SatFlow: Generative Model based Framework for Producing High Resolution Gap Free Remote Sensing Imagery.

Anonymous Authors[1]

Abstract

Frequent, high-resolution remote sensing imagery is crucial for agricultural and environmental monitoring. Satellites from the Landsat collection offer detailed imagery at 30m resolution but with lower temporal frequency, whereas missions like MODIS and VIIRS provide daily coverage at coarser resolutions. Clouds and cloud shadows contaminate about 55% of the optical remote sensing observations, posing additional challenges. To address these challenges, we present SatFlow, a generative model based framework that fuses low-resolution MODIS imagery and Landsat observations to produce frequent, high-resolution, gap-free surface reflectance imagery. Our model, trained via Conditional Flow Matching, demonstrates better performance in generating imagery with preserved structural and spectral integrity. Cloud imputation is treated as an image inpainting task, where the model reconstructs cloud-contaminated pixels and fills gaps caused by scan lines during inference by leveraging the learned generative processes. Experimental results demonstrate the capability of our approach in reliably imputing cloud-covered regions. This capability is crucial for downstream applications such as crop phenolonogy tracking, environmental change detection etc.,

1. Introduction

High spatial and temporal resolution remote sensing imagery enables a wide range of agricultural and environmental monitoring applications, including phenology mapping, yield forecasting, and meteorological disaster prediction (Bolton et al., 2020; Gillespie et al., 2007; Huber et al., 2024). Optical remote sensing imagery provides rich spectral information with strong interpretability. The Landsat program, operational since 1972, provides decades of Earth observation data at 30 m spatial resolution, enabling detailed land surface monitoring over an extended period. However, infrequent revisit intervals (10-16 days) and data gaps caused by cloud cover during imaging and the Scan Line Corrector failure in Landsat 7 pose significant challenges to consistent monitoring (Zhu et al., 2012). Cloud contamination is of particular concern, affecting up to 55% of optical remote sensing observations over land globally (King et al., 2013), leading to substantial loss of clear-sky scenes and limiting subsequent image analysis. These issues are especially acute in agricultural regions, where landscapes are highly dynamic during growing season and high temporal frequency is critical for capturing rapid changes in vegetation growth and phenological transitions. On the other hand, the Moderate Resolution Imaging Spectroradiometer (MODIS) instruments aboard NASA's Terra (launched in 1999) and Aqua (launched in 2002) satellites provide near-daily global coverage at resolutions ranging from 250m to 1km (Xiong et al., 2009). While this temporal fidelity is ideal for tracking short-term changes, the coarse resolution is insufficient for capturing field-level agricultural details or fine-grained ecosystem processes. Nevertheless, the MODIS record, spanning over two decades, forms an invaluable resource for environmental applications, including forest cover change monitoring, urban expansion mapping, and wildfire impact assessment (Liu et al., 2024; Schneider et al., 2010). Integrating MODIS's rich temporal information with Landsat's fine spatial detail offers an opportunity to generate a spatiotemporally enhanced long-term dataset that can inform a broad range of land surface and environmental monitoring and modelling applications.

Several approaches have been investigated to achieve such spatiotemporal integration. Established fusion methods—such as the Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) (Gao et al., 2006), the SpatioTemporal Adaptive fusion of High-resolution satellite sensor Imagery (STAIR) (Zhu et al., 2010), and the Highly Integrated STARFM (HISTARFM) (Zhu et al., 2016)—blend temporally frequent but coarse imagery with sparse but fine-resolution observations. While these methods have demonstrated improvements, they often encounter challenges in

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

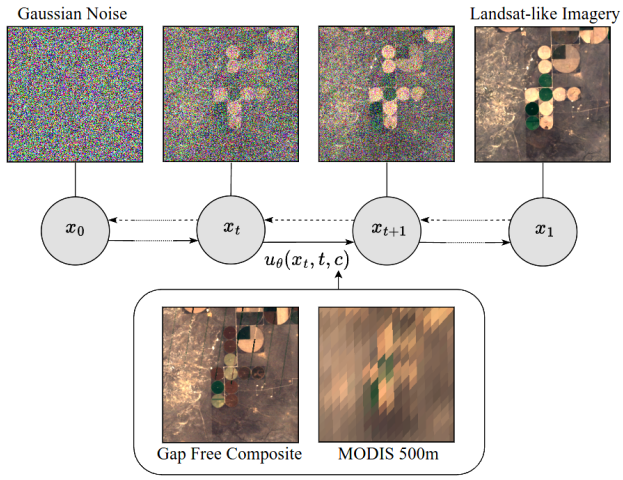

Figure 1. The framework integrates MODIS and Landsat observations through conditional flow matching to downscale MODIS imagery (500m) to Landsat resolution (30m).

heterogeneous landscapes and during periods of rapid land-cover change. STARFM and its variants are limited by the need to manually select one or more suitable pairs of coarse and high-resolution images for each fusion task, which poses challenges for automation at scale.

