N2V2 - Fixing Noise2Void Checkerboard Artifacts with Modified Sampling Strategies and a Tweaked Network Architecture

Anonymous ECCV WS-BIC submission

Paper ID 11

Abstract. In recent years, neural network based image denoising approaches have revolutionized the analysis of biomedical microscopy data. Self-supervised methods, such as Noise2Void (N2V), are applicable to virtually all noisy datasets, even without dedicated training data being available. Arguably, this facilitated the fast and widespread adoption of N2V throughout the life sciences. Unfortunately, we observed that the blind-spot training underlying N2V can lead to rather visible checkerboard artifacts, thereby reducing the quality of final predictions considerably. In this work, we present two modifications to the vanilla N2V setup that both help to reduce the unwanted artifacts considerably. Firstly, we propose a modified network architecture *i.e.*, using *BlurPool* instead of *MaxPool* layers throughout the used U-Net, rolling back the residual-U-Net to a non-residual U-Net, and eliminating the skip connections at the uppermost U-Net level. Additionally, we propose new replacement strategies to determine the pixel intensity values that fill in the elected blind-spot pixels. We validate our modifications on a range of microscopy and natural image data. Based on added synthetic noise from multiple noise types and at varying amplitudes, we show that both proposed modifications push the current state-of-the-art for fully self-supervised image denoising.

1 Introduction

Fluorescence microscopy is one of the major drivers for discovery in the life sciences. The quality of possible observations is limited by the optics of the used microscope, the chemistry of used fluorophores, and the maximum light exposure tolerated by the imaged sample. This necessitates trade-offs, frequently leading to rather noisy acquisitions as a consequence of preventing ubiquitous effects such as photo toxicity and/or bleaching. While the light efficiency in fluorescence microscopy can be optimized by specialized hardware, *e.g.*, by using Light Sheet or Lattice Light Sheet microscopes, software solutions that restore noisy or distorted images are a popular additional way to free up some of the limiting photon budget.

Algorithmic image restoration is the reconstruction of clean images from corrupted versions as they were acquired by various optical systems. A plethora of recent work shows that CNNs can be used to build powerful content-aware

Input N2V [7] **N2V2 (Ours)** GT

Fig. 1: Self-supervised denoising of noisy data (left). Results obtained with N2V [7] (here shown without residual connection and with sampling without the center point) are subject to clearly visible checkerboard artifacts (2nd column). Our proposed method, N2V2, visibly reduces these artifacts, leading to improved quality results (3rd column, here shown with median center pixel replacement). The last column shows ground truth (not available to either method)

image restoration (CARE) methods [18,19,20,21,10,3]. However, when using supervised CARE approaches, as initially proposed in [19], pairs of clean and distorted images are required for training the method. For many applications in the life-sciences, imaging such clean ground truth data is either impossible or comes at great extra cost, often rendering supervised approaches as being practically infeasible [7].

Hence, self-supervised training methods like Noise2Void (N2V) by Krull *et al.* [7], which operate exclusively on single noisy images, are frequently used in life-science research [1,7,9,8,15]. Such *blind-spot approaches* are enabled by excluding/masking the center (blind-spot) of a network's receptive field and then training the network to predict the masked intensity. These approaches collectively assume that the noise to be removed is pixel-wise independent (given the signal) and that the true intensity of a pixel can be predicted after learning a content-aware prior of local image structures from a body of noisy data [7].

More recently, methods that can sample the space of diverse interpretations of noisy data were introduced [14,13]. While these approaches show great performance on denoising and even artifact removal tasks, the underlying network architectures and training procedures are space and time demanding [13] and can typically not be used on today's typical consumer workstations and laptops. Hence, comparatively small blind-spot networks like N2V are available via consumer solutions such as *Fiji* [17,4], *ZeroCostDL4Mic* [6], or the *BioImage.IO Model Zoo* [12], and are therefore still the most commonly used self-supervised denoising methods.

Still, one decisive problem with blind-spot approaches such as Noise2Void is that checkerboard artifacts can commonly be observed (see Figure 1 for an illustrative example). Hence, in this work we present Noise2Void v2 (N2V2), a variation of N2V that addresses the problem with checkerboard artifacts by a series of small but decisive tweaks.

More concretely, the contributions of our work are: (*i*) showcasing and inspecting the short-comings of Noise2Void, (*ii*) introduction of blind-spot pixel

replacement strategies, (*iii*) proposal of an adapted U-Net architecture that replaces *max-pooling* layers with *max-blur-pooling* layers and omits the top-most skip-connection. (*iv*) a systematic evaluation of our proposed replacement strategies and architectural changes on the BSD68 dataset [7], the Mouse, Convallaria and Flywing datasets from [14,15] and two newly added salt and pepper (S&P) noise regimes. (*v*) proposal of a new variation on the Convallaria dataset from [15] that addresses what we believe to be non-ideal setup choices.

2 Related Work

The original CARE work by Weigert *et al.* [19] steered our field away from more established and non-trained denoising methods towards modern data-driven deep denoising methods. When supervised, such methods train neural networks on pairs of low-quality and high-quality images that are pixel-perfectly aligned and contain the exact same objects (or 'scene').

Such pairs need to be carefully acquired at the microscope, typically by varying parameters such as exposure time and illumination intensity. In certain modalities, *e.g.*, cryo-TEM, acquisition of high-exposure images is impossible and even the acquisition of pairs of noisy images is undesirable [3].

However, if such pairs of independently noisy images are available, Noise2Noise training [10] can be applied and high quality predictions are still achievable. Later, Buchholz *et al.* [2], extended these ideas to full cryo-ET workflows [11].

Still, clean ground truth data or a second set of independently noisy images is typically not readily available. This motivated the introduction of self-supervised methods such as Noise2Void [7] and Noise2Self [1]. The simplicity and applicability of these methods makes them, to-date, the de-facto standard approach used by many microscopists on a plethora of imaging modalities and biological samples. All such blind-spot approaches exploit the fact that for noise which is independent per pixel (given the signal), the intensity value of any given pixel can in principle be estimated from examining the pixels image context (surrounding). This is precisely what content-aware image restoration approaches do. Pixel-independent noise, instead, can by definition not be predicted, leading to a situation where the loss minimizing prediction does, in expectation, coincide with the unknown signal at the predicted pixel [7,1,10].

