生成模型合成数据已准备好用于图像识别了吗?

GISLab 2025 年暑期短课程

陈振源

浙江大学地球科学学院

2025

bili_sakura@zju.edu.cn

目录

图像分类：概述

图: 图像分类的基本流程。

背景：深度学习下的图像分类

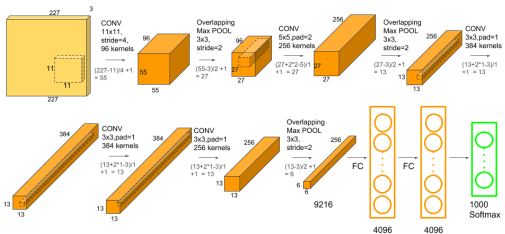
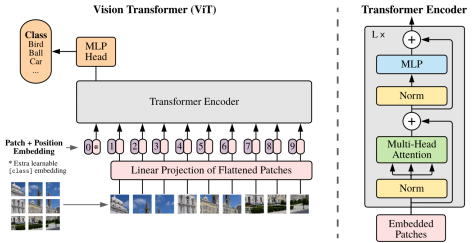


图: 左：AlexNet 在 ILSVRC-2010 (Berg, Deng, and Fei-Fei, 2010)　右：AlexNet 网络结构 (Krizhevsky, Sutskever, and Hinton, 2012)。

Berg, et al. Large Scale Visual Recognition Challenge 2010, 2010.
Krizhevsky, et al. ImageNet Classification with Deep Convolutional Neural Networks, NeurIPS, 2012.

图像分类模型结构演进

- **2012: AlexNet**, **2016: ResNet**
- **2021: ViT**
- **2021: Swin Transformer**
  (Liu et al., 2021) (Dosovitskiy et al., 2021)
- **2021: CLIP-ViT**
  (Radford et al., 2021)
- **2022: MAE-ViT**
  (He et al., 2022)
- **2022: CoCa-ViT**
  (Yu et al., 2022)



Vision Transformer 概览
(Dosovitskiy et al., 2021)。

Liu, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows, ICCV, 2021.
Dosovitskiy, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR, 2021.
Radford, et al. Learning Transferable Visual Models From Natural Language Supervision, ICML, 2021.
He, et al. Masked Autoencoders Are Scalable Vision Learners, CVPR, 2022.
Yu, et al. CoCa: Contrastive Captioners Are Image-Text Foundation Models. TMLR. 2022.

RemoteCLIP：遥感领域视觉-语言基础模型

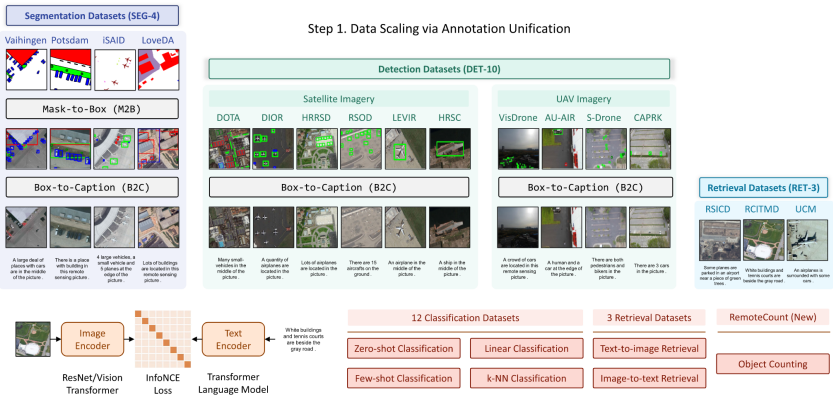


图: RemoteCLIP (Liu et al., 2024)：面向遥感的视觉-语言基础模型。

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

传统数据增强方法

- **几何变换：旋转**、**翻转**（水平/垂直）、**缩放**、**平移**、**裁剪**
- **颜色扰动：**调整亮度、对比度、饱和度和色调
- **噪声注入：**向图像中添加随机噪声
- **Cutout** (DeVries and Taylor, 2017)
- **CutMix** (Yun et al., 2019)
- **Copy-Paste** (Ghiasi et al., 2021)

还有一项系统性研究《How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers》(Steiner et al., 2022)。

DeVries, et al. Improved Regularization of Convolutional Neural Networks with Cutout, arXiv, 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.
Steiner, et al. How to Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers. TMLR. 2022.

