Efficient Online Training for Zero-Shot Time-Lapse Microscopy Denoising and Super-Resolution

Ruian He, Ri Cheng, Xinkai Lyu, Weimin Tan*, Bo Yan*

School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
rahe16@fudan.edu.cn, rcheng22@m.fudan.edu.cn, xklv23@m.fudan.edu.cn, wmtan@fudan.edu.cn, byan@fudan.edu.cn

Abstract

In time-lapse microscopy, inherent noise significantly limits imaging sensitivity and increases measurement uncertainty. Due to the scarcity of clean data, zero-shot approaches have emerged as highly data-efficient solutions for microscopy denoising. However, existing methods typically process video frames independently, resulting in long training times and issues such as temporal noise and over-smoothing. In this paper, we introduce MDSR-Zero, a zero-shot online learning method designed for plug-and-play noise suppression and super-resolution of microscopy videos. Our approach leverages an efficient online training strategy that reuses denoising models from previous frames. By treating the video as a continuous stream, our model significantly reduces training time and ensures temporally consistent denoising. Additionally, we propose a novel loss function tailored for denoising in the context of super-resolution, which enhances the detail in the denoised results. Extensive experiments on both synthetic and real-world noise demonstrate that our method achieves state-of-the-art performance among zero-shot denoising approaches and is competitive with self-supervised methods. Notably, our method can reduce training time by up to 10x compared to the previous SOTA method.

Introduction

Living organisms function through precisely coordinated cellular and subcellular activities, both spatially and temporally. Observing and analyzing these activities under microscopy is essential for understanding biological processes and providing evidence for medical treatment. (Ma et al. 2024) However, the low signal-to-noise ratio (SNR) creates obstacles for microscopy analysis. This is because low-light conditions and short exposure times are widely applied in biological imaging for longer observations and lesser degradation of specimens (Li et al. 2023b). For example, in fluorescent microscopy, the low photon yield of fluorescent indicators and their low concentration in labeled cells lead to a scarcity of photons at the source (Hirano et al. 2022). To get better imaging with less noise, researchers use high-power lasers or longer exposure times, which sacrifice imaging speed or even sample health, to get satisfactory imaging (Meiniel, Olivo-Marin, and Angelini 2018). As indicated in the literature (Chenouard et al. 2014), lower SNR levels make detection and tracking methods break down. Therefore, microscopy analysis calls for denoising methods that can effectively restore clean signals.

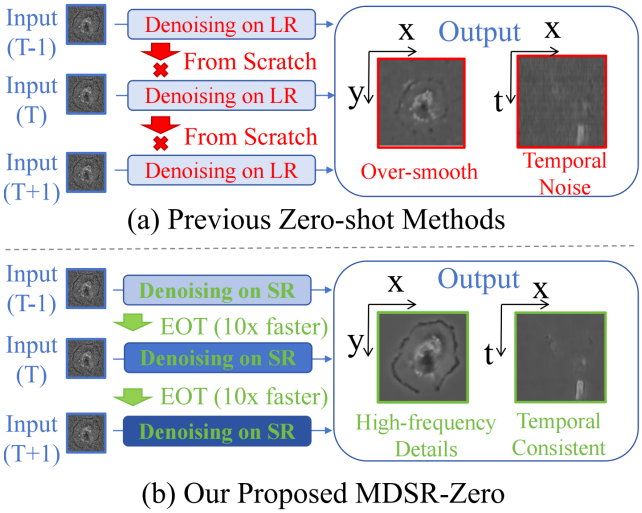


Figure 1: Comparison of the frameworks. Previous method (Lequyer et al. 2022) needs training from scratch and suffered from the over-smooth problem and temporal noise, while our method can recover high-frequency and time-consistent details with 10x fast speed.

Deep denoising methods (Lehtinen et al. 2018; Sheth et al. 2021) provide a plausible solution to enhance the microscopy imaging quality. Recent advances (Lecoq et al. 2021; Li et al. 2021; Zhang et al. 2023) for microscopy imaging enhancement are self-supervised, which focus on learning from the noisy data itself and can perform blind denoising. However, self-supervised methods still require extra training data for a specified microscopy. Zero-shot methods (Lequyer et al. 2022; Mansour and Heckel 2023a; Qiao et al. 2024) further reduce the needs of training data by learning denoising models directly from single images. These methods greatly improve imaging quality with limited data and can facilitate microscopy analysis with high data efficiency.

Though zero-shot methodologies demonstrate superior performance and efficiency in image denoising tasks, they

*Corresponding authors: Bo Yan, Weimin Tan.

are yet to be adapted for time-lapse microscopy videos. As shown in Fig. 1, the frame-by-frame denoising manner discards the trained models from previous frames and requires training from scratch for every frame, which limits the denoising performance and increases the denoising time. Furthermore, these methods often produce overly smooth and temporal-inconsistent results. The smoothing effect, while effective in reducing visible noise, has the unintended consequence of blurring essential details and textures in the video. Moreover, the methods perform denoising frame-by-frame without considering the consistency of the video. Thus, the temporal noise is still significant. Such smoothing effects and temporal noise are particularly problematic in applications where fine features and subtle variations are crucial, such as in biological analysis.

To address these challenges, our contributions can be summarized as follows: **(1)** We introduce MDSR-Zero, a novel zero-shot method for denoising and super-resolution of microscopy videos, which significantly enhances time efficiency while achieving state-of-the-art performance. Fig. 2 demonstrate our advantage in performance and efficiency. Compared to N2F (Lequyer et al. 2022), the SOTA method in microscopy denoising, our method only needs 1/10 time. **(2)** We propose Efficient Online Training (EOT), an innovative strategy that accelerates the denoising process by reusing trained models from previous frames and effectively addresses temporal noise. **(3)** Our loss function integrates denoising and super-resolution within a single model through a self-supervised framework, enabling the recovery of high-frequency details.