An interesting extension of N2V was introduced by Krull *et al.* [8]. Their method, called *Probabilistic Noise2Void* (PN2V), does not only predict a single (maximum likelihood) intensity value per pixel, but instead an entire distribution of plausible pixel intensity values (prior). Paired with an empirical (measured) noise-model [8,15], *i.e.*, the distributions of noisy observations for any given true signal intensity (likelihood), PN2V computes a the posterior distribution of possible predicted pixel intensities and returns, for example, the MMSE of that posterior.

A slightly different approach to unsupervised image denoising was proposed by Prakash*et al.* [14,13]. Their method is called (Hierarchical) DivNoising and employs a variational auto-encoder (VAE), suitably paired with a noise model

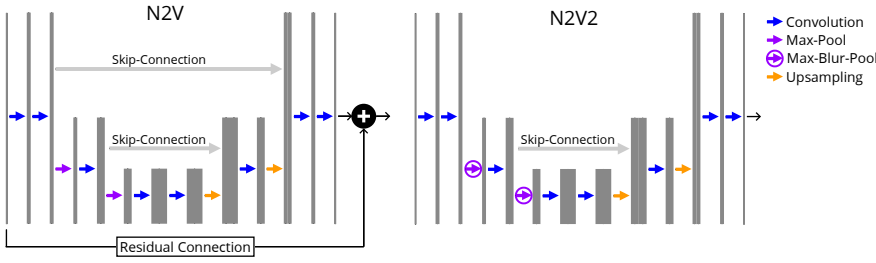

Fig. 2: *Left:* the N2V architecture in [7] is a standard U-Net [16] with a residual connection. *Right:* our N2V2 architecture drops the residual connection, removes the top-most skip-connection and replaces the max-pooling layers with max-blur-pooling layers [21]

of the form described above [8,15], that can be used to sample diverse interpretations of the noisy input data. Note that this is quite different from being able to predict or sample different intensities per pixel. DivNoising is sampling different multi-pixel structures that are (*i*) in-line with the noisy observation, and (*ii*) consistent with the learned content-aware structural prior given the full body of noisy training data.

2.1 Particularities of the Publicly Available Convallaria Dataset

Self-supervised denoising methods are built to operate on data for which no high-quality ground truth exists. This makes them notoriously difficult to evaluate quantitatively, unless when applied on data for which ground truth is obtainable.

To enable a fair comparison between existing and newly proposed methods, several benchmark datasets have been made available over the years. One example is the *Convallaria* data, first introduced by Lalit *et al.* [15]. This dataset consists of 100 noisy short exposure fluorescence acquisitions of the same 1024×1024px field of view of the same sample. The corresponding ground truth image used to compare against was created by pixel-wise averaging of these 100 independently noisy observations.

In later work [15,14], the proposed methods were trained on 95 of the individual noisy images, while the remaining 5 images have been used for validation purposes. For the PSNR values finally reported in these papers, the predictions of the top left 512×512 pixels of all 100 noisy are compared to the corresponding part of the averaged ground truth image. In this paper we refer to this dataset and associated train/validation/test sets as *Convallaria_95*.

We are convinced that training self-supervised image denoising methods on 95 noisy observations of the exact same field of view is leading to slightly misleading results (that overestimate the performance to be expected form the tested method in cases where only one noisy image per sample exists. Also note that in cases where already as few as two noisy observations per sample are available,

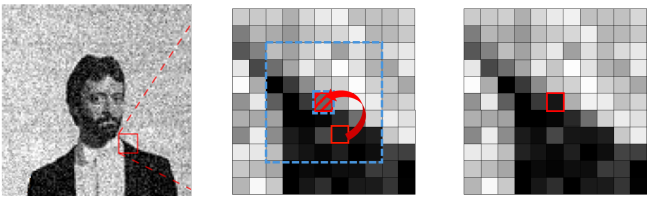

Fig. 3: The original Noise2Void replacement strategy in [7] chooses a random pixel from the center pixel's local neighborhood, which may lead to artifacts like checkerboard patterns in denoised images (see Fig. 5)

a network can be trained via Noise2Noise [10]. With 95 such instances available, one could even average those and use the average as ground truth for fully supervised training CARE training [19].

Hence, we propose here to use the Convallaria data differently, namely by selecting one of the 100 images and tiling it into 64 tiles of 128 × 128px. Of these tiles, 56, 4, and 4 are then used for training, validation and testing respectively. See the supplementary material for more information. We refer to this data and train/validation/test split as *Convallaria_1* Please see Section 4 for a thorough evaluation of achievable denoising results when using *Convallaria_95* versus *Convallaria_1*.

3 Method

As can be seen in Fig. 1, denoising predictions from a vanilla N2V model can exhibit considerable amounts of unwanted checkerboard artifacts. After investigating this phenomenon on several datasets, our hypothesis is that these artifacts are caused by two effects in the vanilla N2V design: (*i*) missing high-frequency suppression techniques to counteract strongly noisy pixel values that really stick out with respect to their close neighbors, and (*ii*) an amplification of this effect due to N2V's self-supervised input replacement scheme (blind-spots). Below we describe to measures we introduce in N2V2 to mitigate these problems.

3.1 A Modified Network Architecture for N2V2

The default Noise2Void configuration employs a *residual* U-Net with 2×2 max-pooling layers throughout the encoder [7]. We propose to change this architecture in three decisive ways by (*i*) removing the residual connection and instead use a regular U-Net, (*ii*) removing the top-most skip-connection of the U-Net to further constrain the amount of high-frequency information available for the final decoder layers, and (*iii*) replacing the standard max-pooling layers by max-blurpool layers [21] to avoid unwanted shifting of structures. In Figure 2 we highlight all proposed architectural changes we propose for N2V2.

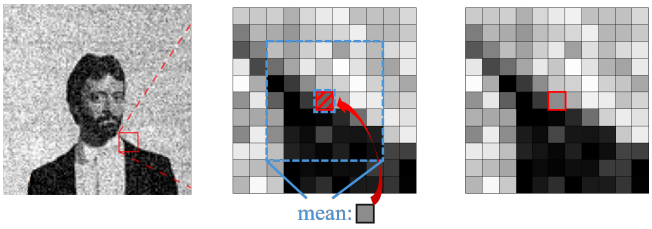

Fig. 4: The average center pixel replacement strategy calculates the mean or median of the pixel's local neighborhood while excluding the center pixel itself

3.2 New Sampling Strategies to Cover Blind-Spots

As mentioned before, self-supervised denoising methods introduce blind-spots, effectively asking the network to perform content-aware single pixel inpainting [7,1,8,15].

During training, a self-supervised loss is employed that compares measured (and left out) pixel values with the corresponding pixel values predicted by the trained network (to which only the local neighborhood of the respective blind-spot pixels is given).