Advances in machine learning and deep generative models, including Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and diffusion-based approaches (Ho et al., 2020) have shown promise in image synthesis and super-resolution tasks (Wang et al., 2019; Lim et al., 2017; Rombach et al., 2022). While GANs can yield highly realistic imagery, they may suffer from training instability and spectral inconsistencies (Dhariwal & Nichol, 2021). Few works have applied generative models to remote sensing domain (Xiao et al., 2024; Khanna et al., 2024) and these typically require a large number of inference steps, as noted by Zou et.al.(2024). While Zou et.al. proposed an efficient diffusion approach for cloud imputation, it is limited to static landscapes and it can not be adapted to dynamic agricultural environments. Our novel framework integrates MODIS observations for contextual information while gap-filling high-resolution imagery. Beyond GANs and diffusion, our work utilizes Conditional flow matching (Lipman et al., 2023; Tong et al., 2024), a growing class of generative models that allow for exact likelihood estimation and often exhibit more stable training. The key contributions of our work are: (1) We present a novel approach for downscaling coarser-resolution MODIS imagery using a generative model to synthesize Landsat-like imagery. (2) We propose a gap-filling strategy that leverages the learned generative process to fill missing pixels in Landsat observations caused by cloud cover and scan lines. (3) We integrate the model into a pipeline to generate high-resolution, gap-free Landsat-like imagery at regular intervals.

2. Methodology

2.1. Flow Matching Formulation

The primary objective is to generate gap-free surface reflectance images given the conditioning factors, which include corresponding low-resolution MODIS imagery and a gap-free composite of previously acquired Landsat images. Our framework builds on conditional flow matching (Lipman et al., 2023; Tong et al., 2024), which generalizes continuous normalizing flows (Grathwohl et al., 2019; Chen et al., 2018) by directly regressing the vector fields for transforming between noise and data distributions. The goal of flow matching, similar to diffusion models (Ho et al., 2020; Rombach et al., 2022), is to generate samples that lie in the data distribution through an iterative process. We refer to the starting random gaussian noise distribution as $x(0)$ and the gap-free Landsat data distribution as $x(1)$, where the generative modeling task is to transform the initial noisy input x_0 to the target distribution x_1, through a learned process that is guided by the conditioning factors c (illustrated in Figure 2) .

2.2. Training

To learn a model that can transform $x(0)$ to $x(1)$, we model a time-varying vector field $u(t) : [0, 1] \times \mathbb{R}$, defined by the following ordinary differential equation: $u(t) = dx(t)/dt$. and a probability path $p(t) : [0, 1] \times \mathbb{R}$. Intuitively, this vector field defines the direction and magnitude by which to move a sample in x_0 so that it arrives at its corresponding location in x_1 by following the probability path p over time. We aim to approximate the true vector field u using a neural network $u_\theta(x_t, t, c)$, parameterized by weights θ. The flow matching objective is to minimize the difference between the predicted vector field $u_\theta(x_t, t, c)$ and the true vector field u_t, as expressed in Equation (1):

$$\min_\theta \mathbb{E}_{t, x_t \sim p(x_t|x_0,t)} \left[\|u_\theta(x_t, t, c) - u_t\|^2 \right]. \quad (1)$$

However, this objective is intractable as there is no closed form representation for the true vector field $u(t)$. Instead, similar to approaches that leverage simple linearized paths for training (Liu et al., 2023; Pooladian et al., 2023), we model the vector field u_t and the probability path $p : [0, 1] \times R$ between x_0 and x_1 with standard deviation σ as shown in Equations (2) and (3).

$$u(t) = x_1 - x_0 \quad (2)$$

$$x_t \sim \mathcal{N}((1 - t) \cdot x_0 + t \cdot x_1, \sigma^2) \quad (3)$$

Equation (3) defines the probability path as a Gaussian distribution centered at a linear interpolation between x_0 and x_1 at time t. Equation (2) defines the target vector field simply as the difference vector pointing from the starting point x_0 to the end point x_1. The training procedure to approximate this vector field is outlined in Algorithm 1.

