Self-Supervised Image Prior Learning with GMM from a Single Noisy Image Supplementary Material

Haosen Liu[1,2], Xuan Liu[1], Jiangbo Lu[2], Shan Tan[1*]
[1]Huazhong University of Science and Technology, [2]SmartMore Corporation
{haosen.liu0803, jiangbo.lu}@gmail.com, {liuxuan99, shantan}@hust.edu.cn

1. Proof of Eq. (15)

The $\mathbf{D}_k$-related terms in 'M-Step' is a quadratic optimization problem with orthogonality constraints, which holds the form as

$$\begin{aligned} \min_{\mathbf{X}} \quad & \mathrm{Tr}\left(\mathbf{W}\mathbf{X}^T\mathbf{A}\mathbf{X}\right) \\ s.t. \quad & \mathbf{X}^T\mathbf{X} = \mathbf{I}, \end{aligned} \tag{S1}$$

where $\mathbf{W}$ is a positive diagonal matrix, $\mathbf{A}$ is a positive definite matrix that can be decomposed as

$$\mathbf{A} = \mathbf{D}\mathbf{\Lambda}\mathbf{D}^T, \tag{S2}$$

where $\mathbf{D}$ is an orthogonal matrix and $\mathbf{\Lambda}$ denotes a diagonal matrix with eigenvalues $\{\lambda_s\}_{s=1}^S$ as its diagonal elements. Without loss of generality, we assume that eigenvalues are sorted as $\lambda_1 > \lambda_2 > ... > \lambda_S > 0$ and that diagonal elements of $\mathbf{W}$ are sorted as $0 < w_1 < w_2 < ... < w_S$.

The Lemma 1 in [8] indicates that the local minimizer of the problem presented in Eq. (S1) has to satisfy the orthogonality constraint $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ and the optimal condition

$$\mathbf{G}\mathbf{X}^T - \mathbf{X}\mathbf{G}^T = \mathbf{0}, \tag{S3}$$

where $\mathbf{G}$ is the gradient of the objective function in Eq. (S1) with respect to $\mathbf{X}$. Based on this Lemma, it can be proved that $\mathbf{X} = \mathbf{D}$ is the solution to the problem shown in Eq. (S1).

Proof. For the objective function shown in Eq. (S1), its gradient with respect to $\mathbf{X}$ is

$$\mathbf{G} = 2\mathbf{A}\mathbf{X}\mathbf{W}. \tag{S4}$$

Besides, for any matrix $\mathbf{X}$, it can be written as

$$\mathbf{X} = \mathbf{D}\mathbf{M}. \tag{S5}$$

Substituting Eqs. (S2), (S4)-(S5) into the optimal condition Eq. (S3), one can obtain

$$\mathbf{\Lambda}\mathbf{M}\mathbf{W}\mathbf{M}^T = \mathbf{M}\mathbf{W}\mathbf{M}^T\mathbf{\Lambda}. \tag{S6}$$

If we denote $\mathbf{B} = \mathbf{M}\mathbf{W}\mathbf{M}^T$, $\mathbf{B}$ has to be a diagonal matrix to satisfy Eq. (S6) since $\mathbf{\Lambda}$ is a diagonal matrix with different diagonal elements.

On the other hand, if $\mathbf{X}$ satisfies the orthogonality constraint $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, the matrix $\mathbf{M}$ will also be an orthogonal matrix. Therefore, $\mathbf{M}\mathbf{W}\mathbf{M}^T$ can be regarded as the eigenvalue decomposition of the diagonal matrix $\mathbf{B}$. Note that, the eigenvectors of a diagonal matrix with different diagonal elements have to be $\{\mathbf{e}_s\}_{s=1}^S$, where $\mathbf{e}_s$ denotes the vector whose s-th element is 1 and other elements are 0. That is to say, for any matrix $\mathbf{X}$ satisfying the orthogonality constraint and the optimal condition, its corresponding matrix $\mathbf{M}$ must hold the form as

$$\mathbf{M} = [\mathbf{e}_{p_1}, \mathbf{e}_{p_2}, ...\mathbf{e}_{p_S}], \tag{S7}$$

where p_s is the s-th element of a vector $\mathbf{p}$ consisting of any arrangement from 1 to S.

