This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Generalizable Unsupervised Microscopy Video Denoising via Weighted SpatioTemporal Sampling

Mary Aiyetigbo[1] Wanqi Yuan[1] Feng Luo[1] Xin Li[2]
Tong Ye[1] Nianyi Li[1]
[1]School of Computing, Clemson University
[2]Department of Computer Science, University at Albany - State University of New York

Abstract

Medical video denoising is essential for improving image quality and enhancing the reliability of clinical data. However, the limited availability of annotated datasets, the variability of noise patterns, and the absence of ground-truth reference frames present significant challenges for traditional supervised denoising approaches. On the other hand, current unsupervised video denoising methods often struggle to balance noise removal and motion preservation, leading to either excessive smoothing that degrades fine details or insufficient denoising that leaves residual noise. To address these challenges, we propose STS-UVD, a novel unsupervised video denoising (UVD) method that removes noise while preserving motion integrity. By refining the optical flow, our method ensures temporal consistency without compromising important motion details. Additionally, STS-UVD demonstrates strong generalization across different noise conditions and datasets, making it a robust solution for medical video analysis. Extensive experiments validate its effectiveness in enhancing video quality while maintaining structural and temporal coherence.

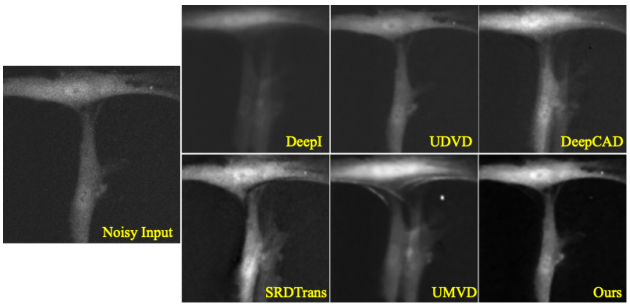


Figure 1. Comparison of denoising results on fluorescence microscopy data with fast-moving cells. The leftmost image represents the noisy input, while the remaining images show results from various denoising methods, including DeepInterpolation (DeepI) [15], UDVD [28], DeepCAD [18], SRDTrans [19], UMVD [1], and our proposed approach (Ours). Our method effectively removes noise while preserving fine structural details and maintaining temporal consistency. In contrast, other state-of-the-art methods exhibit varying levels of blurring, structure loss, or artifacts, particularly in highly dynamic regions.

1. Introduction

Medical imaging is essential for clinical diagnosis and biomedical research, allowing clinicians and researchers to visualize internal anatomical structures non-invasively [4, 6, 22]. Modalities such as fluorescence microscopy, calcium imaging, computed tomography (CT), and positron emission tomography (PET) frequently exhibit significant noise, originating from hardware limitations, environmental factors, and inherent imaging characteristics [11, 24, 27, 40]. Such noise severely compromises image quality, hindering accurate clinical interpretation and reliable automated analysis. In medical videos, where temporal dynamics are critical—such as tracking cellular motion or monitoring physiological changes, noise can severely degrade visual quality, obscure fine details, and hinder clinical interpretation and automated analysis [40]. Effective medical video denoising is thus essential to ensure accurate diagnostics and reliable quantitative assessments, particularly in dynamic scenarios where noise can mask critical spatiotemporal information.

Current state-of-the-art (SOTA) denoising methods, primarily driven by deep learning, have achieved notable success in general video denoising tasks [9, 20, 21, 30, 31]. Techniques leveraging architectures like convolutional neural networks with skip connections [26] and vision transformers [8] have significantly improved denoising quality. However, these traditional algorithms struggle when applied to medical videos. Medical imaging often features unique noise characteristics such as Poisson noise or photon-shot noise in fluorescence microscopy—that differ from Gaussian noise assumptions in natural video denoising [40]. Moreover, the need for paired noisy and clean data in supervised training poses a significant barrier, as acquiring

such ground-truth data in medical contexts is challenging due to ethical, practical, and technical limitations [3, 12]. Consequently, these methods tend to oversmooth critical structural details or fail to adapt to the complex noise distributions and motion patterns inherent in medical videos.

To overcome these limitations, unsupervised denoising techniques that exploit inherent spatial and temporal redundancies have emerged. Approaches such as Noise2Void, UDVD, DeepCAD, and DeepInterpolation [3, 12, 13, 15, 18, 28] implement blind-spot strategies, reconstructing masked pixels or frames using neighboring context to avoid direct noise learning. Others, including Noise2SR, SRDTrans, and adaptive spatiotemporal filtering methods [1, 2, 5, 10, 16, 19, 33], generate subsampled noisy pairs as network inputs and targets, thus eliminating dependence on clean data. Despite their merits, many existing methods struggle with scenarios involving complex or rapid motion, such as in fluorescence microscopy videos [19, 37].

In this paper, we introduce a novel Weighted Spatio-Temporal Sampling (STS) framework specifically tailored for medical video denoising. Our approach dynamically adapts to varying motion patterns by recurrently optimizing and refining optical flow estimates, effectively balancing noise reduction with the preservation of critical structural details. Extensive evaluations on multiple medical imaging datasets demonstrate that our method consistently outperforms existing SOTA methods, offering superior denoising performance and enhanced generalizability across diverse medical imaging modalities and noise conditions.

2. Related Work

In this section, we review the relevant literature on both supervised and unsupervised video denoising. Note that we do not survey the large literature on image denoising algorithms [14, 17, 23, 36, 38, 39] as our method is primarily concerned with video denoising techniques.

Supervised Video Denoising. Recent advancements in video denoising have been driven by the superior performance of supervised deep learning-based approaches [9, 17, 23, 31, 39]. FastDVDnet [31] utilizes UNet [26] blocks to process five successive frames, denoising the central frame. RCD [41] enhances CNN-based denoising by generating multi-level denoised outputs in a single pass. Meanwhile, transformer-based models like VRT [20] and RVRT [21] employ attention mechanisms to capture long-range dependencies, further improving denoising performance. However, these supervised methods face significant challenges in medical imaging applications. They require paired clean/noisy datasets for training, which are often unavailable. Additionally, these models are typically optimized for specific noise distributions, relying on predefined hyperparameters tailored to particular noise characteristics. This specialization limits their adaptability in medical imaging, where noise is often a complex mixture of Gaussian and Poisson components, with unknown distributions due to the lack of ground truth.

