Unsupervised Microscopy Video Denoising

Mary Aiyetigbo[1] Alexander Korte[1] Ethan Anderson[1] Reda Chalhoub[2]
Peter Kalivas[2] Feng Luo[1] Nianyi Li[1]

[1]School of Computing, Clemson University
[2]Department of Neuroscience, The Medical University of South Carolina

Abstract

In this paper, we introduce a novel unsupervised network to denoise microscopy videos featured by image sequences captured by a fixed location microscopy camera. Specifically, we propose a DeepTemporal Interpolation method, leveraging a temporal signal filter integrated into the bottom CNN layers, to restore microscopy videos corrupted by unknown noise types. Our unsupervised denoising architecture is distinguished by its ability to adapt to multiple noise conditions without the need for pre-existing noise distribution knowledge, addressing a significant challenge in real-world medical applications. Furthermore, we evaluate our denoising framework using both real microscopy recordings and simulated data, validating our outperforming video denoising performance across a broad spectrum of noise scenarios. Extensive experiments demonstrate that our unsupervised model consistently outperforms state-of-the-art supervised and unsupervised video denoising techniques, proving especially effective for microscopy videos. The project page is available at https://maryaiyetigbo.github.io/UMVD/

1. Introduction

In biomedical research, microscopy imaging have enabled scientists to observe and analyze biological structures at cellular and molecular levels [1, 8, 9]. These videos, however, are characteristically captured at fixed locations, limiting the field of view to specific areas of interest within the neural landscape. This fixed location feature, while beneficial for focused studies, introduces unique challenges in video denoising, leading to persistent noise patterns that are difficult to differentiate from actual biological signals [3, 4, 21, 23, 35].

In particular, the noise in microscopy videos, emanating from diverse sources such as photon shot noise, background fluorescence, and detector electronics, not only varies in type (Poisson and Gaussian) but also fluctuates in intensity across different frames due to variations in lighting conditions, specimen responses, and camera sensitivity. Traditional video denoising techniques often rely on detecting and leveraging motion between frames to distinguish noise from signal [2, 10–13, 34].

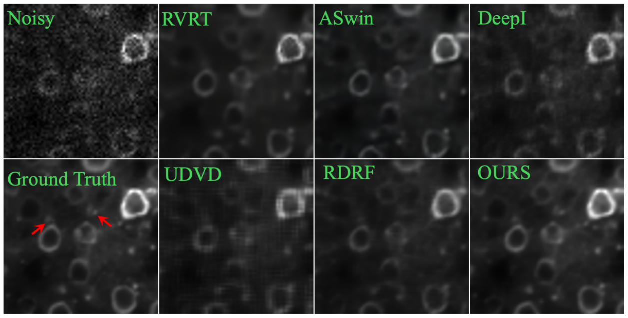

Figure 1. Denoising results of our method, supervised (RVRT [15] ASwin [17]), and unsupervsied SOTAs (Deepinterpolation (DeepI) [12], UDVD [24], RDRF [30]) on simulated two-photon calcium imaging. Our technique showcases superior denoising capabilities. The red arrow indications in the clean image highlight specific signals that were either not adequately reconstructed by other methods in their noise-reduced images or were excessively smoothed

In the context of microscopy videos with fixed locations, the absence of extensive movement can limit the effectiveness of these approaches, making it difficult to apply conventional denoising techniques designed for more dynamic videos. Furthermore, the varying frame rates in microscopy videos, especially calcium imaging recordings, add an additional layer of difficulty, as they require adaptive processing techniques capable of handling these temporal variations without compromising the integrity of the neural signals being observed. Therefore, creating effective denoising methods for microscopy videos that can account for static frames while accurately isolating and eliminating noise is significantly important for biological and neurological research.

Many state-of-the-art (SOTA) video denoising methods [5, 14–17, 27, 28, 31, 33] rely heavily on supervised learning, which require noisy/clean pairs during training. Although they present good denoising performance on videos captured by regular RGB cameras, the supervised video denoising methods show limited generalization ability in medical imaging, where obtaining clean reference images is unrealistic and almost impossible. Various unsupervised approaches address this challenge by optimizing the noisy frames and exploring the motion between frames [2, 10–13, 34]. Nevertheless, SOTA unsupervised video denoising networks, i.e. UDVD [24], Wang et al. [30], have inherent limitations. Firstly, they employ the backbone network architectures of the Multi-Frame2Frame [7, 28] and Laine et al. [11], which extensively exploits information from neighboring pixels, both spatially and temporally, to reconstruct the missing data and eliminate noise. These approaches prove effective primarily for low frame rate videos or videos featuring fast-moving objects or view changes, and easy to overfit to noise when the input videos exhibit slow-moving objects with consistent backgrounds. However, it is common for microscopy videos, e.g., one-photon calcium imaging, containing static but fluctuating noisy background with only a few intermediate signal spikes throughout the recording, thus can pose extreme challenges to these motion-based denoising techniques, as shown in Fig. 1.

In this paper, we present an unsupervised denoising method using DeepTemporal Interpolation, tailored for microscopy videos, which can effectively denoise videos with varied noise types and intensities. Our method consists of two main components: a feature generator ($\mathcal{G}_\phi$) and a Denoiser ($\mathcal{D}_\theta$), as shown in Fig. 2. Distinct from conventional unsupervised methods, which directly use adjacent frames to interpolate the missing central frame, such as DeepInterpolation [12] and Zheng et al. [34], or utilize neighboring pixels for central pixel estimation as seen in UDVD [24] and RDRF[30], our technique employs a temporal filter on the feature maps produced by a sequence of CNN layers ($\mathcal{G}_\phi$) prior to processing by the Denoiser ($\mathcal{D}_\theta$). Our overall pipeline is illustrated in Fig. 2. The major contributions of our work include:

- We propose a novel unsupervised video denoising approach tailored for microscopy videos. Instead of making complex modifications to the network architecture, we apply a temporal filter to the feature map generated by the feature generator which is fed into the denoiser, effectively mitigating the overfitting challenge commonly encountered in unsupervised models.
- Through comprehensive evaluations using both real and simulated medical imaging datasets, our model outperforms supervised and unsupervised denoising methods.
- Our technique also represents a significant advancement in denoising videos containing non-Gaussian noise types, an area where many existing CNN-based models struggle.

