Benchmarking Self-supervised Learning for Denoising Voltage Imaging Data

Ioan Leolea[1]

Supervisors: Nergis Tömen[1]**, Alejandro Castaneda Garcia**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Ioan Leolea
Final project course: CSE3000 Research Project
Thesis committee: Nergis Tömen, Alejandro Castaneda Garcia, Chirag Raman

Abstract

Voltage imaging is a powerful technique for observing fast neural activity, but it often produces images with a high level of noise, making analysis difficult. Deep learning methods have shown promise in denoising such data, but most require large datasets containing both clean and noisy image pairs, which are hard to obtain in real-world settings. To this end, several self-supervised approaches that rely solely on noisy images have been proposed in the literature. In this paper, three self-supervised denoising models–Noise2Void, AP-BSN, and DeepVID v2–are evaluated on both synthetic and real voltage imaging datasets. For the synthetic data, the performance is assessed using PSNR and SSIM, while for the real data, the temporal signal-to-noise ratio (tSNR), a metric well-suited to voltage imaging, is used. Results show that the self-supervised models are effective at denoising both synthetic and real image datasets. In particular, models which use the temporal information of the videos, such as DeepVID v2, obtain the best results.

1 Introduction

All brain functions, such as cognition, emotion, behavior, movement, and sensation, are the result of electrical signals generated by neurons organized in systems of circuits. A central goal of neuroscience is to find out how these neuronal circuits give rise to brain functions. One way in which the mechanisms of these circuits can be studied is by examining the behavior of large neuron populations in the cerebral cortex using optical imaging methods. Complications in this procedure arise from the fact that electrical responses, which take place on a millisecond timescale, need to be analyzed. Namely, the current state-of-the-art method for capturing neuronal activity, calcium imaging, can't provide data on inhibitory and excitatory signals, which take place at all times in most neurons.

Voltage imaging is a relatively new technique that enables capturing rapid neural activity with high spatial and temporal precision using genetically encoded voltage indicators in fluorescence microscopy. The advantage of voltage imaging over calcium imaging is its temporal resolution. The method is fast enough to capture all the stages of neural spikes (including inhibitory and excitatory events) of each neuron in a circuit [1].

Because of the low photon yield, the videos obtained using this method have a low signal-to-noise ratio (SNR). In turn, this impedes accurately localizing the neurons' positions and lowers the temporal precision of neural spike detection. Increasing the SNR experimentally through longer exposure or stronger light excitation offers limited improvements. Therefore, the preferred alternative is increasing the SNR through image denoising.

Image denoising is an essential image processing task, which aims to remove the noise and separate the clean image. Before the rise in popularity of deep learning, the state-of-the-art in microscopy denoising consisted of total variation-based methods, non-local methods, sparse filtering, and variance-stabilizing transforms [2]. More recently, convolutional neural network (CNN)-based methods have surpassed the performance of traditional algorithms [3, 4].

Despite the promising performance of deep learning-based methods, they have limited utility in a real-world setting, since training these models requires a large number of pairs of noisy and clean images of the same instance. This is especially problematic for voltage imaging because significant changes in the frames happen on a millisecond timescale, so obtaining clean images is very complicated.

To solve this problem, several self-supervised methods that require only noisy images for training have been proposed in the last five years [5]. More recently, self-supervised models specialized in denoising microscopy images [6] have emerged. Self-supervised methods have shown promising results on general datasets. Though, due to the recency of the voltage imaging method and scarcity of such datasets, little research has been done on the effectiveness of these models on voltage imaging data.

This work aims to benchmark a wide variety of self-supervised models on both real and synthetic voltage imaging data, using suitable metrics, in order to answer the following question:

How do different self-supervised denoising models–Noise2Void, AP-BSN, and DeepVID v2–compare in their effectiveness at denoising synthetic and real-world voltage imaging data, as measured by PSNR, SSIM, and temporal SNR?

2 Prerequisites and related work

In this section, prerequisites regarding the structure of noise in voltage imaging videos, deep learning concepts and techniques commonly used in self-supervised denoising will be presented.

