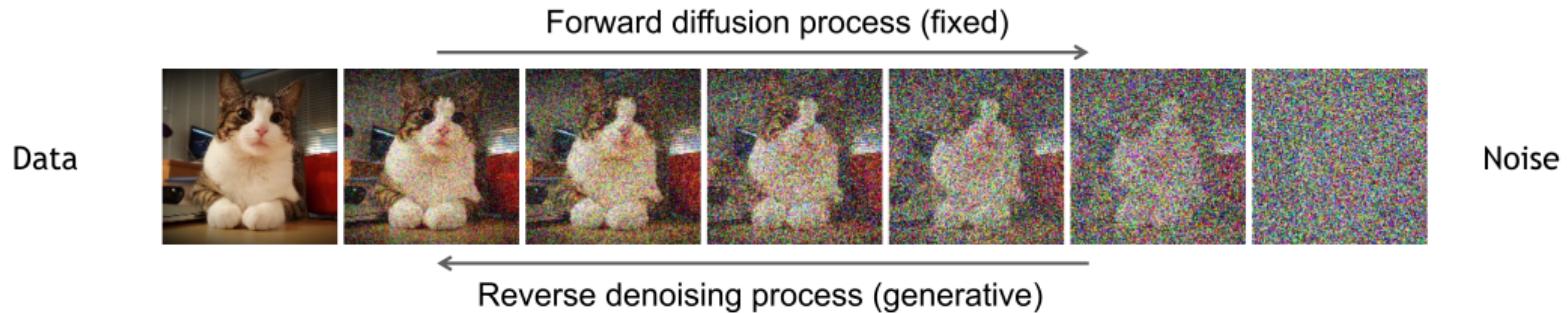


Figure: Diffusion models generate data through iterative denoising (Sohl-Dickstein et al., 2015; Ho, Jain, and Abbeel, 2020).

Diffusion Models: Forward and Reverse Processes

Forward (Diffusion) Process:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$$

$$\text{Equivalently, } \mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Reverse (Denoising) Process:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$$

where \mathbf{x}_0 is the data, β_t is the noise schedule, and $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$. $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Diffusion models generate data by learning to reverse a gradual noising process. (Sohl-Dickstein et al., 2015; Ho, Jain, and Abbeel, 2020)

Diffusion Models: Training and Inference

Training Objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0, \epsilon, t} \left[\left\| \epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|^2 \right]$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$.

Inference (Sampling):

- ▶ Start from pure noise: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- ▶ For $t = T, \dots, 1$:
 - ▶ Predict noise: $\epsilon_{\theta}(\mathbf{x}_t, t)$
 - ▶ Compute mean: $\mu_{\theta}(\mathbf{x}_t, t)$
 - ▶ Sample: $\mathbf{x}_{t-1} \sim \mathcal{N}(\mu_{\theta}(\mathbf{x}_t, t), \Sigma_{\theta}(\mathbf{x}_t, t))$
- ▶ Repeat until \mathbf{x}_0 (generated sample)

Training: Minimize the simplified objective (Ho, Jain, and Abbeel, 2020).

Inference: Iteratively denoise from random noise to generate data.

