Direct Unsupervised Denoising

Benjamin Salmon and Alexander Krull
University of Birmingham
Birmingham B15 2TT, UK
brs209@student.bham.ac.uk, a.f.f.krull@bham.ac.uk

Abstract

Traditional supervised denoisers are trained using pairs of noisy input and clean target images. They learn to predict a central tendency of the posterior distribution over possible clean images. When, e.g.*, trained with the popular quadratic loss function, the network's output will correspond to the minimum mean square error (MMSE) estimate. Unsupervised denoisers based on Variational AutoEncoders (VAEs) have succeeded in achieving state-of-the-art results while requiring only unpaired noisy data as training input. In contrast to the traditional supervised approach, unsupervised denoisers do not directly produce a single prediction, such as the MMSE estimate, but allow us to draw samples from the posterior distribution of clean solutions corresponding to the noisy input. To approximate the MMSE estimate during inference, unsupervised methods have to create and draw a large number of samples – a computationally expensive process, rendering the approach inapplicable in many situations. Here, we present an alternative approach that trains a deterministic network alongside the VAE to directly predict a central tendency. Our method achieves results that surpass the results achieved by the unsupervised method at a fraction of the computational cost.*

1. Introduction

The prevalence of noise in biomedical imaging makes denoising a necessary step for many applications [14]. Deep learning has proven itself to be the most powerful tool for this task, as is evidenced by a growing body of research [27]. Although deep learning-based approaches typically require large amounts of training data, recent advances in unsupervised deep learning [20, 19, 25] have shown that this requirement need not be a barrier to their use. Unlike with supervised deep learning-based denoisers, which are trained with pairs of corresponding noisy and noise-free images, users of unsupervised methods can train their models with the very data they want to denoise.

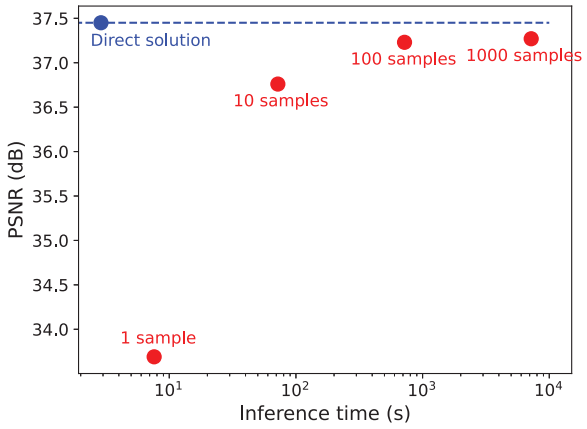

Figure 1. **Our Direct Denoiser outperforms unsupervised VAE-based denoising (HDN) [19], while requiring only a fraction of the computational cost:** In red, the time to draw 1, 10, 100 and 1000 samples from HDN's learned denoising distribution plotted against the PSNR (higher is better) of the per-pixel mean of these samples. Additionally, in blue, the time to take a single solution from our Direct Denoiser is plotted against its PSNR. These results are from denoising the *Convallaria* dataset.

The performance of unsupervised deep learning-based denoisers is now approaching and even sometimes matching the performance of their supervised counterparts [20, 19, 25], however, these two methods are fundamentally different in the way they do inference. By training a Variational AutoEncoder (VAE), unsupervised methods approximate a posterior distribution over the clean images that could underlie a noisy input image. This distribution will be referred to as the *denoising distribution*. Random samples from the denoising distribution then constitute the infinite possible solutions to a denoising problem. Supervised and self-supervised learning methods, on the other hand, offer a single prediction that compromises between all possible solutions. This is usually a central tendency of the denoising distribution and the specific central tendency that is predicted depends on the loss function used. For example,

a supervised method trained with the mean squared error (MSE) loss function will predict the mean, which is also known as the minimum mean squared error (MMSE) estimate. A model trained with the mean absolute error (MAE) loss function will predict the pixel-wise median, which is known as the minimum mean absolute error (MMAE) estimate.

