VECTOR NONLOCAL EUCLIDEAN MEDIAN: FIBER BUNDLE CAPTURES THE NATURE OF PATCH SPACE

CHEN-YUN LIN, ARIN MINASIAN, XIN JESSICA QI, AND HAU-TIENG WU

ABSTRACT. We extensively study the rotational group structure inside the patch space by introducing the fiber bundle structure. The rotational group structure leads to a new image denoising algorithm called the *vector non-local Euclidean median* (VNLEM). The theoretical aspect of VNLEM is studied, which explains why the VNLEM and the traditional non-local mean/non-local Euclidean median (NLEM) algorithm work. The numerical issue of the VNLEM is improved by taking the orientation feature in the commonly applied scale-invariant feature transform (SIFT), and a theoretical analysis of the robustness of the orientation feature in the SIFT is provided. The VNLEM is applied to an image database of 1,361 images and a comparison with the NLEM is provided. Different image quality assessments based on the error-sensitivity or the human visual system are applied to evaluate the performance. The results confirmed the potential of the VNLEM algorithm.

1. INTRODUCTION

Image denoising is a long-lasting challenge in the image processing field. Much effort has been invested in this problem in the past decades. While there are several approaches to handle this problem, like the variational approach, the wavelet approach, the partial differential equation approach, etc (see [1, 21] for an overall survey), we focus on the idea of nonlocal filtering and its associated theoretical analysis in this paper.

Based on the idea that pixels spatially far apart in an image can be similar or even the same, Buades et al. pioneered the nonlocal mean (NLM) filters [2] to denoise a noisy image. The motivation for the NLM could be summarized by taking the patch space into account [23, 4, 31, 32, 34, 6, 7, 30, 29, 45]. For the i-th pixel on a given image I, we could associate it with a patch P_i of size $q \times q$, where q is the patch size determined by the user. Mathematically, a patch is the restriction of the image on a subset around i, so that the i-th pixel of I is the central pixel of P_i. Denote the set of all patches as $\mathcal{X}_I$. The main assumption is that $\mathcal{X}_I$ is located on, or could be well approximated by, a low dimensional geometric object, like a manifold. By viewing a $q \times q$ patch as a q^2-dim vector, the manfiold is a subset of the Euclidean space $\mathbb{R}^{q^2}$, and we could endow an induced Riemannian metric on the manifold from $\mathbb{R}^{q^2}$. Under this assumption, two patches with a similar intensity, while they might be far apart in the image, are close in the intrinsic geometry of the manifold.

Under this low-dimensional and nonlinear patch space structure, the NLM algorithm is introduced [1, 34]. In brief, in the NLM algorithm, to denoise a given pixel i, we find the $m \in \mathbb{N}$ nearest neighboring patches of the patch P_i, denoted as $\mathcal{N}_i$, with respect to the Euclidean distance and then denoise the i-th pixel by evaluating the mean of all central pixels of those patches in $\mathcal{N}_i$. It has been well known that the NLM algorithm leads to better edge preservation [21], and this improvement is directly related to the diffusion process on the nonlinear geometric structure [34]. The NLM algorithm can be understood as reducing the noise influence on the patch space via the diffusion process [34]. Several generalizations of NLM follow based on this diffusion idea. By noting that the mean operator is sensitive to the outliers, the authors in [6] considered replacing the mean in the NLM by the median, which leads to the nonlocal Euclidean median (NLEM). In brief, after finding the neighbors of P_i, the i-th pixel is denoised by evaluating the median of all central pixels of those patches in $\mathcal{N}_i$. It is shown in [6] that the NLEM could tolerate more noises inside the noisy patches. This idea has been applied to the single-channel blind source separation problem to reconstruct the "wave-shape function" and decompose the fetal electrocardiogram signal from the maternal abdominal electrocardiogram signal [37]. Furthermore, by noticing that the mean operator is equivalent to minimizing a functional based on the L^2 norm and the median operator is equivalent to minimizing a functional based on the L^1 norm, and by the need of enhancing the sparsity structure, the nonlocal patch regression (NLPR) is considered in [7], which replaces the L^1 norm in the associated functional by the L^p

norm, where $0 < p < 1$. We mention that the above model and algorithm have been applied to different fields, like the medical imaging problem [5] and the inpainting problem [18, 29, 45, 48].

As successful as the patch space model and those diffusion-based algorithms are, there are structures in the patch space that we can consider to further improve the algorithm and theoretical problems we need to answer. From the model perspective, there are structures in the patch space not considered in the past, particularly the rotational group structure. Since the central pixel of a patch is fixed after rotation, two patches could be viewed the same, or *rotationally invariant*, if they are the same up to a rotation. Therefore, in the patch space model, we could take the rotational group into account. From the theoretical viewpoint, to the best of our knowledge, a study explaining why neighbors could be well approximated from the noisy patches in NLM/NLEM/NLPR was not available. Also, a discussion and explanation of how the patch size should be chosen is lacking. Furthermore, in the literature, the denoising performance is commonly evaluated by the "error-based" measurements, like the signal-to-noise ratio (SNR) or the peak SNR, but it has been well known that those error-based quantities might capture only partial information of the image quality, and more needs to be considered.

In this paper, we aim to advance the progress on these problems. We take the rotational group into account, and model the patch space by a fiber bundle. In this model, the set of rotationally invariant patches is modeled by a fiber that is diffeomorphic to $SO(2)$, and the collection of the set of rotationally invariant patches (or the orbits coming from the $SO(2)$ action), denoted as $\mathcal{X}_I/SO(2)$, is parametrized by a manifold, which is the base manifold of the fiber bundle. We then generalize the NLM/NLEM/NLPR algorithm by taking the fiber bundle structure into account, and call the new algorithm the vector nonlocal Euclidiean median (VNLEM). In the VNLEM, the rotationally invariant distance (RID) associated with the fiber bundle structure is considered so that two rotationally invariant patches will have RID equal to 0. With the RID, we could determine the neighboring patches, and then evaluate the median value of the central pixels of all neighboring patches. Note that this leads to a dimensional reduction of the patch space, since we work with a 1-dim lower base manifold. Hence, we get more samples for the denoise purpose in the VNLEM, when compared with the NLM/NLEM/NLPR. From the theoretical perspective, we study how accurate we could estimate the neighborhood from the noisy patches, and provide a quantification in Theorem 11. In brief, we show that with high probability, which depends on the patch size and the noise level, we could accurately determine the neighborhood from the noisy patches under the RID or the Euclidean distance. By noting that the probability we could determine the correct neighbors depends on the patch size, we could explain why the patch space approach leads to a better denoising result compared with the pixel-based NLM or NLEM algorithm. On the other hand, we also discuss that the patch size cannot be too large, or the patch space will be too "complicated" so that the diffusion algorithm might fail. From the algorithmic perspective, we need to handle the numerical problem for the VNLEM. Note that with the RID distance, the base manifold is no longer embedded in the ordinary Euclidean space, and the ordinary fast nearest neighbor search algorithms cannot be applied. As far as we know, there is no fast algorithm available to determine the neighbors under the RID metric. Our solution is via a relaxation step. We consider the commonly applied scale-invariant feature transform (SIFT) features to estimate candidates for the neighbors. Then we run the RID to determine the true neighbors. To guarantee the applicability of this relaxation, in addition to discussing the relationship between the RID distance and the neighbors determined by the SIFT features, we show that the SIFT features are robust to noise. Finally, in addition to the ordinary error-based measurements, we consider image quality measurements that take the human visual system into account to evaluate the performance of the proposed VNLEM algorithm. The result is reported on a large scale image database consisting of 1,361 images.

The paper is organized as the following. In Section , we introduce a rotationally invariant distance since the patch space allows a canonical rotation action. Then, we propose a principal bundle model for the patch space of an image and provide both continuous and discrete versions of our model. In Section , we give the VNLEM algorithm and discuss how we deal with numerical issues to make the algorithm computationally affordable. One main step is to use orientations in the SIFT algorithm to approximate rotation angles between patches. In Section , we show that with high probability, we could accurately determine nearest neighbors of a clean patch through finding nearest neighbors of the associated noisy patch in the noisy patch space. In Section , we provide a mathematical definition of the orientation feature in the SIFT and show why such approximations are reliable when patches are noisy. The performance evaluation measurements are summarized in Section . In Section , we show our numerical results. In Appendix , we give a brief review

of diffusion geometry which is used for dimension reduction to help find good nearest neighbors for image denoising. In Appendix , we discuss a possible model under which we could approximate a patch space by a manifold .

TABLE 1. Table of commonly used notation throughout the paper

$\|\cdot\|$	ℓ^2 norm
$\Phi_t^{(m)}$	DM with the diffusion time $t > 0$ and the first m non-trivial eigenvectors
$D_t^{(m)}(\cdot,\cdot)$	diffusion distance (DD)
$SO(2)$	rotation group
$O \in SO(2)$	rotation matrix
$d_{\mathtt{RID}}(\cdot,\cdot)$	rotationally invariant distance
$\iota^{(\mathtt{c})} : \mathbb{R}^2 \to \mathbb{R}$	continuous clean image
$p_{\mathbf{x}}^{(\mathtt{c})} : \mathbb{R}^2 \to \mathbb{R}$	continuous round clean patch centered at $\mathbf{x} \in \mathbb{R}^2$ with radius r
$\iota^{(\mathtt{n})} : \mathbb{R}^2 \to \mathbb{R}$	continuous noisy image
$p_{\mathbf{x}}^{(\mathtt{n})} : \mathbb{R}^2 \to \mathbb{R}$	continuous round noisy patch centered at $\mathbf{x} \in \mathbb{R}^2$ with radius r
$\mathcal{X}^{(\mathtt{c})}$	the patch space of image $\iota^{(\mathtt{c})}$
$I^{(\mathtt{c})} \in \mathbb{R}^{N\times N}$	discrete clean image of size $N \times N$
$\xi(\cdot)$	$\xi(\cdot) \sim \mathcal{N}(0,1)$ i.i.d. Gaussian white noise
$I^{(\mathtt{n})} = I^{(\mathtt{c})} + \sigma\xi$	discrete noisy image of size $N \times N$
$P_i^{(\mathtt{c})} \in \mathbb{R}^{q\times q}$	clean patch of size $q \times q$ centered at i-pixel
$P_i^{(\mathtt{n})} = P_i^{(\mathtt{c})} + \sigma\xi$	noisy patch of size $q \times q$ centered at i-pixel
$\mathcal{X}_I^{(\mathtt{c})}$	the patch space of image $I^{(\mathtt{c})}$
$\mathcal{X}_I^{(\mathtt{n})}$	the patch space of image $I^{(\mathtt{n})}$
$G(\mathbf{s})$	Gaussian function of scale 1 in the SIFT algorithm
$L(\mathbf{x}) = L_p(\mathbf{x})$	Gaussian smoothed patch of patch p
$L^{(\mathtt{c})}$	Gaussian smoothed clean patch
$L^{(\mathtt{n})}$	Gaussian smoothed noisy patch
$\theta^{(c)*}$	orientation assignment of a clean patch
$\theta^{(n)*}$	orientation assignment of a noisy patch
$C_c^\infty(\mathbb{R}^2)$	the space of C^∞ functions with compact supports defined on $\mathbb{R}^2$
$\mathcal{D}'(\mathbb{R}^2)$	the space of distributions defined on $\mathbb{R}^2$

2. Mathematical Model

We start from recalling the patch space commonly used in the image processing field [23, 11, 4, 31, 32, 34, 6, 7, 30, 29, 45]. Take a grayscale image I of size N-pixels wide and N-pixels long, which is a real function defined on $\mathbb{Z}_N^2$, where $\mathbb{Z}_N = \{1, 2, \ldots, N\}$. Usually we represent I as a $N \times N$ matrix with real entries. In general, we could consider an image with different width and length, but to simplify the discussion, we limit our focus on square images in this paper. Call a point $(\alpha, \beta) \in \mathbb{Z}_N^2$ the (α, β)-th pixel of the image, and $I(i, j)$ is the intensity of the grayscale image I. Take an odd natural number q. For each pixel $(\alpha, \beta) \in \mathbb{Z}_N^2$, we associate it with a patch $P_{(\alpha,\beta)} \in \mathbb{R}^{q\times q}$, which is defined as

$$P_{(\alpha,\beta)}(k,l) := \begin{cases} I((\alpha+k-(q-1)/2, \beta+l-(q-1)/2)) & \text{when } ((\alpha+k-(q-1)/2, \beta+l-(q-1)/2)) \in \mathbb{Z}_N^2 \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

for $k, l = 1, \ldots, q$. Specifically, we use the notation $P_{(\alpha,\beta)}(c)$ to indicate the central point of the (α, β)-th patch, $P_{(\alpha,\beta)}((q+1)/2, (q+1)/2)$.

To express the notation in a compact format, we stack the columns of the matrix I into a vector $I^\vee \in \mathbb{R}^{N^2}$, where the superscript $\vee$ means the *vector form*, that is, the $((\ell-1)N+1)$-th to the (ℓN)-th entries in $I^\vee$ is the ℓ-th column of I, where $\ell = 1, \ldots, N$. Similarly, denote $P^\vee_{(\alpha,\beta)} \in \mathbb{R}^{q^2}$ to be the vector form of the (α,β)-th patch. When there is no danger of confusion, we ignore the superscript $\vee$ and use the notation I to represent the grayscale image in the matrix form and in the column form interchangeably, and denote $P_i := P^\vee_{(\alpha,\beta)} \in \mathbb{R}^{q^2}$, where $i = (\alpha-1)q + \beta$. We have the following definition for the discretized patch space. For a given grayscale image $I \in \mathbb{R}^{N\times N}$ and the patch size q, where q is an odd integer number, the discretized patch space is defined as

$$\mathcal{X}_I := \{P_i\}_{i=1}^{N^2} \subset \mathbb{R}^{q^2}. \tag{2}$$

Besides the general consensus that we could approximate the patch space by the manifold model [23, 31, 32, 34, 6, 7, 29, 45], the structure of the patch space is less discussed, except the discussion of the Klein bottle in [4, 30]. Inspired by the fact that two image patches might be the same up to a rotation, in [33], the rotation structure naturally existed in the patch space was taken into account. In addition to [33], the same orientation idea was considered in [49, 19, 20, 35, 46], and has been applied to the neuroimaging analysis [28, 20].

In this section, we introduce a fiber bundle structure for the patch space. To make clear how to incorporate the $SO(2)$ group into the model, and how to numerically rotate a patch, we start from a continuous setup. We define the continuous patch space as a topological space with the $SO(2)$ group structure. Then we discuss how to obtain the discrete model from a continuous one. Recall the definition of a fiber bundle with the group structure [12].

Definition 1 (Fiber bundle with group structure G)**.** *Let F and M be manifolds. A fiber bundle E with fiber F over M consists of a topological space E together with a map $\pi : E :\to M$ satisfying the local triviality condition. Let G be a Lie group, for example the rotation group $SO(2)$. Let the map $. : G \times F \to F$ be a smooth left action of F. That is, the map (the action of G) $. : (x, g) \mapsto g.x$ from $G \times F$ to F satisfies*

$$e.g = g \text{ for all } g \in G, \tag{3}$$

where e is the neutral element of G and

$$(gh).x = g.(h.x) \text{ for all } x \in M \text{ and } g, h \in G. \tag{4}$$

This group G is often called the gauge group or the transition group. If the fiber is equal to the structure group, then (π, E, M, G) is called a principal G bundle. (See, for example, Definition 5.7 in [16]). This definition is equivalent to the definition of principal bundle requiring G as a right action on E. (See, for example, Propositions 5.5 and 5.6 in [16]).

Let $\iota^{(c)} : \mathbb{R}^2 \to \mathbb{R}$ be a L^2 function that represents an image. The round patches of a finite radius from the image $\iota^{(c)}$ is defined below.

Definition 2 (Continuous patch and continuous patch space)**.** *Fix a L^2 image $\iota^{(c)}$ and $r > 0$. A patch centered at $\mathbf{x} \in \mathbb{R}^2$, denoted as $p^{(c)}_{\mathbf{x}}$, is defined as*

$$p^{(c)}_{\mathbf{x}}(\mathbf{s}) = \iota^{(c)}(\mathbf{x}+\mathbf{s})\psi(\mathbf{s}), \tag{5}$$

where $\mathbf{s} \in \mathbb{R}^2$ and $\psi \in C^\infty_c$ defined as

$$\psi(\mathbf{s}) = \begin{cases} 1 & \text{when } \|\mathbf{s}\|_{\mathbb{R}^2} \le 3r/2 \\ 0 & \text{when } \|\mathbf{s}\|_{\mathbb{R}^2} > 2r \end{cases}. \tag{6}$$

The (continuous) patch space associated with $\iota^{(c)}$ is denoted as

$$\mathcal{X}^{(c)} := \{p^{(c)}_{\mathbf{x}} : \mathbf{x} \in \mathbb{R}^2\} \subset L^2(\mathbb{R}^2). \tag{7}$$

The superscript "(c)" in the definition indicates that the image is clean. In the literature, it is common to assume that the patch space $\mathcal{X}^{(c)}$ is located on, or could be approximated by, a low dimensional manifold [23, 34, 6, 30, 29]. We will make this assumption in this paper. To further capture the structure of the patch

space, that is, two image patches might be the same up to a rotation, we could consider a $SO(2)$ action on the patch space. For $O \in SO(2)$ and $\mathbf{s} \in \mathbb{R}^2$ expressed by a column vector, we define the action on $p_{\mathbf{x}}$ as

$$(O.p_{\mathbf{x}}^{(\mathtt{c})})(\mathbf{s}) = p_{\mathbf{x}}^{(\mathtt{c})}(O^{-1}\mathbf{s}). \tag{8}$$

This is a left group action since for any $O_1, O_2 \in SO(2)$,

$$O_2.(O_1.p_{\mathbf{x}}^{(\mathtt{c})})(\mathbf{s}) = O_1.p_{\mathbf{x}}^{(\mathtt{c})}(O_2^{-1}\mathbf{s}) = p_{\mathbf{x}}^{(\mathtt{c})}(O_1^{-1}O_2^{-1}\mathbf{s}) = (O_2O_1).p_{\mathbf{x}}^{(\mathtt{c})}(\mathbf{s}) \tag{9}$$

Since each fiber, including a patch and its rotated patches, can be identified as S^1, and $SO(2)$ is diffeomorphic to S^1, the patch space $\mathcal{X}^{(\mathtt{c})}$ could be viewed as a principal $SO(2)$ bundle.

By identifying patches up to a rotation, we have the quotient space $\mathcal{X}^{(\mathtt{c})}/SO(2)$. We make the following assumption

Assumption 3. *The patch space $\mathcal{X}^{(\mathtt{c})}$ is a subset of the fiber bundle $E = \mathcal{X}^{(\mathtt{c})}$ with $\pi : E \to M$ and a left $SO(2)$ group action so that the quotient space $M := \mathcal{X}^{(\mathtt{c})}/SO(2)$ is a manifold. The $SO(2)$ action preserves the fiber that is diffeomorphic to $SO(2)$.*

We consider the rotationally invariant distance (RID) to measure the similarity between patches.

Definition 4. *Let $p_{\mathbf{x}_1}^{(\mathtt{c})}, p_{\mathbf{x}_2}^{(\mathtt{c})}$ be two patches in $\mathcal{X}^{(\mathtt{c})}$. The rotational invariant distance is defined as*

$$d_{RID}(p_{\mathbf{x}_1}^{(\mathtt{c})}, p_{\mathbf{x}_2}^{(\mathtt{c})}) = \min_{O \in SO(2)} \left(\int_{\mathbb{R}^2} \left| p_{\mathbf{x}_1}^{(\mathtt{c})}(\mathbf{s}) - O.p_{\mathbf{x}_2}^{(\mathtt{c})}(s) \right|^2 d\mathbf{s} \right)^{1/2}. \tag{10}$$

The minimum in the RID could be achieved since $SO(2)$ is compact. Also note that the definition is equivalent to

$$d_{\mathtt{RID}}(p_{\mathbf{x}_1}^{(\mathtt{c})}, p_{\mathbf{x}_2}^{(\mathtt{c})}) = \min_{O_1, O_2 \in SO(2)} \left(\int_{\mathbb{R}^2} \left| O_1.p_{\mathbf{x}_1}^{(\mathtt{c})}(\mathbf{s}) - O_2.p_{\mathbf{x}_2}^{(\mathtt{c})}(\mathbf{s}) \right|^2 d\mathbf{s} \right)^{1/2}, \tag{11}$$

since

$$\begin{aligned} \int_{\mathbb{R}^2} |O_1.p_{\mathbf{x}_1}^{(\mathtt{c})}(\mathbf{s}) - O_2.p_{\mathbf{x}_2}^{(\mathtt{c})}(\mathbf{s})|^2 d\mathbf{s} &= \int_{\mathbb{R}^2} |p_{\mathbf{x}_1}^{(\mathtt{c})}(O_1^{-1}\mathbf{s}) - p_{\mathbf{x}_2}^{(\mathtt{c})}(O_2^{-1}\mathbf{s})|^2 d\mathbf{s} \\ &= \int_{\mathbb{R}^2} |p_{\mathbf{x}_1}^{(\mathtt{c})}(\mathbf{s}) - p_{\mathbf{x}_2}^{(\mathtt{c})}(O_2^{-1}O_1\mathbf{s})|^2 d\mathbf{s} \end{aligned} \tag{12}$$

by a change of variables and the fact that $O_1, O_2 \in SO(2)$.

