Noise2Void - Learning Denoising from Single Noisy Images

Alexander Krull, Tim-Oliver Buchholz, Florian Jug

MPI-CBG/PKS (CSBD), Dresden, Germany

CVPR 2018

CONTENS

- Introduction
- Methods
- Experiments

Introduction

- A noisy image = signal + noise
  - $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}$
- **Traditional CNN:** need pairs of noisy input and clean target images $(\boldsymbol{x}, \boldsymbol{s})$
  - cannot be trained without clean target images
- **N2N:** need pairs of noisy input and noisy target images $(\boldsymbol{x}, \boldsymbol{x}')$
  - $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}, \boldsymbol{x}' = \boldsymbol{s} + \boldsymbol{n}'$
  - the acquisition of such pairs with constant $\boldsymbol{s}$ is only possible for static scenes
- **N2V:** need single noisy images as input and target $(\boldsymbol{x})$
- **Assumptions:**
  - (i) the signal $\boldsymbol{s}$ is not pixel-wise independent
  - (ii) the noise $\boldsymbol{n}$ is conditionally pixel-wise independent given the signal $\boldsymbol{s}$

Methods

- **Image Formation:**
- The generation of an image $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}$ as a draw from the joint distribution

$$p(\boldsymbol{s}, \boldsymbol{n}) = p(\boldsymbol{s})p(\boldsymbol{n}|\boldsymbol{s})$$

- The pixels $\boldsymbol{s}_i$ of the signal are not statistically independent

$$p\left(\boldsymbol{s}_i|\boldsymbol{s}_j\right) \neq p\left(\boldsymbol{s}_i\right)$$

- Noise are conditionally independent when the signal given

$$p(\boldsymbol{n}|\boldsymbol{s}) = \prod_i \boldsymbol{p}\left(\boldsymbol{n}_i|\boldsymbol{s}_i\right)$$

- We assume the noise to be zero-mean

$$\mathbb{E}\left[\boldsymbol{n}_i\right] = 0, \quad \mathbb{E}\left[\boldsymbol{x}_i\right] = \boldsymbol{s}_i$$

Traditional CNN

- We can also see our CNN as a function that takes a patch $x_{RF}(i)$ as input and outputs a prediction $\hat{s}_i$ for the single pixel i located at the patch center.

$$f\left(x_{\mathrm{RF}}(i); \boldsymbol{\theta}\right) = \hat{\boldsymbol{s}}_i$$

- Use these pairs to tune the parameters $\boldsymbol{\theta}$ to minimize pixel-wise loss

$$\arg\min_{\boldsymbol{\theta}} \sum_j \sum_i L\left(f\left(\boldsymbol{x}_{\mathrm{RF}}^j(i); \boldsymbol{\theta}\right) = \hat{\boldsymbol{s}}_i^j, \boldsymbol{s}_i^j\right)$$

- Consider the standard MSE loss

$$L\left(\hat{s}_i^j, s_i^j\right) = \left(\hat{s}_i^j - s_i^j\right)^2$$

Noise2Noise

- We start out with noisy image pairs $\left(x^j, x'^j\right)$

$$\boldsymbol{x}^j = \boldsymbol{s}^j + \boldsymbol{n}^j \text{ and } \boldsymbol{x'}^j = \boldsymbol{s}^j + \boldsymbol{n'}^j$$

- As in traditional training, we tune our parameters to minimize a loss

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

- The training will still converge to the correct solution because the expected value of the noisy input is equal to the clean signal.