Related Work

Self-Supervised Image Denoising In scenarios where paired degraded-clean images are not available, researchers have turned to utilizing the inherent information within degraded images themselves. Noise2Noise (Lehtinen et al. 2018) first introduces the concept of training directly on pairs of noisy images. To create these pairs directly from noisy images, later methods such as subsampling from original images (Huang et al. 2021), generating pairs of differently noisy images (Moran et al. 2020), or using data augmentation techniques (Pang et al. 2021) are employed. On the other hand, Noise2Void (Krull, Buchholz, and Jug 2019) and Noise2Self (Batson and Royer 2019) initially introduced the concept of the Blind Spot Network (BSN), which leverages the different perception fields of the neural network. BSN is followed by improvements like Laine19 (Laine et al. 2019) and Blind2Unblind (Wang et al. 2022).

Self-Supervised Video Denoising Self-supervised video denoising methods also fall into two primary categories: those based on Noise2Noise (Lehtinen et al. 2018) principles and those utilizing BSN (Batson and Royer 2019). N2N-based methods, such as Frame2Frame (Ehret et al. 2019) and Multi-Frame2Frame (Dewil et al. 2021), operate under the assumption that adjacent video frames represent the same scene. (Zheng, Pang, and Ji 2023) apply R2R (Pang et al. 2021) to construct training samples for video denoising. DeepInterpolate (Lecoq et al. 2021), DeepCAD (Li et al. 2021) and DeepCAD-RT (Li et al. 2023b) further promote Frame2Frame (Ehret et al. 2019) for calcium and fluorescence microscopy videos. The second category extends image blind spot networks to videos. UDVD (Sheth et al. 2021) and later works (Wang et al. 2023) take a stack of frames into a blind spot network. DeepSeMi (Zhang et al. 2023) follows Laine19 (Laine et al. 2019) to build a 3D blind spot network for microscopy denoising.

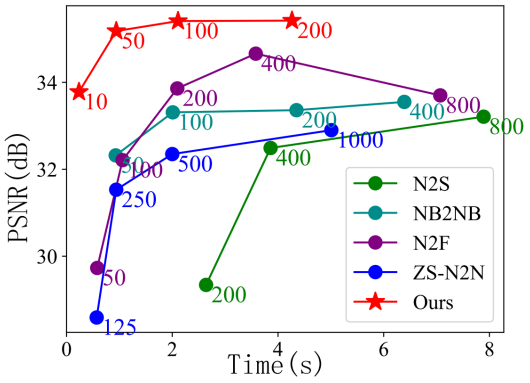


Figure 2: Efficiency comparison of denoising methods. The numbers under the data point indicate the training iterations. The time is calculated for denoising one frame, including training and inferencing. PSNR is averaged under three levels of Poisson noise on CTC. Our method demonstrates ultra-high time efficiency and superior denoising performance.

Zero-Shot Denoising Zero-shot image denoising methods, like S2S(Quan et al. 2020) and ZS-N2N(Mansour and Heckel 2023a), are proposed to improve data efficiency, which takes only one single image for training and testing. It is also helpful in microscopy imaging, where the data is scarce. Noise2Fast (Lequyer et al. 2022) and SRD-Trans (Li et al. 2023a) promote Neighbor2Neighbor (Huang et al. 2021) for zero-shot image denoising of microscopy. ZSDeconv (Qiao et al. 2024) explores zero-shot denoising and super-resolution methods for fluorescence microscopy. However, ZSDeconv needs a pre-defined point-spread function (PSF) and focuses on single images, while our work performs blind restoration and focuses on time-lapse video.

Methodology

Formulation of Microscopy Restoration

In microscopy images, the noise is a crucial degradation which hinder observation and analysis. It is mainly from the statistical fluctuations of photons sensed at a given exposure level (Meiniel, Olivo-Marin, and Angelini 2018). For fluorescence microscopy, the signal $F(r,t)$ at position r and time t is given by the convolution of the system's PSF $H(r)$ and the fluorescence source distribution (Dertinger et al. 2009). The formulation can be expressed as:

$$F(\mathbf{r},t) = H(\mathbf{r}) * S(\mathbf{r},t) + N(\mathbf{r},t) \quad (1)$$

where $*$ is a 2D convolution operator. $S(\mathbf{r},t)$ is the original brightness of the sample, and $N(\mathbf{r},t)$ is a time-dependent

and space-dependent noise.

To reduce the effect of noise, the restoration models methods predict the clean observation of the microscopy by performing the following transformation.

$$f_\theta^{denoise}(F(\mathbf{r},t)) = H(\mathbf{r}) * S(\mathbf{r},t) \tag{2}$$

where $f_\theta^{denoise}$ is the denoising model with parameter θ. The zero-shot denoising models (Lequyer et al. 2022; Mansour and Heckel 2023a) make the assumption that the subsamples F_1 and F_2 of the image F are isolated as independent emitters of photons and can be taken as independent emitters:

$$F_1(\mathbf{r},t) = H(\mathbf{r}) * S(\mathbf{r},t) + N_1(\mathbf{r},t) \tag{3}$$

$$F_2(\mathbf{r},t) = H(\mathbf{r}) * S(\mathbf{r},t) + N_2(\mathbf{r},t) \tag{4}$$

Since in expectation over such noisy instances, and assuming zero mean noise, training a network in a supervised manner to map a noisy image to another noisy image is equivalent to mapping it to a clean image (Lehtinen et al. 2018). And the denoising function is trained to minimize the empirical risk:

$$\theta_t = \underset{\theta}{\operatorname{argmin}}\, \mathrm{E}[||f_\theta(F_1(\mathbf{r},t)) - F_2(\mathbf{r},t)||_2^2] \tag{5}$$

where θ_t is the trained wight at frame t. However, the time-dependent noise $N(\mathbf{r},t)$ fluctuates in different frames and will cause the denoising model to learn a different weight for the same sample illuminance. As shown in Fig. 1, frame-by-frame denoising will cause severe temporal noise. Making the temporal average on trained model weights is a plausible approach to reduce the time-dependent noise.