2. Related Work

Supervised Video Denoising Networks Supervised deep learning-based approaches [7, 13, 19, 28, 32] have demonstrated superior performance in video denoising. DVDNet [27] leverages UNet [22] blocks to process five successive neighboring frames, aiming to denoise the central frame with optical flow to compensate for motion explicitly. FastDVDnet [28] improved on [27] and performs implicit motion compensation through the architecture design. Vision Transformer-based methodologies such as those presented in VRT [14], RVRT [15], ASwin [17] and Song et al. [26] leverage transformer architectures to exploit temporal information in the video, thereby enhancing their noise reduction efficiency. A notable limitation in these supervised techniques is the need for training datasets comprising noisy and clean pairs, which may not always be feasible in real-world settings. Moreover, these models are generally tailored for specific noise types. This potentially constrains their versatility when confronted with different noise conditions, resulting in diminished performance on noise patterns not encountered during training.

Unsupervised Video Denoising Networks Unsupervised video denoising methods mitigate the challenge of requiring noisy/clean video pairs. Noise2Noise [13] demonstrates comparable image denoising performance with supervised one by learning the underlying clean image from pairs of noisy images. Frame2Frame [7] extended this concept to video denoising by treating sequential frames as pairs of noisy images and employing optical flow techniques to estimate motion. However, the performance of this method can be limited by inaccuracies in motion estimation. Dewil et al. [6] extends single-image approach in [7] by introducing a self-supervised fine-tuning framework which adopts FastDVDnet [28] as the backbone network. The Noise2Void method [10] introduced a 'blind-spot' technique which estimates each noisy pixel by considering neighboring pixels without including the noisy pixel itself. Laine et al. [11] enforced the blind-spot concept through architectural design. The state-of-the-art (SOTA) unsupervised video denoising method, i.e. UDVD [24], integrated the blind-spot network (BSN) in [11] into the FastDVDnet [28] framework. Wang et al. [30] increases the receptive field of the BSN and also includes a transformer network to capture temporal information. These methods have the limitation of overfitting noise when neighboring frames have very similar structures and require complex training. Zheng et al. [34] developed an unsupervised loss function during training to provide an unbiased estimator. However, this approach requires prior knowledge of noise distribution, which might not be available when denoising real videos.

The most related to our approach is DeepInterpolation

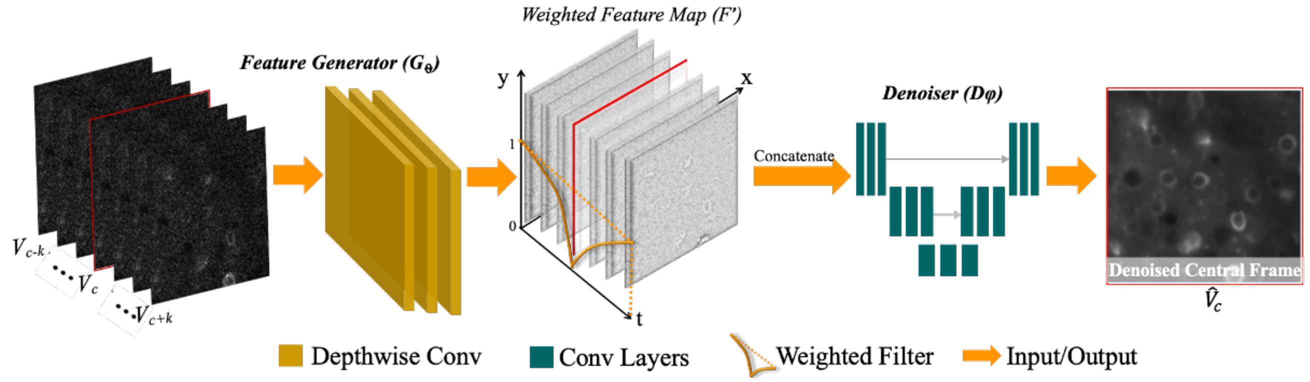

Figure 2. Our DeepTemporal Interpolation Pipeline. The feature generator $\mathbf{F}_\phi$ extracts distinctive features using three depthwise convolutional layers, and a temporal filter enhances denoising accuracy by adjusting feature map weights based on spatial-temporal proximity to the central frame, overcoming challenges in handling high frame rate videos and slow-moving objects.

[12], which uses U-Net as a backbone network. It denoises videos by omitting the central frame and leverages neighboring frames' similarity for interpolation. However, this method struggles when handling scenarios with high noise intensity. In contrast, our innovative model seamlessly integrates a temporal filter into the feature map generated by stacked CNN layers. By allocating higher weights to distant frame feature maps and lower weights to proximate ones, our model addresses the rapid convergence to noisy content seen in high frame rate videos due to redundancy in nearby frames, especially when the noise level is high. Consequently, our method showcases more resilient denoising performance than DeepInterpolation and other unsupervised approaches in UDVD and RDRF.