Algorithm 1 Conditional Flow Matching Training

Require: initial parameters θ, learning rate α
1: **repeat**
2: Sample a batch of final states x_1, corresponding conditions c, initial states $x_0 \sim \mathcal{N}(0, I)$ and $t \sim [0, 1]$.
3: Compute the true vector fields: $u_t = x_1 - x_0$
4: Sample $x_t \sim \mathcal{N}\big((1-t) \cdot x_0 + t \cdot x_1,\ \sigma^2\big)$
5: Compute the loss:
$$L_{CFM}(\theta) = \tfrac{1}{2}\|u_\theta(x_t, t, c) - u_t\|^2$$
6: Update parameters: $\theta \leftarrow \theta - \alpha \nabla_\theta L_{CFM}(\theta)$
7: **until** converged

In algorithm 1, x_1 represents ground-truth Landsat imagery, while the conditioning factors c consist of two components: (1) MODIS observations acquired on the same date as x_1, providing coarse-resolution spectral information, and (2) a gap-free composite constructed from previously captured Landsat images of the same scene, providing high-resolution spatial context. Ideally, the model has to learn to synthesize Landsat-like high-resolution imagery by jointly leveraging the spatial structure from the composite and the spectral characteristics from MODIS data. To achieve this, we employ two key strategies during the training process: (1) MODIS inputs are randomly masked with a probability of 50 %, and (2) the gap-free composite is randomly selected from multiple available composites of the same scene (illustrated in Figure **??**). This augmentation approach encourages the model to disentangle and effectively utilize both information sources - learning to preserve spatial details from the composite while imparting the spectral information from MODIS observations when available. During inference, if MODIS observations are unavailable, the model generates plausible, unconditional spectral signatures, while still conforming to the spatial characteristics of the scene dictated by the Landsat composite.

2.3. Inference

To generate a sample from the target distribution x_1, given conditioning factors c, we integrate the learned vector field $u_\theta(x_t, t, c)$ over time. Specifically, starting from an initial sample $x_0 \sim \mathcal{N}(0, I)$, we follow the trajectory defined by Equation (4):

$$x_1 = x_0 + \int_0^1 u_\theta(x_t, t, c)dt \tag{4}$$

In practice, we approximate this integral using a discrete-time numerical scheme (Chen et al., 2018; Lipman et al., 2023). Forward Euler approach is employed for simplicity and computational efficiency as shown in Algorithm 2.

Algorithm 2 Conditional Flow Matching Inference

Require: conditions c, time step dt, initial $x_0 \sim \mathcal{N}(0, I)$,
1: **for** $t = 0$ to 1 **step** dt **do**
2: $x_{t+dt} = x_t + u_\theta(x_t, t, c) \cdot dt$
3: **end for**
output x_1

In Algorithm 2, starting from a random noise distribution, the model iteratively updates x_t using the vector field $u_\theta(x_t, t, c)$ to produce the final high-resolution Landsat-like imagery x_1. Clouds and scanlines imputation approach in our framework is inspired by the image inpainting methodology investigated by Lugmayr et al. (2022) in the context of diffusion models. Given clouds or scan line contaminated images and their corresponding quality assessment mask m (where $m_i = 1$ indicates cloudy/missing pixels and $m_i = 0$ indicates clear pixels), we introduce a composite update strategy that relies on the learned vector field $u_\theta(x_t, t, c)$ to reconstruct the unknown pixels and for the known pixels, uses a direct interpolation with the observed values as shown in Algorithm 3.

Algorithm 3 Cloud Imputation and Scan Lines Filling

Require: cloudy images x_1^*, cloud mask m, conditions c, time step dt, initial states $x_0 \sim \mathcal{N}(0, I)$,
1: Compute: $u = x_1^* - x_0$
2: **for** $t = 0$ to 1 **step** dt **do**
3: $x_{t+dt} = x_t + (u_\theta(x_t, t, c) \cdot m + u \cdot (1 - m)) \cdot dt$
4: **end for**
output x_1

This composite strategy ensures physical consistency by respecting the known data where available while leveraging the learned generative processes to fill in missing regions. Algorithms 2 and 3 demonstrate the versatility of the model in both generating high-resolution imagery and performing gap filling of the acquired imagery.

