Semantic-Adaptive Diffusion for Dynamic Spatiotemporal Fusion

Supplementary Material

A. Overview of SA-STF

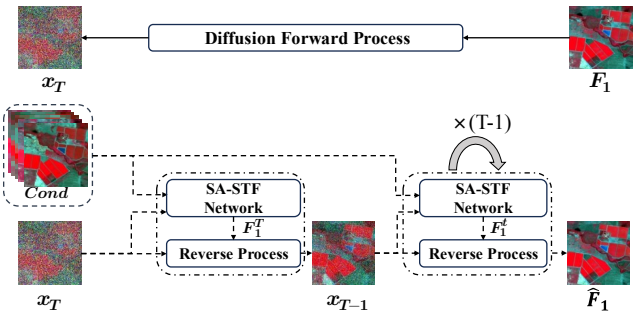

Figure 1. Overview of SA-STF

In SA-STF, the residual x_{res} and random noise are progressively injected into the fine-resolution image F_1 during the forward process, which causes the image to gradually degrade as the time step t increases. When $t = T$, the fine-resolution image is degraded into a coarse image x_T, which is a linear combination of residual and noise. During this process, the SA-STF Network is trained to learn the intermediate variable F_1^t at each diffusion step. During the reverse process from x_t to x_{t-1}, the network takes the degraded image x_t and the conditional inputs $\{C_1, F_0, C_0, F_2, C_2\}$ to predict the intermediate variable F_1^t for reconstruction. The predicted F_1^t is then substituted into the reverse diffusion step to generate x_{t-1}. By iterating this reverse step T times, the model gradually refines x_T and ultimately achieves the target image $\hat{F}_1$.

B. Reverse Process

The original RDDM [17] reverse process is defined as:

$$x_{t-1} = x_t - (\bar{\alpha}_t - \bar{\alpha}_{t-1})\, x_{\text{res}}^{\theta} - \left(\bar{\beta}_t - \bar{\beta}_{t-1}\right) \varepsilon_{\theta}, \quad (1)$$

where x_{res}^{θ} and ε_{θ} denote the residual and noise, respectively, which are typically predicted by two independent networks. Although noise prediction can enhance the diversity of generated results, it may introduce instability during training. By analyzing the forward diffusion process, we observe that both the residual and noise can be expressed in terms of F_1. Based on this observation, we construct a neural network to reconstruct the intermediate variable F_1^t (corresponding to F_1, when $t = 0$) from the degraded image x_t at each diffusion step t. Next, we explain the rationale for introducing F_1^t to formulate the reverse reconstruction process from the perspective of the forward diffusion. The forward diffusion process at time t is defined as:

$$x_t = F_1^t + \bar{\alpha}_t x_{\text{res}} + \bar{\beta}_t \varepsilon, \quad (2)$$

where C_1 denotes the low-resolution image at the prediction time, x_{res} is the residual, and ε represents noise. According to the definition of the residual and Eq. 2, the residual and noise can be expressed as:

$$\begin{aligned} x_{\text{res}} &= C_1 - F_1^t, \\ \varepsilon &= \frac{1}{\bar{\beta}_t}\left(x_t - F_1^t - \bar{\alpha}_t(C_1 - F_1^t)\right). \end{aligned} \quad (3)$$

Substituting Eq. 3 into Eq. 1, x_{t-1} can be expressed solely in terms of x_t, F_1^t, and C_1:

$$\begin{aligned} x_{t-1} &= \frac{\bar{\beta}_{t-1}}{\bar{\beta}_t} x_t + \gamma_t F_1^t - \lambda_t (C_1 - F_1^t), \\ \gamma_t &= 1 - \frac{\bar{\beta}_{t-1}}{\bar{\beta}_t}, \\ \lambda_t &= \bar{\alpha}_{t-1} - \frac{\bar{\beta}_{t-1}}{\bar{\beta}_t}\bar{\alpha}_t. \end{aligned} \quad (4)$$

The above derivation clearly demonstrates that, in the reverse diffusion process, the residual and noise components can be jointly modeled via the intermediate image F_1^t, thereby providing a theoretical foundation for using a single network to perform reverse reconstruction.

C. Performance Analysis under Dynamic Changes

To evaluate the performance of the methods under dynamic changes, Fig. 2 visualizes the boxplots of RMSE, SAM, and SSIM computed over the changed regions of the three test sets. SA-STF exhibits the narrowest box widths and the most favorable median values, indicating its superior overall performance. The RMSE and SSIM boxplots indicate that SA-STF is able to accurately reconstruct high-frequency textures in areas with substantial semantic or phenological changes. Moreover, SA-STF's superior SAM scores further confirm its ability to preserve spectral consistency and prevent spectral distortion in change regions.