To require the original fluorescence signals, the restoration models should further predict the inverse PSF function of the microscope to restore the spatial resolution.

$$f_\theta^{SR}(F(\mathbf{r},t)) = H_\theta^{-1}(\mathbf{r})\, f_\theta^{denoise}(F(\mathbf{r},t)) \quad = S(\mathbf{r},t) \tag{6}$$

where H_θ^{-1} is a deconvolution kernel, which is traditionally estimated by isolating emitters. The spatial resolution is determined by the number of photons that can be collected (Fernández-Suárez and Ting 2008), according to the equation:

$$\Delta\mathrm{loc} \approx \frac{FWHM}{\sqrt{N}} \tag{7}$$

where $\Delta\mathrm{loc}$ is the localization precision, $FWHM$ is the full width at half maximum of the PSF, and N is the number of collected photons. With a fixed PSF function, a larger number of collected photons (N) will make localization more precise. However, the denoising models subsample the image F into independent emitters F_1 and F_2 at a lower resolution. Therefore, the collected photons are reduced, which constrains the localization precision. As shown in Fig. 1, denoising on low-resolution will cause an over-smoothing problem. Discriminating emitters at higher resolutions may address the problem.

MDSR-Zero Framework

As depicted in Fig. 1, we propose a novel framework designed for zero-shot blind denoising and 2× super-resolution of time-lapse microscopy videos. Our framework

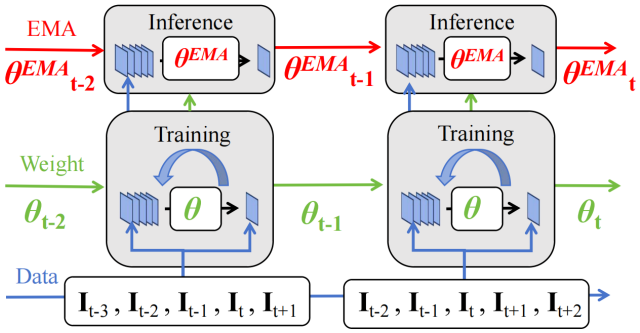


Figure 3: Efficient Online Training. The blue arrows are data flow, the green arrows are the running weights and the red arrows are EMA weights.

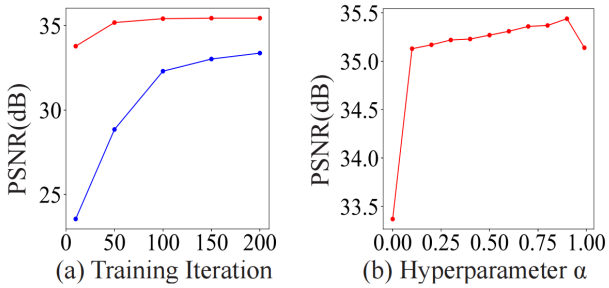


Figure 4: Effect of different hyperparameters. We show the PSNR averaged over 3 levels of Poisson noise on CTC. Red for EOT, Blue for Frame-by-frame.

boosts the training process of zero-shot denoising methods and addresses the problems of over-smoothing and temporal noise. This framework includes **(1) Efficient Online Training** designed for plug-and-play denoising of videos, which features low-temporal noise and faster processing speed, as well as **(2) Denoising on super-resolution** specifically tailored for zero-shot microscopy denoising and super-resolution, by discriminating neighbor emitters from super-resolution to solve the over-smoothing problem.

Efficient Online Training

We propose Efficient Online Training (EOT) to address the temporal noise and boost the training process, as shown in Fig. 3. Exponential Moving Average (EMA) has been validated for training supervised models in classification and segmentation tasks (Cai et al. 2021). However, it is under-explored for video denoising. EOT models the video frames as a data flow and the model as a running weight flow and an EMA weight flow. The weights are passed from frame to frame. The model in frame t first uses the parameters from the previous frame θ_{t-1} as the initial weight and optimize the denoising loss function over the current frame.

$$\theta_t = \arg\min_\theta \sum_{\mathbf{r}} L(f_\theta(F_1(\mathbf{r},t)), F_2(\mathbf{r},t)) \tag{8}$$

where L is the loss function, and F_1, F_2 are the independent noise samples. After training for certain iterations, we

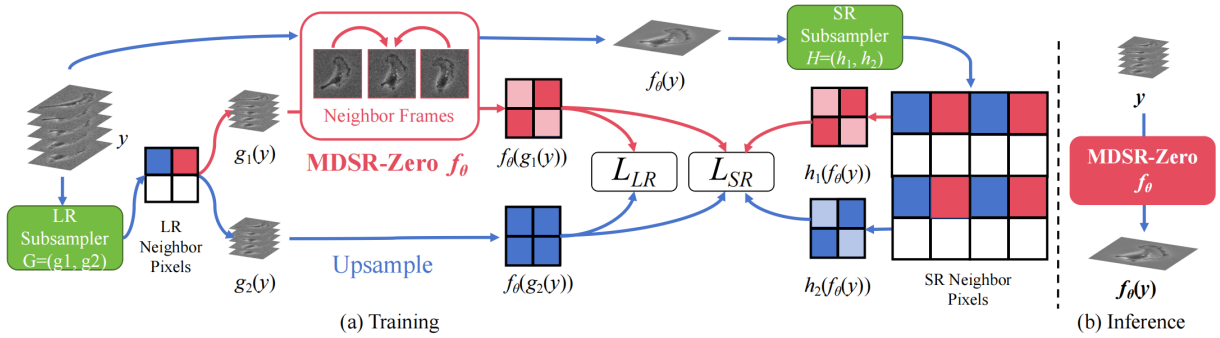


Figure 5: Proposed training target of zero-shot video denoising and super-resolution. The red and blue colors indicate the route for different subsamples. The gradient flow in the blue arrows is disabled.

