ANALYSIS DICTIONARY LEARNING BASED ON NESTEROV'S GRADIENT WITH APPLICATION TO SAR IMAGE DESPECKLING

Jing Dong *Wenwu Wang*

Centre for Vision, Speech and Signal Processing, University of Surrey, UK
Email: j.dong@surrey.ac.uk w.wang@surrey.ac.uk

ABSTRACT

We focus on the dictionary learning problem for the analysis model. A simple but effective algorithm based on Nesterov's gradient is proposed. This algorithm assumes that the analysis dictionary contains unit ℓ_2 norm atoms and trains the dictionary iteratively with Nesterov's gradient. We show that our proposed algorithm is able to learn the dictionary effectively with experiments on synthetic data. We also present examples demonstrating the promising performance of our algorithm in despeckling synthetic aperture radar (SAR) images.

Index Terms— Analysis model, analysis dictionary learning, Nesterov's gradient

1. INTRODUCTION

Dictionary design is an important problem in sparse representation. Recent studies have shown that dictionaries learned from a set of training signals have the potential to fit the signals better than the analytical dictionaries [1], [2]. Many dictionary learning algorithms such as MOD [3], K-SVD [1], and SimCO [4] are established on the synthesis model, where a signal is represented as a linear combination of a few atoms (signal components) from the dictionary. There is an alternative model in sparse representation – analysis model where the product of the dictionary and the signal is sparse. Learning dictionaries for the analysis model, however, has received less attention, with only a few activities emerging recently, such as [2], [5], [6], [7]. These algorithms introduce various constraints on the learning process which make them quite complicated.

In this paper, we focus on the analysis dictionary learning (ADL) problem and propose a new algorithm based on Nesterov's gradient method [8], which is very simple compared with the existing ADL algorithms, such as [2], [6]. Nesterov's gradient method improves the gradient-based methods by achieving the optimal convergence rate $O(\frac{1}{t^2})$ with t being the iteration number, and has been applied to many problems, such as nonnegative matrix factorization [9] and image deblurring [10].

We introduce our proposed algorithm in Section 2. Simulations presented in Section 3 demonstrate the performance of the proposed algorithm on recovering the ground-truth dictionary with synthetic data and despeckling SAR images. Finally, Section 4 concludes the paper.

2. THE PROPOSED ALGORITHM

Given a set of training signals $\mathbf{Y} \in \mathbb{R}^{m\times n}$, ADL aims to learn a dictionary $\mathbf{\Omega} \in \mathbb{R}^{p\times m}$ for which the analysis representations of $\mathbf{Y}$ are sparse. This problem can be cast as

$$\min_{\mathbf{X},\mathbf{\Omega}} \|\mathbf{X}-\mathbf{\Omega}\mathbf{Y}\|_F^2 \quad \text{s.t.} \quad \forall i,\ \|\mathbf{X}_{:,i}\|_0 = p-l \tag{1}$$

where $\mathbf{X}_{:,i}$ is the ith column of $\mathbf{X} \in \mathbb{R}^{p\times n}$. The ℓ_0 quasi-norm $\|\cdot\|_0$ counts the number of non-zeros of a vector and l is the cosparsity. However, the problem (1) has many trivial solutions, for example, $\mathbf{\Omega} = \mathbf{0}$. In order to exclude such trivial solutions, we assume the rows of $\mathbf{\Omega}$ have unit ℓ_2-norm. In this case, the ADL problem can be rewritten as

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{\Omega}} \|\mathbf{X}-\mathbf{\Omega}\mathbf{Y}\|_F^2 \\ \text{s.t.} \quad &\forall i,\ \|\mathbf{X}_{:,i}\|_0 = p-l, \\ &\forall j,\ \|\mathbf{\Omega}_{j,:}\|_2 = 1 \end{aligned} \tag{2}$$

where $\mathbf{\Omega}_{j,:}$ denotes the jth row of $\mathbf{\Omega}$.