用于数据增强的生成模型

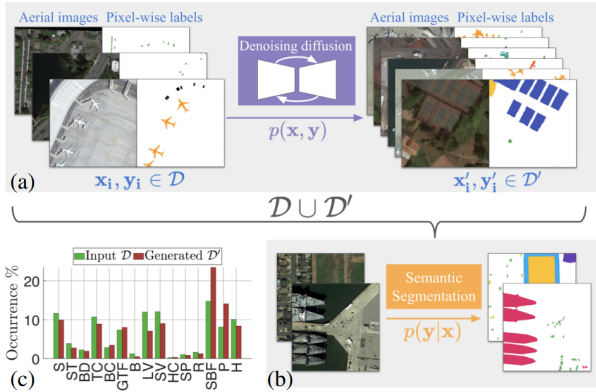


图: SatSyn (Toker et al., 2024) 提出了一种生成模型（扩散模型），可同时生成卫星分割的图像和对应掩码。该合成数据集用于数据增强，在卫星语义分割任务中相比其他增强方法带来了显著的定量提升。

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

生成的文本-图像数据集提升图像分类

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.

由生成模型合成的文本-图像数据集可显著提升图像分类性能，见 (He et al., 2023)。

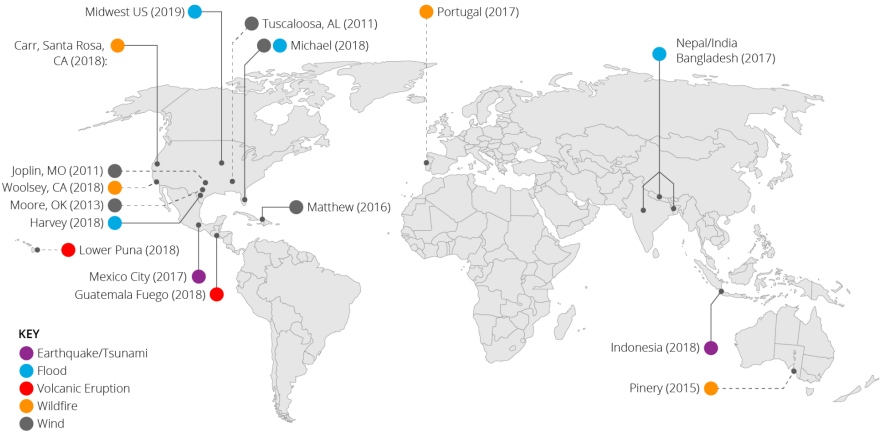
He, et al. Is synthetic data from generative models ready for image recognition?, ICLR, 2023.

xBD：大规模灾害损失数据集

灾前影像（上）与灾后影像（下）。从左到右依次为：哈维飓风、乔普林龙卷风、下普纳火山喷发、巽他海峡海啸。
影像来源：DigitalGlobe。
xBD (Gupta et al., 2019)

Gupta, et al. xBD: A Dataset for Assessing Building Damage from Satellite Imagery, arXiv, CVPRW, 2019.

xBD：全球灾害类型覆盖



xBD 数据集在全球范围内涵盖的灾害类型及事件。
xBD (Gupta et al., 2019)

Gupta, et al. xBD: A Dataset for Assessing Building Damage from Satellite Imagery, arXiv, CVPRW, 2019.

场景分类与超分辨率的基线模型

场景图像分类：

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

文本到图像生成的其他数据集

文本到图像生成：

- **RSICD** (Lu et al., 2018)：遥感图像描述数据集，包含 10,921 张图像，每张图像有 5 条描述。
- **RSICap** (Hu et al., 2025)：高质量数据集，包含 2,585 个人工标注的图像-文本对。
- **UCM-Captions** (Qu et al., 2016)：基于 UC Merced 土地利用数据集，包含 2,100 张图像，每张配有 5 条描述。

Lu, et al. Exploring Models and Data for Remote Sensing Image Caption Generation. TGRS. 2018.
Hu, et al. RSGPT: A remote sensing vision language model and benchmark. ISPRS. 2025.
Qu, et al. Deep semantic understanding of high resolution remote sensing image, CITS, 2016.

附录

生成建模

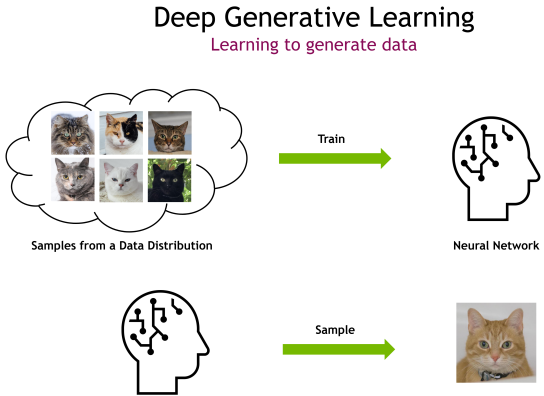


图: 生成建模示意图 (Vahdat，Arash, Song, and Meng, 2023)。

Vahdat，Arash et al. CVPR 2023 Tutorial Denoising Diffusion-based Generative Modeling: Foundations and Applications, 2023.