While the ability of unsupervised methods to produce diverse solutions can in some circumstances be beneficial for downstream processing [20], users oftentimes require only a single solution such as the MMSE estimate. If they are to approximate this from an unsupervised learning-based denoiser, they must process their image many times and average many possible sampled solutions, leading to a significant computational overhead. For example, the authors of [20, 19, 25] average 100 or 1000 samples per image to obtain their MMSE estimate. Such an approach requires substantial computational effort, and is not likely to be economically and ecologically reasonable for labs regularly analyzing terabytes of data.

This paper presents an alternative route to estimating the central tendencies from an unsupervised denosier; one that requires noisy images to be processed only once. We do so by training an additional deterministic convolutional neural network (CNN), termed *Direct Denoiser*, that directly predicts MMSE or MMAE solutions and is trained alongside the VAE. It uses noisy training images as input and the sampled predictions from the VAE as training targets. Lacking a probabilistic nature, this network will minimize its MSE or MAE loss function by predicting the mean or pixel-wise median of the denoising distribution. The result is a denoising network with the evaluation times of a supervised approach and the training data requirements of an unsupervised approach.

In summary, we propose an extension to unsupervised deep learning-based denoisers that dramatically reduces inference time by estimating a central tendency of the learned denoising distribution in a single evaluation step. Moreover, we show these estimates to be more accurate than those obtained by averaging even up to 1000 samples from the denoising distribution. Figure 1 shows how much shorter inference time is with our proposed approach, and how much higher the quality of results are.

The remainder of the paper is structured as follows. In Section 2, we give a brief overview of related work, concentrating on different approaches to denoising. In Section 3, we provide a formal introduction to the unsupervised VAE-based denoising approach, which is the foundation of our method. In Section 4, we describe the training of the Direct Denoiser. We evaluate our approach in Section 5, showing that we consistently outperform our baseline at a fraction of the computational cost. Finally, in Sections 6 and 7 we discuss our results and give an outlook on the expected impact of our work and future perspectives.

2. Related Work

2.1. Supervised denosing

Traditional supervised deep learning-based methods (*e.g.* [30, 28]) rely on paired training data consisting of corresponding noisy and clean images. These methods view denoising as a regression problem, and usually train a UNet [23] or variants of the architecture to learn a mapping from noisy to clean. The most commonly used loss function for this purpose is the sum of pixel-wise quadratic errors (L2 or MSE), which directs the network to predict the MMSE estimate for the noisy input.

The approach's requirement for clean training images greatly limits its applicability, particularly for scientific imaging applications, where often no clean data can be obtained. In 2018, Lehtinen *et al.* [16] had the insight that training of equivalent quality can be achieved by replacing the clean training image with a second noisy image of the same content; a training method termed *Noise2Noise*. In practice, such image pairs can often be acquired by recording two images in quick succession. By using the L2 loss and assuming that the imaging noise is zero-centered, the network is expected to minimize the loss to its noisy training target by converging to the same MMSE estimate as in supervised training.

While Noise2Noise and traditional supervised methods are state-of-the-art with respect to the quality of their results, their requirement for paired training data makes them inapplicable in many situations. In contrast, our method requires only unpaired noisy data, which is available for any denoising task, making it directly applicable in situations where supervised methods are not.

2.2. Self-supervised denoising

Self-supervised methods have been introduced to enable denoising with unpaired noisy data. Here we focus on *blind-spot* approaches (*e.g.* [12, 2, 17, 22]), which mask individual pixels in the input image and use them as training targets. These methods rely on the assumption that imaging noise is pixel-wise independent given an underlying signal. By effectively forcing the network to predict each pixel value from its surroundings, blind-spot approaches can learn to denoise images without the need for paired noisy-clean data. Like supervised methods, self-supervised denoisers (when used with L2 loss) predict an MMSE estimate for each pixel, albeit based on less information, since the corresponding input pixel cannot be used during prediction. As a result, the quality of the output can be worse than supervised methods. The blind-spot approach has been improved to reintroduce the lost pixel information during inference [21, 15], achieving improved quality in some situa-

tions. In [4], Broaddus *et al*. extended the method to allow for the removal of structured noise.

Our method also does not require paired data, but we do not follow the self-supervised blind-spot paradigm. As a consequence, we do not have to address the loss of pixel information.