2.1 Noise distribution in voltage imaging

In voltage imaging videos, three different types of noise are usually present [2]:

- **Dark noise**, caused by the thermal agitation of the electrons in the camera's detector.
- **Photon noise**, which results from the fluctuations in the number of photons detected by the camera.
- **Readout noise**, introduced by imperfections in the electronics of the camera's output amplifier.

The first two types of noise are described by Poisson processes in literature, while the last one is described by a Gaussian noise model. A widely used and realistic noise distribution in cases where all three aforementioned types of noise are present is the Poisson-Gaussian noise model [7]. Mathematically this model is described for each pixel coordinate by

$$P\left(y(s)=g\right)=\frac{e^{-\lambda_s}}{\sqrt{2\pi\sigma^2}}\times\sum_{p=0}^{+\infty}\frac{\lambda_s^p}{p!}e^{-\frac{(\gamma p-g)^2}{2\sigma^2}}$$

where $y(s)$ represents the pixel value at a coordinate s in the image, g is a gray level value and σ is the standard deviation of the noise model, λ_s is the Poisson random process of intensity parameter and γ is a gain constant that modulates how much the Poisson noise affects the model.

A method commonly used in self-supervised denoising models is assuming this Poisson-Gaussian prior distribution, estimating its parameters (or knowing them beforehand) and then performing the training and prediction using both the dataset and this information [8, 9]. However, the drawback of such models is that they perform very well on synthetic data, but fail to generalize on real world data [10].

2.2 Blind-Spot Networks

Among the self-supervised models used for image denoising, the most popular approach is that of utilizing a Blind-Spot Network [5].

The general denoising problem is described by a noisy image $\boldsymbol{x}$ generated from a clean image $\boldsymbol{s}$ and some noise $\boldsymbol{n}$ on top of it, in other words $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}$. To solve the problem, a way to compute a prediction $\hat{\boldsymbol{s}}$ as close as possible to the original image $\boldsymbol{s}$ needs to be found.

In deep-learning terms, the solution is training a CNN to learn a mapping from $\boldsymbol{x}$ to $\boldsymbol{s}$. As a more illustrative view on the problem, when predicting an image from a given input image, every predicted pixel $\hat{\boldsymbol{s}}_i$ is influenced by certain pixels in the input image $\boldsymbol{x}_{RF(i)}$, this is called the receptive field of the pixel. Thus the network can be seen as a function that maps each receptive field to a predicted pixel, which can be defined by:

$$f(\boldsymbol{x}_{RF(i)}, \boldsymbol{\theta}) = \hat{\boldsymbol{s}}_i$$

where $\boldsymbol{\theta}$ represents the parameters of the model

The first ever Blind-Spot Network for denoising was proposed in [11]. The idea behind this type of network is that since no pairs of noisy-clean images on which the network can be trained are available, one possibility is to use pairs of the same noisy image for training. An important consideration in this method is that the noise distribution is pixel-wise independent. Using a classical CNN with this approach would result in the model learning the identity function. The reason for it is that, for every target pixel, its receptive field would already contain that pixel and the model will learn to simply output the original image at all times. To solve this issue, certain pixels are blocked or neglected in the input image during training.

The two general approaches to perform this procedure are the following:

- Randomly selecting pixels from the original image during training and substituting them with another randomly chosen pixel from their vicinity (i.e. blocking them). This modified image becomes the input of the network and a special loss function is then computed only on the blocked pixels for both the input and the target image.
- Using convolutional filters that have zero entries at all times on certain positions and a usual loss function.

The second method has the advantage that it is significantly more computationally efficient, and therefore it is used predominantly in literature. These ideas have paved the way for more advanced self-supervised denoising models which use this blueprint together with more complex techniques.

2.3 Subsampling

In the context of image processing, subsampling refers to a method which randomly selects a subset of pixels from regions of an image and uses them to generate a smaller image. This is typically done by sliding a window over the image with a stride of s, and sampling one pixel from each window position to form smaller images.

This technique can potentially offer a speed-up in the runtime of models as the dimensions of the images used are scaled down, but more importantly, when used in image denoising, along with Blind-Spot Networks, it has two potential uses:

- A noisy image is subsampled into two smaller images, which effectively should be similar to each other, that are then treated as noisy - noisy pairs of the same instance in a model such as Noise2Noise [12]. Such a strategy can be found in the Neighbor2Neighbor model [10].
- The key assumption in the classical Blind-Spot Network is that the noise distribution is pixel-wise independent. This is not the case for real-world images, as the noise in this case is highly spatially correlated. Consequently, certain models, such as AP-BSN [13], use subsampling techniques to reduce the spatial correlation of the noise in images and afterwards input them into a Blind-Spot Network.