get the model's parameter θ_t and calculate the Exponential Moving Average (EMA) weight θ_t^{EMA} as follows:

$$\theta_t^{EMA} = \alpha\theta_{t-1}^{EMA} + (1-\alpha)\theta_t \tag{9}$$

where θ_{t-1}^{EMA} is the EMA weight from the previous frame and α is the hyperparameter to control the decay of the historical weight. If it is the first update for EMA weight, $\theta_0^{EMA} = \theta_0$. Finally, θ_t^{EMA} is used for inference.

As shown in Fig. 4, we demonstrate two main merits for EOT's effect on zero-shot denoising on videos. **(1) Accelerate the training process.** Our model converges faster than the one without EOT (Frame-by-Frame). The EMA model trained with 10 iterations has the same performance as the non-EOT model with 10× more iterations. **(2) Improve the performance.** EOT can help address the temporal noise by averaging the model weight. With the same iteration, the model with EOT achieves better performance than the model without EOT.

Denoising on Super-Resolution

We propose a novel loss function for denoising on super-resolution, as shown in Fig. 5. Our method first involves randomly subsampling video frames using a 2x2 grid and choosing neighbor pixels to get independent noise samples. Similar to (Huang et al. 2021), there are 4 pairs of neighbor pixels and 8 combinations by swapping g_1 and g_2. The input frames will be subsampled to a half resolution before being fed into the network. The LR self-supervision (L_{LR}) can be expressed as:

$$L_{LR} = \|f_\theta\left(g_1(y)\right) - \text{Upsample}(g_2(y))\|_2^2 \tag{10}$$

where Upsample is an image interpolation method, and we use bilinear interpolation here. L_{LR} aims to supervise the reconstruction of the low frequency, and the subsampler along with the upsampler plays as a low-pass filter of the input.

In addition to the LR self-supervision, we propose a SR self-supervision that runs beyond the original resolution of the input. The input y is first fed into the model to get the SR results in $f_\theta(y)$. Then, the results are subsampled with an SR subsampler for neighbor pixels in higher resolution. The neighbor pixels construct another pair, $h_1(f_\theta(y))$ and $h_2(f_\theta(y))$. h_1 and h_2 derivative from g_1 and g_2 and the relation can be expressed as follows.

$$h_i(y) = \text{PixelShuffle}(g_i(\text{PixelUnshuffle}(y))) \tag{11}$$

where PixelShuffle and PixelUnshuffle are first proposed by (Shi et al. 2016) to upsample and downsample images by concatenating neighbor pixels. Then, we regularize the difference in L_{LR} to discriminate the objects in the higher resolution by optimizing over the loss function L_{SR}.

$$L_{SR} = \|f_\theta\left(g_1(y)\right) - \text{Upsample}(g_2(y)) - \Delta_{SR}\|_2^2 \tag{12}$$

where Δ_{SR} is the difference of the SR neighbors, $\Delta_{SR} = h_1(f_\theta(y)) - h_2(f_\theta(y))$.

Different from that in (Huang et al. 2021), the subsampling of L_{SR} is performed at a higher resolution, which improves the localization precision of the emitters. We suppose f_θ is optimal and the output $f_\theta(y)$ is a clean HR image. Δ_{SR} represents the gap between the observations of neighbor photon emitters in HR. If the emitters cannot discriminate in HR and the gap is zero, Δ_{SR} vanishes, and L_{SR} becomes a special case of L_{LR}. However, if the emitters can discriminate and the gap is non-zero, Δ_{SR} serves as a correction of L_{LR} to prevent over-smoothing.

The duplication of L_{LR} in L_{SR} is also necessary to ensure the plus and minus with Δ_{SR} is correct. As the subsamplers have a positional relationship, $f_\theta(g_1(y))$ corresponds to $h_1(f_\theta(y))$ and $Upsample(g_2(y))$ to $h_2(f_\theta(y))$. Minimizing $||\Delta_{SR}||$ will cause a smoothing effect.

The final loss function is a weighted sum of the two losses: $L = L_{LR} + \gamma L_{SR}$, and we set $\gamma = (2 * \text{Epoch})/\text{Total Epochs}$, which gradually increase during training.

Experiments

Experimental Details

Implementation Details We implement our framework with Pytorch. We use a UNet (Ronneberger, Fischer, and Brox 2015) with 17 convolutional layers and a PixelShuffle (Shi et al. 2016) layer at the end. The output is of 2× the size of the input. The number of input channels is 5 (neighbor 5 frames concatenated), and the number of output channels is 1. The input frames are randomly cropped into 128 × 128

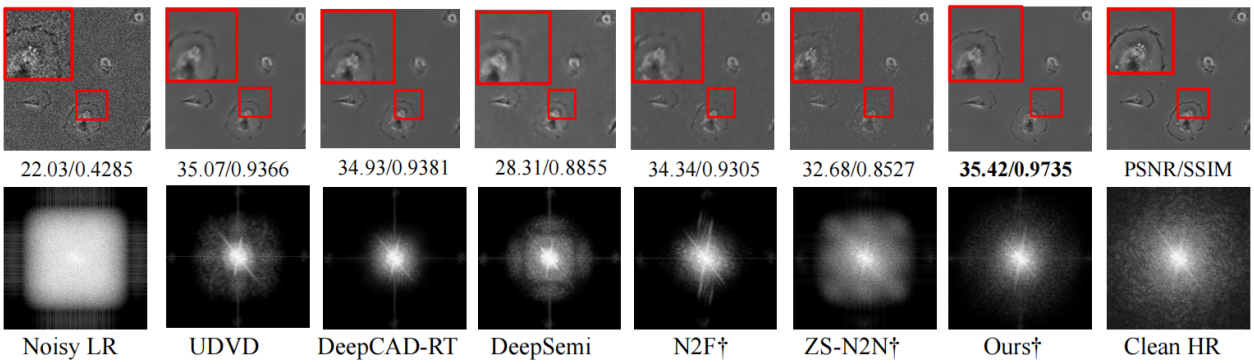


Figure 6: Qualitative comparison in synthetic datasets. We show the denoising results in the first row and the corresponding Fourier Spectrum in the second row. † indicates the zero-shot setting.