2.3. Unsupervised VAE-based denoising

Unsupervised VAE-based denoising methods [20] form the backbone of our method. Like in self-supervised methods, training requires only noisy images. However, their training and inference procedures differ greatly from self-supervised approaches. We discuss this class of methods in detail in Section 3.

2.4. Knowledge distillation

Knowledge distillation [9] is the process of training a smaller *student* network using a large *teacher* network or an ensemble [5] of teachers. The goal of this approach is to reduce the computational effort required during inference and enable more efficient employment of a powerful model. Surprisingly, the student model can achieve better results compared to being trained on the data directly. A survey of the topic can be found in [7].

The approach of training our Direct Denoiser with the output of another network can be seen as knowledge distillation. However, in our case the Direct Denoiser is not intended as a smaller replacement of the VAE, but as a model with a faster inference procedure.

3. Background

3.1. The denoising task

A noisy observation, $\mathbf{x}$, of a signal, $\mathbf{s}$, can be thought of as sampled from an observation likelihood, or *noise model*, $p_{\text{NM}}(\mathbf{x}|\mathbf{s})$. A noise model describes the random, unwanted variation that is added to a signal when it is recorded. The goal of denoising is to estimate the $\mathbf{s}$ that parameterized the noise model from which a known $\mathbf{x}$ was sampled.

3.2. Unsupervised denoising

It was Prakash *et al*. [20] who proposed doing so via variational inference, using a VAE [11] to approximate the posterior distribution $p(\mathbf{s}|\mathbf{x})$. They improved their approach with a more powerful architecture that could also handle mild forms of structured noise in [19]. Salmon and Krull then presented an alternative approach to tackling structured noise in [25], but it unfortunately cannot yet be applied in realistic settings.

To understand how unsupervised denoising works, we must give a brief explanation of the VAE [11]. For a full introduction, see [6].

Given a tractable prior distribution $p_\theta(\mathbf{z})$ and a likelihood $p_\theta(\mathbf{x}|\mathbf{z})$, the marginal distribution $p_\theta(\mathbf{x})$ could be learnt by minimizing the objective

$$-\log p_\theta(\mathbf{x}) = -\log \int_{\mathbf{z}} p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z}. \tag{1}$$

However, this integral is often intractable for high dimensional $\mathbf{x}$. VAEs instead approximate $p_\theta(\mathbf{x})$ by minimizing the following upper bound,

$$\begin{aligned} &-\log p_\theta(\mathbf{x}) + D_{KL}[q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[-\log p_\theta(\mathbf{x}|\mathbf{z}))] + D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})) \parallel p_\theta(\mathbf{z})], \end{aligned} \tag{2}$$

where θ and ϕ are learnable parameters and D_{KL} is the always positive Kullback-Leibler (KL) divergence [13]. Here, an approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ is introduced and optimized to diverge as little as possible from the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$.

The authors of DivNoising [20], Hierarchical DivNoising (HDN) [19] and AutoNoise [25] adapt the VAE for denoising by incorporating a known explicit noise model into this objective, directing the decoder of the VAE to map the latent variable $\mathbf{z}$ to estimates of the signal $\mathbf{s}$,

$$\begin{aligned} &-\log p_\theta(\mathbf{x}) + D_{KL}[q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[-\log p_{\text{NM}}(\mathbf{x}|\mathbf{s}))] + D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})) \parallel p_\theta(\mathbf{z})], \end{aligned} \tag{3}$$

where $\mathbf{s} = g_\theta(\mathbf{z})$.

3.3. Inference in unsupervised denoising

After minimizing this new denoising objective, the signal underlying a given $\mathbf{x}$ is estimated by first encoding $\mathbf{x}$ with $q_\phi(\mathbf{z}|\mathbf{x})$, sampling a $\mathbf{z}$ and mapping that sample to an estimate of the signal with $g_\theta(\mathbf{z})$. These solutions are samples from an approximation of the posterior $p(\mathbf{s}|\mathbf{x})$, which we refer to as the *denoising distribution*.