生成模型发展时间线

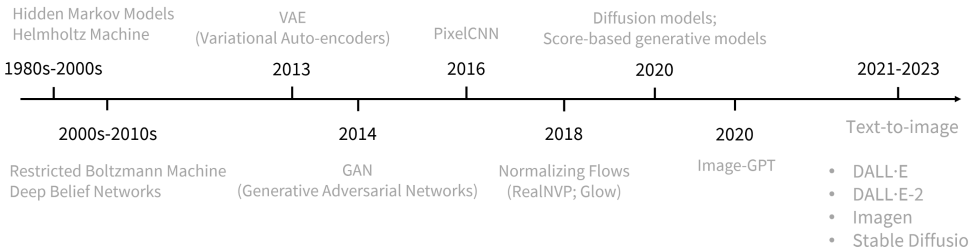


图: 生成模型关键发展历程时间线 (Deng, 2024)。

背景：扩散模型

去噪扩散模型包含两个过程：

- 正向扩散过程：逐步向输入添加噪声
- 反向去噪过程：通过去噪学习生成数据

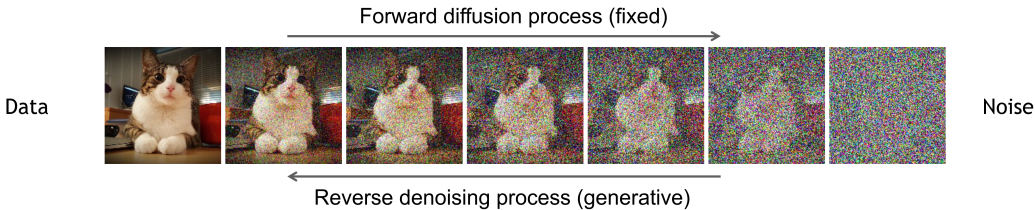


图: 扩散模型通过迭代去噪生成数据 (Sohl-Dickstein et al., 2015; Ho, Jain, and Abbeel, 2020)。

Sohl-Dickstein, et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML, 2015.
Ho, et al. Denoising Diffusion Probabilistic Models, NeurIPS, 2020.

扩散模型：正向与反向过程

正向（扩散）过程：

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

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

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

反向（去噪）过程：

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

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

其中 $\mathbf{x}_0$ 为原始数据，β_t 为噪声调度，$\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$。$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$。

扩散模型通过学习逆转逐步加噪过程来生成数据。(Sohl-Dickstein et al., 2015; Ho, Jain, and Abbeel, 2020)

Sohl-Dickstein, et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML, 2015.
Ho, et al. Denoising Diffusion Probabilistic Models, NeurIPS, 2020.

扩散模型：训练与推理

训练目标：

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

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

推理（采样）：

- 从纯噪声开始：$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- 对 $t = T, \ldots, 1$：
  - 预测噪声：$\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$
  - 计算均值：$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$
  - 采样：$\mathbf{x}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$
- 重复直到得到 $\mathbf{x}_0$（生成样本）

训练：最小化简化目标 (Ho, Jain, and Abbeel, 2020)。
推理：通过迭代去噪从随机噪声生成数据。

Ho, et al. Denoising Diffusion Probabilistic Models, NeurIPS, 2020.