Most ADL algorithms alternate between two stages: analysis sparse coding and dictionary update, initializing with a random and normalized $\mathbf{\Omega}$. The first stage calculates $\mathbf{X}$ for the given dictionary $\mathbf{\Omega}$. In the dictionary update stage, $\mathbf{\Omega}$ is updated assuming known and fixed $\mathbf{X}$. We also apply this framework in our proposed algorithm. Nesterov's gradient method [8] is applied in the dictionary update stage, and thus the algorithm is referred to as NeADL. The procedure of the NeADL algorithm is presented in Algorithm 1. The details are given in the following subsections.

2.1. Analysis Sparse Coding

The goal of analysis sparse coding is to obtain the sparse representations of the training signals based on a given dictionary. The exact representations $\mathbf{X}$ can be calculated directly by simply multiplying the signals by the dictionary, that is

$$\mathbf{X} = \mathbf{\Omega}\mathbf{Y} \tag{3}$$

Algorithm 1 NeADL

Input: $\mathbf{Y}$, p, l
Output: $\hat{\mathbf{\Omega}} = \mathbf{\Omega}^{k+1}$
Initialization:
Initialize the iteration counter $k = 1$ and the analysis dictionary $\mathbf{\Omega}^k$. Perform the following steps.
Main Iterations:
1. Calculate $\mathbf{X}^k$ according to equation (3): $\mathbf{X}^k = \mathbf{\Omega}^k\mathbf{Y}$
2. Apply hard thresholding operation according to equation (4): $\hat{\mathbf{X}}^k = HT_l(\mathbf{X}^k)$
3. Update the analysis dictionary according to Algorithm 2: $\mathbf{\Omega}^{k+1} \leftarrow \mathbf{\Omega}^k$
4. Increase the iteration counter: $k = k + 1$
5. If the stopping criterion is satisfied, quit the iteration. Otherwise, go back to step 1.

However, the representations obtained by equation (3) may not be sparse since the initial dictionary is an arbitrary one.

A hard thresholding operation is therefore applied to enforce the cosparsity

$$\hat{\mathbf{X}} = HT_l(\mathbf{X}) \tag{4}$$

where $HT_l(\mathbf{X})$ is an operator that sets the smallest l elements (in magnitude) of each column of $\mathbf{X}$ to 0. In doing so, the sparsity constraint can be enforced.

2.2. Dictionary Update

The dictionary update stage aims to find the solution of the constrained optimisation problem

$$\min_{\mathbf{\Omega}} \|\mathbf{X} - \mathbf{\Omega}\mathbf{Y}\|_F^2 \quad \text{s.t.} \quad \forall j, \|\mathbf{\Omega}_{j,:}\|_2 = 1. \tag{5}$$

We can solve the following unconstrained optimization problem

$$\min_{\mathbf{\Omega}} \|\mathbf{X} - \mathbf{\Omega}\mathbf{Y}\|_F^2 \tag{6}$$

and then normalize the rows of the optimal solution $\mathbf{\Omega}$ to address problem (5). In the following subsections, we prove that problem (6) satisfies the pre-conditions for Nesterov's gradient method [8] to be applied, and then we give the algorithm of the dictionary update stage in detail.

2.2.1. Convexity

Denote $f(\mathbf{\Omega}) = \|\mathbf{X} - \mathbf{\Omega}\mathbf{Y}\|_F^2$ defined on $\mathbb{R}^{p\times m}$. Obviously, the domain of f is a convex set. Given any two matrices $\mathbf{\Omega}_1$, $\mathbf{\Omega}_2 \in \mathbb{R}^{p\times m}$ and $\theta \in \mathbb{R}$ with $0 \le \theta \le 1$, we have