Each sample from the denoising distribution is unique, allowing users to examine the uncertainty involved in their denoising problem. However, a single consensus solution is often preferred. The authors of [20, 19, 25] chose to calculate the per pixel mean of 100 or 1000 samples, deriving the minimum mean square error (MMSE) estimate of the denoising distribution, to get a consensus solution for measuring denoising performance. Taking so many samples requires many forward passes of the denoiser and incurs a potentially prohibitive computational overhead for large datasets.

Our method extends the high quality denoising performance and minimal training requirements of VAE-based denoisers by allowing them to directly and efficiently produce MMAE and MMSE results without repeated sampling.

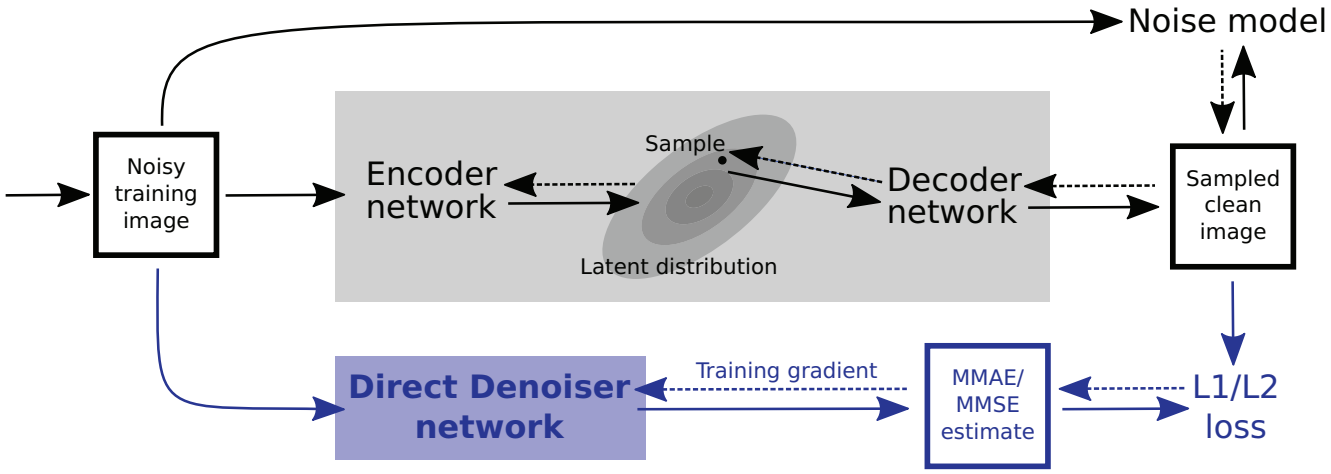

Figure 2. **Training scheme:** We train our novel *Direct Denoiser* (blue) along side a Variational AutoEncoder (VAE) [20, 19]. The processing of data is shown with solid arrows and the backward propagation of gradients required for training is shown with dashed arrows. The VAE encoder takes a noisy image as input and predicts the parameters of a distribution in latent space, a sample is drawn from here and mapped to a possible clean image by the decoder network. The reconstruction loss is computed using a pre-trained noise model. Our Direct Denoiser is trained using noisy images as input and the clean image samples (predicted by the VAE) as target. Since individual samples differ for the same input, there is no unique correct solution for this task. As a consequence, by using an L_2 loss, the Direct Denoiser will learn to predict the expected value, *i.e.*, the MMSE solution. Using an L_1 loss leads to predicting the pixel-wise median. We block gradients from passing through the sampled clean image to prevent the VAE changing its outputs.

4. Method

When given samples from a probability distribution, we are often interested in what a representative value of those samples is. In the case of unsupervised denoising, we are interested in a representative image from the denoising distribution. A common value to choose for this is the central tendency of the distribution [29], a point which minimizes some measure of deviation from all of the samples.