Consider an isotropic homogeneous generalized Gaussian random field Φ with a finite variance defined on $\mathbb{R}^2$ to model the noise [17, Chapter III.5].[1] The noisy image is defined as

$$\iota^{(\mathtt{n})} = \iota^{(\mathtt{c})} + \Phi \in \mathcal{D}'(\mathbb{R}^2), \tag{13}$$

where we assume that the spectral measure of Φ [17, p.264] is the same as $\sigma^2 d\xi$, $\sigma > 0$, and $d\xi$ is the Lebesgue measure on $\mathbb{R}^2$. The superscript "(n)" in the definition indicates that the image is noisy. By the same way as that in Definition 2, the noisy patch centered as $\mathbf{x}$ is denoted as

$$p_{\mathbf{x}}^{(\mathtt{n})} = p_{\mathbf{x}}^{(\mathtt{c})} + \psi\Phi, \tag{14}$$

and the noisy patch space is denoted as $\mathcal{X}^{(\mathtt{n})}$. Note that since Φ is a generalized random field and the cut-off function ψ in (2) is in $C_c^\infty(\mathbb{R}^2)$, the noisy patches are well-defined. However, in general the RID cannot be defined for two noisy patches in the continuous setup, since the noisy patches are distributions.

[1] Recall the definition of Φ. For any $\phi \in C_c^\infty(\mathbb{R}^2)$, $\Phi(\phi)$ is a random variable with mean 0 and finite variance. For functions $\phi_1(x), \ldots, \phi_m(x) \in C_c^\infty(\mathbb{R}^2)$, any vector $v \in \mathbb{R}^2$, and any rotation or reflection O of $\mathbb{R}^2$, the m-dimensional random variables

$$(\Phi(\phi_1(x)), \ldots, \Phi(\phi_m(x))), \quad (\Phi(\phi_1(x+v)), \ldots, \Phi(\phi_m(x+v))), \text{ and } (\Phi(\phi_1(O.x)), \ldots, \Phi(\phi_m(O.x)))$$

are identically distributed.

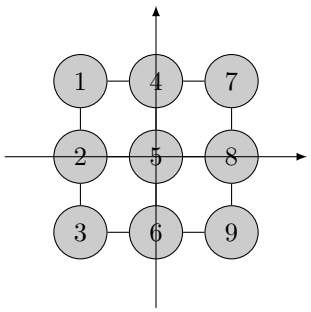


FIGURE 1. Given $q = 3$, we consider a 3×3 grid centered at the origin. Here we denote $\mathbf{x}_a = a$, where $a = 1, \cdots, 9$.

We use a mollifier to create discrete patches from the continuous patches $p_{\mathbf{x}}^{(\mathsf{c})}$ or $p_{\mathbf{x}}^{(\mathsf{n})}$. Note that if an image is regular enough, like continuous, then the discretization could be easily achieved by evaluating the image at the designed grid points. For a L^2 or more general image, however, we need a mollifier to achieve this discretization. Note that we could consider a more general model for an image, like a distribution, but to simplify the discussion we focus on the L^2 image.

First, consider the following discretization map.

Definition 5. *Let η be a mollifier on $\mathbb{R}^2$, that is,*

- *$\eta \in C_c^\infty$ with the unitary L^2 norm;*
- *$\lim_{\epsilon\to 0} \eta^\epsilon(\mathbf{y}) = \delta$ in the weak sense, where $\eta^\epsilon(\mathbf{y}) := \frac{1}{\epsilon^2}\eta(\frac{\mathbf{y}}{\epsilon})$ and δ is the Dirac delta measure.*

For a fixed $\epsilon > 0$ and a set of grid points $\mathcal{G} := \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^2$, we consider a discretization map

$$\begin{aligned} \mathcal{D}_{\mathcal{G}}^\epsilon : L^2(\mathbb{R}^2) &\to \mathbb{R}^n \\ f &\mapsto \begin{bmatrix} f \star \eta_{\mathbf{x}_1}^\epsilon(\mathbf{x}_1) \\ \vdots \\ f \star \eta_{\mathbf{x}_n}^\epsilon(\mathbf{x}_n) \end{bmatrix} \in \mathbb{R}^n, \end{aligned} \tag{15}$$

where $\eta_{\mathbf{x}_i}^\epsilon(\mathbf{y}) := \eta^\epsilon(\mathbf{x}_i - \mathbf{y})$ and $\star$ means the convolution.

In this work, to fulfill the conventional definition of a discrete patch, we consider the discrete patch to be defined on a square that inscribes the D_r, which is a disk centered at the origin with radius $r > 0$. For a fixed odd integer q, we consider a $q \times q$ square sampling grid $\mathcal{G}_q := \{\mathbf{x}_a\}_{a=1}^{q^2} \subset \mathbb{R}^2$, where $\mathbf{x}_a = \left(\frac{r}{\sqrt{2}} + (\alpha-1)\frac{\sqrt{2}r}{q-1}, -\frac{r}{\sqrt{2}} + (\beta-1)\frac{\sqrt{2}r}{q-1}\right)^T \in \mathbb{R}^2$ and (α, β) is the associated index such that $a = (\alpha-1)q + \beta$. See Figure 1 for example. A discrete patch corresponding to $p_{\mathbf{x}}^{(\mathsf{c})}$ is defined as

$$P_{\mathbf{x}}^{(\mathsf{c})} := \mathcal{D}_{\mathcal{G}_q}^\epsilon p_{\mathbf{x}}^{(\mathsf{c})} \in \mathbb{R}^{q^2}, \tag{16}$$

where ϵ is assumed to be much smaller than $\sqrt{2}r/(q-1)$ and the superscript (c) indicates that the image is clean. Note that the distance between two vertically or horizontally consecutive grid points is $\sqrt{2}r/(q-1)$. See Figure 1 for an example. The discrete grayscale image associated with a continuous L^2 image $\iota^{(\mathsf{c})}$ is

$$I^{(\mathsf{c})} := \mathcal{D}_{\mathcal{G}_N}^\epsilon \iota^{(\mathsf{c})} \in \mathbb{R}^{N^2}, \tag{17}$$

where $\mathcal{G}_N = \{\mathbf{x}_i\}_{i=1}^{N^2}$ denotes a uniform sampling grid.

Putting these definitions together, the i-th patch associated with the discrete grayscale image are related by

$$P_i^{(\mathsf{c})} := \mathcal{D}_{\mathcal{G}_q}^\epsilon p_{\mathbf{x}_i}^{(\mathsf{c})}, \tag{18}$$

and we denote the discrete patch space associated with I by

$$\mathcal{X}_I^{(\mathsf{c})} := \{P_i^{(\mathsf{c})}\}_{i=1}^{N^2}. \tag{19}$$

We would like to define $SO(2)$ actions on a discretized patch. Recall that for a given discretized patch $P_{\mathbf{x}}^{(\mathsf{c})}$, the numerical rotation of $P_{\mathbf{x}}^{(\mathsf{c})}$ by $O \in SO(2)$ is carried out by

$$\mathcal{D}^{\epsilon}_{\mathcal{G}_q}(O.\mathcal{I}P_{\mathbf{x}}^{(\mathsf{c})}) = \mathcal{D}^{\epsilon}_{\mathcal{G}_q}(O.\mathcal{I}\mathcal{D}^{\epsilon}_{\mathcal{G}_q}p_{\mathbf{x}}^{(\mathsf{c})}), \tag{20}$$

where $\mathcal{I}$ is the selected interpolation operator. Here $\mathcal{I}$ could be viewed as a deconvolution operator trying to recover $p_{\mathbf{x}}^{(\mathsf{c})}$ from $\mathcal{D}^{\epsilon}_{\mathcal{G}_q}p_{\mathbf{x}}^{(\mathsf{c})}$. Suppose $\mathcal{I}\mathcal{D}^{\epsilon}_{\mathcal{G}_q}$ is the identity operator, then the numerical rotation of $P_{\mathbf{x}}^{(\mathsf{c})}$ by $O \in SO(2)$ becomes $\mathcal{D}^{\epsilon}_{\mathcal{G}_q}(O.p_{\mathbf{x}}^{(\mathsf{c})})$. In general, however, this is not true, unless the function $\iota^{(\mathsf{c})}$ has a special structure so that we can find $\mathcal{I}$. As we utilize interpolation to rotate a discrete patch, the discrepancy between the numerical rotation of $\mathcal{D}^{\epsilon}_{\mathcal{G}_q}p_{\mathbf{x}}^{(\mathsf{c})}$ and $\mathcal{D}^{\epsilon}_{\mathcal{G}_q}(O.p_{\mathbf{x}}^{(\mathsf{c})})$ depends on the rotational angle, the underlying function, and the selected interpolation algorithm. In other words, the discretization and rotation operations are not interchangeable.

Since the numerical rotation performance is not the main focus of this work, to simplify the discussion, we further assume that the numerical impact of numerical rotation of a discrete patch is negligible, and hence $\mathcal{I}\mathcal{D}^{\epsilon}_{q}$ is the identity operator. Thus, we have the following definition.

Definition 6. *We define the $SO(2)$ group action on a discrete patch $P_{\mathbf{x}}^{(\mathsf{c})}$ by*

$$O.P_{\mathbf{x}}^{(\mathsf{c})} := \mathcal{D}^{\epsilon}_{q}(O.p_{\mathbf{x}}^{(\mathsf{c})}), \text{ for any } O \in SO(2). \tag{21}$$

Next, we discuss the discretization procedure for the noisy patches. Indeed, since the mollifier is a function in C_c^{∞}, the noisy image $\iota^{(\mathsf{n})}$ could be discretized as

$$\begin{aligned} \mathcal{D}^{\epsilon}_{\mathcal{G}_N} : D' &\to \mathbb{R}^{N^2} \\ \iota^{(\mathsf{n})} &\mapsto I^{(\mathsf{n})} := \begin{bmatrix} \iota^{(\mathsf{n})} \star \eta^{\epsilon}_{\mathbf{x}_1}(\mathbf{x}_1) \\ \vdots \\ \iota^{(\mathsf{n})} \star \eta^{\epsilon}_{\mathbf{x}_{N^2}}(\mathbf{x}_{N^2}) \end{bmatrix} = I^{(\mathsf{c})} + \begin{bmatrix} \Phi(\eta^{\epsilon}_{\mathbf{x}_1}) \\ \vdots \\ \Phi(\eta^{\epsilon}_{\mathbf{x}_{N^2}}) \end{bmatrix}, \end{aligned} \tag{22}$$

where $\left[\Phi(\eta^{\epsilon}_{\mathbf{x}_1}) \dots \Phi(\eta^{\epsilon}_{\mathbf{x}_{N^2}})\right]^T$ is a Gaussian random vector by the definition of Φ. Precisely, by the assumption of Φ in (13) and the chosen ϵ in the discretization operator, $\Phi(\eta^{\epsilon}_{\mathbf{x}_i})$ is a Gaussian random variable, $\mathbb{E}\Phi(\eta^{\epsilon}_{\mathbf{x}_i}) = 0$ for $i = 1, \dots, N^2$, and $\Phi(\eta^{\epsilon}_{\mathbf{x}_1}), \dots, \Phi(\eta^{\epsilon}_{\mathbf{x}_{N^2}})$ are uncorrelated and hence independent since we have

$$\begin{aligned} \mathbb{E}[\Phi(\eta^{\epsilon}_{\mathbf{x}_i})\Phi(\eta^{\epsilon}_{\mathbf{x}_j})] &= \int |\hat{\eta^{\epsilon}}(\xi)|^2 e^{i2\pi(\mathbf{x}_i - \mathbf{x}_j)\cdot\xi}\sigma^2 d\xi \\ &= \sigma^2 \int \eta^{\epsilon}(\mathbf{y})\eta^{\epsilon}(\mathbf{x}_i - \mathbf{x}_j - \mathbf{y})d\mathbf{y} = \sigma^2\delta_{ij}, \end{aligned} \tag{23}$$

where δ_{ij} is the Kronecker delta and $i, j = 1, \dots, N^2$. The i-th patch associated with the noisy discrete grayscale image is

$$P_i^{(\mathsf{n})} := \mathcal{D}^{\epsilon}_{\mathcal{G}_q}p_{\mathbf{x}_i}^{(\mathsf{n})} = P_i^{(\mathsf{c})} + \sigma\xi_i, \tag{24}$$

where ξ_i is a Gaussian random vector, and $\xi_i(a) \sim \mathcal{N}(0,1)$ and $\mathbb{E}(\xi_i(a)\xi_i(b)) = \delta_{ab}$ for all $a, b = 1, \dots, q^2$. It is clear that for two non-overlapping patches $P_i^{(\mathsf{n})}$ and $P_j^{(\mathsf{n})}$, the associated noises ξ_i and ξ_j are independent. The discrete patch space associated with the noisy image $I^{(\mathsf{n})}$ is denoted by

$$\mathcal{X}_I^{(\mathsf{n})} := \{P_i^{(\mathsf{n})}\}_{i=1}^{N^2} \subset \mathbb{R}^{q^2}. \tag{25}$$

Again, we assume that the error incurred by the numerical rotation is negligible, and have the following definition for the noisy patches.

Definition 7. *We define the $SO(2)$ group action on a discrete patch $P_i^{(\mathtt{n})}$ by*

$$O.P_i^{(\mathtt{n})} := \mathcal{D}^{\epsilon}_{\mathcal{G}_q}(O.p^{(\mathtt{n})}_{\mathbf{x}_i}), \text{ for any } O \in SO(2). \tag{26}$$

Note that due to the isotropic assumption of the homogeneous random field Φ, the distribution of the noise in a noisy patch is fixed after rotation. The RID in the discrete setup is thus defined as the following.

Definition 8. *The rotation invariant distance between two patches, P_i and P_j, which could be clean or noisy, is defined as*

$$d_{RID}(P_i, P_j) := \min_{O\in SO(2)} \|P_i - O.P_j\|, \tag{27}$$

where $\|\cdot\|$ denotes the ℓ^2 norm.

Before closing this section, we show an example to demonstrate the benefit of introducing the frame bundle structure to the patch space. In Fig. 3, we show the 49 nearest neighbors of the patch P indicated by the white box shown in Fig. 2 determined by the RID and L^2 distances respectively. Clearly, with the RID, more nearest neighbor patches that are similar to P are identified. In other words, we reduce the dimension of the patch space by wiping out the fiber associated with the rotationally invariant patches.

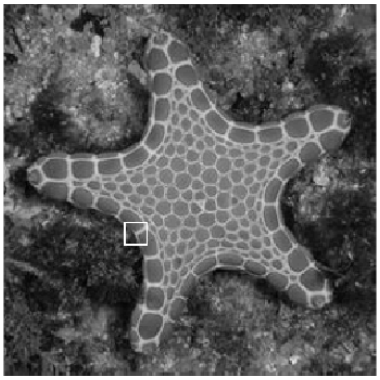


FIGURE 2. Starfish. The selected patch P is indicated by the white box.

3. Vector non-local Euclidean median Algorithm

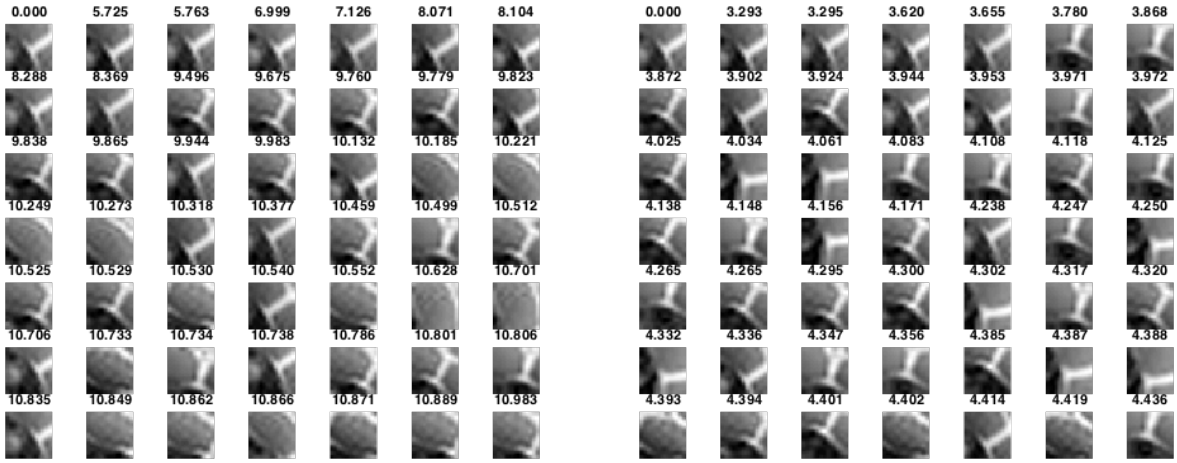
Take a clean grayscale image denoted as $I^{(\mathtt{c})} \in \mathbb{R}^{N\times N}$. Assume that the image has been normalized to be of mean 0 and standard deviation 1; that is,

$$\mu_I := \frac{1}{N^2}\sum_{i=1}^{N^2} I^{(\mathtt{c})}(i) = 0 \quad \text{and} \quad \sigma_I := \left(\frac{1}{N^2}\sum_{i=1}^{N^2}(I^{(\mathtt{c})}(i) - \mu_I)^2\right)^{1/2} = 1. \tag{28}$$

Take the associated noisy image $I^{(\mathtt{n})}$ defined in (22), the noisy patches $P_i^{(\mathtt{n})}$ defined in (24) with $\sigma > 0$, and the noisy patch space defined in (25). We now introduce the VNLEM algorithm.

The goal of the *denoising problem* is finding an algorithm that will recover $I^{(\mathtt{c})}$ from $I^{(\mathtt{n})}$ as accurately as possible. We will come back to the notion of accuracy in Section . In this paper, we consider the following *vector nonlocal Euclidean median* (VNLEM) algorithm, which is a generalization of the NLM, the NLEM [6], and the NLPR [7]. The basic idea is to combine the fiber bundle structure underlying the patch space in order to improve the performance of the NLM and NLEM algorithms. With the RID, define the affinity matrix $W \in \mathbb{R}^{N^2\times N^2}$ by

$$W_{ij} = \exp(-d^2_{\mathtt{RID}}(P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})})/\epsilon), \tag{29}$$



(a) L^2 distance (b) RID

FIGURE 3. Left: the first 49 nearest neighbors of the patch P shown in Figure 2 with respect to the L^2 distance, including P. The patch P is shown in the left top subfigure, and the L^2 distance is shown on the top of each patch. Right: the first 49 nearest neighbors of the patch P with respect to the RID, including P. The patch P is shown in the left top subfigure, and the RID is shown on the top of each patch.

where $i, j = 1, \ldots, N^2$ and $\epsilon > 0$ is the pre-determined bandwidth. For each patch $P_i^{(\mathtt{n})}$, we identify its $N_1 \in \mathbb{N}$ nearest neighbors in the sense of the RID, and we represent this set as $N_{\mathtt{RID}}(i)$. The denoised image, denoted as $\tilde{I}^{(\mathtt{VNLEM})} \in \mathbb{R}^{N\times N}$, is calculated by

$$\tilde{I}^{(\mathtt{VNLEM})}(i) = \tilde{P}_i^{(\mathtt{VNLEM})}(c), \quad \text{where } \tilde{P}_i^{(\mathtt{VNLEM})} := \operatorname*{argmin}_{P\in\mathbb{R}^{q\times q}} \sum_{P_j\in N_{\mathtt{RID}}(i)} W_{ij}\|P - O_{ij}.P_j^{(\mathtt{n})}\|^\gamma, \tag{30}$$

where $i = 1, \ldots, N^2$, $0 < \gamma \leq 1$, and O_{ij} is the rotation that achieves $d^2_{\mathtt{RID}}(P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})})$. Note that when $\gamma = 1$, this is equivalent to taking the median over $\{P_k^{(\mathtt{n})}(c)\}_{k\in N_{\mathtt{RID}}(i)}$. When $0 < \gamma < 1$, this is equivalent to the NLPR proposed in [7]. We call the algorithm VNLEM with $0 < \gamma \leq 1$.

