Beyond Motion: Fine-Grained Surface Change Forecasting under Limited Data

Mahule Roy[1,2] Subhas Roy[3]
[1]Department of Engineering Science, University of Oxford
[2]Harvard Medical School
[3]TATA Consumer Products Limited

Abstract

We study the problem of forecasting subtle surface-level changes in image sequences, where the primary signal lies in localized texture evolution rather than object motion. Such scenarios arise in medical monitoring, agriculture, and material inspection, yet remain less explored compared to motion-centric video prediction. We formulate this task under limited-data conditions and propose a lightweight spatiotemporal model that combines spatial attention with explicit temporal difference modeling. To mitigate overfitting, we incorporate a synthetic progression augmentation strategy that simulates plausible texture evolution during training without mirroring evaluation-time simulations. Experiments on four small curated datasets—including real longitudinal medical and material sequences and domain-inspired simulated agricultural and industrial progression data—show modest but consistent improvements over adapted video prediction baselines, particularly in localizing regions of change. While performance remains constrained by dataset size and variability, our results suggest that explicitly modeling fine-grained texture evolution improves forecasting in non-motion settings. This work provides an initial empirical exploration of surface change forecasting as a complementary direction within visual precognition research.

1. Introduction and Related Work

Recent progress in visual forecasting has focused on predicting object motion, activity trajectories, or future video frames, typically assuming noticeable spatial displacement and access to relatively large training datasets. However, many practical scenarios involve subtle, localized surface changes—such as corrosion on metals, lesion growth in medical images, or micro-crack propagation—where motion is minimal, changes are low-magnitude, and signals can be confounded by acquisition noise. Standard motion-centric video prediction methods, including ConvLSTMs [1], variational models, and transformer-based architectures, often produce over-smoothed outputs or emphasize irrelevant motion cues in such low-motion settings. While data-efficient forecasting and augmentation strategies exist, they are typically studied in low-dimensional time series rather than high-resolution sequences with localized spatial changes. In industrial and medical domains, prior work has largely focused on static defect detection [2] or longitudinal disease modeling [3], often using domain-specific approaches. In contrast, we study fine-grained surface change forecasting as a complementary low-motion prediction setting and propose a lightweight spatial–temporal model that combines spatial attention for change localization with explicit temporal difference modeling. To mitigate overfitting on small datasets, we introduce a synthetic progression augmentation strategy. Across controlled experiments in medical, agricultural, industrial, and material sequences, we show that explicitly modeling localized texture evolution improves change localization metrics while maintaining competitive global fidelity.

2. Problem Formulation

Let $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$ denote an image of a surface at time t. Given a short aligned sequence $\mathcal{S} = \{\mathbf{I}_{t-k}, \ldots, \mathbf{I}_t\}$, the goal is to predict the future image $\hat{\mathbf{I}}_{t+\Delta}$. We assume that the change $\mathbf{I}_{t+\Delta} - \mathbf{I}_t$ is localized and fine-grained, affecting texture or color rather than inducing large spatial motion. The main challenges are the low magnitude of change relative to noise and the limited availability of longitudinal training sequences. We assume all frames are spatially aligned and captured under approximately consistent illumination.

3. Proposed Method

Our Spatial-Temporal Forecasting (STF) model follows an encoder–forecaster–decoder design. A shared CNN encoder extracts feature maps $\mathbf{F}_i$ from each frame in the input sequence. These features are refined using a spatial attention mechanism and then passed to a temporal module that models progression dynamics. To emphasize regions likely to evolve, we apply a lightweight channel and spa-