Let $\mathbf{X} \in \mathfrak{N}^{w \times h}$ be a patch in a given input image with intensity range $\mathfrak{N}$. W.l.o.g., let x_i be a single pixel in $\mathbf{X}$. As loss for a given patch $\mathbf{X}$, Noise2Void [7] starts with proposing

$$\mathfrak{L}_{\text{N2V-naïve}}\big(\mathbf{X}, \mathrm{f}\left(\cdot\right)\big) = \big(\mathrm{f}\left(\mathbf{X} \setminus \{x_i\}\right)_i - x_i\big)^2 \quad , \tag{1}$$

where $\mathbf{X} \setminus \{x_i\}$ denotes the exclusion of pixel x_i from $\mathbf{X}$. This exclusion operation would be computational inefficient when being implemented naïvely in convolutional networks. Krull *et al.* have therefore proposed not to *exclude* x_i, but rather to *replace* x_i's value and thereby hiding the true intensity of blind-spot pixels:

$$\mathfrak{L}_{\text{N2V}}\big(\mathbf{X}, \mathrm{f}\left(\cdot\right)\big) = \big(\mathrm{f}\left(\mathrm{r}\left(\mathbf{X}\right)\right)_i - x_i\big)^2 \quad , \tag{2}$$

where $\mathrm{r}\left(\mathbf{X}\right)$ assigns a new value to x_i in $\mathbf{X}$. While Eq. (2) can be evaluated efficiently compared to Eq. (1), it turns out that the choice of $\mathrm{r}\left(\mathbf{X}\right)$ is more sensitive than originally believed, with some choices leading to emphasized visual artifacts like the ones shown in Figure 1.

Default N2V Pixel Sampling Strategies (uwCP and uwoCP). In [7], Krull *et al.* analyze different blind-spot pixel replacement methods and settle for one default method in their public implementation[1]. This default method, called UPS, is ubiquitously used by virtually all users world-wide and samples a pixel x_j uniformly at random from a small neighborhood $\mathbf{N} \subset \mathbf{X}$ of size $w' \times h'$

[1] https://github.com/juglab/n2v

around a blind-spot pixel x_i (including x_i itself). We refer to this replacement technique as uwCP, and illustrate it in Figure 3.

The first obvious observation is that with probability of $1/(w' \cdot h')$, j will equal i, *i.e.*, no replacement is happening. In these cases, the best solution to any model $\mathrm{f}(\cdot)$ will be the identity, which is clearly not intended for denoising tasks. Therefore, in Probabilistic Noise2Void (PN2V) [8], the available implementation[2] started using a slightly altered sampling strategy that excludes the center pixel from being sampled, *i.e.*, $i \neq j$, which we refer to as uwoCP.

Blind-Spot Replacement Strategies for N2V2. In contrast to the blind-spot replacement strategies via sampling from $\mathbf{N}$, we propose to compute replacement strategies computed from the entire pixel neighborhood $\mathbf{N}$. Specifically, we propose $\mathrm{r}_{\mathrm{mean}}(\mathbf{N}) = \mathrm{mean}(\mathbf{N} \setminus \{x_i\})$ and $\mathrm{r}_{\mathrm{median}}(\mathbf{N}) = \mathrm{median}(\mathbf{N} \setminus \{x_i\})$ as replacement strategies, and refer to them as *mean* and *median* replacement strategies, respectively.

Note that the exclusion of the center pixel is important in order to fully remove any residual information about the blind-spot pixels to be masked. Please refer to Figure 4 for a visual illustration.

4 Evaluation

We evaluate our proposed pixel replacement strategies and the architectural changes on multiple datasets and perform different ablation studies. The covered datasets with their experiment details are described in Section 4.1. Evaluation metrics are listed in Section 4.2. Results on data with S&P noise are given in Section 4.3. Complementary results with other noise types are given in Section 4.4. In Section 4.5, we finally shed light on aspects on generalization and evaluation in scenarios where only single noisy recordings are available.

4.1 Dataset Descriptions and Training Details

All dataset simulation and method evaluation code, together with the used training configurations, is publicly available on GitHub[3].

General Settings In all network trainings, we use an Adam optimizer with a learning rate scheduler which halves the learning rate after hitting a plateau with a patience of 10. In each patch, 0.198% of pixels are randomly chosen as blind-spots and subject to pixel replacement with a neighborhood of size $w' = h' = 5$.

[2] https://github.com/juglab/pn2v

[3] https://github.com/****censored-for-review**********

	Method	Mouse SP3	Mouse SP6	Mouse SP12
	Input	20.03	18.72	17.76
Fully self-supervised	N2V as in [7]	21.32	20.69	20.99
	N2V w/ uwoCP as in [8]	35.17	34.24	33.49
	N2V w/o res, w/ uwoCP	35.44	34.89	34.12
	N2V w/o res w/ mean	35.29	34.71	33.66
	N2V w/o res w/ median	35.23	35.07	33.45
	N2V2 w/ uwCP	35.74	35.32	34.19
	N2V2 w/ uwoCP	**35.91**	35.47	34.52
	N2V2 w/ mean	35.51	35.01	34.17
	N2V2 w/ median	35.81	**35.50**	**34.54**
Self-supervised	PN2V [8]	29.67	N/A	N/A
	DivNoising [14]	36.21	N/A	N/A
Supervised	CARE [19]	<u>37.03</u>	N/A	N/A

Table 1: Quantitative results on data with simulated salt and pepper noise. Results are given in dB of averaged PSNR on test data. Overall best is <u>underlined</u>. Best fully self-supervised is in **bold**

BSD68 An evaluation on natural images is done with the BSD68 dataset as used in the original Noise2Void paper [7]. For training, we use the same 400 natural gray scale images of size 180 × 180px from [20]. From those, 396 are used as training data and 4 for validation as described in Noise2Void. BSD68 networks are of depth 2 with 96 initial feature maps and are trained for 200 epochs, with 400 steps per epoch, a batch size of 128, and an initial learning rate of 0.0004.

Convallaria We evaluate on the fluorescence imaging dataset Convallaria by [15]. Due to its specialities as described in Section 2.1, we call it Convallaria_95. Additionally, we introduce the Convallaria_1 dataset where the input corresponds to only one single noisy observation of 1024 × 1024px and the corresponding ground truth is the average of the 100 noisy Convallaria observations. This image pair is divided into non-overlapping patches of 128 × 128px, resulting in 64 patches. These patches are shuffled and 56, 4, and 4 patches are selected as training, validation and test data respectively (see Supplementary Figure S3). We train Convallaria_95 and Convallaria_1 networks with depth 3, with 64 initial feature maps, and for 200 epochs, with 10 steps per epoch, a batch size of 80, and an initial learning rate of 0.001.

