Spatiotemporal Alignment for Remote Sensing Image Recovery via Terrain-Aware Diffusion

Zhenyu Yu , Haoran Jiang , Pei Wang * , Zizhen Lin , Yong Xiang

Posted Date: 27 February 2026

doi: 10.20944/preprints202602.1915.v1

Keywords: remote sensing; diffusion model; DEM conditioning; terrain awareness; image inpainting

Article

Spatiotemporal Alignment for Remote Sensing Image Recovery via Terrain-Aware Diffusion

Zhenyu Yu [1], Haoran Jiang [2], Pei Wang [3,*], Zizhen Lin [4] and Yong Xiang [5]

[1] Foshan University
[2] Anhui University
[3] Kunming University of Science and Technology
[4] Kunming Institute of Physics
[5] Deakin University
* Correspondence: peiwang@kust.edu.cn

Abstract

Remote sensing imagery is essential for environmental monitoring but often suffers from large gaps due to clouds, sensor failure, or acquisition gaps. Existing interpolation and generative methods struggle to maintain spatial, spectral, and temporal coherence. We present *AlignDiff*, a diffusion-based framework that formulates reconstruction as a **spatiotemporal alignment problem**. It employs a **three-way strategy**: (1) **spatial alignment** via DEM conditioning, (2) **semantic alignment** through prompt-based modulation, and (3) **distributional alignment** with a VGG-Adapter enforcing feature-level consistency. Experiments on Landsat-8 and EarthNet2021 show that *AlignDiff* surpasses state-of-the-art baselines on spatial and temporal completion, enabling scalable, reliable satellite image recovery.

Keywords: remote sensing; diffusion model; DEM conditioning; terrain awareness; image inpainting

1. Introduction

Remote sensing imagery is a critical data source for environmental monitoring, agricultural planning, and disaster response [1,2]. However, satellite observations are frequently disrupted by cloud cover, sensor failure, or acquisition gaps [3,4], especially in high-resolution or high-frequency scenarios. These disruptions yield temporally and spatially incomplete datasets that reduce analysis accuracy, complicate multi-source fusion, and undermine reliable long-term monitoring. In climate-vulnerable and data-scarce regions, such deficiencies further exacerbate global inequities in access to environmental intelligence.

Traditional solutions, including interpolation and deep learning-based image completion, focus on pixel-level plausibility [5,6] but often fail to preserve structural, spectral, or temporal consistency, especially in semantically complex or geographically diverse regions [7]. Moreover, many approaches overlook physical priors such as terrain morphology or acquisition geometry and rely heavily on large labeled datasets, leading to reconstructions that diverge from geophysical reality and limiting scientific and operational utility.

To address these challenges, we reformulate satellite image reconstruction as a *spatiotemporal alignment problem*, aiming to infer content consistent with both spatial landscape and temporal dynamics. This perspective emphasizes context-aware reasoning and geophysically grounded alignment, moving beyond standard completion paradigms.

We present ***AlignDiff***, a diffusion-based framework that treats remote sensing image reconstruction as a spatiotemporal alignment task. *AlignDiff* integrates Digital Elevation Models (DEMs) as terrain priors and uses location–time prompts to guide generation, while a lightweight alignment adapter enforces distributional consistency between generated and observed regions. Our main **contributions** are:

 doi:10.20944/preprints202602.1915.v1

- Casting reconstruction as a unified **spatiotemporal alignment problem** for simultaneous spatial and temporal recovery;
- Proposing ***AlignDiff***, combining terrain priors and prompt-based conditioning for physically consistent completion;
- Designing an **alignment adapter** to improve generalization across terrains and missing patterns.

