Multi-Temporal Mean-Reverting SDE with Preconditioned Diffusion for Remote Sensing Cloud Removal

Zilv Liu[1], Jianfang Shen[2*] and Ziqi Huang[3]

[1,2,3] Guangdong University of Technology, Guangzhou 510006, China
tysjf@gdut.edu.cn

Abstract. The presence of clouds is a long-standing source of data loss in the remote sensing (RS) images and it reduces the accuracy of subsequent Earth observation processes. Recently, diffusion models and stochastic differential equations (SDEs) have been considered as a generative restoration method, but two issues are still not addressed in multi-temporal RS applications: 1) the formulation does not take into account the spatio-temporal redundancy and cross-modal clues already available in multi-sensor acquisitions, and 2) the convergence of the model is unstable when noise level changes significantly in heavily occluded areas. We suggest a Preconditioned Mean-Reverting SDE (PMR-SDE) that will be used to solve cloud removal problems in multi-temporal situations. Rather than considering the restoration process as a generic inversion of generation, we consider it as a mean-reversing process with a drift target based on the input rather than a fixed statistic of the dataset. Specifically, the reverse path is initiated by a cloudy state towards an uncloudy reference which is determined by the Sentinel-1 SAR features that can penetrate the cloud, giving the structure guidance even where the optical signal is missing. This conditioning is implemented using a Spatio-Temporal Dual-Branch Attention (ST-DBA) module that combines non-uniform time series of optical data with SAR backscatter using parallel attention streams in space and time. To overcome this problem of instability in gradient due to vanishing ratio of signal-to-noise, we also construct Instance-Adaptive Preconditioning Control (APC) mechanism which scales the denoiser inputs/outputs dynamics according to local estimates of priors. PMR-SDE has competitive performance in terms of recovery of spatial texture and spectral fidelity on the SEN12MS-CR-TS benchmark, and it has a better effect under heavy cloud cover.

Keywords: Remote sensing cloud removal, Diffusion model, Stochastic differential equation

1 Introduction

Optical remote sensing has been used as a tool in the Earth observation process, such as land-cover mapping and disaster response [1,2]. In practice though, it is not possible to avoid cloud and aerosol obstruction of satellite-mounted optical sensors (the most common examples are Sentinel-2 and Landsat). Clouds reflect the light very much at both visible and near-infrared bands, hence the signal that is received by the sensor on the ground is either significantly reduced or lost. This results in holes in the data, which have detrimental effects on the automated interpretation system. The main challenge in preprocessing of RS images is how to eliminate the interference of clouds without distorting the geometry and spectral signature of the terrain underneath.

The initial attempt to eliminate clouds was based on the concept of atmospheric scattering that predicts transmission maps and ambient light by using physical concepts [1,3]. The Dark Channel Prior (DCP) [2] provided an effective method of calculating the transmission matrices in dehazing and later developments have been applied to remote sensing sensors [4,5]. Physics-based methods are weak because they depend on a priori statistics of scenes, such as the presence of one local patch with almost zero intensity. These assumptions fail when it comes to bright man-made objects in urban areas or when there is absolutely no cloud cover at all, and may lead to apparent color distortions and distortion of geometry. The problem is exacerbated by high resolution images where surface reflectance data is not stationary.

In the recent past, diffusion models [6] and their continuous-time SDE formulation [7,8] have taken hold in image restoration such as remote sensing cloud removal [9,10]. There is a strand of parallel research that investigates Ornstein-Uhlenbeck (mean-reverting) processes to regularize the generative path, where EMRDM [10] is a representative example used in RS dehazing. These methods all share the common design pattern of fixing the reversion target at a constant amount- either the empirical mean of clean images across the entire dataset or a baseline clear reference which does not respond to the input scene in particular. This is fairly effective with light or moderate cover of clouds, but when optical observations are significantly blocked, absence of sample-specific structural constraints enables the reverse process to wander and generate textures that may appear plausible but geometrically or spectrally inconsistent with the actual land surface. Table 1 outlines the major differences between our approach and EMRDM.