Mouse We further conduct evaluations based on the ground truth Mouse dataset from the DenoiSeg paper [5], showing cell nuclei in the developing mouse skull. The dataset consists of 908 training and 160 validation images of size 128 × 128px, with another 67 test images of size 256 × 256px. From this data, we

	Method	Flywing G70	Mouse G20	BSD68
	Input	17.67	22.52	21.32
Fully self-supervised	N2V as in [7]	25.20	34.12	27.70
	N2V w/ uwoCP as in [8]	25.04	33.94	27.37
	N2V w/o res, w/ uwoCP	25.24	34.20	26.95
	N2V w/o res w/ mean	25.54	34.49	28.25
	N2V w/o res w/ median	**25.57**	34.41	27.49
	N2V w/ bp w/ uwCP	25.30	34.17	27.69
	N2V w/o sk w/ uwCP	25.49	34.63	27.88
	N2V2 w/ uwCP	25.42	34.65	28.04
	N2V2 w/ uwoCP	25.49	34.59	27.97
	N2V2 w/ mean	25.48	34.61	28.31
	N2V2 w/ median	25.46	**34.74**	**28.32**
Self-supervised	PN2V [8]	24.85	34.19	N/A
	DivNoising [14]	25.02	34.13	N/A
Supervised	CARE [19]	<u>25.79</u>	<u>35.11</u>	<u>29.06</u>

Table 2: Quantitative results: simulated Gaussian noise. Results are given in dB of averaged PSNR on test data. Overall best is <u>underlined</u>. Best fully self-supervised is in **bold**

simulate Mouse_G20 by adding Gaussian noise with zero-mean and standard deviation of 20. Furthermore, we simulate Mouse_sp3, Mouse_sp6 and Mouse_sp12, three datasets dominated by Salt&Pepper noise. More specifically, we apply Poisson noise directly to the ground truth intensities, then add Gaussian noise with zero-mean and standard deviation of 10, and clip these noisy observations to the range $[0, 255]$. Then, we randomly select $p\%$ of all pixels ($p \in [3, 6, 12]$) and set them to either 0 or 255 with a probability of 0.5. We train networks on the Mouse dataset with depth 3, with 64 initial feature maps, and for 200 epochs, with 90 steps per epoch, a batch size of 80 and an initial learning rate of 0.001.

Flywing Finally, we report results on the Flywing dataset from the DenoiSeg [5], showing membrane labeled cells in a flywing. We follow the data generation protocol described in [14], *i.e.*, we add zero-mean Gaussian noise with a standard deviation of 70 to the clean recordings of the dataset. The data consists of 1 428 training and 252 validation patches of size 128×128px, with additional 42 images of size 512×512px for testing. On the flywing dataset, we train networks with of depth 3, with 64 initial feature maps, and for 200 epochs, with 142 steps per epoch, a batch size of 80 and an initial learning rate of 0.001.

Data Augmentation All training data is 8-fold augmented by applying three 90 deg rotations and flipping. During training, random 64×64 crops are selected from the provided training patches as described in [7].

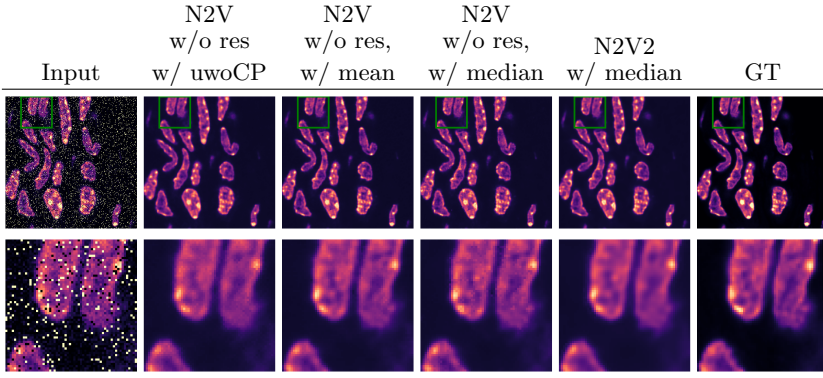

Fig. 5: Qualitative results on Mouse SP12 dataset. This data is dominated by 12% S&P noise, as can be seen in the input image (1st). The results of the N2V method without residual connection and sampling without the center point (2nd, PSNR 33.10) show checkerboard artifacts. These remain when using median replacement (4th, PSNR 32.43), are reduced in the results when using mean replacement (3rd, PSNR 33.01) and eliminated in the N2V2 results with median replacement (5th, PSNR 33.34)

4.2 Evaluation Metrics

We compute peak signal-to-noise ratio (PSNR) in all conducted experiments, evaluated with respect to the corresponding high-SNR images. For the BSD68 dataset, the target range of the PSNR computation is set to $[0, 255]$. For all other datasets, the range is obtained by computing the min and max values of each corresponding ground truth image. We finally report PSNR values averaged over the entire test data.

4.3 Results on Mouse SP3, SP6, and SP12 (Salt&Pepper noise)

The results for the salt and pepper datasets are shown in Table 1. First of all, we see the striking impact of excluding the center pixel from the replacement sampling for S&P noise: while N2V as in [7] can barely increase the PSNR, we see clearly improved results when excluding the center pixel from random sampling for replacement. In addition, a non-residual U-Net further improves the result compared to the residual U-Net that is used by default in the Noise2Void configuration. In a similar line, also our other architecture adaptations yield increased PSNR values. While the proposed replacement strategies mean and median do not result in better quantitative results, we are surprised to see that the mean replacement strategy clearly reduces checkerboard artifacts qualitatively as can be seen in Figure 5. We finally observe that the best fully self-supervised results in the medium and high noise regime are obtained by combining both architecture and replacement adaptations.

