2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)
April 3-7, 2020, Iowa City, Iowa, USA

Fully Unsupervised Probabilistic Noise2Void

Mangal Prakash[1,2*], *Manan Lalit*[1,2*], *Pavel Tomancak*[2],
Alexander Krull[1,2†], *Florian Jug*[1,2†]

[1]Center for Systems Biology Dresden (CSBD)
[2]Max-Planck Institute of Molecular Cell Biology and Genetics
[*]equal contribution [†]joint supervision

ABSTRACT

Image denoising is the first step in many biomedical image analysis pipelines and Deep Learning (DL) based methods are currently best performing. A new category of DL methods such as Noise2Void or Noise2Self can be used fully unsupervised, requiring nothing but the noisy data. However, this comes at the price of reduced reconstruction quality. The recently proposed Probabilistic Noise2Void (PN2V) improves results, but requires an additional noise model for which calibration data needs to be acquired. Here, we present improvements to PN2V that (i) replace histogram based noise models by parametric noise models, and (ii) show how suitable noise models can be created even in the absence of calibration data. This is a major step since it actually renders PN2V fully unsupervised. We demonstrate that all proposed improvements are not only academic but indeed relevant.

Index Terms— unsupervised denoising, deep learning, microscopy, noise model, gaussian mixture model, bootstrapping

1. INTRODUCTION

With the advent of Deep Learning (DL), the field of biomedical image denoising has recently taken rapid strides [1, 2, 3, 4, 5, 6, 7]. Today, Content-Aware image REstoration (CARE) methods are leading the field due to their *content awareness* – learning a strong prior on the visual nature of the data to be reconstructed [1, 8, 9, 10].

While CARE was initially proposed using pairs of noisy and clean (ground truth) images during training, several ways to circumvent this requirement have been proposed. Noise2Noise [11] shows how corresponding noisy image pairs can lead to virtually the same results. Self-supervised models, like Noise2Void (N2V) [5] and Noise2Self [12] show how even the requirement for a second noisy image can be avoided. These methods can train directly on the body of data to be denoised, making them extremely useful for practical applications. However, self-supervised methods are known to perform less well than models trained using paired training data [5, 6].

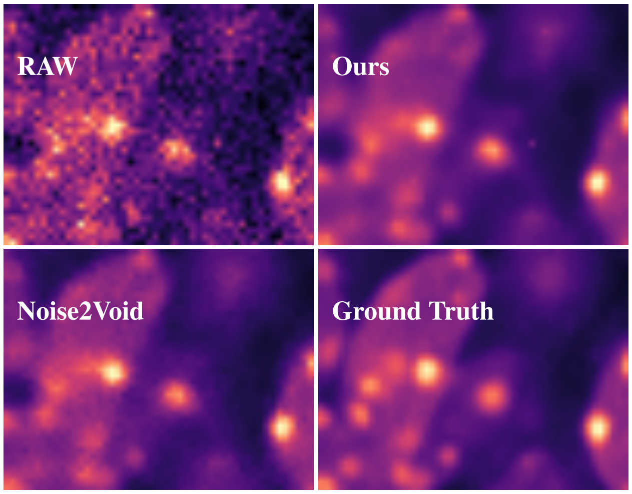

Fig. 1: Our proposed GMM bootstrapping approach does not require paired training or calibration data, but achieves superior results compared to other fully unsupervised methods.

The recently proposed Probabilistic Noise2Void (PN2V) [6] shows how using sensor specific noise models can improve the quality of self-supervised denoising, bringing it close to traditional paired training. A PN2V noise model is computed from a sequence of noisy calibration images and characterizes the distribution of noisy pixel around their respective ground truth signal value. In the context of PN2V, noise models are a collection of histograms [6].

In this work we make three major contributions. (i) We improve PN2V by introducing parametric noise models based on Gaussian Mixture Models (GMM) and show why they perform better than histogram based representations. (ii) We show how to bootstrap a suitable noise model, even in the absence of calibration data. This renders PN2V fully unsupervised, where nothing besides the data to be denoised is required for the method to be applied. (iii) All calibration data and corresponding noisy image data is made publicly available together with the code (`github.com/juglab/ppn2v`).