Under the manifold assumption, we could apply the diffusion map (DM) algorithm to further improve the VNLEM algorithm. We summarize the DM and the theory behind it in Appendix . With the affinity matrix W, the graph Laplacian and the associated transition matrix A could be established. By taking the top K eigenvalues and eigenvectors of A, we then embed each patch into a low-dimensional space and calculate the diffusion distance (DD) to evaluate the true neighbors of each patch. For each patch $P_i^{(\mathtt{n})}$, we identify its $N_2 \in \mathbb{N}$ nearest neighbors in the sense of DD. We represent this set of nearest neighbors of $P_i^{(\mathtt{n})}$ as $N_{\mathrm{DD}}(i)$. Based on the robustness property of the DM [13, 15], this step acts as an additional filtering procedure to dismiss the patches in the initial nearest neighbors set determined by the RID. The denoised image, denoted as $\tilde{I}^{(\mathtt{VNLEM\text{-}DD})} \in \mathbb{R}^{N\times N}$, is calculated by

$$\tilde{I}^{(\mathtt{VNLEM\text{-}DD})}(i) = \tilde{P}_i^{(\mathtt{VNLEM\text{-}DD})}(c), \quad \text{where } \tilde{P}_i^{(\mathtt{VNLEM\text{-}DD})} := \operatorname*{argmin}_{P\in\mathbb{R}^{q\times q}} \sum_{P_j^{(\mathtt{n})}\in N_{\mathtt{DD}}(i)} W_{ij}\|P - O_{ij}.P_j^{(\mathtt{n})}\|^\gamma \tag{31}$$

and $i = 1, \ldots, N^2$.

While it seems a straightforward generalization of the NLM/NLEM/NLPR by replacing the Euclidean distance by the RID, there are numerical issues we have to handle, and we discuss three of them below.

First, note that the established W is a dense matrix, which is not feasible to handle when the image size is large. Furthermore, obtaining the pairwise distances between all pairs of patches is a computationally intense and time-consuming task. One practical solution to this issue is to only consider the nearest neighbors of any given patch when forming the affinity matrix. By finding a pre-assigned number of nearest neighbors, we could simultaneously reduce the computational time and the memory required to save W. However, to the best of our knowledge, there is no available efficient nearest neighbor searching algorithm for the RID.

To handle this numerical issue, we consider the *search window* scheme [6] by limiting our algorithm to consider only patches that are within a given search window centered around the reference patch. Precisely, for a patch $P_i^{(\mathtt{n})}$, we consider the search window of size $(2N_2+1)\times(2N_2+1)$ that is centered at the

$$S_i := \{P_j^{(\mathtt{n})} | \, j \in \{1,\dots,N^2\}, \text{ the difference of } i \text{ and } j \text{ is bounded by } N_2 \text{ in both } x \text{ and } y \text{ axes}\}, \tag{32}$$

where $N_2 \in \mathbb{N}$ so that $(2N_2+1)\times(2N_2+1) > N_1$. That is, when we establish W, we only consider the $(2N_2+1)\times(2N_2+1)$ patches whose centers are within N_2 pixels away from the center of $P_i^{(\mathtt{n})}$ in both the x-axis and y-axis. With this search window, we form the affinity matrix by the following:

$$W_{ij} = \begin{cases} \exp(-d^2_{\mathtt{RID}}(P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})})/\epsilon), & \text{for } P_j^{(\mathtt{n})} \in S_i \\ 0, & \text{otherwise} \end{cases}. \tag{33}$$

Note that finding the RID between two given patches incurs huge computational costs in its general form. Also, in general the patch is square and the size is limited, like 11×11 or 13×13, performing a direct numerical rotation might lead to a non-negligible error and deviate the estimated RID. To alleviate these two troubles and facilitate the derivation of the affinity matrix, we use the scale invariant feature transform (SIFT) [27] to approximate the RID. We mention that the central moments are used in [49, 19, 20] and the curvelet transform is used in [46] to capture the rotational feature.

SIFT is an algorithm to extract the local features in an image. The particular feature extracted by the SIFT that we have interest in is the orientation feature. In short, for each pixel, based on the local image gradient direction, an orientation angle is calculated and assigned as the local feature. We will use this feature orientation to approximate the RID distances between the patches. Denote the orientation for the local feature centered in $P_i^{(\mathtt{n})}$ as $\theta_i^{(\mathtt{n})}$. The relative angle between $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$ achieving the RID is approximated by $\theta_i^{(\mathtt{n})} - \theta_j^{(\mathtt{n})}$. We then rotate $P_j^{(\mathtt{n})}$ by $R_{\theta_i^{(\mathtt{n})}}.R^{-1}_{\theta_j^{(\mathtt{n})}}.P_j^{(\mathtt{n})}$, where $R_\theta \in SO(2)$ means the rotation by θ degrees.

Note that the SIFT provides an approximation of the angular relationship between two patches, which allows us to approximate the RID between two patches. However, it might not be accurate. To improve the accuracy, we perform the exhaustive search over a small range centered around the estimated angular relationship $\theta_i^{(\mathtt{n})} - \theta_j^{(\mathtt{n})}$:

$$\theta_{ij} := \underset{\theta\in\{\theta:|\theta-(\theta_i^{(\mathtt{n})}-\theta_j^{(\mathtt{n})})|<\theta_l\}}{\operatorname{argmin}} \|U_k P_i^{(\mathtt{n})} - R_\theta.U_k P_j^{(\mathtt{n})}\|, \tag{34}$$

where $U_k : \mathbb{R}^{q^2} \to \mathbb{R}^{k^2q^2}$ is the chosen upsampling operator that increases the sampling rate of the patch $P_i^{(\mathtt{n})}$ by $k \in \mathbb{N}$ times and $\theta_l > 0$ is the parameter chosen by the user, and hence

$$\tilde{d}_{\mathtt{RID}}(P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})}) := \|P_i^{(\mathtt{n})} - R_{\theta_{ij}}.P_j^{(\mathtt{n})}\|, \tag{35}$$

which is used as an approximation of the RID distance. Note that U_k is applied to improve the accuracy of numerical rotation. To estimate the value of the clean image at pixel i, we use the following affinity weights in (31):

$$W'_{ij} = \exp(-\tilde{d}^2_{\mathtt{RID}}(P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})})/\epsilon), \tag{36}$$

where $j \in S_i$. Note that this method for finding RID offers a trade-off between computational time and accuracy of the result. With the estimated RID and the estimated affinity matrix W', we could run the DM and get the estimated DD.

The proposed algorithms, after taking the above modifications into account, are summarized in Algorithm 1. We call the modified denoising scheme in (30) based on the approximated RID (36) the *VNLEM* algorithm, and call the modified denoising scheme in (31) based on the estimated DD the *VNLEM with DD* (VNLEM-DD). We denote $\tilde{I}^{(\texttt{VNLEM})}$ as the denoised image by VNLEM and $\tilde{I}^{(\texttt{VNLEM-DD})}$ as the denoised image by VNLEM with DD.

Algorithm 1 Vector non-local Euclidean median algorithm.

Input : Noisy image $I^{(\mathtt{n})}$, patch size $q \in \mathbb{N}$, the number of nearest neighbors $N_1 \in \mathbb{N}$, the search window size $N_2 \in \mathbb{N}$, the kernel bandwidth $\epsilon > 0$, the DM embedding dimension $m \in \mathbb{N}$, the diffusion time $t > 0$, and the power $0 < \gamma \leq 1$.
Output : Denoised image $\tilde{I}$.
[pre-1] Pad the image array with a border of $\lceil q/2 \rceil$ pixels.
[pre-2] Create the patch space $\mathcal{X}^{(\mathtt{n})} := \{P_i^{(\mathtt{n})}\}_{i=1}^{N^2} \subset \mathbb{R}^{q^2}$, where the center of $P_i^{(\mathtt{n})}$ is $I^{(\mathtt{n})}(i)$.
[pre-3] Find SIFT orientation feature for each patch and form an affinity matrix W using these orientations from the search window S_i of size $(N_2+1) \times (N_2+1)$ according to equation (33).
[VNLEM. Step 1] For each i, find N_1 nearest neighbours from S_i according to W.
[VNELM. Step 2] Find the more accurate estimation of RID, $\tilde{d}_{\texttt{RID}}$ in (35), and form $N_{\texttt{RID}}(i)$ that contains $\lceil N_1/2 \rceil$ patches that are closer to patch i according to $\tilde{d}_{\texttt{RID}}$, where $\lceil x \rceil$ means the smallest integer greater than or equal to $x \in \mathbb{R}$.
[VNLEM. Step 3] For each i, set $\tilde{I}^{(\texttt{VNLEM})}(i)$ to be the center point of

$$\underset{P \in \mathbb{R}^{q \times q}}{\operatorname{argmin}} \sum_{P_j^{(\mathtt{n})} \in N_{\texttt{RID}}(i)} W'_{ij} \|P - O_{ij}.P_j^{(\mathtt{n})}\|^{\gamma}.$$

[VNLEM-DD. Step 1] Form the eigenvalue decomposition of $D^{-1}W$, where $D \in \mathbb{R}^{N^2 \times N^2}$ is the diagonal matrix determined by $D_{ii} = \sum_{j=1}^{N^2} W_{ij}$.
[VNLEM-DD. Step 2] Embed $P_i^{(\mathtt{n})}$ into $\mathbb{R}^m$ by $\Phi_t^{(m)}(P_i^{(\mathtt{n})}) = (\lambda_2^t \phi_2(i), \ldots, \lambda_{m+1}^t \phi_{m+1}(i))$ and evaluate the DD between patches.
[VNLEM-DD. Step 3] For each $P_i^{(\mathtt{n})}$, find N_1 nearest neighbours in terms of DD.
[VNLEM-DD. Step 4] Find $\lceil N_1/2 \rceil$ closest patches with respect to $\tilde{d}_{\texttt{RID}}$ among the N_1 patches from the previous step to form $N_{\texttt{DD}}(i)$.
[VNLEM-DD. Step 5] For each i, set $\tilde{I}^{(\texttt{VNLEM-DD})}(i)$ to be the center point of

$$\underset{P \in \mathbb{R}^{q \times q}}{\operatorname{argmin}} \sum_{P_j^{(\mathtt{n})} \in N_{\texttt{DD}}(i)} W'_{ij} \|P - O_{ij}.P_j^{(\mathtt{n})}\|^{\gamma}.$$

4. Theoretical Analysis

In this section, we provide a theoretical analysis to study the proposed VNLEM algorithm. The first part concerns how the VNLEM algorithms work. Precisely, we claim that the clean patch neighbors could be accurately evaluated from the noisy patch neighbors with high probability. This theorem also explains why the traditional nonlocal mean/median algorithm work. The second part concerns how accurate the orientation feature determined by the SIFT could help us to accurately approximate the RID.

In this section, we show that through finding nearest neighborhoods of noisy patches, it is with high probability that we would find "correct" nearest neighborhoods of clean patches as well.

Note that the rotation group action on patches can be expressed as

$$O.P_j^{(\mathtt{n})} = O.P_j^{(\mathtt{c})} + \sigma O.\xi_j, \tag{37}$$

where $O \in SO(2)$. When two patches $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$ do not overlap, the noises of $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$ are independent. However, when $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$ overlap, the associated noises are not independent, and we need to control the dependence. To achieve this, we introduce the following sets that are associated with $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$:

$$K_{\mathtt{0}}(O) := \{(a,b)|\, a,b \in \{1,\cdots,q^2\} \text{ such that } \xi_i(a) = [O.\xi_j](b)\}, \tag{38}$$

which is associated with the overlapped pixels of $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$ and is dependent on the rotation O, but whose cardinality does not depend on O;

$$K_{\mathtt{S}}(O) := \{(a,b)|\, a,b \in \{1,\cdots,q^2\} \text{ such that } a \neq b,\ \xi_i(a) = O.\xi_j(b) \text{ and } \xi_i(b) = [O.\xi_j](a)\}, \tag{39}$$

which is associated with the "swapped" pixel indices after rotation and depends on $O \in SO(2)$; and

$$K_{\mathtt{I}}(O) := \{a \in \{1,\cdots,q^2\}|\, \xi_i(a) = [O.\xi_j](a)\}, \tag{40}$$

which is associated with the overlapped pixels of $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$ with "identical indices" after rotation and depends on $O \in SO(2)$. By definition, the cardinalities of $K_{\mathtt{S}}(O)$ and $K_{\mathtt{I}}(O)$ both depend on O.

Note that $K_{\mathtt{S}}(O) \subset K_{\mathtt{0}}(O)$ and $K_{\mathtt{I}}(O) \subset K_{\mathtt{0}}(O)$, and when $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$ do not overlap, $K_{\mathtt{0}}(O)$, $K_{\mathtt{S}}(O)$ and $K_{\mathtt{I}}(O)$ are all empty sets. Also note that $K_{\mathtt{I}}(O)$ would only have at most one element. We mention that for the NLEM, since there is no rotation, $K_{\mathtt{I}}(O)$ and $K_{\mathtt{S}}(O)$ will be empty.

To better illustrate the sets $K_{\mathtt{0}}(O)$, $K_{\mathtt{S}}(O)$, and $K_{\mathtt{I}}(O)$, see Figures 4 to 7 when the rotation is of 180 degree. It is easier to visualize the overlap in the continuous setup. Let $\mathbf{x}$ be a point in the overlap region of $p_i^{(\mathtt{n})}$ and $p_j^{(\mathtt{n})}$. If there exists $O \in SO(2)$ so that after rotation the point would be in the same relative position in $p_i^{(\mathtt{n})}$ and $O.p_j^{(\mathtt{n})}$, then the distances from $\mathbf{x}$ to the boundaries of $p_i^{(\mathtt{n})}$ and $p_j^{(\mathtt{n})}$ must be the same. See Figure 4 for an illustration. Therefore, when two patches have overlap, only the points on the line segment connecting the intersection points of the boundary circles. In the discrete setup, the same consideration holds. See Figure 5 for an illustration. Therefore, for each rotation action, there would be at most one pixel pixel being lined up, and hence $|K_{\mathtt{I}}(O)| \leq 1$.

On the other hand, if $K_{\mathtt{S}}(O)$ is not empty, there exist two points $\mathbf{x}$ and $\mathbf{y}$ in the intersection so that their corresponding rotated points $\mathbf{x}'$ and $\mathbf{y}'$ are at the same relative positions but swapped. This is only possible when the rotation angle is π in the continuous setup. See Figure 6. Note that the overlap region is symmetric about the centre $\mathbf{x}$. When the rotation angle is π in the discrete setup, the overlapping region, except $\mathbf{x}$, are points in $K_{\mathtt{S}}(O)$, and hence $|K_{\mathtt{S}}(O)| \leq |K_{\mathtt{0}}(O)|$. See Figure 7. Clearly, in general we have a rough bound $|K_{\mathtt{0}}(O)| \leq q(q-1)$.

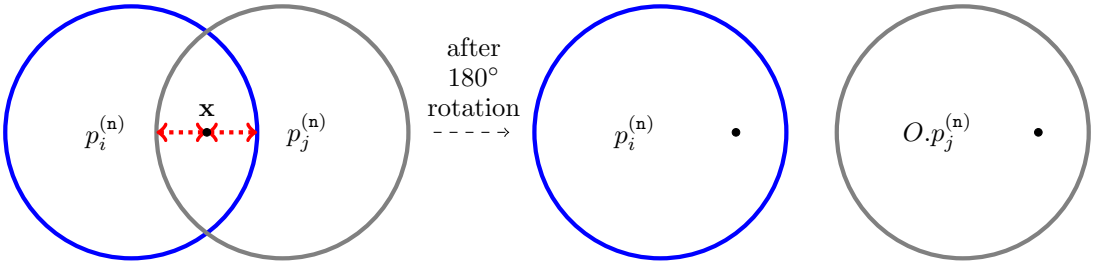


FIGURE 4. Illustration of two overlapping patches under the continuous setup.

Lemma 9. *Fix $O \in SO(2)$. Take two patches $P_i^{(n)}, P_j^{(n)} \in \mathcal{X}_I^{(n)} \subset \mathbb{R}^{q^2}$, where $\mathcal{X}_I^{(n)}$ is defined in (25). Then, we have*

$$\mathbb{E}(\|P_i^{(n)} - O.P_j^{(n)}\|^2) = \|P\|^2 + 2\sigma^2(q^2 - |K_I(O)|) \tag{41}$$

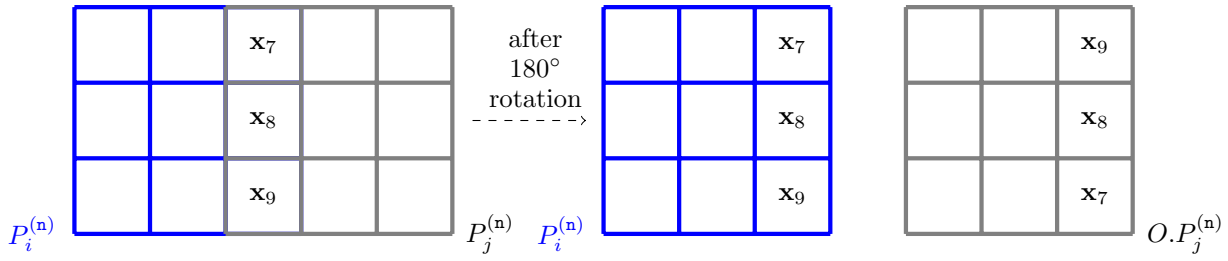


FIGURE 5. Illustration of the set $K_{\mathtt{I}}(O)$ of two overlapping patches under the discrete setup. In this example there are three overlapped pixels. Clearly, $K_{\mathtt{I}}(O) = \{8\}$ since $P_i^{(\mathrm{n})}(8) = O.P_j^{(\mathrm{n})}(8)$.

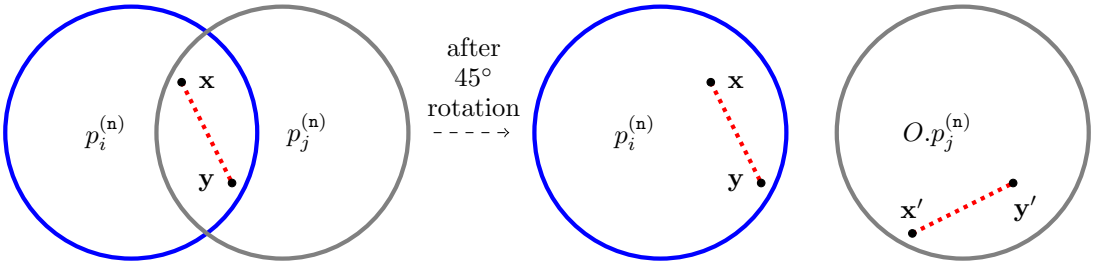


FIGURE 6. Illustration of the set $K_{\mathtt{S}}(O)$ two overlapping patches under the continuous setup. In this case, $K_{\mathtt{S}}(O)$ is empty.