For samples from a learned denoising distribution, $p(\hat{\mathbf{s}}|\mathbf{x})$, over possible solutions $\hat{\mathbf{s}}$ for a noisy input image $\mathbf{x}$, this would be

$$\hat{\mathbf{s}}^* = \arg\min_{\mathbf{y}} \mathbb{E}_{\hat{\mathbf{s}}|\mathbf{x}}[L(\mathbf{y}, \hat{\mathbf{s}})], \tag{4}$$

where L is some per-pixel loss function. If L is the L_1 loss,

$$L(\mathbf{y}, \hat{\mathbf{s}}) = 1/n \sum_{i}^{n} |y_i - \hat{s}_i|, \tag{5}$$

then $\hat{\mathbf{s}}^*$ corresponds to the pixel-wise median of the distribution, *i.e.*, the MMAE estimate. Here, n denotes the number of pixels and y_i and $\hat{s}_i$ denote i^{th} pixel values. For the L_2 loss,

$$L(\mathbf{y}, \hat{\mathbf{s}}) = 1/n \sum_{i}^{n} (y_i - \hat{s}_i)^2, \tag{6}$$

$\hat{\mathbf{s}}^*$ will be the arithmetic mean, *i.e.*, the MMSE.

The authors of [20, 19, 25] estimated $\hat{\mathbf{s}}^*$ using a large number of samples from their denoising distribution. We propose instead training a CNN to directly predict a central tendency.

Let h_η be our Direct Denoiser with parameters η and $p(\hat{\mathbf{s}}|\mathbf{x})$ be a denoising distribution. The following objective,

$$\arg\min_{\eta} \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{\hat{\mathbf{s}}|\mathbf{x}}[L(h_\eta(\mathbf{x}), \hat{\mathbf{s}})]], \tag{7}$$

where L is either the L_1 or L_2 loss, would train h_η to predict either the pixel-wise median or mean of $p(\hat{\mathbf{s}}|\mathbf{x})$, respectively. After training an unsupervised denoiser according to [20, 19, 25], we could train our Direct Denoiser with Eq. 7 by sampling noisy images $\mathbf{x}$ from a training set and then running them through the unsupervised denoiser to obtain possible clean solutions $\hat{\mathbf{s}}$ from the denoising distribution.

We however find that it is possible to train both models simultaneously. Let $f_{\theta,\phi}$ represent a VAE with the loss function in Equation 3, where $\hat{\mathbf{s}} \sim f_{\theta,\phi}(\mathbf{x})$ is a sample from the denoising distribution.

A single training step for simultaneously optimizing an unsupervised denoiser and an accompanying Direct Denoiser is as follows:

1. Pass a noisy training image $\mathbf{x}$ to the unsupervised denoiser and sample a possible solution $\hat{\mathbf{s}}$.

2. Update the parameters (θ, ϕ) towards minimizing the loss function in Equation 3.

3. Pass the same $\mathbf{x}$ to the Direct Denoiser, calculating $h_\eta(\mathbf{x})$.

4. Update the parameters η to minimize $L(h_\eta(\mathbf{x}), \hat{\mathbf{s}})$, where L is the L_1 or L_2 loss function.

5. Repeat until convergence.

A visual representation of this training scheme can be found in Figure 2.

5. Experiments

Our Direct Denoiser was trained alongside HDN [19], using six datasets of intrinsically noisy microscopy images that come with known ground truth signal. Each dataset can be found in [19], as can details of their size, spatial resolution and train, validation and test splits. Note that for the *Struct. Convallaria* dataset, we adapted HDN into $\text{HDN}_{3\text{-}6}$, making it capable of handling structured noise.

Denoising Performance To evaluate denoising performance, we compare the Peak Signal-to-Noise Ratio (PSNR) of our Direct Denoiser's direct solutions to the PSNR of HDN's consensus solutions. The consensus solutions were produced by averaging samples of size 1, 10, 100 and 1000, reporting both their per-pixel median and mean. The Direct Denoiser's solutions were reported from a network trained with an L_1 loss and a network trained with an L_2 loss. Results are in Table 1. Visual results from the same experiment can be seen in Figure 3.

Inference Times We also compared inference time to denoising performance. Specifically, the total time for HDN to generate 1, 10, 100 and 1000 samples for all 100 images in the *Convallaria* test set was measured, then plotted against the PSNR of the mean of those samples, averaged over all 100 images. On the same plot, the total time for our Direct Denoiser to produce single solutions for each image is plotted against their average PSNR. Each test image consisted of 512×512 pixels.

