REMOVING STRUCTURED NOISE WITH SELF-SUPERVISED BLIND-SPOT NETWORKS

Coleman Broaddus[1,2] *Alexander Krull*[1,2] *Martin Weigert*[1,2,3] *Uwe Schmidt*[1,2] *Gene Myers*[1,2]

[1]Max Planck Institute of Molecular Cell Biology and Genetics (MPI-CBG), Dresden, Germany
[2]Center for Systems Biology Dresden (CSBD), Germany
[3]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Removal of noise from fluorescence microscopy images is an important first step in many biological analysis pipelines. Current state-of-the-art supervised methods employ convolutional neural networks that are trained with clean (ground-truth) images. Recently, it was shown that self-supervised image denoising with *blind spot networks* achieves excellent performance even when ground-truth images are not available, as is common in fluorescence microscopy. However, these approaches, *e.g.* Noise2Void (N2V), generally assume pixel-wise independent noise, thus limiting their applicability in situations where spatially correlated (structured) noise is present. To overcome this limitation, we present *Structured Noise2Void* (StructN2V), a generalization of blind spot networks that enables removal of structured noise without requiring an explicit noise model or ground truth data. Specifically, we propose to use an extended *blind mask* (rather than a single pixel/blind spot), whose shape is adapted to the structure of the noise. We evaluate our approach on two real datasets and show that StructN2V considerably improves the removal of structured noise compared to existing standard and blind-spot based techniques.

***Index Terms*—** image denoising, deep learning, CNN, self-supervised, structured noise

1. INTRODUCTION

Removal of noise in fluorescence microscopy images is often an important step in many imaging projects to facilitate visualization, and further downstream processing such as image segmentation, detection or tracking [1, 2, 3, 4]. This applies especially to biological microscopy, where the signal-to-noise ratio (SNR) is often a compromise of different requirements, especially in live-cell imaging (phototoxicity, temporal resolution, *etc.*).

Machine learning based image denoising has recently started to outperform strong engineered methods (such as NLM [5] or BM3D [6]), in both image quality and processing speed (*e.g.* [7]). The gap between engineered and learned methods has widened even more since the advent of deep learning (DL). In particular, convolutional neural networks (CNNs) have shown remarkable improvements for many image restoration tasks, including image denoising [8, 9, 10, 11]. Image restoration based on CNNs has also shown substantially improved results for fluorescence microscopy images [4].

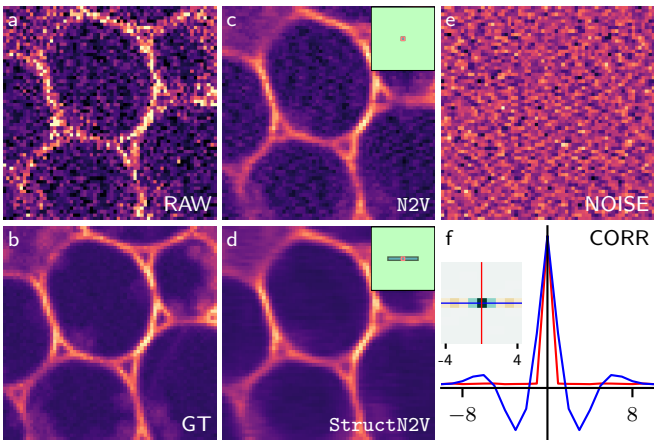

Fig. 1: For noisy microscopy images of *C. majalis* (a), our approach StructN2V (d) can remove horizontal noise artifacts, which N2V (c) is unable to do (ground truth shown in (b)). Autocorrelation (f) of a pure noise image (e) reveals the horizontal shape of the structured noise pattern. Insets in (c) and (d) depict the blind masks used during training of the CNNs.

However, learned methods have the disadvantage that they need to be trained, which is typically done via *supervised learning*, requiring pairs of noisy *input* images (with low SNR) and desired output images (with high SNR), called *ground truth* (GT). High-SNR ground truth images are typically acquired by increasing the light intensity or exposure time. However, many samples are too fragile or sensitive to handle the higher light intensities required to reduce noise.

Lehtinen *et al.* [12] have shown that neural networks for image denoising can also be trained with pairs of only low-SNR images, using one image as training input and one as ground truth. However, applicability of the method is limited for microscopy. To achieve the best restoration quality, it is crucial that each image pair of input and ground truth is very well aligned or registered, depicting the same im-

In Proceedings of the *IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Iowa City, Iowa, April 2020. doi:10.1109/ISBI45749.2020.9098336

age structures, and only differs in terms of corruption (*i.e.*, noise). In practice, fast biological processes can elude the repeated exposures required to record such data, rendering the method impractical. Furthermore, when attempting to denoise already acquired datasets, it is often impossible to collect additional images of similar characteristics for training.