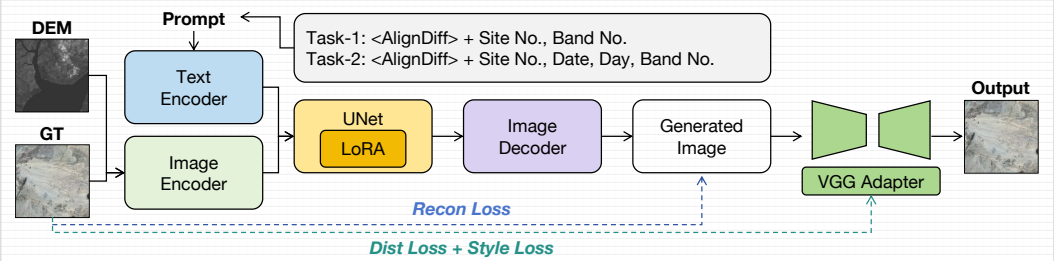

Figure 1. The *AlignDiff* pipeline integrates DEM-guided diffusion generation, LoRA-based prompt adaptation, and VGG-based distribution alignment for reconstructing missing satellite data across spatial and temporal gaps.

2. Proposed Framework: *AlignDiff*

We propose ***AlignDiff***, a unified diffusion-based generative framework that reconstructs missing satellite observations through three complementary alignment mechanisms. (1) **spatial alignment** via terrain-aware priors; (2) **distribution alignment** to enforce perceptual consistency; and (3) **semantic alignment** through prompt-conditioned generation.

2.1. Problem Formulation and Task Setting

Let $x \in \mathbb{R}^{C\times H\times W}$ be the complete image and x_{obs} its partially observed version with missing regions $\mathcal{M} \subset \Omega$. The goal is to generate $\hat{x}$ that fills $\mathcal{M}$ while preserving spatial, spectral, and contextual consistency. We learn the conditional distribution

$$p(\hat{x} \mid x_{\text{obs}}, a_{\text{spatial}}, a_{\text{dist}}, a_{\text{semantic}}) \tag{1}$$

where a_{spatial} encodes DEM-based terrain priors, a_{dist} enforces distributional alignment, and a_{semantic} provides prompt-based semantic conditioning.

We address two tasks: **Task-1 (Spatial Completion)**: reconstruct missing regions across spatially distinct tiles at a fixed time t; **Task-2 (Temporal Completion)**: recover observations at missing timepoints $\{t_k\}$ for a fixed location s using semantics and context.

2.2. Three-Way Alignment Strategy

To address the complex and heterogeneous nature of missing patterns in satellite imagery, *AlignDiff* integrates three alignment modules: spatial alignment with terrain priors, distribution alignment for perceptual fidelity, and semantic alignment via prompt-aware adaptation. Each module contributes complementary inductive biases, as detailed below.

2.2.1. Spatial Alignment with DEM Conditioning.

To incorporate physical topography and enforce terrain-aware realism, we adopt a ControlNet-style branch [8] that conditions the diffusion process on Digital Elevation Models (DEMs). This spatial alignment guides generation in mountainous, coastal, and varied terrain regions.

Let c_{spatial} denote the DEM-based conditioning input. The conditional denoising process is formulated as:

$$p(\hat{x} \mid c_{\text{spatial}}) = \prod_{t=1}^{T} p(x_t \mid x_{t-1}, c_{\text{spatial}}) \tag{2}$$

 doi:10.20944/preprints202602.1915.v1

where x_t denotes the latent image at timestep t. The spatial reconstruction is supervised via a pixel-wise MSE loss:

$$\mathcal{L}_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2 \tag{3}$$

2.2.2. Distribution Alignment with VGG Adapter.

To correct brightness imbalance, texture distortion, and spectral inconsistencies, we introduce a VGG-based adapter that aligns feature distributions between generated and ground-truth images.

Let $\phi(\cdot)$ denote features extracted from a pretrained VGG-19 network. We minimize the Maximum Mean Discrepancy (MMD) between the empirical feature distributions:

$$\mathcal{L}_{\text{dist}} = \|\mathbb{E}_x[\phi(x)] - \mathbb{E}_{\hat{x}}[\phi(\hat{x})]\|^2 \tag{4}$$

Additionally, we include a multi-level style loss to regularize fine-grained texture via Gram matrix alignment:

$$\mathcal{L}_{\text{style}} = \sum_{l=1}^{L} \frac{1}{4C_l^2 H_l^2 W_l^2} \left\| G_{\hat{x}}^l - G_x^l \right\|_F^2 \tag{5}$$

where G^l is the Gram matrix of VGG features at layer l.