2.4. Model Architecture

The model architecture employs a U-Net (Ronneberger et al., 2015) design augmented with ResNet-style blocks (He et al., 2016) and self-attention layers (Vaswani et al., 2017; Bello et al., 2019), as illustrated in Figure 2. Conditioning information comprising MODIS observations and a gap-free Landsat composite - is concatenated along the channel dimension with the current state x_t. The current time step t

and ancillary metadata which includes day of year (DOY), sensor type (TM/OLI) and MODIS availability flag are encoded via learned embeddings and injected into the network at multiple resolutions. The network processes the inputs through a series of downsampling and upsampling stages linked by skip connections. Residual connections help stabilize training, and self-attention mechanisms capture both local and global dependencies. The network outputs the vector field $u_\theta(x_t, t, c)$ with six channels, corresponding to the multi-spectral dimensions of the Landsat data.

2.5. Overall Framework

We integrate the trained generative model into a pipeline to produce gap-free high resolution imagery at regular intervals. The framework processes two complementary data streams: daily MODIS imagery and Landsat observations (Landsat 5-9) with varying revisit times. The pipeline comprises of three key components: (1) Pre-processing of MODIS imagery: A temporal interpolation module that fills cloud-contaminated pixels in the MODIS time series using clear observations from adjacent days; (2) A gap-filling module that fills the clouds and scan-lines in the acquired Landsat scenes utilizing the trained model (Algorithm 3) (3) Finally, Landsat-like imagery are synthesized by the model at regular intervals by fusing the processed MODIS observations and gap-filled Landsat scenes. Since MODIS sensors (Aqua and Terra) acquire global imagery on a near-daily basis (as opposed to Landsat's 2–6 observations per month), temporal interpolation allows short gaps to be reconstructed with minimal discrepancy. Gap-filling module leverages spatial context from the Landsat composite and spectral information from temporally rich MODIS observations. The hierarchical design of the framework enables robust spatio-temporal fusion. A similar approach can be adapted to integrate other remote sensing data sources (e.g., VIIRS, Sentinel-2, SAR) for broader applicability.

3. Experiments

3.1. Dataset

The dataset for training the model was derived from Landsat and MODIS satellite imagery, spanning the period from years 2000 to 2024 across the Contiguous United States (CONUS). The years 2012 and 2015 were chosen to be excluded from the training set for validation, as these years represent contrasting dry and wet conditions respectively. The study utilized Level 2 processed surface reflectance data from Landsat 5, 7, 8, and 9 missions (Crawford et al., 2023) and the MODIS Bidirectional Reflectance Distribution Function (BRDF)-corrected MCD43A4 product (Schaaf et al., 2002; Lucht et al., 2000). The MCD43A4 product integrates data from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensors aboard the Aqua (launched in 2002) and Terra (launched in 1999) satellites, which observe Earth's surface at different times during the day (Link et al., 2017), providing daily global coverage at $500m$ resolution. Using stratified sampling, $\mathbf{20,000}$ locations were sampled across the contiguous United States based on the Cropland Data Layer (CDL) of year 2020 provided by USDA-NASS, covering diverse land cover and crop types. For each sampled location, imagery was obtained from four different dates where the cloud cover was less than 10% amounting to $\mathbf{80,000}$ data points for training. The dataset included: (1) Landsat surface reflectance imagery of size 256×256 pixels at $30m$ resolution, containing the six spectral bands (Red, Blue, Green, near-infrared (NIR), and two shortwave infrared bands (SWIR1 and SWIR2)); (2) corresponding MODIS imagery, resampled and aligned to match Landsat's spatial resolution; and (3) gap-free composites generated by stacking temporally preceding Landsat scenes. These composites were created by applying quality assessment masks to eliminate clouds, scan lines, and cloud shadows, followed by mosaicking the remaining clear pixels to ensure continuous spatial coverage. The data processing and collection workflow was implemented on Google Earth Engine. Prior to training, the reflectance values are normalized to lie within $[-1, 1]$ range using the scaling coefficients computed over the training set.

