Invited Paper

Unsupervised video denoising for structured noise via Fourier masking

Haosen Liu and Edmund Y. Lam*

Department of Electrical and Computer Engineering
The University of Hong Kong, Pokfulam, Hong Kong, China

ABSTRACT

We present an unsupervised video denoising method designed to remove structured noise, such as stripes or other random patterns common in microscopy and remote sensing, without needing clean training data. Unlike existing spatial-domain approaches, our method operates in the Fourier domain, leveraging the observation that Fourier coefficients of structured noise exhibit independence. By randomly masking and replacing Fourier coefficients across frames, the model learns to recover clean video content. Results show effective noise suppression without prior noise modeling, offering a practical solution for real-world imaging systems where clean references are unavailable.

Keywords: Video denoising, unsupervised learning, structured noise

1. INTRODUCTION

Noise is a common interference in imaging systems, affecting both the visual quality of images and the performance of subsequent downstream tasks. Its origins are diverse: due to the quantum nature of photons, the photoelectric signals acquired by imaging sensors exhibit Poisson statistics;[1] thermal agitation and the readout process introduce Gaussian noise. Beyond these spatially independent noise sources, spatially correlated structured noise also exists. For example, crosstalk[2] can cause the noise at each pixel to correlate with its neighbors. Moreover, imaging systems employing row-wise readout[3] or line scanning[4,5] often exhibit stripe noise that extends across entire rows or columns.

Research on denoising algorithms has continued for decades to mitigate the impact of such noise on image quality. Some approaches are model-based,[6,7] explicitly modeling video priors and noise statistics to construct denoising methods. Others are learning-based,[8–10] directly learning a mapping from noisy to clean videos using collected datasets. However, each paradigm has its limitations. Model-based methods typically rely on simplified models and often struggle with complex real-world scenarios. For learning-based methods, collecting large quantities of high-quality paired data becomes a new challenge: real dynamic scenes are rarely repeatable, making it difficult to acquire paired noisy and clean video data.

To harness the power of deep learning while circumventing the data acquisition bottleneck, increasing attention has been turned to unsupervised denoising. For instance, DeepInterpolation[11] trains a video interpolation model directly from noisy videos to achieve unsupervised video denoising. DeepCAD[12] splits each original noisy video into two sub-videos containing odd and even frames, respectively, and uses them as paired training data. Other methods, such as UDVD[13] and RDRF,[14] construct blind-spot networks to perform unsupervised video denoising. However, these approaches have limitations in real-world scenarios with structured noise. DeepInterpolation does not leverage information from the center frame during denoising, which severely limits its performance. DeepCAD assumes that adjacent odd and even frames are similar, making it unsuitable for dynamic scenes with fast motion. UDVD and RDRF assume spatially independent noise, and thus are ineffective in the presence of structured noise.

In this work, we develop an effective unsupervised video denoising method specifically for structured noise. Motivated by our previous observation that although structured noise exhibits strong spatial correlation, its Fourier coefficients can be independent,[15] we formulate our method in the Fourier domain rather than the

*Corresponding author: elam@eee.hku.hk

Optics, Photonics, and Digital Technologies for Imaging Applications IX, edited by Peter Schelkens, Tomasz Kozacki, Proc. of SPIE Vol. 14085, 140850W · 2026 Published by SPIE
0277-786X · doi: 10.1117/12.3106007

spatial domain. The core of our approach is a Fourier masking strategy: we randomly mask coefficients of a center frame with those from neighboring frames to construct the input for unsupervised training, while the hidden coefficients serve as the training target. To compensate for potential misalignment between the generated training pairs, a regularization term is incorporated into the loss function to enhance the content consistency. Experimental results demonstrate that our method effectively eliminates various types of spatially correlated noise without requiring ground-truth videos or prior knowledge of the noise statistics.

2. METHOD

We consider a sequence of $2T+1$ consecutive noisy frames $\mathbf{x} = [x_{-T}, \ldots, x_0, \ldots, x_T]$, where $x_t = z_t + n_t$ for $t = -T, \ldots, T$. Our goal is to train a network $\hat{z}_0 = f_\theta(\mathbf{x})$ to predict the clean center frame z_0 in an unsupervised manner. We assume that the noise in different frames is independent and that the noise in each frame follows the model

$$n = h * \eta, \tag{1}$$

where h is a correlation kernel and η is independent and identically distributed zero-mean noise. By choosing different h, this model can represent a wide range of structured noise types,[6,16] including crosstalk and stripe noise. In our previous work,[15] we showed that for such noise, when the image is sufficiently large, its Fourier coefficients asymptotically follow independent Gaussian distributions. This observation motivates us to develop an unsupervised denoising algorithm in the Fourier domain.