Using our GPU (an NVIDIA GeForce RTX 3090 Ti), generating a single 512×512 solution from HDN's denoising distribution takes 0.076 seconds, using 2207MiB of the GPU's memory. Our Direct Denoiser takes 0.029 seconds at 1909MiB to do the same. Processing one image with either model uses the full capacity of the GPU's parallelism, so we saw no speed improvements by processing more than one image at a time.

If a consensus solution from HDN with PSNR approaching that of the the Direct Denoiser requires sampling 1000 solutions, inference with the proposed method is $2621\times$ faster.

Training Times and Memory Usage Finally, the additional training time incurred by co-training HDN with the Direct Denoiser was examined. The authors of HDN [19] train their network for 200,000 steps for all datasets, using a batch size of 64 and image patch size of 64×64. Using our GPU, training HDN alone takes 0.27 seconds per step for 15 hours total, using 13GiB of GPU memory. Training both HDN and the Direct Denoiser takes 0.34 seconds per step for 18.9 hours total, using 15GiB of GPU memory. Note that smaller virtual batches can be used as in [19] to reduce memory consumption. For the proposed method to be a net time saving, inference would have to take 3.9 hours less. Using our hardware and inference image resolution, time is saved when the inference test set consists of 185 images with 512×512 resolution.

Network Architecture and Training The Direct Denoiser used in these experiments was a UNet [23] with approximately 12 million parameters, while the unsupervised denoiser was the same Hierarchical VAE [26] used in [19] with approximately 7 million parameters. We chose to give our UNet more parameters than the Hierarchical VAE to ensure the former had the capacity to learn the full relationship between noisy images and solutions generated by the latter. This may not have been necessary, and training a Direct Denoiser with a lower computational demand would be an interesting topic for future research.

Our UNet had a depth of four, with a residual block [8] consisting of two convolutions followed by a ReLU activation function [1] at each level. Downsampling was performed by convolutions with a stride of two, and upsampling by nearest neighbor interpolation [24] followed by a single convolution with stride one. All convolutions had a kernel size of 3. The number of filters was 32 at the first level and that number doubled at each subsequent level. Skip connections were merged by concatenating the skipped features with the features from the previously level and passing the two through a residual block.

Training followed the same procedure described in [19], with the only difference being that our Direct Denoiser had its own Adamax optimizer [10] with an initial learning rate of 3e-4 that reduced by a factor of 0.5 when validation loss had plateaued for 10 epochs.

6. Discussion

Solutions from our Direct Denoiser consistently scored a higher PSNR than consensus solutions of 1000 samples from HDN. Table 1 shows HDN's PSNRs converging towards our direct prediction result with increased sample size. It seems that solutions from our Direct Denoiser are sometimes equivalent to averaging sample sizes orders of magnitude larger than the largest samples size we used in our experiment. Moreover, by looking at the inference times reported in Figure 1, the time required to take such a sample size would be impractical for large datasets.

7. Conclusions

We have demonstrated that an extension of the unsupervised denoising approach–the Direct Denoiser–can be used to dramatically speed up inference time, while at the same time improving performance when compared the standard inference procedure with up to 1000 sampled images.

Dataset	Number of samples (HDN) 1	10	100	1000	Direct
Convallaria	33.69 / 33.69	36.59 / 36.76	37.17 / 37.23	37.19 / 37.27	**37.50** / 37.45
Confocal Mice	35.43 / 35.43	37.30 / 37.42	37.58 / 37.68	37.62 / 37.69	**37.77** / 37.75
2 Photon Mice	31.21 / 31.21	32.63 / 32.68	32.86 / 32.87	32.89 / 32.89	**33.55** / 33.54
Mouse Actin	31.62 / 31.62	33.52 / 33.66	33.87 / 33.92	33.91 / 33.95	34.22 / **34.28**
Mouse Nuclei	33.48 / 33.48	36.24 / 36.44	36.79 / 36.89	36.81 / 36.90	36.87 / **36.93**
Struct. Convallaria	29.02 / 29.02	30.88 / 31.00	31.22 / 31.27	31.27 / 31.29	31.58 / **31.64**

Table 1. **PSNR of consensus solutions from HDN [19] compared to direct solutions from our novel *Direct Denoiser*.** HDN's consensus solutions were obtained by taking samples of varying sizes from its denoising distribution and calculating both their per-pixel median and their per-pixel mean. The Direct Denoiser's solutions were obtained from a single pass of a network trained under an L_1 loss and a single pass of a network trained under an L_2 loss. PSNRs are reported as an average over all images in each test set, and are presented as the median/mean consensus for HDN and as the solution from the L_1/L_2 network for the Direct Denoiser. Best results are printed in bold.