3 Models selected for the experiment

This section presents representative self-supervised denoising models benchmarked on synthetic and real-world voltage imaging data.

3.1 Noise2Void

This is the paper that introduced the first version of the Blind-Spot Network. It uses the U-Net architecture [14], but differs by randomly masking pixels during training, as discussed in the previous section. In this model, for a predicted pixel $\hat{\boldsymbol{s}}_i$, the only input pixels affecting it are the ones in a square neighborhood (i.e. a $2N+1 \times 2N+1$ region centered at the pixel) with the exception of the input pixel $\boldsymbol{x}_i$. Consequently, the training of the model entails minimizing the following empirical loss function:

$$\arg\min_{\theta} \sum_j \sum_i L\left(f\left(\tilde{\boldsymbol{x}}^j_{RF(i)}; \boldsymbol{\theta}\right), \boldsymbol{x}^j_i\right)$$

where the superscript j refers to a particular image in the training set. The model is used as a baseline for future comparisons as, despite its age, it still performs well for certain types of noise.

3.2 AP-BSN

This model is inspired by [15], particularly through the use of pixel-shuffling down-sampling (PD), which is combined with a Blind-Spot Network.

Pixel-shuffling down-sampling is a subsampling method that preserves the noise distribution (i.e. mean and variance) of the original image, while decreasing its spatial correlation and allowing it to be used as input in a Blind-Spot Network. The subsampling procedure, as described in [15], involves the following steps:

1. Find the smallest subsampling stride s that makes the downsampled spatial correlated sub-images match the pixel-independent noise distribution.
2. Pixel-shuffle the image into a mosaic y_s.
3. Denoise y_s using the denoising model.
4. Refill each sub-image with noisy blocks separately and pixel-shuffle up-sample them.
5. Denoise each refilled image again using the denoising model and average them to obtain the texture details T.
6. Obtain the flat regions of the image F using a noise estimating model.
7. Combine T and F to get the final image.

The problem AP-BSN aims to solve is the inherent trade-off in the classical PD between the pixel-wise independence assumption and the reconstruction quality: for low stride factors s, image structure is preserved, but the spatial correlation is not significantly reduced, whereas for high stride factors s, the opposite occurs. The solution proposed in the paper is using different stride factors during training and inference time, while also using a Blind-Spot Network as the denoiser.

Subsampling-based models like AP-BSN have not yet been evaluated on voltage imaging data. Therefore, it is important to compare their performance with models that use alternative approaches to determine whether this technique is well-suited for such data. AP-BSN was chosen for benchmarking as it is the model using subsampling techniques which consistently achieves the best denoising results in the literature on real-world RGB data. Thus, out of the models of this type, it is assumed that it will perform best on voltage imaging data too.

3.3 DeepVID v2

The model is built on top of the usual Blind-Spot Network architecture, but has two additional features specialized for denoising voltage imaging videos:

1. It leverages the temporal information in the videos by using multiple consecutive frames (a total of $N = 2N_0+1$) centered around the frame to be denoised. These frames help the network better understand motion and dynamic changes.
2. It addresses the common denoising problem of over-smoothed images by adding an edge extraction side branch. This branch first computes a local mean frame from a separate series of $M = 2M_0 + 1$ frames, applies Gaussian smoothing, and extracts directional edges using four Sobel filters (at 0°, 45°, 90°, and 135°). These edge maps are then fed into the main denoising branch as extra input channels to help preserve fine structures in the image.

This model was chosen for benchmarking as it is the follow-up version of DeepVID, a voltage imaging denoising model used in the widely adopted VolPy pipeline [16]. It is an already established model, which is known to perform well on this type of data. Comparing the other models to it can reveal whether non-specialized methods can achieve denoising performance comparable to the state of the art.

4 Experimental setup and results

The experiment entails training the models presented in the previous section on two datasets, one with synthetic voltage imaging data and the other with real-world voltage imaging data, and then evaluating their predictions on the respective datasets using specific metrics. The training and prediction were performed on a computer with 32 GB RAM and an NVIDIA GeForce RTX 5080 GPU.