The main point of this paper is to substitute the fixed reversion target with a dynamic one, which results in Preconditioned Mean-Reverting SDE (PMR-SDE) of multi-temporal cloud removal. The difference lies in the drift term. EMRDM [10] has a forward SDE of the form $dx= \kappa(\mu - \mathrm{x})dt + \mathrm{g}(t)\, dW$, where is an empirical mean that is spatially uniform and computed on the training set- a quantity that does not contain any per-sample geometric information. PMR-SDE replaces μ with a function$\Phi(X, x_{SAR})$ that depends on the input multi-temporal optical stack X and a co-registered Sentinel-1 SAR image x_SAR, giving $dx = \kappa(\Phi(X, x_{SAR}) - x)dt +$

$g(t)dW$. Since SAR pulses can penetrate clouds, x_SAR offers a structural reference that is still reliable when the optical channels are completely blocked. The function is not hand-made but it is learned by a Spatio-Temporal Dual-Branch Attention (ST-DBA) module which combines irregular optical time series with the SAR anchor using parallel spatial and temporal attention branches to align the heterogeneous feature spaces of the two modalities. We further propose an Instance-Adaptive Preconditioning Control (APC) scheme that scales the input-output dynamics of the denoiser according to locally estimated priors. This stabilizes training when the effective signal-to-noise ratio becomes zero during the later steps of diffusion, a regime in which conventional preconditioning tends to lead to gradient fluctuation and color artifacts.

Our main contributions are:

(1) A continuous-time cloud removal model based on a cross-modally conditioned mean-reverting SDE. By binding the drift target to input-specific SAR-optical fusion, instead of a dataset-wide average, one can suppress hallucinations and retain geographic authenticity [7,8].

(2) Spatio - Temporal Dual-Branch Attention (ST-DBA) module. It processes multi-temporal optical sequences and SAR backscatter in a joint manner. For the purpose of modality gap bridging, temporal positional encodings and domain - specific normalization are used so as to let the network route clean structural information through thick cloud.

(3) A Instance-Adaptive Preconditioning Control (APC) policy that scales noise variance based on scene-specific priors during extreme diffusion steps, enhancing the stability of training and fine-detail recovery.

Table 1. Key design distinctions between EMRDM and the proposed PMR-SDE framework.

Design Aspect	**EMRDM [10]**	**PMR-SDE (Ours)**
Reversion Target	Static empirical image mean	SAR-conditioned $\Phi(X, xSAR)$
Cross-modal Structural Guidance	None	Sentinel-1 SAR
Multi-temporal Fusion	Convolutional aggregation	ST-DBA with temporal encoding
Preconditioning	Fixed rescaling	Instance-Adaptive Control (APC)

2 Related Work

2.1 Physics-Based Cloud Removal Methods

The conventional cloud removal is based on the models of atmospheric scattering, which express the radiance as a function of the surface reflectance in terms of transmission and airlight [1,3]. The Dark Channel Prior (DCP) [2] gave an estimate of the transmission map by way of practical heuristic, and subsequent research was used to make it suitable with multispectral remote sensing [4,5]. The primary drawback is that it depends on hand - crafted scene priors. For example, the idea of a dark pixel for each patch is not very strong. In built - up areas where there are bright rooftops or under heavy cloud cover, these assumptions can result in color problems and shape distortion.

2.2 Discriminative and Generative Deep Learning Methods

The data-driven approaches have since become the norm. The early CNN-based methods such as AOD-Net [11] learned transmission map end-to-end; FFA-Net [12] and MSBDN [13] introduced multi-scale feature aggregation. Non-negative matrix factorization [9] and Vision Transformers [14,15] introduced longer-range spatial modeling. One of the well-known drawbacks of purely regression-based models (L1 or MSE loss) is that they tend to average over plausible outputs, resulting in over-smoothed textures.

The generative Adversarial Networks [16,17] made the perceptual quality better at the expense of training instability and inconsistency in geometry. The more stable alternative to generate is diffusion models [6], and the continuous-time SDE formulation [7,8] can also make the path of generation regularized by the drift design. Nevertheless, when the clouds are totally opaque and no light reaches the sensor then even a regularized diffusion process may hallucinate textures which are visually plausible but geometrically inaccurate. We are different because we base our work on the idea of having a SAR-based structural reference as the target of the drift and thus the reverse process will be constrained within a manifold that is related to the real scene.

2.3 Multi-Temporal and Multi-Modal Fusion

Multi-temporal and multi-modal fusion is a natural solution to the information gap that dense clouds leave. U-TAE [18] employs temporal self-attention to build cloud-free composites; UnCRtainTS [19] weights frames by estimated cloud density. On the side of multi-modality, CR-TS Net [20] and DSen2-CR [21] provided early foundations on optical-SAR fusion and STGAN [22] studied spatiotemporal generative networks. These approaches have evident temporal reasoning but their regression-style decoders are still unable to generate high-frequency detail. The motivation behind PMR-SDE is the combination of the generative ability of SDEs with the structural reliability of SAR-guided temporal fusion.