Type	UDVD	DeepCAD-RT	DeepSeMi	N2S †	NB2NB †	N2F †	ZS-N2N †	Ours†
P20	34.47/0.9242	**35.03/0.9355**	29.66/0.9124	29.07/0.8553	33.11/0.9098	33.21/0.9191	32.74/0.8880	34.80/0.9295
P30	**35.64**/0.9378	35.38/**0.9382**	29.86/0.9159	29.55/0.8715	33.40/0.9167	33.98/0.9240	33.38/0.8923	35.53/0.9360
P40	35.80/0.9387	35.48/**0.9392**	30.65/0.9195	29.39/0.8688	33.57/0.9204	34.40/0.9267	33.82/0.8967	**35.95**/0.9391
avg.	35.30/0.9336	35.30/**0.9377**	30.06/0.9160	29.34/0.8652	33.36/0.9156	33.86/0.9232	33.31/0.8923	**35.42**/0.9348
G20	35.43/0.9332	35.69/0.9407	31.55/0.9236	29.10/0.8503	33.69/0.9232	35.07/0.9320	34.40/0.9018	**36.58/0.9424**
G30	**35.32**/0.9350	35.09/**0.9367**	28.61/0.9138	29.49/0.8682	33.28/0.9135	33.64/0.9220	33.24/0.8913	35.20/0.9304
G40	32.47/0.8719	**34.75/0.9342**	28.69/0.9043	32.54/0.9155	29.26/0.8432	32.78/0.9017	32.35/0.8843	34.22/0.9194
avg.	34.40/0.9134	35.17/**0.9372**	29.62/0.9139	29.28/0.8539	33.25/0.9128	33.75/0.9231	33.33/0.8925	**35.33**/0.9307
MPG20	33.39/0.9187	36.03/0.9447	25.25/0.8918	29.46/0.8617	34.05/0.9287	35.98/0.9380	35.50/0.9147	**37.38/0.9502**
MPG30	33.20/0.9187	35.87/0.9428	29.10/0.8990	28.89/0.8433	33.89/0.9264	35.40/0.9341	34.84/0.9074	**36.77/0.9456**
MPG40	33.15/0.9158	35.28/0.9398	30.23/0.8990	29.33/0.8666	33.69/0.9230	34.64/0.9285	34.18/0.9026	**36.15/0.9407**
avg.	33.25/0.9177	35.73/0.9424	28.19/0.8966	29.23/0.8572	33.87/0.9260	35.34/0.9335	34.84/0.9083	**36.77/0.9455**

Table 1: PSNR in dB and SSIM scores with Poisson (P), Gaussian (G) and mixed Poisson-Gaussian (MPG) noise. **Bold** indicates the best results and underline indicates the second best results. We test with three different noise levels and calculate the average.

patches when training. The optimizer is Adam (Kingma and Ba 2014), and the learning rate is set to 3e-4. Following previous zero-shot denoising works (Lequyer et al. 2022), our model is randomly initialized and trained from scratch for each video sequence.

Baselines We divide the baseline models into 2 categories: **(1) Self-supervised denoising methods.** We compare SOTA methods in microscopy video denoising, UDVD (Sheth et al. 2021), DeepCAD-RT (Li et al. 2023b), and DeepSeMi (Zhang et al. 2023). The self-supervised models are trained on the whole video sequence with corresponding default settings before testing. **(2) Zero-shot denoising methods.** We compare with SOTA zero-shot denoising methods, such as Noise2Fast (N2F) (Lequyer et al. 2022) and Zero-shot Noise2Noise (ZS-N2N) (Mansour and Heckel 2023b). We also adapt Neighbor2Neighbor (NB2NB) (Huang et al. 2021) and Noise2Self (Batson and Royer 2019) to zero-shot settings for wider comparison, as performed in (Lequyer et al. 2022).

Synthetic Noise

In the context of microscopy, there are two main sources of noise (Meiniel, Olivo-Marin, and Angelini 2018). The readout noise from the equipment follows the Gaussian distribution $\mathcal{N}$. The dark noise and the photon noise inherent to the sample follow the Poisson distribution $\mathcal{P}$. The unified formulation of noise can be expressed as:

$$F(\mathbf{r}) = \gamma \mathcal{P}\left(\lambda S(\mathbf{r})\right) + n, n \sim \mathcal{N}(0, \sigma^2) \tag{13}$$

where λ is the intensity parameter of Poisson noise, and σ is for Gaussian noise. γ is the normalization factor, which is set to $1/\lambda$. In our experiments, we test three types of noise: Gaussian noise, Poisson noise and mixed Poisson-Gaussian (MPG). To assess the model's generalization capabilities, we vary the noise levels, utilizing $\sigma = 20, 30, 40$ for Gaussian noise and $\lambda = 20, 30, 40$ for Poisson noise. We then evaluate the performance using metrics such as PSNR and SSIM, and calculate the average values of the levels.