2.2.3. Semantic Alignment with Prompt.

To condition the generation process on high-level semantics—such as location, seasonality, or band type—we introduce a lightweight adaptation module using Low-Rank Adaptation (LoRA) [9]. A task-specific token `<AlignDiff>` is prepended to the text encoder, enabling fine-grained prompt control across domains.

For each task, the injected prompt is structured as:

Task 1: `<AlignDiff>Region, Band`

Task 2: `<AlignDiff>Region, Date, Day, Band`

LoRA adapts the attention weight matrix $W \in \mathbb{R}^{d\times d}$ by injecting a low-rank perturbation:

$$W' = W + AB, \quad A \in \mathbb{R}^{d\times r}, \quad B \in \mathbb{R}^{r\times d} \tag{6}$$

This design ensures efficient parameter tuning while maintaining generalization and cross-domain adaptability.

2.3. Unified Loss Function and Training

To jointly optimize spatial realism, statistical fidelity, and semantic control, we formulate a unified training objective that balances all three alignment mechanisms. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{style}}\mathcal{L}_{\text{style}} \tag{7}$$

Here, λ_{dist} and λ_{style} are hyperparameters used to control the influence of perceptual distribution and style consistency. In practice, we set $\lambda_{\text{dist}} = 1.0$ and $\lambda_{\text{style}} = 100.0$ based on validation performance.

3. Experiment

3.1. Implementation Details

Dataset Description. **Task 1 – Landsat-8 Multi-Region.** We use Landsat-8 imagery and DEM data from 10 diverse regions covering various climates and land covers. Images (30 m resolution) are filtered in Google Earth Engine (cloud $< 1\%$) and tiled into 512×512 patches with 50% overlap. **Task 2 – EarthNet2021.** EarthNet2021 [10] offers 200k+ Sentinel-2 sequences for spatiotemporal modeling. Samples with $> 5\%$ invalid pixels are removed, and splits follow the official IID/OOD protocol. **Masking.** We adopt a unified masking strategy: (1) remove cloudy pixels, (2) randomly mask 10–50%

 doi:10.20944/preprints202602.1915.v1

valid pixels to simulate missing data, and (3) store indices for reproducible evaluation across both tasks.

Experimental Settings. Experiments ran on a single NVIDIA A100 (80 GB). The diffusion model was trained at 512×512 resolution with learning rate 5×10^{-5}. Inference used DDIM (50 steps, guidance 0.9, classifier-free scale 1.0, $\eta = 1.0$). STCNN baselines used batch size 16, learning rate 1×10^{-4}, 100 epochs. An 80:20 train-test split was applied for both tasks.

Evaluation Metrics. We evaluate *AlignDiff* using these metrics: Root Mean Square Error (RMSE) [11] and Mean Absolute Error (MAE) [12] for pixel-level accuracy, Peak Signal-to-Noise Ratio (PSNR) [13] and Structural Similarity Index (SSIM) [14] for image quality and structural similarity, and Learned Perceptual Image Patch Similarity (LPIPS) [15] for perceptual similarity.

3.2. Comparison with Existing Methods

Task-1: Spatial Completion across Diverse Regions. We compare with state-of-the-art methods including natural image inpainting models, unconditioned diffusion models, and terrain-aware diffusion baselines. **Palette** [16] and **LaMa** [17], originally designed for natural image inpainting, demonstrate limited transferability to the satellite domain. Despite LaMa showing improved robustness among the two, both methods fail to preserve geophysical structure in large or topographically varied missing areas. **Stable Diffusion (SD)** [18], used as an unconditioned diffusion baseline, achieves the highest SSIM (0.5402 ± 0.012) among all models. However, its low PSNR (17.1599 ± 0.84) and high RMSE (0.1448 ± 0.007) reflect blurred reconstructions and spectral distortion, due to the absence of terrain- or time-specific conditioning. **ControlNet** [8], adapted with DEM guidance, significantly improves generation quality, yielding PSNR of 21.1847 ± 0.72 and RMSE of 0.0873 ± 0.005. This demonstrates the value of integrating physical priors into the generation process.