SpatioTemporal Sampling in Unsupervised Video Denoising. SOTA self-supervised and unsupervised approaches are largely inspired by Noise2Void [12], employing blind spot masking to facilitate denoising without direct supervision. This typically follows a uniform spatial sampling strategy, where masked regions are denoised by inferring clean signals from surrounding pixels. For video denoising, this strategy extends to the temporal domain, where methods attempt to adapt blind-spot networks to video sequences. Frame2Frame [9] treats consecutive frames as noisy pairs, later improved by Multi-Frame2Frame [7], which leverages multiple frames with FastDVDnet [31] as its backbone. Laine et al. [13] refined this approach by incorporating blind spots directly into the network structure. The state-of-the-art unsupervised video denoising method, UDVD [28], further integrates blind-spot network structure in [13] into Multi-Frame2Frame architecture [7]. RDRF [35] implements a denser receptive field and integrates temporal information from both local and distant neighbors. DeepInterpolation [15] omits the central frame from the input and estimates the denoised central frame using neighboring frames, while DeepCAD [18] forms paired batches, where one batch is used as input and the other as the target, by omitting frames from the input stack. SRDTrans [19] extends this by incorporating a spatial redundancy-based sampling strategy, extracting adjacent orthogonal training pairs. In contrast, UMVD [1] avoids omitting frames and instead employs a fixed linear temporal filter that assigns smaller weights to frames close to the reference while giving higher weights to distant frames to prevent identity noise mapping, setting the central frame's weight to zero.

While these methods perform well in scenarios with slow motion and consistent views across frames, they fail to maintain temporal consistency in videos with fast-moving objects, leading to information loss, motion artifacts, and blurriness in high-speed recordings. Our method introduces a dynamic spatiotemporal kernel that incorporates explicit optical flow information across frames, allowing for adaptive motion-aware filtering. This enables a more direct and efficient denoising process, minimizing information loss and preserving texture integrity. By leveraging an optimized spatial and temporal kernel, our approach effectively balances noise reduction and detail preservation, ensuring robust performance in unsupervised video denoising, particularly for medical video applications where maintaining both spatial and temporal consistency is crucial.

3. STS-based Unsupervised Video Denoising

Given a noisy video sequence $\mathbf{I} \in \mathbb{R}^{T\times H\times W\times C}$ where T, H, W and C are the video length, height, width and

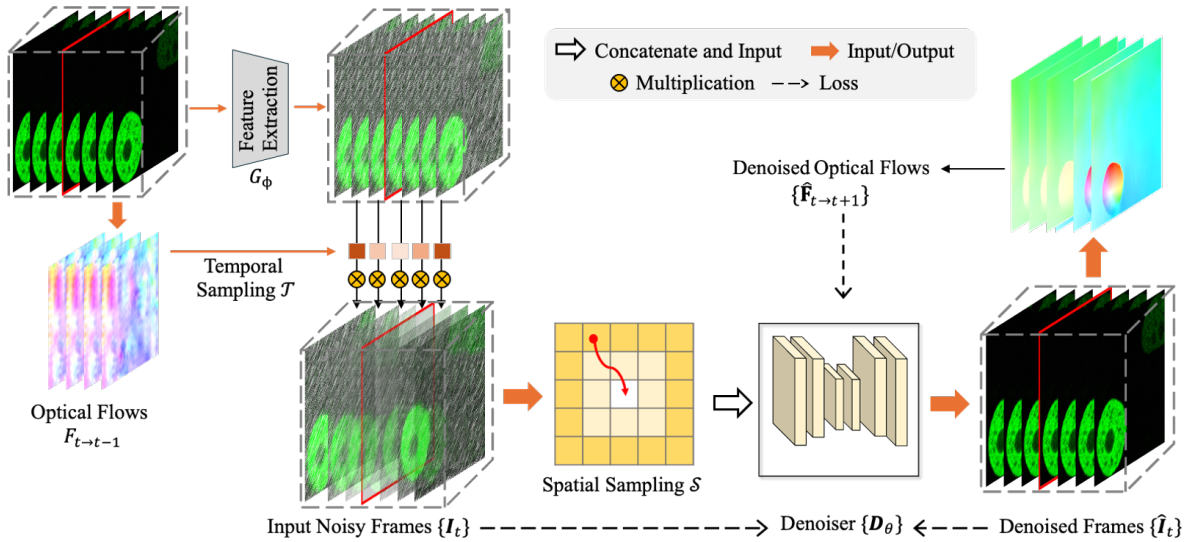


Figure 2. Our STS-UVD network processes consecutive frames in a batch to denoise the central frame. The SpatioTemporal Sampling (STS) module leverages spatial-temporal proximity to the central frame to denoise noisy video, significantly enhancing the generalization ability of the STS-UVD pipeline.

channel, respectively, the goal of STS-UVD is reconstruct a denoised video $\hat{\mathbf{I}} \in \mathbb{R}^{T\times H\times W\times C}$. Our method takes two inputs: video contaminated with unknown noise distribution and the optical flow maps computed using neighboring noisy frames. We denote the input frame as $\{\mathbf{I}_t\}_1^T$, where T is the total number of frames in the input video, and the initial optical flow maps as $\{\mathbf{F}_{t\to t+1}\}$. Our weighted SpatioTemporal Sampling (STS) based unsupervised video denoising method utilizes the coarse $\{\mathbf{F}_{t\to t+1}\}$ to determine the hyperparameters of the temporal sampling kernel $\mathcal{T}$. In each iteration, we concatenate each noisy frame $\mathbf{I}_t$ with its contiguous neighboring frames, denoted as $\{\mathbf{I}_{t+i}\}_{-N/2}^{+N/2}$. These frames are first processed through a shallow feature extraction module G_ϕ, consisting of three layers of group convolutions with 21 channels each, capturing essential low-level features. Next, we apply the initialized temporal sampling kernel $\mathcal{T}$ to refine temporal information, followed by spatial sampling kernel $\mathcal{S}$, which enhances spatial consistency in the sampled batch. The resulting features are then passed through the denoising network to reconstruct the clean frame $\hat{\mathbf{I}}_t$. After each epoch, we employ an optical flow consistency check to the denoised video and use the loss to update the network weights in the next round. This recurrent optimization will stop when the optical flow-based consistency loss does not change anymore. A comprehensive visualization of our pipeline is provided in Figure 2.