3 Methodology

We model the cloud degradations as a continuous stochastic process whose atmospheric disturbances behave as a rich, non-stationary noise distribution. The general architecture of the proposed PMR-SDE framework is depicted in Fig. 1.

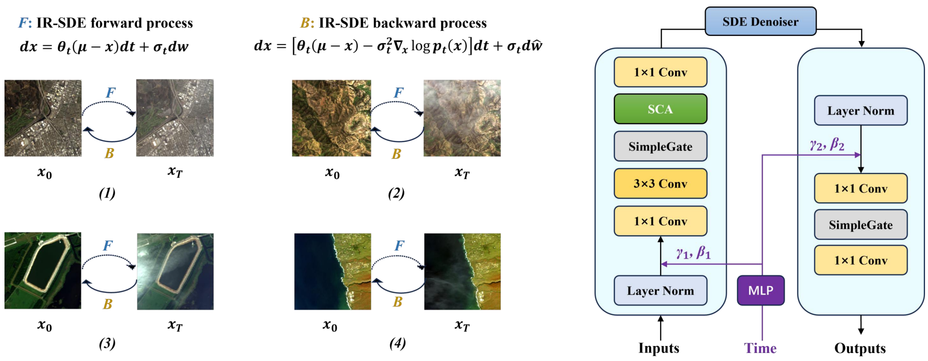

Fig. 1. The proposed Preconditioned Mean-Reverting SDE (PMR-SDE) framework. In the continuous reverse SDE path, the process of building up clean textures on land surfaces is gradually done by using multi-temporal optical series (256x256x4: RGB + NIR bands) and Sentinel-1 SAR structural anchors (256x256x2: VV+VH polarizations), where the results are then released as a restored optical image at the same resolution.

3.1 Denoising Diffusion Probabilistic Models as Foundation

We extend the Denoising Diffusion Probabilistic Model (DDPM) [6]. In case of a clean image x_0, forward process is a Markov chain $x_0 \rightarrow x_1 \rightarrow \cdots \rightarrow x_T$ with variance schedule β_t. The marginal at step t admits a closed form: $xt = \sqrt{(\bar{\alpha_t})}x_0 + \sqrt{(1 - \bar{\alpha_t})}\epsilon$, where$\bar{\alpha_t}$= ∏(i=1 to t) $(1-\beta_i)$ and ϵ ~ N$(0, I)$. A denoising network ϵ_θ is trained using the simplified MSE objective:

$$\mathcal{L}_{\mathrm{DDPM}}(\theta) = \mathbb{E}_{t,\, x_0,\, \epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] \tag{1}$$

3.2 Cross-Modally Conditioned Mean-Reverting SDE

We generalize the discrete formulation to continuous time [7], regularizing the drift with an Ornstein-Uhlenbeck (OU) process OU whose equilibrium depends on the input, rather than diffusing towards pure Gaussian noise as in standard score-based models. Denote by XX the multi-temporal optical stack and by the co-registered SAR image. The forward SDE is:

$$dx(t) = \kappa(\Phi(\mathcal{X}, x_{\mathrm{SAR}}) - x(t))\, dt + g(t)\, dW(t) \quad (2)$$

with reversion rate κ > 0, diffusion coefficient g(t), and a standard Wiener process W(t). The conditioning function $\Phi(\mathcal{X}, x_{\mathrm{SAR}})$ is implemented by the ST-DBA module (Section 3.3). Specifically, given optical frames X = {x^1, …, x^N}and the SAR anchor,

$$\Phi(\mathcal{X}, x_{\mathrm{SAR}}) = \mathrm{ST-DBA}\left(\left[\mathrm{E}_{\mathrm{opt}}\left(x^{(1)}\right), \ldots, \mathrm{E}_{\mathrm{opt}}\left(x^{(N)}\right), \mathrm{E}_{\mathrm{SAR}}\left(x_{\mathrm{SAR}}\right)\right]\right) \quad (3)$$

Where $\mathrm{E}_{\mathrm{opt}}$ and $\mathrm{E}_{\mathrm{SAR}}$ are shallow modality-specific encoders which transform both modalities into a common latent space, and ST-DBA carries out spatial-temporal fusion (Eqs. 57). This coupling of the SDE drift with the neural architecture guarantees that is a data-computed structural prior and not an abstract free variable.