CTC (Maška et al. 2023) has time-lapse microscopy videos of various types of cells and nuclei shot through multiple microscopes. We select a clean dataset, PhC-C2DH-U373, and add synthetic noise for evaluating the quantitative performance. The dataset is composed of two videos. Every video has 114 frames of 696×520. The input frames are downsampled with 2× bilinear interpolation to the dataset, and the noise is added to the video frames. The bilinear interpolation simulates a simple PSF with a triangular 2D convo-

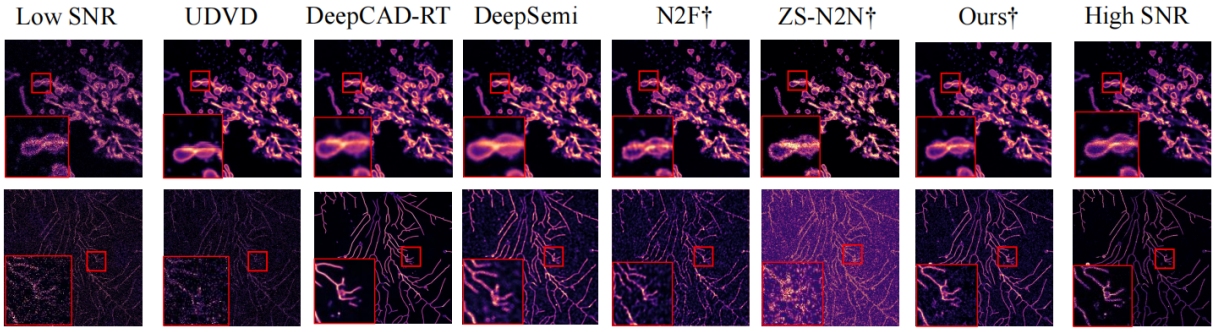


Figure 7: Qualitative comparison on the DeepSeMi dataset. The first row is from T4, and the second row is from EGFP.

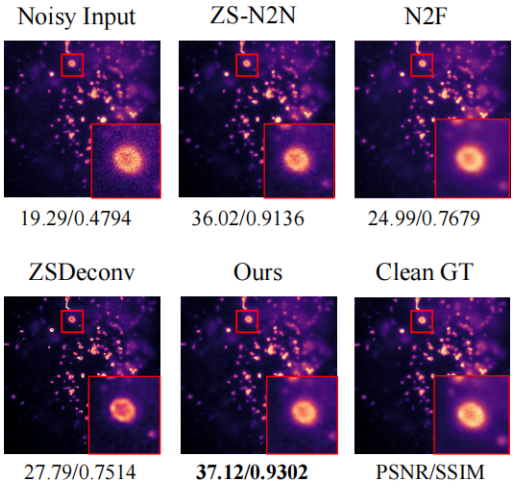


Figure 8: Qualitative comparison on the ZSDeconv dataset.

lution kernel. The output of the denoising methods (except for our method) is subsequently upsampled using $2\times$ bilinear interpolation to keep align with our upsampling function.

Tab. 1 presents the quantitative results. Our approach outperforms other zero-shot methods, even comparable with the self-supervised methods, which require extra training. When it comes to complicated noise like MPG, the zero-shot methods are more adapted than a single noise. Furthermore, our model performs surprisingly well on mixed Poisson-Gaussian noise, surpassing all other methods. We owe the performance advantage to our training strategy, eliminating both temporal and spatial noise.Fig. 6 showcases the qualitative comparison on Poisson noise at $\lambda = 40$. Our method produces a cleaner image with less loss of detail compared to other zero-shot methods. We also visualize with the tools of the Fourier Spectrum, as depicted in the second row of Fig. 6. The center of the Fourier Spectrum indicates the low frequency, while the corners indicate the high frequency. Previous denoising methods focus on subsampling at the current resolution, which could limit the restoration frequency. Our method can provide realistic details similar to clean HR.

Methods	EGFP	T4
UDVD	26.86/0.6919	21.72/0.2319
DeepCAD-RT	26.95/**0.8052**	**25.65/0.5529**
DeepSeMi	22.64/0.5034	21.51/0.2675
N2S †	9.19/0.0256	7.53/0.0189
NB2NB †	26.62/0.6680	10.08/0.0478
N2F †	22.16/0.5762	20.82/0.2462
ZS-N2N †	24.32/0.5778	9.77/0.0292
Ours †	**27.29**/0.6861	23.37/0.2778

Table 2: Quantitative comparison on DeepSeMi dataset. **Bold** indicates the best results and underline indicates the second best.

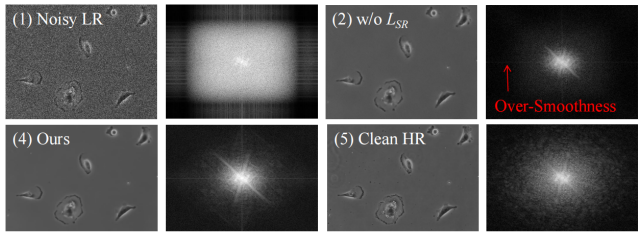
Real-World Noise

DeepSemi dataset. We conduct experiments on two video datasets (T4 and EGFP) from the DeepSemi paper (Zhang et al. 2023), which contains live confocal imaging under low and high SNR conditions. 100 consecutive frames are selected from each dataset. Fig. 7 shows the qualitative comparison. Our method can generate cleaner results than previous zero-shot methods. ZS-N2N failed to generate clean results on EGFP, probably due to the strong noise in the constructed image pairs. In contrast, our method performs averaging over frames, which is less affected. We also provide quantitative evaluation in Tab. 2. The high SNR data are used as GT to calculate PSNR and SSIM. Our method achieves a higher metric than other zero-shot methods and is only worse than DeepCAD-RT (Li et al. 2023b).