$$\mathbb{E}\left[\boldsymbol{x}_i\right] = \boldsymbol{s}_i$$

Noise2Void

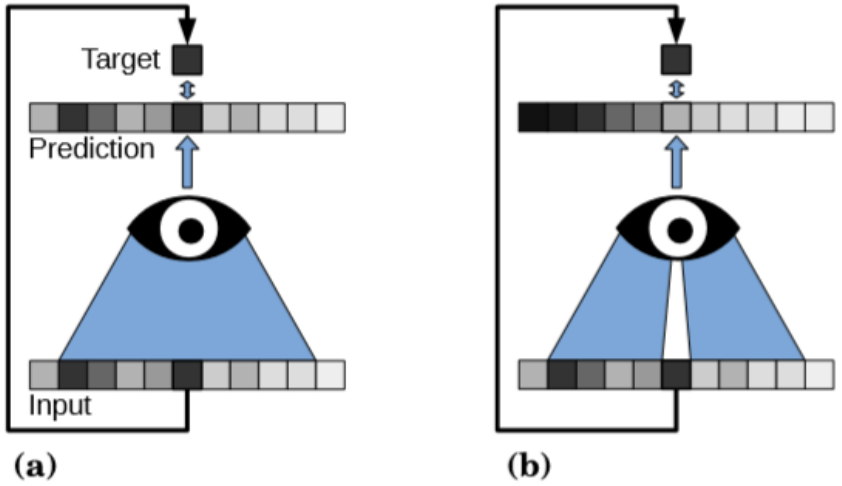


Figure 2: A conventional network versus our proposed blind-spot network. **(a)** In the conventional network the prediction for an individual pixel depends an a square patch of input pixels, known as a pixel's *receptive field* (pixels under blue cone). If we train such a network using the same noisy image as input and as target, the network will degenerate and simply learn the identity. **(b)** In a *blind-spot network*, as we propose it, the receptive field of each pixel excludes the pixel itself, preventing it from learning the identity. We show that blind-spot networks can learn to remove pixel wise independent noise when they are trained on the same noisy images as input and target.

- The input and the target are from a single noisy training image $\boldsymbol{x}_j$
- The receptive field $\tilde{\boldsymbol{x}}^j_{\mathrm{RF}}(i)$ of this network to have a blind-spot in its center.
- We can train it by minimizing the empirical risk