AlignDiff improves upon ControlNet by introducing a VGG-based adapter for distribution alignment, achieving the best PSNR (23.04 ± 0.69), RMSE (0.0713 ± 0.004), and MAE (0.0500 ± 0.002). Despite slightly lower SSIM, qualitative results (Figure 2) show more coherent and terrain-consistent outputs, confirming the benefit of combining geospatial priors with perceptual alignment.

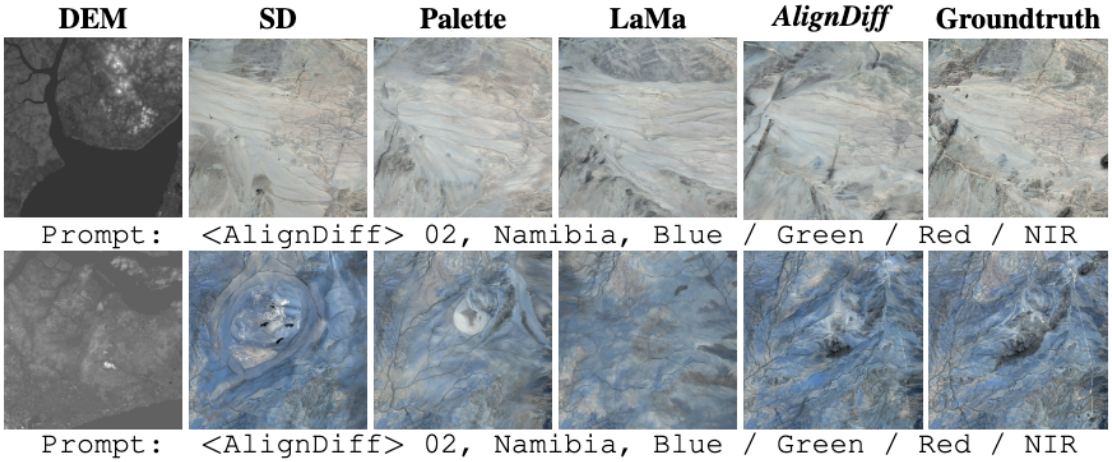

Figure 2. Comparison for Task-1, addressing missing data in specific regions over a fixed time period.

Task-2: Spatiotemporal Completion under Structured Constraints. In this setting, we evaluate the model's ability to restore missing observations across both spatial and temporal dimensions. We consider two reconstruction scenarios: (1) temporal prediction from prior frames, and (2) spatial reconstruction from DEM priors. The results are averaged across 100 globally sampled locations with varied seasonal and spectral conditions.

Under the temporal setting, interpolation-based methods exhibit poor performance, with PSNR below 12.0 and high RMSE above 0.25, failing to exploit semantic or physical consistency. STCNN improves moderately (PSNR=14.5317 ± 0.65), but still lacks high-frequency details. AutoEncoder [19]

 doi:10.20944/preprints202602.1915.v1

yields strong metrics (SSIM=0.6090 $\pm$ 0.013, PSNR=18.2038 $\pm$ 0.78), yet suffers from hallucination and texture mismatch in regions with significant temporal gaps.

In spatial reconstruction, diffusion models demonstrate superior stability. ControlNet achieves SSIM of 0.3787 $\pm$ 0.011 and PSNR of 22.7866 $\pm$ 0.73, substantially better than unconditioned models (SD: SSIM=0.2819 $\pm$ 0.012). However, ControlNet alone fails to ensure distributional consistency, often introducing spectral shifts and texture inconsistencies.

AlignDiff addresses these issues holistically by combining spatial priors, semantic prompts, and perceptual alignment. It outperforms all baselines in every metric—SSIM (0.5704 $\pm$ 0.011), PSNR (24.3429 $\pm$ 0.71), RMSE (0.0642$\pm$ 0.003), MAE (0.0479 $\pm$ 0.002), and LPIPS (0.0469 $\pm$ 0.002). Compared to ControlNet, this represents a 50.68% improvement in SSIM and 11.56% reduction in RMSE. The integration of alignment modules enables *AlignDiff* to handle complex missing patterns across seasons, land types, and spectral bands—making it a robust candidate for real-world satellite observation recovery.