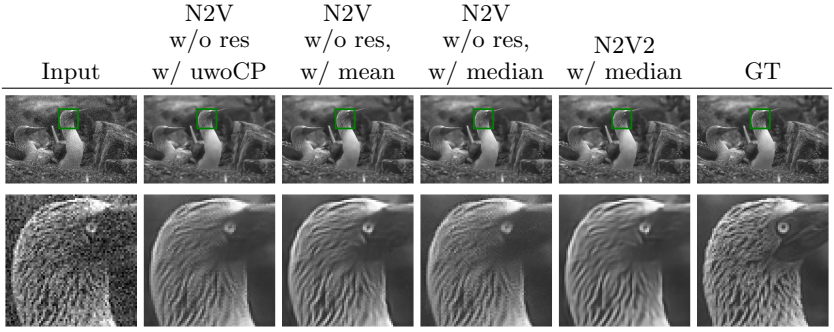

Fig. 6: Qualitative results on the BSD68 dataset. After applying the trained N2V model without residual connection and sampling without the center point to the noisy input (1st), the result shows undesirable checkerboard artifacts (2nd, PSNR 29.01).These remain when using median replacement (4th, PSNR 28.79), are reduced with mean replacement (3rd, PSNR 29.27), and eliminated when using N2V2 with median replacement (5th, PSNR 29.24) method

4.4 Evaluation Flywing G70, Mouse G20, BSD68

We report results for the datasets with simulated Gaussian noise in Table 2. In contrast to the results for simulated salt and pepper noise, we interestingly see that results do not improve simply by excluding the center pixel from the window for sampling replacement. Also, not using a residual U-Net only yields slight improvements for the microscopy datasets and none for the natural image dataset BSD68, where PSNR even drops. However, the alternative replacement strategies mean and median lead to improved PSNR values, as well as the architecture adaptations bp sk. Combining both adaptations leads to the best self-supervised results for the Mouse G20 and BSD68 datasets.

This is in line with qualitative results shown in Figure 6 for the BSD dataset, where we clearly see checkerboard artifacts in the N2V standard setting, but significantly cleaner predictions with the proposed adaptations. Additional qualitative results for given in the supplementary material section S1.

4.5 Evaluation of Real Noisy Data: Convallaria_95 and Convallaria_1

As displayed in Table 3, both the median replacement strategy as well as the N2V2 architecture adaptations improve the results for both Convallaria datasets. This can also be seen in the qualitative example in Figure 7. N2V2 with median replacement strategy yields the best fully self-supervised results for both cases. Interestingly, according to PSNR values, the mean replacement method does not improve when compared to the baseline N2V performance.

Comparing the two columns in Table 3, a considerable difference in PSNR is apparent, with the denoising results when using the reduced Convallaria_1

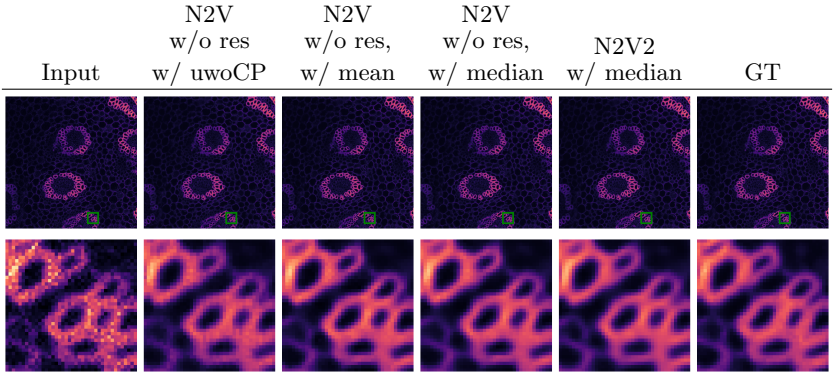

Fig. 7: Qualitative results on the convallaria dataset. After applying the trained N2V model without residual connection and sampling without the center point to the noisy input (1st), the result shows undesirable checkerboard artifacts (2nd, PSNR 35.78). These are eliminated when using mean (3rd, PSNR 35.91) and median replacement (4th, PSNR 36.39 and as well with the N2V2 with median replacement (5th, PSNR 36.37)

dataset being poorer. This leads to two possible interpretations, namely (*i*) having 95 noisy images of the same field of view allows for better results of the self-supervised denoising methods or (*ii*) results are poorer on the hold-out tiles of the Convallaria_1 test set because they represent parts of the field of view that were not seen during training. However, judging by Table 4, which displays a comparison of the results on the train vs the test tiles, this seems not to be the case. A similar conclusion is suggested by Figure 8, showing a qualitative comparison between denoised train and test tiles. Please also refer to the supplementary material section S2 for additional qualitative results obtained for the whole slide.

5 Discussion & Conclusions

In this work, we introduced N2V2, an improved setup for the self-supervised denoising method Noise2Void by Krull *et al.* [7]. N2V2 which is build around two complementary contributions: (*i*) a new network architecture, and (*ii*) modified pixel value replacement strategies for blind-spot pixels.

We showed that N2V2 reduces previously observed checkerboard artifacts, which have been responsible for reduced quality of predictions from Noise2Void. While we observed in qualitative examples that the mean replacement strategy is overall more successful than the median replacement strategy, we did not find this trend consistently in all quantitative results. Nonetheless, we have shown that only changing the architecture or only switching to one of our sampling strategies does already lead to improved results. Still, the combination of both

	Method	Convallaria_95	Convallaria_1
	Input	29.40	25.81
Fully self-supervised	N2V as in [7]	35.89	31.43
	N2V w/ uwoCP as in [8]	35.58	31.24
	N2V w/o res, w/ uwoCP	35.76	31.27
	N2V w/o res w/ mean	35.90	31.34
	N2V w/o res w/ median	**36.39**	**31.77**
	N2V2 w/ uwCP	36.26	31.45
	N2V2 w/ uwoCP	36.31	31.51
	N2V2 w/ mean	36.27	31.48
	N2V2 w/ median	36.36	31.28
Self-supervised	PN2V [8]	36.47	N/A
	DivNoising [14]	<u>36.90</u>	N/A
Supervised	CARE [19]	36.71	N/A

Table 3: Quantitative results on real Convallaria data. Results are given in dB of averaged PSNR on test data. Overall best is <u>underlined</u>. Best fully self-supervised is in **bold**

	Method	Convallaria_1 train	Convallaria_1 test
	Input	25.21	25.81
Fully self-supervised	N2V w/o res, w/ uwoCP	30.52	31.27
	N2V w/o res, w/ mean	31.37	31.34
	N2V w/o res, w/ median	31.35	31.77
	N2V2 w/ mean	31.10	31.48
	N2V2 w/ median	31.06	31.28

Table 4: Quantitative results on the Convallaria_1 train and test sets. Results are given in dB of averaged PSNR

yields best overall denoising results (measured by means of PSNR to ground clean truth images).