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

- Accuracy will be slightly impaired compared to a normal network

Noise2Void

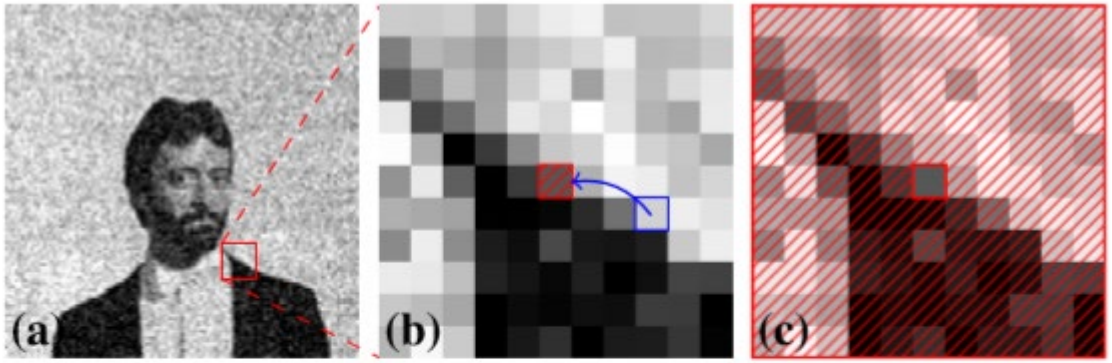
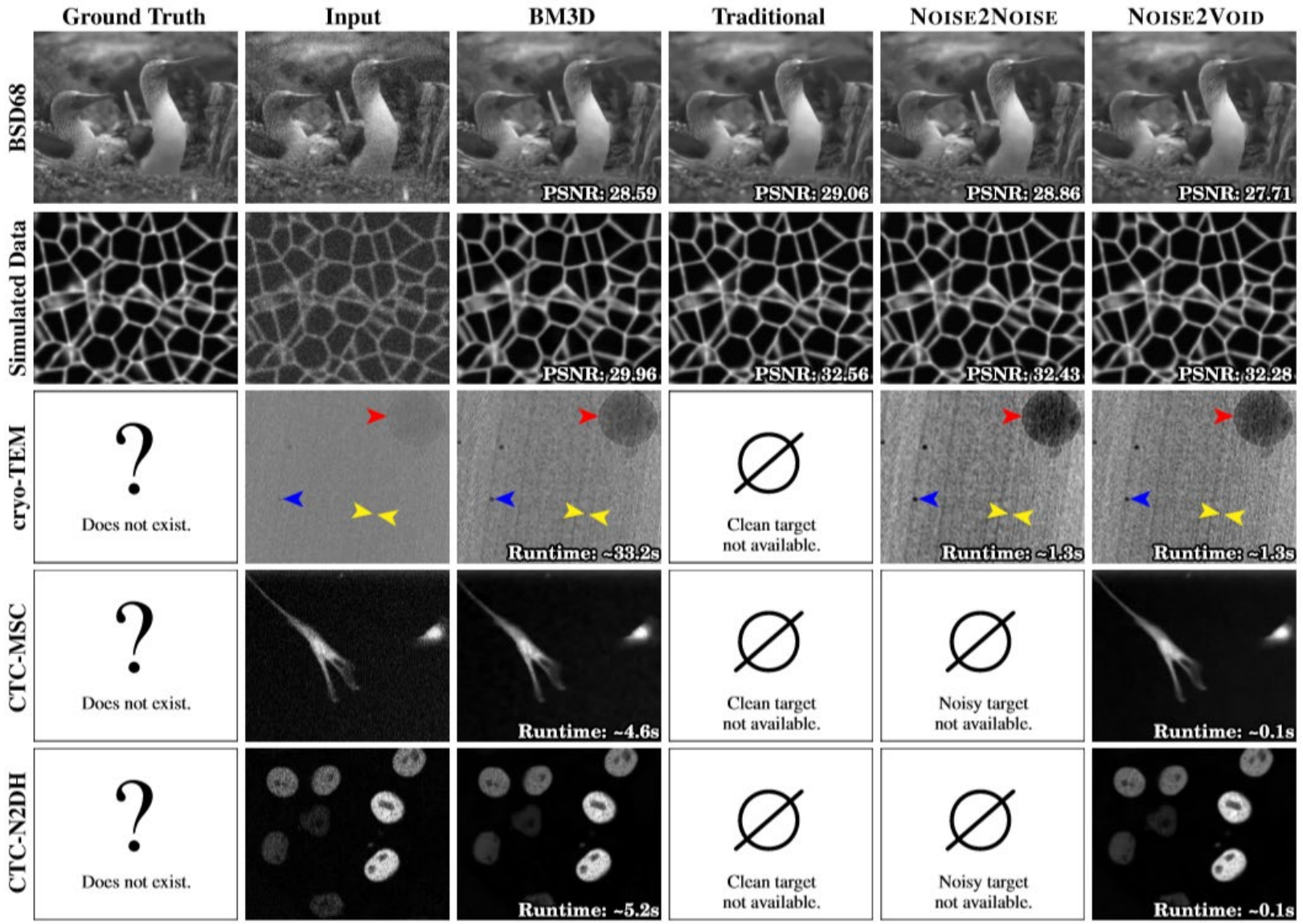