D. Dataset

The CIA dataset was collected in southern New South Wales, Australia (34.0034°E, 145.0675°S), a representative irrigated agricultural region dominated by rice cultivation. The LGC dataset was collected in northern New South Wales, Australia (149.2815°E, 29.0855°S), capturing the complete sequence of flooding events and the subsequent vegetation recovery. The AHB dataset was collected

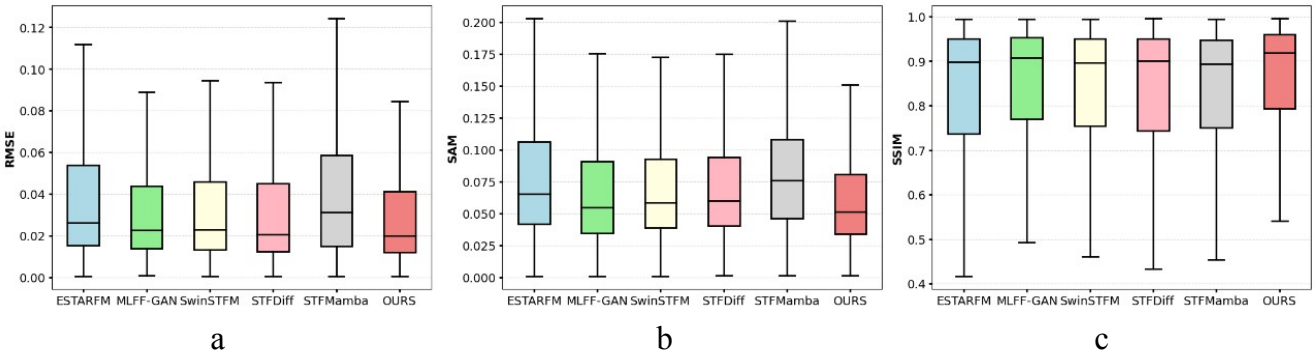

Figure 2. Performance analysis under dynamic changes. a. RMSE Boxplot; b. SAM Boxplot; c. SSIM Boxplot;

from Ar Horqin Banner, Inner Mongolia Autonomous Region, China (43.3619°N, 119.0375°E), a typical rural landscape characterized by mixed farming and pastoral activities. The region exhibits notable spatial heterogeneity, with diverse land cover types such as cropland, grassland, forest, and bare soil, as well as distinct seasonal phenological dynamics. In all three datasets, each image was divided into patches of size 256×256. Specifically, the CIA dataset contains 1,287 training patches and 117 testing patches; the LGC dataset includes 2,527 training patches and 361 testing patches; and the AHB dataset comprises 2,907 training patches and 327 testing patches. The dataset splits are summarized in Table 3.

E. Network Parameters

Table 2 describes the network architecture used in our framework. It lists each module, including the Noise Encoder, Shallow Fusion, Deep Fusion, De-Res Block and Denoising Decoder, along with their corresponding block names, convolution kernel sizes, input and output channels, and feature map sizes at each stage.

F. Classify results

The goal of STF is to generate high-quality time-series imagery for downstream tasks like land cover classification and agricultural assessment. We evaluate fusion quality by performing K-means [5] classification on both ground truth and fused images across the three datasets, using OA, Precision, Recall, and F1 Score metrics [22]. The classification results and quantitative evaluations on the CIA, LGC, and AHB datasets are shown in Fig. 3, Fig. 4, Fig. 5 and Table 1. Traditional and deep learning-based approaches often exhibit limited generalization across diverse datasets. In contrast, SA-STF consistently achieves the best performance across all evaluation metrics, highlighting its strong generalization capability to downstream tasks. This performance gain is attributed to the temporal feature alignment and semantic-adaptive fusion modules, which explicitly model land cover dynamics and adaptively transfer high-frequency information from reference images. Therefore, even under challenging conditions with complex surface changes, SA-STF maintains high spectral fidelity and spatiotemporal consistency.