978-1-5386-9330-8/20/$31.00 ©2020 IEEE

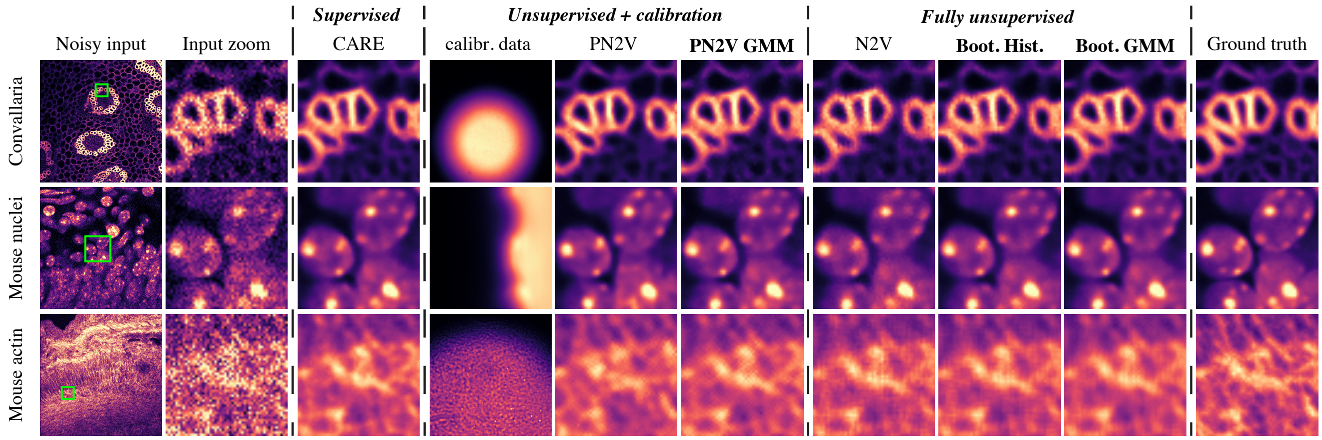

Fig. 2: A visual comparison of results obtained by CARE, N2V, PN2V, and our proposed methods (bold). We distinguish three families of methods: fully supervised (CARE), unsupervised but requiring additional calibration data (PN2V, our PN2V GMM), and fully unsupervised (N2V, PN2V using our bootstrapped histogram and GMM based noise models). The leftmost column in the unsupervised+calibration category shows the average of all available calibration images used for PN2V and PN2V GMM (see main text). Note that results of our fully unsupervised methods reach very similar quality to methods requiring either clean GT, or additional calibration data.

Methods	*Convallaria*	Mouse nuclei	Mouse actin
CARE	**36.71±0.026**	**36.58±0.019**	**34.20±0.021**
PN2V	**36.51±0.025**	36.29±0.007	33.78±0.006
PN2V GMM	36.47±0.031	**36.35±0.018**	**33.86±0.018**
N2V	35.73±0.037	35.84±0.015	33.39±0.014
Boot. Hist.	36.19±0.016	36.31±0.013	33.61±0.016
Boot. GMM	**36.70±0.012**	**36.43±0.014**	**33.74±0.012**

Table 1: Comparision of the denoising performance of all tested methods. Mean PSNR and ±1 standard error over five repetitions of each experiment are shown. Names of our proposed methods are shown in bold. Bold numbers indicate the best performing method in its respective category (supervised, unsupervised + calibration, and fully unsupervised; from top to bottom, separated by dashed lines).