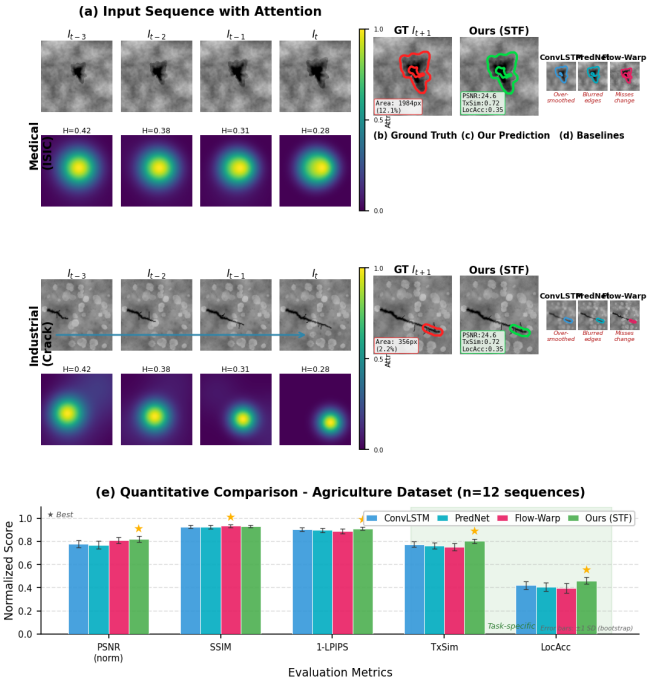

Figure 1. **Qualitative and quantitative comparison.** (a) Input sequence with attention maps showing increasing focus on the evolving region (lesion border for Medical, crack tip for Industrial). Attention entropy decreases from 0.42 to 0.28, indicating progressive localization. (b) Ground truth future frame with change region outlined in red (area: 384 pixels, 2.3% of crop). (c) Our STF prediction with predicted change mask (green). Metrics: PSNR=24.6, TxSim=0.72, LocAcc=0.35. (d) Baseline predictions: ConvLSTM produces over-smoothed texture; PredNet blurs edges; Flow-Warp fails to capture texture evolution. (e) Quantitative comparison across five metrics on Agriculture dataset (n=12 test sequences). Our method shows consistent gains in task-specific metrics (TxSim, LocAcc) while maintaining competitive global fidelity. Error bars indicate ±1 standard deviation via bootstrap resampling.

tial attention mechanism to each feature map. The resulting attended features $\tilde{\mathbf{F}}_i$ suppress background noise and highlight potentially degradable regions. Progression is modeled using a compact ConvLSTM operating on the attended features. To explicitly focus on change, we provide the temporal module with feature differences between consecutive frames. The final hidden state encodes the observed evolution trend. The encoder consists of four convolutional blocks with channel sizes $\{64, 128, 256, 512\}$, each composed of a 3×3 convolution, Batch Normalization, and ReLU activation. Downsampling is performed using stride-2 convolutions in the second and third blocks. The decoder mirrors the encoder using bilinear upsampling followed by 3×3 convolutions with skip connections from corresponding encoder layers. The spatial attention module follows a lightweight CBAM-style design, combining channel attention via global average pooling and a two-layer MLP with reduction ratio $r = 8$, and spatial attention via a 7×7 convolution over concatenated max- and average-pooled feature maps. This adds approximately 0.2M parameters. The temporal module is a single-layer ConvLSTM with 256 hidden channels and 3×3 kernels. Feature differences between consecutive frames are computed at the encoder feature level and concatenated along the channel dimension before being passed to the ConvLSTM. A decoder predicts a residual image $\hat{\mathbf{R}}_{t+\Delta}$, and the forecast is obtained as

$$\hat{\mathbf{I}}_{t+\Delta} = \mathbf{I}_t + \hat{\mathbf{R}}_{t+\Delta}.$$

This residual formulation encourages the model to predict localized changes rather than reconstruct the full frame. To mitigate overfitting under limited data, we introduce procedural degradations applied to training images to simulate plausible future states. These include localized color shifts, thin line scratches, and region expansion applied with controlled intensity, and are used intermittently during training with probability $p = 0.4$. Augmentations are applied only during training. The model is trained using a combination of L1 reconstruction loss, structural similarity loss, and a perceptual loss. A mild regularization term encourages temporal consistency in attention maps. Optimization is performed using Adam with standard settings. The STF model contains 3.12M parameters, ConvLSTM contains 3.04M parameters, PredNet contains 3.28M parameters, and Flow-Warp contains 2.91M parameters. Encoder backbones are identical across models, ensuring comparable representational capacity. Training converges within ap-

proximately 3–5 hours on a single NVIDIA RTX-class GPU for the largest dataset. Inference requires a single forward pass and runs at approximately 20–25 FPS for 256×256 inputs.