4.1 Datasets

For reproducibility, the two datasets were taken from literature related to voltage imaging.

The synthetic noise dataset was taken from the CellMincer paper [17]. In the paper, four datasets at different noise levels were created using the Optosynth tool. For the analysis in this paper, only the dataset containing the most noisy images was used. The dataset consists of 10 sets of 7,000 images resembling neuron populations generated using Optosynth. Five sets contain clean images, while the other five were obtained by adding Poisson-Gaussian noise.

The real voltage imaging frames dataset was taken from [18]. It consists of 13 videos of 15,000 frames each. The dataset consists of images of neurons from different regions of the hippocampi of mice and it does not contain any ground truth. Due to the large size of the full dataset, only two videos (the files "00_02.tif" and "00_03.tif") were used for training and evaluation.

Both datasets contain only grayscale images, as is generally the case for voltage imaging videos.

4.2 Metrics

One of the datasets does not have any ground truth. Therefore the metrics used for the two datasets differ. In the formulas below x represents the ground truth image and y represents the denoised image, both of which have size $n \times m$. Moreover, the final reported values for the metrics presented below are computed for each pair of images in the dataset and then averaged.

Synthetic dataset metrics

As this synthetic dataset contains ground truth images, metrics computed between a predicted image and a clean image are used.

Structural Similarity Index Measure (SSIM) evaluates the perceptual quality of images by comparing structural information between a denoised image and the ground truth.

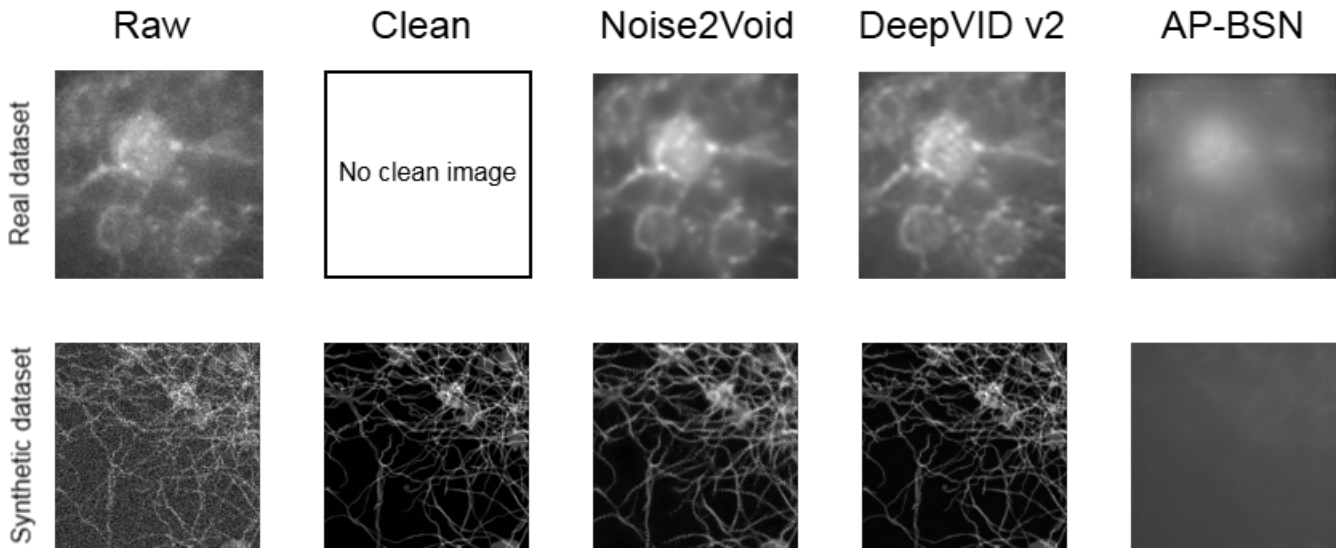

Figure 1: Visual comparison of denoising performance across models. Each row shows a different dataset: the top row corresponds to the real dataset, and the bottom row to the synthetic dataset. For each dataset, the noisy input, denoised outputs from each model, and (where available) the ground truth are shown.