2. PROPOSED APPROACHES AND METHODS

Histogram based noise models, as originally suggested for PN2V, are built from a stack of calibration images $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m$. The imaged structures in this sequence can be arbitrary but must be static. Such images can, for example, be recorded by imaging the back illuminated half opened field diaphragm (see Fig. 2). In order to minimize the effects of vibrations and sample drift, we recommend to acquire calibration data in defocus. We call the average signal $\boldsymbol{s} = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{x}^j$ ground truth (GT). It is an established protocol to average multiple static but noisy acquisitions to obtain a corresponding GT image [2]. By discretizing each GT pixel signal s_i and corresponding noisy observations x_i^j, a histogram can be created for each GT signal covering all corresponding noisy observations. The normalized set of histograms constitutes the camera noise model used in PN2V [6], describing the distribution of noisy pixel values $p(x_i|s_i)$ that are to be expected for each GT signal.

GMM based noise models describe the distribution of noisy observations x_i for a GT signal s_i as the weighted average of K normal distributions:

$$p(x_i|s_i) = \sum_{k=1}^{K} \alpha_k(s_i) f\left(\mu_k(s_i), \sigma_k^2(s_i)\right), \tag{1}$$

where $f\left(\mu_k(s_i), \sigma_k^2(s_i)\right)$ is the probability density function of the the normal distribution. We define each component's weight $\alpha_k(s_i)$, mean $\mu_k(s_i)$, and variance $\sigma_k^2(s_i)$ as a function of the signal s_i. To ensure all weights are positive and sum to one we define

$$\alpha_k(s_i) = \exp\left(g_k^{\alpha}(s_i)\right) / \sum_{k'=1}^{K} \exp\left(g_{k'}^{\alpha}(s_i)\right), \tag{2}$$

where $g_{k'}^{\alpha}(s_i)$ is a polynomial of degree n. To ensure that our distributions are always centered around the true signal s_i, we define $\mu_k(s_i) = s_i + g_k^{\mu}(s_i) - \sum_{k'=1}^{K} \alpha_{k'}(s_i) g_{k'}^{\mu}(s_i)$, where $g_k^{\mu}(s_i)$ is again a polynomial of degree n. Finally, to ensure numerical stability, we define the variance $\sigma_k^2(s_i) = \max(g_k^{\sigma}(s_i), c)$, where $c = 50$ is a constant, and $g_k^{\sigma}(s_i)$ is again a polynomial of degree n. Hence, our GMM based noise model is fully described by the $3 \times K \times n$ long vector of polynomial coefficients $\boldsymbol{a}$. We use a maximum likelihood approach to fit the parameters to our calibration data, optimizing for

$$\underset{\boldsymbol{a}}{\arg\max} \sum_{i,j} \log p(x_i^j|s_i), \tag{3}$$

where $p(x_i^j|s_i)$, is the GMM as described in Eq. 1. We use numerical optimization, see Section 3.

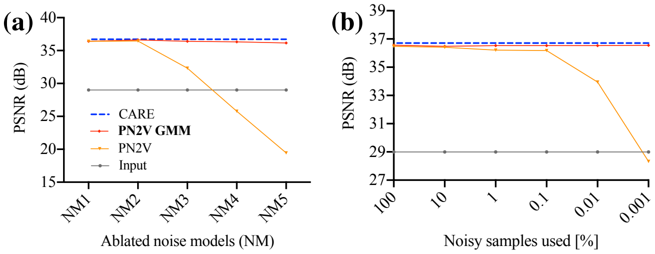

Fig. 3: Ablation studies on *Convallaria* data. Denoising performance of PN2V with histogram and linear GMM noise models is shown. **(a)** The five noise models we tested are deduced from subsets of the available calibration data, such that: the entire range of signals used in the *Convallaria* data is covered (NM1), only the lower 40% are covered (NM2), the lower 25% (NM3), 15% (NM4) and 9% (NM5). **(b)** This case is obtained by reducing the fraction of available noisy calibration pixels from NM1, via random subsampling.

Gaussians	Two coefficients	Three coefficients
1	36.56±0.022	36.34±0.040
2	36.48±0.020	36.35±0.014
3	36.47±0.031	36.31±0.022

Table 2: Testing a variety of GMM hyper-parameters. We tested GMMs using one, two, and three Gaussians, each using linear ($n = 2$) and quadratic ($n = 3$) parametrizations (see Section 2), to denoise the *Convallaria* data. The table always shows the mean PSNR and standard error over 5 repetitions.