3.2. Setup

The model was trained to minimize the Mean Squared Error (MSE) loss between the predicted and target vector fields, following the procedure outlined in section 2.2. We adopted the AdamW optimizer (Loshchilov & Hutter, 2019) with a base learning rate of $1e\text{-}4$. A cosine learning rate schedule (Loshchilov & Hutter, 2017) with $6,000$ warmup steps was employed to improve convergence and mitigate potential instabilities during the early stages of training. Each training spans 120 epochs and was conducted on two NVIDIA RTX A6000 GPUs, each processing a batch of 16 images. We further applied gradient accumulation over 4 steps, effectively increasing the batch size without exceeding GPU memory limits. All training runs employed a standard deviation of $\sigma = 0.001$ in Algorithm 1 to define the probability path. Influence of alternative choices for standard deviation (σ) were not investigated in our work.

We validated our method on a dataset comprising 2,500 held-out scenes from 2012 and 2015. First, we evaluated downscaling quality (from 500m MODIS to 30m Landsat resolution), comparing the model's predicted Landsat-like outputs with the actual high-resolution Landsat images. we also trained the same model architecture with conditional diffusion methodology as outlined by Zou et al. (2024) to compare the performance with conditional flow matching. A sigmoid noise schedule rescaled to a zero terminal signal-to-noise ratio (SNR) is implemented for the diffusion model, as

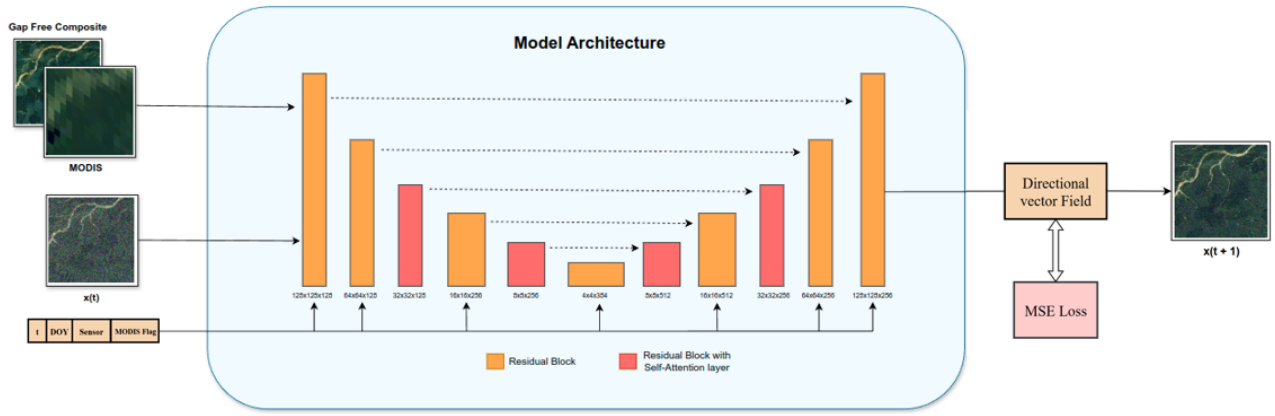

Figure 2. The Conditioning input are concatenated along the channel dimension with the current state x_t. The current time step t and metadata are encoded via learned embedding and integrated into the network at multiple resolutions. The network predicts the vector field $u_\theta(x_t, t, c)$ and MSE loss is computed between the predicted and target vector fields.

it demonstrated superior performance. To assess gap-filling performance by synthetically masking clean Landsat imagery with varying cloud coverage levels (10%–75%) using randomly generated cloud masks (Czerkawski et al., 2023). These artificial gaps are filled by the model as outlined in Algorithm 3, enabling direct comparisons against the known ground truth reflectance values. To our knowledge, no publicly available benchmarks are suitable for evaluating our data-fusion and cloud imputation approaches.

3.3. Evaluation Metrics

To assess the quality of the generated surface reflectance images, we employ the following metrics:

1. Spectral Information Divergence (SID): Spectral Information Divergence (Chang, 2000) is an information-theoretic metric introduced to measure discrepancies between two spectral signatures. In our evaluation, we compute SID across all six spectral bands (Red, Green, Blue, NIR, SWIR1, SWIR2) between the generated and original Landsat imagery. Lower SID values indicate that the reconstructed spectrum closely matches the reference. The SID between two spectral signatures p and q is given by:

$$SID(p,q) = \sum_{i=1}^{N} p_i \log\left(\frac{p_i}{q_i}\right) + \sum_{i=1}^{N} q_i \log\left(\frac{q_i}{p_i}\right) \quad (5)$$

where p_i and q_i represent the normalized reflectance values for band i in the generated and reference images respectively.