Application in Remote Sensing Image Generation: Text2Earth

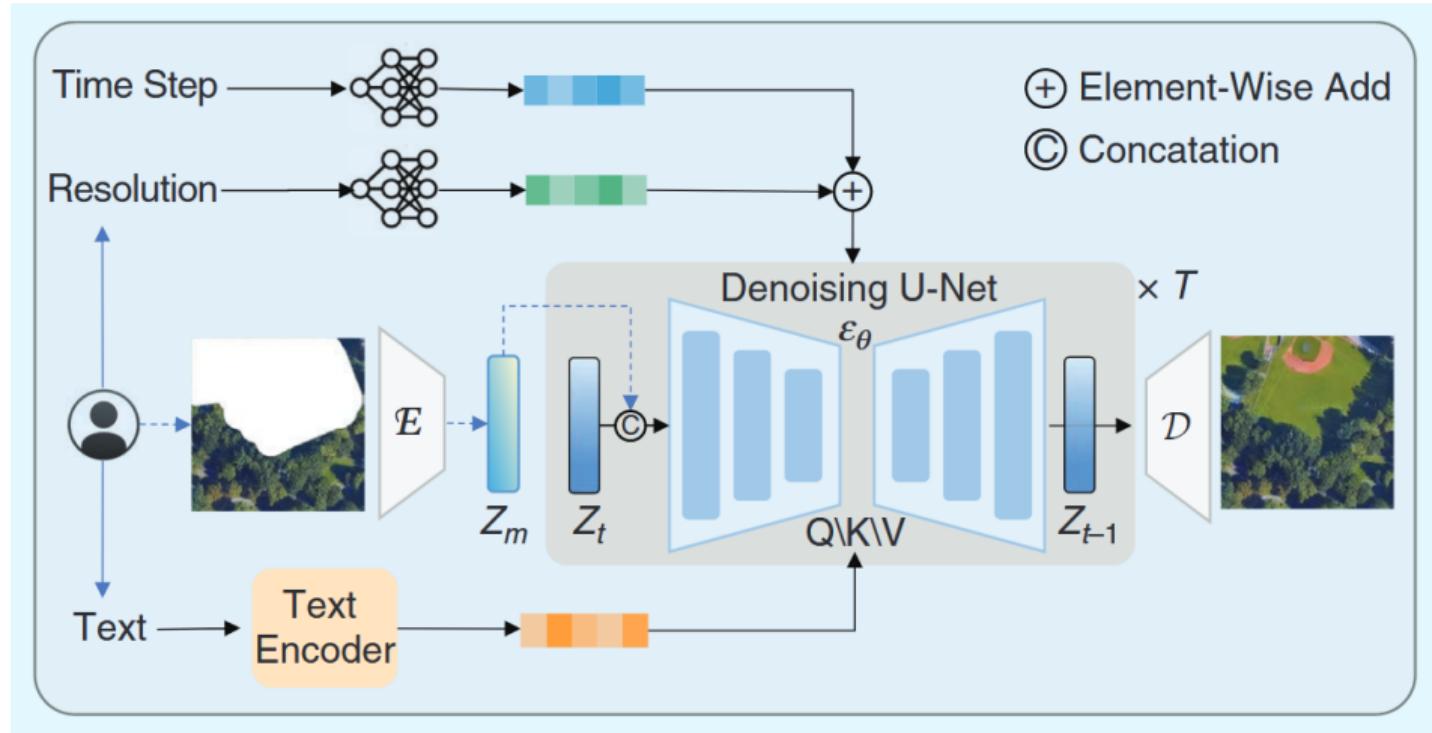


Figure: Text2Earth: Foundation model for text-driven Earth observation (Liu et al., 2025).

Text2Earth: Example Results

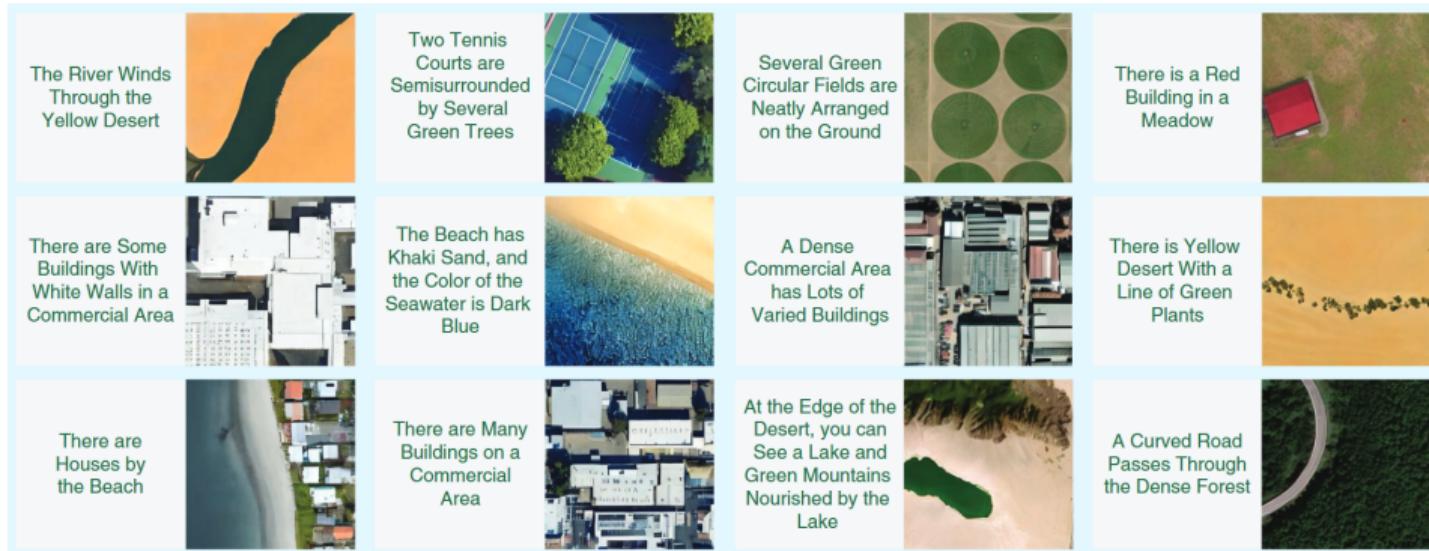


Figure: Example results generated by Text2Earth (Liu et al., 2025).