3.3. Ablation Study

The proposed VGG-Adapter serves as a style alignment module to mitigate distributional discrepancies between generated outputs and reference data. As shown in Figure 4, omitting the adapter results in artifacts such as overexposed regions and brightness shifts, particularly in areas with complex land cover and elevation variability. Quantitatively, the mean brightness of reconstructed images *without* the adapter reaches $\mu = 123.63 \pm 6.02$, significantly deviating from the reference ($\mu = 95.19 \pm 3.87$). In contrast, *AlignDiff with VGG-Adapter* restores this value to $\mu = 95.27 \pm 3.82$, closely matching the ground truth distribution.

Table 1 (averaged over 100 sites) further corroborates this improvement across all metrics. Notably, the red band—which is highly sensitive to spectral shifts—shows the most significant gain in PSNR (+11.01 dB) and RMSE reduction (-0.12), as illustrated in Figure 3. These results confirm that the VGG-Adapter contributes both perceptual and statistical alignment, and plays a critical role in enhancing output consistency across diverse geographies.

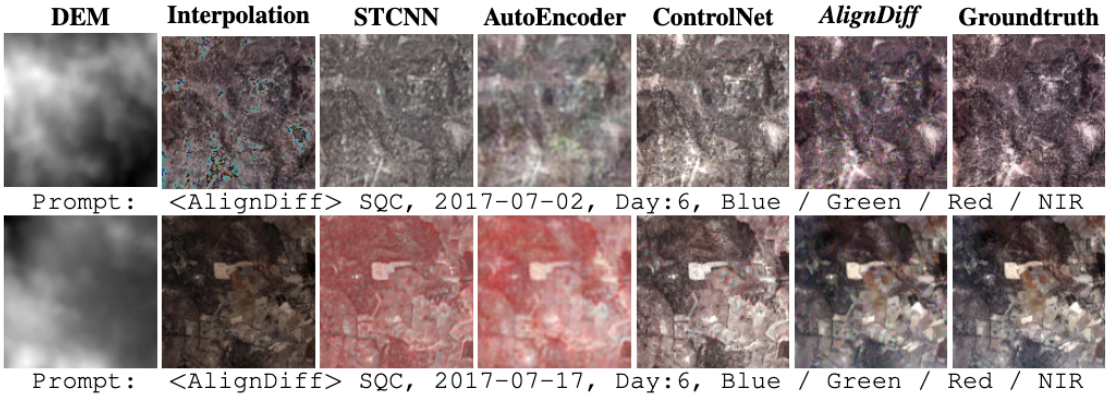

Figure 3. Comparison for Task-2 with selected missing data days. The above images are the results after brightness adjustment and gamma correction, with a coefficient of 1.2. The original image was used for calculating the evaluation metrics.

 doi:10.20944/preprints202602.1915.v1

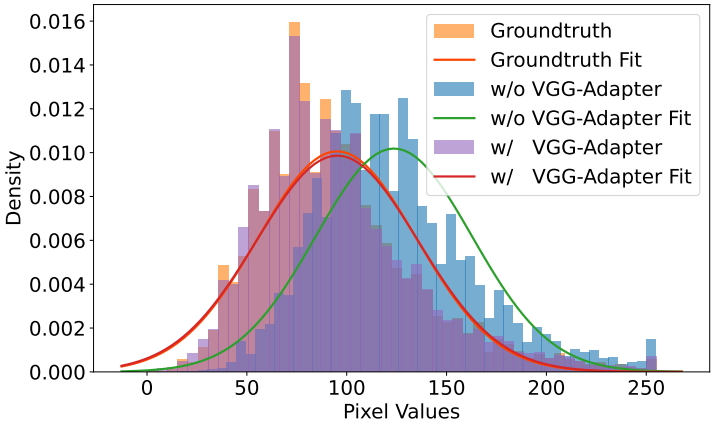

Figure 4. Ablation study on the VGG-Adapter.

Table 1. Ablation study on the VGG-Adapter module. **Bold** indicates the best result; Underline denotes the second-best.