We believe our approach will become the default way of producing central tendencies from unsupervised denoising models with the increase in speed potentially allowing an easy adaptation by the community.

While we have evaluated our method only for MSE and MAE loss functions, we believe the approach could also be used with other loss functions such as *Tukey's biweight loss* [3], which might allow us to find regions of high probability density or even the *maximum a posteriori* estimate.

Recent work in image restoration has suggested the use of more sophisticated perceptual loss functions (see *e.g.* [18]). These types of loss functions would likely only be possible in a supervised setting with clean training data and would be unlikely to succeed with Noise2Noise or self-supervised methods. However, since the training targets sampled by our VAE are essentially clean images, they should be compatible with different types of complex loss functions, opening the door to using perceptual loss with noisy unpaired data.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 5

[2] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision, 2019. 2

[3] Vasileios Belagiannis, Christian Rupprecht, Gustavo Carneiro, and Nassir Navab. Robust optimization for deep regression. In *Proceedings of the IEEE international conference on computer vision*, pages 2830–2838, 2015. 6

[4] C. Broaddus, A. Krull, M. Weigert, U. Schmidt, and G. Myers. Removing structured noise with self-supervised blind-spot networks. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 159–163, 2020. 3

[5] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000. 3

[6] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 3

[7] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015. 3

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 3rd international conference on learning representations, iclr 2015. *arXiv preprint arXiv:1412.6980*, 9, 2015. 5

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014. 3

[12] 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*, pages 2129–2137, 2019. 2

[13] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 3

[14] Romain F Laine, Guillaume Jacquemet, and Alexander Krull. Imaging in focus: an introduction to denoising bioimages in the era of deep learning. *The International Journal of Biochemistry & Cell Biology*, 140:106077, 2021. 1

[15] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. In *Advances in Neural Information Processing Systems*, pages 6968–6978, 2019. 2

[16] 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*, pages 2965–2974, 2018. 2

[17] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12064–12072, 2020. 2

[18] Aamir Mustafa, Aliaksei Mikhailiuk, Dan Andrei Iliescu, Varun Babbar, and Rafał K Mantiuk. Training a task-specific