Applying Itô's lemma gives a closed-form transition density:

$$q_t(x(t) \mid x(0)) = \mathcal{N}(x(t);\ x(0)\, e^{-\kappa t} + \Phi(\cdot)(1 - e^{-\kappa t}), \sigma^2(t)\, \mathbb{I}) \quad (4)$$

with $\sigma^2(t) = \int_0^t$ e^{-2κ(t−s)} $g(s)^2$ ds. As t increases, the mean tends to Φ(·) not to zero or a constant of the dataset, thus the path is still bound to the geometry of the scene during its course.

For the reverse process, Andersons theorem and score identity ∇_x log qt = −(x(t) − E[x(t)|x(0)])/$\sigma^2(t)$ allows us to parameterize the score via the denoising network ϵ_θ(x(t), x_{SAR}, M_t, t) as ∇_xlog qt ≈ $-\epsilon\theta/\sigma$(t). The SDE in reverse-time becomes:

$$dx(t) = \left[\kappa\Phi(\mathcal{X}, x_{\mathrm{SAR}}) - \kappa x(t) + \frac{g(t)^2}{\sigma(t)}\epsilon_\theta(x_t, x_{\mathrm{SAR}}, M_t, t)\right] d\bar{t} + g(t)\, d\bar{W}(t) \quad (5)$$

The drift is clearly divided into a deterministic SAR-conditioned pull to $\Phi(\cdot)$and a learned stochastic correction, with the sampling path constrained by the geometric constraints encoded in $\Phi(\cdot)$.

3.3 Spatio-Temporal Dual-Branch Attention Module

The ST-DBA module (Fig. 2) implements the conditioning function $\Phi(\cdot)$ by fusing multi-temporal optical observations with the SAR anchor.

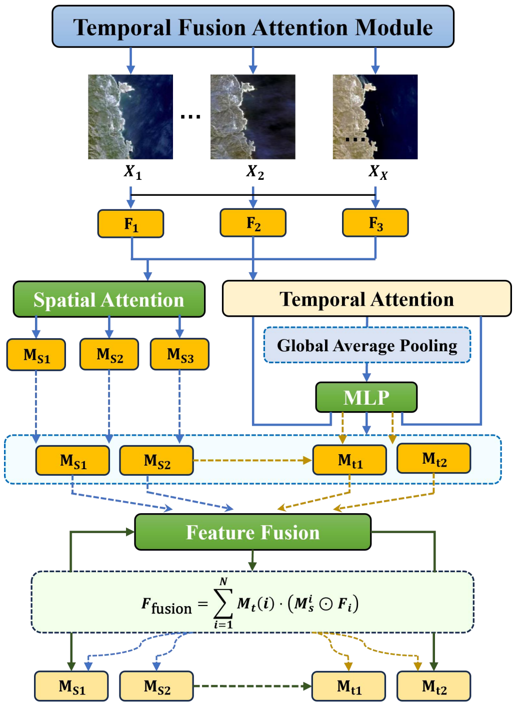

Fig. 2. Architecture of the Spatio-Temporal Dual-Branch Attention (ST-DBA) module. Multi-temporal optical features and SAR features are processed through parallel spatial and temporal attention streams and fused into the conditioning signal $\Phi(X, x_{SAR})$.

Combining optical reflectance with microwave SAR backscatter is not trivial since the SAR data has multiplicative speckle noise whose statistics are very different to optical imagery. We manage this using domain-specific shallow encoders: the SAR encoder has an Instance Normalization (IN) layer that suppresses localised speckle variance and brings the feature distribution closer to the optical branch [23].

The acquisition time of the real time series is not uniform, thus we add Temporal Positional Encodings (TPE) $P\Delta_t$ to the optical features. This is done by computing the day difference between the current frame and the target frame [24]. We then have a set of $F = \{f^1, \ldots, f^N, f_{\text{SAR}}\}$ where two attention branches are used in parallel:

1. Spatial Attention Branch: Within-frame textures are captured and speckle is suppressed. For each $f^i \in f$, spatial weights are obtained through convolution and Softmax:

$$M_s{}^i = \frac{\exp\left(\text{Conv}(f^i)\right)}{\sum_k \exp\left(\text{Conv}(f^k)\right)} \tag{6}$$

2. Temporal Attention Branch: It is trained to have a temporal weight of the frame that is global and weights down highly clouded dates, increasing the SAR prior. The

multi-layer perceptron (MLP) on the globally average-pooled features generates the temporal weight vector $M_t = [\gamma_1, \ldots, \gamma_N, \gamma_{SAR}]$:

$$M_t = \text{Softmax}\left(\text{MLP}\big(\text{GlobalAvgPool}(\mathcal{F})\big)\right) \tag{7}$$

The fused feature is assembled by element-wise gating:

$$f_{\text{fused}} = \sum_{i=1}^{N+1} \gamma_i \cdot \left(f^i \odot M_s{}^i\right) \tag{8}$$

where $\odot$ is element-wise multiplication. The dual gating removes non-stationary cloud noise and transmits unaltered spatial structure into the SDE drift.

