Riesz-Quincunx-Unet Variational Auto-Encoder for Satellite Image Denoising

Duy H. Thai[1], Xiqi Fei[1], Minh Tri Le[1], Andreas Züfle[2], Konrad Wessels[1]
[1]George Mason University, Department of Geography and Geoinformation Science, USA
[2]Emory University, Department of Computer Science, USA

***Abstract*—Multiresolution deep learning approaches, such as the U-Net architecture, have achieved high performance in classifying and segmenting images. However, these approaches do not provide a latent image representation and cannot be used to decompose, denoise, and reconstruct image data. The U-Net and other convolutional neural network (CNNs) architectures commonly use pooling to enlarge the receptive field, which usually results in irreversible information loss. This study proposes to include a Riesz-Quincunx (RQ) wavelet transform, which combines 1) higher-order Riesz wavelet transform and 2) orthogonal Quincunx wavelets (which have both been used to reduce blur in medical images) inside the U-net architecture, to reduce noise in satellite images and their time-series. In the transformed feature space, we propose a variational approach to understand how random perturbations of the features affect the image to further reduce noise. Combining both approaches, we introduce a hybrid RQUNet-VAE scheme for image and time series decomposition used to reduce noise in satellite imagery. We present qualitative and quantitative experimental results that demonstrate that our proposed RQUNet-VAE was more effective at reducing noise in satellite imagery compared to other state-of-the-art methods. We also apply our scheme to several applications for multi-band satellite images, including: image denoising, image and time-series decomposition by diffusion and image segmentation.**

Keywords: Quincunx wavelet, high order Riesz transform, image time series decomposition, variational auto-encoder, deep neural networks, Sentinel-2, Unet

I. Introduction

The temporal frequency of medium resolution, optical satellite imagery, such as Landsat 8 and 9 and Sentinel2A& B, has increase significant in the past four years from one observation every 16 days with Landsat 8 to an observation every 2.9 days on average [16], [25]. Such time series of multi-spectral satellite data enables many novel applications such as global agriculture monitoring and land cover change at the appropriate resolution of land management impacts that has not been possible with low resolution (250-500m GSD) time-series such as MODIS [14], [28], [37]. Moreover, the harmonized Landsat Sentinel2 products open up new avenues for real-time monitoring of a wider variety of change phenomenon [5]. There is a wide variety of time series analysis methods for change detection, which focus primarily on the temporal domain while the spatial domain has been largely neglected (for review see [46]). The availability of hyper-temporal medium resolution imagery allows new application in the spatial domain, such as semantic segmentation with UNET [24] to track objects of change through time.

Change detection methods are hampered by significant noise in the time series that remains despite various processing efforts to reduce the noise [46]. This noise is firstly caused by clouds, and cloud shadows that result in data gaps even when they are correctly detected and serious change artifacts are caused when some clouds or their edges remain undetected [26]. Second, atmospheric variability, notably water vapor and aerosols that cause variability in top of atmosphere reflectance, despite best efforts at atmospheric correction [39], [45]. Third, BRDF variation due to variation in sun-sensor geometry and non-Lambertian reflectance properties of the target. Fourth, variations in spectral reflectance due to seasonal vegetation phenology, which become hard to model when clouds result in missing data that makes the time-series irregular and unpredictable. These issues are often addressed by multi-date composites or temporal interpolation to fill in clouds, shadows and other missing data. Despite efforts to reduce the impact of spatio-temporal noise in the satellite time series, it often limits the accuracy of timely land cover change detection over very large areas using either conventional time-series analysis [42], as well as the new generation of machine learning methods [46].

Various cutting-edge techniques have been developed for image decomposition that can be used to remove noise from conventional images or videos, for example, wavelet smoothing techniques [36], [8], [2] and regularization methods [27], but these methods require parameter selection that vary greatly between datasets and cannot optimally represent varying signals over space and time commonly found in earth observation images. On the other hand, a neural network (UNet) can automatically learn optimal local representation of signal and noise [24], [32], but incurs high computational cost and requires prohibitively large sets of domain-specific training data. The approach proposed in this paper combines the strengths of convolutional neural networks and conventional smoothing techniques. For this purpose we introduce a novel non-subsampled high order Riesz-Quincunx wavelet with variational auto-encoder UNet (RQUNet-VAE) as a hybrid combination of deterministic wavelet expansion (two-dimensional Riesz transformation [36], Quincunx wavelet) and a variational version of a convolutional neural network [22]. The proposed

Manuscript received ??, 2022; revised ??.

RQUNet-VAE approach uses framelet decomposition [43] to map an image into a sparse feature space and leverages UNet [24] to enable learning of the frames of the feature space. The rationale is that such a hybrid method should provide better feature representation and artifact reduction compared to conventional approaches. Therefore, while previous studies used UNet-VAE for image segmentation [15], we use it to mimic the properties of latent factor model in our RQUNet-VAE decomposition.

To implement this concept, we first introduce a non-subsampled version of high-order Riesz wavelet expansion [36], [34], [35] with Quincunx sampling [8]. Quincunx sampling is used to reduce redundancy of an expansion while high order Riesz transform is used to increase the directional property of wavelet expansion. The hybrid model reduces computational cost with deterministic bases as predefined parameters instead of letting all bases be learnable parameters.