Band	w/o VGG-Adapter					w/ VGG-Adapter				
	SSIM↑	PSNR↑	RMSE↓	MAE↓	LPIPS↓	SSIM↑	PSNR↑	RMSE↓	MAE↓	LPIPS↓
Blue	0.3842 ± 0.024	17.8211 ± 1.102	**0.1295 ± 0.012**	**0.1047 ± 0.009**	0.1244 ± 0.008	0.5662 ± 0.020	24.1039 ± 0.954	0.0662 ± 0.006	0.0489 ± 0.004	0.0469 ± 0.005
Green	**0.3861 ± 0.025**	17.7244 ± 1.031	0.1302 ± 0.013	0.1055 ± 0.010	0.1217 ± 0.007	0.5774 ± 0.018	**24.1578 ± 0.822**	**0.0647 ± 0.005**	**0.0479 ± 0.004**	0.0478 ± 0.006
Red	0.3613 ± 0.029	15.1034 ± 1.354	0.1759 ± 0.015	0.1598 ± 0.012	0.1369 ± 0.009	**0.5829 ± 0.019**	**26.4456 ± 1.021**	**0.0503 ± 0.005**	**0.0387 ± 0.003**	**0.0388 ± 0.004**
NIR	0.3771 ± 0.022	17.0143 ± 0.889	0.1407 ± 0.011	0.1173 ± 0.008	**0.1181 ± 0.006**	0.5524 ± 0.023	23.0387 ± 0.978	0.0741 ± 0.006	0.0555 ± 0.004	0.0511 ± 0.005
AVG	0.3772 ± 0.025	16.9160 ± 0.874	0.1441 ± 0.012	0.1218 ± 0.009	0.1253 ± 0.007	0.5697 ± 0.020	24.4365 ± 0.844	0.0638 ± 0.006	0.0478 ± 0.004	0.0462 ± 0.005

Table 2. Performance comparison on **Task-1**. **Bold** indicates the best, underline the second-best result.

Method	SSIM↑	PSNR↑	RMSE↓	MAE↓	LPIPS↓
SD	**0.5412 ± 0.023**	17.2635 ± 1.084	0.1452 ± 0.011	0.0916 ± 0.009	0.3378 ± 0.013
Palette	0.4350 ± 0.025	18.7321 ± 1.021	0.1218 ± 0.009	0.0821 ± 0.007	0.3424 ± 0.014
LaMa	0.4924 ± 0.018	20.8812 ± 0.994	0.0915 ± 0.008	0.0641 ± 0.006	0.2905 ± 0.012
ControlNet	0.4898 ± 0.020	21.2276 ± 0.935	0.0871 ± 0.007	0.0589 ± 0.005	**0.2872 ± 0.011**
AlignDiff	0.4563 ± 0.017	**23.1072 ± 0.879**	**0.0708 ± 0.006**	**0.0496 ± 0.004**	0.3417 ± 0.010

Table 3. Performance comparison on **Task-2**. The upper part uses the timestep as input; the lower uses DEM guidance. **Bold** denotes best, underline second-best.

Method	SSIM↑	PSNR↑	RMSE↓	MAE↓	LPIPS↓
Interpolation	0.5254 ± 0.028	11.9225 ± 0.845	0.2534 ± 0.022	0.2023 ± 0.019	0.3380 ± 0.017
STCNN	0.4047 ± 0.021	14.5317 ± 0.779	0.1877 ± 0.015	0.1498 ± 0.012	0.6835 ± 0.026
Autoencoder	**0.6090 ± 0.030**	**18.2038 ± 0.935**	**0.1230 ± 0.010**	**0.0981 ± 0.009**	**0.2487 ± 0.014**
STCNN	0.2232 ± 0.020	13.9769 ± 0.788	0.2107 ± 0.017	0.1723 ± 0.014	0.4208 ± 0.021
Autoencoder	0.2514 ± 0.024	14.7913 ± 0.801	0.2040 ± 0.016	0.1684 ± 0.013	0.3073 ± 0.019
SD	0.2819 ± 0.022	16.1943 ± 0.832	0.1550 ± 0.013	0.1335 ± 0.011	0.1468 ± 0.012
ControlNet	0.3787 ± 0.019	22.7866 ± 0.741	0.0726 ± 0.006	0.0570 ± 0.005	0.0721 ± 0.006
AlignDiff	**0.5704 ± 0.018**	**24.3429 ± 0.695**	**0.0642 ± 0.005**	**0.0479 ± 0.004**	**0.0469 ± 0.005**