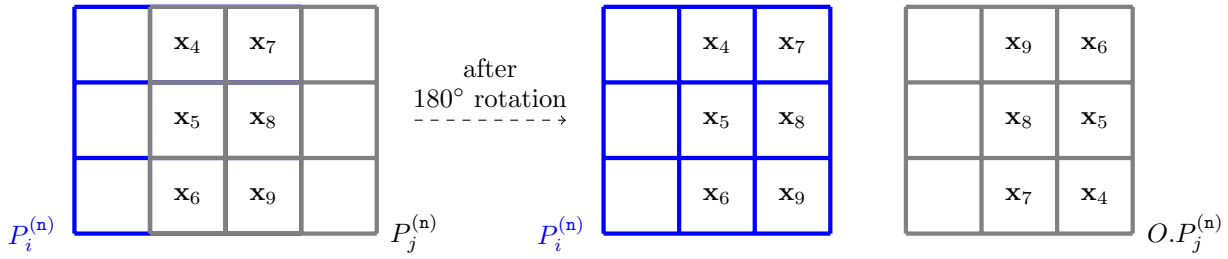


FIGURE 7. Illustration of the set $K_{\mathtt{S}}(O)$ of two overlapping patches under the discrete setup. In this example there are six overlapped pixels. By definition, we have $K_{\mathtt{S}}(O) = \{(4,9),(9,4),(5,8),(8,5),(6,7),(7,6)\}$ and $|K_{\mathtt{S}}(O)| = 6$.

and

$$\begin{aligned}\mathrm{Var}(\|P_i^{(\mathrm{n})} - O.P_j^{(\mathrm{n})}\|^2) = 8\sigma^2 &\left(\|P\|^2 - \sum_{(a,b)\in K_{\mathtt{O}}(O)} P(a)P(b) + \sum_{a\in K_{\mathtt{I}}(O)} P(a)^2\right) \\ &+ 4\sigma^4\left(2q^2 + |K_{\mathtt{S}}(O)| + |K_{\mathtt{O}}(O)| - 3|K_{\mathtt{I}}(O)|\right),\end{aligned} \tag{42}$$

where $P := P_i^{(\mathrm{c})} - O.P_j^{(\mathrm{c})}$. *Particularly, when* O *is the identity, we have*

$$\mathbb{E}(\|P_i^{(\mathrm{n})} - P_j^{(\mathrm{n})}\|^2) = \|P\|^2 + 2\sigma^2 q^2 \tag{43}$$

and

$$\mathrm{Var}(\|P_i^{(n)} - P_j^{(n)}\|^2) = 8\sigma^2 \left(\|P\|^2 - \sum_{(a,b)\in K_{\mathtt{O}}} P(a)P(b) \right) + 4\sigma^4 \left(2q^2 + |K_{\mathtt{O}}|\right) . \tag{44}$$

Remark 10. *Before proving the Lemma, we have some comments. First, since* $2\sigma^2(q^2 - |K_{\mathtt{I}}(O)|) > 0$, $\|P_i^{(n)} - O.P_j^{(n)}\|^2$ *is a biased estimator of* $\|P_i^{(c)} - O.P_j^{(c)}\|^2$. *Second, by the lemma, if* $P_i^{(n)}, P_j^{(n)} \in \mathcal{X}_I^{(n)} \subset \mathbb{R}^{q^2}$ *do not overlap, we have*

$$\mathbb{E}(\|P_i^{(n)} - O.P_j^{(n)}\|^2) = \|P\|^2 + 2\sigma^2 q^2 \text{ and } \mathrm{Var}(\|P_i^{(n)} - O.P_j^{(n)}\|^2) = 8\sigma^2\|P\|^2 + 8\sigma^4 q^2 . \tag{45}$$

Third, when O *is the identity, the Lemma could be applied to study the NLEM. Finally, since the sign of* $\sum_{(a,b)\in K_{\mathtt{O}}} P(a)P(b)$ *is not controlled, this term leads to the complicated behavior of the* L^2 *distance or RID when two patches overlap. Indeed, when two patches overlap, depending on the clean patches' structure, the RID estimator from the noisy patches might be biased toward overlapped patches. Thus, if the overlapped patches are included in the denoising process, the search of the nearest neighboring patches might be biased to the "local patches" that have overlaps.*

Proof. Since i, j are fixed, to simplify the notation, we write

$$P_i^{(\mathtt{n})}(a) - O.P_j^{(\mathtt{n})}(a) = P_i^{(\mathtt{c})}(a) - O.P_j^{(\mathtt{c})}(a) + \sigma(\xi_i - O.\xi_j) = P(a) + \sigma(\xi(a) - \xi'(a)) ,$$

where $\xi' := O.\xi_j$. Hence, we can express the ℓ^2 norm of $P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}$ as

$$\begin{aligned} \|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|^2 &= \|P\|^2 + 2\sigma P^T(\xi - \xi') + \sigma^2\|\xi - \xi'\|^2 \\ &= \|P\|^2 + 2\sigma P^T(\xi - \xi') + \sigma^2 \left(\|\xi\|^2 + \|\xi'\|^2 - 2\xi^T\xi'\right) . \end{aligned} \tag{46}$$

By the assumption, $\xi(a), a = 1, \ldots, q^2$ are i.i.d. Gaussian random variables and as well as $\xi'(a), a = 1, \ldots, q^2$. We also know that $\xi(a)$ and $\xi'(a)$ are independent Gaussian random variables when $a \notin K_{\mathtt{I}}(O)$. By a direct calculation, we have $\mathbb{E}(\xi^T\xi') = |K_{\mathtt{I}}(O)|$. Combining this with the facts that

$$\mathbb{E}(\|\xi\|^2) = \mathbb{E}(\|\xi'\|^2) = q^2 \text{ and } \mathbb{E}(\xi(a) - \xi'(a)) = 0 , \tag{47}$$

we obtain

$$\mathbb{E}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|^2) = \|P\|^2 + 2\sigma^2(q^2 - |K_{\mathtt{I}}(O)|). \tag{48}$$

To compute the variance, we write $\mathrm{Var}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|^2)$ as the following by expanding (46):

$$\begin{aligned} &\mathrm{Var}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|^2) \\ &= \mathrm{Var}\left(2\sigma \sum_{a=1}^{q^2} P(a)(\xi(a) - \xi'(a))\right) + \mathrm{Var}\left(\sigma^2 \sum_{a=1}^{q^2} (\xi(a) - \xi'(a))^2\right) \\ &\quad + 2\mathrm{Cov}\left(2\sigma \sum_{a=1}^{q^2} P(a)(\xi(a) - \xi'(a)), \sigma^2 \sum_{a=1}^{q^2} (\xi(a) - \xi'(a))^2\right) \\ &= 4\sigma^2 \cdot (\mathtt{I}) + \sigma^4 \cdot (\mathtt{II}) + 4\sigma^3 \cdot (\mathtt{III}), \end{aligned} \tag{49}$$

where $(\mathtt{I}) := \mathrm{Var}\left(\sum_{a=1}^{q^2} P(a)(\xi(a) - \xi'(a))\right)$, $(\mathtt{II}) := \mathrm{Var}\left(\sum_{a=1}^{q^2} (\xi(a) - \xi'(a))^2\right)$, and $(\mathtt{III}) := \mathrm{Cov}\left(\sum_{a=1}^{q^2} P(a)(\xi(a) - \xi'(a)), \sum_{a=1}^{q^2} (\xi(a) - \xi'(a))^2\right)$. We compute $(\mathtt{I}), (\mathtt{II})$, and $(\mathtt{III})$ below.

$$
\begin{aligned}
(\mathtt{I}) &= \mathrm{Var}\left(\sum_{a=1}^{q^2} P(a)(\xi(a)-\xi'(a))\right) \\
&= \sum_{a=1}^{q^2} P^2(a)\mathrm{Var}(\xi(a)-\xi'(a)) + \sum_{a,b\in\{1,\cdots,q^2\},a\neq b} P(a)P(b)\mathrm{Cov}\left(\xi(a)-\xi'(a),\xi(b)-\xi'(b)\right) \\
&= 2\|P\|^2 - 2\sum_{a,b\in\{1,\cdots,q^2\},a\neq b} P(a)P(b)\mathrm{Cov}(\xi(a),\xi'(b)) \qquad (50)
\end{aligned}
$$

where the second equality comes from a direct expansion, and the last equality holds since

$$
\mathrm{Cov}\left(\xi(a)-\xi'(a),\xi(b)-\xi'(b)\right) = -\left[\mathrm{Cov}\left(\xi(a),\xi'(b)\right)+\mathrm{Cov}\left(\xi(b),\xi'(a)\right)\right], \qquad (51)
$$

which comes from the fact that $\{\xi(a)\}_{a=1}^{q^2}$ are independent and $\{\xi'(a)\}_{a=1}^{q^2}$ are independent. Since only overlapped pixels lead to non-zero $\mathrm{Cov}(\xi(a),\xi'(b))$, we have

$$
\sum_{a,b\in\{1,\cdots,q^2\},a\neq b} P(a)P(b)\mathrm{Cov}(\xi(a),\xi'(b)) = \sum_{(a,b)\in K_{\mathtt{0}}(O)} P(a)P(b) - \sum_{a\in K_{\mathtt{I}}(O)} P(a)^2, \qquad (52)
$$

where we subtract $\sum_{a\in K_{\mathtt{I}}(O)} P(a)^2$ since $a\neq b$. As a result,

$$
(\mathtt{I}) = 2\|P\|^2 - 2\left(\sum_{(a,b)\in K_{\mathtt{0}}(O)} P(a)P(b) - \sum_{a\in K_{\mathtt{I}}(O)} P(a)^2\right). \qquad (53)
$$

Next,

$$
\begin{aligned}
(\mathtt{II}) &= \mathrm{Var}\left(\sum_{a=1}^{q^2}(\xi(a)-\xi'(a))^2\right) = \mathrm{Var}\left(\|\xi\|^2+\|\xi'\|^2-2\xi^T\xi'\right) \\
&= \mathrm{Var}(\|\xi\|^2)+\mathrm{Var}(\|\xi'\|^2)+4\mathrm{Var}\left(\sum_{a=1}^{q^2}\xi(a)\xi'(a)\right)+2\mathrm{Cov}\left(\sum_{a=1}^{q^2}\xi(a)^2,\sum_{b=1}^{q^2}\xi'(b)^2\right) \\
&\quad -4\mathrm{Cov}\left(\sum_{a=1}^{q^2}\xi(a)^2,\sum_{b=1}^{q^2}\xi(b)\xi'(b)\right)-4\mathrm{Cov}\left(\sum_{a=1}^{q^2}\xi'(a)^2,\sum_{b=1}^{q^2}\xi(b)\xi'(b)\right). \qquad (54)
\end{aligned}
$$

We now calculate $(\mathtt{II})$ term by term. By a direct expansion, we have

$$
\begin{aligned}
\mathrm{Var}\left(\sum_{a=1}^{q^2}\xi(a)\xi'(a)\right) &= \sum_{a=1}^{q^2}\mathrm{Var}(\xi(a)\xi'(a)) + \sum_{a,b\in\{1,\cdots,q^2\},a\neq b}\mathrm{Cov}\left(\xi(a)\xi'(a),\xi(b)\xi'(b)\right) \\
&= \Big[\sum_{a\notin K_{\mathtt{I}}(O)}\mathrm{Var}(\xi(a))\mathrm{Var}(\xi'(a)) + \sum_{a\in K_{\mathtt{I}}(O)}\mathrm{Var}(\xi(a)^2)\Big] + \sum_{(a,b)\in K_{\mathtt{S}}(O)}\mathrm{Var}(\xi(a)\xi(b)) \\
&= [(q^2-|K_{\mathtt{I}}(O)|)+2|K_{\mathtt{I}}(O)|]+|K_{\mathtt{S}}(O)| \\
&= q^2+|K_{\mathtt{I}}(O)|+|K_{\mathtt{S}}(O)|, \qquad (55)
\end{aligned}
$$

where the second equality holds since $\mathrm{Cov}\left(\xi(a)\xi'(a),\xi(b)\xi'(b)\right)\neq 0$ only when $(a,b)\in K_{\mathtt{S}}(O)$ and the third equality holds since $\mathrm{Var}(\xi(a)^2)=2$ and $\mathrm{Var}(\xi(a)\xi(b)) = \mathbb{E}\xi(a)^2\mathbb{E}\xi(b)^2 = 1$ due to the independence. Similarly, we have by a direct calculation

$$\text{Cov}\left(\sum_{a=1}^{q^2}\xi(a)^2,\sum_{b=1}^{q^2}\xi'(b)^2\right)=\sum_{(a,b)\in K_{\mathtt{0}}(O)}\text{Cov}\left(\xi(a)^2,\xi'(b)^2\right)=2|K_{\mathtt{0}}(O)|\,. \tag{56}$$

By the linearity of the covariance,

$$\begin{aligned}&\text{Cov}\left(\sum_{a=1}^{q^2}\xi(a)^2,\sum_{b=1}^{q^2}\xi(b)\xi'(b)\right)\\&=\sum_{a=1}^{q^2}\text{Cov}(\xi(a)^2,\xi(a)\xi'(a))+\sum_{a,b\in\{1,\dots,q^2\},a\neq b}\text{Cov}(\xi(a)^2,\xi(b)\xi'(b))\\&=2|K_{\mathtt{I}}(O)|+\sum_{a\neq b,\,b\in K_{\mathtt{I}}(O)}\text{Cov}(\xi(a)^2,\xi(b)\xi'(b))+\sum_{a\neq b,\,b\notin K_{\mathtt{I}}(O)}\text{Cov}(\xi(a)^2,\xi(b)\xi'(b))=2|K_{\mathtt{I}}(O)|\,,\end{aligned} \tag{57}$$

where

$$\sum_{a\neq b,\,b\in K_{\mathtt{I}}(O)}\text{Cov}(\xi(a)^2,\xi(b)\xi'(b))=\sum_{a\neq b,\,b\in K_{\mathtt{I}}(O)}\text{Cov}(\xi(a)^2,\xi(b)^2)=0 \tag{58}$$

and

$$\sum_{a\neq b,\,b\notin K_{\mathtt{I}}(O)}\text{Cov}(\xi(a)^2,\xi(b)\xi'(b))=0, \tag{59}$$

since $\mathbb{E}(X)=\mathbb{E}(X^3)=0$ for $X\sim\mathcal{N}(0,1)$. To be more precise, when $a\neq b$ and $b\in K_{\mathtt{I}}(O)$, $\text{Cov}(\xi(a)^2,\xi(b)^2)=0$ due to the independence assumption; when $a\neq b$ and $b\notin K_{\mathtt{I}}(O)$, $\xi(b)$ and $\xi'(b)$ are independent, so

$$\text{Cov}(\xi(a)^2,\xi(b)\xi'(b))=\mathbb{E}(\xi(a)^2\xi(b)\xi'(b))-\mathbb{E}\xi(a)^2\mathbb{E}(\xi(b)\xi'(b))=\mathbb{E}(\xi(a)^2\xi(b)\xi'(b)), \tag{60}$$

which is 0 no matter $\xi(a)^2$ is independent of $\xi(b)\xi'(b)$ or not. Note that by our assumption, $\xi(a)$ is dependent on $\xi(b)$ (or $\xi'(b)$) if and only if $\xi(a)$ is the same as $\xi(b)$ (or $\xi'(b)$).

Similarly, we have

$$\text{Cov}\left(\sum_{a=1}^{q^2}\xi'(a)^2,\sum_{b=1}^{q^2}\xi(b)\xi'(b)\right)=2|K_{\mathtt{I}}(O)|. \tag{61}$$

By substituting (55), (56), (57), and (61) into (54), we obtain

$$(\mathtt{II})=8q^2+4(|K_{\mathtt{S}}(O)|+|K_{\mathtt{0}}(O)|-3|K_{\mathtt{I}}(O)|). \tag{62}$$

Finally, we have

$$\begin{aligned}(\mathtt{III})&=\text{Cov}\left(\sum_{a=1}^{q^2}P(a)(\xi(a)-\xi'(a)),\sum_{a=1}^{q^2}(\xi(a)-\xi'(a))^2\right)=\sum_{a,b=1}^{q^2}P(a)\mathbb{E}(\xi(a)-\xi'(a))(\xi(b)-\xi'(b))^2\\&=\sum_{a,b=1}^{q^2}P(a)\mathbb{E}[\xi(a)\xi(b)^2-2\xi(a)\xi(b)\xi'(b)+\xi(a)\xi'(b)^2-\xi'(a)\xi(b)^2+2\xi'(a)\xi(b)\xi'(b)-\xi'(a)\xi'(b)^2]=0,\end{aligned} \tag{63}$$

since $\text{Cov}(X,Y^2)=0$ for independent $X,Y\sim\mathcal{N}(0,1)$ and $\text{Cov}(X,YZ)=0$ for independent $X,Y,Z\sim\mathcal{N}(0,1)$.

Combining equations (53), (62), and (63), we conclude that

$$\begin{aligned}(64)\quad &\mathrm{Var}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|^2)\\ &= 8\sigma^2\left(\|P\|^2 - \sum_{(a,b)\in K_{\mathtt{O}}(O)} P(a)P(b) + \sum_{a\in K_{\mathtt{I}}(O)} P(a)^2\right) + 4\sigma^4\left(2q^2 + |K_{\mathtt{S}}(O)| + |K_{\mathtt{O}}(O)| - 3|K_{\mathtt{I}}(O)|\right).\end{aligned}$$

□

With this Lemma, we are ready to show that the RID between two clean patches could be well approximated by noisy patches; particularly, if two clean patches are close enough, their RID can be well approximated by the associated noisy patches.

Theorem 11. *Take two patches $P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})} \in \mathcal{X}_I^{(\mathtt{n})} \subset \mathbb{R}^{q^2}$. Suppose that $d_{RID}(P_i^{(\mathtt{c})}, P_j^{(\mathtt{c})}) < \epsilon$ and that $\sigma q < \epsilon$. Then*

$$(65)\qquad \Pr\left(d_{RID}(P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})}) < 2\epsilon\right) > \left(1 + 8\sigma^2\frac{2\epsilon^2 + \sigma^2(2q^2-q)}{\left(3\epsilon^2 - 2\sigma^2(q^2-1)\right)^2}\right)^{-1},$$

which increases when q decreases.

Suppose that $d_{RID}(P_i^{(\mathtt{c})}, P_j^{(\mathtt{c})}) > 2\epsilon$. Then

$$(66)\qquad \Pr\left(d_{RID}(P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})}) > \epsilon\right) > \left(1 + 8\sigma^2\frac{2\|P\|^2 + \sigma^2(2q^2-q)}{\left(3\|P\|^2/4 + 2\sigma^2(q^2-1)\right)^2}\right)^{-1},$$

which increases when q increases.

In particular, when the patches $P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})} \in \mathcal{X}_I^{(\mathtt{n})} \subset \mathbb{R}^{q^2}$ have no overlap, we have if $d_{RID}(P_i^{(\mathtt{c})}, P_j^{(\mathtt{c})}) < \epsilon$, then

$$(67)\qquad \Pr\left(d_{RID}(P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})}) < 2\epsilon\right) > \left(1 + 8\sigma^2\frac{\epsilon^2 + \sigma^2q^2}{(3\epsilon^2 - 2\sigma^2q^2)^2}\right)^{-1};$$

if $d_{RID}(P_i^{(\mathtt{c})}, P_j^{(\mathtt{c})}) > \epsilon$, then

$$(68)\qquad \Pr\left(d_{RID}(P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})}) > \epsilon\right) > \left(1 + 8\sigma^2\frac{\|P\|^2 + \sigma^2q^2}{\left(3\|P\|^2/4 + 2\sigma^2q^2\right)^2}\right)^{-1}.$$

Remark 12. *When two patches $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$ are disjoint, the bound (67) suggests that the smaller patch size is better. However, the bound (68) suggests the opposite. We thus need to choose a suitable patch size q that balances (67) and (68).*

When two patches $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$ overlap, the analysis of choosing the patch size becomes very complicated. It would depend on the image, the minimiser rotation O, and how the patches overlap. See (73) for example. The worst bounds (65) and (67) we have in Theorem 11 also suggest that q should not be too small but also not too large. In practice, we found that an odd value q between 7 and 15 leads to a good performance, but the optimal q depends on the image.

We mention that if O is the identity in Theorem 11, the same argument explains why the L^2 distance between two clean patches could be well approximated by noisy patches, and hence better understand the NLEM algorithm.