Bootstrapped PN2V allows us to address the scenarios where no calibration data is available, *e.g.*, data that was acquired without denoising in mind. We propose the following bootstrapping procedure. First, we train and apply the unsupervised N2V [5] on the body of available noisy images $\boldsymbol{x}^j$. Then, we treat the resulting denoised images $\hat{\boldsymbol{s}}^j$ as if they were the GT, henceforth calling them pseudo ground truth. We can now use the corresponding noisy x_i^j and denoised $\hat{s}_i^j$ pixel values to either construct a histogram or learn a GMM based noise model.

3. EXPERIMENTS AND RESULTS

Datasets: We acquired three datasets (Fig. 2) which are made publicly available: (i) *Convallaria* data, available online as part of PN2V, consisting of 100 calibration images (diaphragm images, as previously explained) and 100 noisy images of a *Convallaria* section, (ii) mouse skull nuclei dataset consisting of 500 calibration images (showing the edge of a fluorescent slide) and 200 noisy realizations of the same static mouse skull nuclei, and (iii) mouse actin data consisting of 100 calibration images (diaphragm images with only the sample mounting medium in field of view) and 100 noisy realizations of the same static actin sample. The *Convallaria* and mouse actin datasets are acquired on a spinning disc confocal microscope while the mouse skull nuclei dataset is acquired with a point scanning confocal microscope.

Gaussians	NM1	NM2	NM3	NM4	NM5
1	36.56	36.03	35.98	35.85	35.78
3	36.47	36.58	36.37	36.20	36.08

Table 3: Denoising performance of PN2V GMM with linear noise models using one versus three Gaussians. For each case, five noise models were derived from different subsets of the available calibration data (see Fig. 3). We report the mean PSNR over 5 repetitions for each setup.

Implementation and training details: All evaluated training schemes are based on the implementation from [6] and use the same network architecture: a U-Net [13] with depth 3, 1 input channel, and 64 feature channels in the first layer. All networks are trained with ADAM [14] with initial learning rate of 0.001, a patch size of 100, a batch size of 1, a virtual batch size of 20 and the standard learning rate scheduler as used in [6]. Training is done for 200 epochs, each consisting of 5 steps. We use the *N2V* and *CARE* (traditional supervised training) implementations from [6].

With *PN2V*, we will refer to the version with the original histogram based noise model, derived form the available calibration data. As in [6], for each dataset, we use a $B \times B$ bin discretization, where B is an integer determined in an empirically optimal manner for which the denoising performance (PSNR) of histogram based PN2V is maximized. The minimum and maximum bins are set to the minimum and maximum values present in the data to be denoised.

Whenever we use our proposed GMM noise model, we will label results with *PN2V GMM*. As long as not stated differently, all GMM noise models use $K = 3$ Gaussians and $n = 2$ coefficients per parameter, and are trained on the available calibration data. Starting from a random initialization, optimization is performed using ADAM with learning rate 0.1, using a batch size of 25000 and 4000 iterations for mouse skull nuclei and mouse actin datasets, and a batch size of 250000 and 2000 iterations for the *Convallaria* data.

For bootstrapped PN2V (histogram and GMM based), we use the same setup as for PN2V but naturally taking the bootstrapped noise models instead. They are referred to as *Boot. Hist.* and *Boot. GMM* respectively. For the latter, we disregard the top and bottom 0.5% percentile of the pseudo GT pixels during noise model training, as we empirically observe that their N2V predictions can be often unreliable.

Comparing different training schemes: For each dataset and denoising method, we repeated each experiment 5 times and then compared the denoised images in terms of peak-signal-to-noise-ratio (PSNR) to available GT images. Results can be seen in Fig. 2, as well as Table 1. We also evaluated the structural similarity (SSIM) score for all datasets and made them available at `github.com/juglab/ppn2v/wiki`.