2. Structural Similarity Index Measure (SSIM): Structural Similarity Index Measure (Wang et al., 2004) is computed over local 11×11 pixel windows. For each window pair x and y:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (6)$$

where μ_x, μ_y are the mean intensities of windows x and y respectively, σ_x^2, σ_y^2 are their variances, and σ_{xy} is the covariance between the windows. The final SSIM score is obtained by averaging across all windows and RGB bands, higher values indicating greater structural similarity.

3. Peak Signal-to-Noise Ratio (PSNR): Peak Signal-to-Noise Ratio is useful for evaluating the pixel-wise accuracy, with a typical range of 20 to 40 dB for acceptable image reconstruction. PSNR is computed as:

$$PSNR = 10 \cdot \log_{10}\left(\frac{1}{MSE}\right) \quad (7)$$

where MSE is the mean squared error between the generated and reference normalized reflectance values, calculated across all six spectral bands.

3.4. Quantitative Comparisons

Table 1 summarizes the effect of the number of inference steps on the model performance. Notably, even with 3 inference steps the model achieves a decent baseline (SSIM = 0.738; SID = 0.039), illustrating the efficiency of linearized paths in Conditional Flow Matching. Performance steadily improves as the number of steps increases, with diminishing returns beyond 50 steps (SSIM: 0.912 at 50 steps vs.

0.908 at 100 steps). We thus select 50 steps to balance computational cost and accuracy.

Table 1. Performance Metrics vs. Number of Inference Steps

STEPS	1	3	5	10	50	100
SID	0.285	0.039	0.0216	0.0194	0.018	0.012
SSIM	0.651	0.738	0.862	0.895	0.912	0.908
PSNR	23.3	28.5	29.7	29.9	31.8	30.5

We evaluate our CFM approach against a conditional diffusion method (Zou et al., 2024) and a traditional remote sensing fusion baseline (STARFM). For comparision, we chose number of inference steps as 50 for both CFM and diffusion models. Table 2 shows that CFM outperforms alternatives in terms of SID, SSIM, and PSNR. These gains translate directly to higher-quality reconstructions in both downscaling (500m MODIS to 30m Landsat) and cloud gap-filling scenarios.

Table 2. Comparison with Baseline Methods on Held-Out Scenes

METHOD	SID	SSIM	PSNR
STARFM	0.0481	0.852	28.6
DIFFUSION	0.0271	0.891	30.0
CFM	**0.0186**	**0.912**	**31.8**

Lastly, we assess cloud imputation accuracy under different cloud coverage (10%, 25%, 50%, and 75%). Table 3 demonstrates the efficacy of multi-sensor fusion: adding MODIS consistently yields lower SID and higher SSIM. This advantage becomes more pronounced as cloud coverage increases. For instance, at 75% coverage, our method with MODIS exhibits lower SID and higher SSIM compared to the scenario without MODIS data, emphasizing the importance of leveraging coarse daily observations in heavily occluded conditions.

Table 3. Performance vs. Cloud Cover (%) With and Without MODIS Input

CLOUD COVER (%)	WITH MODIS		WITHOUT MODIS	
	SID	SSIM	SID	SSIM
10	0.015	0.960	0.028	0.932
25	0.032	0.921	0.056	0.884
50	0.068	0.875	0.098	0.821
75	0.071	0.812	0.167	0.723

Together, these findings indicate that (1) our approach offers a robust framework for combining data from multiple sensors, (2) linearized flows enable faster, more efficient inference,, and (3) incorporating MODIS observations in gap-filling further enhances resilience to occlusions by providing additional temporal and spectral context.

Figure 3. Example of artifacts introduced by Quality Assessment misclassification. The images on the left show the original cloudy Landsat image, and the images on the right show resulting artifacts in the gap-filled output.