The SSIM score ranges from -1 to 1, with values closer to 1 indicating greater structural similarity. SSIM was chosen as it better reflects perceptual image quality by comparing structural information.

The formula for SSIM between two standardized image patches x and y is:

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where:

- μ_x, μ_y are the means of x and y,
- σ_x^2, σ_y^2 are the variances of x and y,
- σ_{xy} is the covariance between x and y,
- C_1 and C_2 are small constants to stabilize the division.

Scale-invariant Peak Signal-to-Noise Ratio (PSNR) measures the quality of reconstructed images. Higher PSNR values indicate better denoising performance and greater similarity to the ground truth. This version of PSNR standardizes pixel values before computing the metric. The metric was chosen as it is easy to compute and sensitive to even small differences between the denoised image and the ground truth, which complements SSIM well. Moreover, it is a widely used metric in the image processing and microscopy communities [2]. The formula used to calculate it is:

$$PSNR = 10 \cdot \log_{10}\left(\frac{RANGE^2}{MSE}\right)$$

where $RANGE$ is defined as

$$RANGE = \frac{MAX(x) - MIN(x)}{\sigma_x}$$

and MSE is defined as:

$$MSE = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - y_{ij})^2$$

Real dataset metrics

The metrics used for data without ground truth are generally scarce and uninformative, but in the context of denoising voltage imaging, because of the temporal dependence of the frames, a metric called **Temporal Signal to Noise Ratio (tSNR)** can be computed. This metric is calculated by selecting a window (in this case with size of 7 frames) and calculating the ratio between the average and the standard deviation of the frames in the window:

$$TSNR = 10 \cdot \log_{10}\left(\frac{\mu}{\sigma}\right)$$

Temporal Signal to Noise Ratio is commonly used in voltage imaging denoising papers [19] and shows how similar frames close to each other in the video are.

4.3 Results

Since the evaluated models are self-supervised, they can't overfit on the available data. Consequently, they were trained on all of the noisy images for both the synthetic and real image datasets. Moreover, each model was trained using the hyperparameters suggested in their respective papers, as fine-tuning them for these datasets is not an option due to the long training times.

Figure 1 shows a comparison between the predictions of the denoising models, along with the ground truth and noisy images from each of the two datasets. Table 1 highlights the obtained PSNR and SSIM values after training and inference for each model on the synthetic dataset.

Model	PSNR (dB)	SSIM
Noise2Void	26.662	0.877
AP-BSN	13.130	0.074
DeepVID v2	33.193	0.970

Table 1: Quantitative denoising results for each model on the synthetic dataset

For the real image dataset, a baseline value of the tSNR was computed for the noisy dataset in order to see any potential improvements. Table 2 shows the obtained tSNR values after training and inference for each model on the real dataset.

Model	tSNR (dB)
Baseline	5.36
Noise2Void	5.44
AP-BSN	6.41
DeepVID v2	5.82

Table 2: Quantitative denoising results for each model on the real dataset

For the synthetic image dataset, results from Table 1 show that DeepVID v2 performed best in terms of both PSNR and SSIM by a large margin. The reason this model gave the best results is its use of the temporal image stacks. This feature gives more information about the current frame by analyzing previous and future frames from the video. This increases the performance on these metrics for the dataset, as frames are quite similar to each other. The main differences between frames are the patterns of the noise, not the positions or appearance (e.g. more intense lighting when a signal happens) of the neurons.

For the real image dataset, the best tSNR value was obtained for AP-BSN. The metric measures how different denoised frames that are close to each other in the video compare in terms of overall structure. As a result, if images in the temporal pack are very similar to one another, the value of tSNR will be high. In the case of AP-BSN, the model oversmoothed image details, making all frames look very similar. As a result, the tSNR was high despite the poor visual quality of the denoised images (see top right of Figure 1). By manually checking the denoised images and comparing them to the originals, the conclusion is that DeepVID v2 once again performs the best as it removes part of the noise, while still keeping the details of the image intact. Moreover, the same model has also performed best in terms of tSNR outside of AP-BSN. This is again due to the model's use of both previous and future frames, which naturally leads to a higher tSNR, by virtue of utilizing more temporal information.