3.4 Adaptive Preconditioning Control Strategy

Training a diffusion model across the full range of noise levels poses a scaling challenge: as $t \to T_{\max}$, the signal-to-noise ratio (SNR) goes to zero and gradient sizes can be very variable that results in heavy-cloud regions being biased spectrally and artifacts.

We respond to this with a Preconditioning control of Instance-Adaptive (APC) scheme. APC does not employ a global mean channel in order to rescale, but it takes the local prior x_{prior} of the least cloudy optical frame of the current multi-temporal sequence and stabilizes it through the average pooling globally. The denoiser f_θ is then wrapped to predict the clean target x_0 directly rather than unscaled noise. The preconditioned input and output are:

$$x_t{}^{\text{scaled}} = c_{\text{in}}(t) \cdot \left(x_t - \sqrt{1 - \sigma^2(t)}\, x_{\text{prior}}\right) \tag{9}$$

$$x_0{}^{\text{pred}} = c_{\text{skip}}(t)\, x_t + c_{\text{out}}(t)\, f_\theta\big(x_t{}^{\text{scaled}}, t, \Phi(\cdot)\big) \tag{10}$$

with time-dependent coefficients $c_{\text{in}}(t)$, $c_{\text{skip}}(t)$, $c_{\text{out}}(t)$ obtained from the variance schedule $\sigma(t)$ such that the input to the network stays unit variance at each diffusion step. The noise estimate used for score matching is then recovered as:

$$\epsilon_\theta(x_t, t) = \frac{x_t - e^{-\kappa t} x_0{}^{\text{pred}} - \left(1 - e^{-\kappa t}\right) x_{\text{prior}}}{\sigma(t)} \tag{11}$$

It should be noted that x_{prior} is not the only local scaling anchor within APC; it is a mean-reverting equilibrium of the reverse SDE (Eq. 5) and it is still controlled by SAR-conditioned $\Phi(X, x_{SAR})$. This dynamic variance bound will ensure the stability of the gradient landscape as well as the recovery of fine spatial textures in case of extreme noise levels.

One practical limitation is that when all the frames in the multi-temporal sequence are densely clouded, the least-cloudy frame may itself carry a large amount of occlusion, which degrades x_{prior} as a rescaling reference. In these all-cloud conditions, the SAR-conditioned drift $\Phi(X, x_{SAR})$ in Eq. 5 carries a heavier share of the guidance, partially offsetting the degraded optical prior. Learning a prior estimator from multi-modal cues is left for future investigation.

4 Experiments and Analysis

4.1 Dataset and Implementation Details

We learn and test on SEN12MS-CR-TS [20], a multi-temporal benchmark of multi-modal RS restoration. The samples are co-registered triplets: Sentinel-1 SAR (VV and VH polarizations), cloudy Sentinel-2 optical (RGB and near-infrared bands) and a clear-sky ground truth, gathered in globally distributed areas. We use the official division - 30 regions to train (with winter and summer subsets, 25,234 triplets altogether) and 12 disjoint regions to test. All patches are 256x256 pixels.

The denoising network ϵ_θ is a U-Net with channel dimensions [64, 128, 256, 512] and four scales in space. The encoder block has two layers of 3x3 convolutions with GroupNorm (32 groups) and SiLU followed by 2x2 max pooling. Bottleneck is at 16x16 resolution (512 channels) with 4 head self attention. Decoder blocks are similar to the encoder as they use transposed-convolution upsampling and skip connections. FiLM is used to encode time t as sinusoidal embeddings and input into each residual block. We have AdamW (initial LR 10^{-4}, cosine annealing to 10^{-6}, weight decay 10^{-4}), training for 100 epochs on four RTX 3090 GPUs with batch size 16. The forward SDE uses $\kappa = 1.0$. Augmentation consists of horizontal/vertical flips and 90-degree rotations. At inference, the reverse path uses N_sample = 50 steps (Full) or a truncated 15 steps (Fast). The training objective is a weighted sum:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} + \lambda_{\text{score}}\mathcal{L}_{\text{score}} \tag{12}$$

with $\lambda_{\text{score}} = 1.0$, $\lambda_{\text{MSE}} = 10.0$, and $\lambda_{\text{SSIM}} = 5.0$, chosen empirically to balance distribution matching with radiometric accuracy. The continuous time t is drawn uniformly from $(\epsilon_{\text{time}}, \mathrm{T}_{\max}]$ where $\epsilon_{\text{time}} = 10^{-3}$ and $\mathrm{T}_{\max} = 1.0$.