Table 1. Quantitative evaluations of classification performance

Dataset	Method	OA↑	Precision↑	Recall↑	F1↑
CIA	ESTARFM [41]	0.6729	0.6415	0.6449	0.6421
	MLFF-GAN [27]	0.5727	0.5736	0.5717	0.5715
	SwinSTFM [1]	0.5745	0.5404	0.5420	0.5381
	STFDiff [10]	0.6585	0.6353	0.6363	0.6342
	STFMamba [38]	0.4968	0.4786	0.4752	0.4743
	OURS	**0.7121**	**0.6862**	**0.6866**	**0.6862**
LGC	ESTARFM[41]	0.6091	0.6149	0.6429	0.6171
	MLFF-GAN[27]	0.6129	0.6141	0.6597	0.6322
	SwinSTFM[1]	0.6040	0.6103	0.6391	0.6227
	STFDiff[10]	0.6319	0.6454	0.6562	0.6504
	STFMamba[38]	0.6403	0.6497	0.6752	0.6608
	OURS	**0.6561**	**0.6620**	**0.6840**	**0.6718**
AHB	ESTARFM[41]	0.5266	0.5152	0.5060	0.5083
	MLFF-GAN[27]	0.5182	0.4959	0.4921	0.4932
	SwinSTFM[1]	0.5436	0.5174	0.5168	0.5167
	STFDiff[10]	0.5330	0.5173	0.5116	0.5131
	STFMamba[38]	0.5227	0.4912	0.4931	0.4920
	OURS	**0.5668**	**0.5498**	**0.5440**	**0.5459**

Table 2. Network Structure

Module		Kernel Size	In-channel	Out-channel	Feature Size
Noise Encoder	Block1	3×3	6	64	256×256
	Block2	3×3	64	64	256×256
	Block3	3×3	64	128	128×128
	Block4	3×3	128	256	64×64
	Block5	3×3	256	512	32×32
	Block6	3×3	512	512	16×16
Shallow Fusion	Block1	3×3	6	64	256×256
	Block2	3×3	64	64	256×256
	Block3	3×3	64	128	128×128
	Block4	3×3	128	256	64×64
	Block5	3×3	256	512	32×32
Deep Fusion	TFA1	1×1	512	512	16×16
	TFA2	1×1	512	512	16×16
De-Res Block	TE-Block1	3×3	512	512	16×16
	TE-Block	3×3	512	512	16×16
Denoising Decoder	Block1	3×3	1024	512	16×16
	Block2	3×3	1024	512	32×32
	Block3	3×3	768	256	64×64
	Block4	3×3	384	128	128×128
	Block5	3×3	192	64	256×256
	Block6	3×3	64	6	256×256

Table 3. Training and testing dataset.

Dataset	CIA			LGC			AHB		
Time	Time1	Time2	Time3	Time1	Time2	Time3	Time1	Time2	Time3
Training	20011007	20011016	20011101	20040416	20040502	20040705	20160607	20160826	20161013
	20011016	20011101	20011108	20040502	20040705	20040806	20160826	20161013	20170525
	20011101	20011108	20011124	20040705	20040806	20040822	20161013	20170525	20170610
	20011108	20011124	20011203	20040806	20040822	20041025	20170525	20170610	20170728
	20011124	20011203	20020104	20040822	20041025	20041126	20170610	20170728	20170829
	20011203	20020104	20020111	20041025	20041126	20050302	20170728	20170829	20180512
	20020104	20020111	20020212	20041126	20050302	20050403	20170829	20180512	20181003
	20020111	20020212	20020221				20180512	20181003	20181019
	20020212	20020221	20020309						
	20020221	20020309	20020316						
	20020309	20020316	20020401						
Testing	20020410	20020417	20020426	20041228	20050113	20050129	20150504	20150621	20150707

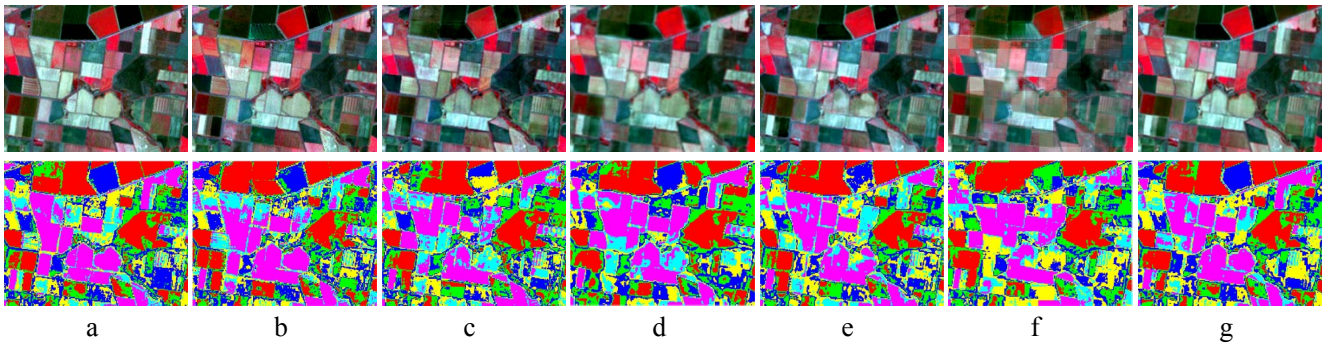

Figure 3. Classification results on the CIA dataset. a. Ground Truth (GT); b. ESTARFM[41]; c. MLFF-GAN[27]; d. SwinSTFM[1]; e. STFDiff[10]; f. STFMamba[38]; g. OURS.