Application in Remote Sensing Image Generation: CRS-Diff

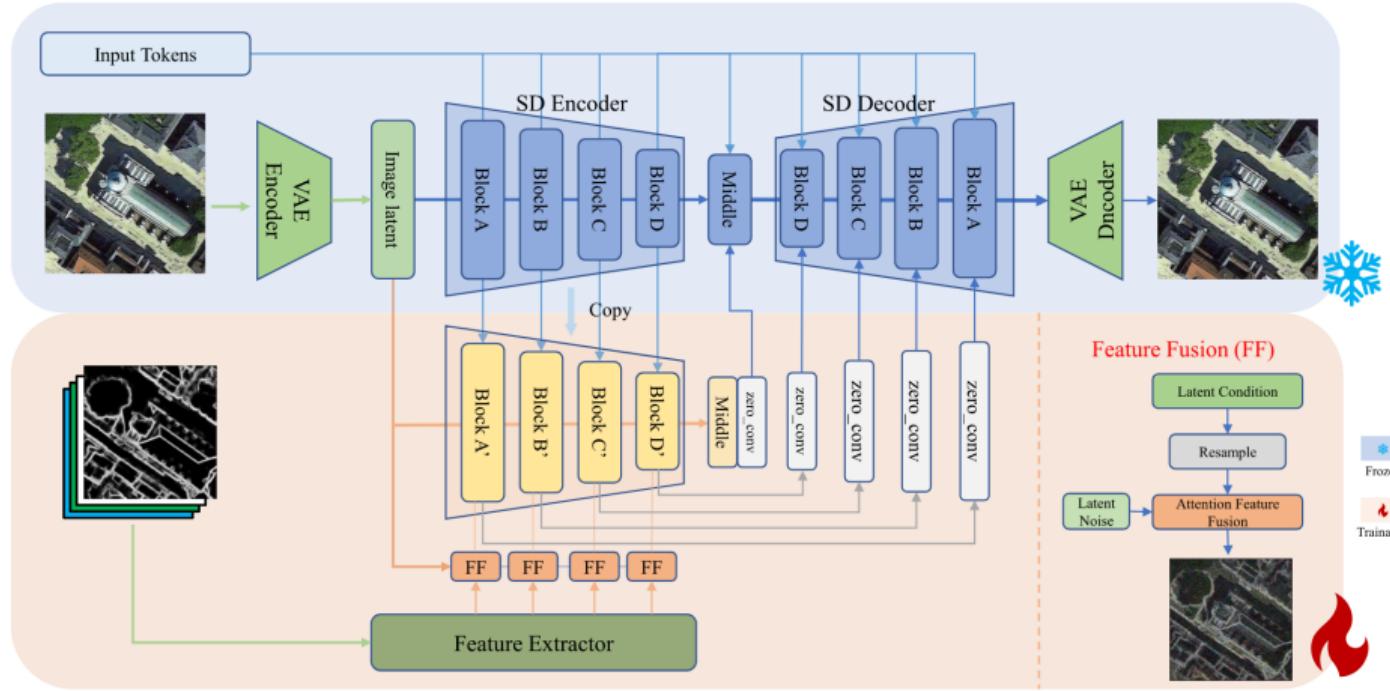


Figure: CRS-Diff: Controllable remote sensing image generation framework (Tang, Li, et al., 2024).