遥感图像生成应用：Text2Earth

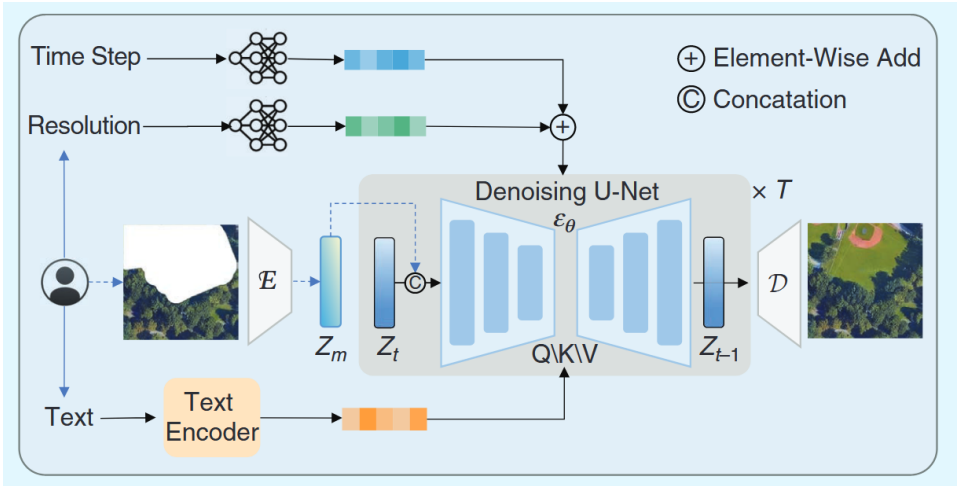


图: Text2Earth：面向文本驱动地球观测的基础模型 (Liu et al., 2025)。

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

Text2Earth：生成结果示例

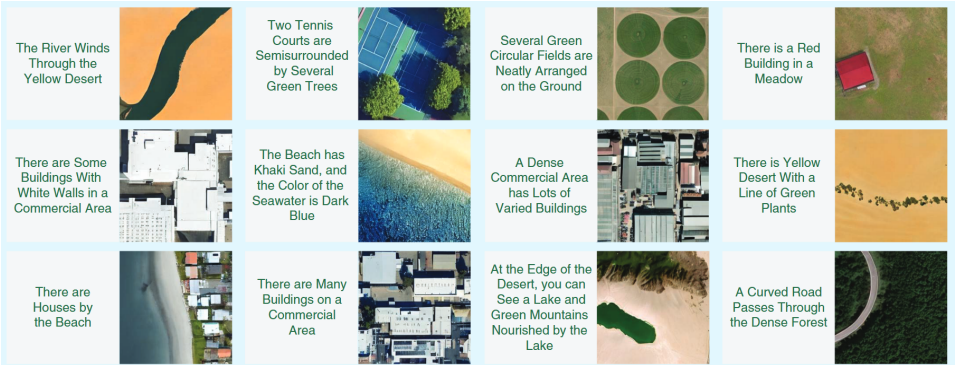


图: Text2Earth 生成的示例结果 (Liu et al., 2025)。

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

遥感图像生成应用：CRS-Diff

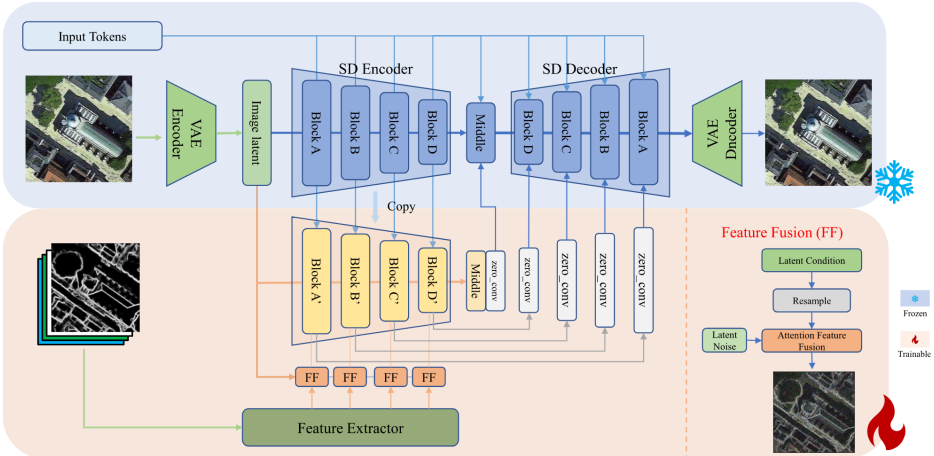


图: CRS-Diff：可控遥感图像生成框架 (Tang, Li, et al., 2024)。

Tang, et al. CRS-Diff: Controllable Remote Sensing Image Generation with Diffusion Model. TGRS. 2024.