3.1. Weighted SpatioTemporal Sampling (STS)

Weighted Temporal Sampling $\mathcal{T}$. Given a noisy frame $\mathbf{I}_t$ along with its contiguous neighboring frames and their corresponding optical flow maps $\mathbf{F}_{t\to t+1}$, we first apply shallow feature extraction on $\mathbf{I}_t$, obtaining $G_\phi(\mathbf{I}_t)$. We then perform temporal sampling using a weighted kernel $\mathcal{T}$. In our approach, $\mathcal{T}$ plays an essential role in leveraging the motion information from optical flow to enhance feature alignment and denoising. This kernel is specifically designed to modulate the pixel values' weights across frames based on their temporal distance from the central frame in the sequence. Specifically, $\mathcal{T}$ assigns weights along the temporal dimension to the modified frame sequence, such that the minimum weight is assigned to the central frame and the maximum weight is assigned to the farthest frames. This weighting scheme is crucial for preventing the network from simply mapping the noise directly from one frame to another, especially in temporally adjacent frames. The temporal weights are computed using the following function:

$$\omega_t = \begin{cases} \kappa + (1-\kappa)\left(1 - \left(\frac{t}{C}\right)^L\right) & \text{if } t \le C, \\ \kappa + (1-\kappa)\left(1 - \left(\frac{2C-t}{C}\right)^L\right) & \text{otherwise,} \end{cases} \tag{1}$$

where ω_t is the weight for frame $\mathbf{I}_t$, t is the index of the frame, C is the index of the central frame, L is a parameter controlling the curvature of the weighting function as shown in Figure 3 (b), and κ is the minimum weight, which diminishes towards zero as the training progresses. The dynamic adjustment of the minimum value κ is controlled by the current epoch of training, ensuring that as the model learns more about the video content, the influence of the central frame reduces, allowing for a more balanced and effective denoising process:

$$\kappa = \max\left(0, \kappa_0\left(1 - \frac{\text{epoch}}{\text{max epoch}}\right)\right). \tag{2}$$

In equation 2, κ_0 is the initial value of κ, "epoch" is the current training epoch and "max epoch" is the total number of epochs planned for training. It is important to note that higher κ_0 will encourage the network taking more information from the immediate neighbors, *i.e.* of smaller temporal receptive fields, while lower κ_0 encourages considering faraway frames. To overcome the disadvantages introduced by the uniform temporal sampling strategy, adopted by UDVD and UMVD, we utilize the averaging optical flows $\{\mathbf{F}_{t\to t+1}\}_1^N$, computed using RAFT [32], to calculate κ_0 in Eqn 3, so that our network can adaptively adjust the shape of $\mathcal{T}$ and better handle videos with different frame rates and object motion speeds.

$$\kappa_0 = 0.2 \cdot \left(1 - e^{-\eta \times \overline{M}^2}\right), \tag{3}$$

where η is a scaling parameter that controls the rate at which κ_0 increases with $\overline{M}$, ensuring that κ_0 remains within the desired range. Specifically, for small $\overline{M}$, κ_0 approaches 0, while for large $\overline{M}$, it converges to 0.2. The value of $\overline{M}$ is the average optical flow magnitude as defined in Eqn 4.

$$\overline{M} = \frac{1}{N \cdot W \cdot H} \sum_{i=1}^{N} \sum_{x=1}^{W} \sum_{y=1}^{H} \sqrt{\mathbf{F}^2_{i\to i+1}(x,y,1) + \mathbf{F}^2_{i\to i+1}(x,y,2)}, \tag{4}$$

where (x, y) is the pixel coordinates. We apply the weights $\{\omega_t | t = 1, ..., N\}$ to the input frames through elementwise multiplication as shown in Eqn 5.

$$\mathbf{I}'_t = \omega_t \odot G_\phi(\mathbf{I}_t) \quad \forall t \in 1, ..., N. \tag{5}$$

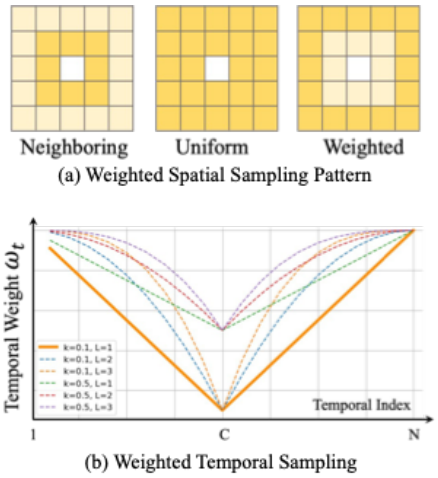
This ensures that frames distant from the central frame exert stronger influence, thus effectively mitigating the direct transfer of noise and enhancing the overall quality of the denoised video.

Weighted Spatial Kernel $\mathcal{S}$. Next, we apply our weighted spatial kernel to individual temporally sampled frames. The primary concept involves selectively modifying a subset of pixels in each frame to enhance denoising efficiency while preserving essential details. Our spatial sampling begins by randomly selecting 10-20% of pixels, denoted as the set $\{\mathbf{p}\}$. Each chosen pixel $\mathbf{p}$ is then replaced with neighboring pixel $\mathbf{q}$ from its neighborhood, $\Omega_\mathbf{p}$ within a 5×5 window:

$$\tilde{\mathbf{I}}_t(\mathbf{p}) = \mathbf{I}'_t(\mathbf{q}). \tag{6}$$

The selection of this replacement pixel is guided by a probability distribution that prioritizes faraway pixels while deemphasizing those closer to the center. The probability of choosing a neighboring pixel $\mathbf{q} \in \Omega_\mathbf{p}$ for replacement is defined as follows:

$$\mathrm{P}(\mathbf{q}|\mathbf{p}) = \frac{e^{-\alpha \cdot d(\mathbf{p},\mathbf{q})}}{\sum_{\mathbf{q}' \in \Omega_p} e^{-\alpha \cdot d(\mathbf{p},\mathbf{q}')}}, \tag{7}$$