5 Discussion

For the synthetic image dataset, the results generally align with the initial hypotheses. Noise2Void is the oldest model and is not specifically designed to handle voltage imaging datasets or account for the time-dependence of frames. Nonetheless, it scored reasonably well in terms of PSNR and SSIM. On the other hand, DeepVID v2 achieved the best performance, as expected, because it leverages the temporal dependence of frames and addresses known denoising issues such as oversmoothing. Surprisingly, AP-BSN yielded very poor results. In many papers, it consistently ranks among the best models for PSNR and SSIM on RGB image datasets, but in the current evaluation, it performed the worst by a large margin. Moreover, its output images are oversmoothed and appear grayed out. This could be due to the model being primarily designed for real-world RGB image denoising. It is possible that the subsampling procedure used by AP-BSN is simply ill-suited for simple images with very small areas of interest–such as the dendrites of neurons–leading to a loss of essential structure.

For the real image dataset, the highest tSNR value was achieved by AP-BSN, but this appears to result from oversmoothing rather than true denoising effectiveness. DeepVID v2 shows a noticeable improvement in tSNR compared to the baseline, which is expected, as it is the only model that uses temporal stacks of frames. Noise2Void, on the other hand, shows no significant improvement. Interestingly, for this dataset, the images denoised by AP-BSN resemble the original raw images more closely than those from the synthetic dataset. This seems to support the hypothesis that AP-BSN performs better on images with higher resolution or larger areas of interest. This conclusion is further supported by its strong performance on real-world RGB datasets with large images, as shown in [13].

The analysis presented here has several limitations. First, the number of videos used from the real dataset is small due to computational constraints, which may affect the generalizability of the results. Second, the absence of ground truth data in the real dataset means that tSNR, while useful, is an imperfect proxy for actual denoising performance–evidenced by the fact that the worst-performing model achieved the highest tSNR. Third, AP-BSN was evaluated using the hyperparameters recommended in the original paper, but the images in this evaluation were grayscale, not RGB. This mismatch could partly explain the reduced performance, although different sets of hyperparameters were not tested due to the long training times of the model.

6 Conclusions and Future Work

To address the research question of how effective self-supervised denoising methods are for cleaning voltage imaging datasets, models were trained and evaluated on both synthetic and real datasets using a variety of metrics.

Results obtained from the synthetic dataset, which includes ground truth, show that self-supervised denoising models can be highly effective on voltage imaging data. Interestingly, simpler models that avoid subsampling and pixel shuffling

tend to perform better, suggesting that these techniques may be ill-suited for such datasets.

For the real image dataset, lacking ground truth, the models demonstrated modest improvements in the tSNR metric, indicating potential effectiveness. However, the model that achieved the highest tSNR produced the poorest visual results, calling into question the validity of tSNR as a reliable quality metric for denoising in this context.

To obtain more comprehensive insights into the effectiveness of self-supervised denoising methods for voltage imaging, further research is needed. In particular, this work does not explore Blind-Spot Network–based models that incorporate subsampling or prior distribution assumptions, other than AP-BSN. A challenge in evaluating such models lies in their often outdated or incompatible codebases. Additionally, more recent architectures–such as those based on Transformers or Generative Adversarial Networks (GANs)–are promising for this task but were not included due to time constraints.

Another limitation of this paper is the reliance on a real dataset without ground truth. More robust evaluation could be achieved by using voltage imaging datasets that contain both high-SNR (clean) and low-SNR (noisy) recordings of the same scene. Unfortunately, such datasets are rare, as the underlying biological processes are too fast to capture without introducing significant noise.

7 Responsible research

In this section, the commitment to transparency, reproducibility, and the ethical context of the research is outlined to ensure the work meets high standards of scientific integrity.

7.1 Reproducibility

To ensure that the findings can be independently verified and built upon, the reproducibility of the experiments has been prioritized.

- **Public Datasets:** The evaluation relies exclusively on publicly available datasets from prior scientific literature. The synthetic data is from the Cellmincer study [17], and the real-world voltage imaging data is from [18]. The use of established, open datasets ensures that other researchers can perform direct comparisons using the same source material.
- **Methodological Transparency:** The specific models evaluated and the metrics used for assessment have been clearly detailed. These metrics are standard in the fields of image processing and voltage imaging analysis.
- **Hardware and Environment:** The computational hardware used for training and inference has been specified to provide a baseline for performance replication.
- **Open Source Commitment:** Code developed for training the models and the final model checkpoints can be found at https://github.com/ileolea317/voltage_imaging_self_supervised_denoising.