Proof. Suppose $d_{\mathtt{RID}}(P_i^{(\mathtt{c})}, P_j^{(\mathtt{c})}) < \epsilon$ and $\sigma q < \epsilon$. Suppose $O \in SO(2)$ is such that $d_{\mathtt{RID}}(P_i^{(\mathtt{c})}, P_j^{(\mathtt{c})}) = \|P_i^{(\mathtt{c})} - O.P_j^{(\mathtt{c})}\| = \|P\|^2$. We start by preparing a bound. Applying Lemma 9 and the bounds of $|K_{\mathtt{S}}(O)|$ and $|K_{\mathtt{I}}(O)|$, we have

$$
\begin{aligned}
&\frac{\mathrm{Var}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|)}{\left(4\epsilon^2 - \mathbb{E}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|)\right)^2} \\
&= \frac{8\sigma^2\left(\|P\|^2 - \sum_{(a,b)\in K_{\mathtt{0}}, a,b\notin K_{\mathtt{I}}} P(a)P(b)\right) + 4\sigma^4\left(2q^2 + |K_{\mathtt{0}}| + |K_{\mathtt{S}}(O)| - 3|K_{\mathtt{I}}(O)|\right)}{\left(4\epsilon^2 - \|P\|^2 - 2\sigma^2(q^2 - |K_I|)\right)^2} \\
&\leq \frac{16\sigma^2\epsilon^2 + 8\sigma^4(2q^2 - q)}{\left(3\epsilon^2 - 2\sigma^2(q^2-1)\right)^2},
\end{aligned} \tag{69}
$$

since

$$
\left|\sum_{(a,b)\in K_{\mathtt{0}}} P(a)P(b)\right| \leq \|P\|^2,\ |K_{\mathtt{0}}| \leq q(q-1),\ |K_{\mathtt{S}}| \leq q(q-1),\ \text{and } |K_{\mathtt{I}}| \leq 1. \tag{70}
$$

Recall the one-sided Chebychev's inequality for a random variable X with a finite second moment: $\Pr(X \geq \mathbb{E}X + a) \leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)+a^2}$, for $a > 0$. Applying the one-sided Chebyshev's inequality and the inequality above, we obtain

$$
\begin{aligned}
&\Pr\left(d_{\mathtt{RID}}(P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})}) < 2\epsilon\right) \geq \Pr\left(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|^2 < (2\epsilon)^2\right) \\
&> \left(1 + \frac{\mathrm{Var}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|)}{\left(4\epsilon^2 - \mathbb{E}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|)\right)^2}\right)^{-1} \geq \left(1 + \frac{16\sigma^2\epsilon^2 + 8\sigma^4(2q^2-q)}{\left(3\epsilon^2 - 2\sigma^2(q^2-1)\right)^2}.\right)^{-1}.
\end{aligned} \tag{71}
$$

When patches $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$ do not overlap,

$$
\frac{\mathrm{Var}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|)}{\left(4\epsilon^2 - \mathbb{E}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|)\right)^2} = \frac{8\sigma^2\|P\|^2 + 8\sigma^4 q^2}{(4\epsilon^2 - \|P\|^2 - 2\sigma^2 q^2)^2}, \tag{72}
$$

which implies (67).

Now, suppose $d_{\mathtt{RID}}(P_i^{(\mathtt{c})}, P_j^{(\mathtt{c})}) > 2\epsilon$. For any $O \in SO(2)$, we apply the assumption $d_{\mathtt{RID}}(P_i^{(\mathtt{c})}, P_j^{(\mathtt{c})}) > 2\epsilon$ and Lemma 9. We obtain

$$
\begin{aligned}
&\frac{\mathrm{Var}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|)}{\left(\mathbb{E}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|) - \epsilon^2\right)^2} \\
&\leq \frac{8\sigma^2\left(\|P\|^2 - \sum_{(a,b)\in K_{\mathtt{0}}, a,b\notin K_{\mathtt{I}}} P(a)P(b)\right) + 4\sigma^4\left(2q^2 + |K_{\mathtt{0}}| + |K_{\mathtt{S}}| - 3|K_{\mathtt{I}}|\right)}{\left(3\|P\|^2/4 + 2\sigma^2(q^2 - |K_I|)\right)^2},
\end{aligned} \tag{73}
$$

where $P = P_i^{(\mathtt{c})} - O.P_j^{(\mathtt{c})}$. Due to the bounds shown in (70), we can further obtain the following bound which is independent of the rotation O:

$$
\frac{\mathrm{Var}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|)}{\left(\mathbb{E}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|) - \epsilon^2\right)^2} \leq \frac{16\sigma^2\|P\|^2 + 8\sigma^4(2q^2-q)}{\left(3\|P\|^2/4 + 2\sigma^2(q^2-1)\right)^2}.
$$

Applying the one-sided Chebyshev's inequality, we can obtain a universal lower bound

$$\Pr\left(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|^2 > \epsilon^2\right) > \left(1 + \frac{\mathrm{Var}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|)}{\left(\mathbb{E}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|) - \epsilon^2\right)^2}\right)^{-1} \geq \left(1 + 4\sigma^2 \frac{4\|P\|^2 + \sigma^2\left(4q^2 - 2q\right)}{\left(3\|P\|^2/4 + 2\sigma^2(q^2-1)\right)^2}\right)^{-1}. \tag{74}$$

Since the lower bound (74) holds for any rotation O, we therefore have

$$\Pr\left(d_{\mathtt{RID}}(P_i^{(\mathtt{n})}, P_j^{(\mathtt{n})}) > \epsilon\right) > \left(1 + 4\sigma^2 \frac{4\|P\|^2 + \sigma^2\left(4q^2 - 2q\right)}{\left(3\|P\|^2/4 + 2\sigma^2(q^2-1)\right)^2}\right)^{-1}. \tag{75}$$

When patches $P_i^{(\mathtt{n})}$ and $P_j^{(\mathtt{n})}$ do not overlap,

$$\frac{\mathrm{Var}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|)}{\left(\mathbb{E}(\|P_i^{(\mathtt{n})} - O.P_j^{(\mathtt{n})}\|) - \epsilon^2\right)^2} = \frac{8\sigma^2\|P\|^2 + 8\sigma^4}{(\|P\|^2 - \epsilon^2 + 2\sigma^2 q^2)^2} \tag{76}$$

which implies (68). □

Some discussions are needed for this Theorem. The quantity $\sigma^2 q^2$ could be understood as the "total energy" of the added noise, and the condition $\sigma^2 q^2 < \epsilon^2$ means that the RID estimated from two noisy patches is controlled by the square root of the energy of the noise. With this energy viewpoint, we could apply the technique developed in [15, Theorem 2.1]. However, we carried out the proof in the above way to emphasize the main purpose of the RID, and to find the true neighbors and the dependence on the patch size.

Second, we mention that when $d_{\mathtt{RID}}(P_i^{(\mathtt{c})}, P_j^{(\mathtt{c})}) < \epsilon$ and $P_i^{(\mathtt{c})}$ is of size $q \times q$, then the difference of the central pixels of $P_i^{(\mathtt{c})}$ and $P_j^{(\mathtt{c})}$ is bounded by ϵ in the worst case. Indeed, we have

$$d^2_{\mathtt{RID}}(P_i^{(\mathtt{c})}, P_j^{(\mathtt{c})}) = \|P_i^{(\mathtt{c})} - O.P_j^{(\mathtt{c})}\|^2 = \sum_{a=1, a\neq c}^{q^2} |P_i^{(\mathtt{c})}(a) - (O.P_j^{(\mathtt{c})})(a)|^2 + |P_i^{(\mathtt{c})}(c) - P_j^{(\mathtt{c})}(c)|^2 < \epsilon^2, \tag{77}$$

where $O \in SO(2)$ is the rotation that achieves the RID, and the second equality holds since $(O.P_j^{(\mathtt{c})})(c) = P_j^{(\mathtt{c})}(c)$. In practice, $|P_i^{(\mathtt{c})}(c) - P_j^{(\mathtt{c})}(c)|$ could be smaller than ϵ when $\sum_{a=1,a\neq c}^{q^2} |P_i^{(\mathtt{c})}(a) - (O.P_j^{(\mathtt{c})})(a)|^2$ is not zero. When O is the identity, the same argument holds for the NLEM algorithm, and this explains why we could have a better denoising result by using the patches.

SIFT is a method for extracting features that are invariant to image scale and rotation [27]. The idea was originally from [3, 9, 26], and became popular after [27]. For a given image, SIFT detects points of interest, called keypoints, under the scale-space model [24, 25]. It then assigns the *orientation feature* at each keypoint. Since the centre point of each patch is our point of interest, we only use orientation feature assignments in the SIFT algorithm and skip the keypoint detection. For more details and different variations of SIFT, we refer the interested reader to [26] for a review. In this subsection, we give theoretical proofs to show why orientation assignments in SIFT is robust to noise and hence can be used to approximate rotation angles between patches.

Definition 13. *Let $p : \mathbb{R}^2 \to \mathbb{R}$ be a round patch of an image $I : \mathbb{R}^2 \to \mathbb{R}$ defined in Definition 2. The Gaussian smoothed patch, denoted as L_p, is defined as the convolution of the Gaussian function $G(\mathbf{s}, 1)$ with the patch p*

$$L_p(\mathbf{x}) = G * p(\mathbf{x}) = \iint_{\mathbb{R}^2} G(\mathbf{s}, 1) p(\mathbf{x} - \mathbf{s}) d\mathbf{s}, \tag{78}$$

where the Gaussian $G(\mathbf{s}, \ell) := \frac{1}{2\pi\ell^2} e^{-\frac{\|\mathbf{s}\|^2}{2\ell^2}}$ and $\ell > 0$ denotes the scale in the SIFT algorithm.

We only consider the case when $\ell = 1$. Similar argument could be carried over for $\ell \neq 1$. From now on, we write $G(\mathbf{s}, 1)$ as $G(\mathbf{s})$ to simplify the notation. Suppose L_p and $L_{O.p}$ denote the Gaussian smoothed patches of p and its rotated patch $O.p$ by the angle ϕ. Then since $O \in SO(2)$ and $G(\mathbf{s})$ is rotationally invariant, we have

$$L_{O.p}(\mathbf{x}) = L_p(O^{-1}\mathbf{x}) \tag{79}$$

and the orientation features of p and $O.p$ will differ by the angle ϕ as well.

Denote S^1 to be the unit circle in $\mathbb{R}^2$ with the metric induced from the canonical metric of $\mathbb{R}^2$. We now give a mathematical definition of the orientation feature in the SIFT algorithm.

Definition 14 (Orientation)**.** *Given a patch $p : D_r \to \mathbb{R}$, let Ψ be the map*

$$\begin{aligned} \Psi : D_r &\to S^1 \\ \mathbf{x} &\mapsto \theta, \end{aligned} \tag{80}$$

where θ is an angle between $\nabla L_p(\mathbf{x})$ and the positive x-axis. For a fixed positive number $\delta < \pi$, the orientation feature is defined as an angle $\theta^* \in S^1$ such that

$$\iint_{\Psi^{-1}(N_{\theta_*})} \|\nabla L(\mathbf{x})\| d\mathbf{x} = \max_{\theta'} \iint_{\Psi^{-1}(N_{\theta'})} \|\nabla L(\mathbf{x})\| d\mathbf{x}, \tag{81}$$

where $N_{\theta'} := \{\theta |\, d_{S^1}(\theta, \theta') < \delta\}$ and d_{S^1} is the distance with respect to the canonical metric on S^1.

In general, there might be more than one orientation feature associated with one patch. To simplify the discussion, we assume that there is only one orientation feature.

Assumption 15. *Take $\epsilon > 0$ and $\delta > 0$ associated with the orientation feature. Assume*

$$\iint_{\Psi^{-1}(N_{\theta_*})} \|\nabla L(\mathbf{x})\| d\mathbf{x} > \iint_{\Psi^{-1}(N_{\theta'})} \|\nabla L(\mathbf{x})\| d\mathbf{x} + \epsilon \pi r^2 \tag{82}$$

for any $\theta' \in S^1$ such that $d_{S^1}(\theta', \theta_*) > \delta$.

With the orientation features of patches, we could define the following "SIFT distance" between patches.

Definition 16. *The "SIFT distance" between patches $\tilde{p}$ and p is defined as*

$$d_{\mathtt{SIFT}}(\tilde{p}, p) := \|\tilde{p} - R(\tilde{\theta})R(\theta)^{-1}.p\|, \tag{83}$$

where $\tilde{\theta} \in S^1$ and $\theta \in S^1$ are the orientation feature of $\tilde{p}$ and p, and $\|\cdot\|$ denotes the L^2 norm.

First of all, note that the "SIFT distance" is not really a distance, but it is intimately related to the RID. If $\tilde{p} = R(\phi).p$, where $\phi \in S^1$ and $R : S^1 \to SO(2)$ is a diffeomorphic map, then it is clear that the orientation features of $\tilde{p}$ and p, denoted as $\tilde{\theta}$ and θ respectively, are related by $R(\tilde{\theta})R(\theta)^{-1} = R(\phi)$. However, the reverse is not true. If two patches have the same orientation features, it does not imply that they are the same. In practice, suppose the orientation features of $\tilde{p}$ and p are $\tilde{\theta}$ and θ respectively, we have

$$d_{\mathtt{RID}}(\tilde{p}, p) \leq \|\tilde{p} - R(\tilde{\theta})R(\theta)^{-1}.p\|_{L^2}. \tag{84}$$

In other words, two patches that are determined to be neighbors by the RID will be determined to be neighbors by the $d_{\mathtt{SIFT}}$ defined in (35). Note that the SIFT distance could be further improved to better approximate the RID, like (35). Thus, orientation features under SIFT can be used to estimate rotation angles between patches, or to "nonlinearly filter out" the impossible neighbors.

Next, we show why the orientation feature is robust to noise. We first study gradients of Gaussian smoothed patches. While the noise is modeled by the generalized random process $\psi\Phi$ in (14), to make the calculation succinct, we consider the following simplified model:

$$p_i^{(\mathtt{n})}(\mathbf{x}) = p_i^{(\mathtt{c})}(\mathbf{x}) + \sigma \xi_i(\mathbf{x}) \tag{85}$$

be a noisy patch defined on D_r, where $\sigma > 0$, $p_i^{(\mathtt{c})}$ denotes the patch of the clean image supported on D_r, and ξ_i are i.i.d. standard random normal variables for all $\mathbf{x} \in D_r$.[2] To further simplify notation, we denote the Gaussian smoothed noisy and clean patches, $p_i^{(\mathtt{n})}$ and $p_i^{(\mathtt{c})}$ defined in (14), by

$$L^{(\mathtt{n})}(\mathbf{x}) := L_{p_i^{(\mathtt{n})}}(\mathbf{x}), L^{(\mathtt{c})}(\mathbf{x}) := L_{p_i^{(\mathtt{c})}}(\mathbf{x}), \text{ and } \xi = \xi_i \tag{86}$$

respectively.

Lemma 17. *We have*

$$\mathbb{E}\left[\|\nabla L^{(\mathtt{n})}(\mathbf{x}) - \nabla L^{(\mathtt{c})}(\mathbf{x})\|^2\right] = \frac{\sigma^2}{4\pi} \tag{87}$$

and

$$\mathrm{Var}\left[\|\nabla L^{(\mathtt{n})}(\mathbf{x}) - \nabla L^{(\mathtt{c})}(\mathbf{x})\|^2\right] = \frac{\sigma^4}{16\pi^2}\left(1 + \frac{27}{16\pi}\right). \tag{88}$$

Proof. Note that

$$\|\nabla L^{(\mathtt{n})}(\mathbf{x}) - \nabla L^{(\mathtt{c})}(\mathbf{x})\|^2 = \sigma^2\left(\left(\iint_{\mathbb{R}^2} \xi(\mathbf{s}) G_x(\mathbf{x}-\mathbf{s}) d\mathbf{s}\right)^2 + \left(\iint_{\mathbb{R}^2} \xi(\mathbf{s}) G_y(\mathbf{x}-\mathbf{s}) d\mathbf{s}\right)^2\right), \tag{89}$$

where we denote $\nabla L = [L_x\ L_y]^T$. Since $\xi(\mathbf{s}) \sim \mathcal{N}(0,1)$ are i.i.d., we have

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla L^{(\mathtt{n})}(\mathbf{x}) - \nabla L^{(\mathtt{c})}(\mathbf{x})\|^2\right] \\
=&\sigma^2 \mathbb{E}\left[\iint_{\mathbb{R}^2} G_x(\mathbf{x}-\mathbf{s})\xi(\mathbf{s}) d\mathbf{s} \iint_{\mathbb{R}^2} G_x(\mathbf{x}-\mathbf{t})\xi(\mathbf{t}) d\mathbf{t}\right] \\
&+ \sigma^2 \mathbb{E}\left[\iint_{\mathbb{R}^2} G_y(\mathbf{x}-\mathbf{s})\xi(\mathbf{s}) d\mathbf{s} \iint_{\mathbb{R}^2} G_y(\mathbf{x}-\mathbf{t})\xi(\mathbf{t}) d\mathbf{t}\right] \\
=&\sigma^2\left(\iint_{\mathbb{R}^2} \mathbb{E}[\xi(\mathbf{s})^2] G_x(\mathbf{x}-\mathbf{s})^2 + \mathbb{E}[\xi(\mathbf{s})^2] G_y(\mathbf{x}-\mathbf{s})^2 d\mathbf{s}\right) \\
=&\sigma^2\left(\iint_{\mathbb{R}^2} \frac{x^2}{4\pi} e^{-(x^2+y^2)} + \frac{y^2}{4\pi} e^{-(x^2+y^2)} d\mathbf{s}\right) = \frac{\sigma^2}{4\pi}.
\end{aligned} \tag{90}$$

To obtain the variance, we evaluate $\mathbb{E}\left(\|\nabla L^{(\mathtt{n})}(\mathbf{x}) - \nabla L^{(\mathtt{c})}(\mathbf{x})\|^4\right)$.

[2]Note that in the general model with the patch defined in (14), the calculation is the same while the cut-off function ψ will come into play and the calculation will be tedious. For example, in (89) the noise term becomes $\Phi(\psi G_x)$ and $\Phi(\psi G_y)$, and the expectation and variance will be similar to the result, but the expression will not be explicit.

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla L^{(\mathtt{n})}(\mathbf{x})-\nabla L^{(\mathtt{c})}(\mathbf{x})\|^4\right]\\
=&\sigma^4\mathbb{E}\left[\left(\iint_{\mathbb{R}^2} G_x(\mathbf{x}-\mathbf{s})\xi(\mathbf{s})d\mathbf{s}\right)^4\right]+\sigma^4\mathbb{E}\left[\left(\iint_{\mathbb{R}^2} G_y(\mathbf{x}-\mathbf{s})\xi(\mathbf{s})d\mathbf{s}\right)^4\right]\\
&+2\sigma^4\mathbb{E}\left[\left(\iint_{\mathbb{R}^2} G_x(\mathbf{x}-\mathbf{s})\xi(\mathbf{s})d\mathbf{s}\right)^2\left(\iint_{\mathbb{R}^2} G_y(\mathbf{x}-\mathbf{s})\xi(\mathbf{s})d\mathbf{s}\right)^2\right]\\
=&\sigma^4\left[\left(3\iint_{\mathbb{R}^2}\mathbb{E}[\xi(\mathbf{s})^2]G_x(\mathbf{x}-\mathbf{s})^2d\mathbf{s}\right)^2+\iint_{\mathbb{R}^2}\mathbb{E}[\xi(\mathbf{s})^4]G_x(\mathbf{x}-\mathbf{s})^4d\mathbf{s}\right]\\
&+\sigma^4\left[\left(3\iint_{\mathbb{R}^2}\mathbb{E}[\xi(\mathbf{s})^2]G_y(\mathbf{x}-\mathbf{s})^2d\mathbf{s}\right)^2+\iint_{\mathbb{R}^2}\mathbb{E}[\xi(\mathbf{s})^4]G_y(\mathbf{x}-\mathbf{s})^4d\mathbf{s}\right]\\
&+2\sigma^4\left(\iint_{\mathbb{R}^2}\mathbb{E}[\xi(\mathbf{s})^2]G_x(\mathbf{x}-\mathbf{s})^2d\mathbf{s}\right)\left(\iint_{\mathbb{R}^2}\mathbb{E}[\xi(\mathbf{s})^2]G_y(\mathbf{x}-\mathbf{s})^2d\mathbf{s}\right)\\
&+4\sigma^4\left(\iint_{\mathbb{R}^2}\mathbb{E}[\xi(\mathbf{s})^2]G_x(\mathbf{x}-\mathbf{s})G_y(\mathbf{x}-\mathbf{s})d\mathbf{s}\right)+2\sigma^4\iint_{\mathbb{R}^2}\mathbb{E}[\xi(\mathbf{s})^4]G_x(\mathbf{x}-\mathbf{s})^2G_y(\mathbf{x}-\mathbf{s})^2d\mathbf{s}\\
=&\sigma^4\left(\frac{1}{8\pi^2}+\frac{27}{256\pi^3}\right). \qquad (91)
\end{aligned}
$$

Therefore,

$$
\mathrm{Var}\left[\|\nabla L^{(\mathtt{n})}(\mathbf{x})-\nabla L^{(\mathtt{c})}(\mathbf{x})\|^2\right]=\frac{\sigma^4}{16\pi^2}\left(1+\frac{27}{16\pi}\right). \qquad (92)
$$

□

Corollary 18. *With high probability, we have that* $\|\nabla L^{(\mathtt{n})}(\mathbf{x})\|\approx\|\nabla L^{(\mathtt{c})}(\mathbf{x})\|$. *More precisely, for* $k>0$

$$
\Pr\left(\left|\|\nabla L^{(\mathtt{n})}(\mathbf{x})\|-\|\nabla L^{(\mathtt{c})}(\mathbf{x})\|\right|<\sigma\sqrt{\frac{1+1.25k}{4\pi}}\right)>1-\frac{1}{1+k^2}. \qquad (93)
$$

Proof. By Lemma 17 and the one-sided Chebyshev inequality, we can obtain

$$
\Pr\left(\|\nabla L^{(\mathtt{n})}(\mathbf{x})-\nabla L^{(\mathtt{c})}(\mathbf{x})\|^2<\frac{\sigma^2}{4\pi}+k\frac{\sigma^2}{4\pi}\left(1+\frac{27}{16\pi}\right)\right)>1-\frac{1}{1+k^2}. \qquad (94)
$$

Since $\left|\|\nabla L^{(\mathtt{n})}(\mathbf{x})\|-\|\nabla L^{(\mathtt{c})}(\mathbf{x})\|\right|<\|\nabla L^{(\mathtt{n})}(\mathbf{x})-\nabla L^{(\mathtt{c})}(\mathbf{x})\|$, and $\sqrt{1+\frac{27}{16\pi}}<1.25$, we obtain (93). □

Proposition 19. *Let* $p^{(\mathtt{n})}$ *be the associated noisy patch of a clean patch* $p^{(\mathtt{c})}$. *Denote their orientation assignments by* $\theta^{(c)*}\in S^1$ *and* $\theta^{(n)*}\in S^1$, *respectively. Suppose* σ *is small and* k *satisfies* $\sigma\sqrt{\frac{1+1.25k}{4\pi}}<\epsilon$, *where* $\epsilon>0$. *With the probability higher than* $1-\frac{1}{1+k^2}$, *we have*

$$
d_{S^1}(\theta^{(c)*},\theta^{(n)*})<\delta, \qquad (95)
$$

where δ *is the number in Definition 14.*

Proof. By Corollary 18, the probability that

$$
\|\nabla L^{(\mathtt{c})}\|-\sigma\sqrt{\frac{1+1.25k}{4\pi}}<\|\nabla L^{(\mathtt{n})}\|<\|\nabla L^{(\mathtt{c})}\|-\sigma\sqrt{\frac{1+1.25k}{4\pi}} \qquad (96)
$$

is higher than $1-\frac{1}{1+k^2}$.