(a) Weighted Spatial Sampling Pattern

(b) Weighted Temporal Sampling

Figure 3. Illustration of the Weighted SpatioTemporal sampling patterns. (a) $\mathcal{S}$. Neighboring: samples nearby regions. Uniform: uses equal sampling probability, and Weighted: our proposed approach (darker colors indicate higher sampling rates). (b) $\mathcal{T}$: Curves representing $\mathcal{T}$ for distinct curvature parameter values L and minimum weight κ.

where $d(\mathbf{p}, \mathbf{q})$ represents the l_1 distance between the pixel $\mathbf{p}$ and a neighboring pixel $\mathbf{q}$, and α is a scaling factor that controls the emphasis on edge pixels. We set α as 3, ensuring that pixels closer to the edges (smaller $d(\mathbf{p}, \mathbf{q})$) have a higher probability of being chosen. By adaptively modifying a portion of the pixels in each frame, our weighted spatial kernel, as illustrated in Figure 3 (a) effectively balances noise reduction with detail preservation, making it a crucial component in our video denoising strategy.

After applying the weighted Spatial Kernel to temporally sampled frames, we get:

$$\tilde{\mathcal{I}} = \mathcal{S}(\{\tilde{\mathbf{I}}_t\}_{t=1}^N). \tag{8}$$

We decided to keep the spatial kernel $\mathcal{S}$ exclusively 2D without including the temporal dimension due to two reasons: 1) Noise patterns are generally more consistent within the spatial domain of a frame rather than across time, which is particularly true for sensor-induced noise. Introducing temporal components into $\mathcal{S}$ could potentially dilute this spatial consistency, especially given the *limited frame batches* processed in each optimization iteration, and could inadvertently lead the network away from accurately learning the true color and details of individual pixels. 2) Many SOTAs, such as Noise2Void and UDVD, already implemented spatial denoising mechanisms, whether implicitly or explicitly. Our goal with the STS kernel, particularly the temporal kernel $\mathcal{T}$, is to complement these existing methods by enhancing temporal information processing without disrupting their inherent spatial denoising capabilities.

3.2. Backbone Denoiser D_θ

The denoising network D_θ produces a denoised central frame, denoted as $\hat{\mathbf{I}}_C$:

$$\hat{\mathbf{I}}_C = \mathcal{D}_\theta(\tilde{\mathcal{I}}). \tag{9}$$

We then calculate the optical flows $\{\hat{\mathbf{F}}_{t\to t+1}\}_1^N$ using the denoised frames and their corresponding immediate neighboring frames using RAFT [32]. In our denoising framework, the integration of the proposed weighted SpatioTemporal Sampling (STS) kernel with a five-layer U-Net architecture [26], has shown remarkable efficacy in restoring videos contaminated with various strengths of noise. Our denoiser network consists of two encoder blocks and three decoder blocks, followed by three 1×1 convolution layers with 348, 96, and C filters, respectively, progressively reducing the number of channels to match the input image dimension. Each encoder block comprises five 3×3 convolution layers, each with 48 filters. The decoder blocks, responsible for progressive feature restoration, consist of six convolution layers with 96 filters. Skip connections between encoder and decoder layers facilitate the retention of fine details, reducing information loss during reconstruction. By leveraging this architecture, combined with the STS kernel's adaptive spatiotemporal weighting, our model achieves robust noise suppression while maintaining spatial and temporal consistency as shown in section 5.4.

4. Recurrent Optimization

Our recurrent training strategy for the denoising network $\mathcal{D}_\theta$ aims at optimizing the network parameters to enhance the prediction of both the denoised frames $\{\hat{\mathbf{I}}_t\}_1^N$ (Eq. 9) and their corresponding optical flow maps $\{\hat{\mathbf{F}}_{t\to t+1}\}_1^N$. To effectively guide the network during training, we employ two loss terms. One is the pixel-wise L1 loss between $\hat{\mathbf{I}}_t$ and the reference frame $\mathbf{I}_t^{ref}$, and the second loss term is a consistency loss based on the optical flow estimation of the denoised frames.

$$\mathcal{L}_1 = \sum_{\mathbf{p}\in\mathbf{I}_t} |\hat{\mathbf{I}}_t(\mathbf{p}) - \mathbf{I}_t^{ref}(\mathbf{p})|. \tag{10}$$

$$\mathcal{L}_2 = \frac{1}{N}\sum_{t=1}^{N-1} \left\| W(\hat{\mathbf{I}}_{t+1}, \hat{\mathbf{F}}_{t\to t+1}) - \hat{\mathbf{I}}_t \right\|^2. \tag{11}$$

$W(\hat{\mathbf{I}}_{t+1}, \hat{\mathbf{F}}_{t\to t+1})$ denotes the warping of frame $\hat{\mathbf{I}}_{t+1}$ using the optical flow map $\hat{\mathbf{F}}_{t\to t+1}$, and $\mathcal{L}_2$ is the consistency error across the batch of frames. The final loss function is:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2. \tag{12}$$

Considering that there are no clean reference frames to train the model, we update the reference frame $\mathbf{I}_t^{ref}$ recurrently using the output denoised frame $\mathbf{I}_t$ from the previous epoch. We stop training once the consistency loss $\mathcal{L}_2$ reaches convergence, in this case when it is below a small threshold value δ. These steps are particularly beneficial in an unsupervised setting, as it helps to accentuate the network's focus on reducing noise while preserving essential details, thereby facilitating more effective learning.

5. Experiments

5.1. Datasets

We evaluated our method on two datasets. The first is simulated two-photon (2P) calcium imaging data generated using the Neural Anatomy and Optical Microscopy (NAOMi) simulator [29], which realistically models calcium imaging of the mouse visual cortex. It provides synchronized ground truth and introduces realistic noise through per-pixel Poisson photon sampling and electronic amplification variability, making it a valuable benchmark for evaluating performance under complex noise conditions. We generated two distinct two-photon calcium imaging recordings with varying signal-to-noise ratios (SNRs). The first recording has a field of view (FOV) of $150 \times 150\,\mu m^2$ with 1000 frames, simulating a scenario with lower neural activity. The second recording contains a larger FOV of $500 \times 500\,\mu m^2$ containing 1500 frames, modeling higher neural activity with increased noise levels (lower SNR).