Next, this Riesz-Quincunx wavelet is integrated into the skip-connections of the deep neural network UNet-VAE [13], [15], [7] for learning new bases from the training dataset. Our rationale of using wavelet expansion is that signals extracted from the UNet-VAE encoder, at the skip-connection level, contains both the main signal and details (as well as noise) of an input image which are separated into scaling and wavelet coefficients. Truncation of these coefficients eliminates small wavelet coefficients which contain noise and detailed texture. By decoding the remaining coefficients back into image space, we obtain a denoised version of the original image.

The theoretical framework of RQUNet-VAE is based on Hankel matrix algebra [20], framelet decomposition [43], [44] and proximal operators [21]. Framelet decomposition uses isotropic family-matrix convolution to combine all channels of a multi-band image in a convolutional operation with learnable frames. Furthermore, there is a connection between framelet decomposition and sampling with a finite rate of innovation [40] via Hankel matrix theory and annihilating filter. We prove that RQUNet-VAE also relates to latent factor model whose loss function is defined by Kullback-Leibler divergence [38].

Finally we demonstrate how to apply our proposed RQUNet-VAE to satellite image time series, that is, sequences of images of the same area. Reducing noise satellite image time series is challenging due to the severe background noise resulting from spectral variability caused by changing environmental conditions due to atmospheric and seasonal variability, remaining small clouds, as well as variable sun-sensor geometry [25], [45]. The level of noise is very different from that of conventional videos which contains slow motion changes between subsequent images that can be removed with time delay embedding, for example. In contrast, the satellite time series are constituted by discrete frames that are independently capture several days apart, causing large variability in reflectance properties of the background and objects.

To test the effectiveness of our proposed concept, the RQUNet-VAE was applied to image decomposition, image denoising and segmentation of satellite images and their time-series. Our experimental results show that our hybrid method provides better feature representation and artifact reduction than traditional approaches. The objectives of this paper are:

1) introduce RQUNet-VAE as a generalized wavelet expansion approach;
2) extend RQUnet-VAE to a diffusion process [23] and enable spectral decomposition [9];
3) apply it to image denoising and segmentation in noisy environment for multi-band satellite images and their time-series.

Organization of the paper is: Section II provides the RQUNet-VAE expansion, including mathematical properties and image or time-series decomposition; Section III gives numerical examples and comparisons of image denoising and segmentation in noisy environments for multi-band satellites images. Section IV gives the conclusions, Mathematical background, proofs, and additional experiments are are provided in our supplemental material (SM).

II. Riesz-Quincunx-UNet variational auto-encoder (RQUNet-VAE)

A. Notations and Definitions:

The following list provides an overview of notations used throughout this work.

- continuous coordinate in the spatial domain: $x = (x_1\,, x_2) \in \mathbb{R}^2$,
- discrete coordinate in the spatial domain: $k = (k_1\,, k_2)\,, m = (m_1\,, m_2) \in \mathbb{Z}^2$,
- Fourier coordinate: $\omega = (\omega_1\,, \omega_2) \in [-\pi\,, \pi]^2$,
- Complex number: $j = \sqrt{-1}$,
- Image domain: $\Omega = \{1\,, \ldots\,, n_1\} \times \{1\,, \ldots\,, n_2\}$, $|\Omega| = n_1 \times n_2$,
- Images and stacks of images:
  - a gray-scale image $\underline{f} = \begin{pmatrix} f_1 & \ldots & f_{n_2} \end{pmatrix} \in \mathbb{R}^{|\Omega|}\,, f_i \in \mathbb{R}^{n_1}$,
  - a multi-channel image $\underline{\underline{f}} = \left\{\underline{f_1}\,, \ldots\,, \underline{f_P}\right\} \in \mathbb{R}^{|\Omega| \times P}\,, \underline{f_i} \in \mathbb{R}^{|\Omega|}$,
  - a set of observed images $\mathfrak{F} := \underline{\underline{\underline{f}}} = \left\{\underline{\underline{f_i}}\right\}_{i=1}^{T} \in \mathbb{R}^{|\Omega| \times P \times T}$,
- Distribution and Lebesgue density: $\underline{f} \overset{\text{i.i.d.}}{\sim} \mathbb{P}\left(\underline{f}\right) := \mathbb{P}\left(\mathrm{d}\underline{f}\right) = \mathtt{p}\left(\underline{f}\right) \mathrm{d}\underline{f}$,
- Continuous Fourier transform of a continuous function $a \in L_2(\mathbb{R}^2)$ is:

$$a(x) \xleftrightarrow{\mathcal{F}} \widehat{a}(\omega) = \int_{\mathbb{R}^2} a(x) e^{-j\langle x\,, \omega\rangle_{\ell_2}} \mathrm{d}x\,,$$

and its discrete version is computed via Poisson summation formulae:

$$a[k] := a(x) \mid_{x=k \in \mathbb{Z}^2}$$

$$\xleftrightarrow{\mathcal{F}} \widehat{A}\left(e^{j\omega}\right) = \sum_{k \in \mathbb{Z}^2} a[k] e^{-j\langle k\,, \omega\rangle_{\ell_2}} = \sum_{k \in \mathbb{Z}^2} \widehat{a}(2\pi k + \omega)\,.$$

Additional definitions can be found in the supplemental material.