Recent *self-supervised* denoising approaches have alleviated these issues of training data acquisition to a large extent. Krull *et al*. [13, 14] (N2V) and Batson and Royer [15] proposed methods to train denoising CNNs without requiring training pairs. During training, these methods use the same noisy image as both input and ground truth. To enable training and prevent the network from simply learning the identity, they follow a particular training procedure: they calculate the loss only at individual pixels, which have been previously masked (replaced with random values) in the input image. The masked pixels are referred to as *blind spots*.

Self-supervised methods do not require the collection of additional training data, making high-quality image denoising via CNNs accessible to a wide range of bioimage data incompatible with supervised training. However, this strategy comes with a caveat. It is based on the assumption that the noise in each pixel of a given image is generated independently. More formally, the noise in any two pixels is assumed conditionally independent given the true noise-free image. While this typically holds true for the most dominant sources of noise (such as Poisson shot noise and additive read out noise), it poses a limitation in practical microscopy. In practice, due to the acquisition process in many microscopes the noise can be highly correlated among neighboring pixels (see Fig. 1) and causes N2V and other self-supervised methods to produce poor results [13].

Here, we introduce *Structured Noise2Void* (StructN2V) to address this problem. To that end, we go beyond the assumption of pixel-wise (conditionally) independent noise and demonstrate how N2V can be extended, by masking not individual pixels but a larger area during training. Our method yields substantially improved restoration quality for images with locally correlated noise. That way, we facilitate the practical application of self-supervised CNN denoising on an even wider array of real bioimage datasets.

2. METHOD

Each of the recent self-supervised denoising methods [13, 15, 16] is trained[1] such that the same noisy image is used both as input and prediction target. However, in order to prevent the model from learning the identity function, pixels in the model output on which the loss is evaluated (*active pixels*) must be *hidden* in the input.[2] This forces the model to predict

[1]Note that we only consider CNN training here.

[2]While in theory optimal performance is achieved by using a single *active* target pixel, in practice there are typically multiple (spread-out) masked pixels to enable more efficient training.

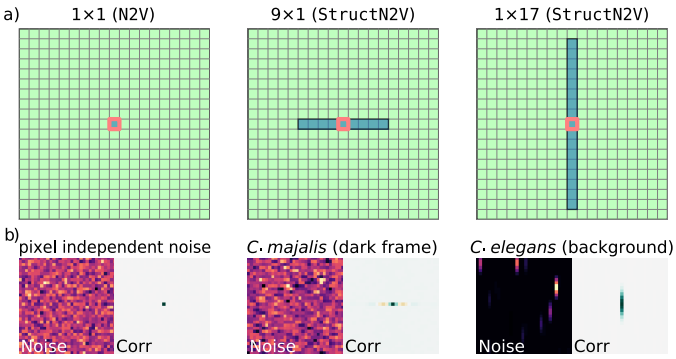

Fig. 2: (a) Masking schemes for N2V (left) and two variants of StructN2V (middle, right). Shown are the *active pixel* (red), the hidden *blind mask* (blue) and the pixels available (green) to the network to predict the value of the active pixel. (b) Example crops of pure noise images and their spatial autocorrelation (Corr) for pixel independent (Gaussian) noise (left), a dark frame of dataset *C. majalis* (middle), and a background frame of dataset *C. elegans* (right). Note that the spatial extent of the autocorrelation pattern in each case suggests the corresponding blind mask depicted directly above in (a).

their values from the surrounding pixels. While a reasonable prediction is possible for a structured signal of interest, the best prediction for the noise, which is assumed (conditionally) pixel-wise independent and zero centered, is its expected value (*i.e.*, zero). Hence, there is a denoising effect.

A pixel can be hidden by designing a special network architecture and receptive field [16]. An architecture-independent alternative is pixel *masking* [13, 15], whereby the loss is only active for some randomly selected pixels, whose values in the input image have been replaced (masked) with random values drawn from a local neighborhood [13] or uniform distribution [15]. While designing a special architecture has practical advantages, it does lack flexibility in hiding more than a single pixel, which is what we are proposing below.