3. Unsupervised Video Denoising

Our goal is to generate denoised video $\{\hat{\mathbf{V}}_t | t = 1, 2, ...T\}$ given a sequence of noisy video input, $\{\mathbf{V}_t | t = 1, 2, ...T\}$, where T is the total number of frames in the input video. In each iteration, a subset of contiguous frames $\{\mathbf{V}_t | t = 1, ...c, ...N\}$ is taken as input, with N indicating the batch frame count and c denoting the central frame's index. These frames are then passed through the feature generator $\mathbf{G}_\phi$, producing the corresponding feature maps $\{\mathbf{F}_t | t = 1, ..., N\}$. Then, the temporal filter $\{\gamma_t\}_{t=1}^N$ weights these feature maps $\{\mathbf{F}_t\}_{t=1}^N$, assigning diminished values to features nearer the central frame compared to those more distant. Next, the resulting weighted feature maps are concatenated and fed into the Denoiser $\mathbf{D}_\theta$, producing the denoised central frame $\hat{\mathbf{V}}_c$. A comprehensive visualization of our pipeline is provided in Fig. 2.

3.1. Feature Generator G_ϕ

Given a batched frames $\{\mathbf{V}_t\}_{t=1}^N$, the feature generator $\mathbf{G}_\phi$ produces feature maps $\{\mathbf{F}_t\}_{t=1}^N$ aligned with the input frames. We utilize depth-wise convolution layers to expedite the training phase, eliminating the need to process each frame individually through $\mathbf{G}_\phi$. Unlike standard 2D convolutions that apply a multi-channel filter to the entire input depth, allowing for channel mixing, depthwise convolution maintains the separation of input channels and generates the output feature maps independently. Specifically, our feature generator integrates three depth-wise convolution layers. Each depthwise group generates feature map $\mathbf{F}_t \in \mathbb{R}^{C \times H \times W}$ for $\mathbf{V}_t$, where C is the number of output channels, while H and W represent the input image's height and width, respectively.

3.2. DeepTemporal Interpolation

Upon generating the feature maps $\mathbf{F}_t$ at the initial stage, we introduce the DeepTemporal Interpolation method that applies a set of weighted parameters to the feature maps. This filter plays a crucial role in enhancing denoising accuracy. Unlike DeepInterpolation [12], we do not need to remove the central frame from the input. Instead, our temporal filter is distinct in its ability to adjust the weight of each feature map based on its temporal proximity by setting the values of the central frame's feature map to zero, while diminishing weights are assigned to the feature maps of close neighboring frames. This unique weighting allows our approach to manage the intricate temporal similarities within high frame rate and slow-motion videos more effectively, addressing the shortcomings of existing models. This strategy prevents the model from mapping the inherent noise patterns in the target and surrounding frames. Notably, the temporal filter offers a new paradigm for video denoising, overcoming the limitations of current methods in handling high frame rate videos and slow-moving objects, which is common in microscopy videos. The temporal filter weights the feature map using Eq 1.

$$\gamma(.) = \left[\left\{ \frac{k-i}{k} \right\}_{i=0}^{k-1}, \left\{ \frac{i}{k} \right\}_{i=1}^{k} \right], \tag{1}$$

where $k = \lfloor M/2 \rfloor$, and $\lfloor . \rfloor$ denotes the floor function.

We apply the weights $\{\gamma_t | t = 1, ..., N\}$ to the feature maps by performing an elementwise multiplication. This ensures that the weight of feature map $\mathbf{F}_c$ at index c is zero, and the weights of the first and last feature map in the batch ($\mathbf{F}_1$ and $\mathbf{F}_N$) are one. This operation is depicted in Eq 2.

$$\mathbf{F}'_t = \gamma_t \odot \mathbf{F}_t \quad \forall t \in 1, ..., N. \tag{2}$$

Through this innovative application of the temporal filter, our method is able to provide robust denoising for high frame rate and slow-motion videos, effectively overcoming the challenges that limit current state-of-the-art approaches.

3.3. Denoiser D_θ

Given the weighted feature maps $\{\mathbf{F}'_t | t = 1, ..., c, ..., N\}$, the Denoiser D_θ generates a denoised central frame $\hat{\mathbf{V}}_c$ as in Eq 3. The Denoiser D_θ model is a U-Net [22] architecture with three encoders and two decoder layers. The concatenation of $\{\mathbf{F}'_t\}_{t=1}^{N}$ serves as the input to D_θ, enabling it to explore temporal correlations within adjacent frames.

$$\hat{\mathbf{V}}_c = \mathcal{D}_\theta([\mathbf{F}'_1, ...\mathbf{F}'_N]). \tag{3}$$

All convolutional layers in the network utilize ReLU activation functions except the output layer. It is important to note that the integration of skip connections between input and output layers in U-Net inherently amplifies the network's sensitivity to high-frequency details present in the input tensor, encompassing the inherent noise. Consequently, assigning a value of $\gamma_c = 0$ is imperative to guarantee the exclusion of any original central frame information from being transferred to D_θ. We found that even marginal values for γ_c can prompt the U-Net to rapidly converge towards the noise during training.

4. Training Details.

In our implementation, we take a stack of $N = 7$ contiguous frames. We adopt training protocols consistent with other unsupervised image and video denoising techniques, such as Laine et al. [11] and UDVD [24]. Specifically, the input tensor is partitioned into overlapping patches of size 128×128 in spatial dimensions, from which we predict the denoised central patches. This procedure, which is a form of data augmentation, significantly augments the number of training samples, thereby enhancing the denoising performance. It is essential to note that while we utilize these cropped patches during training, the full-resolution tensor was used for denoising during the inference phase.

