1. Detailed Proof of Downsampled Invariance Loss

Proposition 1. Let \( x \) be a normalized zero-mean noisy image conditioned on \( y \), \( \mathbb{E}[x|y] = y \). Let \( d \) be any downsampling operation and \( d_s(x) \) be a set of downsampled pixels of \( x \) with a stride of \( s \). Assume that downsampled subimage \( d_s(x) \) has zero pixel-wise correlation and \( f_M \) is a blind-spot network. Then, the following inequality holds.

\[
\mathbb{E}_x \| f(x) - x \|^2 + \| x - y \|^2 \leq \mathbb{E}_x \| f(x) - x \|^2 + 2\sqrt{m}s^2 \mathbb{E}_{d_s(x)} \| d_s(f(x)) - f_M(d_s(x)) \|^2 \frac{1}{2}. \tag{1}
\]

Proof. We follow similar steps with the supplementary material of [5]. Self-supervised loss can be decomposed as

\[
\mathbb{E}_x \| f(x) - x \|^2 = \mathbb{E}_{x,y} \| f(x) - y \|^2 + \| x - y \|^2 - 2\langle f(x) - y, x - y \rangle. \tag{2}
\]

Then, Proposition 1 is equivalent to that the third term \( \langle f(x) - y, x - y \rangle \) is upper-bounded by the rightmost term in Eq. (1). \( \langle f(x) - y, x - y \rangle \) can be formulated as

\[
\mathbb{E}_{x,y} \langle f(x) - y, x - y \rangle = \mathbb{E}_y \mathbb{E}_x [y \sum_j (f(x)_j - y_j)(x_j - y_j)] \tag{3}
\]

\[
= \sum_j \mathbb{E}_y (\mathbb{E}_x |y \langle f(x)_j - y_j \rangle)(x_j - y_j) \tag{4}
\]

\[
- \mathbb{E}_{x|y} \langle f(x)_j - y_j \rangle \mathbb{E}_{x|y} \langle x_j - y_j \rangle \tag{5}
\]

\[
= \sum_j \mathbb{E}_y [\text{Cov}(f(x)_j - y_j, x_j - y_j)] \tag{6}
\]

\[
= \sum_j \mathbb{E}_y [\text{Cov}(f(x)_j, x_j|y)]. \tag{7}
\]

Eq. (5) holds since \( \mathbb{E}_{x|y} \langle x_j - y_j \rangle = 0 \) by the zero-mean noise assumption. Let \( J \) be a subset of the image sampled by a random downsampling operation \( d_s(x) \). Then we have the equation,

\[
\sum_j \mathbb{E}_y [\text{Cov}(f(x)_j, x_j|y)] = \frac{m}{|J|} \mathbb{E}_J \sum_j \mathbb{E}_y [\text{Cov}(f(x)_j, x_j|y)], \tag{8}
\]

since every pixel has the chance of selecting \( |J|/m = 1/s^2 \). On the right-hand side, the covariance term can be upper-bounded as

\[
\frac{1}{|J|} \sum_j \mathbb{E}_y [\text{Cov}(f(x)_j - f_M(d_s(x))_j, x_j|y)] \tag{9}
\]

\[
= \frac{1}{|J|} \sum_j \mathbb{E}_y [\text{Cov}(f(x)_j - f_M(d_s(x))_j, x_j|y)] \tag{10}
\]

\[
\leq \frac{1}{|J|} \sum_j \mathbb{E}_y [\text{Var}(f(x)_j - f_M(d_s(x))_j|y) \frac{1}{2} \cdot \text{Var}(x_j|y) \frac{1}{2}] \tag{11}
\]

\[
\leq \left( \frac{1}{|J|} \sum_j \mathbb{E}_y [\text{Var}(f(x)_j - f_M(d_s(x))_j|y) \cdot \text{Var}(x_j|y)] \right)^{\frac{1}{2}} \tag{12}
\]

\[
\leq \left( \frac{1}{|J|} \sum_j \mathbb{E}_y [\text{Var}(f(x)_j - f_M(d_s(x))_j|y)] \cdot 1 \frac{1}{2} \right] \tag{13}
\]

\[
= \left( \frac{1}{|J|} \sum_j \mathbb{E}_y [\text{Var}(f(x)_j - f_M(d_s(x))_j)] \right) \frac{1}{2} \tag{14}
\]

\[
= \left( \frac{1}{|J|} \sum_j \mathbb{E}_y [\text{Var}(f(x)_j - f_M(d_s(x)))] \right) \frac{1}{2} \tag{15}
\]