$$\begin{aligned}
&f(\theta\mathbf{\Omega}_1 + (1-\theta)\mathbf{\Omega}_2) - \theta f(\mathbf{\Omega}_1) - (1-\theta) f(\mathbf{\Omega}_2) \\
&= \|\mathbf{X} - (\theta\mathbf{\Omega}_1 + (1-\theta)\mathbf{\Omega}_2)\mathbf{Y}\|_F^2 \\
&\quad - \theta\|\mathbf{X} - \mathbf{\Omega}_1\mathbf{Y}\|_F^2 - (1-\theta)\|\mathbf{X} - \mathbf{\Omega}_2\mathbf{Y}\|_F^2 \\
&= \theta(\theta-1) tr\big((\mathbf{\Omega}_1\mathbf{Y} - \mathbf{\Omega}_2\mathbf{Y})^T(\mathbf{\Omega}_1\mathbf{Y} - \mathbf{\Omega}_2\mathbf{Y})\big) \\
&= \theta(\theta-1)\|\mathbf{\Omega}_1\mathbf{Y} - \mathbf{\Omega}_2\mathbf{Y}\|_F^2 \\
&\le 0
\end{aligned}$$

Therefore, we have

$$f(\theta\mathbf{\Omega}_1 + (1-\theta)\mathbf{\Omega}_2) \le \theta f(\mathbf{\Omega}_1) + (1-\theta) f(\mathbf{\Omega}_2) \tag{7}$$

In other words, $f(\mathbf{\Omega})$ is a convex function.

2.2.2. Lipschitz continuity and Lipschitz constant

The gradient of the function f is

$$\nabla_{\mathbf{\Omega}} f(\mathbf{\Omega}) = -2(\mathbf{X} - \mathbf{\Omega}\mathbf{Y})\mathbf{Y}^T \tag{8}$$

For any matrices $\mathbf{\Omega}_1, \mathbf{\Omega}_2 \in \mathbb{R}^{p\times m}$, we have

$$\begin{aligned}
&\|\nabla_{\mathbf{\Omega}} f(\mathbf{\Omega}_1) - \nabla_{\mathbf{\Omega}} f(\mathbf{\Omega}_2)\|_F^2 \\
&= \| -2(\mathbf{X} - \mathbf{\Omega}_1\mathbf{Y})\mathbf{Y}^T + 2(\mathbf{X} - \mathbf{\Omega}_2\mathbf{Y})\mathbf{Y}^T\|_F^2 \\
&= 4\|(\mathbf{\Omega}_1 - \mathbf{\Omega}_2)\mathbf{Y}\mathbf{Y}^T\|_F^2 \\
&= 4\|(\mathbf{\Omega}_1 - \mathbf{\Omega}_2)\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\|_F^2 \\
&= 4tr\Big(\big((\mathbf{\Omega}_1 - \mathbf{\Omega}_2)\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\big)^T\big((\mathbf{\Omega}_1 - \mathbf{\Omega}_2)\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\big)\Big) \\
&= 4tr\big(\mathbf{U}^T(\mathbf{\Omega}_1 - \mathbf{\Omega}_2)^T(\mathbf{\Omega}_1 - \mathbf{\Omega}_2)\mathbf{U}\mathbf{\Sigma}^2\big) \\
&\le 4\sigma_1^2 tr\big(\mathbf{U}^T(\mathbf{\Omega}_1 - \mathbf{\Omega}_2)^T(\mathbf{\Omega}_1 - \mathbf{\Omega}_2)\mathbf{U}\big) \\
&= 4\sigma_1^2 tr\big((\mathbf{\Omega}_1 - \mathbf{\Omega}_2)^T(\mathbf{\Omega}_1 - \mathbf{\Omega}_2)\mathbf{U}\mathbf{U}^T\big) \\
&= 4\sigma_1^2\|\mathbf{\Omega}_1 - \mathbf{\Omega}_2\|_F^2
\end{aligned} \tag{9}$$

where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the SVD decomposition of $\mathbf{Y}\mathbf{Y}^T$ and σ_1 is the largest singular value, i.e. $\|\mathbf{Y}\mathbf{Y}^T\|_2$. Therefore, we have

$$\|\nabla_{\mathbf{\Omega}} f(\mathbf{\Omega}_1) - \nabla_{\mathbf{\Omega}} f(\mathbf{\Omega}_2)\|_F^2 \le 4\sigma_1^2\|\mathbf{\Omega}_1 - \mathbf{\Omega}_2\|_F^2, \tag{10}$$

implying that $\nabla f(\mathbf{\Omega})$ is Lipschitz continuous and the Lipschitz constant is $2\sigma_1$, i.e. $L = 2\|\mathbf{Y}\mathbf{Y}^T\|_2$ [8].