In the case of spatially extended structured noise, the neural network will be able to predict the noise contribution to the value of an *active* pixel from neighboring noisy pixels. Removal of such noise will thus fail. To reduce this effect for a given pixel, we suggest to additionally hide (neighboring) pixels that contain information about the noise of the active pixel. Hence, we propose to use an extended *blind mask* of pixels that are replaced by random values in the input image. We still only have the loss active for individual pixels at the center of the mask. This amounts to decoupling pixels that are active in the loss (*active pixels*) from those that are hidden in the input (*blind mask*, *cf.* Fig. 2). We call our approach StructN2V. Note that StructN2V includes N2V as a special case, when the blind mask and the active pixel coincide (*cf.* Fig. 2).

Evidently, hiding additional pixels makes the signal of in-

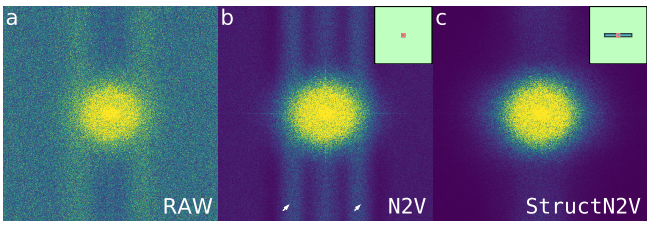

Fig. 3: Comparison of Fourier power spectrum of *C. majalis* images (a), denoised with N2V (b) and StructN2V (c). While N2V and StructN2V both reduce noise across the spectrum, vertical bands in the spectrum (denoted by white arrows) correspond to horizontal noise artifacts not removed by N2V. Insets in (b) and (c) depict the blind masks used during training.

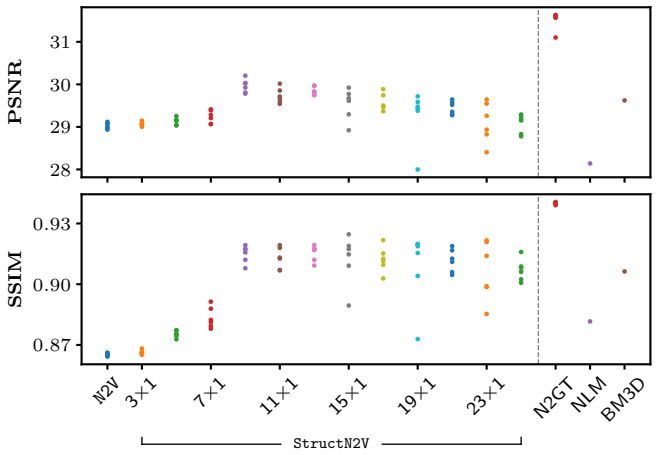

Fig. 4: Comparison of denoising results for dataset *C. majalis* via PSNR and SSIM [17]. For StructN2V we use (horizontal) blind masks of size $w\times1$ with varying width w. The performance of StructN2V initially improves with increased w before it plateaus / gradually decreases for $w \geq 9$. For all learned methods (N2V, StructN2V, N2GT) we depict the results of 5 different training runs using different random seeds (individual dots). Note that all methods to the right of the gray line are optimized with knowledge of the ground truth.

terest less predictable as well. Hence, a tradeoff exists between reducing the amount of reconstructed structured noise and the ability to better restore the underlying signal. Therefore our approach is only practical if the length scale of correlations of the noise is smaller than that of the signal.

Selecting the blind mask. A suitable blind mask is crucial to achieve superior results with StructN2V (over N2V). Determining a good mask automatically is challenging, however, since the shape of the mask is a tradeoff: it should be as small as possible, but include pixels whose noise value is (highly) predictive for the noise at a given active pixel. To guide the mask selection, it is helpful to compute the spatial *autocorrelation of the noise* (Fig. 1f, Fig. 2): the autocorrelation (Corr) pattern is typically spatially concentrated, and its support is often well approximated by a neighborhood of rectangular size $w\times h$, which we then choose as the blind mask (Fig. 2). A pure noise image for this analysis can be obtained by selecting an "empty" area of the image, or by acquiring a so-called dark frame (*i.e.*, with the shutter closed). Of course, this assumes that the structured noise is caused or dominated by the readout process. In some cases, the structured noise is already visible in the raw data, or can be greatly enhanced by removing only the pixel-independent noise, *e.g.* with N2V (*cf.* Fig. 1c).

3. EXPERIMENTS

We evaluate our StructN2V approach on two fluorescence microscopy datasets, one (*C. majalis*) acquired in 2D with a camera-based spinning disk microscope, the other (*C. elegans*) in 3D with a laser-scanning confocal microscope.