CRS-Diff: Example Results

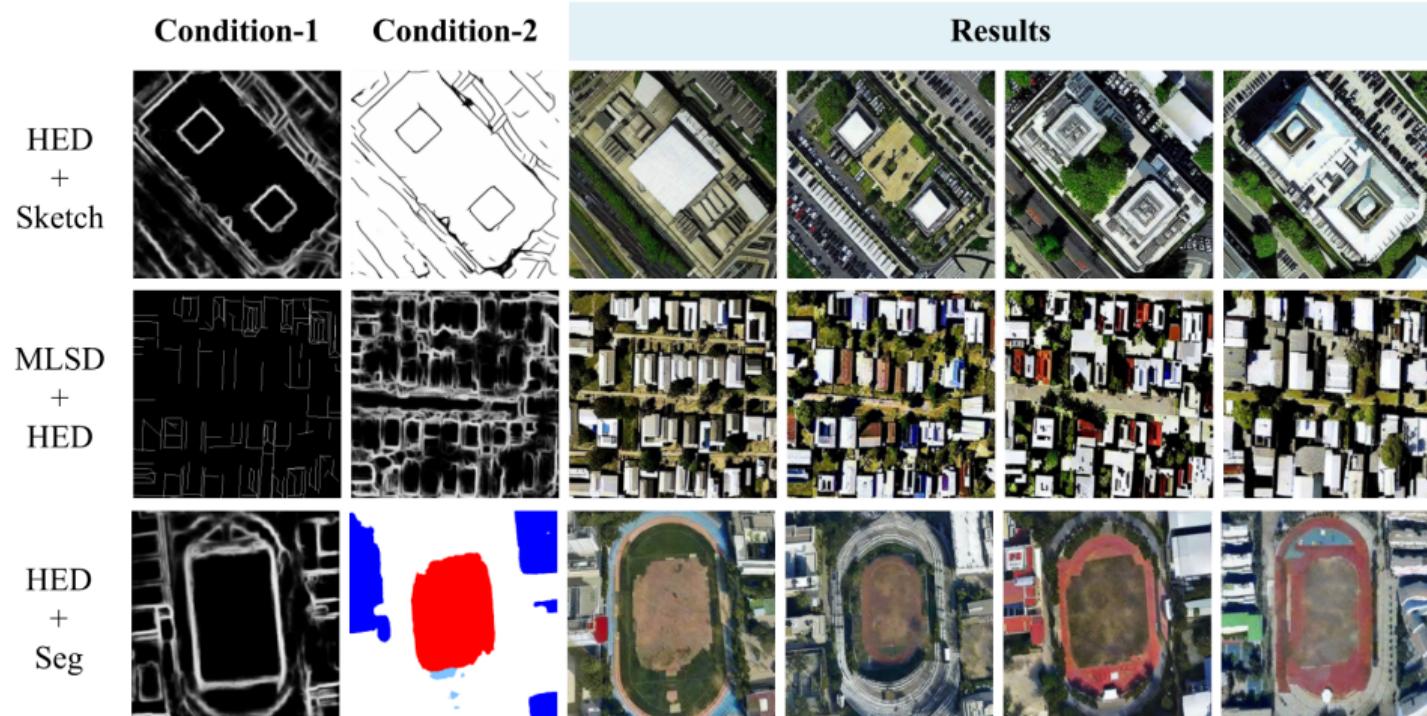


Figure: Example results generated by CRS-Diff (Tang, Li, et al., 2024).

DiffusionSat: Framework Overview

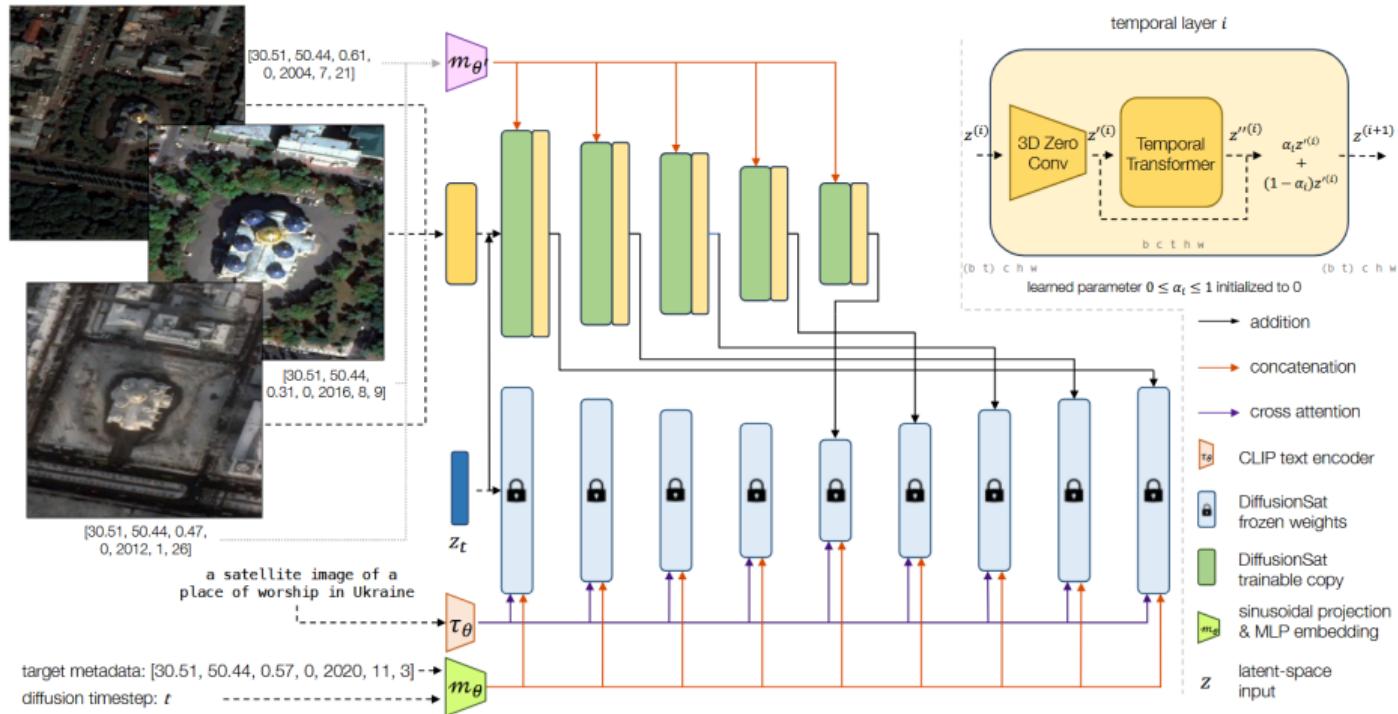


Figure: DiffusionSat: A generative foundation model for satellite imagery (Khanna et al., 2024).