In our proposed Fourier masking strategy, we randomly replace some Fourier coefficients of the center frame with those from neighboring frames to construct a frame $\tilde{x}_0$, which is

$$\tilde{x}_0 = g(\mathbf{x}) = \mathscr{F}^{-1}\left(\mathscr{F}(x_0)\cdot(1-m_2) + \left(\mathscr{F}(x_{-1})\cdot m_1 + \mathscr{F}(x_1)\cdot(1-m_1)\right)\cdot m_2\right), \tag{2}$$

where m_1 and m_2 are binary mask matrices whose entries are drawn from Bernoulli distributions with probabilities p_1 and p_2, respectively. In our implementation, we set $p_1 = 0.2$ and $p_2 = 0.5$. The mask m_2 controls which Fourier coefficients are replaced, and m_1 determines whether the replacement comes from the previous frame x_{-1} or the next frame x_1. We then use the masked Fourier coefficients of the original center frame, $m_2\mathscr{F}(x_0)$, as the supervision signal for training. The corresponding loss function is

$$L\big(f_\theta(\tilde{\mathbf{x}})\big) = \left\|\left(\mathscr{F}(f_\theta(\tilde{\mathbf{x}})) - \mathscr{F}(x_0)\right)\cdot m_2\right\|_1, \tag{3}$$

where $\tilde{\mathbf{x}} = [x_{-T}, \ldots, \tilde{x}_0, \ldots, x_T]$ and $\|\cdot\|_1$ denotes the ℓ_1 norm. Under our noise assumptions, the noise in the input $\tilde{\mathbf{x}}$ and the supervision signal are independent. Based on the theoretical analysis in our previous work,[15] the optimal solution of this objective would be $\hat{z}_0^\star = z_0$.

However, the center frame in the network input is $\tilde{x}_0$ rather than x_0. Ideally, the denoised output should be $\tilde{z}_0$ (the clean version of $\tilde{x}_0$), not z_0. When consecutive frames are nearly identical, z_0 and $\tilde{z}_0$ are very close, so making the network output approximate z_0 is acceptable. But when there is significant misalignment between consecutive frames (e.g., due to fast motion), the network may produce an output inconsistent with the input. To address this, we estimate the discrepancy between z_0 and $\tilde{z}_0$ and correct the supervision signal accordingly. More specifically, we obtain the corrected target via

$$\hat{x}_0 = x_0 + g(\hat{\mathbf{z}}) - \hat{z}_0, \tag{4}$$

where $\hat{\mathbf{z}} = [\hat{z}_{-T}, \ldots, \hat{z}_0, \ldots, \hat{z}_T]$ and $\hat{z}_t$ is the estimate of z_t with the network under training. Notably, the gradient backpropagation is stopped to stabilize the training process while correcting the target. This leads to a regularization term as

$$R_{\text{align}}\big(f_\theta(\tilde{\mathbf{x}})\big) = \left\|\left(\mathscr{F}(f_\theta(\tilde{\mathbf{x}})) - \mathscr{F}(\hat{x}_0)\right)\cdot m_2\right\|_1. \tag{5}$$

Furthermore, we observe small periodic high-frequency artifacts in smooth regions of the output when this regularization term is applied. To suppress them, we adopt a total variation (TV)[17] regularization term, i.e.,

$$R_{\text{TV}}\big(f_\theta(\tilde{\mathbf{x}})\big) = \|\nabla f_\theta(\tilde{\mathbf{x}})\|_1, \tag{6}$$