3.1. *C. majalis*

We acquired 2D images of fluorescently labeled membranes of a fixed *C. majalis* (lilly of the valley) sample. All 100 recorded images (1024×1024 pixels) show the same region of interest and only differ in their noise. The average of these 100 images is considered as ground truth (GT) image[3]. We train and compare N2V, StructN2V, and N2GT[4] while using the same neural network architecture for all methods, only the training procedure is particular to each method.

CNN architecture. We use a U-Net [18] with two levels, using two 3×3 convolutional layers followed by 2×2 max pooling per level. We learn 16 convolution kernels at level one and double (halve) that number after each pooling (up-sampling).

Training. We normalize the input images by applying an (outlier-robust) percentile-based normalization, such that most pixel values are in the range $[0, 1]$. We use the entire dataset for both training and evaluation, since there is no distinction between train and test images for N2V and StructN2V. However, note that the performance of N2GT is thus higher (as one would expect in practice), and therefore represents an upper bound. We train all methods on 1600 random patches of size 256×256 for 600 epochs (batch size 4) with the Adam optimizer [19] at a fixed learning rate of 2×10^{-5}. For N2V and StructN2V, we randomly sample 2% of *active* pixels in each training patch.

For our StructN2V method, we estimate the mask shape via the correlation of the noise (*cf.* Fig. 1f) as described in Section 2. To that end, we use a horizontal mask of size $w\times1$, whose width w we vary from small ($w = 3$ pixel) to large

[3]This is possible because we observe that the noise has spatially constant mean and is uncorrelated between acquisitions.

[4]*Noise-to-Ground-Truth*, *i.e.* conventional supervised training with the GT images as target.

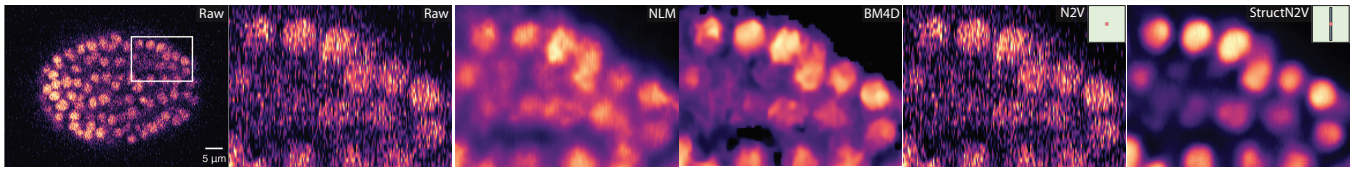

Fig. 5: Qualitative comparison for the 3D *C. elegans* dataset. We show a single slice of the raw stack and the result of NLM [5], BM4D [6], N2V [13] and StructN2V (ours). While N2V completely fails to remove any noise, StructN2V yields a perceptually noise-free image. Insets in N2V and StructN2V depict the blind masks used for training.

($w = 25$ pixel, *cf.* Fig. 2). In Fig. 4, we show denoising performance as a function of mask size, and also compare against N2V, N2GT, and standard methods. We find that StructN2V provides an improvement over N2V, although some masks are clearly better than others (with the mask 9×1 performing best, *cf.* Table 1). Importantly, the results of StructN2V show substantially less structured artifacts compared to N2V, as can be inferred from the Fourier power spectrum (Fig. 3): whereas the spectrum of both the raw (left) and N2V restoration (middle) exhibit vertical stripes at a distinct frequency range (*i.e.* indicating horizontal correlations), they are absent in the StructN2V output (right).

Additionally, Table 1 summarizes the best results of all methods (with a mask of size 9×1 for StructN2V). Note that NLM, BM3D and N2GT were all tuned/trained to yield optimal results using ground truth data (not available to N2V and StructN2V), hence the results denote upper bounds on what is possible in practice. Still, StructN2V performs better than all compared methods (apart from N2GT, which is expected).

3.2. *C. elegans*

We additionally demonstrate that StructN2V can substantially outperform N2V on the publicly available 3D dataset of a developing *C. elegans* (roundworm) embryo from [3], which was acquired with a laser-scanning confocal microscope. The data is a time series with 190 frames of size $32\times512\times708$ (ZYX). The images exhibit visually striking vertical correlations of unknown origin along the Y dimension. Fig. 5 shows that N2V completely fails to remove any of the noise, most likely due to the strong vertical correlation of the noise that renders it easy to predict a pixel's value from its immediate neighbor. In contrast, StructN2V yields smooth and perceptually noise-free images.