DiffusionSat: Super-Resolution Results

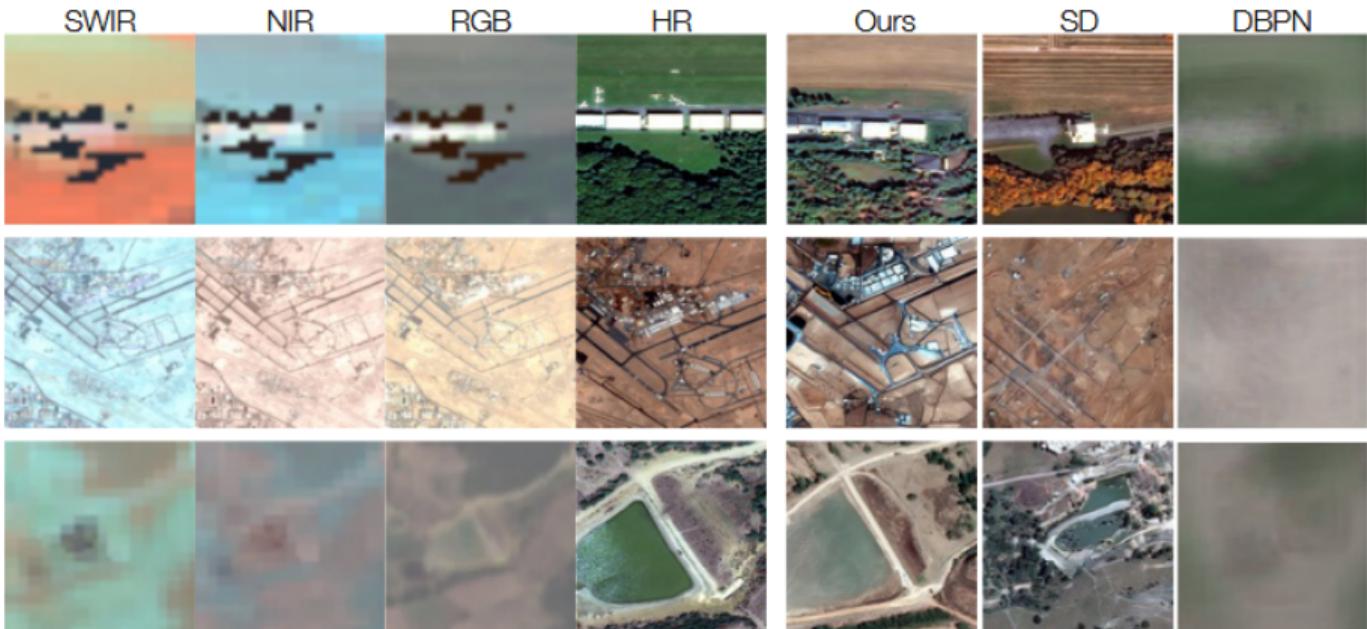


Figure: Example results: DiffusionSat for multi-spectral super-resolution (Khanna et al., 2024).

DiffusionSat: Inpainting Results



Figure: Example results: DiffusionSat for remote sensing image inpainting (Khanna et al., 2024).

Q&A

Project Assignment: Research Topic

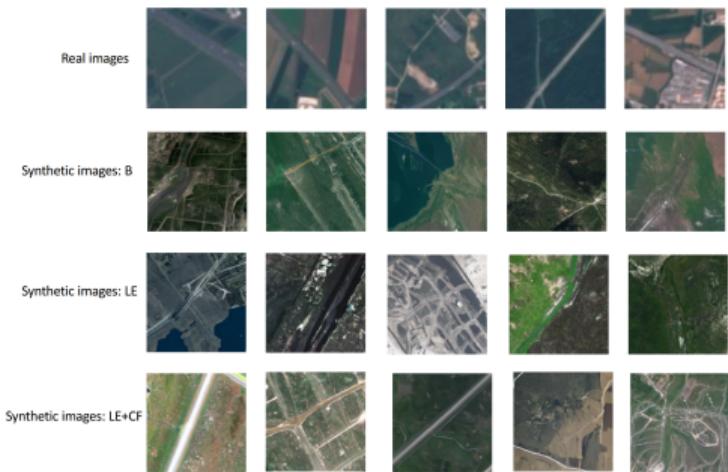
Main Question:

Is synthetic data from generative models ready for image recognition?