2.2.3. Dictionary update algorithm

As the optimization problem (6) is convex and the Lipschitz constant is $L = 2\|\mathbf{Y}\mathbf{Y}^T\|_2$, Nesterov's gradient method can be applied to solve the unconstrained problem [8]. The algorithm of the dictionary update stage using Nesterov's gradient method is presented in Algorithm 2. Notice that $\nabla_{\mathbf{\Omega}} f(\mathbf{Z}_t)$ in Step 1 is calculated by substituting $\mathbf{\Omega}$ with $\mathbf{Z}_t$ into equation (8), i.e. $\nabla_{\mathbf{\Omega}} f(\mathbf{Z}_t) = -2(\mathbf{X} - \mathbf{Z}_t\mathbf{Y})\mathbf{Y}^T$.

3. SIMULATION RESULTS

We illustrate the ability of the NeADL algorithm for recovering the ground-truth dictionary with some synthetic data and also present the application of the algorithm in SAR image despeckling.

3.1. Experiments With Synthetic Data

In this experiment, we randomly generated a dictionary and a set of training data based on this dictionary with fixed cosparsity level l. The entries of the ground-truth analysis dictionary

Algorithm 2 Dictioanry Update using Nesterov's gradient method

Input: $\mathbf{\Omega}^k$, $\mathbf{Y}$, $\hat{\mathbf{X}}^k$
Output: $\mathbf{\Omega}^{k+1} = \mathbf{\Omega}_t$
Initialization:
Initialize $\mathbf{Z}_1 = \mathbf{\Omega}_0 = \mathbf{\Omega}^k$, $\alpha_1 = 1$, $L = 2\|\mathbf{Y}\mathbf{Y}^T\|_2$, $t = 1$. Perform the following steps.
Main Iterations:
1. $\mathbf{\Omega}_t = \mathbf{Z}_t - \frac{1}{L}\nabla_{\mathbf{\Omega}} f(\mathbf{Z}_t)$.
2. $\alpha_{t+1} = \frac{1+\sqrt{4\alpha_t^2+1}}{2}$.
3. $\mathbf{Z}_{t+1} = \mathbf{\Omega}_t + \frac{\alpha_t - 1}{\alpha_{t+1}}(\mathbf{\Omega}_t - \mathbf{\Omega}_{t-1})$.
4. If the stopping criterion is satisfied, go to step 6. Otherwise, increase the iteration counter: $t = t + 1$ and go back to step 1.
6. Normalize the rows of $\mathbf{\Omega}_t$ and then $\mathbf{\Omega}^{k+1} = \mathbf{\Omega}_t$

$\mathbf{\Omega} \in \mathbb{R}^{p\times m}$ were generated from a Gaussian distribution with zero mean and unit variance and the rows of $\mathbf{\Omega}$ were normalized. Then the training signals were generated based on $\mathbf{\Omega}$ with a fixed cosparsity. We used the parameters $p = 20$, $m = 15$, $l = 10$, $n = 30000$. The NeADL algorithm was performed for 300 iterations (counted by k) with 100 iterations (counted by t) of Nesterov's gradient method for the dicitionary update in each iteration.

Following the experiments in [2], we used the percentage of the recovered atoms to measure the performance of the algorithm for recovering the ground-truth dictionary. A row $\boldsymbol{\omega}_j$ in the true dictionary $\mathbf{\Omega}$ is regarded as recovered if [2]

$$\min_i (1 - |\hat{\boldsymbol{\omega}}_i \boldsymbol{\omega}_j^T|) < 0.01 \tag{11}$$

where $\hat{\boldsymbol{\omega}}_i$ are the atoms of the trained dictionary.