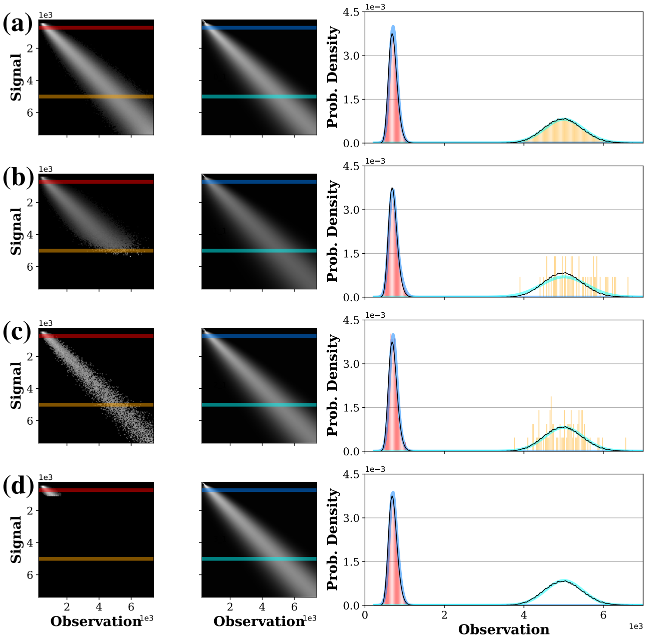

Fig. 4: Noise models for *Convallaria* data. Left column shows histogram based noise models, the center column their respective GMM based equivalent. Rightmost column shows noise models for specific signals (colors, histograms shown as vertical lines). For comparison, full calibration data histogram is always included as black curve. **(a)** Noise models trained on full calibration data. **(b)** Bootstrapped noise models. **(c)** Noise models trained on sub-sampled (0.1%) calibration data (Fig. 3b). **(d)** Noise models trained on reduced available calibration data (NM5 from Fig. 3a).

Naturally, the fully supervised CARE networks, trained on clean ground truth images, show the best performance on all datasets. On the mouse actin dataset, PN2V using our GMM based noise model derived from high quality calibration data outperforms all other methods. Notably, on the other two datasets, our fully unsupervised bootstrapped approach provides superior results and is remarkably close to CARE. For a discussion of these results, see Section 4.

Ablation and parameter study: Next we compare the robustness of histogram and GMM based noise models with respect to increasingly imperfect calibration data, using the *Convallaria* dataset as an example. These *ablation studies* consist of two scenarios, where (i) the available calibration data covers less and less of the range of signals in the data to be denoised, and (ii) the amount of available calibration pixels decreases successively. Figure 4 (c,d) shows example noise models that are derived from ablated calibration data. Evidently, for both ablation tests PN2V GMM performance is more robust compared to PN2V (see Fig. 3).

We also investigated the sensitivity of GMM noise models with respect to the chosen hyper parameters. We performed a parameter study, varying the number of Gaussian kernels K and polynomial coefficients n, using the *Convallaria* dataset with full available calibration data. Results are summarized in Table 2. While these tests suggest that the simple linear model (one Gaussian, two coefficients) is slightly preferable, the performance of all configurations remains superior to N2V (see Table 1). We additionally measured the performance of a linear noise model using 1 Gaussian and 3 Gaussians with imperfect calibration data (see Table 3). We observe that a noise model with 3 Gaussians leads to more stable results.

4. DISCUSSION

We presented a GMM based variation of PN2V noise models and showed that they can achieve higher reconstruction quality even with imperfect calibration data (Fig. 3). Additionally, we introduced a novel bootstrapping scheme, which allows PN2V to be trained fully unsupervised using only the data to be denoised (Fig. 4(b)). Our results (Table 1) show that the denoising quality of bootstrapped PN2V is quite close to fully supervised CARE [1] and significantly outperforms N2V [5]. Hence, if calibration data for a given microscope is unavailable, bootstrapping offers an excellent alternative.

Interestingly, at times, bootstrapped GMM based noise models even outperform models derived from calibration data. A possible reason for such good performance is that the distribution of pseudo GT signals used in bootstrapping corresponds well to the distribution of signals in the data to be denoised. The distribution of GT signals in the calibration data however, can be quite different.