We employ the l_2 loss function in Eq 4 to minimize the difference between the model's output $\hat{\mathbf{V}}_c$ and the reference central noisy frame $\mathbf{V}_C$

$$\mathcal{L} = ||\hat{\mathbf{V}}_c - \mathbf{V}_c||_2^2. \tag{4}$$

5. Experiments

We carried out a series of evaluations to assess the efficiency of our method. Our results show that our approach surpasses many SOTA video denoising techniques, including both supervised and unsupervised methods. Additionally, our ablation studies offer valuable understanding regarding the importance of key components within our methodology.

5.1. Datasets

We evaluated our method across diverse datasets, comprising both real noisy sequences in the case of calcium imaging and those infused with synthetic noise. For the natural video, sequences were contaminated with a range of noise types and intensities to simulate various challenging conditions. For medical imaging, we used real-world noisy videos primarily from fluorescence microscopy and calcium imaging. We used the recordings from neural recordings of freely moving mice for one-photon imaging. To evaluate with two-photon imaging, we simulated realistic calcium imaging, which also has the ground truth clean imaging.

Real-world Dataset Our real-world dataset comprises one-photon calcium imaging recordings procured locally from freely behaving transgenic mice (Drd1-Cre and Drd2-Cre) during cocaine/sucrose self-administration experiments. This calcium imaging technique, critical in neuroscience, allows for real-time visualization of neuronal activity. However, these recordings are often degraded by noise, making denoising an essential preprocessing step. For these recordings, single-channel epifluorescent miniaturized miniscopes were employed, capturing the fine details of the neuronal activity. Given that this dataset possesses complex noise levels and lacks reference ground-truth clean videos, it provides a valuable practical demonstration of our model's application to real-world scenarios.

For fluorescence microscopy, we utilized the microscopy recordings of live cells obtained from [29]. We used the GOWT1 cell recording (Fluo-C2DL-MSC) and mesenchymal stem cell recording video (Fluo-N2DH-GOWT1). Similar to the one-photon calcium imaging, this data has no ground truth clean video.

We simulated realistic two-photon calcium imaging data using neural anatomy and optical microscopy (NAOMi) [25] to perform the qualitative evaluation on realistic imaging with ground truth. We generated two datasets with 1,000 and 1,500 frames, respectively, with the field of view (FOV) of 150x150 μm^2 and 500x500 μm^2 respectively.

High Frame Rate RGB Video Dataset. To validate our model's generalization capabilities across a broad spectrum of video content, we extend our evaluation beyond microscopy imaging to include color video data captured with RGB cameras. This approach ensures a comprehensive evaluation under diverse noise types and intensities, showcasing the model's adaptability to real-world applications. To achieve this, we corrupted the video sequence with synthetic noise, employing three noise types: Gaussian, Poisson, and Impulse (Salt-and-Pepper). We used the LIVE YouTube High Frame Rate (LIVE-YT-HFR) [18] datasets to evaluate on high frame rate RGB video. This dataset features a wide array of video sequences, each captured at various frame rates per second (fps) and subject to different compression levels. Specifically, our evaluation concentrated on sequences captured at 120fps. This demonstrates the robustness of our model across various high frame rate video scenarios, further affirming its efficacy and wide-ranging applicability in video denoising tasks.

To simulate varying intensities of Gaussian noise, we used standard deviation values (σ) of 30, 50, and 90. Intensities of Poisson noise (photon shot noise) were varied by setting the maximum event count (λ) to 30, 50, and 90. For Impulse noise, which appears as random white or black pixels due to sudden signal changes, we used pixel ratios (α) of 0.2, 0.3, and 0.4 to indicate different noise levels.

5.2. Experimental Setup

We benchmarked our DeepTemporal Interpolation against SOTA unsupervised video denoising techniques DeepInterpolation [12], UDVD [24], RDRF [30], and also against the supervised paradigm, RVRT [15], and ASwin [17]. All models were implemented using the PyTorch framework [20] and NVIDIA A100 GPU was used for execution. We adopted the ADAM optimizer with default parameters. Over a span of 25 epochs, each iteration employed an input stack as the mini-batch, targeting the central frame's denoising. An early stopping strategy was used to mitigate potential overfitting, halting the training process in the absence of notable loss decrement. The inception learning rate was fixed at $1e-3$, and was systematically halved every 10 epochs.

For the supervised video denoising methods, i.e. RVRT [14] and ASwin [17], we use their publicly available pre-trained model, which was already trained under various noise conditions to generate the denoising results. Specifically, for ASwin, we used the model trained for blind denoising of real-world noisy video. For UDVD [24], we train individual video sequences using the **UDVD-S** model. The training process for **UDVD-S** followed a blind denoising approach, minimizing the mean squared error (MSE) without any assumptions regarding the noise parameters. Similar to UDVD, we also trained each video sequence separately for DeepInterpolation [12]. For RDRF [30], we generated the denoising result for the color video data using the pre-trained model provided by the authors. We specifically pass the noisy input to the model without giving any prior information about the noise parameters. However, we trained the model from scratch on each video recording for the microscopy dataset. All implemented models were trained and evaluated on the same datasets, ensuring consistency in our comparative analysis.

It's worth noting that while specific unsupervised image/video denoising methodologies, including Laine et al. [11], UDVD [24] RDRF [30], Zheng et al. [34] boost their efficacy by introducing random noise to clean reference frames in the training phase, we deliberately refrained from employing such strategies. Utilizing these techniques can result in overfitting to a particular noise characteristic, often Gaussian, and a specific noise intensity. This, in turn, degrades the model's adaptability to real-world, varied noise situations. Consequently, our training strategy only utilizes the inherent noisy video frames as input, avoiding the addition of any noise during training.

