DiffusionSat: A Generative Foundation Model for Satellite Imagery

Michael Dushkoff

DiffusionSat

DiffusionSat – Background

- Satellite Imagery
  - Rarely contains labels
  - Metadata
    - Latitude
    - Longitude
    - Timestamp
    - Ground sampling distance (GSD)

Dataset	Image	Caption	Metadata
fmow		a fmow satellite image of a car dealership in United States of America	lon: -76.781 lat: 17.98 gsd: 0.941 cloud_cover: 0 year: 2010 month: 10 day: 6
satlas		a satlas satellite image of 26 ms buildings	lon: 78.995 lat: 85.048 gsd: 2 cloud_cover: 0 year: 2013 month: 6 day: 22
spacenet		a spacenet satellite image of 144 buildings covering an area of 9280.166 squared meters in Rio	lon: -43.636 lat: -22.892 gsd: 0.793 cloud_cover: 0 year: 1980 month: 0 day: 0

DiffusionSat – Background

- Generative Tasks:
  - Generate satellite image from numerical metadata
    - (Single image generation)
  - 3D control signal conditioning
    - Super resolution
    - Temporal prediction
    - In-painting

DiffusionSat – Background

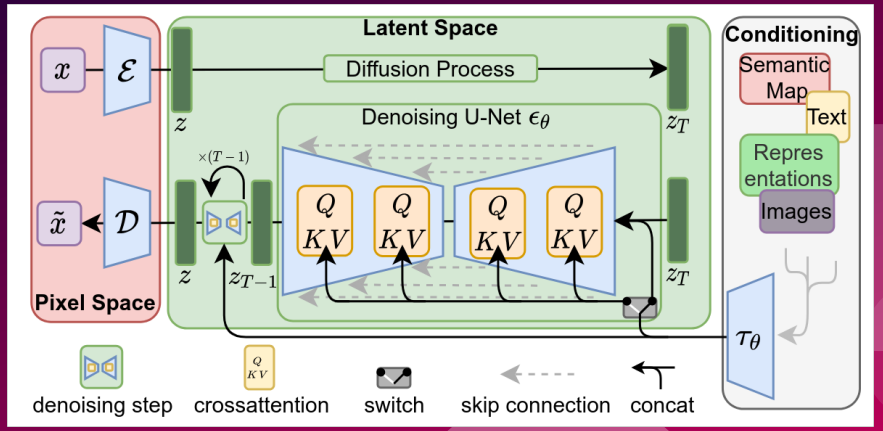
- **Diffusion models**
  - Learn data distribution from samples
  - Add noise progressively to an image, then denoise
  - Denoising model can be conditioned

DiffusionSat – Background

- **Diffusion models**
  - Forward Diffusion Process:
    - Noisy input: $x_t=\alpha_t x+\sigma_t \epsilon$
      - α_t, σ_t – Noise scheduling parameters over time *t*
      - $\epsilon \sim \mathcal{N}(0,I)$ – Gaussian distributed noise
  - Denoising process:
    - $E_{x\sim pData,\epsilon\sim\mathcal{N}(0,I)}[||y-\epsilon_\Theta(x_t;t,c)||^2]$
    - y - Target

DiffusionSat – Background

- Latent Diffusion Models (LDM)
  - Large input images = Big memory requirement
  - Perform diffusion in latent space instead
  - Variational Autoencoder (VAE)
    - Ɛ – Encoder
    - D – Decoder
  - Denoising UNet
    - Conditioned with τ_θ using cross-attention