The second dataset is fluorescence microscopy video recordings of live cells from Cell Tracking Challenge [34], we used the Fluo-C2DL-MSC recording, containing two video sequences with frame resolutions of 992×832 and 1200×782, respectively, and the Fluo-N2DH-GOWT1, which consists of two video sequences, each with frame resolution of 1024×1024. These recordings capture rapid cell movement, providing a challenging case to assess the robustness of our approach in high-motion biological imaging scenarios. Particularly for Fluo-C2DL-MSC whose cells have highly irregular shapes and move really fast, which leads to low cell overlaps between consecutive frames.

5.2. Experiment Setup.

We benchmarked the STS-UVD against current state-of-the-art unsupervised video denoising techniques: UDVD [28], SRDTrans [19], DeepCAD [18], DeepInterpolation [15], and UMVD [1]. To ensure a fair comparison, we trained each model on each dataset individually. Our architectures were implemented using the PyTorch framework [25] and executed on an NVIDIA A100 GPU. ADAM optimizer, with β_1 and β_2 set to 0.5 and 0.9 respectively, and weight decay of 0.0005, was utilized for minimizing the loss function. We set "max epoch" to 30, η to 0.0003, and L to 1. In each training iteration, the model processes a stack of inputs as mini-batch, focusing on denoising the central frame in the batch. The initial learning rate was set to 0.0001 and consistently reduced by half after every 10 epochs.

	DeepI [15]		DeepCAD [18]		UDVD [28]		SRDTrans [19]		UMVD [1]		Ours	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
$150\,\mu m^2$	30.020	0.911	36.606	0.971	31.889	0.899	41.176	0.984	38.871	0.985	43.970	0.992
$500\,\mu m^2$	34.222	0.907	29.130	0.885	35.775	0.932	34.595	0.924	32.933	0.909	37.501	0.952
Average	32.121	0.909	32.868	0.928	33.832	0.916	<u>37.886</u>	<u>0.954</u>	35.902	0.947	**40.736**	**0.972**

Table 1. Quantitative comparison of our method with SOTA video denoising techniques on simulated two-photon calcium imaging datasets with varying fields of view (FOV). Text highlighted in **bold** signifies the highest value, while <u>underlined</u> text denotes the second highest.

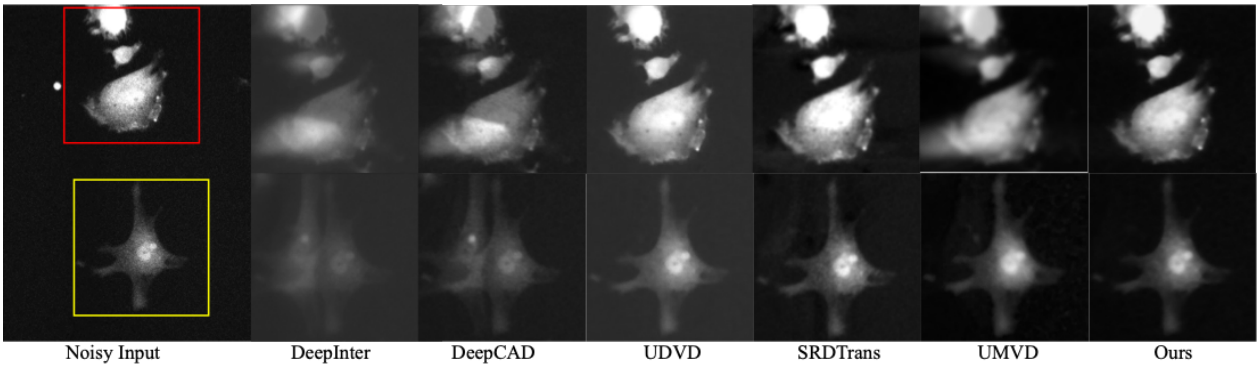


Figure 4. Visual Comparison on Fluo-C2DL-MSC fluorescence microscopy recording. *Best viewed when zoomed in.*

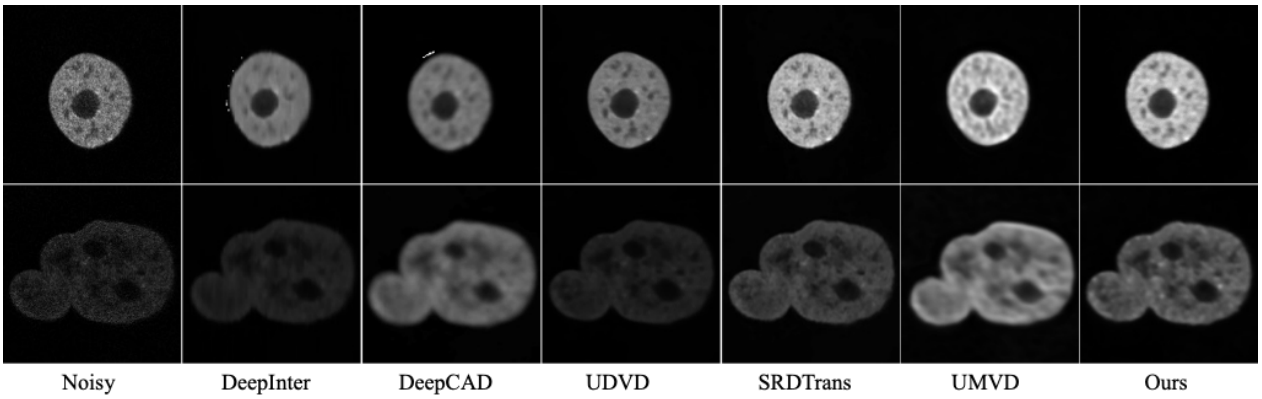


Figure 5. Visual Comparison on Fluo-N2DH-GOWT1 fluorescence microscopy recording. *Best viewed when zoomed in.*

5.3. Quantitative Evaluation

To comprehensively evaluate the effectiveness of our proposed video denoising methodology, we employed two widely recognized benchmark metrics in image and video processing: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). The quantitative evaluation was conducted on simulated two-photon (2P) calcium imaging data, as it provides access to clean ground truth for objective assessment. Table 1 compares our method against DeepI, DeepCAD, UDVD, SRDTrans, and UMVD across two fields of view (FOVs): $150\mu m^2$ and $500\mu m^2$. Our approach consistently achieves the highest PSNR and SSIM, demonstrating superior noise suppression and structural preservation. Specifically, we achieve 43.97 dB PSNR, 0.992 SSIM for $150\mu m^2$ and 37.501 dB PSNR, 0.952 SSIM for $500\mu m^2$, outperforming all baselines. On average, our method attains 40.736 dB PSNR, 0.972 SSIM, significantly surpassing SRDTrans. While other methods tend to over-smooth details (DeepI, DeepCAD) or fail to suppress noise effectively (UDVD, UMVD), our approach strikes a better balance between denoising performance and structure preservation, making it highly effective for fluorescence microscopy imaging.