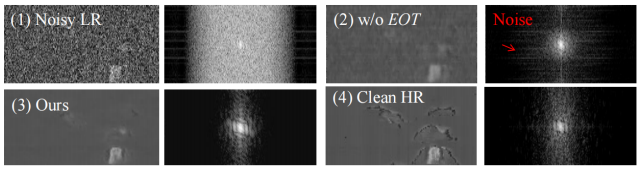
ZSDeconv dataset. We also conducted experiments on a single noisy image from ZSDeconv paper (Qiao et al. 2024). The paper provides a pair of wide-field images of Lysosome. We compare with zero-shot methods, such as ZS-N2N (Mansour and Heckel 2023b), N2F (Lequyer et al. 2022) and ZSDeconv (Qiao et al. 2024). We show the qualitative comparison in Fig. 8. Our method can generate more similar results as clean GT with higher PSNR and SSIM, while ZSDeconv generate an over-smoothing result.

Ablation Study

Fig. 9 (a) shows the output of our model with different loss functions and corresponding Fourier Spectrum. We can draw from the figure that high frequency is replaced with



(a) Ablation on L_{SR}



(b) Ablation on *EOT*

Figure 9: Visualization for the ablation study.

noise in the noisy LR. The LR self-supervision will cause the model to only focus on the restoration of low frequency and produce overly smooth results. The trade-off between noise reduction and detail retention poses a great challenge. Therefore, we propose the denoising-on-super-resolution to address both problems and recover details. With SR self-supervision, the model can restore more high-frequency details of the image.

Fig. 9 (b) visualize the temporal noise and the effectiveness of EOT. The noisy LR has a strong temporal noise along the time axis. The denoising network without EOT is separately trained in a frame-by-frame fashion. The results of the frame-by-frame method still have noticeable high-frequency noise. It is because the model cannot handle the temporal noise. Our model addresses the temporal noise by averaging the model weight over the frames, which produces a time-consistent denoising.

Tab. 3 shows quantitative experiments of the ablation study. We use the same network for all experiments. We evaluate in PhC-C2DH-U373 dataset with Poisson and Gaussian noise conditions, and show the average metrics of noise levels of 20, 30, and 40. The model performance will improve with L_{SR} with and without EOT. Tab. 3 also With EOT, our model improves around 2dB in PSNR under both Gaussian and Poisson noise.

We also note that Fig. 4 shows the experiment with different iterations and hyperparameter α. We can draw the conclusion that longer training will improve the performance, but cost more time. The selection of α is also important. It depends on the motion significance of the objects. For our experiments, $\alpha = 0.9$ is the best choice with the PhC dataset.

Efficiency Analysis

Tab. 4 shows the comparison of the computational costs. The time is computed for training and inferencing a single frame at 348×256 resolution. Therefore, the time is not comparable for self-supervised methods like UDVD, DeepCAD-RT and DeepSeMi, which are trained on the whole sequence in advance. For fair comparison, we also test N2F and ZS-N2N under the same network and iteration numbers as our model. All time costs are tested on a RTX 3090 GPU. Comparing to self-supervised models, our method does not need an extra training process and can perform denoising plug-and-play. Compared to the zero-shot methods, our method can achieve better performance with the same training and inference time or save time for the same denoising performance. Fig. 2 shows more comparison with the zero-shot method with different epochs. Our method can adapt to different computing power by changing the training iterations.

L_{LR}	L_{SR}	EOT	Gaussian	Poisson
✓			31.97/0.8712	32.58/0.8803
✓		✓	35.20/0.9303	35.34/0.9347
✓	✓		33.27/0.8827	33.37/0.8854
✓	✓	✓	**35.33/0.9307**	**35.42/0.9348**

Table 3: Ablation study of our model. **Bold** text indicates the best results.

Models	Param(M)↓	Time(s)↓	PSNR ↑	SSIM ↑
UDVD	2.124	-	35.30	0.9336
DeepCAD-RT	1.020	-	35.30	0.9377
DeepSeMi	0.068	-	30.06	0.9160
N2F	0.259	2.095	33.86	0.9232
N2F*	0.852	6.169	34.66	0.9302
ZS-N2N	0.021	11.73	33.31	0.8923
ZS-N2N*	0.853	9.370	26.15	0.4315
Ours (200)	0.884	4.265	**35.42**	**0.9348**
Ours (100)	0.884	2.109	35.41	0.9324
Ours (50)	0.884	0.941	35.18	0.9241
Ours (10)	0.884	0.235	33.78	0.8645

Table 4: Efficiency analysis. The time is for denoising a single frame. The time for the self-supervised methods is discarded, since extra training on the whole sequence is needed. For fair comparison, we also test N2F and ZS-N2N with the same network and epochs as our model, denoted as *. We also show average PSNR and SSIM metrics on PhC with Poisson noise.

Conclusion

In this paper, we introduce MDSR-Zero, an efficient training strategy for zero-shot denoising and super-resolution model of microscopy videos. Our approach uses an efficient online training (EOT) strategy that reuses denoising models from previous frames, treating the video as a continuous stream to reduce training time and ensure consistent denoising. We also propose a novel loss function for joint super-resolution and denoising, improving detail in the results. Extensive experiments show that our method achieves state-of-the-art performance among zero-shot denoising models and competes with SOTA self-supervised methods.

Acknowledgements

This work is supported by NSFC (grant nos. 62472102 to B.Y. and 62372117 to W.T.) and by the Natural Science Foundation of Shanghai (24ZR1490400 to W. T.). The computations in this research were performed using the CFFF platform of Fudan University.

References

Batson, J.; and Royer, L. 2019. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, 524–533. PMLR.

Cai, Z.; Ravichandran, A.; Maji, S.; Fowlkes, C.; Tu, Z.; and Soatto, S. 2021. Exponential moving average normalization for self-supervised and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 194–203.