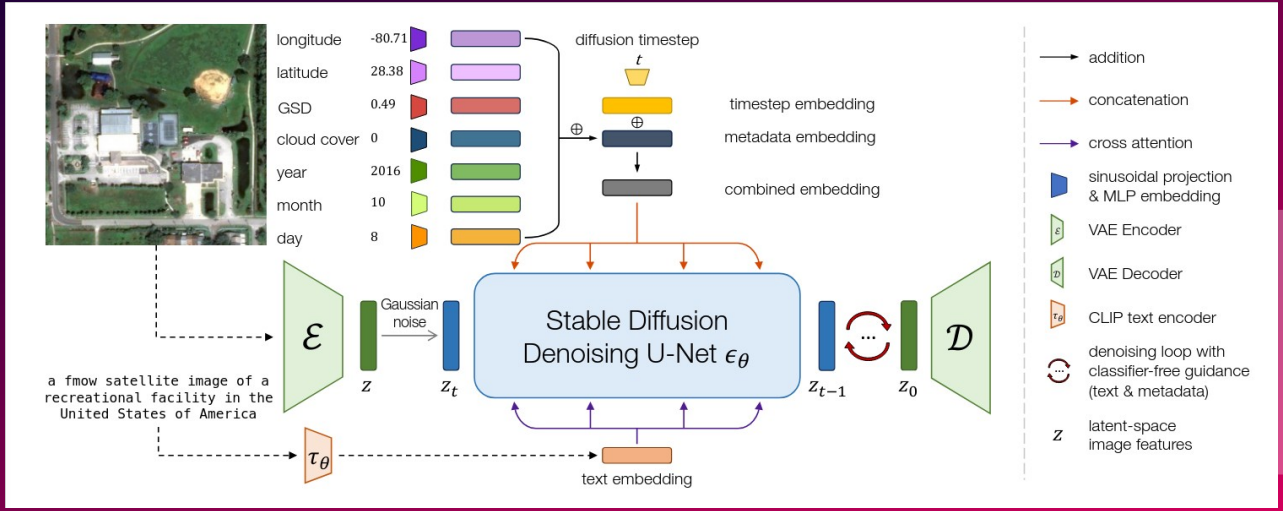


DiffusionSat – Problems

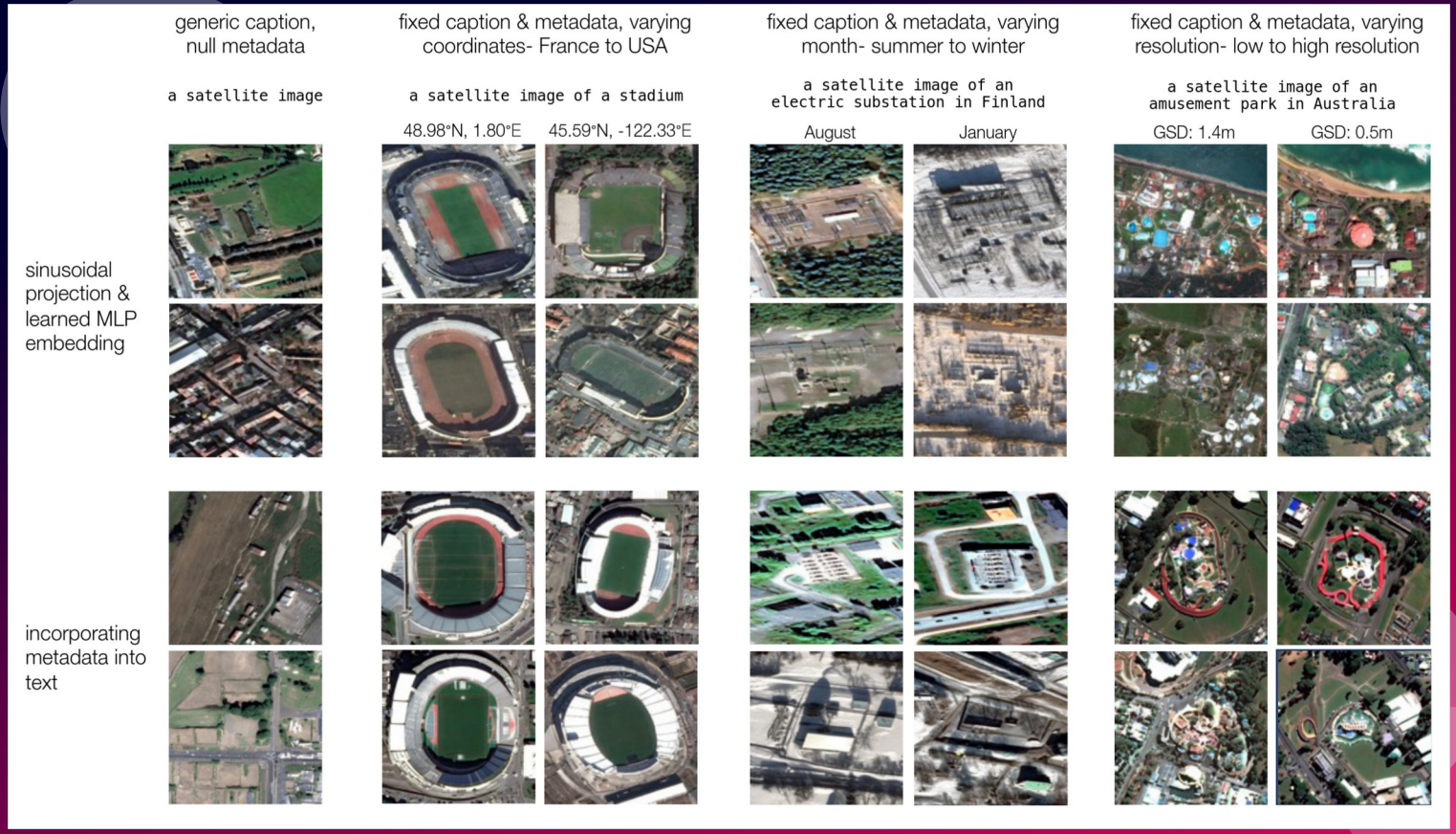
- Temporal Generation
  - Satellite images have long time variations
    - Can’t use 2D UNet with channel frame for time
    - Can’t use 3D convolution (long/inconsistent time differences)
  - Global time exists across different locations
    - Similar patterns at same point in time
    - Patterns conditioned on location