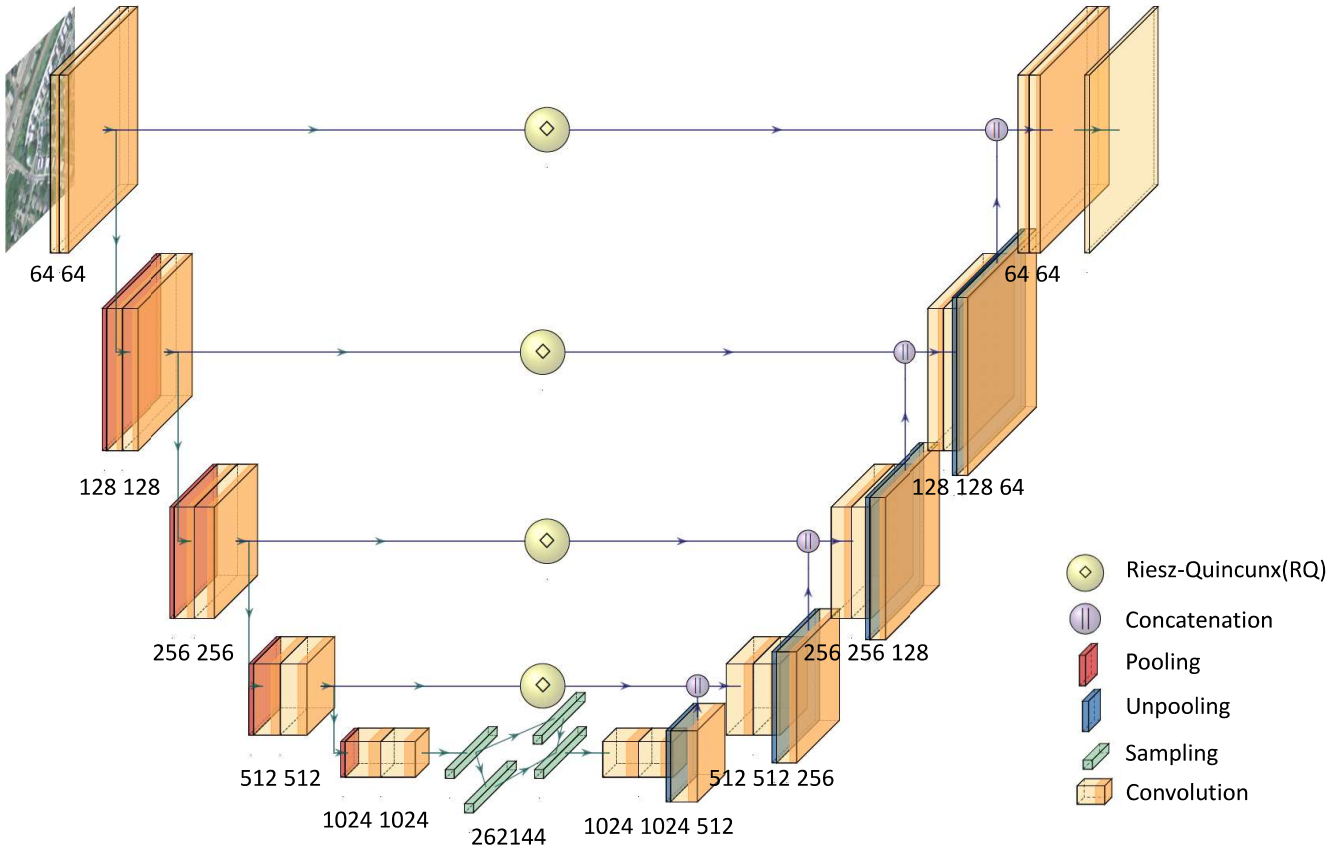


Fig. 1. RQUNet-VAE architecture.

B. RQUNet-VAE Architecture Overview

The network architecture of our proposed RQUNet-VAE is illustrated in Figure 1. The network utilizes a UNet [24] as our primary backbone architecture with the modified skip-connection signals, between the encoder and decoder paths, and bottom layer. The network introduces the Variational Autoencoder (VAE) in the bottom layer to learn the signal distribution in the latent space during training on large datasets. The VAE layer combined with the UNet backbone allows the model to learn and retain much of the input information, similar to generative models, to perform image reconstruction. The reconstructed image then can go through a convolution layer, or a classifier, to produce the final classification or segmentation output. The primary objective of our proposed network is to remove certain levels of noise in the input images in order to produce improved, quality reconstructed images using the Riesz-Quincunx (RQ) scheme in the skip-connections, before performing image classification. Since the RQ scheme is computational heavy, we only applied the RQ computation on a pre-trained network to produce predictions on input images.

C. N-th order Riesz Quincunx non-subsampled wavelet

Before proposing our RQUNet-VAE expansion, we firstly introduce framelet decomposition [43] in the following proposition.