4. Conclusion

We propose *AlignDiff*, a diffusion-based framework for remote sensing image reconstruction under missing data. It addresses both spatial completion and temporal recovery by leveraging Digital Elevation Models (DEMs) as structural priors for terrain-consistent generation. A VGG-Adapter mitigates distributional shifts via perceptual alignment, improving realism and spectral fidelity. Experiments on global Landsat-8 imagery and the EarthNet2021 benchmark show consistent gains over Stable Diffusion, ControlNet, and AutoEncoder across SSIM, PSNR, RMSE, MAE, and

 doi:10.20944/preprints202602.1915.v1

LPIPS. *AlignDiff* provides a scalable, generalizable, and geophysically grounded solution for satellite image recovery, especially in persistently data-scarce regions.

Acknowledgments: This work was supported in part by the Australian Research Council under grant DP250102634.

References

1. Yu, Z.; IDRIS, M.Y.I.; Wang, P.; Qureshi, R. CoTextor: Training-Free Modular Multilingual Text Editing via Layered Disentanglement and Depth-Aware Fusion. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems Creative AI Track: Humanity, 2025.
2. Yu, Z.; Idris, M.Y.I.; Wang, P.; Xia, Y.; Xiang, Y. Forgetme: Benchmarking the selective forgetting capabilities of generative models. *Engineering Applications of Artificial Intelligence* **2025**, *161*, 112087.
3. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing* **2016**, *114*, 24–31.
4. Yu, Z.; Idris, M.Y.I.; Wang, P. Visualizing Our Changing Earth: A Creative AI Framework for Democratizing Environmental Storytelling Through Satellite Imagery. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems Creative AI Track: Humanity, 2025.
5. Li, Z.L.; Wu, H.; Duan, S.B.; Zhao, W.; Ren, H.; Liu, X.; Leng, P.; Tang, R.; Ye, X.; Zhu, J.; et al. Satellite remote sensing of global land surface temperature: Definition, methods, products, and applications. *Reviews of Geophysics* **2023**, *61*.
6. Zhang, W.; Wang, Y.; Ni, B.; Yang, X. Fully context-aware image inpainting with a learned semantic pyramid. *Pattern Recognition* **2023**, *143*, 109741.
7. Gui, S.; Song, S.; Qin, R.; Tang, Y. Remote sensing object detection in the deep learning era—a review. *Remote Sensing* **2024**, *16*, 327.
8. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3836–3847.
9. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* **2021**.
10. Requena-Mesa, C.; Benson, V.; Reichstein, M.; Runge, J.; Denzler, J. EarthNet2021: A large-scale dataset and challenge for Earth surface forecasting as a guided video prediction task. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1132–1142.
11. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research* **2005**, *30*, 79–82.
12. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development* **2014**, *7*, 1247–1250.
13. Hore, A.; Ziou, D. Image quality metrics: PSNR vs. SSIM. In Proceedings of the 2010 20th international conference on pattern recognition. IEEE, 2010, pp. 2366–2369.
14. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **2004**, *13*, 600–612.
15. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.
16. Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; Norouzi, M. Palette: Image-to-image diffusion models. In Proceedings of the ACM SIGGRAPH 2022 conference proceedings, 2022, pp. 1–10.
17. Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; Lempitsky, V. Resolution-robust large mask inpainting with fourier convolutions. In Proceedings of the Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 2149–2159.
18. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
19. Rifai, S.; Vincent, P.; Muller, X.; Glorot, X.; Bengio, Y. Contractive auto-encoders: Explicit invariance during feature extraction. In Proceedings of the Proceedings of the 28th international conference on international conference on machine learning, 2011, pp. 833–840.

 doi:10.20944/preprints202602.1915.v1