5.3. Quantitative Evaluation

To evaluate our proposed method, we used two metrics in the domain of image and video denoising: the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) As shown in Table 1, our method outperforms the unsupervised methods, RDRF and UDVD, across various noise categories on the RGB video dataset. While the supervised RVRT and ASwin models demonstrate good denoising performance on Gaussian noise, their performances significantly drop relatively on other noise types, especially in scenarios with high noise levels. Additionally, the findings in Table 2 demonstrate that our method surpasses supervised and unsupervised techniques on the simulated two-photon calcium imaging dataset. This corroborates our model's efficacy in denoising across various datasets and noise scenarios, showcasing its broad applicability and effectiveness.

5.4. Qualitative Comparison

The visual comparison of our model against other models on microscopy datasets as shown in Figures 1, 3, and 5, highlights the denoising effectiveness on two-photon calcium imaging. Our model demonstrates an exceptional ability to remove noise while retaining crucial signal details in calcium imaging. A detailed inspection in Figure 3 reveals that our technique successfully recovers intrinsic high-frequency information, often lost or overly smoothed in other methods. Similarly, the comparison results of one-photon imaging in Figure 6 show our model's proficiency in denoising and preserving finer details. Additionally, Figure 4 illustrates the superiority of our approach in denois-

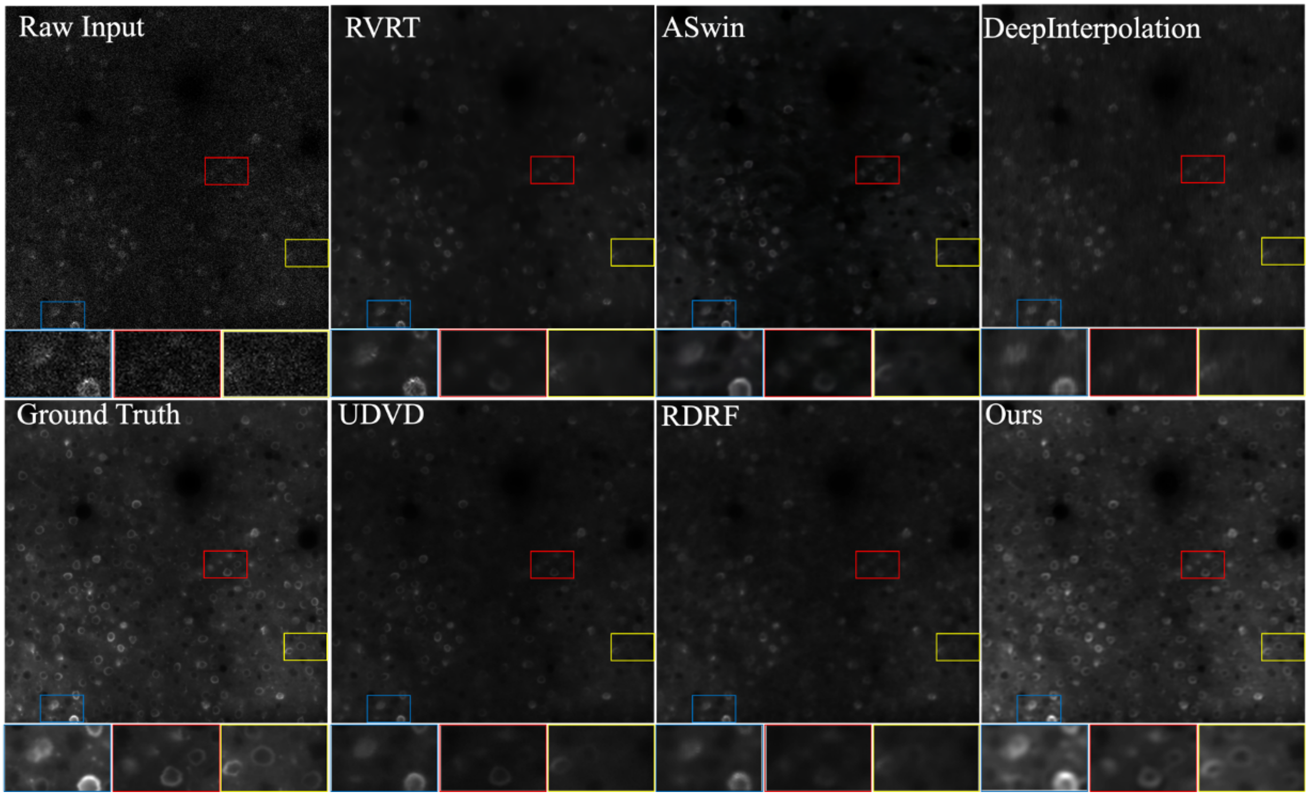

Figure 3. This figure presents a comparative analysis of denoising performance between our approach and that of other established methods, including RVRT, Aswin, DeepInterpolation, UDVD, and RDRF, on Two-Photon (2P) calcium imaging data simulated using NAOMi to produce realistic imaging with a ground truth reference. The outcomes illustrate that our model successfully denoises the video data while preserving relevant signals and avoiding obscuring critical information. Conversely, the denoising outputs from alternative methods reveal their limitations in accurately recovering underlying signals.