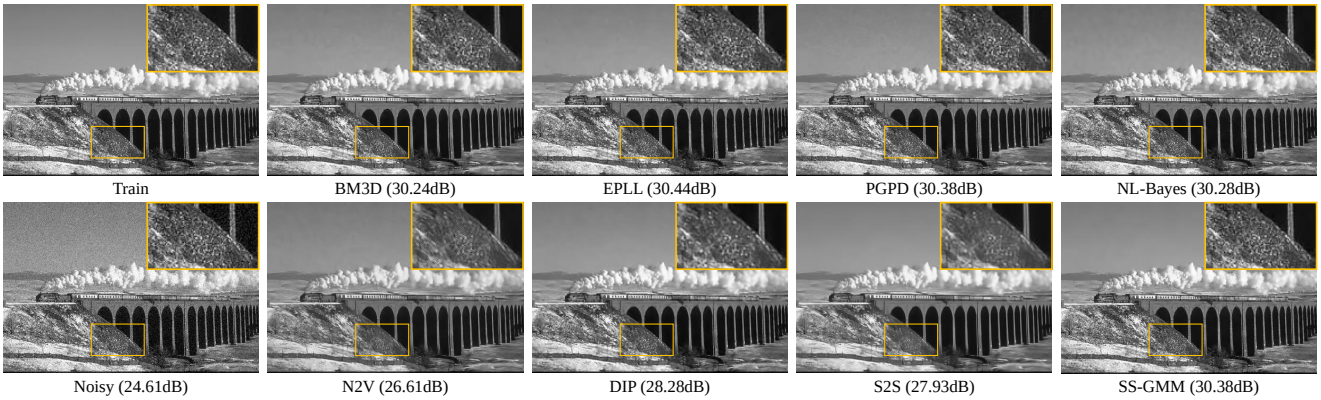
Furthermore, substituting Eqs. (S2), (S5), (S7) into Eq. (S1), the original optimization problem can be converted as

$$\min_{\mathbf{p}} \quad \sum_{s=1}^{S} w_s \lambda_{p_s}. \tag{S8}$$

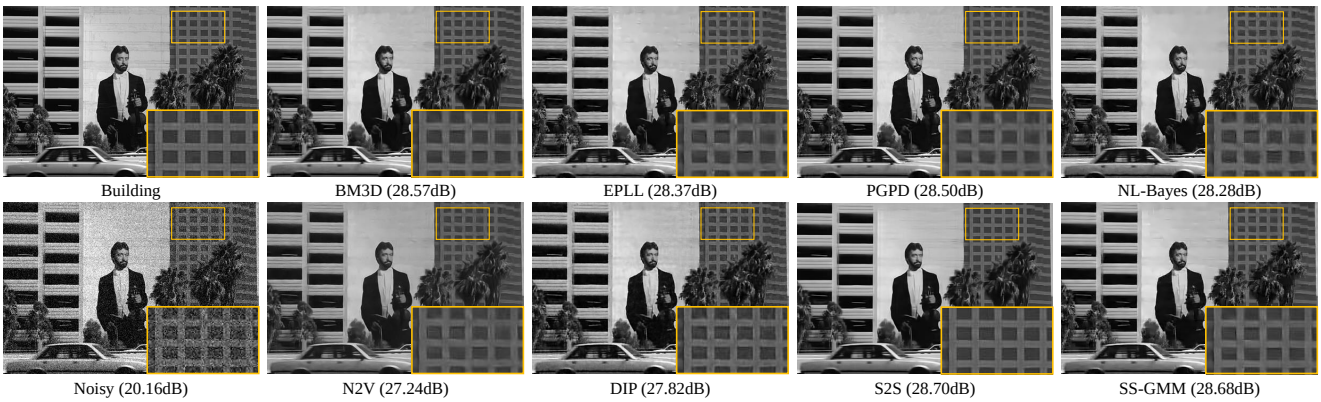
As one can see, this is essentially a sorting problem aimed at finding a specific order of $\{\lambda_s\}_{s=1}^S$ that holds the minimum objective function value. Since there are $\lambda_1 > \lambda_2 > ... > \lambda_S > 0$ and $0 < w_1 < w_2 < ... < w_S$, it can be easily observed that the solution to Eq. (S8) is $\mathbf{p} = [1, 2, ...S]$. In this case, $\mathbf{M}$ is the identity function and thus there is $\mathbf{X} = \mathbf{D}$. The proof is done. □