4.2 Evaluation Metrics

The results are reported in the form of four metrics, namely Root Mean Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) and Spectral Angle Mapper (SAM). RMSE and PSNR represent the accuracy of pixels, SSIM is used to measure the fidelity of structure and contrast, and SAM is a measurement of per-pixel spectral angular deviation (lower values indicate higher quality regardless of brightness).

4.3 Quantitative and Qualitative Evaluation

Table 2 shows results for N=33N=3 multi-temporal reconstruction on SEN12MS-CR-TS, together with per-patch inference latency and number of parameters to make the efficiency trade-off explicit.

Table 2. Quantitative multi-temporal image reconstruction results and efficiency profiles on the SEN12MS-CR-TS benchmark (N=3). Best results are highlighted in bold.

Model	RMSE↓	PSNR(dB)↑	SSIM↑	SAM↓	Params(M)	Latency(s)↓
Least cloudy	0.079	—	0.815	12.204	—	—
DSen2-CR [21]	0.060	26.04	0.810	12.147	12.4	0.012
STGAN [22]	0.057	25.42	0.818	12.548	44.2	0.035
CR-TS Net [20]	0.051	26.68	0.836	10.657	15.8	0.018
U-TAE [18]	0.051	27.05	0.849	11.649	22.1	0.024
UnCRtainTS [19]	0.051	27.84	0.866	10.160	31.4	0.029
Ours (PMR-SDE, Fast)	**0.049**	**27.91**	**0.869**	**10.195**	**38.6**	**0.078**
Ours (PMR-SDE, Full)	**0.048**	**28.02**	**0.875**	**10.152**	**38.6**	**0.294**

Ours-Full obtains 28.02 dB PSNR and 0.875 SSIM, which is in favor of the assumption that multi-temporal restoration can be cast as a conditioned mean-reverting process that contributes to the reduction of spatial and structural distortion. One interesting outcome is the SAM score: Ours-Full at 10.152 is just a bit better than the best regression baseline UnCRtainTS (10.160). This is interesting since generative models tend to sacrifice spectral fidelity in order to have high-frequency detail; here it seems like the SAR-anchored drift keeps the spectral signature grounded. The PSNR gain on UnCRtainTS is small under normal conditions (+0.18 dB), but the difference increases significantly under heavy cloud (Section 4.5: +1.11 dB over EMRDM), confirming that the SDE-based design is most effective when optical information is least available.

The price of the iterative sampling is seen in the latency column: Ours-Full will be 0.294 s on a patch, which is about ten times as long as the single-forward approaches. In order to provide an alternative that is more time-saving, Ours-Fast minimizes the discretization to 15 and reaches 0.078 (73.5% decrease) with still better results than U-TAE and CR-TS Net at all metrics (27.91 dB PSNR, 0.869 SSIM). This provides two points of operation; one point of quality and the other of speed. However, the generative overhead would still be larger than the discriminative ones, and it would take some work to put it into edge deployment: model quantization, neural ODE solvers or distillation are obvious choices.