7.2 Ethical Considerations

This research is computational in nature and focuses on improving a data processing technique for basic scientific inquiry. While the direct work of developing a denoising algorithm does not pose ethical issues, the context in which the data is generated and the potential impact of the research are acknowledged.

- **Use of Animal Data:** The real-world dataset used in this paper was derived from experiments involving mice, as documented in the original publication [18]. This work relies on secondary data, and it is operated under the assumption that the original researchers adhered to all institutional and national guidelines for the ethical treatment and welfare of laboratory animals.
- **Societal Impact:** The primary goal of improving voltage imaging analysis is to accelerate the understanding of neural circuits. Enhanced denoising techniques can lead to more accurate data, which in turn can advance neuroscience and contribute to a better understanding of brain function and pathologies. The application of this work is believed to be firmly rooted in a positive contribution to scientific and medical knowledge.

References

[1] Thomas Knöpfel and Chenchen Song. "Optical voltage imaging in neurons: moving from technology development to practical tool". In: *Nature Reviews Neuroscience* 20.12 (2019), pp. 719–727.

[2] William Meiniel, Jean-Christophe Olivo-Marin, and Elsa D. Angelini. "Denoising of Microscopy Images: A Review of the State-of-the-Art, and a New Sparsity-Based Method". In: *IEEE Transactions on Image Processing* 27.8 (2018), pp. 3842–3856.

[3] Rini Thakur, R.N. Yadav, and Lalita Gupta. "State-of-Art Analysis of Image Denoising Methods using Convolutional Neural Networks". In: *IET Image Processing* 13 (Oct. 2019).

[4] Kai Zhang et al. "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising". In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3142–3155.

[5] Dan Zhang et al. *Unleashing the Power of Self-Supervised Image Denoising: A Comprehensive Review*. 2024.

[6] Minho Eom et al. "Statistically unbiased prediction enables accurate denoising of voltage imaging data". In: *bioRxiv* (2022).

[7] Ajay Kumar Boyat and Brijendra Kumar Joshi. "A Review Paper: Noise Models in Digital Image Processing". In: *CoRR* abs/1505.03489 (2015).

[8] Alexander Krull et al. "Probabilistic Noise2Void: Unsupervised Content-Aware Denoising". In: *Frontiers in Computer Science* 2 (Feb. 2020).

[9] Jaeseok Byun, Sungmin Cha, and Taesup Moon. *FBI-Denoiser: Fast Blind Image Denoiser for Poisson-Gaussian Noise*. 2021.

[10] Tao Huang et al. "Neighbor2Neighbor: A Self-Supervised Framework for Deep Image Denoising". In: *IEEE Transactions on Image Processing* 31 (2022), pp. 4023–4038.

[11] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. “Noise2Void - Learning Denoising from Single Noisy Images”. In: *CoRR* abs/1811.10980 (2018).

[12] Jaakko Lehtinen et al. “Noise2Noise: Learning Image Restoration without Clean Data”. In: *CoRR* abs/1803.04189 (2018).

[13] Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. *AP-BSN: Self-Supervised Denoising for Real-World Images via Asymmetric PD and Blind-Spot Network*. 2022.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *CoRR* abs/1505.04597 (2015).

[15] Yuqian Zhou et al. “When AWGN-based Denoiser Meets Real Noises”. In: *CoRR* abs/1904.03485 (2019).

[16] Changjia Cai et al. “VolPy: automated and scalable analysis pipelines for voltage imaging datasets”. In: *bioRxiv* (2020).

[17] Brice Wang et al. “Robust self-supervised denoising of voltage imaging data using CellMincer”. In: *bioRxiv* (2024).

[18] Yosuke Bando et al. “Real-time Neuron Segmentation for Voltage Imaging”. In: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, Dec. 2023, pp. 813–818.

[19] C. Liu et al. “DeepVID v2: Self-Supervised Denoising with Decoupled Spatiotemporal Enhancement for Low-Photon Voltage Imaging”. In: *bioRxiv* (2024). Preprint.