Proposition II.1. *Given an image* $\underline{f} \in \mathbb{R}^{n_1 \times n_2}$*, all wavelet filter banks* $\underline{\Phi}, \underline{\tilde{\Phi}} \in \mathbb{R}^{n_2 d_1 \times d_2}$ *(see definitions of wavelet filter banks in Section 4 of the supplemental material) and local basis* $\Xi, \tilde{\Xi} \in \mathbb{R}^{n_1 \times d}$ *(see Equation 5 in the supplemental material for details) satisfy the unity conditions*

$$\tilde{\Xi}\Xi^T = \sum_{i=1}^{d} \tilde{\xi}_i \xi_i^T = Id_{n_1 \times n_1}\,, \;\; \underline{\Phi}\,\underline{\tilde{\Phi}}^T = \sum_{i=1}^{d_2} \tilde{\phi}_i \phi_i^T = Id_{n_2 d_1 \times n_2 d_1} \tag{1}$$

then, a framelet decomposition is:

$$\begin{aligned} \underline{f} &= \mathscr{H}^{\dagger}_{d_1|n_2}\left(\underline{\tilde{\Xi}} \underline{c_f} \underline{\tilde{\Phi}}^T\right) \\ &= \frac{1}{d_1}\sum_{s=1}^{d_2}\sum_{l=1}^{d}\sum_{i=1}^{n_2} \begin{pmatrix} \left\langle f_k\,, \mathfrak{C}_{\phi_s^i}(\xi_l)\right\rangle_{\ell_2} \mathfrak{C}_{\tilde{\phi}_s^1}\left(\tilde{\xi}_l\right) \\ \cdots \\ \left\langle f_k\,, \mathfrak{C}_{\phi_s^i}(\xi_l)\right\rangle_{\ell_2} \mathfrak{C}_{\tilde{\phi}_s^{n_2}}\left(\tilde{\xi}_l\right) \end{pmatrix} \\ &= \frac{1}{d_1}\sum_{s=1}^{d_2} \begin{pmatrix} \mathfrak{C}_{\tilde{\phi}_s^1}\left(\tilde{\Xi} c_{f,s}\right) & \cdots & \mathfrak{C}_{\tilde{\phi}_s^{n_2}}\left(\tilde{\Xi} c_{f,s}\right) \end{pmatrix} \end{aligned} \tag{2}$$

where $\mathscr{H}^{\dagger}_{d_1|n_2}$ *is the extended Hankel matrix of image* $\underline{f}$ *as described in the SM. Framelet coefficients are:*

$$\underline{c_f} := \begin{pmatrix} c_{f,1} & \cdots & c_{f,d_2} \end{pmatrix} = \Xi^T \mathscr{H}_{d_1|n_2}\left(\underline{f}\right)\underline{\Phi}\,. \tag{3}$$

Proof. For self-containment, we provide a proof of Proposition II.1 in Section 5.1. in the SM. □

Inspired by Proposition II.1, we introduce the proposed RQUNet-VAE expansion in the following.

a) Riesz-Quincunx wavelet expansion:: To have wavelet expansion, firstly we provide a definition of frame by isotropic polyharmonic Bspline and N-th order Riesz transform, see [36]. In particular, an isotropic polyharmonic B-splines basis is defined in the Fourier domain as: $x \in \mathbb{R}^2$,

$$\beta_\gamma(x) \xleftrightarrow{\mathcal{F}} \widehat{\beta}_\gamma(\omega) = \frac{\widehat{V}^{iso}(e^{j\omega})^{\frac{\gamma}{2}}}{\|\omega\|_{\ell_2}^{\gamma}} \tag{4}$$

with a 2D localization operator:

$$\widehat{V}^{\text{iso}}(e^{j\omega}) = \frac{10}{3}$$

$$- \frac{1}{3}\left[4\cos\omega_1 + 4\cos\omega_2 + \cos(\omega_1+\omega_2) + \cos(\omega_1-\omega_2)\right].$$

Its dual function is defined via an auto-correlation function: $k \in \mathbb{Z}^2$,

$$\tilde{\beta}_\gamma(x) \xleftrightarrow{\mathcal{F}} \widehat{\tilde{\beta}}_\gamma(\omega) = \frac{\widehat{\beta}_\gamma(\omega)}{\widehat{A}(e^{j\omega})},$$

$$a(k) \xleftrightarrow{\mathcal{F}} \widehat{A}(e^{j\omega}) = \sum_{m\in\mathbb{Z}^2} \widehat{\beta}_{2\gamma}(2\pi m + \omega). \tag{5}$$

An impulse response of the L-th order Riesz transform is:

$$\mathscr{R}^l\{\delta\}(x) \xleftrightarrow{\mathcal{F}} \widehat{\mathscr{R}^l}(\omega) = (-j)^L \sqrt{\frac{L!}{n!(L-l)!}} \frac{\omega_1^l \omega_2^{L-l}}{\left(\omega_1^2+\omega_2^2\right)^{\frac{L}{2}}},$$

for $l = 0, \dots, L$ and where $\delta(\cdot)$ is the Dirac delta function. Secondly, wavelet expansion is defined as follows: Given a 2D dyadic sampling matrix $\underline{\mathrm{D}} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, we have $\widehat{\rho}_0 = [0,0]^{\mathrm{T}}, \widehat{\rho}_1 = [1,0]^{\mathrm{T}}$ and

$$\underline{\mathrm{D}}^i = \begin{cases} 2^{\frac{i}{2}}\mathrm{Id}_2, & i \text{ is even} \\ 2^{\frac{i-1}{2}}\underline{\mathrm{D}}, & i \text{ is odd} \end{cases}, \quad \left|\det \underline{\mathrm{D}}^i\right| = 2^i.$$