DiffusionSat – Architecture (Basic)

- Based on Stable Diffusion
  - Coordinates+time encoded using sinusoid projection & MLP embedding
  - Text metadata encoded with CLIP embedding

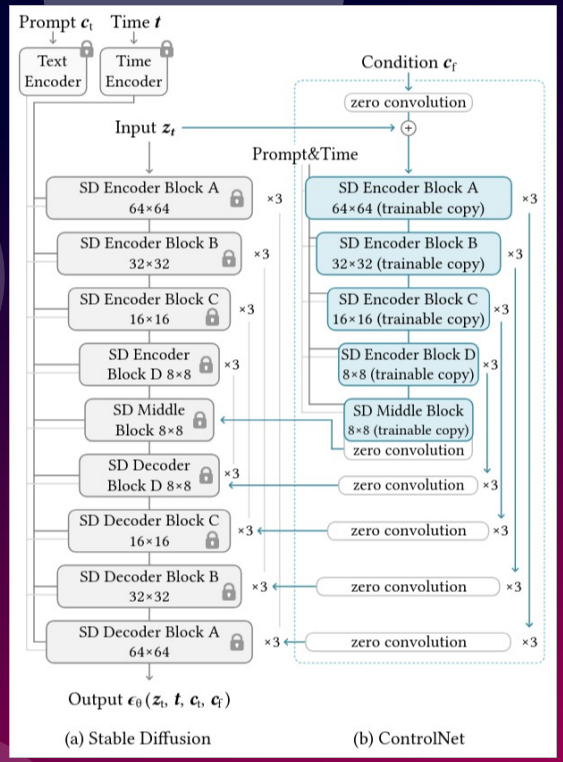


DiffusionSat – Single Image Generation Task

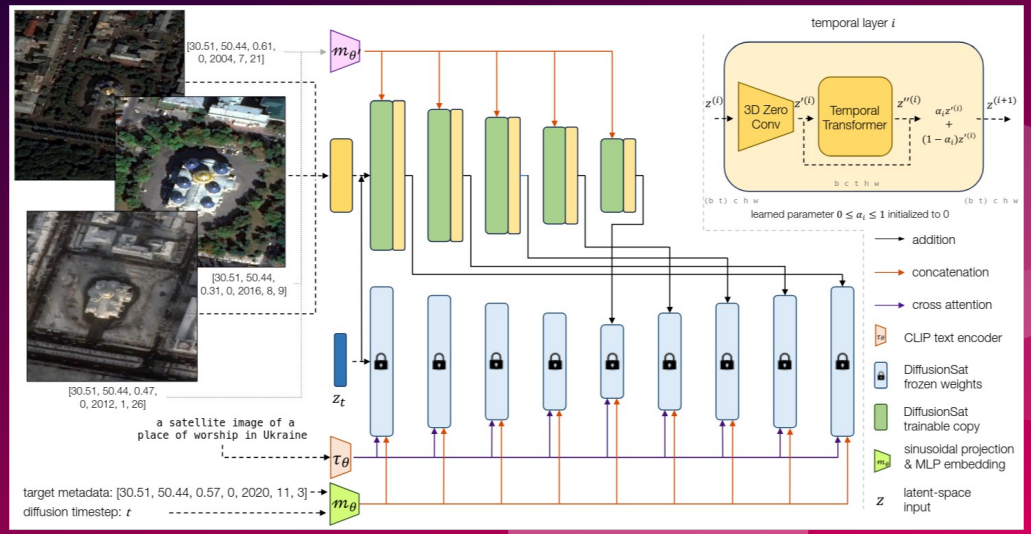


DiffusionSat – Architecture (Temporal)