5.4. Qualitative Comparison

We present a qualitative evaluation of our method against SOTA unsupervised video denoising techniques. In high-speed fluorescence microscopy from the Fluo-N2DH-MSC dataset where cells exhibit rapid motion, several existing methods, including DeepI, DeepCAD, and UMVD struggle with temporal inconsistencies, resulting in ghosting artifacts and duplicate structures, as shown in Figure 4. These artifacts occur because these models do not account for motion, leading to misaligned reconstructions for fast-moving

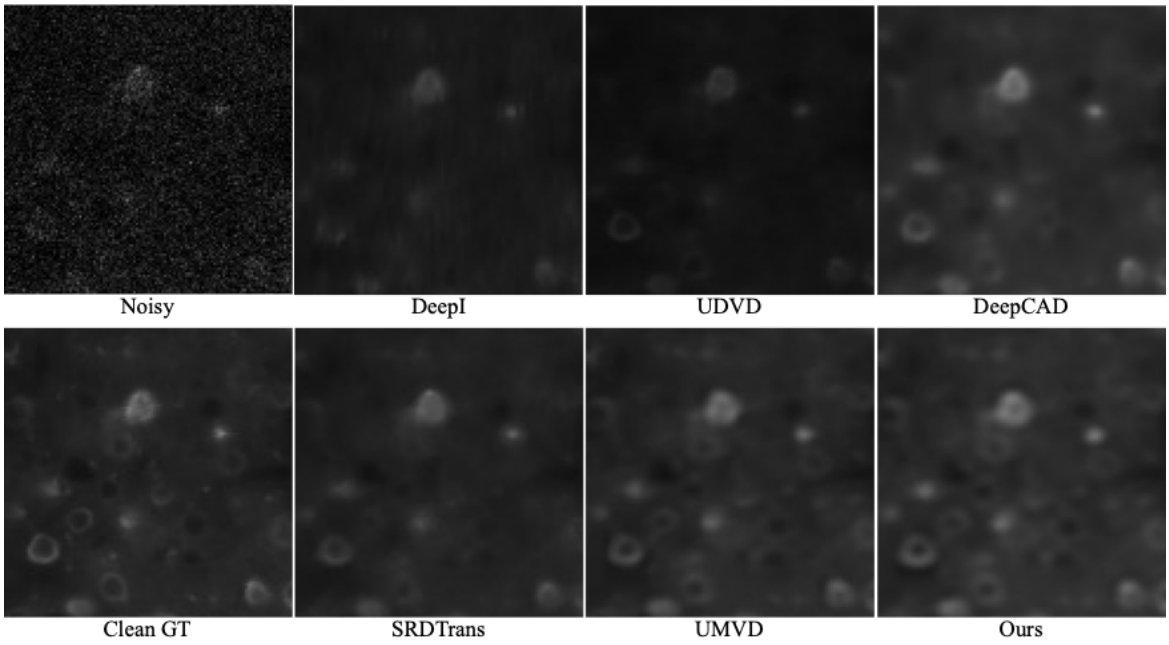


Figure 6. Visual Comparison on two-photon calcium imaging with FOV of $500\,\mu m^2$. *Best viewed when zoomed in.*

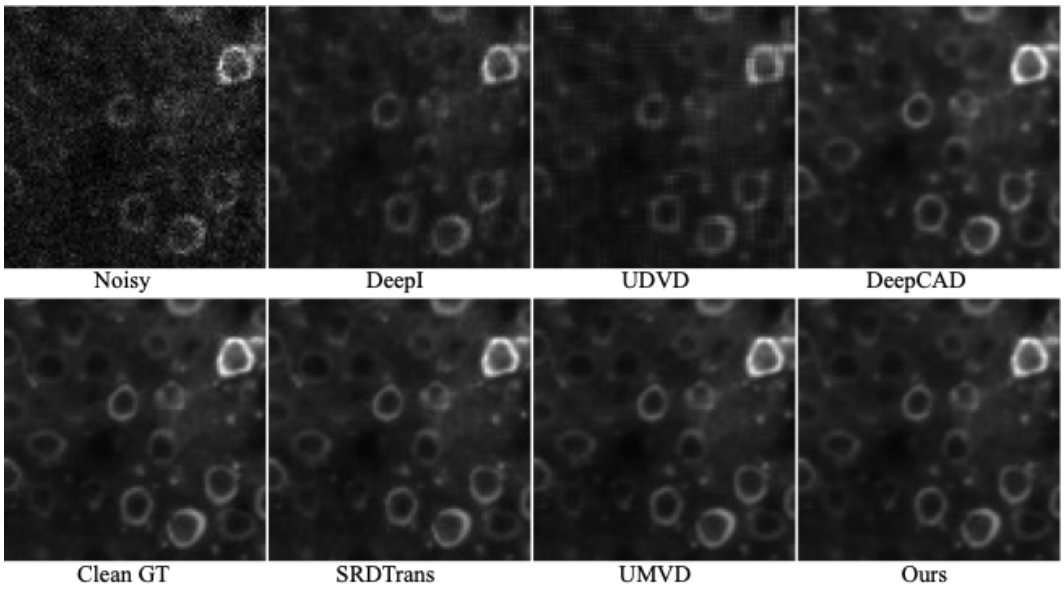


Figure 7. Visual Comparison on two-photon calcium imaging with FOV of $150\,\mu m^2$. *Best viewed when zoomed in.*

cellular components. Additionally, SRDTrans introduces noticeable background artifacts. In contrast, our method leverages optical flow-based temporal regularization to enforce motion coherence, effectively mitigating motion artifacts while preserving fine structural details.