In Eq. (10), the equality holds since \( x_j \) is excluded in BSN and downsampled surroundings have no correlation with \( x_j \) by the assumption. Note that the Inequality (11) is derived from the Cauchy-Schwarz inequality, and the Inequality (12) is derived from Jensen’s inequality. Also, the Inequality (13) holds by the fact that \( \text{Var}(x) \leq \mathbb{E}[x^2] \), and by
the assumption that input \( x \) is normalized i.e., \( \text{Var}(x_j|y) \leq \text{Var}(x_j) = 1 \).

By the Proposition 1, we use Eq. (15) as downsampling invariance loss,
\[
\mathcal{L}_{\text{inv}} = \sqrt{\frac{s^2}{m}} |d_s(f(x)) − sg(f_M(d_s(x))))|_2, \quad (16)
\]
where \( sg \) is the stop gradient operation. \( f_M(d_s(x)) \) is introduced to Eq. (10) since it has zero correlation with \( x_j \). Therefore, we regard it as a constant and adopt a stop-gradient operation in the loss function. Lastly, we replace the root mean squared error with mean absolute difference in downsampling invariance loss as
\[
\mathcal{L}_{\text{inv}} = \frac{s^2}{m} |d_s(f(x)) − sg(f_M(d_s(x))))|_1. \quad (17)
\]

2. Analysis of Downsampling Ratio in Loss Functions

We conduct extensive experiments to analyze the effects of the downsampling ratios in \( \mathcal{L}_{\text{inv-RS}} \) and \( \mathcal{L}_{\text{blind}} \). Figure 1 shows the PSNR of C-BSN\(_{a/b}\) on SIDD validation dataset [1], where \( a \) is the stride of RS in the downsampling invariance loss and \( b \) is the stride of S2B in the blind loss.

Using strides less than 4 in the blind loss leads to suboptimal performance, showing that reducing spatial correlation of masked network input is crucial. Regarding the strides of RS, the performance tends to decrease as the stride increases over 3, while C-BSN with \( a = 1 \) fails to denoise the image. Although the performance is maximized with C-BSN\(_{3/4}\), the performance gap is marginal and falls within the range of variation caused by the randomness of the training process. Therefore, we adopt C-BSN\(_{2/5}\) as a baseline, consistent with AP-BSN [3].

3. Ablation on Downsampler of Blind Loss

We conduct an additional ablation study on the downsampler of blind loss. We follow the same setting as Section 4.3 in the paper. Table 1 reports PSNR and SSIM of the network with different downsampler in the blind loss. Regardless of downsampling operations, models trained with small strides show poor performance, which is consistent with the result of Figure 1. Space2batch, with a stride of 5, achieves the highest PSNR and SSIM compared to the other two downsamplers. Therefore, we employ S2B as the downsampling function for the blind loss.

4. More Visualized Results

We present more visual comparisons on SIDD [1] validation and NIND [2]. We compare C-BSN with other self-supervised methods, CVF-SID (T) [4], CVF-SID (S\(^2\)), AP-BSN [3], AP-BSN (R\(^3\)) [3], which aim to remove real-world noise. We use official code from the authors’ GitHub with the pre-trained model. The denoised results of various scenes are illustrated in Figure 2.

For NIND, we use C-BSN\(^\dagger\) which is trained on the test set directly. Figure 3 shows the noisy images from NIND and its denoised outputs. We mark ROI with red boxes for each image and present noisy-denoised pairs of cropped patches.
References


Figure 2. Visual comparison of denoised images on SIDD validation [1]. We provide PSNR and SSIM in the upper left of the images. All images are upsampled by 2 with the nearest neighbor for better comparison. Best viewed in pdf.
Figure 3. C-BSN$^3$ results of NIND [2] samples. (Left) Real noisy images from NIND. (Right) Enlarged noisy-Denoised image pairs.