Chenouard, N.; Smal, I.; De Chaumont, F.; Maška, M.; Sbalzarini, I. F.; Gong, Y.; Cardinale, J.; Carthel, C.; Coraluppi, S.; Winter, M.; et al. 2014. Objective comparison of particle tracking methods. *Nature methods*, 11(3): 281–289.

Dertinger, T.; Colyer, R.; Iyer, G.; Weiss, S.; and Enderlein, J. 2009. Fast, background-free, 3D super-resolution optical fluctuation imaging (SOFI). *Proceedings of the National Academy of Sciences*, 106(52): 22287–22292.

Dewil, V.; Anger, J.; Davy, A.; Ehret, T.; Facciolo, G.; and Arias, P. 2021. Self-supervised training for blind multi-frame video denoising. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2724–2734.

Ehret, T.; Davy, A.; Morel, J.-M.; Facciolo, G.; and Arias, P. 2019. Model-blind video denoising via frame-to-frame training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11369–11378.

Fernández-Suárez, M.; and Ting, A. Y. 2008. Fluorescent probes for super-resolution imaging in living cells. *Nature reviews Molecular cell biology*, 9(12): 929–943.

Hirano, M.; Ando, R.; Shimozono, S.; Sugiyama, M.; Takeda, N.; Kurokawa, H.; Deguchi, R.; Endo, K.; Haga, K.; Takai-Todaka, R.; et al. 2022. A highly photostable and bright green fluorescent protein. *Nature Biotechnology*, 40(7): 1132–1142.

Huang, T.; Li, S.; Jia, X.; Lu, H.; and Liu, J. 2021. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14781–14790.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krull, A.; Buchholz, T.-O.; and Jug, F. 2019. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2129–2137.

Laine, S.; Karras, T.; Lehtinen, J.; and Aila, T. 2019. High-quality self-supervised deep image denoising. *Advances in Neural Information Processing Systems*, 32.

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

Lehtinen, J.; Munkberg, J.; Hasselgren, J.; Laine, S.; Karras, T.; Aittala, M.; and Aila, T. 2018. Noise2Noise: Learning image restoration without clean data. In *International Conference on Machine Learning*, 80. PMLR.

Lequyer, J.; Philip, R.; Sharma, A.; Hsu, W.-H.; and Pelletier, L. 2022. A fast blind zero-shot denoiser. *Nature Machine Intelligence*, 4(11): 953–963.

Li, X.; Hu, X.; Chen, X.; Fan, J.; Zhao, Z.; Wu, J.; Wang, H.; and Dai, Q. 2023a. Spatial redundancy transformer for self-supervised fluorescence image denoising. *Nature Computational Science*, 1067–1080.

Li, X.; Li, Y.; Zhou, Y.; Wu, J.; Zhao, Z.; Fan, J.; Deng, F.; Wu, Z.; Xiao, G.; He, J.; et al. 2023b. Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit. *Nature Biotechnology*, 41(2): 282–292.

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

Ma, C.; Tan, W.; He, R.; and Yan, B. 2024. Pretraining a foundation model for generalizable fluorescence microscopy-based image restoration. *Nature Methods*, 1–10.

Mansour, Y.; and Heckel, R. 2023a. Zero-shot noise2noise: Efficient image denoising without any data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14018–14027.

Mansour, Y.; and Heckel, R. 2023b. Zero-Shot Noise2Noise: Efficient Image Denoising without any Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14018–14027.

Maška, M.; Ulman, V.; Delgado-Rodriguez, P.; Gómez-de Mariscal, E.; Nečasová, T.; Guerrero Peña, F. A.; Ren, T. I.; Meyerowitz, E. M.; Scherr, T.; Löffler, K.; et al. 2023. The Cell Tracking Challenge: 10 years of objective benchmarking. *Nature Methods*, 1–11.

Meiniel, W.; Olivo-Marin, J.-C.; and Angelini, E. D. 2018. Denoising of microscopy images: a review of the state-of-the-art, and a new sparsity-based method. *IEEE Transactions on Image Processing*, 27(8): 3842–3856.

Moran, N.; Schmidt, D.; Zhong, Y.; and Coady, P. 2020. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12064–12072.

Pang, T.; Zheng, H.; Quan, Y.; and Ji, H. 2021. Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2043–2052.

Qiao, C.; Zeng, Y.; Meng, Q.; Chen, X.; Chen, H.; Jiang, T.; Wei, R.; Guo, J.; Fu, W.; Lu, H.; et al. 2024. Zero-shot learning enables instant denoising and super-resolution in optical

fluorescence microscopy. *Nature Communications*, 15(1): 4180.

Quan, Y.; Chen, M.; Pang, T.; and Ji, H. 2020. Self2Self With Dropout: Learning Self-Supervised Denoising From Single Image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. 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*, 234–241. Springer.

Sheth, D. Y.; Mohan, S.; Vincent, J. L.; Manzorro, R.; Crozier, P. A.; Khapra, M. M.; Simoncelli, E. P.; and Fernandez-Granda, C. 2021. Unsupervised deep video denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1759–1768.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Wang, Z.; Liu, J.; Li, G.; and Han, H. 2022. Blind2unblind: Self-supervised image denoising with visible blind spots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2027–2036.

Wang, Z.; Zhang, Y.; Zhang, D.; and Fu, Y. 2023. Recurrent Self-Supervised Video Denoising with Denser Receptive Field. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7363–7372.

Zhang, G.; Li, X.; Zhang, Y.; Han, X.; Li, X.; Yu, J.; Liu, B.; Wu, J.; Yu, L.; and Dai, Q. 2023. Bio-friendly long-term subcellular dynamic recording by self-supervised image enhancement microscopy. *Nature Methods*, 20(12): 1957–1970.

Zheng, H.; Pang, T.; and Ji, H. 2023. Unsupervised deep video denoising with untrained network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3651–3659.