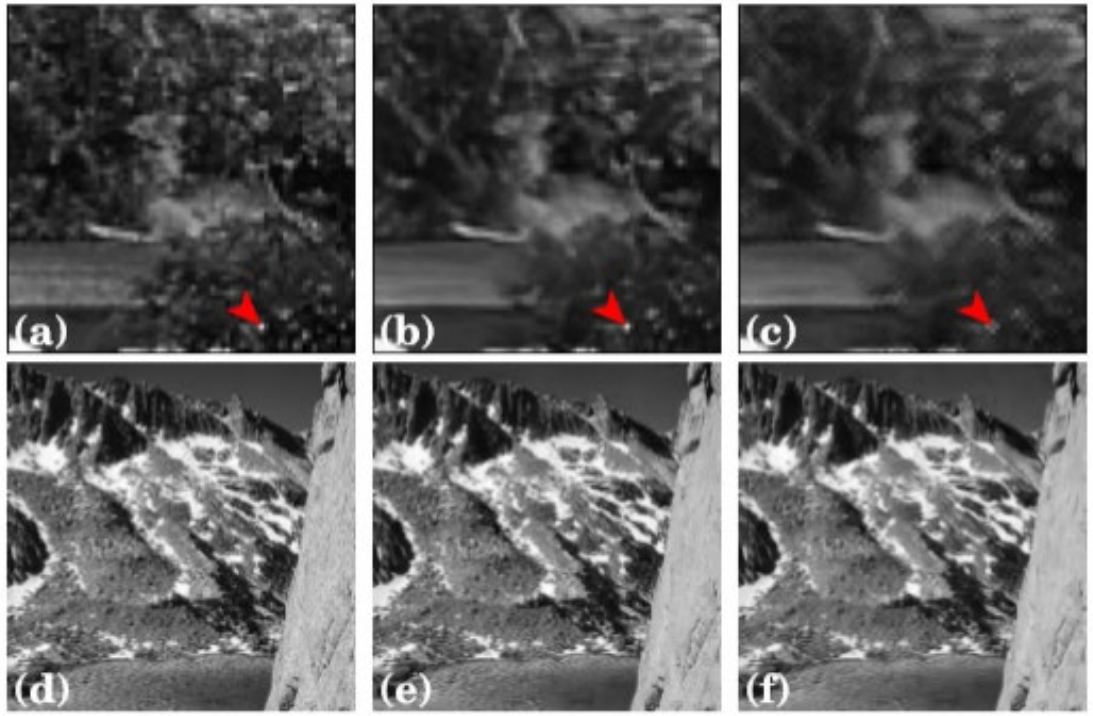
Figure 3: Blind-spot masking scheme used during NOISE2VOID training. **(a)** A noisy training image. **(b)** A magnified image patch from (a). During N2V training, a randomly selected pixel is chosen (blue rectangle) and its intensity copied over to create a blind-spot (red and striped square). This modified image is then used as input image during training. **(c)** The target patch corresponding to (b). We use the original input with unmodified values also as target. The loss is only calculated for the blind-spot pixels we masked in (b).

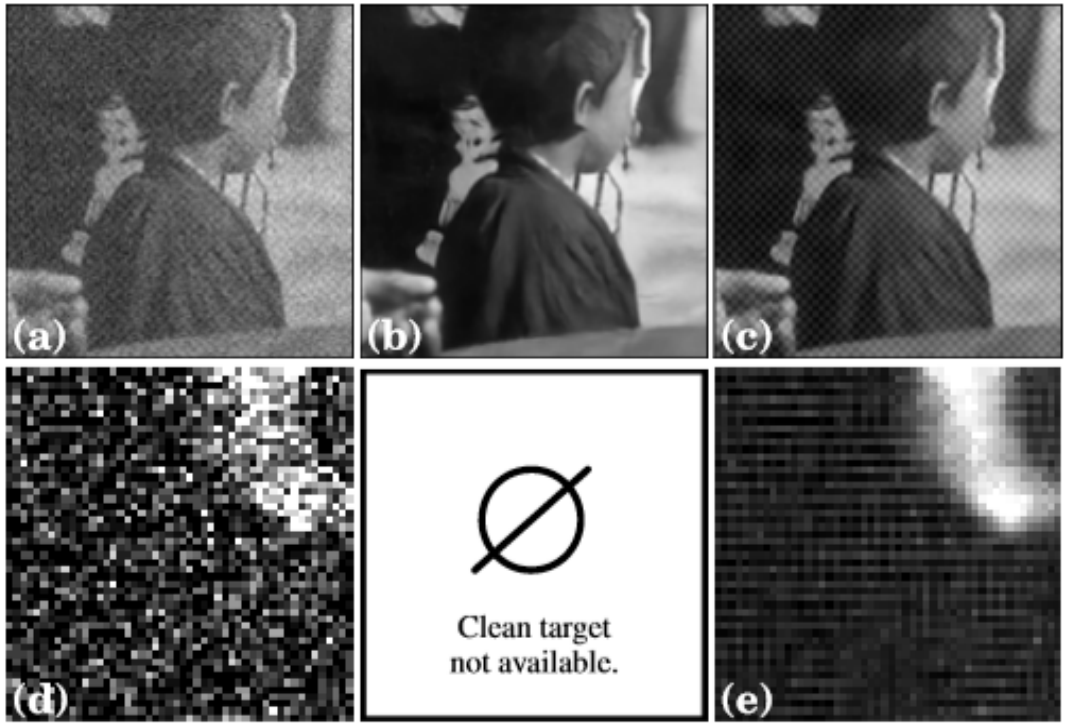
- If we implement the naïve training scheme, it is not efficient: We have to process an entire patch to calculate the gradients for a single output pixel
- Replace the value in the center of each input patch with a randomly selected value from the surrounding area
- We can now simultaneously calculate the gradients for all of them, while ignoring the rest of the predicted image
- A specialized loss function that is zero for all but the selected pixels