Another important observation is that the residual connections, which are by default used in N2V, can lead to worse results. This is most apparent in the S&P noise setting, where denoising with a residual U-Net and sampling replacement with inclusion of the center pixel (uwCP) yields at best a marginal improvement over the input. Additionally, we saw similar behaviour when combining the residual U-Net with the average replacement method (not shown).

When using a residual U-Net, pixels altered by a huge amount of noise appear at times to be strongly biased by the residual input and denoising is therefore negatively effected. Without residual connections, on the other hand, this bias is removed and performance therefore improved.

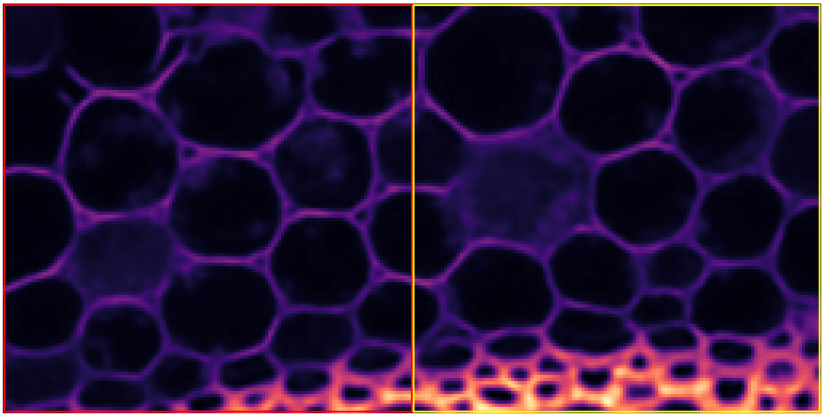

Fig. 8: Does prediction on training data impact N2V quality? Predictions on data which used for training N2V (left) and hold-out data (right)

Additionally, we have introduced a modified Convallaria data set (*Convallaria_1*), now featuring (*i*) a clean split between train, validation and test sets, and offering (*ii*) a more realistic scenario to test self-supervised denoising methods. The newly proposed dataset includes only one noisy input image instead the previously used 99 noisy acquisitions of the same field of view of the same sample. We strongly urge future methods to be evaluated on this improved Convallaria setup.

As a final point of discussion, we note that since we decided to train all N2V and N2V2 setups much longer than in previous publications (*e.g.*, [8]), even the baselines we have simply re-run now outperform the corresponding results as reported in the respective original publications. This indicates that original training times were chosen too low and urges all future users of self-supervised denoising methods to ensure that their training runs have indeed converged before stopping them[4].

We have presented an improved version of N2V, namely N2V2, a self-supervised denoising method leading to denoising results of improved quality on virtually all biomedical microscopy data. At the same time, N2V2 is equally elegant, does not require more or additional training data, and is equally computationally efficient as N2V. Hence, we hope that N2V2 will mark an important update of N2V and will continue the success which N2V has celebrated in the past three years.

[4] Note that this is harder to judge for self-supervised compared to supervised methods since loss plots report numbers that are computed between predicted values and *noisy* blind-spot pixel values.

References

1. Batson, J., Royer, L.: Noise2self: Blind denoising by self-supervision. In: International Conference on Machine Learning (ICML). pp. 524–533. PMLR (2019) 2, 3, 6
2. Buchholz, T.O., Jordan, M., Pigino, G., Jug, F.: Cryo-care: content-aware image restoration for cryo-transmission electron microscopy data. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 502–506. IEEE (2019) 3
3. Buchholz, T.O., Krull, A., Shahidi, R., Pigino, G., Jékely, G., Jug, F.: Content-aware image restoration for electron microscopy. Methods in cell biology **152**, 277–289 (2019) 2, 3
4. Buchholz, T.O., Prakash, M., Schmidt, D., Krull, A., Jug, F.: DenoiSeg: Joint denoising and segmentation. In: Computer Vision – ECCV 2020 Workshops. pp. 324–337. Springer International Publishing (2020) 2
5. Buchholz, T.O., Prakash, M., Schmidt, D., Krull, A., Jug, F.: Denoiseg: joint denoising and segmentation. In: European Conference on Computer Vision (ECCV). pp. 324–337. Springer (2020) 8, 9
6. von Chamier, L., Laine, R.F., Jukkala, J., Spahn, C., Krentzel, D., Nehme, E., Lerche, M., Hernández-Pérez, S., Mattila, P.K., Karinou, E., Holden, S., Solak, A.C., Krull, A., Buchholz, T.O., Jones, M.L., Royer, L.A., Leterrier, C., Shechtman, Y., Jug, F., Heilemann, M., Jacquemet, G., Henriques, R.: Democratising deep learning for microscopy with ZeroCostDL4Mic. Nat. Commun. **12**(1), 2276 (Apr 2021) 2
7. Krull, A., Buchholz, T.O., Jug, F.: Noise2void-learning denoising from single noisy images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2129–2137 (2019) 2, 3, 4, 5, 6, 8, 9, 10, 12, 13
8. Krull, A., Vičar, T., Prakash, M., Lalit, M., Jug, F.: Probabilistic noise2void: Unsupervised content-aware denoising. Frontiers in Computer Science **2**, 5 (2020) 2, 3, 4, 6, 7, 8, 9, 13, 14
9. Laine, S., Karras, T., Lehtinen, J., Aila, T.: High-quality self-supervised deep image denoising. Advances in Neural Information Processing Systems (NeurIPS) **32** (2019) 2
10. Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., Aila, T.: Noise2noise: Learning image restoration without clean data. arXiv preprint arXiv:1803.04189 (2018) 2, 3, 5
11. Jiménez de la Morena, J., Conesa, P., Fonseca, Y.C., de Isidro-Gómez, F.P., Herreros, D., Fernández-Giménez, E., Strelak, D., Moebel, E., Buchholz, T.O., Jug, F., Martinez-Sanchez, A., Harastani, M., Jonic, S., Conesa, J.J., Cuervo, A., Losana, P., Sánchez, I., Iceta, M., Del Cano, L., Gragera, M., Melero, R., Sharov, G., Castaño-Díez, D., Koster, A., Piccirillo, J.G., Vilas, J.L., Otón, J., Marabini, R., Sorzano, C.O.S., Carazo, J.M.: ScipionTomo: Towards cryo-electron tomography software integration, reproducibility, and validation. J. Struct. Biol. **214**(3), 107872 (Jun 2022) 3
12. Ouyang, W., Beuttenmueller, F., Gómez-de Mariscal, E., Pape, C., Burke, T., Garcia-López-de Haro, C., Russell, C., Moya-Sans, L., de-la Torre-Gutiérrez, C., Schmidt, D., Kutra, D., Novikov, M., Weigert, M., Schmidt, U., Bankhead, P., Jacquemet, G., Sage, D., Henriques, R., Muñoz-Barrutia, A., Lundberg, E., Jug, F., Kreshuk, A.: BioImage model zoo: A Community-Driven resource for accessible deep learning in BioImage analysis (Jun 2022) 2