4. Limitations

While our proposed framework demonstrates strong performance in generating gap-free daily Landsat-like imagery, there remain several important limitations. First, our gap-filling strategy (Algorithm 3) relies on a mask that distinguishes clear pixels from contaminated ones. In our work, we utilize the quality assessment masks provided by Landsat level-2 processed products. In practice, these masks are prone to misclassification—particularly at cloud edges or shadows—which can introduce artifacts in the reconstructed outputs. As shown in Figure 3, misclassifications can lead to visible artifacts and degraded image quality in the outputs. Advanced cloud and shadow detection algorithms could alleviate these artifacts. Second, preprocessing of daily MODIS imagery involves temporal interpolation for missing or cloudy observations. However, linear or spline interpolation will perform poorly in the presence of cloud cover over extended time periods and extreme events (e.g., wildfires, floods, or snowfall), which may feature abrupt spectral changes. In such scenarios, the reconstruction will not reflect the real-world conditions. Incorporating com-

plementary modalities, such as Sentinel-1 SAR (Synthetic Aperture Radar) data and multiple remote sensing sources, may mitigate this shortcoming. However availability of newer earth observation datasets (e.g., Sentinel missions) is limited to the years after 2015.

5. Conclusion and Future Work

We presented a Conditional Flow Matching (CFM) model that fuses daily coarse-resolution MODIS imagery with Landsat observations to generate gap-free, high-resolution surface reflectance data. We proposed integration of this model into a framework (SatFlow) to produce gap-free, Landsat-like imagery at regular intervals. This capability facilitates the generation of long-term remote sensing datasets, enhancing environmental monitoring and modeling applications. Our experimental results demonstrate that, particularly under high occlusion rates, the combined utilization of MODIS coarse data and Landsat composites allows reliable gap filling. In forthcoming work, we aim to extend the framework to include additional remote sensing sources such as Sentinel-2 optical imagery, VIIRS, and SAR data (Sentinel-1), aiming to further enhance robustness in cloudy or otherwise adverse conditions. We also plan to investigate efficient architectures derived from Vision Transformers (ViT) and Swin-UNet models, with the goal of achieving faster and better performing models capable of scaling to continental or global domains. Finally, we intend to quantify uncertainty in the generated reflectance maps, thereby providing reliability estimates for subsequent remote sensing analyses and decision-making.

Software and Data

The software and dataset will be made available upon completion of review process.

Acknowledgements

...

Impact Statement

This paper presents work whose goal is to enhance earth observation with advanced techniques in Machine Learning. While our work may have various societal implications, we do not identify any that require specific discussion here.

References

Bello, I., Zoph, B., Vasudevan, V., and Le, Q. V. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3286–3295, 2019.

Bolton, D. K., Gray, J. M., Melaas, E. K., Moon, M., Eklundh, L., and Friedl, M. A. Continental-scale land surface phenology from harmonized Landsat 8 and Sentinel-2 imagery. *Remote Sensing of Environment*, 240:111685, 2020. doi: 10.1016/j.rse.2020.111685.

Chang, C.-I. An information-theoretic-based approach to spectral variability, similarity, and discrimination for hyperspectral image analysis. *IEEE Transactions on Information Theory*, 46(5):1927–1932, 2000.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 6571–6583. Curran Associates, Inc., 2018.

Crawford, C. J., Roy, D. P., Arab, S., Barnes, C., Vermote, E., Hulley, G., Gerace, A., Choate, M., Engebretson, C., Micijevic, E., Schmidt, G., Anderson, C., Anderson, M., Bouchard, M., Cook, B., Dittmeier, R., Howard, D., Jenkerson, C., Kim, M., Kleyians, T., Maiersperger, T., Mueller, C., Neigh, C., Owen, L., Page, B., Pahlevan, N., Rengarajan, R., Roger, J.-C., Sayler, K., Scaramuzza, P., Skakun, S., Yan, L., Zhang, H. K., Zhu, Z., and Zahn, S. The 50-year Landsat collection 2 archive. *Science of Remote Sensing*, 8:100103, 2023. doi: 10.1016/j.srs.2023.100103.

Czerkawski, M., Atkinson, R., Michie, C., and Tachtatzis, C. SatelliteCloudGenerator: Controllable cloud and shadow synthesis for multi-spectral optical satellite images. *Remote Sensing*, 15(17):4138, 2023. doi: 10.3390/rs15174138.

Dhariwal, P. and Nichol, A. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Gao, F., Masek, J. G., Schwaller, M., and Hall, F. G. On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, 44(8):2207–2218, 2006.