- ▶ Synthetic data generated by generative models is a trending new way of data augmentation (He et al., 2023; Toker et al., 2024).
- ▶ In this project, we will explore whether such augmentation benefits downstream tasks such as text-image retrieval, image scene classification, and super-resolution in the remote sensing domain.

Background: Why Synthetic Data?

- ▶ Manual data collection and annotation is expensive and time-consuming.
- ▶ Synthetic data from generative models enables large-scale data augmentation.
- ▶ The effectiveness of synthetic data for downstream remote sensing tasks is under active investigation.



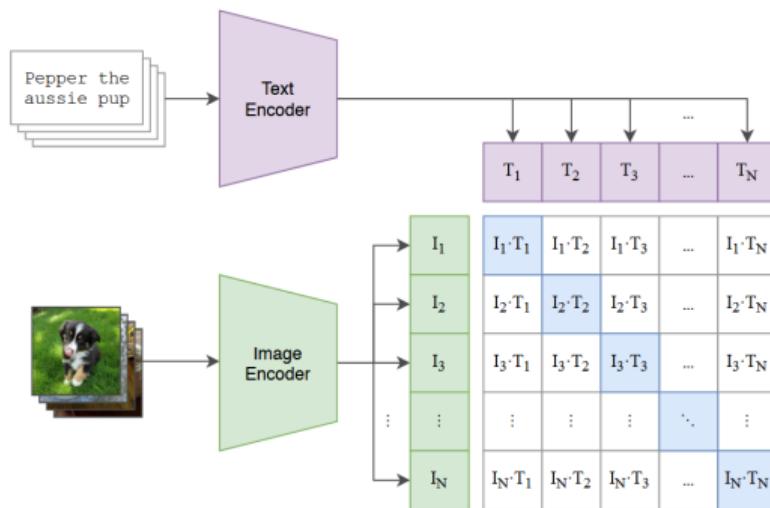
Visualization of different strategies of synthetic data in zero-shot settings (He et al., 2023). Visual illustration of ground-truth real data and synthesized images by different strategies for the zero-shot settings: basic (B), language enhancement (LE), and language enhancement with CLIP-based filtering (LE+CF).

Experimental Settings: Generative Data for Downstream Tasks

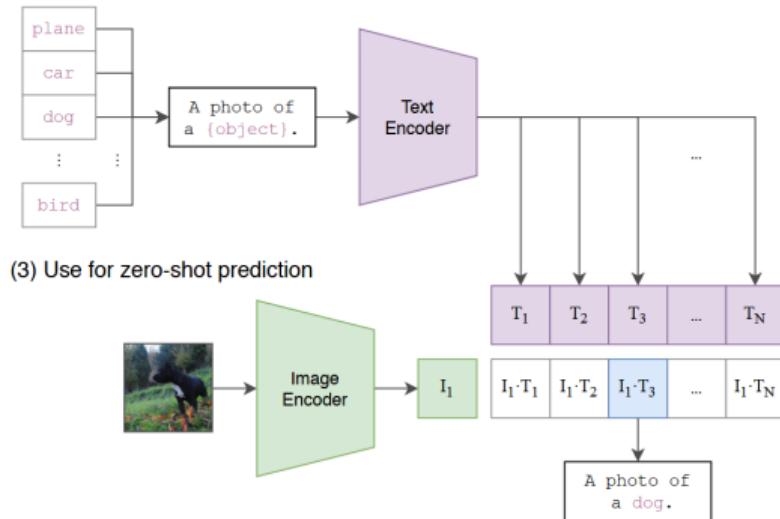
- ▶ We leverage state-of-the-art generative models such as **DiffusionSat** (Khanna et al., 2024) and **Text2Earth** (Liu et al., 2025) for synthetic data augmentation.
- ▶ **Two main strategies:**
 1. **Text-to-image generation:**
 - ▶ Generate text-image pairs.
 - ▶ Enables tasks like image scene classification and image-text retrieval.
 2. **Super-resolution generation:**
 - ▶ Generate paired low-resolution (LR) and high-resolution (HR) images.
 - ▶ Enables super-resolution tasks in remote sensing.

Image Scene Classification with CLIP

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

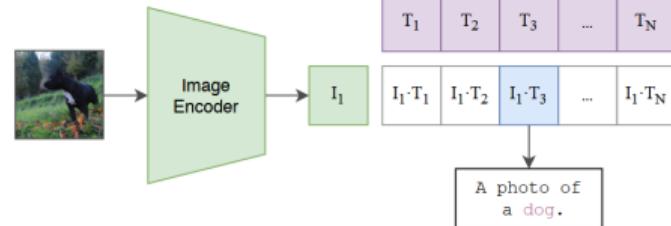


Figure: CLIP (Radford et al., 2021) model overview: joint training of image and text encoders for cross-modal understanding.