Figure 5 presents denoising results on the Fluo-N2DH-GOWT1 fluorescence microscopy dataset. While existing methods tend to produce over-smoothed reconstructions that obscure fine structures, our approach preserves sharper and more detailed cellular features. Notably, although SRDTrans demonstrates relatively strong performance among the state-of-the-art methods, it still fails to fully recover the fine-grained details—particularly evident in the second row. In contrast, our method consistently produces clearer and more structurally detailed reconstructions. For two-photon calcium imaging, shown in Figures 6 and 7, varying signal-to-noise ratios (SNRs) introduce additional challenges. Methods such as DeepI, UDVD, DeepCAD, and SRDTrans exhibit significant blurring and are unable to fully reconstruct all the underlying intricate neural struc-

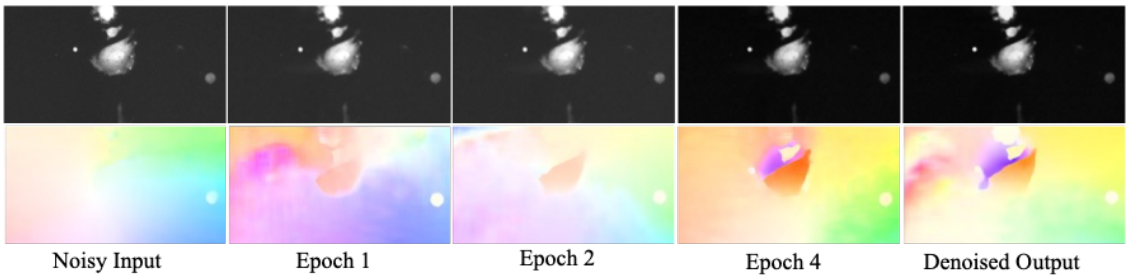


Figure 8. Denoised results and optical flow estimates at different epochs

tures. In contrast, our method achieves superior noise suppression while effectively preserving cellular morphology, making it robust across both high-SNR and low-SNR conditions. The integration of optical flow constraints enables accurate motion tracking, preventing ghosting artifacts and structural duplication, making our method highly effective for real-world microscopy applications.

5.5. Optical Flow Outputs

Figure 8 presents the denoising results at various epochs during the training process, specifically at epochs 1, 2, 4, and the final epoch. For each epoch, the figure displays the denoised image alongside its corresponding optical flow beneath it. The first column shows the noisy input frame and the computed optical flow estimation, while the last column illustrates the denoised frame along with its accurate flow. These results demonstrate that our model effectively produces high-quality denoised outputs and accurately estimates the optical flow for the denoised frames.

5.6. Ablation Study

We performed ablation studies to assess the impact of different design choices in our network architecture. Quantitative experiments were conducted on the simulated 2P calcium imaging dataset with a FOV of $150, \mu m^2$, evaluating the denoising performance of various network variants.

Network architecture. We analyze the contributions of different components within our framework to the overall denoising performance, by evaluating four variations of our network: 1) D_θ – a baseline using only the U-Net backbone for denoising; 2) $\mathcal{S} + D_\theta$ – incorporates a weighted spatial sampling kernel into the U-Net backbone; 3) $\mathcal{T} + D_\theta$ – integrates a weighted temporal kernel with the backbone denoiser and 4) $\mathcal{T} + \mathcal{S} + D_\theta$ – our full pipeline, as detailed in Table 2. To prevent direct noise mapping, we exclude the temporal kernel in the relevant variant in (2) by removing the target frame from the input batch. Figure 9 illustrates the effectiveness of our STS kernel in maintaining temporal consistency and enhancing denoising performance on Fluo-N2DH-MSC fluorescence microscopy. The D_θ-only model directly maps noise, while $\mathcal{S} + D_\theta$, lacking temporal consistency, results in blurry and inaccurate reconstructions. In contrast, our full pipeline ($\mathcal{T}+\mathcal{S}+D_\theta$) preserves fine details and ensures higher-quality denoising.

	D_θ	$\mathcal{S}+D_\theta$	$\mathcal{T}+D_\theta$	$\mathcal{S}+\mathcal{T}+D_\theta$
PSNR	30.440	37.589	42.168	**43.97**
SSIM	0.887	0.976	0.990	**0.992**

Table 2. Ablation study on network variants.

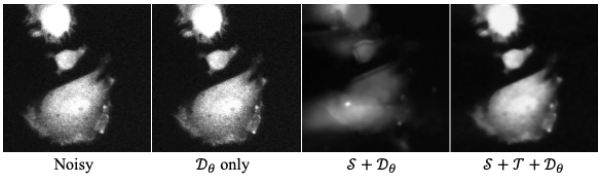


Figure 9. Denoised results on various network variants

6. Conclusion

We presented a novel recurrent unsupervised video denoising pipeline using an explicitly designed Weighted SpatioTemporal Sampling strategy. To better accommodate fast movements, we integrate optical flow estimation to dynamically adjust STS parameters, significantly improving performance in scenarios with high motion. Through extensive experiments, our model demonstrates superior results compared to SOTA methods across diverse medical imaging datasets. The weighted SpatioTemporal Sampling (STS) filter played a critical role in enhancing the denoising quality, emphasizing the importance of spatiotemporal feature adaptation in video denoising. While our approach leverages RAFT-based optical flow for robust motion handling, we acknowledge that its computational cost may pose challenges for extremely high-resolution videos or real-time applications. Future work will explore lightweight flow estimation networks or adaptive sampling strategies to further improve efficiency without compromising performance.

Acknowledgments. This work was supported in part by the South Carolina EPSCoR/IDeA program (SC EPSCoR/IDeA) under award number 25-GA02 and by the National Science Foundation (NSF) under grant IIS-2232298. The authors gratefully acknowledge this support.

