Supplementary Material: Self-supervised Image Denoising with Downsampled Invariance Loss and Conditional Blind-Spot Network

Yeong Il Jang[1] Keuntek Lee[1] Gu Yong Park[1] Seyun Kim[2] Nam Ik Cho[1,3]

[1]Department of ECE, INMC, Seoul National University, Seoul, Korea

[2]Gauss Labs Inc.

[3]IPAI, Seoul National University, Seoul, Korea

{jyicu, leekt000, pgy9134}@snu.ac.kr, seyun.kim@gausslabs.ai, nicho@snu.ac.kr

1. Detailed Proof of Downsampled Invariance Loss

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

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left\|f(\boldsymbol{x})-\boldsymbol{y}\right\|^2 + \|\boldsymbol{x}-\boldsymbol{y}\|^2 \leq \mathbb{E}_{\boldsymbol{x}}\left\|f(\boldsymbol{x})-\boldsymbol{x}\right\|^2 + 2\sqrt{ms^2}\underset{d_s(\boldsymbol{x})}{\mathbb{E}}[\mathbb{E}\left\|d_s(f(\boldsymbol{x}))-f_M(d_s(\boldsymbol{x}))\right\|^2]^{\frac{1}{2}}. \quad (1)$$

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

$$\mathbb{E}_{\boldsymbol{x}}\,||f(\boldsymbol{x})-\boldsymbol{x}||^2 = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\,||f(\boldsymbol{x})-\boldsymbol{y}||^2 + ||\boldsymbol{x}-\boldsymbol{y}||^2 - 2\langle f(\boldsymbol{x})-\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}\rangle. \quad (2)$$

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

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\langle f(\boldsymbol{x})-\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}\rangle \quad (3)$$
$$= \mathbb{E}_{\boldsymbol{y}}\,\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\sum_j (f(\boldsymbol{x})_j - y_j)(x_j - y_j) \quad (4)$$
$$= \sum_j \mathbb{E}_{\boldsymbol{y}}[\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(f(\boldsymbol{x})_j - y_j)(x_j-y_j) - \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(f(\boldsymbol{x})_j - y_j)\,\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(x_j-y_j)] \quad (5)$$
$$= \sum_j \mathbb{E}_{\boldsymbol{y}}[\mathrm{Cov}(f(\boldsymbol{x})_j - y_j, x_j - y_j|\boldsymbol{y})] \quad (6)$$
$$= \sum_j \mathbb{E}_{\boldsymbol{y}}[\mathrm{Cov}(f(\boldsymbol{x})_j, x_j|\boldsymbol{y})]. \quad (7)$$

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

$$\sum_j \mathbb{E}_{\boldsymbol{y}}[\mathrm{Cov}(f(\boldsymbol{x})_j, x_j|\boldsymbol{y})] = \frac{m}{|J|}\,\mathbb{E}_J\sum_j \mathbb{E}_{\boldsymbol{y}}[\mathrm{Cov}(f(\boldsymbol{x})_j, x_j|\boldsymbol{y})], \quad (8)$$

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

$$\frac{1}{|J|}\sum_{j\in J}\mathbb{E}_{\boldsymbol{y}}[\mathrm{Cov}(f(\boldsymbol{x})_j, x_j|\boldsymbol{y})] \quad (9)$$
$$= \frac{1}{|J|}\sum_{j\in J}\mathbb{E}_{\boldsymbol{y}}[\mathrm{Cov}(f(\boldsymbol{x})_j - f_M(d_s(\boldsymbol{x}))_j, x_j|\boldsymbol{y})] \quad (10)$$
$$\leq \frac{1}{|J|}\sum_{j\in J}(\mathbb{E}_{\boldsymbol{y}}[\mathrm{Var}(f(\boldsymbol{x})_j - f_M(d_s(\boldsymbol{x}))_j|\boldsymbol{y})^{\frac{1}{2}} \cdot \mathrm{Var}(x_j|\boldsymbol{y})^{\frac{1}{2}}]) \quad (11)$$
$$\leq (\frac{1}{|J|}\sum_{j\in J}\mathbb{E}_{\boldsymbol{y}}[\mathrm{Var}(f(\boldsymbol{x})_j - f_M(d_s(\boldsymbol{x}))_j|\boldsymbol{y})\cdot \mathrm{Var}(x_j|\boldsymbol{y})])^{\frac{1}{2}} \quad (12)$$
$$\leq (\frac{1}{|J|}\sum_{j\in J}\mathbb{E}_{\boldsymbol{y}}[E[(f(\boldsymbol{x})_j - f_M(d_s(\boldsymbol{x}))_j)^2|\boldsymbol{y}]]\cdot 1)^{\frac{1}{2}}] \quad (13)$$
$$= (\frac{1}{|J|}\sum_{j\in J}\mathbb{E}[(f(\boldsymbol{x})_j - f_M(d_s(\boldsymbol{x}))_j)^2])^{\frac{1}{2}} \quad (14)$$
$$= (\frac{s^2}{m}\,\mathbb{E}[(d_s(f(\boldsymbol{x})) - f_M(d_s(\boldsymbol{x})))^2])^{\frac{1}{2}} \quad (15)$$