Super-Resolution with Real-ESRGAN

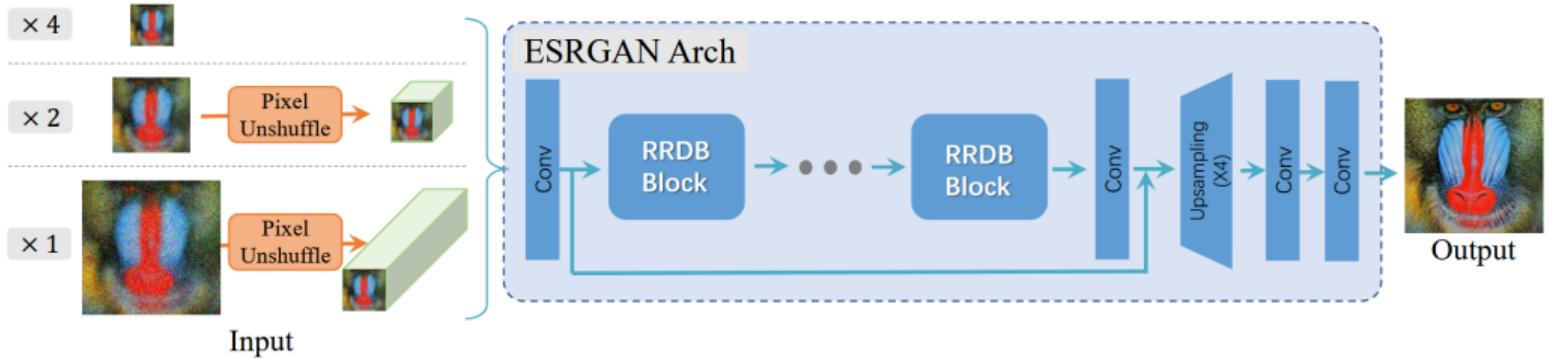


Figure: Real-ESRGAN (Wang et al., 2021) framework: architecture for real-world image super-resolution.

Baseline Models for Scene Classification and Super-Resolution

Scene Image Classification:

- ▶ **CLIP** (Radford et al., 2021)
- ▶ **RemoteCLIP** (Liu, Chen, Guan, et al., 2024)
- ▶ **Git-RSCLIP** (Liu, Chen, Zhao, et al., 2025)

Super-Resolution:

- ▶ **Real-ESRGAN** (Wang, Xie, et al., 2021)
- ▶ **StableSR** (Wang, Yue, et al., 2024)
- ▶ **FaithDiff** (Chen, Pan, and Dong, 2025)
- ▶ **EResShift** (Yue, Wang, and Loy, 2025)

Radford, et al. Learning Transferable Visual Models From Natural Language Supervision, ICML, 2021.

Liu, Chen, Guan, et al. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. TGRS. 2024.

Liu, Chen, Zhao, et al. Text2Earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. GRSM. 2025.

Wang, Xie, et al. Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data, ICCV, 2021.

Wang, Yue, et al. Exploiting Diffusion Prior for Real-World Image Super-Resolution. IJCV. 2024.

Chen, et al. FaithDiff: Unleashing Diffusion Priors for Faithful Image Super-resolution, CVPR, 2025.

Yue, et al. Efficient Diffusion Model for Image Restoration by Residual Shifting. TPAMI. 2025.

Datasets for Text-to-Image Generation

Text-to-Image Generation:

- ▶ **RSICD** (Lu et al., 2018): Remote Sensing Image Captioning Dataset with 10,921 images and five captions per image.
- ▶ **RSICap** (Hu et al., 2025): High-quality dataset with 2,585 human-annotated image-caption pairs.
- ▶ **UCM-Captions** (Qu et al., 2016): Derived from the UC Merced Land Use Dataset, containing 2,100 images with five captions each.

Datasets for Super-Resolution

Super-Resolution:

- ▶ **fMoW**: Paired dataset of Sentinel-2 (10m GSD) (Cong et al., 2022) and fMoW-RGB (0.3m) (Christie et al., 2018).
- ▶ 22,852 high-resolution RSIs (256×256 size) used in (Meng et al., 2024) from Potsdam from the ISPRS 2-D semantic labeling contest (Sherrah, 2016), Toronto dataset (Rottensteiner et al., 2014) and UC Merced data (Yang and Newsam, 2010).

Cong, et al. Functional Map of the World - Sentinel-2 corresponding images. 2022.

Christie, et al. Functional Map of the World, CVPR, 2018.

Meng, et al. A Conditional Diffusion Model With Fast Sampling Strategy for Remote Sensing Image Super-Resolution. IEEE Transactions on Geoscience and Remote Sensing. 2024.

Sherrah, Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery, arXiv, 2016.

Rottensteiner, et al. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing. 2014.

Yang, et al. Bag-of-visual-words and spatial extensions for land-use classification, Association for Computing Machinery, 2010.