It is worth noting that $\mathbf{X} = \mathbf{D}$ is still the minimizer of Eq. (S1) when the assumption is relaxed to $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_S > 0$ and $0 < w_1 \leq w_2 \leq ... \leq w_S$. The corresponding proof can be derived with a similar procedure as the above proof.

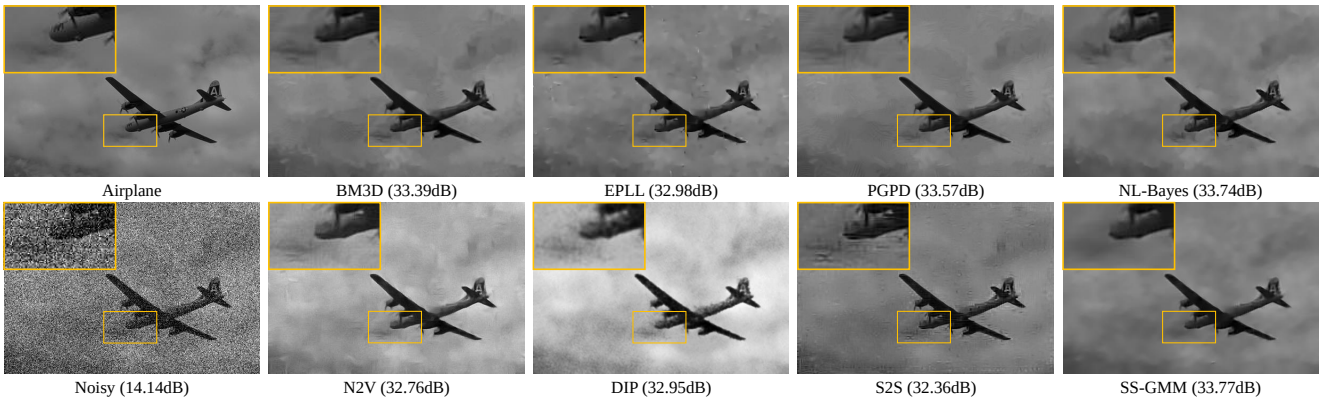
*Corresponding author. This work was supported in part by the National Natural Science Foundation of China (NNSFC), under Grant Nos. 61672253 and 62071197. Part of this work was done when Haosen was interning in SmartMore Co., Ltd.



(a) Visual results of comparison methods on image 'Train' of BSD68 with the noise level of $\sigma = 15$.



(b) Visual results of comparison methods on image 'Building' of BSD68 with the noise level of $\sigma = 25$.



(c) Visual results of comparison methods on image 'Airplane' of BSD68 with the noise level of $\sigma = 50$.

Figure S1. Visual comparisons of different image denoising algorithms.