References

[1] Mary Aiyetigbo, Alexander Korte, Ethan Anderson, Reda Chalhoub, Peter Kalivas, Feng Luo, and Nianyi Li. Unsupervised microscopy video denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2024. 1, 2, 5, 6

[2] Carole Amiot, Catherine Girard, Jocelyn Chanussot, Jérémie Pescatore, and Michel Desvignes. Spatio-temporal multiscale denoising of fluoroscopic sequence. *IEEE transactions on medical imaging*, 35(6):1565–1574, 2016. 2

[3] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pages 524–533. PMLR, 2019. 2

[4] John A Calarco and Aravinthan DT Samuel. Imaging whole nervous systems: insights into behavior from worms to fish. *Nature methods*, 16(1):14–15, 2019. 1

[5] H-Y Cheong, Alexis M Tourapis, Joan Llach, and Jill Boyce. Adaptive spatio-temporal filtering for video denoising. In *2004 International Conference on Image Processing, 2004. ICIP'04.*, pages 965–968. IEEE, 2004. 2

[6] Kevin J Coakley and Jorge Llacer. Use of cross-validation as a stopping rule in emission tomography image reconstruction. In *Medical Imaging V: Image Physics*, pages 226–233. SPIE, 1991. 1

[7] Valéry Dewil, Jérémy Anger, Axel Davy, Thibaud Ehret, Gabriele Facciolo, and Pablo Arias. Self-supervised training for blind multi-frame video denoising. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2724–2734, 2021. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[9] Thibaud Ehret, Axel Davy, Jean-Michel Morel, Gabriele Facciolo, and Pablo Arias. Model-blind video denoising via frame-to-frame training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11369–11378, 2019. 1, 2

[10] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14781–14790, 2021. 2

[11] Na Ji, Jeremy Freeman, and Spencer L Smith. Technologies for imaging neural activity in large volumes. *Nature neuroscience*, 19(9):1154–1164, 2016. 1

[12] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2129–2137, 2019. 2

[13] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[14] Marc Lebrun, Antoni Buades, and Jean-Michel Morel. A nonlocal bayesian image denoising algorithm. *SIAM Journal on Imaging Sciences*, 6(3):1665–1688, 2013. 2

[15] Jérôme Lecoq, Michael Oliver, Joshua H Siegle, Natalia Orlova, Peter Ledochowitsch, and Christof Koch. Removing independent noise in systems neuroscience data using deepinterpolation. *Nature methods*, 18(11):1401–1408, 2021. 1, 2, 5, 6

[16] Seong-Won Lee, Vivek Maik, Jihoon Jang, Jeongho Shin, and Joonki Paik. Noise-adaptive spatio-temporal filter for real-time noise removal in low light level images. *IEEE Transactions on Consumer Electronics*, 51(2):648–653, 2005. 2

[17] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018. 2

[18] Xinyang Li, Guoxun Zhang, Jiamin Wu, Yuanlong Zhang, Zhifeng Zhao, Xing Lin, Hui Qiao, Hao Xie, Haoqian Wang, Lu Fang, et al. Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised denoising. *Nature methods*, 18(11):1395–1400, 2021. 1, 2, 5, 6

[19] Xinyang Li, Xiaowan Hu, Xingye Chen, Jiaqi Fan, Zhifeng Zhao, Jiamin Wu, Haoqian Wang, and Qionghai Dai. Spatial redundancy transformer for self-supervised fluorescence image denoising. *Nature Computational Science*, 3(12):1067–1080, 2023. 1, 2, 5, 6

[20] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 1, 2

[21] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022. 1, 2

[22] Wasim Q Malik, James Schummers, Mriganka Sur, and Emery N Brown. Denoising two-photon calcium imaging data. *PloS one*, 6(6):e20490, 2011. 1

[23] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems*, 29, 2016. 2

[24] Issei Mori, Yoshio Machida, Makoto Osanai, and Kazuhiro Iinuma. Photon starvation artifacts of x-ray ct: their true cause and a solution. *Radiological physics and technology*, 6:130–141, 2013. 1

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted*

Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pages 234–241. Springer, 2015. 1, 2, 5

[27] Joshua D Schaefferkoetter, Jianhua Yan, Therese Sjöholm, David W Townsend, Maurizio Conti, John Kit Chung Tam, Ross A Soo, and Ivan Tham. Quantitative accuracy and lesion detectability of low-dose 18f-fdg pet for lung cancer screening. *Journal of Nuclear Medicine*, 58(3):399–405, 2017. 1

[28] Dev Yashpal Sheth, Sreyas Mohan, Joshua L Vincent, Ramon Manzorro, Peter A Crozier, Mitesh M Khapra, Eero P Simoncelli, and Carlos Fernandez-Granda. Unsupervised deep video denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1759–1768, 2021. 1, 2, 5, 6

[29] Alexander Song, Jeff L Gauthier, Jonathan W Pillow, David W Tank, and Adam S Charles. Neural anatomy and optical microscopy (naomi) simulation for evaluating calcium imaging methods. *Journal of neuroscience methods*, 358:109173, 2021. 5

[30] Matias Tassano, Julie Delon, and Thomas Veit. Dvdnet: A fast network for deep video denoising. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1805–1809. IEEE, 2019. 1

[31] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1354–1363, 2020. 1, 2

[32] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 4, 5

[33] Xuanyu Tian, Qing Wu, Hongjiang Wei, and Yuyao Zhang. Noise2sr: Learning to denoise from super-resolved single noisy fluorescence image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 334–343. Springer, 2022. 2

[34] Vladimír Ulman, Martin Maška, Klas EG Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, et al. An objective comparison of cell-tracking algorithms. *Nature methods*, 14(12):1141–1152, 2017. 5

[35] Zichun Wang, Yulun Zhang, Debing Zhang, and Ying Fu. Recurrent self-supervised video denoising with denser receptive field. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7363–7372, 2023. 2

[36] Jun Xu, Lei Zhang, Wangmeng Zuo, David Zhang, and Xiangchu Feng. Patch group based nonlocal self-similarity prior learning for image denoising. In *Proceedings of the IEEE international conference on computer vision*, pages 244–252, 2015. 2

[37] Zhiwen Yang, Haowei Chen, Ziniu Qian, Yang Yi, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. All-in-one medical image restoration via task-adaptive routing. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 67–77. Springer, 2024. 2

[38] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning patch-wise non-local strengths for image denoising. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2773–2780. IEEE, 2011. 2

[39] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 2

[40] Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. A poisson-gaussian denoising dataset with real fluorescence microscopy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11710–11718, 2019. 1

[41] Zhaoyang Zhang and et al. Real-time controllable denoising for image and video. In *CVPR*, pages 14028–14038, 2023. 2