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 $\mathrm{Var}(x) \leq E[x^2]$, and by the assumption that input $\boldsymbol{x}$ is normalized *i.e.*, $\mathrm{Var}(x_j|\boldsymbol{y}) \leq \mathrm{Var}(x_j) = 1$. $\square$

By the Proposition 1, we use Eq. (15) as downsampled invariance loss,

$$\mathcal{L}_{inv} = \sqrt{\frac{s^2}{m}} ||d_s(f(\boldsymbol{x})) - sg(f_M(d_s(\boldsymbol{x})))||_2, \quad (16)$$

where sg is the stop gradient operation. $f_M(d_s(\boldsymbol{x}))$ 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 downsampled invariance loss as

$$\mathcal{L}_{inv} = \frac{s^2}{m} ||d_s(f(\boldsymbol{x})) - sg(f_M(d_s(\boldsymbol{x})))||_1. \quad (17)$$

2. Analysis of Downsampling Ratio in Loss Functions

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

Using strides less than 4 in the blind loss leads to sub-optimal 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),

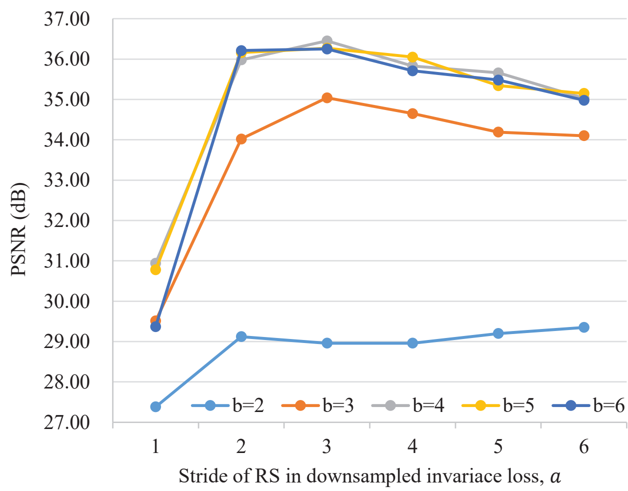


Figure 1. **PSNR of C-BSN$_{a/b}$ on SIDD validation [1],** where a denotes the stride of RS and b denotes the stride of S2B.

Table 1. **Ablation on the downsampler of blind Loss.**

downsampler	stride	PSNR(dB)	SSIM
PD	5	34.83	0.912
	2	29.11	0.715
$S2B$	5	**36.22**	**0.935**
	2	25.93	0.810
RS	5	35.67	0.924
	2	30.54	0.771

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† 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

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018. 2, 4

[2] Benoit Brummer and Christophe De Vleeschouwer. Natural image noise dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 5

[3] Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17725–17734, 2022. 2

[4] Reyhaneh Neshatavar, Mohsen Yavartanoo, Sanghyun Son, and Kyoung Mu Lee. Cvf-sid: Cyclic multi-variate function for self-supervised image denoising by disentangling noise from image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17583–17591, 2022. 2

[5] Yaochen Xie, Zhengyang Wang, and Shuiwang Ji. Noise2same: Optimizing a self-supervised bound for image denoising. *Advances in Neural Information Processing Systems*, 33:20320–20330, 2020. 1