2. Experimental Results

2.1. Implementation of Comparison Methods

The comparison methods include a) the non-learning method: BM3D [2]; b) GMM-related methods: EPLL [10], PGPD [9] and NL-Bayes [4]; c) self-supervised deep learning methods: N2V [3], DIP [7] and S2S [6]. All of these methods are implemented with the source codes and/or trained models provided by their authors. For BM3D and NL-Bayes, all of the default settings are adopted. For EPLL and PGPD, the models trained by their authors are adopted. As [9, 10] indicates, the training data used for the trained EPLL and PGPD are from [5] and the Kodak PhotoCD Dataset (http://r0k.us/graphics/kodak/), respectively. Since

BM3D, EPLL, PGPD and NL-Bayes require the noise level as an extra input, the true noise level is provided to them in advance. For N2V, it is trained by us with the source code on 400 images of the size 180×180 from [1]. A similar setting is adopted in [6] for N2V. For DIP, its iteration number is tuned by hand to achieve the best performance. For each test image, the network is at first trained for 3000 iterations. Then, the best iteration number for each noisy image is selected within 3000 iterations by using the corresponding clean image as guidance. For S2S, it is trained on each noisy image with the default hyper-parameters.

2.2. Implementation of determining L

Given the estimated noise level σ, L is determined by Eq. (19). At first, we sort all of $\tilde{\lambda}_{ks}^{E}$ in ascending order. In this case, calculating Eq. (19) is equivalent to averaging the first $KS-L$ elements in this ordered sequence. Then, we decrease L from $KS-1$ till the σ calculated by Eq. (19) is equal or close enough to that estimated as Sec. 3.1. The whole process is efficient, only taking about 50 milliseconds.

2.3. More Results on Image Denoising

In Fig. S1, the visual results of comparison methods on three images 'Train', 'Building' and 'Airplane' with different noise levels are provided.

2.4. More Analysis

Validation of the noise estimation module. In Tab. S1, we provide a comparison of our proposed method to its modified version that is provided with the true noise level. As Tab. S1 shows, these two versions perform similarly on Set12 at all of the three noise levels. This demonstrates the effectiveness of the self-contained noise estimator of our proposed method.

Table S1. Image denoising results w/o the true noise level.

Dataset	σ known?	$\sigma=15$	$\sigma=25$	$\sigma=50$
Set12	yes	32.19	29.70	26.40
	no	32.18	29.68	26.38

Validation of the determination of constraint level L. In Fig. S2, we plot the PSNR results versus the levels of sparsity constraint on two images 'cameraman' and 'house' of noise level $\sigma=25$. The level of sparsity constraint is calculated as $1-L/(K\cdot S)$, where L is the parameter determining the constraint level, K is the number of Gaussian components and S is the number of eigenvalues in each component.

As Fig. S2 shows, the PSNR results peak at different constraint levels for images 'cameraman' and 'house'. This demonstrates the necessity of choosing different L for different images. Our proposed method provides a solution to the adaptive determination of L. To demonstrate its effectiveness, we plot the PSNR results of our proposed method in Fig. S2 with dashed lines. As one can see, the dashed lines are respectively around the peak regions of the other two curves. This indicates that our proposed method successfully selects the optimal L for both 'cameraman' and 'house'.

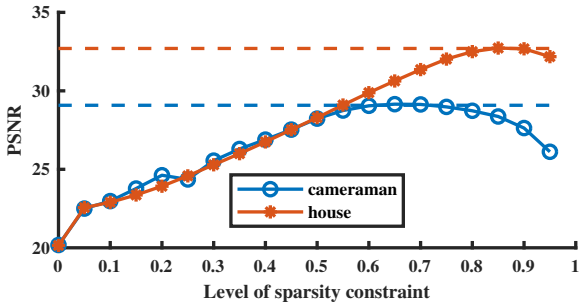


Figure S2. Image denoising results on 'house' and 'cameraman' of noise level $\sigma=25$ with different levels of sparsity constraint. The dashed lines are the PSNR results achieved by our proposed way of determining L.

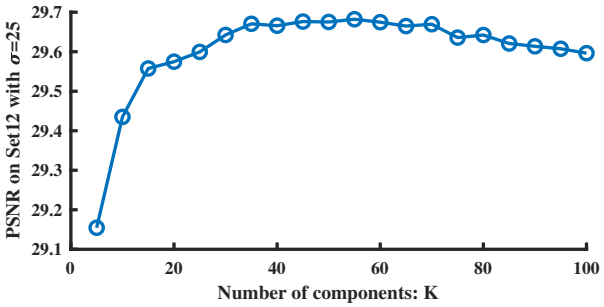


Figure S3. Image denoising results on Set12 of noise level $\sigma=25$ with different numbers of Gaussian components K.