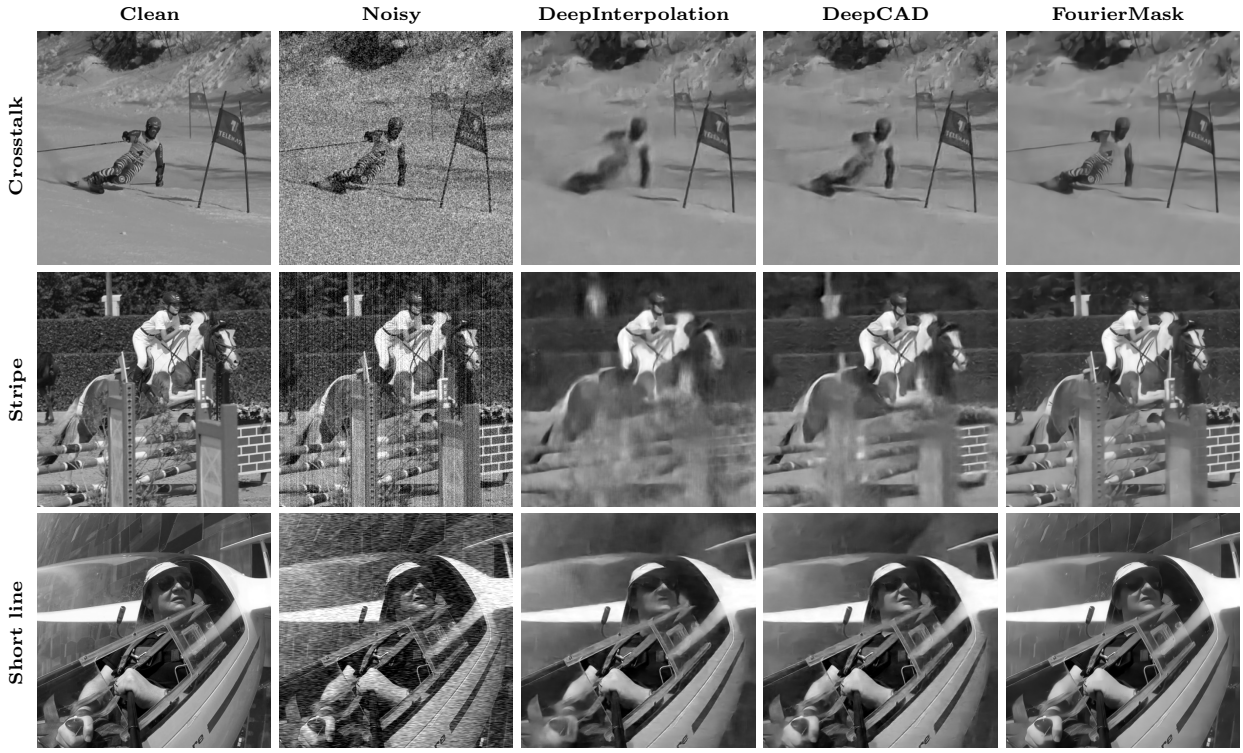

Figure 1. Visual comparison of unsupervised video denoising techniques on three types of structured noise.

Table 1. Average PSNR(dB) and SSIM of unsupervised video denoising techniques on three types of structured noise.

Method	Crosstalk		Stripe		Short line	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DeepInterpolation	25.63	0.7307	25.58	0.7357	25.28	0.7266
DeepCAD	24.19	0.7184	24.29	0.7306	24.15	0.7215
FourierMask (Ours)	29.31	0.8073	30.37	0.8431	30.28	0.8617

where ∇ is the gradient operator. Therefore, the final objective is

$$\mathcal{L} = L + \lambda_{\text{align}} R_{\text{align}} + \lambda_{\text{TV}} \cdot R_{\text{TV}}, \tag{7}$$

where λ_{align} and λ_{TV} are weighting factors. With this combined loss, the network learns to effectively remove structured noise while preserving content consistency and avoiding spurious high-frequency patterns.

3. EXPERIMENTAL RESULTS

To validate the effectiveness of our proposed method (denoted as FourierMask), we compared it with two baseline methods, DeepInterpolation[11] and DeepCAD,[12] on synthetic data. Three types of structured noise are considered: (1) crosstalk noise with only local spatial correlation, (2) stripe noise spanning entire columns, and (3) short-line noise whose correlation range lies between the previous two. Noisy frames are generated by adding synthetic noise to the gray images converted from source images from the DAVIS[18] dataset, which contains 90 video sequences for training and 30 sequences for testing. We set $T = 2$, so the input consists of $2T + 1 = 5$ consecutive frames. The network architecture is the ResNet.[15] We use the Adam[19] optimizer with an initial learning rate of 0.0001, which is halved after $200,000$ steps. The total number of training steps is $300,000$.

For quantitative evaluation, we report the average Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) of the denoising results in Table 1. For visual assessment, representative denoising results are provided in Fig. 1. As Table 1 indicates, FourierMask outperforms the two baselines by more than 5 dB in PSNR. Furthermore, Fig. 1 demonstrates that our method produces substantially sharper results with richer texture and edge details compared to DeepInterpolation and DeepCAD, confirming the superiority of our approach.