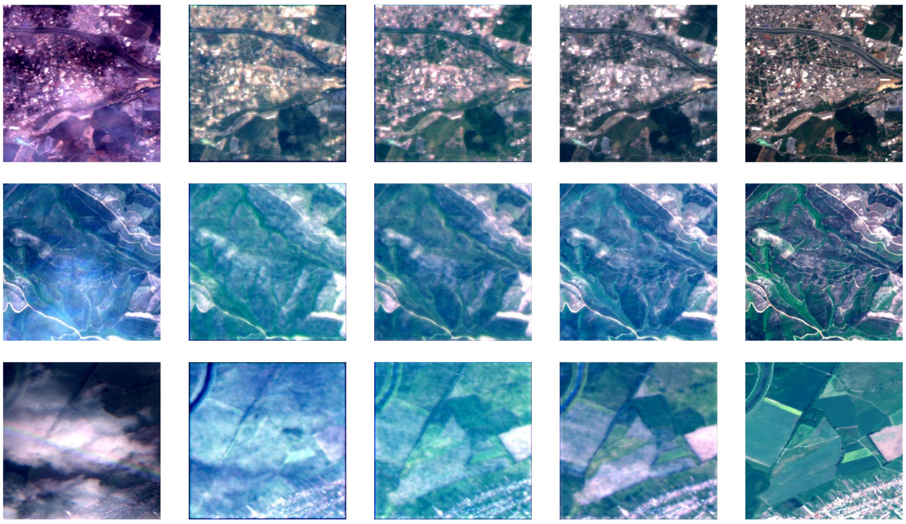

Fig. 3. Multi-temporal cloud removal comparison of qualitative results on SEN12MS-CR-TS. PMR-SDE is able to reconstruct fine-grained ground textures with more fidelity and baseline methods either over-smooth the details or create spatial artifacts.

4.4 Ablation Study and Sensitivity Analysis

Table 3 reports the ablation results.

Table 3. Ablation study and temporal length scaling analysis on the SEN12MS-CR-TS dataset.

Configuration	RMSE↓	PSNR(dB)↑	SSIM↑	SAM↓
Baseline SDE (Vanilla VP-SDE, N=3)	0.059	26.14	0.811	12.312
w/o SAR Modality (x_{SAR} removed)	0.053	27.21	0.835	10.820
w/o ST-DBA (Mean Pooling Fusion)	0.053	27.35	0.854	10.951
w/o APC	0.054	27.02	0.847	11.350
Full Model (Ours, N=3)	**0.048**	**28.02**	**0.875**	**10.152**
Full Model (Ours, N=5)	**0.045**	**28.45**	**0.889**	**9.985**

The biggest change in the entire board happens when switching between the mean-reverting SDE and a standard VP-SDE (pure noise diffusion), which proves that an unconstrained drift is dangerous with dense occlusion. The SAR anchor can be removed, lowering SSIM by 0.875 to 0.835, the most dramatic single-component degradation, which is consistent with the intuition that SAR gives geometric structure where optical channels do not. APC primarily affects RMSE and PSNR negatively, as expected due to its contribution to stabilizing training gradients. Increasing temporal window N=3 to N=5 also provides additional benefits (SSIM 0.889, SAM 9.985) indicating that the model enjoys richer multi-temporal redundancy.

4.5 Robustness under Extreme Heavy-Cloud Scenarios

In order to evaluate the robustness in extreme occlusion, we have generated a heavy-cloud subset of SEN12MS-CR-TS by sampling patches where per-pixel cloud masks are more than 60% spatially occluded (1,063 patches altogether). The UnCRtainTS and EMRDM were retrained on the same training split with their official settings. Table 4 shows the results.

Table 4. Quantitative evaluation under extreme heavy-cloud scenarios (>60% cloud occlusion, N=3).

Model	RMSE↓	PSNR(dB)↑	SSIM↑	SAM↓
UnCRtainTS [19]	0.082	22.15	0.724	14.850
EMRDM [10]	0.078	22.84	0.741	14.210
Ours (PMR-SDE)	**0.069**	**23.95**	**0.812**	**13.104**

The ranking varies significantly in heavy cloud. UnCRtainTS and EMRDM degrade drastically since they do not have good optical references. PMR-SDE has SSIM of 0.812 and outperforms EMRDM with +1.11 dB in PSNR. The cause is structural: the mean-reversion target of EMRDM is a fixed dataset-wide average which does not possess any scene-specific geometry, hence when optical channels are blacked out, there is no anchoring point to reverse the process. SAR-conditioned drift of PMR-SDE is informative irrespective of the thickness of clouds and offers geometric guidance even during total occlusion.

5 Conclusion

We have introduced PMR-SDE, a new framework of continuous time multi-temporal cloud removal where the traditional target that is not constrained by diffusion is substituted with an instance-based SAR conditionally mean-reverting equilibrium. It is easy to understand the concept: in case of blocked optical channels, the structure of the scene geometry can be maintained using a radar-derived reference of structural information as it follows the actual generative path. The ST-DBA module is the implementation of this fusion and attends to the irregular optical time series and SAR backscatter in both space and time and APC scheme is used to stabilize the training process through rescaling the input-output of the denoiser based on local priors. In SEN12MS-CR-TS, PMR-SDE performs equally or better than the current approaches on common metrics and when clouds are dense, the difference between the two methods becomes even more apparent. Its main drawback is the cost of computation; it takes longer to do iterative sampling than one forward regression. Natural next steps include faster solvers and compression of models. We feel that the principle behind this general idea of having a modality which is always available to use as an anchor point for drift is applicable to other inverse problems besides cloud removal in which a particular sensing mode is compromised intermittently.