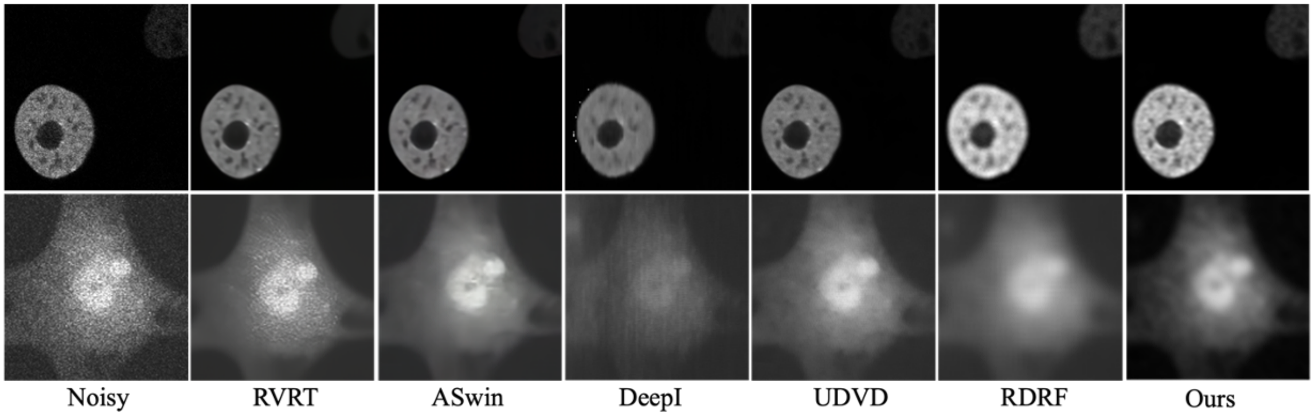

Figure 4. Visual Comparison on Flourescence Microscopy

ing fluorescence microscopy datasets over other supervised and unsupervised techniques, further affirming our model's robust denoising capabilities across various scenarios. Figure 7 shows the extracted traces of two region-of-interests (ROIs) manually drawn from the simulated two-photon calcium imaging.

5.5. Ablation Study

We conduct an ablation study to explore the influence of various components on our model's effectiveness.

Influence of the DeepTemporal Interpolation. This study probes the critical contribution of the Temporal Fil-

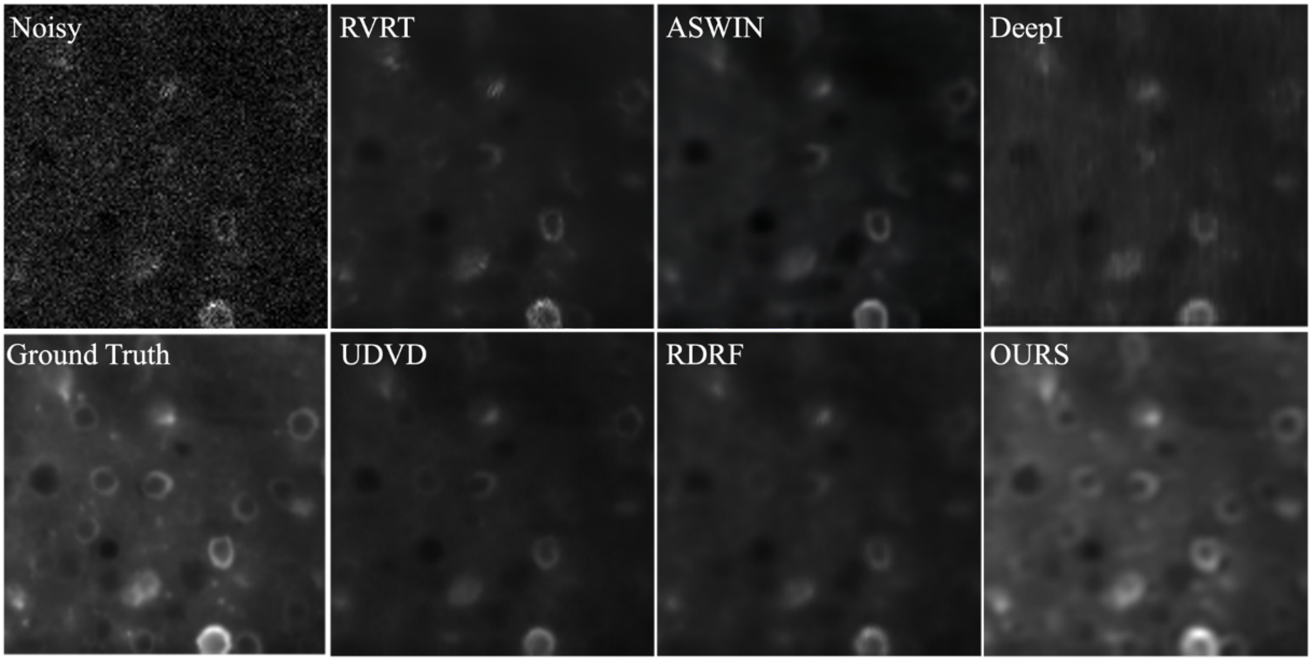

Figure 5. Additional visual comparison of two-photon calcium imaging

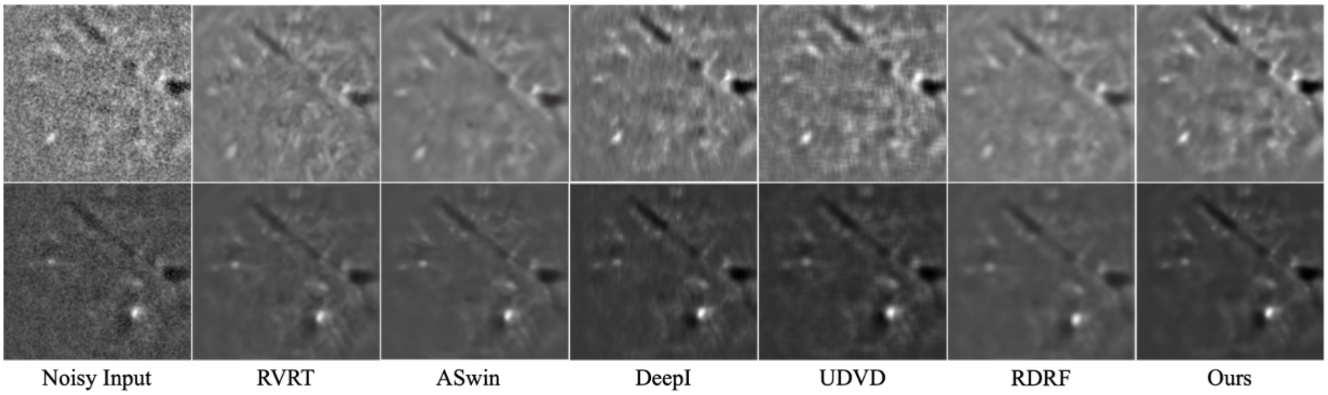

Figure 6. Denoising results on one-photon calcium imaging recordings of freely behaving transgenic mice