Selection of the number of Gaussian components K. In Fig. S3, we plot the average PSNR results on Set12 of noise level $\sigma=25$ versus the numbers of Gaussian components K. As Fig. S3 shows, when K increases from 5 to 100, the average PSNR result increases at first and then decreases. This is the result of two factors. On the one hand, when K increases, the model capacity of the GMM increases so that it can potentially lead to better image denoising results. On the other hand, when K is too large, the effective number N_k of each component tends to be smaller and smaller. This will finally violate the assumption adopted by us that N_k is large enough. As a result, the highest PSNR occurs at a reasonable K instead of the largest K. This means that K is an empirical parameter. Fortunately, we observe that K is not as sensitive as L to the image content. Therefore, we can determine it with only a few images and then apply it to other cases.

Selection of parameter r. In Fig. S4, we plot the av-

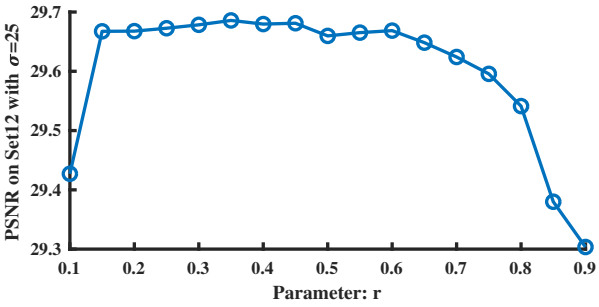


Figure S4. Image denoising results on Set12 of noise level $\sigma = 25$ with different parameters r.

erage PSNR results on Set12 of noise level $\sigma = 25$ versus the parameter r. As shown in Fig. S4, when r changes from 0.3 to 0.7, the average PSNR varies in a small range (29.62dB$\sim$29.68dB), indicating insensitivity of ours to choice of r. In this paper, we choose r for all test cases.

Analysis on estimation error at small eigenvalues. In the Fig. 5 of the main body, we compare histograms of eigenvalues learned by our proposed SS-GMM algorithm and those learned from the clean/noisy 'couple' with the EM-GMM algorithm. As this figure shows, the histogram corresponding to SS-GMM coincides with the curve 'Clean + EM' at most places. They are different only at small eigenvalues. A key question is *how important these small eigenvalues are?* To explore this question, we manually set small eigenvalues ($\lambda < 0.25\sigma^2$) of the GMM trained by EPLL [10] to 0. In this case, the modified model can still achieve 31.84dB, which is similar to the original model (31.83dB), on Set12 when $\sigma = 15$. This demonstrates that the error at small eigenvalues is not important since it will only incur a very minor difference.

References

[1] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2016. 3

[2] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 2

[3] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2Void-Learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2129–2137, 2019. 2

[4] Marc Lebrun, Antoni Buades, and Jean-Michel Morel. A nonlocal Bayesian image denoising algorithm. *SIAM Journal on Imaging Sciences*, 6(3):1665–1688, 2013. 2

[5] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001. 2

[6] Yuhui Quan, Mingqin Chen, Tongyao Pang, and Hui Ji. Self2Self with dropout: Learning self-supervised denoising from single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1890–1898, 2020. 2, 3

[7] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 2

[8] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434, 2013. 1

[9] Jun Xu, Lei Zhang, Wangmeng Zuo, David Zhang, and Xiangchu Feng. Patch group based nonlocal self-similarity prior learning for image denoising. In *Proceedings of the IEEE international conference on computer vision*, pages 244–252, 2015. 2

[10] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *2011 International Conference on Computer Vision*, pages 479–486. IEEE, 2011. 2, 4