In addition, we introduce an operator $\|\mathbf{x}\|_0^\epsilon$ to count the number of the elements in $\mathbf{x} \in \mathbb{R}^p$ that are below the threshold ϵ, i.e.

$$\|\mathbf{x}\|_0^\epsilon = \#\{i : |x_i| < \epsilon, i = 1, 2, ..., p\} \tag{12}$$

where x_i denotes the ith element of $\mathbf{x}$. We can obtain the cosparsities of the training signals in the trained dictionary by applying this operator to their representation vectors. Here we use the threshold $\epsilon = 0.01$. The average cosparsity of all the training signals can be used to measure the quality of the trained dictionary, as the final goal of ADL is to find a dictionary that can sparsely represent the training signals.

The percentage of the recovered atoms and the average cosparsity over iterations (counted by k) are plotted in Fig. 1. The NeADL algorithm can recover the ground-truth dictionary with a high recovery rate (over 75% after 300 iterations). The trained dictionary is able to represent the training samples with an average cosparsity which is close to the reference cosparsity, showing that the dictionary trained by NeADL is effective at representing the training data sparsely.

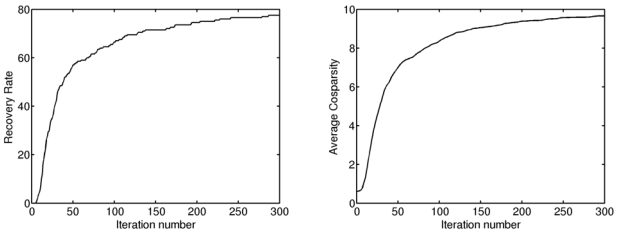

Fig. 1. Recovery rate (left) and average cosparsity (right).

Table 1. The despeckling results.

Image name	Metric	Speckle	NeADL	K-SVD
San Francisco	PSNR	16.32	**21.16**	20.56
	MSSIM	0.18	**0.57**	**0.57**
Stockton	PSNR	12.21	19.60	**19.77**
	MSSIM	0.05	0.41	**0.43**

3.2. SAR Image Despeckling

Speckle removal is a major problem in the analysis of SAR images. We applied our proposed algorithm to the SAR image despeckling problem. We used the images 'San Francisco', and 'Stockton' from the paper [11]. For each speckle image, the NeADL algorithm was employed to learn a dictionary with size 63×49 from 7×7 patches extracted from the image. The cosparsity was set as $l = 42$. The despeckled images were reconstructed with the trained analysis dictionaries using the analysis pursuit algorithm OBG [2]. We compared our method with K-SVD denoising algorithm [12] [1].

The performances of the dspeckling results are quantified by the peak signal-to-noise ratio (PSNR) and the Mean Structural SIMilarity Index (MSSIM) [13] between the original and the despeckled images. Here we use the same set of parameters as originally suggested in [13] to compute the MSSIM. The despeckled results averaged over 5 independent tests are summarized in Table 1. The speckle images and the despeckled images obtained by NeADL and K-SVD in one test are presented in Fig. 2. The results of these two algorithms are very similar in general, as can be observed from Fig. 2. For the image 'San Francisco', the PSNR of the image despeckled by NeADL is higher than that of K-SVD, and the MSSIMs of these two despeckled images are equal. For the image 'Stockton', the despeckling results of K-SVD is slightly better than the NeADL algorithm in both PSNR and MSSIM. The results confirm that the NeADL algorithm is suitable for despeckling SAR images.

[1] For K-SVD denoising, we used the default parameters of the code downloaded from the website `http://www.cs.technion.ac.il/~elad/software/`