Suppose that $d_{S^1}(\theta^{(c)*}, \theta^{(n)*}) > \delta$. Then

$$\begin{aligned}\iint_{\Psi^{-1}(N_{\theta^{(n)*}})} \|\nabla L^{(\mathtt{n})}(\mathbf{x})\| d\mathbf{x} &> \iint_{\Psi^{-1}(N_{\theta^{(c)*}})} \|\nabla L^{(\mathtt{n})}(\mathbf{x})\| d\mathbf{x} + \epsilon\pi r^2 \\ &> \iint_{\Psi^{-1}(N_{\theta^{(c)*}})} \Big(\|\nabla L^{(\mathtt{c})}(\mathbf{x})\| - \sigma\sqrt{\frac{1+1.25k}{4\pi}}\Big) d\mathbf{x} + \epsilon\pi r^2, \end{aligned} \tag{97}$$

where Ψ is defined as (80). On the other hand,

$$\iint_{\Psi^{-1}(N_{\theta^{(n)*}})} \|\nabla L^{(\mathtt{n})}(\mathbf{x})\| d\mathbf{x} < \iint_{\Psi^{-1}(N_{\theta^{(c)*}})} \Big(\|\nabla L^{(\mathtt{c})}(\mathbf{x})\| + \sigma\sqrt{\frac{1+1.25k}{4\pi}}\Big) d\mathbf{x}. \tag{98}$$

Combining the two inequalities and the assumption $\sigma\sqrt{\frac{1+1.25k}{4\pi}} < \epsilon$, we have

$$\iint_{\Psi^{-1}(N_{\theta^{(n)*}})} \|\nabla L^{(\mathtt{c})}(\mathbf{x})\| < \iint_{\Psi^{-1}(N_{\theta^{(c)*}})} \|\nabla L^{(\mathtt{c})}(\mathbf{x})\| d\mathbf{x} + \epsilon\pi r^2 \tag{99}$$

which leads to a contradiction to Assumption 15. Hence $d_{S^1}(\theta^{(c)*}, \theta^{(n)*}) < \delta$. □

Remark 20. *Definition 14 corresponds to the case where there is only one orientation feature in the SIFT algorithm. We may generalize the definition that allows two (or more) orientation features as the following:*

For fixed small positive numbers δ and ϵ, angles $\theta_1^*, \theta_2^* \in S^1$ and $d_{S^1}(\theta_1^*, \theta_2^*) > \delta$ are orientations if

$$\iint_{\Psi^{-1}(N_{\theta_i^*})} \|\nabla L(\mathbf{x})\| d\mathbf{x} > \iint_{\Psi^{-1}(N_{\theta'})} \|\nabla L(\mathbf{x})\| d\mathbf{x} + \epsilon\pi r^2 \tag{100}$$

for any $\theta' \in S^1$ outside of $N_{\theta_1^*}$ and $N_{\theta_2^*}$. For a patch with two orientations, we can prove that with high probability, the orientations of the associated noisy patch will be close to the ones of the clean patch.

5. Image Quality Assessment

Image quality assessment (IQA) is an important subfield in image processing. The goal is to find an index quantifying "how good" an image is, which is suitable for different scenarios. We consider measures of two major categories in this paper to evaluate the VNLEM algorithm. The first category consists of objective measures based on a chosen theoretical model without taking the human visual system (HVS) into account. The second category consists of objective measures based on models taking the HVS into account. Below we summarize these measures. Denote the clean image as $I \in \mathbb{R}^{N\times N}$. We are concerned with how close the noisy observation $I + \sigma\xi \in \mathbb{R}^{N\times N}$ is to I, or the denoised image $\tilde{I} \in \mathbb{R}^{N\times N}$ is to I.

The signal-to-noise ratio (SNR) belongs to the first category, and is given in decibels. By denoting

$$E := \left[\sum_{i=1,\dots,N^2} (\tilde{I}(i) - I(i))^2\right]^{1/2}, \tag{101}$$

the SNR is defined as

$$\mathtt{SNR} = 20\log_{10}\big(\frac{\sigma_I}{E}\big), \tag{102}$$

where σ_I is defined in (28) and assumed to be 1. Clearly, if the denoising algorithm can fully recover the clean image; that is $\tilde{I} = I$, then the SNR is ∞. The peak-signal-to-noise ratio (PSNR) also belongs to the first category, which is given in decibels:

$$\mathtt{PSNR} = 20\log_{10}\big(\frac{p_I}{E}\big), \tag{103}$$

where

$$p_I := \max_{i=1,\dots,N^2} |I(i)|. \tag{104}$$

The SNR gives us a sense of how strong the signal and the noise are, but if the image is rather homogenous, the SNR is not very informative. The PSNR is a lot more content dependent, and it gives us a sense of how well the high-intensity regions of the image is coming through the noise i.e. the contrast. Since the denoising filter can adjust the contrast of the image, the PSNR can be rather helpful in demonstrating the performance of the various denoising filters. While SNR and PSNR are widely applied IQA's in the field, they do not necessarily tell us all aspects of how well the denoising methods performed. For example, they do not readily capture the edge preserving capability of an algorithm.

To capture the edge preservation performance, we consider the third measurement, the *Sobolev index* [42], which also belongs to the first category. Let $\hat{I}$ and $\hat{\tilde{I}}$ denote the discrete Fourier transforms of I and $\tilde{I}$, respectively. The Sobolev index of order κ is then defined by the Sobolev norm, and is given by

$$\texttt{SOB} = \left[\frac{1}{|\Omega|^2} \sum_{\omega \in \Omega} (1 + |\eta_\omega|^2)^\kappa |\hat{I}(\omega) - \hat{\tilde{I}}(\omega)|^2 \right]^{1/2}, \tag{105}$$

where Ω is the lattice of the frequency domain and η_ω is the two-dimensional frequency vector associated with $\omega \in \Omega$.

The SNR, PSN, and the Sobolev norm aim to evaluate how close the denoised image is to the clean image. We further consider the earth mover's distance (EMD) to measure how well we could recover the noise [39, Section 2.2]. The EMD between two probability distributions μ and ν on $\mathbb{R}$ is defined as

$$d_{\texttt{OT}}(\mu, \nu) := \int_{\mathbb{R}} |f_\mu(x) - f_\nu(x)| \, dx \,,$$

where $f_\mu(x) := \int_{-\infty}^{x} d\mu$ is the cumulative distribution function of μ and similarly for f_ν. We will evaluate the EMD to compare how close the distribution of the estimated noise is to the added noise.

The above measurements are designed mainly around the idea of "how well the error is captured", or "error sensitivity" [41]. While they have been widely applied in different problems and provide useful information, it has been well accepted that they do not capture all aspects from the perspective of image quality. Particularly, generally it is not statistically consistent with human observers [47]. Several metrics have been designed in the past decades to faithfully take the HVS into account, and they belong to the second category. These metrics emphasize the importance of luminance, the contrast, and the frequency/phase content. To further evaluate the performance of VNLEM, we consider the state-of-art measurement in this category, the Feature SIMilarity (FSIM) index [47]. The FSIM is based on the model that the HVS perceives an image mainly based on its low-level features, such as edges and zero crossings, and it separates the similarity measurement task into phase congruency and gradient magnitude. Here we summarize the FSIM index. Suppose the dynamical range of the image is $\mathcal{R}$. The definition of FSIM depends on the definition of the phase congruency and gradient magnitude. The phase congruent of I at i, denoted as $P_I(i)$, and the gradient magnitude of I at i, denoted as $G_I(i)$, are defined in [47, Equation (3) and Section II.B]. Similarly, we could define $P_{\tilde{I}}(i)$ and $G_{\tilde{I}}(i)$. The FSIM between I and $\tilde{I}$ is defined as

$$\texttt{FSIM}(I, \tilde{I}) := \frac{\sum_{i=1}^{N^2} S_L(i) P_m(i)}{\sum_{i=1}^{N^2} P_m(i)}, \tag{106}$$

where

$$P_m(i) = \max\{P_I(i), P_{\tilde{I}}(i)\}, \quad S_L(i) := S_P(i) S_G(i), \tag{107}$$

$$S_P(i) := \frac{2P_I(i)P_{\tilde{I}}(i) + T_1}{P_I(i)^2 + P_{\tilde{I}}(i)^2 + T_1}, \quad \text{and } S_G(i) := \frac{2G_I(i)G_{\tilde{I}}(i) + T_2}{G_I(i)^2 + G_{\tilde{I}}(i)^2 + T_2}. \tag{108}$$

Here, we follow [47] and choose $T_1 = 0.85$ and $T_2 = 160$. There are several other measures of this kind in the field, and we refer interested readers to [41, 47] for a review of these indices.

6. Numerical Result

In our numerical experiments, we fix the following parameters for NLEM, VNLEM, and VNLEM-DD for a fair comparison. Fix $q = 13$. We build 13×13 patches around each pixel of the noisy image. We chose $\epsilon = (16.5)^2$, the number of nearest neighbors as $N_1 = 100$, the size of the search window for creating the initial affinity matrix is determined by $N_2 = 10$; that is, 21×21 neighbours of each patch are chosen for the search window. The θ_l in (34) is set to 30 degrees, the upsampling operator U_k is implemented by the bicubic interpolation, and k is set to 2. After building the transition matrix, we choose $m = 30$ to evaluate the DM and DD. Finally, we select $\gamma = 0.1$ for the final denoising step. The Matlab code is available via request.

To compare our results with those of the NLEM algorithm, we also preformed the NLEM denoising with $\epsilon = (6.5)^2$[3], where the search window and patch sizes are chosen to be identical to those selected for our proposed schemes. The kernel bandwidth is chosen to give the best performance for the NLEM algorithm in terms of SNR and PSNR.

In Table 2 we report the different IQA metrics, including SNR, PSNR, RMS, SOB, and FSIM discussed previously as well as the computational time, by running the three denoising algorithms on 1,361 sample images of size 512×512[4]. There are 98 images for animals, 143 images for flowers, 52 images for fruits, 115 images for landscapes, 450 images for faces, 419 images for manmade structures, and 44 miscellaneous images. The SOB metric is applied to the image recovery error. This measure particularly reflects the amount of edge information wiped out due to the denoising process. Therefore, the scheme with a lower SOB index performs the better. For the other indices, the higher the index is, the better the performance is. Under the null hypothesis that the performance of two algorithms is the same, we reject the hypothesis by the Mann-Whitney U test with the p value less than 10^{-4}. Note that based on the overall statistics, VNELM and VNLEM-DD outperform NLEM statistically significantly on all IQA metrics. On the other hand, we cannot distinguish the performance of VNLEM and VNLEM-DD statistically, except on the FSIM index. This result suggests that VNLEM-DD could better recover features sensitive to HVS.

The execution times based on 17 images are 501.8 ± 203.3s, 1489.8 ± 26.1s, and 1619 ± 34.8s for NLEM, VNLEM, and VNLEM-DD respectively. This execution time is obtained on a PC with 8 Gb of RAM using a single core from Intel Corei7 CPU with a clock speed of 3.7 GHz running on Microsoft Windows 7.

Table 2. Summary statistics over 1,361 images of different denoising algorithms evaluated by different image quality assessment metrics. $^{*\#}$: $p < 10^{-8}$. †: $p < 10^{-6}$. a.u.: the arbitrary unit.

	NLEM	**VNLEM**	**VNLEM-DD**
PSNR (dB)	$18.78 \pm 2.92^{*\#}$	$19.49 \pm 2.72^{*}$	$19.62 \pm 2.81^{\#}$
SNR (dB)	$13.33 \pm 2.78^{*\#}$	$14.04 \pm 2.36^{*}$	$14.18 \pm 2.49^{\#}$
RMS$\times$100 (a.u.)	$5.77 \pm 1.58^{*\#}$	$5.35 \pm 1.33^{*}$	$5.24 \pm 1.36^{\#}$
SOB$\times$100 (a.u.)	$5.9 \pm 1.63^{*\#}$	$5.45 \pm 1.37^{*}$	$5.35 \pm 1.4^{\#}$
OT$\times$100 (a.u.)	$0.59 \pm 0.4^{*\#}$	$0.32 \pm 0.16^{*}$	$0.35 \pm 0.22^{\#}$
FSIM$\times$100 (a.u.)	$88.33 \pm 2.98^{*\#}$	$89.64 \pm 2.13^{*\dagger}$	$90.03 \pm 2.13^{\#\dagger}$

Fig. 8 depicts an example of noisy image recovery performed using three different denoising algorithms, namely the NLEM, the VNLEM and the VNLEM-DD algorithms. The original image is of size 512×512. In this figure, we have also presented the denoising error for each scheme. These errors help us identify the amount of details and desired features that are lost in the image recovery process. For this and the consequent

[3]The code is available in `https://www.mathworks.com/matlabcentral/fileexchange/40624-non-local-patch-regression`

[4]The images are collected from :

- Caltech-UCSD Birds-200-2011 collection at : `http://www.vision.caltech.edu/visipedia/CUB-200-2011.html`
- Digital Image Processing, 3rd ed, by Gonzalez and Woods at : `http://www.imageprocessingplace.com/DIP-3E/dip3e_book_images_downloads.htm`
- USC-SIPI image database at: `http://sipi.usc.edu/database/`

examples, we have also reported the PSNR, SNR, SOB, and FSIM values achieved by the denoising process. The PSNR, SNR, SOB, and FSIM all suggest that by taking the rotational fiber structure into account, the proposed algorithm improves the NLEM scheme in terms of the amount of details preserved in the recovered image. A close look at the results of the recovered image suggests that the VNLEM-DD algorithm preserves the most amount of texture features present in the image. The smaller SOB's of VNLEM and VNLEM-DD indicate a better edge preservation. The visual perception improvement is captured by the higher FSIM.

Similar observations can be made in our second example in Fig. 9.[5] The visual perception improvement by reading Fig. 9 is supported by the higher FSIM. By looking at the denoising errors one can notice that the details of the edges are lost in all three schemes. However, while the VNLEM and VNLEM-DD algorithms lead to higher SNR and PSNR, and edge preservation; this fact is quantified by the larger SOB.

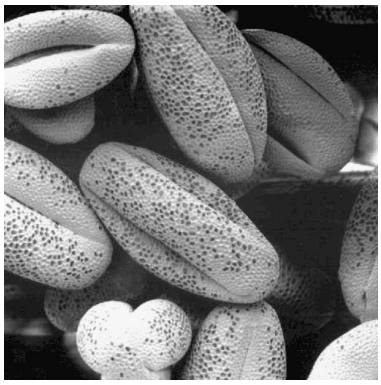
While statistically VNLEM and VNLEM-DD outperform NLEM, there are cases where NLEM outperforms. We now take a closer look into some of these examples. In Fig. 10, we see that the NLEM scheme achieves higher PSNR and SNR values in the recovered image. A second look at the recovery errors reveals that this superior performance comes at the cost of substantial loss of edge details in the results, and this is reflected in the higher SOB metric. Specifically, note that the teeth are better recovered in the VNLEM-DD. This result supports that the PSNR and SNR measures alone cannot throughly represent the performance of a denoising scheme, and IQA's from different perspectives are needed to better quantify the performance. It is also worth noting that the VNLEM-DD algorithm introduces "texture-like artifices" in the areas of the image that do not manifest any distinct feature, for example, the forehead of the portrait shown in Fig. 10. This comes from the following facts. Note that the clean patches associated with this region are concentrated at one point in the high dimensional space $\mathbb{R}^{q^2}$ and the added noise creates a geometric pattern (see, for example, the discussion in [14]) that is irrelevant to the underlying image itself. Thus, the DD provides a deviated neighbors for the median filter. We thus have to be careful when applying VNLEM-DD on images with this type of "flat region".

Another example worth looking into is presented in Fig. 11. This image contains substantial fine details that should be preserved during the recovery. Looking into results of the three different denoising schemes, one can see that in such an image, the NLEM scheme outperforms the VNLEM and VNLEM-DD in terms of SNR and PSNR as well as the level of details kept in the process, like the SOB and FSIM. This example shows that while the VNLEM and VNLEM-DD overall outperforms the NLEM statistically, there are examples where the NLEM performs better.

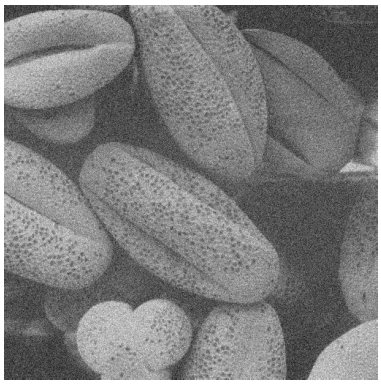
One interesting parameter that influences the image denoising performance is the "image resolution". Note that the "image resolution" is not a well-defined term, and in this example it means the number of pixels in the image – the more pixels there are in an image, the higher the image resolution is. Equivalently, an image with a higher resolution means a denser sampling of the image function. In Fig. 12, we take the starfish image shown in Fig. 9 and show how the image resolution affects the final result. In this figure, we present the outputs of the VNLEM-DD algorithm (left column) and the NLEM algorithm (right columns) for $N = 200, 512, 1024$. It can be clearly seen that in all three cases, the VNLEM-DD algorithm produces a more clean image compared to the NLEM scheme, and the performance of each algorithm increases as the resolution increases.

In this last example, we apply the developed VNLEM to the third-harmonic-generation (THG) microscopy image. The goal of the THG microscopy-based imaging cytometry is to automatically differentiate and count different types of blood cells with less blood *ex vivo*, or even *in vivo* [43]. One of the many strengths of THG is reflecting the granularity of leukocytes, which allows us to apply image processing techniques for the automatic classification. However, the raw data is noisy most of time, and a denoising technique is needed. We now apply the NLEM, VNLEM, and VNLEM-DD to the THG sectioning image of the whole blood smear at 1 hour post blood sampling. The data is provided by Professor Tzu-Ming Liu, Faculty of Health Sciences, University of Macau. The result is shown in Figure 13. Note that since we do not have the "ground truth" for a comparison, we only show the FSM for the quality evaluation purpose. Note that while the result is

[5]The original image can be found at `https://wall.alphacoders.com/big.php?i=109992`.