Non-subsampled scaling and wavelet spaces $\mathcal{V}_i^{\text{ns}} = \text{span}_{m\in\mathbb{Z}^2}\left\{\varphi_i(\cdot - m)\right\}$ and $\mathcal{W}_{il}^{\text{ns}} = \text{span}_{m\in\mathbb{Z}^2}\left\{\psi_{il}(\cdot - m)\right\}$ (for $i = 0, \dots, I; l = 0, \dots, L$) satisfy a multiscale decomposition $\mathcal{V}_{i-1}^{\text{ns}} = \mathcal{V}_i^{\text{ns}} \oplus \sum_{l=0}^{L} \mathcal{W}_{il}^{\text{ns}}, i = 1, \dots, I$ where primal/dual scaling and wavelet functions are:

$$\varphi_i(x) = 2^{-\frac{i}{2}}\beta_\gamma\left(\underline{\mathrm{D}}^{-i}x\right), \ \tilde{\varphi}_i(x) = 2^{-\frac{i}{2}}\tilde{\beta}_\gamma\left(\underline{\mathrm{D}}^{-i}x\right),$$

$$\psi_{il}(x) = 2^{-\frac{i}{2}}\mathscr{R}^l\{\psi\}\left(\underline{\mathrm{D}}^{-i}x\right), \ \tilde{\psi}_{il}(x) = 2^{-\frac{i}{2}}\mathscr{R}^l\{\tilde{\psi}\}\left(\underline{\mathrm{D}}^{-i}x\right).$$

Their Fourier transforms are

$$\widehat{\varphi}_i(\omega) = 2^{\frac{i}{2}}\widehat{\beta_\gamma}\left(\underline{\mathrm{D}}^{i\mathrm{T}}\omega\right), \ \widehat{\tilde{\varphi}}_i(\omega) = 2^{\frac{i}{2}}\widehat{\tilde{\beta}_\gamma}\left(\underline{\mathrm{D}}^{i\mathrm{T}}\omega\right),$$

$$\widehat{\psi_{il}}(\omega) = 2^{\frac{i}{2}}\widehat{\mathscr{R}^l}\left(\underline{\mathrm{D}}^{i\mathrm{T}}\omega\right)\widehat{\psi}\left(\underline{\mathrm{D}}^{i\mathrm{T}}\omega\right),$$

$$\widehat{\tilde{\psi}_{il}}(\omega) = 2^{\frac{i}{2}}\widehat{\mathscr{R}^l}\left(\underline{\mathrm{D}}^{i\mathrm{T}}\omega\right)\widehat{\tilde{\psi}}\left(\underline{\mathrm{D}}^{i\mathrm{T}}\omega\right).$$

Due to a discrete Fourier transform of continuous functions, by Poisson summation we have the following proposition:

Proposition II.2. *The scaling and wavelet functions satisfy the unity condition in the Fourier domain:*

$$\widehat{\tilde{\varphi}}_I^*(\omega)\widehat{\varphi}_I(\omega) + \sum_{i=0}^{I}\sum_{l=0}^{L}\widehat{\tilde{\psi}}_{il}^*(\omega)\widehat{\psi}_{il}(\omega) + \widehat{e}(\omega) = 1 \tag{6}$$

up to a discretization error:

$$\begin{aligned}\widehat{e}(\omega) = {} & \widehat{\tilde{\varphi}}_I^*(\omega)\sum_{k\in\mathbb{Z}^2\setminus\{0\}}\widehat{\varphi}_I(2\pi k+\omega) \\ & + \sum_{m\in\mathbb{Z}^2\setminus\{0\}}\widehat{\tilde{\varphi}}_I^*(2\pi m+\omega)\left(\widehat{\varphi}_I(\omega) + \sum_{k\in\mathbb{Z}^2\setminus\{0\}}\widehat{\varphi}_I(2\pi k+\omega)\right) \\ & + \sum_{i=0}^{I}\sum_{l=0}^{L}\left[\widehat{\tilde{\psi}}_{il}^*(\omega)\sum_{k\in\mathbb{Z}^2\setminus\{0\}}\widehat{\psi}_{il}(2\pi k+\omega)\right. \\ & \left. + \sum_{m\in\mathbb{Z}^2\setminus\{0\}}\widehat{\tilde{\psi}}_{il}^*(2\pi m+\omega)\left(\widehat{\psi}_{il}(\omega) + \sum_{k\in\mathbb{Z}^2\setminus\{0\}}\widehat{\psi}_{il}(2\pi k+\omega)\right)\right].\end{aligned} \tag{7}$$

Proof. We provide a proof of Proposition II.2 in Section 5.2 in SM. □

To compensate the error in Equation 7 in the unity condition (Equation 6), wavelet function at scale $i = 0$ is defined as:

$$\widehat{\psi}_0(\omega) = \frac{1}{\widehat{\tilde{\psi}}_0^*(\omega)}\left(1 - \widehat{\tilde{\varphi}}_I^*(\omega)\widehat{\varphi}_I(\omega) - \sum_{i=1}^{I}\widehat{\tilde{\psi}}_i^*(\omega)\widehat{\psi}_i(\omega)\right);$$

then, we have a wavelet expansion for an image $f \in \ell_2(\mathbb{Z}^2)$:

$$\begin{aligned}f[k] = {} & \sum_{m\in\mathbb{Z}^2}\left\langle f, \tilde{\varphi}_I(\cdot - m)\right\rangle_{\ell_2}\varphi_I(k-m) \\ & + \sum_{i=0}^{I}\sum_{l=0}^{L}\sum_{m\in\mathbb{Z}^2}\left\langle f, \tilde{\psi}_{il}(\cdot - m)\right\rangle_{\ell_2}\psi_{il}(k-m).\end{aligned} \tag{8}$$