CRS-Diff：生成结果示例

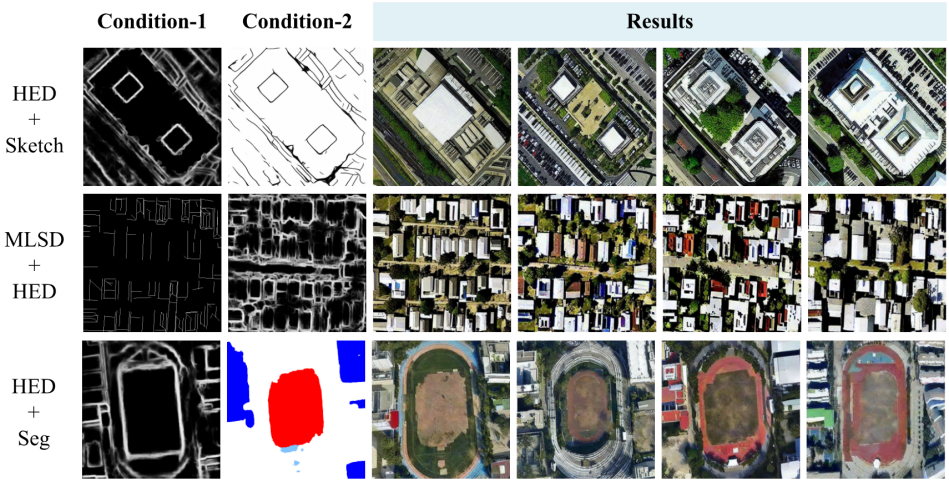


图: CRS-Diff 生成的示例结果 (Tang, Li, et al., 2024)。

Tang, et al. CRS-Diff: Controllable Remote Sensing Image Generation with Diffusion Model. TGRS. 2024.

DiffusionSat：框架概览

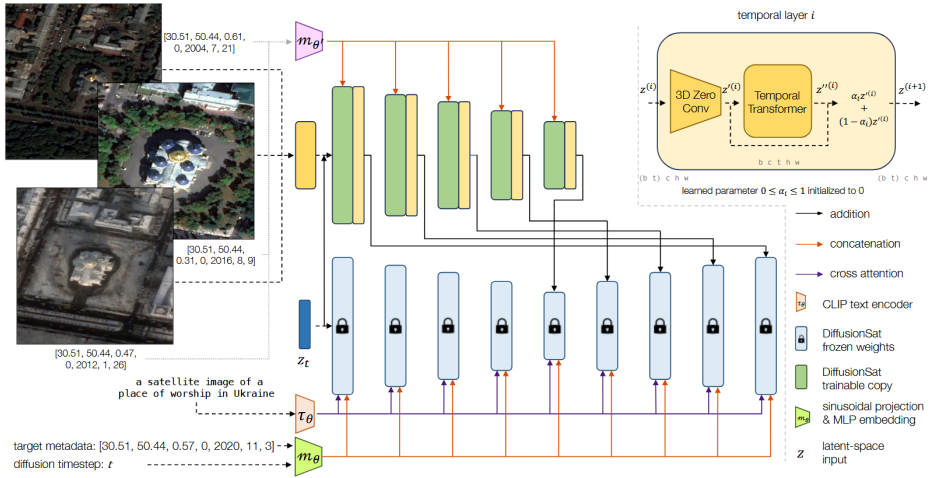


图: DiffusionSat：卫星影像生成基础模型 (Khanna et al., 2024)。

Khanna, et al. DiffusionSat: A Generative Foundation Model for Satellite Imagery, ICLR, 2024.

DiffusionSat：多光谱超分辨率结果

图: DiffusionSat 多光谱超分辨率示例结果 (Khanna et al., 2024)。

Khanna, et al. DiffusionSat: A Generative Foundation Model for Satellite Imagery, ICLR, 2024.

DiffusionSat：遥感图像修复结果

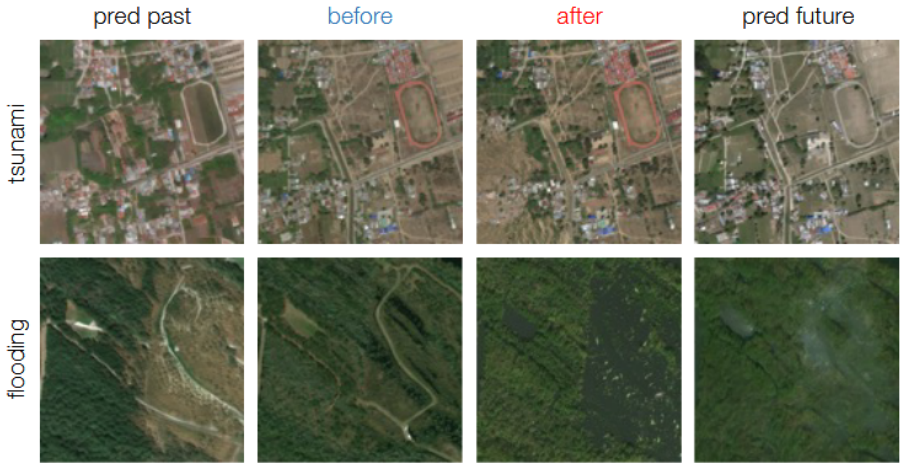


图: DiffusionSat 遥感图像修复示例结果 (Khanna et al., 2024)。

Khanna, et al. DiffusionSat: A Generative Foundation Model for Satellite Imagery, ICLR, 2024.