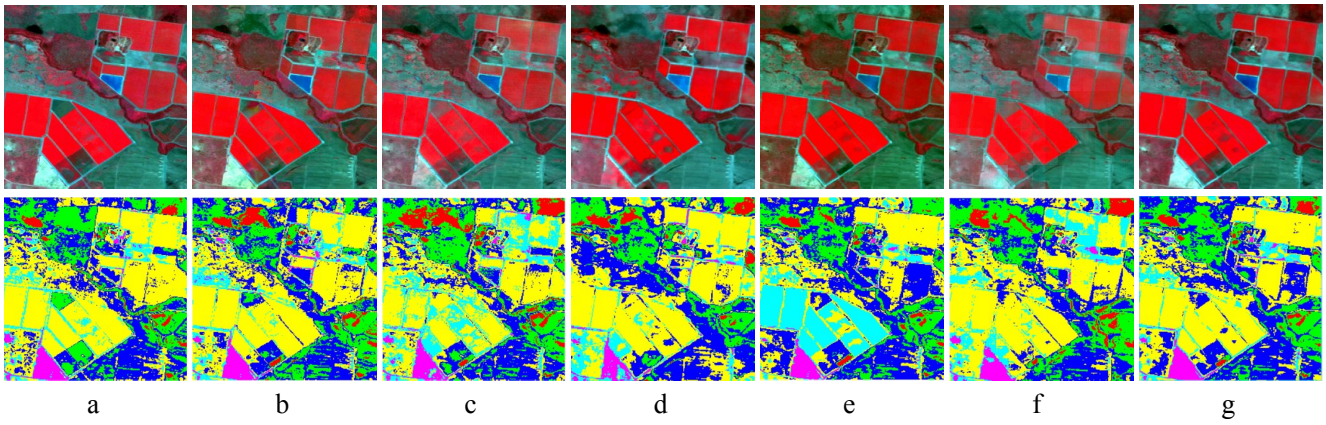

Figure 4. Classification results on the LGC dataset. a. Ground Truth (GT); b. ESTARFM[41]; c. MLFF-GAN[27]; d. SwinSTFM[1]; e. STFDiff[10]; f. STFMamba[38]; g. OURS.

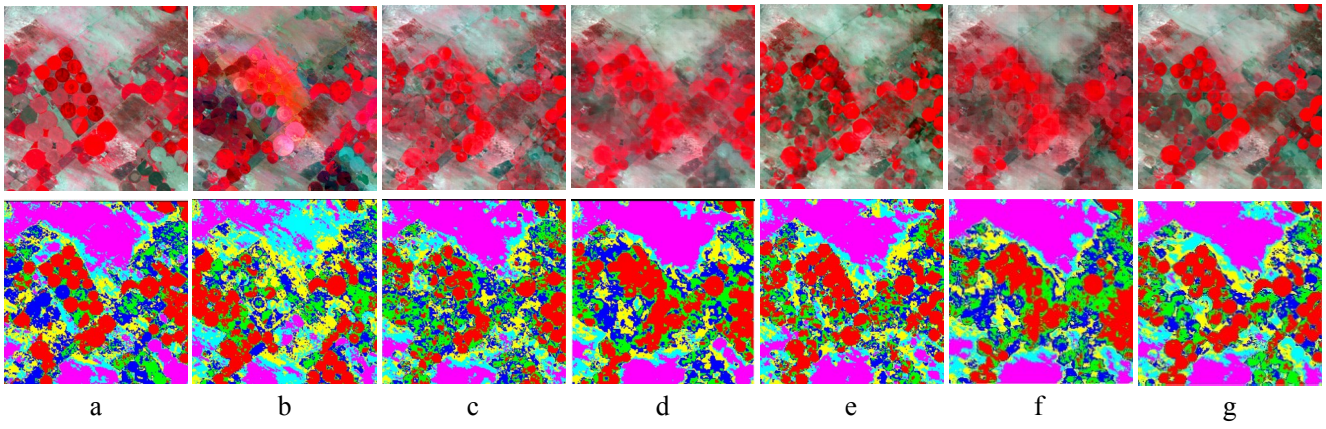

Figure 5. Classification results on the AHB dataset. a. Ground Truth (GT); b. ESTARFM[41]; c. MLFF-GAN[27]; d. SwinSTFM[1]; e. STFDiff[10]; f. STFMamba[38]; g. OURS.