Gillespie, T. W., Chu, J., Frankenberg, E., and Thomas, D. Assessment and prediction of natural hazards from satellite imagery. *Progress in Physical Geography*, 31(5):459–470, oct 2007. doi: 10.1177/0309133307083296.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., and Duvenaud, D. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Huber, F., Inderka, A., and Steinhage, V. Leveraging remote sensing data for yield prediction with deep transfer learning. *Sensors*, 24(3):770, jan 2024. doi: 10.3390/s24030770.

Khanna, S., Liu, P., Zhou, L., Meng, C., Rombach, R., Burke, M., Lobell, D. B., and Ermon, S. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=I5webNFDgQ.

King, M. D., Platnick, S., Menzel, W. P., Ackerman, S. A., and Hubanks, P. A. Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites. *IEEE Transactions on Geoscience and Remote Sensing*, 51(7):3826–3852, 2013. doi: 10.1109/TGRS.2012.2227333.

Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144, 2017. doi: 10.1109/CVPRW.2017.151.

Link, D., Wang, Z., Twedt, K. A., and Xiong, X. Status of the MODIS spatial and spectral characterization and performance after recent SRCA operational changes. In Butler, J. J., Xiong, X., and Gu, X. (eds.), *Earth Observing Systems XXII*, volume 10402, pp. 104022G. International Society for Optics and Photonics, SPIE, 2017. doi: 10.1117/12.2273053.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2303.08369*, 2023.

Liu, Y., Liu, R., Chen, J., Ju, W., Feng, Y., Jiang, C., Zhu, X., Xiao, X., and Gong, P. A global annual fractional tree cover dataset during 2000–2021 generated from realigned MODIS seasonal data. *Scientific Data*, 11:832, 2024. doi: 10.1038/s41597-024-03671-9.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Lucht, W., Schaaf, C. B., and Strahler, A. H. An algorithm for the retrieval of albedo from space using semiempirical BRDF models. *IEEE Transactions on Geoscience and Remote Sensing*, 38(2):977–998, 2000.

Lugmayr, M., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Pooladian, S., Chen, R. T. Q., Rubanova, Y., Polley, D., and Duvenaud, D. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2310.11779*, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.

Schaaf, C. B., Gao, F., Strahler, A. H., Lucht, W., Li, X., Tsang, T., Strugnell, N. C., Zhang, X., Jin, Y., Muller, J.-P., Lewis, P., Barnsley, M., Hobson, P., Disney, M., Roberts, G., Dunderdale, M., Doll, C., d'Entremont, R. P., Hu, B., Liang, S., and Privette, J. L. First operational BRDF, albedo and nadir reflectance products from MODIS. *Remote Sensing of Environment*, 83(1-2):135–148, 2002.

Schneider, A., Friedl, M. A., and Potere, D. Mapping global urban areas using MODIS 500-m data: New methods and datasets based on 'urban ecoregions'. *Remote Sensing of Environment*, 114(8):1733–1746, 2010. doi: 10.1016/j.rse.2010.03.003.

Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. Expert Certification.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Wang, H., Wu, W., Su, Y., Duan, Y., and Wang, P. Image super-resolution using a improved generative adversarial network. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pp. 312–315, 2019. doi: 10.1109/ICEIEC.2019.8784610.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Xiao, Y., Yuan, Q., Jiang, K., He, J., Jin, X., and Zhang, L. EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2023.3341437.

Xiong, X., Chiang, K., Sun, J., Barnes, W., Guenther, B., and Salomonson, V. V. NASA EOS Terra and Aqua MODIS on-orbit performance. *Advances in Space Research*, 43:413–422, feb 2009. doi: 10.1016/j.asr.2008.04.008.

Zhu, X., Chen, J., Gao, F., and Masek, J. G. Combining Landsat and MODIS for a cloud-free, consistent time series. *Remote Sensing of Environment*, 114(11):2623–2635, 2010.

Zhu, X., Liu, D., and Chen, J. A new geostatistical approach for filling gaps in Landsat ETM+ SLC-off images. *Remote Sensing of Environment*, 124:49–60, 2012. doi: 10.1016/j.rse.2012.04.019.

Zhu, X., Gao, F., Liu, Y., and Chen, J. Better monitoring of land cover dynamics using Landsat 8 time series. *Remote Sensing of Environment*, 190:233–241, 2016.

Zou, X., Li, K., Xing, J., Zhang, Y., Wang, S., Jin, L., and Tao, P. DiffCR: A fast conditional diffusion framework for cloud removal from optical satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3365806.