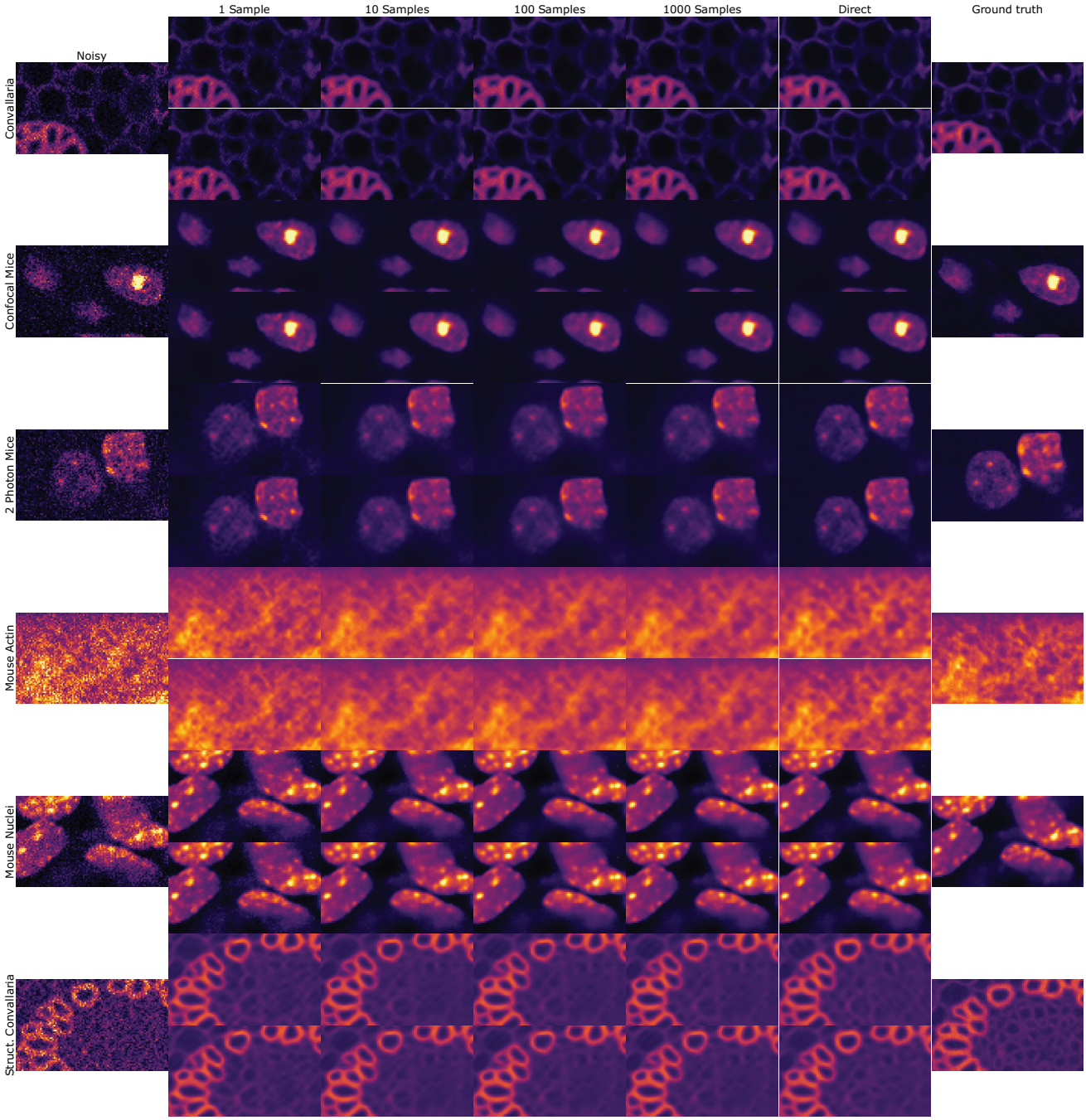

Figure 3. **Visual results:** Cropped images from each dataset showing consensus solutions of varying sample sizes from HDN's denoising distribution with direct solutions from our Direct Denoiser. For each dataset, the top row shows the median of HDN samples and a solution from our L_1 trained Direct Denoiser, while the bottom row shows the mean of HDN samples and a solution from our L_2 trained Direct Denoiser.

image reconstruction loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2319–2328, 2022. 6

[19] Mangal Prakash, Mauricio Delbracio, Peyman Milanfar, and Florian Jug. Interpretable unsupervised diversity denoising and artefact removal. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 4, 5, 6

[20] Mangal Prakash, Alexander Krull, and Florian Jug. Fully unsupervised diversity denoising with convolutional variational autoencoders. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 4

[21] Mangal Prakash, Manan Lalit, Pavel Tomancak, Alexander

Krull, and Florian Jug. Fully unsupervised probabilistic noise2void. *arXiv preprint arXiv:1911.12291*, 2019. 2

[22] Yuhui Quan, Mingqin Chen, Tongyao Pang, and Hui Ji. Self2self with dropout: Learning self-supervised denoising from single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1890–1898, 2020. 2

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 5

[24] Olivier Rukundo and Hanqiang Cao. Nearest neighbor value interpolation. *arXiv preprint arXiv:1211.1768*, 2012. 5

[25] Benjamin Salmon and Alexander Krull. Towards structured noise models for unsupervised denoising. In *European Conference on Computer Vision*, pages 379–394. Springer, 2022. 1, 2, 3, 4

[26] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016. 5

[27] Jingwen Su, Boyan Xu, and Hujun Yin. A survey of deep learning approaches to image restoration. *Neurocomputing*, 487:46–65, 2022. 1

[28] 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*, 15(12):1090–1097, 2018. 2

[29] Herbert Weisberg. *Central tendency and variability*. Number 83. Sage, 1992. 4

[30] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 2