GMM noise models, trained according to Eq. 3, prioritize signals that are abundant in the (pseudo) GT and provide a better fit in these regions compared to others. Figure 4(b) corroborates that our bootstrapped GMM fits well to the true noise distribution for lower signals, which frequently occur in the *Convallaria* data, but fails for higher signals. However, the GMM trained on calibration data (Fig. 4(a)), prioritizes its fit for higher signals, which are frequent in the calibration data, but barely present in the *Convallaria* dataset.

We strongly believe that the methods we propose will help to make high quality DL based denoising an easily applicable tool that does not require the acquisition of paired training data or calibration data. This would facilitate a plethora of projects in cell biology, where the processes to be imaged are very photosensitive or so dynamic that suitable training image pairs cannot be obtained.

5. ACKNOWLEDGEMENTS

The authors thank the LM Facility, Diana Afonso and Jacqueline Tabler from MPI-CBG for kindly sharing samples and expertise, and Matthias Arzt from CSBD/MPI-CBG for helpful discussions. Work was supported by the BMBF under the codes 031L0102 (de.NBI) and 01IS18026C (ScaDS2), and the DFG under code JU3110/1-1 (FiSS).

6. REFERENCES

[1] Martin Weigert, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Akanksha Jain, Benjamin Wilhelm, Deborah Schmidt, Coleman Broaddus, Siân Culley, et al., "Content-aware image restoration: pushing the limits of fluorescence microscopy," *Nature methods*, vol. 15, no. 12, pp. 1090, 2018.

[2] Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard, "A poisson-gaussian denoising dataset with real fluorescence microscopy images," in *CVPR*, 2019.

[3] Tim-Oliver Buchholz, Mareike Jordan, Gaia Pigino, and Florian Jug, "Cryo-care: content-aware image restoration for cryo-transmission electron microscopy data," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 502–506.

[4] Tim-Oliver Buchholz, Alexander Krull, Réza Shahidi, Gaia Pigino, Gáspár Jékely, and Florian Jug, "Content-aware image restoration for electron microscopy," *Methods Cell Biol*, vol. 152, pp. 277–289, 2019.

[5] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug, "Noise2void-learning denoising from single noisy images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2129–2137.

[6] Alexander Krull, Tomas Vicar, and Florian Jug, "Probabilistic noise2void: Unsupervised content-aware denoising," *arXiv preprint arXiv:1906.00651*, 2019.

[7] Chinmay Belthangady and Loic A Royer, "Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction," *Nature methods*, p. 1, 2019.

[8] Liyuan Sui, Silvanus Alt, Martin Weigert, Natalie Dye, Suzanne Eaton, Florian Jug, Eugene W Myers, Frank Jülicher, Guillaume Salbreux, and Christian Dahmann, "Differential lateral and basal tension drive folding of drosophila wing discs through two distinct mechanisms," *Nature communications*, vol. 9, no. 1, pp. 4620, 2018.

[9] Romain F Laine, Kalina L Tosheva, Nils Gustafsson, Robert DM Gray, Pedro Almada, David Albrecht, Gabriel T Risa, Fredrik Hurtig, Ann-Christin Lindås, Buzz Baum, et al., "Nanoj: a high-performance open-source super-resolution microscopy toolbox," *Journal of Physics D: Applied Physics*, vol. 52, no. 16, pp. 163001, 2019.

[10] Wei Ouyang, Andrey Aristov, Mickaël Lelek, Xian Hao, and Christophe Zimmer, "Deep learning massively accelerates super-resolution localization microscopy," *Nature biotechnology*, vol. 36, no. 5, pp. 460, 2018.

[11] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila, "Noise2Noise: Learning image restoration without clean data," in *International Conference on Machine Learning*, 2018.

[12] Joshua Batson and Loic Royer, "Noise2self: Blind denoising by self-supervision," in *International Conference on Machine Learning*, 2019.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[14] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.