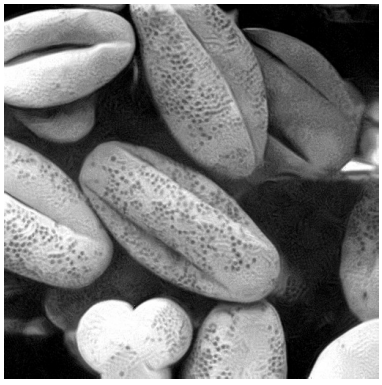


(a) Original image



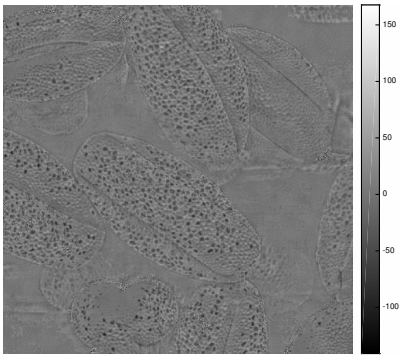
(b) Noisy image, PSNR = 11.47, SNR = 5.24

(c) NLEM, PSNR = 19.37, SNR = 13.14, FSIM = 0.853

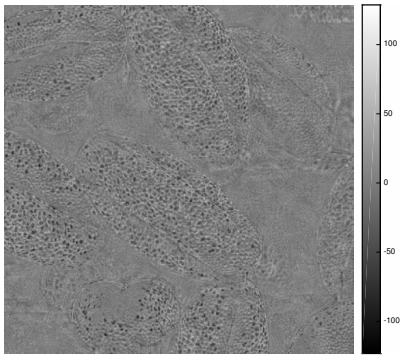


(d) VNLEM, PSNR = 20.22, SNR = 13.99, FSIM = 0.883

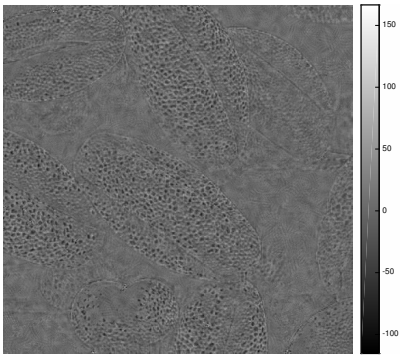
(e) VNLEM-DD, PSNR = 20.07, SNR = 13.84, FSIM = 0.877



(f) NLEM, difference, SOB = 0.166



(g) VNLEM, difference, SOB = 0.059



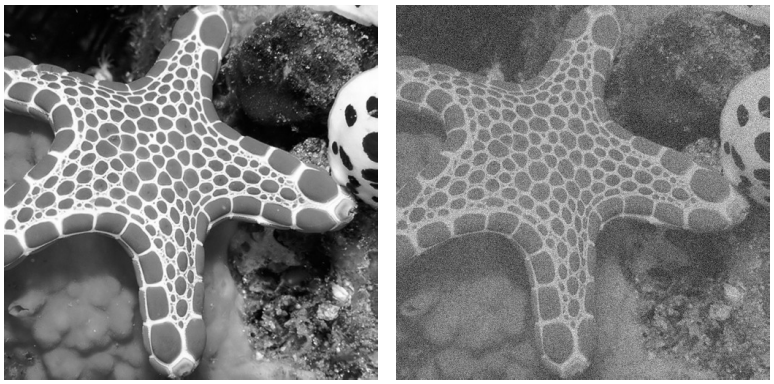
(h) VNLEM-DD, difference, SOB = 0.06

FIGURE 8. Example 1: the beans

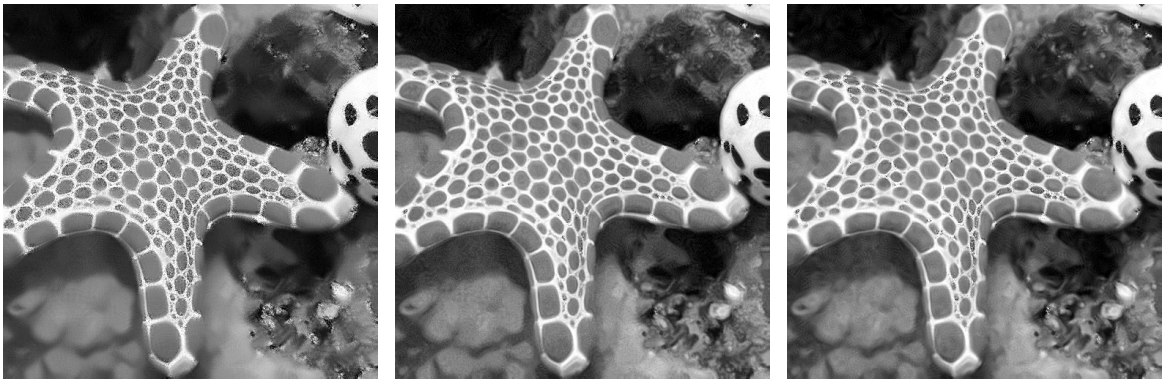
encouraging, a systematical study of the problem, and a systematic comparison of the proposed algorithm with existing algorithms is needed. The result will be reported in a future work.

7. Conclusion and Discussion

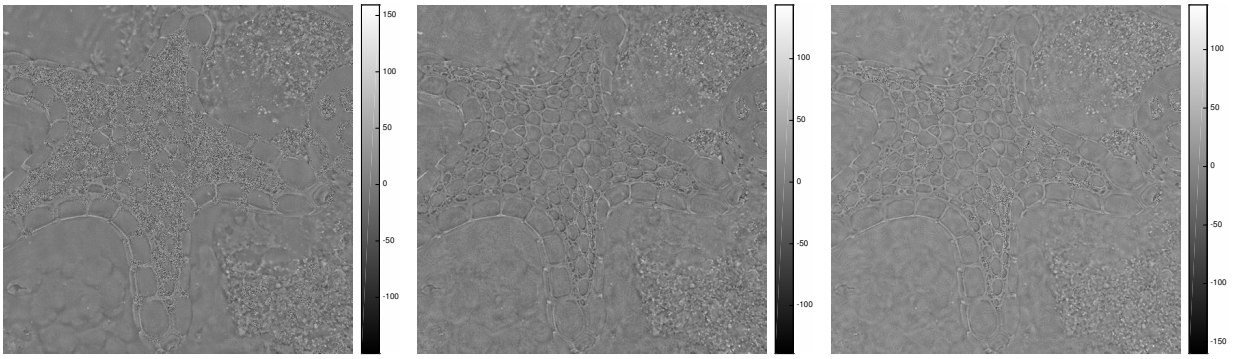
In this work, we propose a fiber bundle structure to model the patch space, and take the fiber structure to generalize the commonly used NLEM algorithm to the VNLEM/VNLEM-DD algorithm. One main benefit of introducing the fiber structure is the dimension reduction. To speed up the VNLEM algorithm and stabilize the numerical rotation on a small patch, different numerical techniques are applied, including the



(a) Original image (b) Noisy image, PSNR = 11.60, SNR = 5.25



(c) NLEM, PSNR = 16.79, SNR = 10.42, FSIM = 0.902 (d) VNLEM, PSNR = 19.68, SNR = 13.32, FSIM = 0.914 (e) VNLEM-DD, PSNR = 19.49, SNR = 13.13, FSIM = 0.918



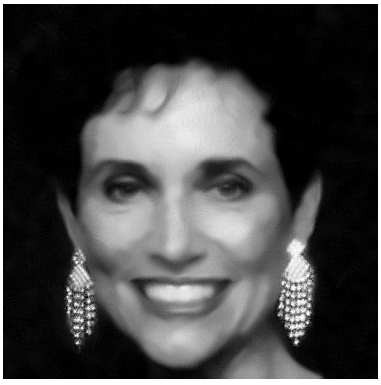
(f) NLEM, difference, SOB = 0.047 (g) VNLEM, difference, SOB = 0.057 (h) VNLEM-DD, difference, SOB = 0.057

FIGURE 9. Example 2: The starfish.

search window and the SIFT features. The numerical simulation provides positive evidence of the potential of the proposed algorithm. In addition to providing the theoretical justification of how the VNLEM and the NLEM work, particularly why we could accurately find nearest neighbors from the noisy patches, we study the stability of the widely applied SIFT algorithm. Both theoretical results support how the proposed VNLEM algorithm works. The potential of the proposed model, algorithms, and the associated theory are statistically supported by a large image database composed of 1,361 images. Below, we discuss the limitations of the current work and several future works.

(a) Original image

(b) Noisy image, PSNR = 13.55, SNR = 5.23



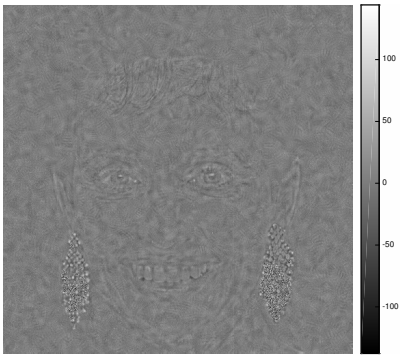
(c) NLEM, PSNR = 25.68, SNR = 17.36, FSIM = 0.928

(d) VNLEM, PSNR = 25.48, SNR = 17.16, FSIM = 0.897

(e) VNLEM-DD, PSNR = 25.54, SNR = 17.21, FSIM = 0.898

(f) NLEM, difference, SOB = 0.043

(g) VNLEM, difference, SOB = 0.039

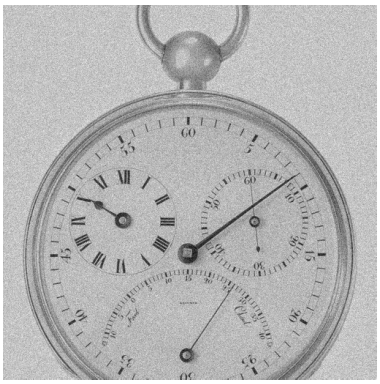


(h) VNLEM-DD, difference, SOB = 0.038

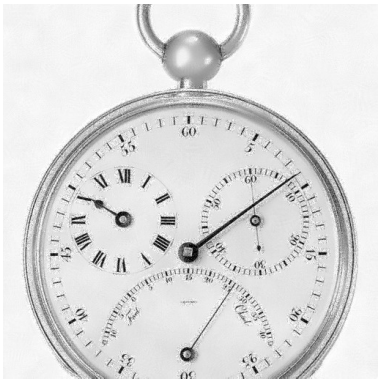
FIGURE 10. Example 3: the lady.

First, the computational complexity needs to be further improved. Note that the main difference between the VNLEM and the NLEM algorithms is the chosen metric. In the NLEM, since the L^2 distance is chosen to evaluate the similarity of two patches, there are several fast algorithms available to evaluate the nearest neighbors. However, in the VNLEM, there does not exist a fast algorithm to determine the nearest neighbors with respect to the RID, to the best of our knowledge. Although we have delegated the problem of evaluating the RID distance to that of evaluating the SIFT distance, the numerical performance still has a significant room for improvement.

(a) Original image

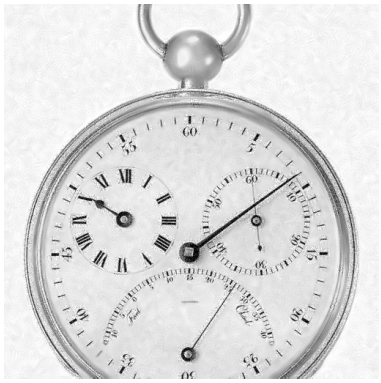


(b) Noisy image, PSNR = 13.55, SNR = 5.23

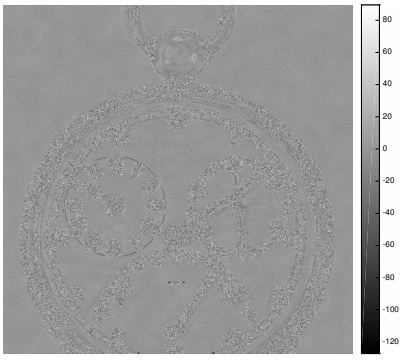


(c) NLEM, PSNR = 25.68, SNR = 17.36, FSIM = 0.952

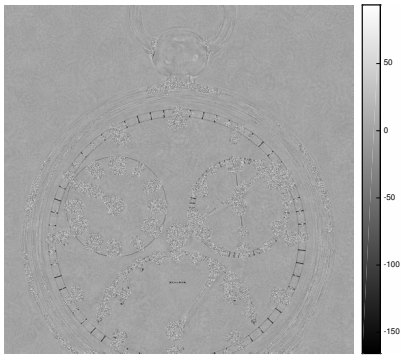
(d) VNLEM, PSNR = 25.48, SNR = 17.16, FSIM = 0.944



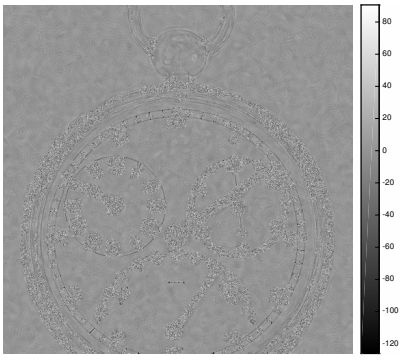
(e) VNLEM-DD, PSNR = 25.54, SNR = 17.21, FSIM = 0.943



(f) NLEM, difference, SOB = 0.033



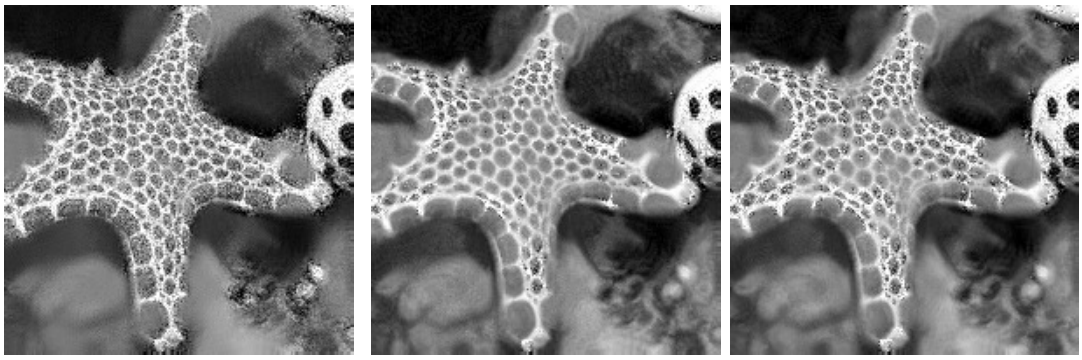
(g) VNLEM, difference, SOB = 0.042



(h) VNLEM-DD, difference, SOB = 0.044

FIGURE 11. Example 4: the clock.

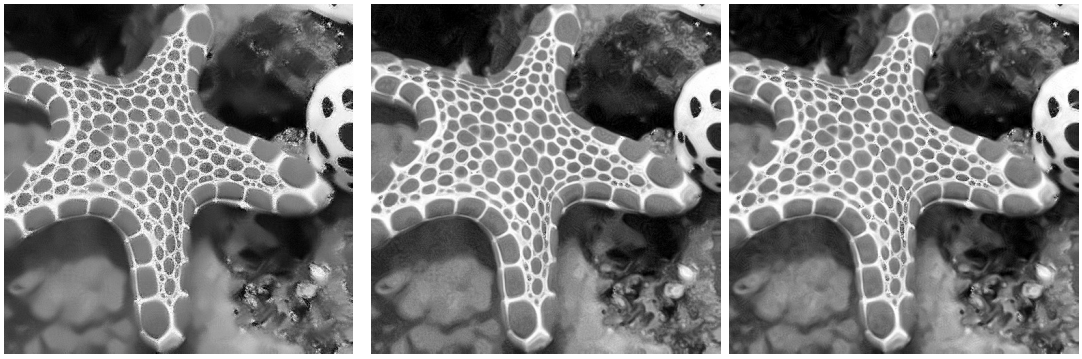
Second, although the manifold model has been widely accepted in the field, and our algorithm is also based on the manifold structure, it is certainly arguable if in general a patch space could be well approximated by a manifold. On one hand, we need to consider a more general model than the fiber bundle; on the other hand, for different problems we may want to better understand its associated manifold structure, if there is any. In other words, we might need different models, and hence different metrics, for different kinds of images. For example, while the RID helps reduce the dimension of the patch space of a "structured" image, its deterministic nature might render it unsuitable for analyzing a "texture" image, since the texture features are stochastic in nature. In short, it might be beneficial to take the metrics designed for the texture



(a) NLEM, $N = 200$, PSNR = 14.90, SNR = 8.05, FSIM = 0.812

(b) VNLEM, $N = 200$, PSNR = 16.79, SNR = 9.93, FSIM = 0.829

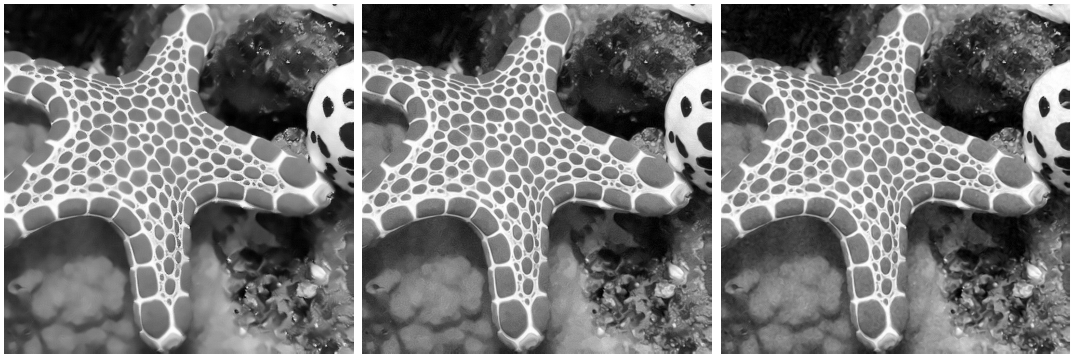
(c) VNLEM-DD, $N = 200$, PSNR = 16.19, SNR = 9.33, FSIM = 0.819



(d) NLEM, $N = 512$, PSNR = 16.79, SNR = 10.43, FSIM = 0.902

(e) VNLEM, $N = 512$, PSNR = 19.68, SNR = 13.32, FSIM = 0.914

(f) VNLEM-DD, $N = 512$, PSNR = 19.49, SNR = 13.13, FSIM = 0.918



(g) NLEM, $N = 1024$, PSNR = 21.11, SNR = 14.78, FSIM = 0.953

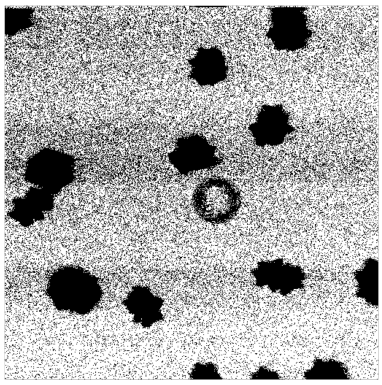
(h) VNLEM, $N = 1024$, PSNR = 22.50, SNR = 16.16, FSIM = 0.959

(i) VNLEM-DD, $N = 1024$, PSNR = 22.94, SNR = 16.60, FSIM = 0.961

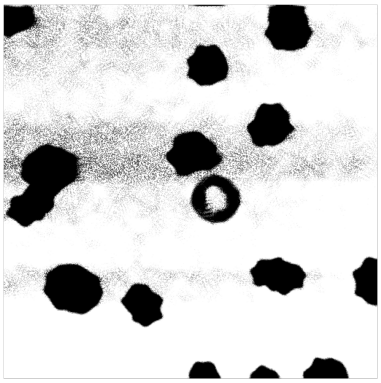
FIGURE 12. The starfish in Figure 9 with the same noise level but different resolutions.

analysis into account. On the other hand, we could consider to segment the given noisy image into different categories, and run the VNLEM on each category. This segmentation step is related to the "multi-manifold model" considered in the literature [44, 40], and could be understood as a generalization of the search window method used in this paper.