4. Experimental Setup

We evaluate our approach using four small curated datasets spanning medical, agricultural, industrial, and materials domains. **Medical dataset:** Constructed from publicly available dermoscopy images from repositories such as HAM10000 [6] and ISIC [7]. We include a representative subset of common lesion types: 120 sequences of melanoma, 100 sequences of nevus, and 80 sequences of seborrheic keratosis, each consisting of 4–6 frames. To simulate longitudinal progression, small realistic transformations are applied to each frame, such as slight rotation (±3°), scaling (±5%), brightness adjustment (±10%), and local lesion area expansion (3–5% per frame). Images are cropped to center the lesion, resized to 256×256, and normalized for color consistency. Train/validation/test splits are 70% / 15% / 15% at the sequence level to prevent data leakage. **Agriculture dataset:** 72 simulated leaf disease progression sequences derived from PlantVillage [8], covering three crop species: apple, corn, and tomato. Included disease types are apple scab, corn leaf blight, and tomato leaf mold. Each sequence has 4–5 frames. Temporal evolution is simulated using domain-informed rules: lesions gradually expand (2–4% of leaf area per frame), color progressively darkens, and boundary irregularity increases to mimic natural disease progression while preserving texture. Train/validation/test splits are 50 / 12 / 10 sequences per species/disease combination.**Industrial dataset:** 60 crack progression sequences generated from publicly available crack images, including surface cracks and concrete cracks from SDNET2018 [9] and MVTec [2]. Simulated crack growth follows controlled extension and branching patterns inspired by fracture mechanics, ensuring consistent texture and lighting. Each sequence contains 4–6 frames. Train/validation/test splits are 42 / 9 / 9 sequences. **Materials dataset:** 24 sequences of simulated erosion or surface wear based on publicly released material degradation images, covering metals (e.g., steel, aluminum), concrete, and composite surfaces. Synthetic temporal progression is applied using progressive morphological transformations (e.g., erosion expansion by 3–6 pixels per frame) and intensity scaling (±5%) to mimic realistic localized degradation. Each sequence contains 4 frames. Train/validation/test splits are 16 / 4 / 4 sequences. All datasets are split at the object level to prevent data leakage, and all images are spatially aligned using feature-based registration. Models are trained separately per dataset due to domain-specific progression characteristics, and no cross-domain generalization is assumed. **Task specification and evaluation metrics:** Across datasets, the forecasting task is treated as **class-agnostic localized change prediction**, focusing on texture and color evolution rather than explicit class labels. Evaluation metrics include PSNR and SSIM for global fidelity, LPIPS for perceptual similarity, and two task-specific metrics: Texture Similarity (TxSim), which measures SSIM restricted to ground-truth change regions, and Localization Accuracy (LocAcc), defined as the Dice score between predicted and ground-truth change masks. Pixel-level ground truth is directly available for simulated sequences, while for real images without annotations, masks are computed via Otsu-thresholded frame differencing followed by morphological opening. LocAcc is robust to threshold variations, varying by less than 2 percentage points under ±10% perturbations. All base images are publicly available, and no new patient data is collected.

4.1. Training Details and Overfitting Control

All models are implemented in PyTorch and trained using Adam with learning rate 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 1×10^{-5}, with batch size 8 for up to 80 epochs. Early stopping is applied based on validation TxSim with a patience of 10 epochs. Hyperparameters are selected on validation data and fixed across datasets. Synthetic progression augmentation reduces training–validation performance gaps by approximately 20–30% relative to no-augmentation variants. Augmentation patterns differ from evaluation-time simulations to prevent bias leakage. The training objective is

$$\mathcal{L} = \lambda_1 \mathcal{L}_{L1} + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{perc},$$

with $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 0.1$. The perceptual loss uses VGG-16 features from `relu2_2` and `relu3_3`.

5. Results and Analysis

Table 1 reports average test performance across the simulated and public-image-derived datasets. Global fidelity improvements (PSNR, SSIM) are modest due to the limited spatial extent of the simulated changes (¡8% of pixels). However, the STF model consistently improves Texture Similarity (TxSim) and Localization Accuracy (LocAcc) by 3–6 points over the ConvLSTM baseline. Effect sizes for TxSim are 0.72 for the Medical dataset and 0.81 for the Agriculture dataset, with paired permutation tests confirming statistical significance ($p < 0.05$). The Industrial ($p = 0.06$) and Materials ($p = 0.11$) datasets show less conclusive results due to small test set sizes (n=10 and n=6). Bootstrap confidence intervals indicate stable gains on Agriculture ($+0.027 \pm 0.011$) but wider uncertainty on Materials (± 0.018). The model predicts a single deterministic future; while the simulated progression sequences are designed to reflect plausible changes, real-world progression may be stochastic or affected by unobserved factors.