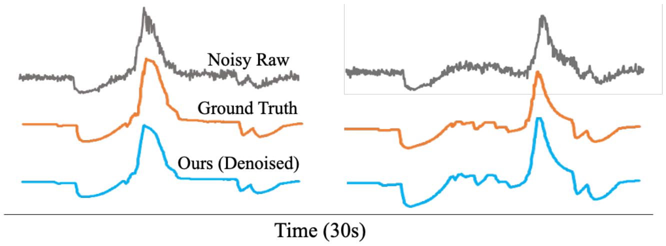

Figure 7. Comparative Traces of Two Example ROIs from the simulated calcium imaging.

ter (TF) in our framework. In the experiment, we feed the output from the feature generator straight to the Denoiser and evaluate the denoising performance using the variant without TF, denoting as $G_\phi + D_\theta$. The results presented in Table 3 underline that the TF plays a substantial role in enhancing denoising efficiency.

Impact of Frame Rates Variation. In our study, we systematically examined the performance of our model across various frame rates, focusing primarily on high-speed videos. We subjected the LIVE-YT-HFR dataset to varying frame rates, including 120, 98, 82, 60, 30, and 24 fps, while introducing different noise types and intensities. Table 4 provides a detailed breakdown of our findings.

Notably, as evidenced by the results, our model showcases consistently superior denoising performance at higher

		Gaussian			Poisson			Impulse		
		$\sigma = 30$	$\sigma = 50$	$\sigma = 90$	$\lambda = 30$	$\lambda = 50$	$\lambda = 90$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$
LIVE-YT-HFR 120fps	RVRT	30.75/0.80	**29.61**/0.83	18.12/0.22	27.68/0.78	27.27/**0.92**	17.64/0.22	23.84/0.63	17.32/0.19	13.98/0.09
	ASwin	15.50/0.40	15.55/0.41	15.90/0.42	15.12/0.39	14.86/0.39	14.29/0.39	15.81/0.42	16.10/0.42	16.18/0.42
	DeepI	30.25/0.87	28.18/0.82	**22.94**/0.71	28.09/0.89	26.36/0.88	23.25/0.86	23.86/0.75	21.15/0.67	19.06/0.61
	UDVD	28.47/0.84	25.66/0.76	22.15/0.65	29.37/0.87	26.24/0.84	22.18/0.74	22.48/0.68	20.24/0.61	18.45/0.57
	RDRF	25.80/0.58	20.08/0.35	15.35/0.18	25.43/0.65	22.39/0.53	18.99/0.41	16.79/0.23	14.61/0.16	13.14/0.12
	Ours	**31.67/0.90**	28.70/**0.85**	22.82/**0.75**	**30.82/0.92**	**28.72**/0.89	**24.76/0.88**	**24.44/0.82**	**21.29/0.74**	**19.23/0.67**

Table 1. **Performance in Denoising Synthetic Noise.** This table presents a comparison of average PSNR/SSIM values of denoised performance on LIVE-YT-HFR datasets. Text highlighted in **bold** signifies the highest value, while underlined text denotes the second highest. Our method demonstrates superior performance in most cases and remains highly competitive with the supervised methods.

	RVRT	ASwin	DeepI	UDVD	RDRF	Ours
PSNR	30.05	25.94	28.64	25.75	28.31	**35.90**
SSIM	0.92	0.90	0.91	0.74	0.92	**0.95**

Table 2. **Denoising performance on simulated 2P Calcium Imaging.** The result shows the average PSNR and SSIM values on the 2P imaging simulated with NAOMi [25]

	Gaussian 30		Poisson 30		Impulse 0.2	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
$G_\phi + D_\theta$	21.23	0.42	21.26	0.49	13.88	0.16
$G_\phi + TF + D_\theta$	**31.67**	**0.90**	**30.82**	**0.92**	**24.32**	**0.82**

Table 3. **Effect of Excluding Temporal Filter (TF) on Denoising Effectiveness.** This table presents a contrast in PSNR/SSIM metrics for our LIVE-YT-HFR-trained model, comparing denoising outcomes with and without including the temporal filter layer.

	Gaussian 30		Poisson 30		Impulse 0.2	
fps	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
120	**31.67**	**0.90**	**30.82**	**0.92**	**24.32**	**0.82**
98	28.83	0.83	28.69	0.88	22.44	0.73
82	29.30	0.85	29.42	0.89	22.67	0.74
60	29.10	0.85	28.90	0.89	22.91	0.75
30	24.29	0.71	24.68	0.76	21.25	0.65
24	23.58	0.68	23.50	0.72	20.43	0.59

Table 4. **Impact of Frame Rate Variation.** This table showcases the PSNR/SSIM metrics across various frame rates from the LIVE-YT-HFR dataset.

frame rates, particularly at 120 fps across all noise types. However, the performance starts to wane with lower frame rate spectrums, especially at 30 and 24 fps. This highlights a potential limitation of our model regarding denoising videos with lower frame rates. This degradation in performance at lower frame rates suggests a possible direction for future improvements and optimizations to our algorithm.