Method	PSNR	SSIM
StructN2V (ours)	$\mathbf{29.73 \pm 0.16}$	$\mathbf{0.913 \pm 0.005}$
N2V [13]	29.03 ± 0.06	0.8653 ± 0.0007
NLM [5]	28.14	0.8816
BM3D [6]	29.62	0.9063
N2GT	31.52 ± 0.19	0.9399 ± 0.0005

Table 1: Denoising results (PSNR and SSIM [17]) on *C. majalis* dataset. Note that we include only the best performing StructN2V mask (9×1). For all learned methods (N2V, StructN2V, N2GT) we display the mean ($\pm$ standard deviation) of 5 different training runs using different random seeds.

Network details. The mask shape for StructN2V was estimated as $1\times17\times1$ (ZYX, *cf.* Fig. 2). For all compared methods (N2V and StructN2V) we use a 3D U-Net [20] with two levels, using two $3\times5\times5$ (ZYX) convolutional layers followed by $1\times2\times2$ max pooling per level. Again we learn 16 convolution kernels at the first level and double (halve) that number after each pooling (up-sampling). We also apply the same image normalization as before and train both methods on 1900 random patches of size $32\times64\times64$ for 100 epochs (batch size 10) using Adam [19] at a fixed learning rate of 4×10^{-6}.

4. DISCUSSION

We proposed StructN2V, an extension of self-supervised blind spot networks, which allows us to effectively remove spatially correlated (*i.e.*, structured) noise without the necessity of acquiring clean ground truth images for training. The efficacy of our method is demonstrated on two fluorescence microscopy datasets that exhibit challenging structured noise. We were able to improve upon N2V and engineered image denoising methods (NLM and BM3D). Our approach relies on selecting an extended *blind mask* with the aim of hiding pixels from the neural network during training, thereby promoting faithful reconstruction of the image signal rather than of the structured noise.

We discussed several heuristics to select a good mask shape, which we found effective in practice. Nevertheless, in future work we would like to devise (semi-)automatic approaches to select appropriate blind mask shapes.

Acknowledgments

The authors thank the LMF of the MPI-CBG, in particular Britta Schroth-Diez for acquiring the *C. majalis* dataset, and Sebastian Bundschuh and Britta for valuable discussions regarding potential sources of structured noise. Finally, we thank our colleagues at the MPI-CBG and CSBD for their input throughout the duration of this work.

5. REFERENCES

[1] Qiang Wu, Fatima Merchant, and Kenneth Castleman, *Microscope image processing*, Elsevier, 2010.

[2] Jérôme Boulanger, Charles Kervrann, Patrick Bouthemy, Peter Elbau, Jean-Baptiste Sibarita, and Jean Salamero, "Patch-based nonlocal functional for denoising fluorescence microscopy image sequences," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 442–454, 2009.

[3] Vladimír Ulman, Martin Maška, Klas EG Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, et al., "An objective comparison of cell-tracking algorithms," *Nature Methods*, vol. 14, no. 12, pp. 1141, 2017.

[4] Martin Weigert, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Akanksha Jain, Benjamin Wilhelm, Deborah Schmidt, Coleman Broaddus, Siân Culley, Mauricio Rocha-Martins, Fabián Segovia-Miranda, Caren Norden, Ricardo Henriques, Marino Zerial, Michele Solimena, Jochen Rink, Pavel Tomancak, Loic Royer, Florian Jug, and Eugene W. Myers, "Content-aware image restoration: Pushing the limits of fluorescence microscopy," *Nature Methods*, vol. 15, no. 12, pp. 1090–1097, 2018.

[5] Antoni Buades, Bartomeu Coll, and J-M Morel, "A non-local algorithm for image denoising," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 2, pp. 60–65.

[6] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, "BM3D image denoising with shape-adaptive principal component analysis," in *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

[7] Uwe Schmidt and Stefan Roth, "Shrinkage fields for effective image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2774–2781.

[8] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[9] Junyuan Xie, Linli Xu, and Enhong Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 341–349.

[10] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[11] Viren Jain and Sebastian Seung, "Natural image denoising with convolutional networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2009, pp. 769–776.

[12] 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 (ICML)*, 2018, pp. 2971–2980.

[13] 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.

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

[15] Joshua Batson and Loic Royer, "Noise2self: Blind denoising by self-supervision," in *International Conference on Machine Learning (ICML)*, 2019, pp. 524–533.

[16] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila, "High-quality self-supervised deep image denoising," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 6968–6978.

[17] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] 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 (MICCAI)*, 2015, pp. 234–241.

[19] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[20] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2016.