References

1. McCartney, E.J.: Optics of the Atmosphere: Scattering by Molecules and Particles. John Wiley & Sons, New York (1976)
2. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Trans. Pattern Anal. Mach. Intell. 33(12), 2341–2353 (2010)
3. Anandakrishnan, J., Sundaram, V.M., Paneer, P.: CERMF-Net: A SAR-optical feature fusion for cloud elimination from Sentinel-2 imagery using residual multiscale dilated network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17, 11741–11749 (2024)
4. Shen, H., Li, H., Qian, Y., Zhang, L., Yuan, Q.: An effective thin cloud removal procedure for remote sensing images. ISPRS J. Photogramm. Remote Sens. 96, 224–235 (2014)
5. Shen, H., Wu, J., Cheng, Q., Aihemaiti, M., Zhang, C., Li, Z.: A spatiotemporal fusion based cloud removal method for remote sensing images with land cover changes. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12(3), 862–874 (2019)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020)
7. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: Proc. ICLR (2021)
8. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Image restoration with mean-reverting stochastic differential equations. In: Proc. ICML, pp. 18857–18870 (2023)
9. Li, X., Wang, L., Cheng, Q., Wu, P., Gan, W., Fang, L.: Cloud removal in remote sensing images using nonnegative matrix factorization and error correction. ISPRS J. Photogramm. Remote Sens. 148, 103–113 (2019)
10. Liu, Y., Li, W., Guan, J., Zhou, S., Zhang, Y.: Effective cloud removal for remote sensing images by an improved mean-reverting denoising model with elucidated design space. In: Proc. IEEE/CVF CVPR, pp. 17851–17861 (2025)
11. Li, B., Peng, X., Wang, Z., Ren, J., Xu, C., Tao, D., Yang, M.-H.: AOD-Net: All-in-one dehazing network. In: Proc. IEEE ICCV, pp. 4770–4778 (2017)
12. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: FFA-Net: Feature fusion attention network for single image dehazing. In: Proc. AAAI, vol. 34(07), pp. 11908–11915 (2020)
13. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.-H.: Multi-scale boosted dehazing network with dense feature fusion. In: Proc. IEEE/CVF CVPR, pp. 2157–2167 (2020)
14. Zhou, H., Chen, Z., Liu, Y., Sheng, Y., Ren, W., Xiong, H.: Physical-priors-guided DehazeFormer. Knowl.-Based Syst. 266, 110410 (2023)
15. Song, Y., He, Z., Qian, H., Du, X.: Vision transformers for single image dehazing. IEEE Trans. Image Process. 32, 1927–1941 (2023)
16. Zhang, X.: Research on remote sensing image dehaze based on GAN. J. Signal Process. Syst. 94(3), 305–313 (2022)
17. Zhang, S., Zhang, X., Shen, L., Fan, E.: GAN-based dehazing network with knowledge transferring. Multimed. Tools Appl. 83(15), 45095–45110 (2024)
18. Sainte Fare Garnot, V., Landrieu, L.: Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In: Proc. IEEE/CVF ICCV, pp. 4872–4881 (2021)
19. Ebel, P., Sainte Fare Garnot, V., Schmitt, M., Wegner, J.D., Zhu, X.X.: UnCRtainTS: Uncertainty quantification for cloud removal in optical satellite time series. In: Proc. IEEE/CVF CVPR, pp. 2086–2096 (2023)
20. Ebel, P., Xu, Y., Schmitt, M., Zhu, X.X.: SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. IEEE Trans. Geosci. Remote Sens. 60, 1–4 (2022)

21. Meraner, A., Ebel, P., Zhu, X.X., Schmitt, M.: Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. ISPRS J. Photogramm. Remote Sens. 166, 333–346 (2020)
22. Sarukkai, V., Jain, A., Uzkent, B., Ermon, S.: Cloud removal from satellite images using spatiotemporal generator networks. In: Proc. IEEE/CVF WACV, pp. 1796–1805 (2020)
23. Huang, X., Wen, D., Li, J., Qin, R.: Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. Remote Sens. Environ. 247, 111936 (2020)
24. Garnot, V.S.F., Landrieu, L., Giordano, S., Chehata, N.: Lightweight temporal self-attention for classifying satellite images time series. In: Advanced Geographic Information Systems (2020)