Different of Number of Input Frames (N) We conducted an ablation study on the number of frames (N) utilized as the input batch in our model. The findings in Table 5 indicate that $N = 7$ yields the optimal performance across different noise types.

	N=3	N=5	N=7	N=9	N=11
Poisson 50	27.35	25.81	**28.72**	26.97	26.23
Impulse 0.2	23.93	24.42	**24.44**	24.07	24.22

Table 5. **Impact of Number of Input Frames.** This table displays the average PSNR values for various numbers of input frames from the LIVE-YT-HFR dataset.

6. Discussion

We presented a novel unsupervised approach to denoise microscopy videos, leveraging the DeepTemporal Interpolation based CNN network architecture. We demonstrate that our unsupervised microsopy video denoising method achieves outperforming performance on various real-world micoscopy dataset and diverse noise types by emphasizing the importance of spatiotemporal features in microscopy videos. In our future work, we plan to enhance our model's efficiency with low frame rate videos. Additionally, we intend to refine our temporal filter to adjust the parameters dynamically based on observed motion speed. We also aim to adaptive accommodate the number of input frames in the denoisng batch according to the speed of moving objects in the videos.

References

[1] Zainab T Al-Sharify, Talib A Al-Sharify, Noor T Al-Sharify, et al. A critical review on medical imaging techniques (ct and pet scans) in the medical field. In IOP Conference Series: Materials Science and Engineering, volume 870, page 012043. IOP Publishing, 2020. 1

[2] Joshua Batson and Loic Royer. Noise2self: Blind denoising

by self-supervision. In International Conference on Machine Learning, pages 524–533. PMLR, 2019. 1, 2

[3] Tamara Berdyyeva, Stephani Otte, Leah Aluisio, Yaniv Ziv, Laurie D Burns, Christine Dugovic, Sujin Yun, Kunal K Ghosh, Mark J Schnitzer, Timothy Lovenberg, et al. Zolpidem reduces hippocampal neuronal activity in freely behaving mice: a large scale calcium imaging study with miniaturized fluorescence microscope. PLoS One, 9(11):e112068, 2014. 1

[4] Christian R Burgess, Rohan N Ramesh, Arthur U Sugden, Kirsten M Levandowski, Margaret A Minnig, Henning Fenselau, Bradford B Lowell, and Mark L Andermann. Hunger-dependent enhancement of food cue responses in mouse postrhinal cortex and lateral amygdala. Neuron, 91(5):1154–1169, 2016. 1

[5] Michele Claus and Jan Van Gemert. Videnn: Deep blind video denoising. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 0–0, 2019. 1

[6] 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

[7] 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. 2

[8] Shah Hussain, Iqra Mubeen, Niamat Ullah, Syed Shahab Ud Din Shah, Bakhtawar Abduljalil Khan, Muhammad Zahoor, Riaz Ullah, Farhat Ali Khan, Mujeeb A Sultan, et al. Modern diagnostic imaging technique applications and risk factors in the medical field: a review. BioMed research international, 2022, 2022. 1

[9] Hany Kasban, MAM El-Bendary, and DH Salama. A comparative study of medical imaging techniques. International Journal of Information Science and Intelligent System, 4(2):37–58, 2015. 1

[10] 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. 1, 2

[11] 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, 4, 5

[12] 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, 3, 5

[13] 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. 1, 2

[14] 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, 5

[15] 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, 5

[16] 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.

[17] Lydia Lindner, Alexander Effland, Filip Ilic, Thomas Pock, and Erich Kobler. Lightweight video denoising using aggregated shifted window attention. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 351–360, 2023. 1, 2, 5

[18] Pavan C. Madhusudana, Xiangxu Yu, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Subjective and objective quality assessment of high frame rate videos. IEEE Access, 9:108069–108082, 2021. 5

[19] 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

[20] 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

[21] Lucas Pinto and Yang Dan. Cell-type-specific activity in prefrontal cortex during goal-directed behavior. Neuron, 87(2):437–450, 2015. 1

[22] 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. 2, 4

[23] Juergen Sawinski, Damian J Wallace, David S Greenberg, Silvie Grossmann, Winfried Denk, and Jason ND Kerr. Visually evoked activity in cortical cells imaged in freely moving animals. Proceedings of the National Academy of Sciences, 106(46):19557–19562, 2009. 1

[24] 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, 4, 5

[25] 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. 4, 8

[26] Mingyang Song, Yang Zhang, and Tunç O Aydın. Tempformer: Temporally consistent transformer for video denoising. In European Conference on Computer Vision, pages

481–496. Springer, 2022. 2

[27] 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, 2

[28] 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

[29] 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. 4

[30] 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. 1, 2, 5

[31] Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, and Jingyu Yang. Supervised raw video denoising with a benchmark dataset on dynamic scenes. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2301–2310, 2020. 1

[32] 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

[33] Zhaoyang Zhang and et al. Real-time controllable denoising for image and video. In CVPR, pages 14028–14038, 2023. 1

[34] Huan Zheng, Tongyao Pang, and Hui Ji. Unsupervised deep video denoising with untrained network. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 3651–3659, 2023. 1, 2, 5

[35] Yaniv Ziv, Laurie D Burns, Eric D Cocker, Elizabeth O Hamel, Kunal K Ghosh, Lacey J Kitch, Abbas El Gamal, and Mark J Schnitzer. Long-term dynamics of ca1 hippocampal place codes. Nature neuroscience, 16(3):264–266, 2013. 1