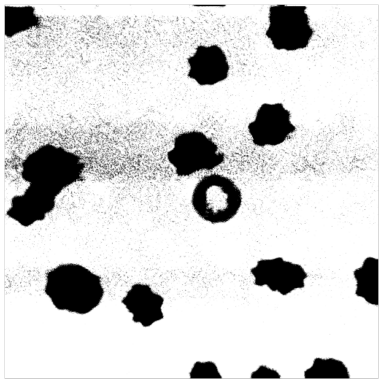
Third, we should consider different structures in the denoising procedure. In addition to taking the rotation group to fibrate the patch space, it is an intuitive generalization to further consider other groups, like the dilation group or even the general linear group. Also, while the current work focuses on grayscale images, the proposed algorithm has the potential to be generalized to colored images. In colored images,



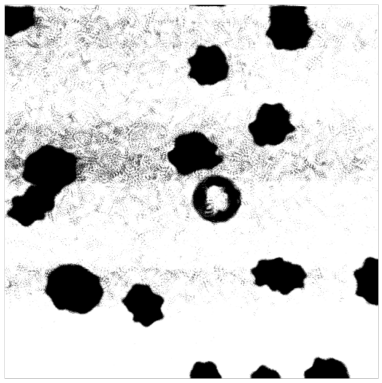
(a) Original image



(b) NLEM, FSIM = 0.947



(c) VNLEM, FSIM = 0.967



(d) VNLEM-DD, FSIM = 0.97

FIGURE 13. The cytometry image. Since the "ground truth" is not available for a comparison, we only show the FSIM for the quality evaluation purpose.

more structures, like the color space, will be taken into account. Furthermore, in practice we would expect to have more than one image from the practical problem. Under the assumption that the noise behavior is similar, it is of great interest to see if we could further improve the denoise performance by denoising multiple available images simultaneously.

Fourth, note that the proposed algorithm could be understood as aiming to reduce the error introduced to the clean image. However, it has been widely argued in the IQA society that simply reducing the error might not lead to the optimal result in all scenarios. It might be more important to take the human perception into account, if the images are meant to be watched by a human being. While the proposed VNLEM provides a satisfactory result by the FSIM evaluation, note that the "features" considered in the FSIM are not used in the algorithm. It is reasonable to expect that by taking these features into account, we could further improve the result.

Fifth, in this paper we focus only on comparing our algorithm with the NLEM to study the corresponding diffusion property and the geometric structure of the underlying patch space model. For the image denoising purpose, there are several other image denoising algorithms available in the field, and we will do a systematic comparison in a upcoming report. For example, while not specifically indicated, the widely used algorithm block-matching and 3-D filtering (BM3D) [10] and its generalizations, for example [22], are also based on the patch space model. We could view the sparsity structure used in BM3D as a different way to design a "metric" to compare different patches.

Last but not least, although we compared the algorithm on a big image database and reported the statistical significance, note that statistical significance does not imply practical significance. Particularly, the included images are not exhaustive. A more systematic comparison is thus needed. In practice, the overall

performance might depend on the problems encountered, and the specific applications, like the cytometry problem, will be discussed in a upcoming research report.

Acknowledgement

Hau-tieng Wu's research is partially supported by Sloan Research Fellow FR-2015-65363 and partially by Connaught New Researcher grant 498992. He would like to thank the valuable discussions with Professor Ingrid Daubechies and Professor Amit Singer. Chen-Yun Lin would like to thank Professor Chiahui Huang for her helpful discussions. The authors would like to thank Professor Tzu-Ming Liu for sharing the cytometry image.

References

[1] B. Coll A. Buades and J. M. Morel. Image denoising methods. a new nonlocal principle. *SIAM Review*, 52(1):113–147, 2010.
[2] A. Buades and B. Coll. A non-local algorithm for image denoising. In *CVPR*, pages 60–65, 2005.
[3] P. Burt and T. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 9(4):532–540, 1983.
[4] G. Carlsson, T. Ishkhanov, V. Silva, and A. Zomorodian. On the Local Behavior of Spaces of Natural Images. *International Journal of Computer Vision*, 76(1):1–12, 2007.
[5] C. Chan, R. Fulton, D. D. Feng, and S. Meikle. Median non-local means filtering for low SNR image denoising: Application to PET with anatomical knowledge. *IEEE Nuclear Science Symposium Conference Record*, pages 3613–3618, 2010.
[6] K. N. Chaudhury and A. Singer. Non-local euclidean medians. *IEEE Signal Processing Letters*, 19(11):745–748, 2012.
[7] K. N. Chaudhury and A. Singer. Non-local patch regression: Robust image denoising in patch space. *ICASSP*, pages 1345–1349, 2013.
[8] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5 – 30, 2006.
[9] J. L. Crowley and R. M. Stern. Fast computation of the difference of low pass transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):212–222, 1984.
[10] K. Dabov, A. Foi, and V. Katkovnik. Image denoising by sparse 3D transformation-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):1–16, 2007.
[11] D. L. Donoho and C. Grimes. Image manifolds which are isometric to euclidean space. *Journal of Mathematical Imaging and Vision*, 23(1):5–24, 2005.
[12] T. Eguchi, P. B. Gilkey, and A. J. Hanson. Gravitation, gauge theories and differential geometry. *Physics Reports*, 66(6):213 – 393, 1980.
[13] N. El Karoui. On information plus noise kernel random matrices. *The Annals of Statistics*, 38(5):3191–3216, 2010.
[14] N. El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.
[15] N. El Karoui and H.-T. Wu. Connection graph Laplacian methods can be made robust to noise. *The Annals of Statistics*, 44(1):346–372, 2016.
[16] J. Gallier. Notes on group actions manifolds, lie groups and lie algebras. `http://www.cis.upenn.edu/~cis610/lie1.pdf`, 2005. Accessed: 2016-11-21.
[17] I. Gel'fand and N. Ya. Vilenkin. *Generalized function theory Vol 4*. Academic Press, 1964.
[18] S. Gepshtein and Y. Keller. Image completion by diffusion maps and spectral relaxation. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 22(8):2983–94, 2013.
[19] S. Grewenig, S. Zimmer, and J. Weickert. Rotationally invariant similarity measures for nonlocal image denoising. *Journal of Visual Communication and Image Representation*, 22(2):117–130, 2011.
[20] N. Guizard, P. Coupe, V. S. Fonov, J. V. Manjon, D. L. Arnold, and D. L. Collins. Rotation-invariant multi-contrast non-local means for MS lesion segmentation. *NeuroImage: Clinical*, 8:376–389, 2015.
[21] P. Jain and V. Taygi. A survey of edge-preserving image denoising methods. *Information Systems Frontiers*, 18:159–170, 2016.
[22] V. Katkovnik, A. Foi, K. Egiazarian, and J. Astola. From local kernel to nonlocal multiple-model image denoising. *International Journal of Computer Vision*, 86(1):1–32, 2010.
[23] A. Lee, K. Pedersen, and D. Mumford. The nonlinear statistics of high-contrast patches in natural images. *International Journal of Computer Vision*, 54:83–103, 2003.
[24] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer/Springer, Boston, 1994.
[25] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1998.
[26] T. Lindeberg. Scale Invariant Feature Transform. *Scholarpedia*, 7(5):10491, 2012. revision #153939.
[27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.
[28] J. V. Manjón, P. Coupé, A. Buades, D. Louis Collins, and M. Robles. New methods for MRI denoising based on sparseness and self-similarity. *Medical Image Analysis*, 16(1):18–27, 2012.

[29] S. Osher, Z. Shi, and W. Zhu. Low dimensional manifold model for image processing. *tech. report, UCLA, Tech. Rep. CAM report 1604*, 2016.
[30] J. A. Perea and G. Carlsson. A klein-bottle-based dictionary for texture representation. *International Journal of Computer Vision*, 107(1):75–97, 2014.
[31] G. Peyre. Image processing with nonlocal spectral bases. *Multiscale Model. Simul.*, 7(2):703–730, 2008.
[32] G. Peyre. Manifold models for signals and images. *Computer Vision and Image Understanding*, 113(2):249–260, 2009.
[33] X. Qi. Vector Nonlocal Mean Filter. Master's thesis, University of Toronto, Toronto, Nov 2015. http://hdl.handle.net/1807/70528.
[34] A. Singer, Y. Shkolnisky, and B. Nadler. Diffusion interpretation of nonlocal neighborhood filters for signal denoising. *SIAM J. Imaging Sciences*, 2(1):118–139, 2009.
[35] S. Sreehari, S. V. Venkatakrishnan, L. F. Drummy, J. P. Simmons, and C. A. Bouman. Rotationally-invariant non-local means for image denoising and tomography. In *International Conference on Image Processing (ICIP)*, 2015.
[36] J. Stark, D.S. Broomhead, M.E. Davies, and J. Huke. Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis: Theory, Methods & Applications*, 30(8):5303–5314, 1997.
[37] L. Su and H.-T Wu. Extract fetal ECG from single-lead abdominal ECG by de-shape short time Fourier transform and nonlocal median. *arXiv:1609.02938*, 2014.
[38] F. Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer Berlin Heidelberg, 1981.
[39] C. Villanic. *Topics in Optimal Transportation*. Graduate Studies in Mathematics, American Mathematical Society, 2003.
[40] X. Wang, K. Slavakis, and G. Lerman. Riemannian Multi-Manifold Modeling. *ArXiv e-prints*, 2014.
[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
[42] D. L. Wilson, A. J. Baddeley, and R. A. Owens. A new metric for grey-scale image comparison. *International Journal of Computer Vision*, 24(1):5–17, 1997.
[43] C.-H. Wu, T.-D. Wang, C.-H. Hsieh, S.-H. Huang, J.-W. Lin, S.-C. Hsu, H.-T. Wu, Y.-M. Wu, and T.-M. Liu. Imaging cytometry of human leukocytes with third harmonic generation microscopy. *Scientific Reports*, 6(11):37210, 2016.
[44] W. Yang, C. Sun, and L. Zhang. A multi-manifold discriminant analysis method for image feature extraction. *Pattern Recognition*, 44(8):1649 – 1657, 2011.
[45] R. Yin, T. Gao, Y. Lu, and I. Daubechies. A tale of two bases: Local-nonlocal regularization on image patches with convolution framelets. *https://arxiv.org/abs/1606.01377*, 2016.
[46] D. Zhang, J. He, and M. Du. Image restoration via patch orientation-based low-rank matrix approximation and nonlocal means. *Journal of Electronic Imaging*, 25(2):023021, 2016.
[47] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.
[48] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin. Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE Transactions on Image Processing*, 21(1):130–144, 2012.
[49] S. Zimmer, S. Didas, and J. Weickert. A rotationally invariant block matching strategy improving image denoising with non-local means. *Mathematical Image Analysis Group*, pages 135 – 142, 2008.

Appendix A. Diffusion Map

To make the paper self-contained, we summarize the DM algorithm here. DM were initially introduced in [8] as a means to extract feature and reduce the dimensionality. This mapping embeds the points from the original data set, which might be high-dimensional, into a low-dimensional Euclidean space so that the geometric properties of the original dataset are less distorted. The coordinates of the embedded points are derived from the eigenvectors and eigenvalues of the *transition matrix* of the graph Laplacian associated with the data set. Below we summarize the embedding procedure. For a detailed algorithm description and a summary of the existing theorems describing the asymptotical behavior of DM, we refer the interested reader to, for example, the online supplementary of [15].

For a give point cloud $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^n$, we construct an affinity graph (V, E, w), where $V := \{x_1, x_2, \ldots, x_n\}$, E is the set of edges that is determined by the user, and $w : E \to \mathbb{R}^+$ is the affinity function defined by the user. Usually w is defined as $w_{ij} = K(\|x_i - x_j\|)$ when $(i,j) \in E$, where K is a chosen kernel, and $w_{ij} = 0$ when $(i,j) \notin E$. With the affinity graph, we have an equivalent expression of the affinity function as the $n \times n$ *affinity matrix*, defined as

$$W_{ij} = \begin{cases} w(i,j) & \text{if } (x_i, x_j) \in E \\ 0 & \text{otherwise} \end{cases} . \tag{109}$$

We then consider the *transition matrix*

$$A = D^{-1}W \tag{110}$$

where D is the *degree matrix* defined as a $n \times n$ diagonal matrix defined as

$$D_{ii} = \sum_{j=1}^{n} W_{ij}. \tag{111}$$

Note that though A may not be symmetric in general, when D is not singular, A is similar to $D^{-1/2}WD^{-1/2}$ which is symmetric and thus diagonalizable. More specifically, there exists a diagonal matrix $\Lambda \in \mathbb{R}^{n\times n}$ and an orthogonal matrix $Q \in O(n)$ such that $D^{-1/2}WD^{-1/2} = Q\Lambda Q^T$, where $\Lambda = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is the matrix of eigenvalues such that $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0$ as $W \geq 0$. Therefore, we can write A as

$$A = U\Lambda V^T \tag{112}$$

where $U = D^{-1/2}Q$ and $V = D^{1/2}Q$ and their column vectors are called right and left eigenvectors of A, respectively. In this work, we assume that D is not singular. With the above preparation, we could define the DM and diffusion distance (DD). Take a diffusion time $t > 0$. The DM $\Phi_t : V \to \mathbb{R}^m$ is defined as

$$\Phi_t^{(m)}(i) = (\lambda_2^t\phi_2(i), \lambda_3^t\phi_3(i), \ldots, \lambda_{m+1}^t\phi_{m+1}(i)), \tag{113}$$

where $\phi_1, \phi_2, \cdots, \phi_n$ are the column vectors of U and $m \in \mathbb{N}$ is determined by the user. We could view $\Phi_t^{(m)}(i)$ as a new feature representing x_i. The DD between x_i and x_j in $\mathcal{X}$ with diffusion time $t > 0$ is then defined as

$$D_t^{(m)}(i,j) := \|\Phi_t^{(m)}(i) - \Phi_t^{(m)}(j)\|. \tag{114}$$

The DD could be view as a new metric on the dataset. It has been shown in [13, 15] that the DD is robust to "big" noise, and hence suitable for us to suppress the influence of inevitable noise in our denoising problem.

Appendix B. Why could we approximate the patch space by a manifold?

While we follow the convention and assume that the patch space could at least be well approximated by a manifold, this assumption certainly deserves more discussion. While this is not the focus of this paper, we mention that the same patch space idea could be applied to study the one dimensional signal; particularly the time series. For example, the same nonlocal median filter idea has been applied to decompose the fetal electrocardiogram signal from the single-lead maternal abdominal electrocardiogram signal [37].

In this section, we provide a review of another viewpoint of "getting a manifold" inside the one dimensional time series. Precisely, we discuss a set of theorems provided in [38] and an associated embedding algorithm in the time series framework, which is exactly the patch space of the one dimensional image. The algorithm is well known as the lag map, and has been extensively applied in several fields, for example, the heart rate variability analysis in the bio-medical field.

From now on, denote M to be a d-dim compact manifold without boundary. For the sake of self-containedness, we recall the following definitions.

Definition 21 (Discrete time dynamics)**.** *By a discrete time dynamics, we mean a diffeomorhism $\varphi : M \to M$ with the time evolution $i \mapsto \varphi^i(x_0)$, $i \in \mathbb{N}$, where x_0 is the starting status.*

Definition 22 (Continuous time dynamics)**.** *By a continuous time dynamics, we mean a smooth vector field $X \in \Gamma(M)$ with the time evolution $t \mapsto \gamma_t(x_0)$, where γ_t is the integral curve with respect to X via x_0.*

To simplify the discussion, in both cases, we denote $\Phi_t(x_0)$ to be the time evolution with time $t \in \mathbb{N}$ or $\mathbb{R}$ with the starting point x_0.

Definition 23 (Observed time series)**.** *Let $\Phi_t(x_0)$ be a dynamics on M. The observation is modeled as a function $f : M \to \mathbb{R}$ and the observed time series is $f(\Phi_t(x_0))$.*

The question we have interest in with respect to the patch space formation is that if we have an observed time series $f(\Phi_t(x_0))$, whether we can recover M. Moreover, can we even recover the dynamics Φ_t? The positive answer and the precise statements are provided in the following two theorems. The proof of these theorems can be found in [38], and the noise analysis could be found in [36]. Below, by generic, we mean an open dense subset of all possible (φ, f). We mention that the theorems hold for non-compact manifolds if f is proper.

Theorem 24 (discrete time dynamics)**.** *For a pair (φ, f), $\varphi : M^d \to M^d$ is the C^2-diffeomorphism and $f \in C^2$, it is generic that the map $\Psi : M \to \mathbb{R}^{2d+1}$ given by*

$$\Psi : x \mapsto (f(x),\, f(\varphi(x)),\, f(\varphi^2(x)) \dots f(\varphi^{2d}(x)))^T \in \mathbb{R}^{2d+1}$$

is an embedding.

Theorem 25 (Continuous time dynamics)**.** *When $X \in C^2(\Gamma M)$ and $f \in C^2(M)$, it is generic that $\Psi : M \to \mathbb{R}^{2d+1}$ given by*

$$\Psi : x \mapsto (f(x),\, f(\gamma_1(x)),\, f(\gamma_2(x)) \dots f(\gamma_{2d}(x)))^T \in \mathbb{R}^{2d+1}$$

is an embedding, where $\gamma_t(x)$ is the flow of X of time t via x.

These theorems tell us that we could embed the manifold into a $(2d+1)$ dimensional Euclidean space if we have access to all dynamical behaviors from all points on the manifold. However, in practice the above model and theorem cannot be applied directly. Indeed, for a given dynamical system, most of time we may only have one or few experiments that are sampled at discrete times; that is, we only have access to one or few $x \in M$. We thus ask the following question. Suppose we have the time series

$$\{f(\Phi_{\ell\alpha}(x))\}_{\ell=0}^{N},$$

where $x \in M$ is fixed and inaccessible to us, $\alpha > 0$ is the sampling period, and $N \gg 1$ is the number of samples, what can we do? We first give the following definition.

Definition 26. *The positive limit set (PLS) of x of a vector field $X \in C^2(\Gamma M)$ is defined as*

$$L_c^+(x) := \{x' \in M |\, \exists\, t_i \to \infty,\ t_i \in \mathbb{R} \text{ such that } \gamma_{t_i}(x) \to x'\}$$

and the PLS of x of a diffeomorphism $\varphi : M \to M$ is defined as

$$L_d^+(x) := \{x' \in M |\, \exists\, n_i \in \mathbb{N} \to \infty,\ \text{such that } \varphi^{n_i}(x) \to x'\}.$$

It turns out that in this case, we should know whether under generic assumptions the topology and dynamics in the PLS of x is determined by $\{f(\Phi_{\ell\alpha}(x))\}_{\ell=0}^{\infty}$. Precisely, we have the following theorem

Theorem 27 (Continuous dynamics with 1 trajectory)**.** *Fix $x \in M$. When $X \in C^2(\Gamma M)$ with flow γ_t passing x, then there exists a residual subset $C_{X,x} \subset \mathbb{R}^+$ such that for all $\alpha \in C_{X,x}$ and diffeomorphism $\varphi := \gamma_\alpha$, the PLS $L_c^+(x)$ for flow γ_t and $L_d^+(x)$ for φ are the same; that is, for all $\alpha \in C_{X,x}$ and for all $q \in L_c^+(p)$, there exists $n_i \in \mathbb{N} \to \infty$ such that $\varphi^{n_i}(x) \to q$.*

This theorem leads to the following corollary, which is what we need to analyze the time series.

Corollary 28. *Take $x \in M$, generic $X \in C^2(\Gamma M)$ and $f \in C^2(M)$, and $a \in \mathbb{R}^+$ satisfying generic conditions depending on X and x. Denote the set*

$$\mathcal{P} := \left\{f(\gamma_{k\alpha}(x)),\, f(\gamma_{k\alpha}(x)), \dots, f(\gamma_{(k+2d)\alpha}(x))\right\}_{k=0}^{\infty}.$$

Then there exists a smooth embedding of M into $\mathbb{R}^{2d+1}$ mapping PLS L_c^+ bijectively to the set $\mathcal{P}$.

While the above model and theorems work well for the one dimensional "image" or time series, to the best of our knowledge, there is no parallel theorem for the higher dimensional statement. In the image processing setup we have interest and the patch space, we could parallel the above setup by viewing an image as an observation of a random field; that is, the temporal one-dimensional axis in the above theorems is replaced by the spatial two dimensional "time". Precisely, given a random field defined on M, an image could be viewed as an observation of the random field on M. Now, the patch space could be viewed as a "two dimensional lag map" defined on the observation, and we would expect that for a suitably chosen patch size, the patch space could be well approximated by a manifold diffeomorphic to M. However, it is not clear at this moment how to justify this statement. We thus conjecture that if an image is generated by this an observation process, then the patch space could be well modeled by a manifold.

Chen-Yun Lin, Department of Mathematics, University of Toronto
Email address: cylin@math.toronto.edu

Arin Minasian, Department of Electrical Engineering, University of Toronto
Email address: arin.minasian@mail.utoronto.ca

Xin Jessica Qi, Department of Mathematics, University of Toronto
Email address: jxin.qi@alum.utoronto.ca

Hau-Tieng Wu, Department of Mathematics, University of Toronto
Email address: hauwu@math.toronto.edu