13. Prakash, M., Delbracio, M., Milanfar, P., Jug, F.: Interpretable unsupervised diversity denoising and artefact removal. In: International Conference on Learning Representations (ICLR) (2022) 2, 3
14. Prakash, M., Krull, A., Jug, F.: Fully unsupervised diversity denoising with convolutional variational autoencoders. In: International Conference on Learning Representations (ICLR) (2021) 2, 3, 4, 8, 9, 13
15. Prakash, M., Lalit, M., Tomancak, P., Krull, A., Jug, F.: Fully unsupervised probabilistic noise2void. In: IEEE International Symposium on Biomedical Imaging (ISBI) (2020) 2, 3, 4, 6, 8
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 4
17. Schroeder, A.B., Dobson, E.T.A., Rueden, C.T., others: The ImageJ ecosystem: Open-source software for image visualization, processing, and analysis. Proteins (2021) 2
18. Weigert, M., Royer, L., Jug, F., Myers, G.: Isotropic reconstruction of 3d fluorescence microscopy images using convolutional neural networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI. pp. 126–134. Springer International Publishing, Cham (2017) 2
19. Weigert, M., Schmidt, U., Boothe, T., Müller, A., Dibrov, A., Jain, A., Wilhelm, B., Schmidt, D., Broaddus, C., Culley, S., et al.: Content-aware image restoration: pushing the limits of fluorescence microscopy. Nature methods **15**(12), 1090–1097 (2018) 2, 3, 5, 8, 9, 13
20. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing (TIP) **26**(7), 3142–3155 (2017) 2, 8
21. Zhang, R.: Making convolutional networks shift-invariant again. In: International conference on machine learning (ICML). pp. 7324–7334. PMLR (2019) 2, 4, 5

N2V2 - Fixing Noise2Void Checkerboard Artifacts with Modified Sampling Strategies and a Tweaked Network Architecture – Supplementary Material –

Anonymous ECCV WS-BIC submission

Paper ID 11

Abstract. In this supplementary document, we provide additional qualitative results to further strengthen our findings as reported in the main paper.

S.1 Additional qualitative results for BSD68

In the main paper, we report quantitative and qualitative results in Section 4.4 for the BSD68 natural images dataset. In Figure S.1 and Figure S.2, we add more qualitative results to further underline the benefits of N2V2.

S.2 Whole slide results on the Convallaria_1 dataset

In the main paper, we report results on the Convallaria_1 dataset in Section 4.5. We discuss that although a clear separation of training data and test data is a sound experimental setup even for self-supervised training scenarios, no clear differences for denoising results have been observed. In Figure S.3, we show additional qualitative results obtained with N2V2 w/ mean by visualizing denoising results on the entire Convallaria slide in Figure S.3. The origin of each patch, *i.e.*, if being used in the training set, validation set, or test set, is further indicated by the colored frame.

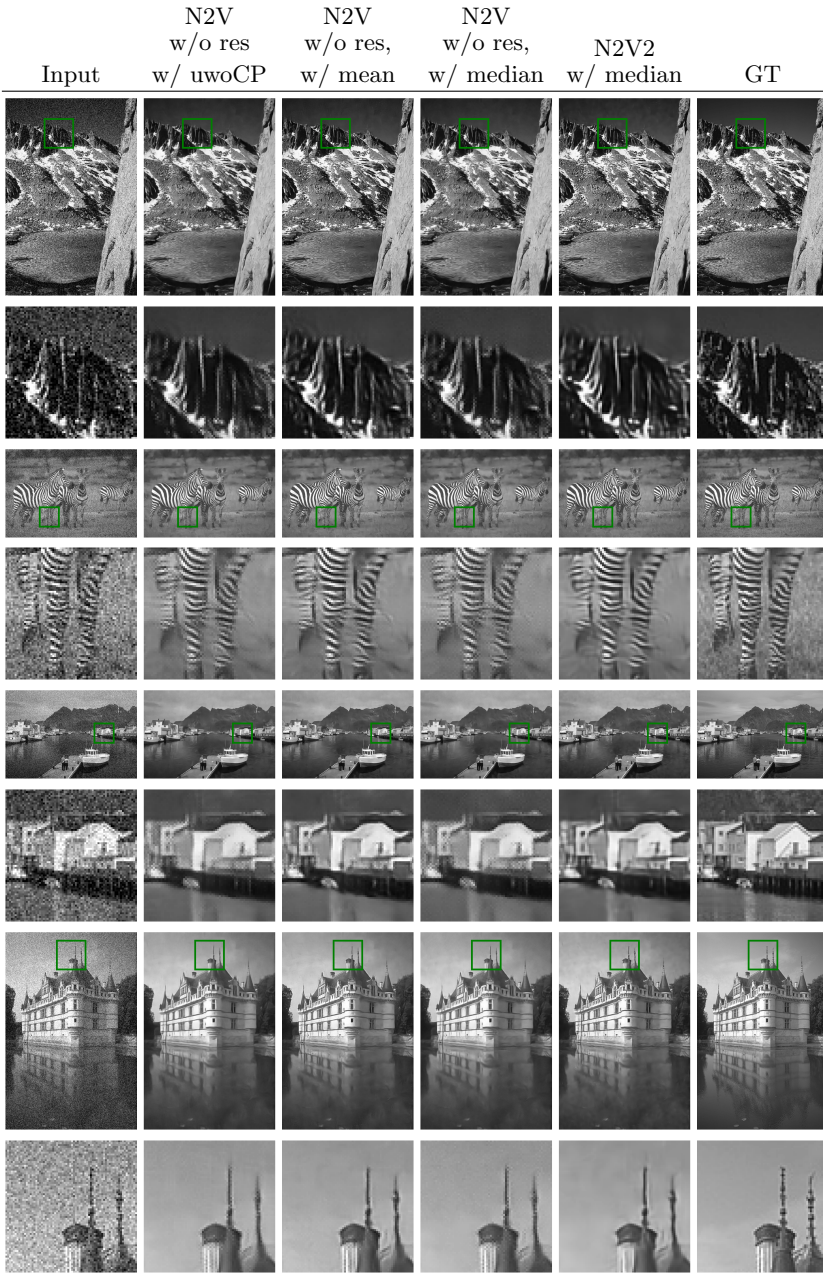

Fig. S.1: Additional qualitative results on the BSD68 dataset. After applying the trained N2V model without residual connection and sampling without the center point to the noisy input (1st), the result shows undesirable checkerboard artifacts (2nd).These remain when using median replacement (4th), are reduced with mean replacement (3rd), and eliminated when using N2V2 with median replacement (5th) method