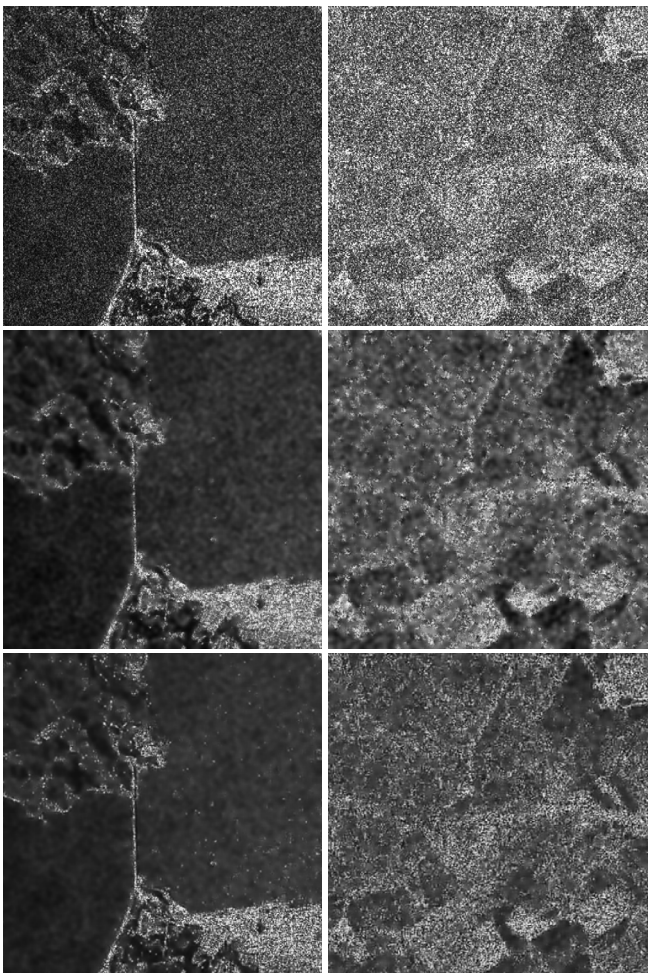

Fig. 2. San Francisco (left column) and Stockton (right column). Speckle images (top), despeckled images by NeADL (middle) and despeckled images by K-SVD (bottom).

4. CONCLUSION

We introduced a new analysis dictionary learning algorithm based on Nesterov's gradient method. The dictionary learning process is formulated as an optimisation problem with the cosparsity and unit ℓ_2-norm constraints of the atoms in the dictionary. The algorithm iteratively solves this problem by thresholding and Nesterov's gradient method followed by the normalization of the rows of the dictionary. Experimental results on synthetic data demonstrated the promising performance in recovery rate and average cosparsity. The algorithm is shown to work well for the despeckling of SAR images.

5. ACKNOWLEDGEMENT

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K014307/1 and the MOD University Defence Research Collaboration in Signal Processing. We thank Prof. Jonathon Chambers for proofreading the manuscript.

6. REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[2] R. Rubinstein, T. Peleg, and M. Elad, "Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model," *IEEE Trans. Signal Process.*, vol. 61, no. 3, pp. 661–677, 2013.

[3] K. Engan, S.O. Aase, and J.H. Hakon, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, 1999, vol. 5, pp. 2443–2446.

[4] W. Dai, T. Xu, and W. Wang, "Simultaneous codeword optimization (SimCO) for dictionary update and learning," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6340–6353, 2012.

[5] M. Yaghoobi, S. Nam, R. Gribonval, and M.E. Davies, "Constrained overcomplete analysis operator learning for cosparse signal modelling," *IEEE Trans. Signal Process.*, vol. 61, no. 9, pp. 2341–2355, 2013.

[6] S. Hawe, M. Kleinsteuber, and K. Diepold, "Analysis operator learning and its application to image reconstruction," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2138–2150, 2013.

[7] S. Ravishankar and Y. Bresler, "Learning overcomplete sparsifying transforms for signal processing," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, 2013, pp. 3088–3092.

[8] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Mathematics Doklady*, vol. 27, no. 2, pp. 372–376, 1983.

[9] N. Guan, D. Tao, Z. Luo, and B. Yuan, "NeNMF: an optimal gradient method for nonnegative matrix factorization," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2882–2898, 2012.

[10] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[11] A. Lapini, T. Bianchi, F. Argenti, and L. Alparone, "Blind speckle decorrelation for sar image despeckling," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1044–1058, 2014.

[12] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Signal Process.*, vol. 15, no. 12, pp. 3736–3745, 2006.

[13] Z. Wang, A. C Bovik, H. R Sheikh, and E. P Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.