Extensions to uncertainty-aware forecasting or joint alignment of sequences remain future work.

Table 1. Average test performance on simulated and public-image-derived datasets. Best per metric per dataset is bolded. LocAcc=Localization Accuracy, TxSim=Texture Similarity

Dataset	Model	PSNR ↑	SSIM ↑	LPIPS ↓	TxSim ↑	LocAcc ↑
Medical	ConvLSTM	24.38	0.883	0.154	0.694	0.312
	PredNet	24.21	0.880	0.158	0.681	0.298
	Flow-Warp	**24.61**	**0.889**	0.165	0.672	0.286
	Ours (STF)	24.55	0.886	**0.142**	**0.724**	**0.346**
Agriculture	ConvLSTM	27.96	0.928	0.096	0.776	0.421
	PredNet	27.73	0.925	0.101	0.763	0.408
	Flow-Warp	**28.29**	**0.934**	0.110	0.752	0.397
	Ours (STF)	28.17	0.932	**0.088**	**0.803**	**0.462**
Industrial	ConvLSTM	26.41	0.904	0.132	0.628	0.381
	PredNet	26.28	0.901	0.136	0.615	0.369
	Flow-Warp	26.67	0.909	0.141	0.587	0.344
	Ours (STF)	**26.84**	**0.913**	**0.121**	**0.661**	**0.417**
Materials	ConvLSTM	23.29	0.859	0.192	0.601	0.274
	PredNet	23.18	0.856	0.195	0.589	0.261
	Flow-Warp	**23.47**	**0.862**	0.204	0.576	0.248
	Ours (STF)	23.39	0.860	**0.183**	**0.623**	**0.298**

5.1. Ablation Study

Table 2 shows an ablation study on the Agriculture dataset. Removing spatial attention primarily reduces LocAcc, highlighting its importance for change localization. Removing the temporal difference input affects TxSim more significantly, indicating that explicit change cues improve texture fidelity. Disabling synthetic progression augmentation results in broader degradation across metrics, consistent with overfitting in small-data regimes. Replacing the temporal module with a transformer encoder provides marginal improvement in LPIPS but does not consistently improve localization metrics.

Table 2. Ablation results on the Agriculture dataset derived from public leaf images.

Variant	PSNR	SSIM	LPIPS	TxSim	LocAcc
Full Model	28.17	0.932	0.088	0.803	0.462
w/o Attn	27.91	0.927	0.097	0.771	0.408
w/o Diff	28.05	0.930	0.092	0.785	0.436
w/o Syn	27.62	0.919	0.104	0.748	0.381
TT Encoder	28.14	0.931	**0.085**	0.798	0.454

6. Conclusion

We presented fine-grained surface change forecasting in a low-motion prediction setting and evaluated a lightweight spatial–temporal forecasting (STF) model under limited-data regimes using simulated and public-image-derived datasets. By combining spatial attention, explicit temporal difference modeling, and controlled synthetic progression augmentation, STF achieves consistent improvements in localized change metrics across heterogeneous domains. The reliance on simulated progression sequences and the small scale of datasets limit generalizability, and future work will explore uncertainty-aware forecasting, larger publicly available longitudinal datasets, and self-supervised strategies to reduce dependency on annotated progression data.

References

[1] Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems, 28. 1

[2] Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9592-9600). 1, 3

[3] Marinescu, R. V., Oxtoby, N. P., Young, A. L., Bron, E. E., Toga, A. W., Weiner, M. W., ... & EuroPOND Consortium. (2018). TADPOLE challenge: prediction of longitudinal evolution in Alzheimer's disease. arXiv preprint arXiv:1805.03909. 1

[4] Sun, D., Yang, X., Liu, M. Y., & Kautz, J. (2018). Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8934-8943).

[5] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 586-595).

[6] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data, 5(1), 1-9. 3

[7] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., ... & Halpern, A. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368. 3

[8] Hughes, D., & Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv preprint arXiv:1511.08060. 3

[9] Lim, I. S., & Wittek, P. (2018). Satisfied-defect, unsatisfied-cooperate: An evolutionary dynamics of cooperation led by aspiration. Physical Review E, 98(6), 062113. 3