4. CONCLUSION

In this paper, we have presented an unsupervised video denoising method tailored for structured noise, including crosstalk, stripe, and short-line patterns. Unlike existing spatial-domain approaches, our method operates in the Fourier domain, where the coefficients of structured noise become approximately independent. The core idea is a random Fourier masking strategy that constructs training pairs from neighboring frames without requiring clean ground-truth data. To address misalignment between consecutive frames, we introduce a consistency-enforcing regularization term, complemented by TV regularization to suppress high-frequency artifacts. Experiments on three types of structured noise demonstrate the superiority of our method over existing unsupervised baselines.

ACKNOWLEDGMENTS

This work is supported in part by the Research Grants Council of Hong Kong (GRF 17201822) and the Theme-based Research Scheme (T45-701/22-R).

REFERENCES

[1] Janesick, J. R., [*Photon transfer*], SPIE press (2007).

[2] Blockstein, L. and Yadid-Pecht, O., "Crosstalk quantification, analysis, and trends in cmos image sensors," *Applied optics* **49**(24), 4483–4488 (2010).

[3] Wang, W., Chen, X., Yang, C., Li, X., Hu, X., and Yue, T., "Enhancing low light videos by exploring high sensitivity camera noise," in [*IEEE/CVF International Conference on Computer Vision*], 4111–4119 (2019).

[4] Chang, Y., Chen, M., Yan, L., Zhao, X.-L., Li, Y., and Zhong, S., "Toward universal stripe removal via wavelet-based deep convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing* **58**(4), 2880–2897 (2019).

[5] Chen, R., Wang, C., Li, Y., Cao, Y., Zhu, S., and Lam, E. Y., "Self-calibrated neuromorphic hyperspectral derivative imaging," *Optica* **13**(4), 587–590 (2026).

[6] Mäkinen, Y., Azzari, L., and Foi, A., "Collaborative filtering of correlated noise: Exact transform-domain variance for improved shrinkage and patch matching," *IEEE Transactions on Image Processing* **29**, 8339–8354 (2020).

[7] Chen, N., Wu, Y., Tan, C., Cao, L., Wang, J., and Lam, E. Y., "Uncertainty-aware Fourier ptychography," *Light: Science & Applications* **14**(1), 236 (2025).

[8] Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L., "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing* **26**(7), 3142–3155 (2017).

[9] Zhang, Y., Chan, S. H., and Lam, E. Y., "Photon-starved snapshot holography," *APL Photonics* **8**(5) (2023).

[10] Zhang, S., Meng, N., and Lam, E. Y., "LRT: An efficient low-light restoration transformer for dark light field images," *IEEE Transactions on Image Processing* **32**, 4314–4326 (2023).

[11] Lecoq, J., Oliver, M., Siegle, J. H., Orlova, N., Ledochowitsch, P., and Koch, C., "Removing independent noise in systems neuroscience data using DeepInterpolation," *Nature methods* **18**(11), 1401–1408 (2021).

[12] Li, X., Zhang, G., Wu, J., Zhang, Y., Zhao, Z., Lin, X., Qiao, H., Xie, H., Wang, H., Fang, L., et al., "Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised denoising," *Nature methods* **18**(11), 1395–1400 (2021).

[13] Sheth, D. Y., Mohan, S., Vincent, J. L., Manzorro, R., Crozier, P. A., Khapra, M. M., Simoncelli, E. P., and Fernandez-Granda, C., "Unsupervised deep video denoising," in [*IEEE/CVF international conference on computer vision*], 1759–1768 (2021).

[14] Wang, Z., Zhang, Y., Zhang, D., and Fu, Y., "Recurrent self-supervised video denoising with denser receptive field," in [*ACM International Conference on Multimedia*], 7363–7372 (2023).

[15] Liu, H., Liu, J., Tan, S., and Lam, E. Y., "Image restoration learning via noisy supervision in the Fourier domain," *arXiv preprint arXiv:2506.00564* (2025).

[16] Fehrenbach, J., Weiss, P., and Lorenzo, C., "Variational algorithms to remove stationary noise: applications to microscopy imaging," *IEEE Transactions on Image Processing* **21**(10), 4420–4430 (2012).

[17] Rudin, L. I., Osher, S., and Fatemi, E., "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena* **60**(1-4), 259–268 (1992).

[18] Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., and Van Gool, L., "The 2017 davis challenge on video object segmentation," *arXiv:1704.00675* (2017).

[19] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," in [*International Conference on Learning Representations*], 1–15 (2014).