Possible Results

- ▶ Results from (He et al., 2023) demonstrate notable accuracy gains on remote sensing benchmarks.

| Dataset | Task | CLIP-RN50 | CLIP-RN50+SYN | CLIP-ViT-B/16 | CLIP-ViT-B/16+SYN |
|-----------------|------|-----------|----------------|---------------|-------------------|
| CIFAR-10 | o | 70.31 | 80.06 (+9.75) | 90.80 | 92.37 (+1.57) |
| CIFAR-100 | o | 35.35 | 45.69 (+10.34) | 68.22 | 70.71 (+2.49) |
| Caltech101 | o | 86.09 | 87.74 (+1.65) | 92.98 | 94.16 (+1.18) |
| Caltech256 | o | 73.36 | 75.74 (+2.38) | 80.14 | 81.43 (+1.29) |
| ImageNet | o | 60.33 | 60.78 (+0.45) | 68.75 | 69.16 (+0.41) |
| SUN397 | s | 58.51 | 60.07 (+1.56) | 62.51 | 63.79 (+1.28) |
| Aircraft | f | 17.34 | 21.94 (+4.60) | 24.81 | 30.78 (+5.97) |
| Birdsnap | f | 34.33 | 38.05 (+3.72) | 41.90 | 46.84 (+4.94) |
| Cars | f | 55.63 | 56.93 (+1.30) | 65.23 | 66.86 (+1.63) |
| CUB | f | 46.69 | 56.94 (+10.25) | 55.23 | 63.79 (+8.56) |
| Flower | f | 66.08 | 67.05 (+0.97) | 71.30 | 72.60 (+1.30) |
| Food | f | 80.34 | 80.35 (+0.01) | 88.75 | 88.83 (+0.08) |
| Pets | f | 85.80 | 86.81 (+1.01) | 89.10 | 90.41 (+1.31) |
| DTD | t | 42.23 | 43.19 (+0.96) | 44.39 | 44.92 (+0.53) |
| EuroSAT | si | 37.51 | 55.37 (+17.86) | 47.77 | 59.86 (+12.09) |
| ImageNet-Sketch | r | 33.29 | 36.55 (+3.26) | 46.20 | 48.47 (+2.27) |
| ImageNet-R | r | 56.16 | 59.37 (+3.21) | 74.01 | 76.41 (+2.40) |
| Average | / | 55.13 | 59.47 (+4.31) | 65.42 | 68.32 (+2.90) |

Table 1: **Main Results on Zero-shot Image Recognition.** All results are top-1 accuracy on test set.
o: object-level. s: scene-level. f: fine-grained. t: textures. si: satellite images. r: robustness.

Figure: Improvement in classification performance when using synthetic data augmentation (He et al., 2023).

Diffusion-based Synthetic Data Augmentation is Superior

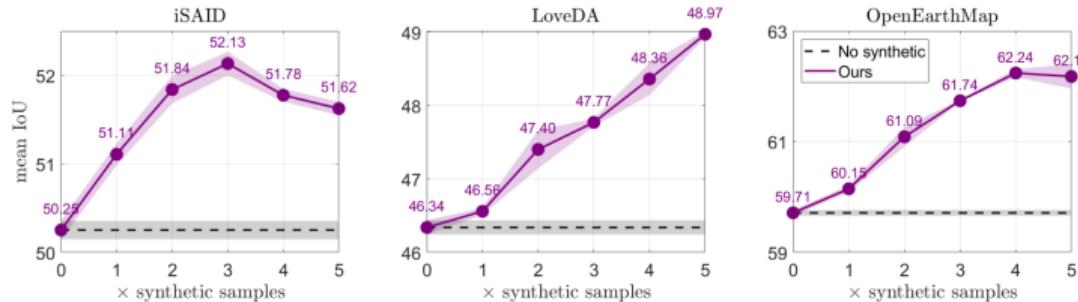


Figure: Table from SatSynth: Diffusion-based synthetic data augmentation achieves higher segmentation accuracy compared to classical augmentation (Toker et al., 2024).

Table: Quantitative comparison: mean IoU on iSAID (Waqas Zamir et al., 2019) dataset.

| No add. | Ours | Cutout (DeVries and Taylor, 2017) | CutMix (Yun et al., 2019) | Copy-Paste (Ghiasi et al., 2021) |
|---------|--------------|-----------------------------------|---------------------------|----------------------------------|
| 50.25 | 51.11 | 50.47 | 50.60 | 50.51 |

Toker, et al. SatSynth: Augmenting Image-Mask Pairs through Diffusion Models for Aerial Semantic Segmentation, CVPR, 2024.

Waqas Zamir, et al. isaid: A large-scale dataset for instance segmentation in aerial images, CVPRW, 2019.

DeVries, et al. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552. 2017.

Yun, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features, ICCV, 2019.

Ghiasi, et al. Simple copy-paste is a strong data augmentation method for instance segmentation, CVPR, 2021.