- Traditional ControlNet



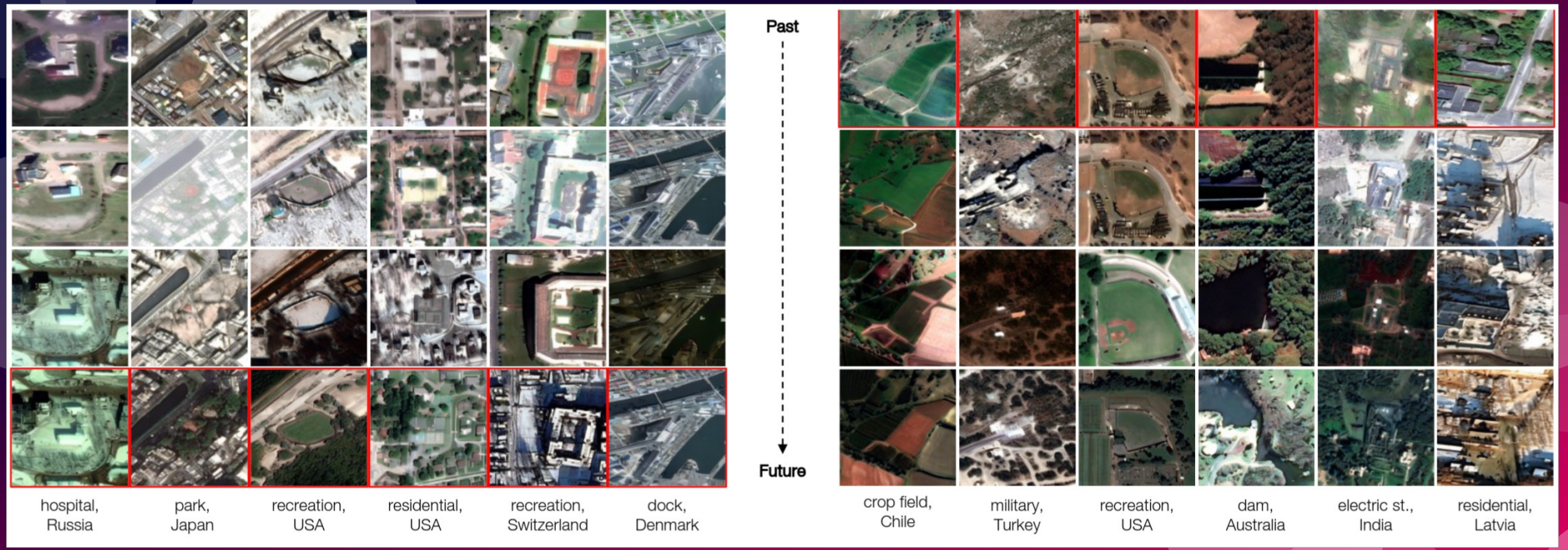
- DiffusionSat 3D ControlNet



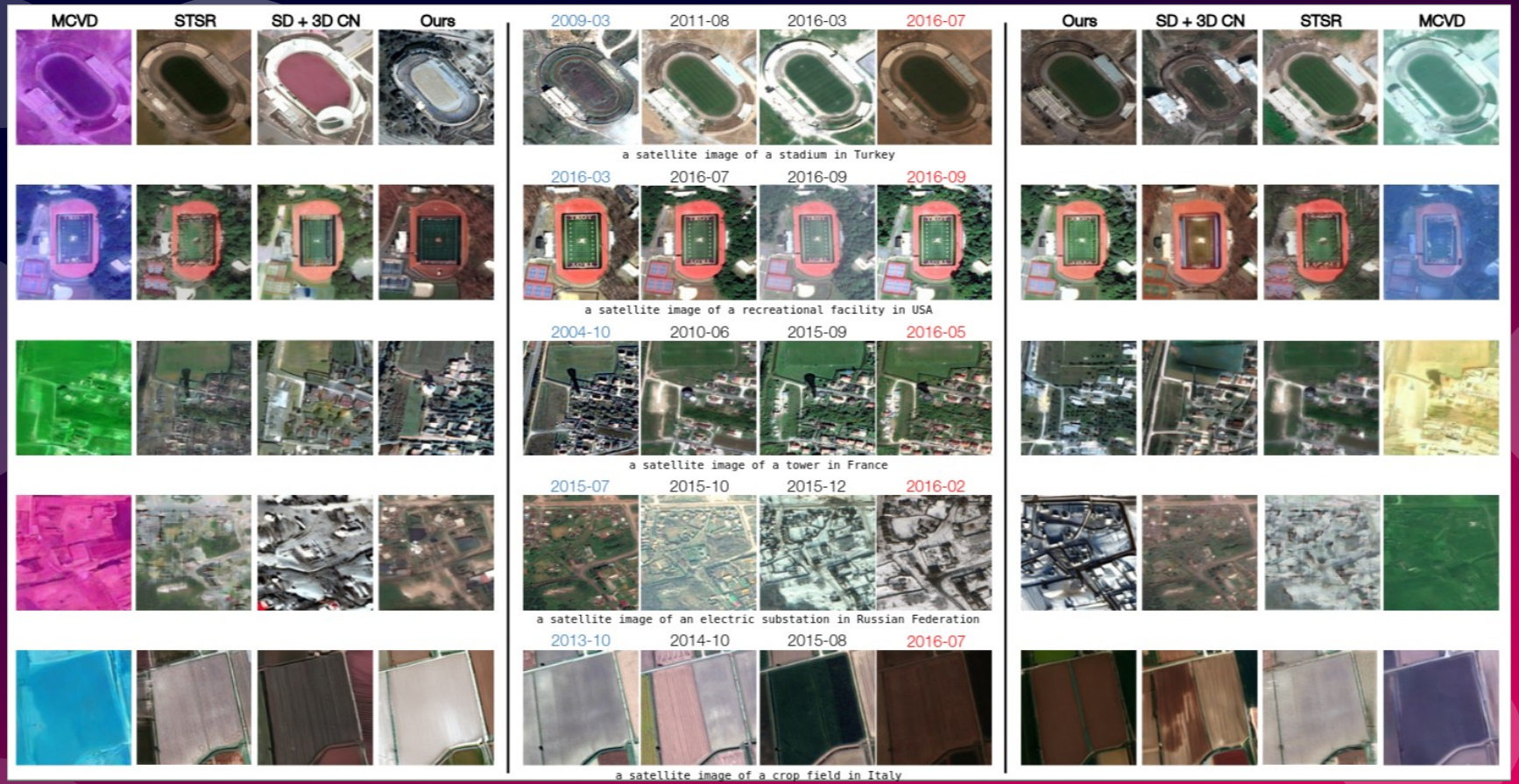
DiffusionSat – Architecture (Temporal)

- ControlNet
  - Injects additional conditions into the blocks of a Unet
  - Pre-trained model parameters are frozen
  - Copy of the model is set to trainable (encoder blocks)
    - Connected to the frozen model via zero convolutions (1x1 convolutions initialized to 0)
    - Prevents harmful noise form influencing hidden states
  - External conditioning vector *c* as input to the trainable model to guide its conditional distribution

DiffusionSat – Temporal Prediction



DiffusionSat – Temporal Prediction



DiffusionSat – Inpainting Task

DiffusionSat – Comparison

Model	$t' > t$			$t' < t$		
	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
STSR (EAD) (He et al., 2021)	0.3657	13.5191	0.4898	0.3654	13.7425	0.4940
MCVD (Voleti et al., 2022)	0.3110	9.6330	0.6058	0.2721	9.5559	0.6124
SD + 3D CN	0.2027	11.0536	0.5523	0.2218	11.3094	0.5342
DiffusionSat + CN	0.3297	13.6938	0.5062	0.2862	12.4990	0.5307
DiffusionSat + 3D CN	**0.3983**	**13.7886**	**0.4304**	**0.4293**	**14.8699**	**0.3937**

Table 4: Sample quality quantitative results on fMoW-temporal validation data. $t' > t$ represents generating an image in the past given a future image, and $t' < t$ is the task for generating a future image given a past image.

DiffusionSat – Key Takeaways

- 3D ControlNet allows for injection of prior information
  - Useful for relating data, steering the distribution based on more information
  - General technique for finer diffusion distribution tuning
    - Own research involves relating spatial data with instrumentation parameters of a microscope device
    - Leverage a similar ControlNet to take in this additional information