Experiments - Denoising Results and Costs



Experiments - Errors and Limitations

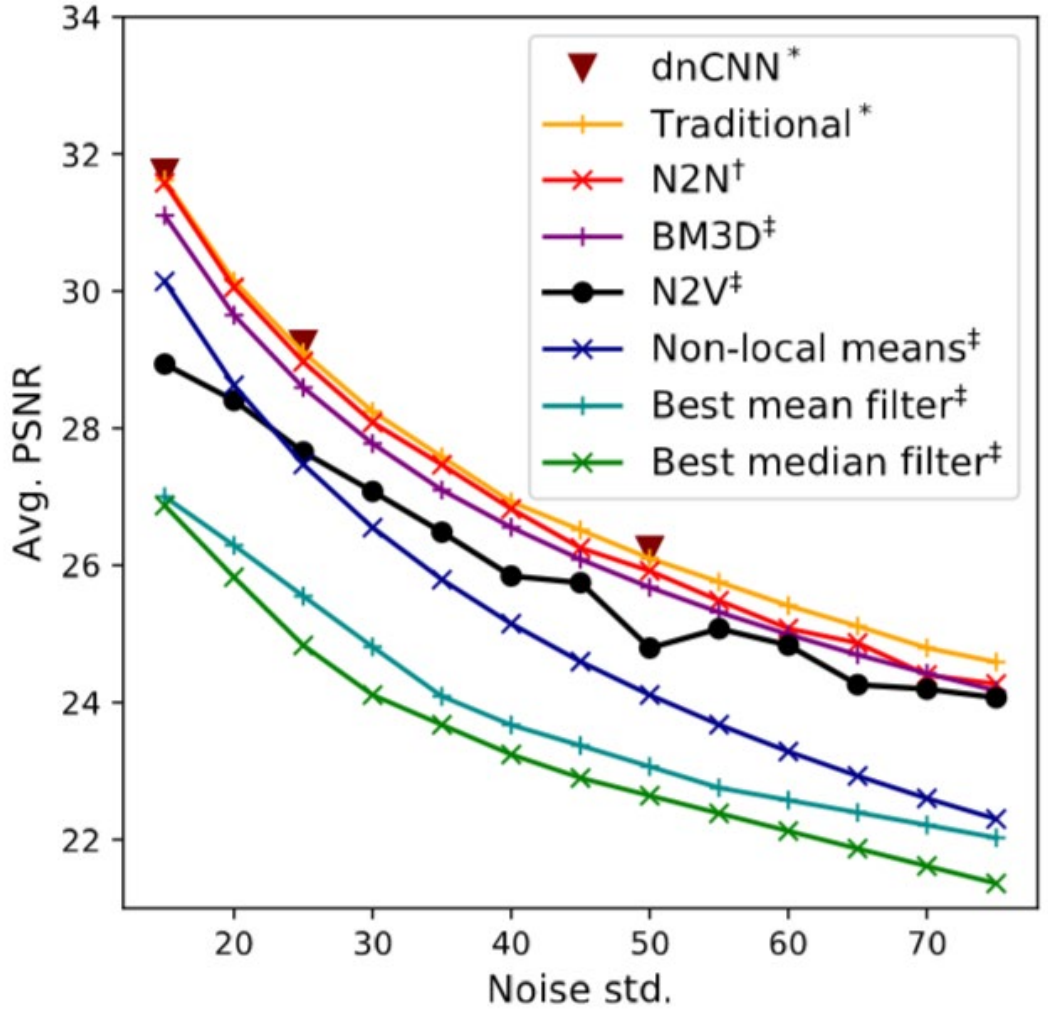




- Images with more high irregularities are difficult to predict
- The more difficult it is to predict a pixel's signal from its surroundings the more errors are expected to appearin N2V predictions

- Structured noise
- The N2V trained CNN removes the unpredictable components of the noise, but reveals the hidden pattern

Experiments - Performance



Mean filter (5x5)

Median filter (5x5)

N2V