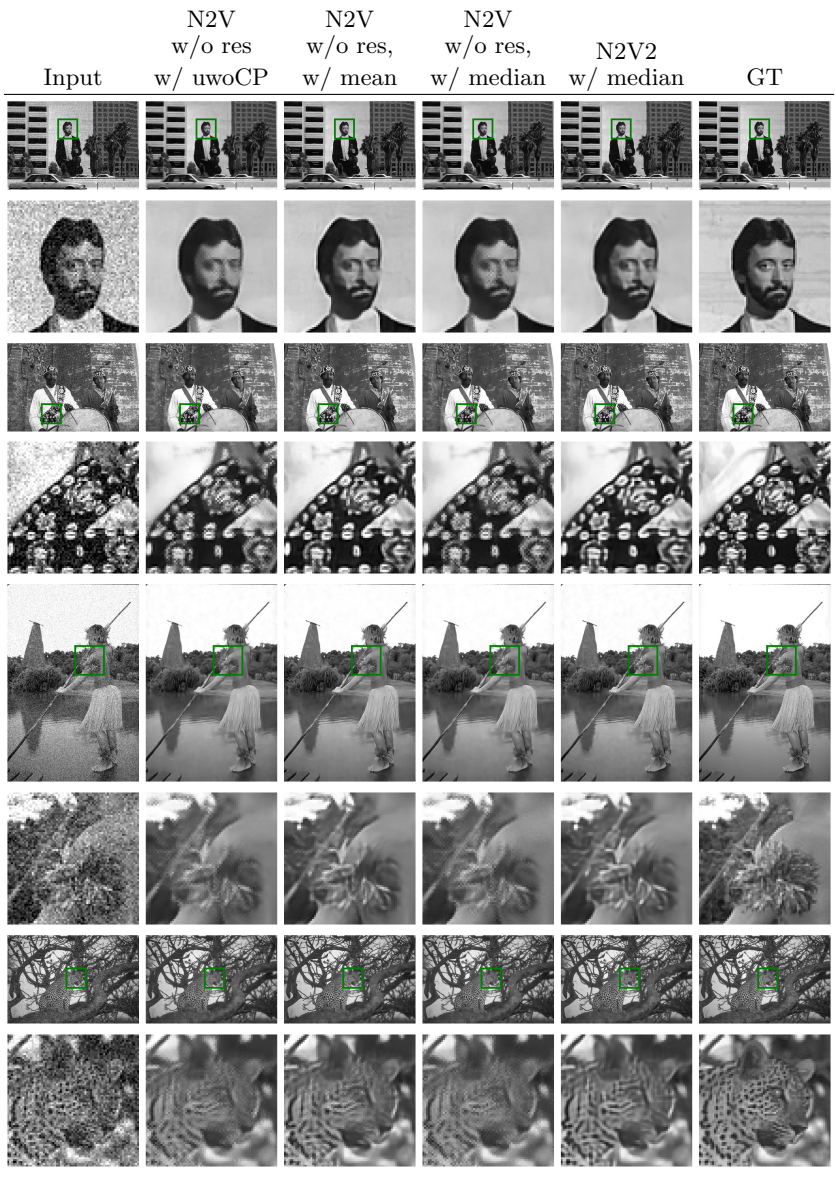

Fig. S.2: Even more qualitative results on the BSD68 dataset. After applying the trained N2V model without residual connection and sampling without the center point to the noisy input (1st), the result shows undesirable checkerboard artifacts (2nd).These remain when using median replacement (4th), are reduced with mean replacement (3rd), and eliminated when using N2V2 with median replacement (5th) method

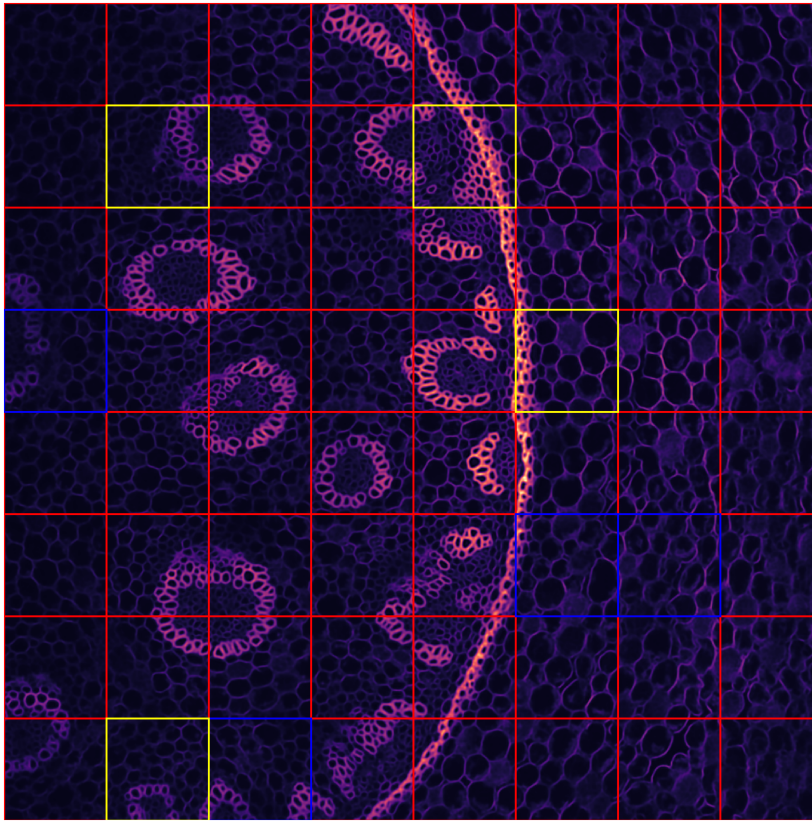

Fig. S.3: What is used for train and test? Red: used for training. Blue: used for validation (*i.e.*, , not for network parameter optimization during training of N2V, but we still create a val-set due to historical reasons). Yellow: only used for testing