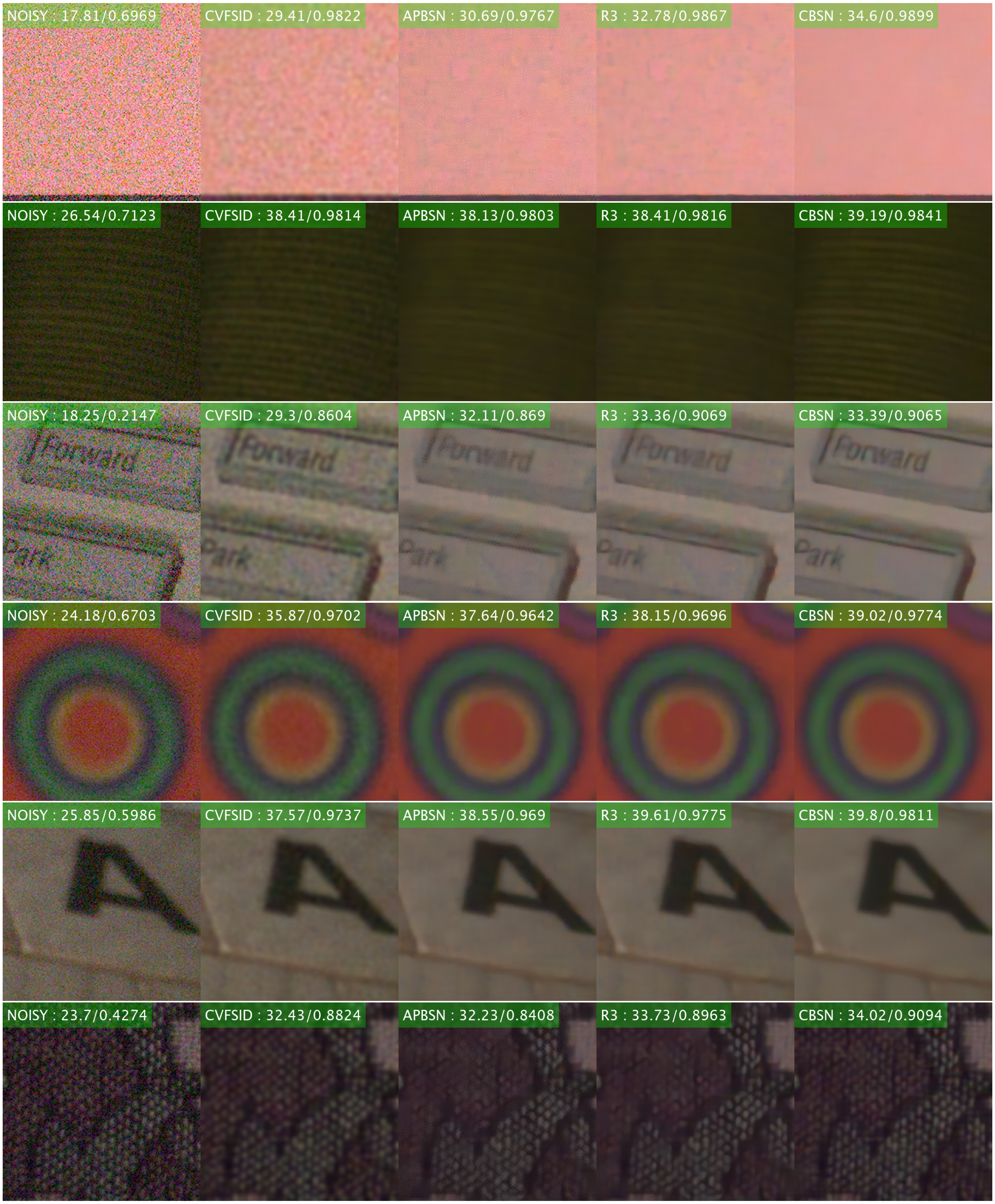
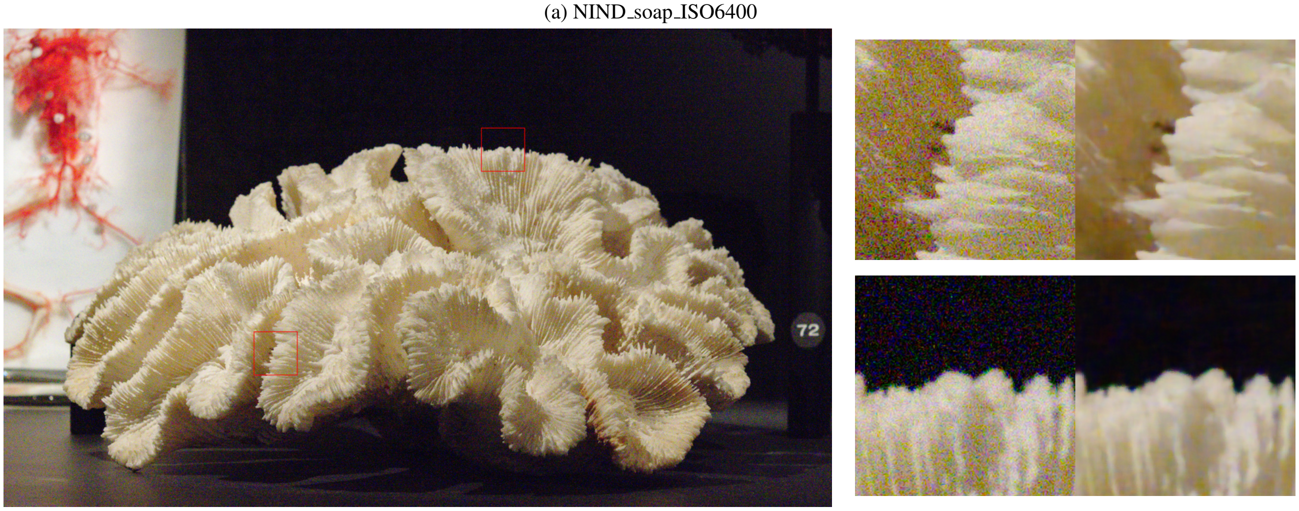


Figure 2. **Visual comparison of denoised images on SIDD validation [1].** We provide PSNR and SSIM in the upper left of the images. All images are upsampled by 2 with the nearest neighbor for better comparison. Best viewed in pdf.

(a) NIND_soap_ISO6400



(b) NIND_MuseeL-coral2_ISOH1

(c) NIND_MVB-LouveFire_ISOH1

Figure 3. **C-BSN† results of NIND [2] samples.** (Left) Real noisy images from NIND. (Right) Enlarged noisy-Denoised image pairs.