Following [34], primal and dual wavelet functions are defined as:

$$\widehat{\psi}(\omega) = 2^{-\frac{1}{2}}\widehat{G}(e^{j\underline{\mathrm{D}}^{-\mathrm{T}}\omega})\widehat{\beta}_\gamma(\underline{\mathrm{D}}^{-\mathrm{T}}\omega),$$

$$\widehat{\tilde{\psi}}(\omega) = 2^{-\frac{1}{2}}\widehat{\tilde{G}}\left(e^{j\underline{\mathrm{D}}^{-\mathrm{T}}\omega}\right)\widehat{\tilde{\beta}}_\gamma(\underline{\mathrm{D}}^{-\mathrm{T}}\omega), \tag{9}$$

where the refinement and highpass filters are

$$h[k] \overset{\mathcal{F}}{\longleftrightarrow} \widehat{H}(e^{j\omega}) = 2^{\frac{1}{2}} \frac{\widehat{\beta}_\gamma(\underline{\mathrm{D}}^{\mathrm{T}}\omega)}{\widehat{\beta}_\gamma(\omega)}\,,$$

$$\tilde{h}[k] \overset{\mathcal{F}}{\longleftrightarrow} \widehat{\tilde{H}}(e^{j\omega}) = \frac{\widehat{A}(e^{j\omega})}{\widehat{A}(e^{j\underline{\mathrm{D}}^{\mathrm{T}}\omega})} \widehat{L}(e^{j\omega})\,,$$

$$g[k] \overset{\mathcal{F}}{\longleftrightarrow} \widehat{G}(e^{j\omega}) = -e^{-j\omega_1} \widehat{H}(e^{-j(\omega+\pi)}) \widehat{A}(e^{j(\omega+\pi)})\,,$$

$$\tilde{g}[k] \overset{\mathcal{F}}{\longleftrightarrow} \widehat{\tilde{G}}(e^{j\omega}) = -e^{-j\omega_1} \frac{\widehat{H}(e^{-j(\omega+\pi)})}{\widehat{A}(e^{j\underline{\mathrm{D}}^{\mathrm{T}}\omega})}\,,$$

and a scaled auto-correlation function is:

$$\widehat{A}(e^{j\mathrm{D}^{\mathrm{T}}\omega}) = \frac{1}{2}\left|\widehat{H}(e^{j\omega})\right|^2 \widehat{A}(e^{j\omega}) + \frac{1}{2}\left|\widehat{H}(e^{j(\omega+\pi)})\right|^2 \widehat{A}(e^{j(\omega+\pi)})\,. \tag{10}$$

b) Non-subsampled Riesz-Quincunx wavelet smoothing and Hankel matrix: Denote $\underline{\mathfrak{C}}_{\underline{\phi}} := \underline{\mathfrak{C}}^*_{\underline{\varphi}_I} \underline{\mathfrak{C}}_{\underline{\tilde{\varphi}}_I}$ where a matrix kernel $\underline{\phi} \in \mathbb{R}^{n_1 \times n_2}$ and its matrix form $\underline{\Phi} := \underline{\tilde{\Phi}}_I \underline{\check{\Phi}}_I^{\mathrm{T}}$ are defined in Equation 3 in the SM where $\underline{\tilde{\Phi}}_I$ and $\underline{\Phi}_I$ are matrix forms of $\underline{\tilde{\varphi}_I}$ and $\underline{\varphi_I}$, respectively. Similarly, we denote wavelet kernel tensors $\underline{\underline{\psi}} = \left\{\psi_p\right\}_{p=1}^P := \left\{\psi_{il}\right\}_{i=0,\dots,I}^{l=0,\dots,L}$ and $\underline{\underline{\tilde{\psi}}} = \left\{\tilde{\psi}_p\right\}_{p=1}^P := \left\{\tilde{\psi}_{il}\right\}_{i=0,\dots,I}^{l=0,\dots,L}$ with $\underline{\psi_p}, \underline{\tilde{\psi}_p} \in \mathbb{R}^{|\Omega|}$ $(p = 1,\dots,P)$ and their matrix form are $\underline{\check{\Psi}} = \begin{pmatrix} \underline{\check{\Psi}_1} & \dots & \underline{\check{\Psi}_P} \end{pmatrix}$ and $\underline{\tilde{\Psi}} = \begin{pmatrix} \underline{\tilde{\Psi}_1} & \dots & \underline{\tilde{\Psi}_P} \end{pmatrix}$ where block element matrices $\underline{\Psi_p}, \underline{\tilde{\Psi}_p} \in \mathbb{R}^{n_2 n_1 \times n_2}$ are also are defined in Equation 3 in the supplemental material. For an image $\underline{f} \in \mathbb{R}^{n_1 \times n_2}$, from proposition II.1 we have the following proposition:

Proposition II.3. *A non-subsampled Riesz Quincunx wavelet has a form of framelet decomposition (2):*

$$\begin{aligned} \underline{f} &= \underline{\mathfrak{C}}_{\underline{\phi}}\left(\underline{f}\right) + \underline{\underline{\mathfrak{C}}}^*_{\underline{\underline{\psi}}} \circ \mathit{prox}_{\mu\mathscr{P}} \circ \underline{\underline{\mathfrak{C}}}_{\underline{\underline{\tilde{\psi}}}}\left(\underline{f}\right) \\ &= n_1 \mathscr{H}^{\dagger}_{n_1|n_2}\left(\mathscr{H}_{n_1|n_2}\left(\underline{f}\right)\left(\underline{\Phi} + \underline{\tilde{\Psi}}\,\underline{\check{\Psi}}^T\right)\right)\,. \end{aligned} \tag{11}$$

Scaling and wavelet filter bank matrices satisfy the unity condition:

$$\underline{\Phi} + \underline{\tilde{\Psi}}\,\underline{\check{\Psi}}^T = \frac{1}{n_1} Id_{n_1 n_2 \times n_1 n_2}\,. \tag{12}$$

Proof. We provide a proof of Proposition II.3 in Section 5.3 in SM. □

A smoothing version of a framelet decomposition (Equation 11) is defined with proximity operators $\mathrm{prox}_{\mu\mathscr{P}}$ (for $\mu > 0$) (as defined in Equation 13 in the supplemental material):

$$\underline{\tilde{f}} = n_1 \mathscr{H}^{\dagger}_{n_1|n_2}\left(\mathscr{H}_{n_1|n_2}\left(\underline{f}\right)\underline{\Phi} + \mathrm{prox}_{\mu\mathscr{P}}\left\{\mathscr{H}_{n_1|n_2}\left(\underline{f}\right)\underline{\tilde{\Psi}}\right\}\underline{\check{\Psi}}^{\mathrm{T}}\right)\,.$$

Its iterative scheme, called generalized intersection algorithm with fixpoints is described in [23].

D. RQUNet-VAE expansion

Given a multi-channel image $\underline{f} = \left\{\underline{f_1}, \dots, \underline{f_P}\right\} \in \mathbb{R}^{|\Omega| \times P}$, we introduce our RQUNet-VAE expansion via mappings for skip-connecting signal, latent variables, and a reconstructed signal. Then, we propose RQUNet-VAE expansion and its functional space for regularization.

a) Mappings:: We introduce filter banks in an encoder $\underline{\underline{\theta^{1(i)}}} \in \mathbb{R}^{d_1 \times d_2 \times P \times 2^i L}$ and $\underline{\underline{\underline{\theta^{2(i)}}}} \in \mathbb{R}^{d_1 \times n_2 \times P \times 2^i L}$ whose matrix forms are $\underline{\Theta^{1(i)}} \in \mathbb{R}^{d_1 d_2 P \times d_2 \times 2^i L}, \underline{\Theta^{1(i)}} \in \mathbb{R}^{d_1 n_2 P \times n_2 \times 2^i L}$ (as defined in Equation 6 in the supplemental material). Similar for filter banks $\left\{\left(\underline{\underline{\tilde{\theta}^{1(i)}}}, \underline{\underline{\underline{\tilde{\theta}^{2(i)}}}}\right)\right\}_{i-0}^{I-1}$ in a decoder. For batch-normalization and dropout layers, we refer the readers to [11], [29].

a.1. Skip-connecting signal (encoder): Given a local basis $\underline{\Xi^{(i)}}$ and an analysis operator at scale $i = 1,\dots,I$:

$$\begin{aligned} \mathscr{T}^{(i)} &:= \mathscr{R}_p \circ \mathscr{B} \circ \mathrm{prox}_{\mathrm{ReLU}} \circ \underline{\underline{\mathfrak{C}}}^{\mathrm{iso}}_{\underline{\underline{\underline{\theta^{2(i)}}}}} \circ \mathrm{prox}_{\mathrm{ReLU}} \circ \underline{\underline{\mathfrak{C}}}^{\mathrm{iso}}_{\underline{\underline{\theta^{1(i)}}}} \\ &: \mathbb{R}^{n_1 \times n_2 \times P} \to \mathbb{R}^{2^{-i} n_1 \times 2^{-i} n_2 \times 2^i L}\,, \end{aligned} \tag{13}$$

we define an iterated mapping:

$$\begin{aligned} \mathscr{C}^{(i)} &= \mathscr{T}^{(i)} \circ \underline{\Xi^{(i-1),\mathrm{T}}} \mathscr{T}^{(i-1)} \circ \dots \circ \underline{\Xi^{(1),\mathrm{T}}} \mathscr{T}^{(1)} \\ &: \mathbb{R}^{n_1 \times n_2 \times P} \to \mathbb{R}^{2^{-(i-1)} n_1 \times 2^{-(i-1)} n_2 \times 2^{(i-1)} L}\,. \end{aligned}$$

Then, we have the following proposition describing a mapping for the skip-connecting signal:

Proposition II.4. *A mapping for skip-connecting signal in RQUnet-VAE is:*

$$\mathscr{C}_{\gamma_c} : \mathbb{R}^{n_1 \times n_2 \times P} \to \left\{\mathbb{R}^{2^{-(i-1)} n_1 \times 2^{-(i-1)} n_2 \times 2^{(i-1)} L}\right\}_{i=1}^{I}\,;$$

$$\underline{\underline{c}} = \mathscr{C}_{\gamma_c}\left(\underline{f}\right) = \left\{\underline{\underline{c^{(i)}}} = \mathscr{C}^{(i)}\left(\underline{f}\right)\right\}_{i=0}^{I-1} = \underline{\mathfrak{C}}_{\underline{\phi}}\left(\underline{\underline{c}}\right) + \underline{\underline{\mathfrak{C}}}^*_{\underline{\underline{\psi}}} \underline{\underline{\mathfrak{C}}}_{\underline{\underline{\tilde{\psi}}}}\left(\underline{\underline{c}}\right) \tag{14}$$

which is equivalent to:

$$\underline{\underline{c^{(i)}}} = \left\{\underline{c_l^{(i)}}\right\}_{l=0}^{2^{i-1}L-1}\,,\ \underline{c_l^{(i)}} = \underline{\mathfrak{C}}_{\underline{\phi}}\left(\underline{c_l^{(i)}}\right) + \underline{\underline{\mathfrak{C}}}^*_{\underline{\underline{\psi}}} \underline{\underline{\mathfrak{C}}}_{\underline{\underline{\tilde{\psi}}}}\left(\underline{c_l^{(i)}}\right)$$

for $i = 0,\dots,I-1$ where unknown filter banks are $\gamma_c := \left\{\left(\underline{\underline{\theta^{1(i)}}}, \underline{\underline{\underline{\theta^{2(i)}}}}\right)\right\}_{i=1}^{I}$.

Proof. We provide a proof of Proposition II.4 in Section 5.4 in SM. □

a.2. Variational term (encoder): We firstly define linear mappings as a perceptron network for latent variable:

$$\begin{aligned} \mathscr{F}^{\mu}(\cdot) &= \underline{W^{\mu}}\mathrm{vec}(\cdot) + b^{\mu}\,, \\ \mathscr{F}^{\sigma}(\cdot) &= \underline{W^{\sigma}}\mathrm{vec}(\cdot) + b^{\sigma} \ :\ \mathbb{R}^{2^{-i-1} n_1 \times 2^{-i-1} n_2 \times 2^i L} \to \mathbb{R}^d\,; \end{aligned} \tag{15}$$

where $\underline{W^{\mu}}, \underline{W^{\sigma}} \in \mathbb{R}^{d \times 2^{-(I+1)} n_1 n_2 L}$ and $b^{\mu}, b^{\sigma} \in \mathbb{R}^{d}$ and d is latent dimension and $\mathrm{vec}(\cdot)$ is a vectorize operation. Denote Hadamard product $\odot$ and model's parameters as:

$$\gamma_c \cup \gamma_s \cup \gamma_m := \left\{ \underline{W^{\mu}}, b^{\mu}, \underline{W^{\sigma}}, b^{\sigma}, \left\{ \left(\underline{\underline{\theta^{1(i)}}}, \underline{\underline{\theta^{2(i)}}} \right) \right\}_{i=1}^{I} \right\};$$

then, we have the following proposition for a variational term:

Proposition II.5. *A latent variable in RQUnet-VAE is sampled from a distribution:*

$$z = \mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right) + \mathscr{S}_{\gamma_s}^{\frac{1}{2}}\left(\underline{\underline{f}}\right) \odot \epsilon, \ \epsilon \overset{i.i.d.}{\sim} \mathcal{N}_d\left(\mathbf{0}_d, Id_d\right) \quad (16)$$

where maps for the mean and variance for the latent variable are:

$$\mathscr{M}_{\gamma_m} := \mathscr{F}^{\mu} \circ \mathscr{R}_p \circ \underline{\Xi^{(I),T}} \mathscr{T}^{(I)} \circ \ldots \circ \underline{\Xi^{(1),T}} \mathscr{T}^{(1)}, \quad (17)$$

$$\mathscr{S}_{\gamma_s}^{\frac{1}{2}} := \mathscr{F}^{\sigma} \circ \mathscr{R}_p \circ \underline{\Xi^{(I),T}} \mathscr{T}^{(I)} \circ \ldots \circ \underline{\Xi^{(1),T}} \mathscr{T}^{(1)}. \quad (18)$$

Proof. We provide a proof of Proposition II.5 in Section 5.5 in SM. □

a.3. Decoder: Firstly, we define a mapping in the decoder:

$$\tilde{\mathscr{T}}^{(i)} = \left(\mathcal{Z}_1^{-1}, \mathrm{prox}_{\mathrm{ReLU}} \circ \underline{\underline{\underline{\underline{\mathfrak{C}}}}}^{\mathrm{iso}}_{\underline{\underline{\underline{\underline{\tilde{\theta}}}}}^{1(i+1)}} \circ \mathrm{prox}_{\mathrm{ReLU}} \circ \underline{\underline{\underline{\underline{\mathfrak{C}}}}}^{\mathrm{iso}}_{\underline{\underline{\underline{\underline{\tilde{\theta}}}}}^{2(i+1)}} \circ \mathscr{P}_{\underline{\Xi}^{(i)}_{\mathrm{aug}}} \right),$$

$$\tilde{\mathscr{T}}^{(0)} = \mathrm{prox}_{\mathrm{ReLU}} \circ \underline{\underline{\underline{\underline{\mathfrak{C}}}}}^{\mathrm{iso}}_{\underline{\underline{\underline{\underline{\tilde{\theta}}}}}^{1(1)}} \circ \mathrm{prox}_{\mathrm{ReLU}} \circ \underline{\underline{\underline{\underline{\mathfrak{C}}}}}^{\mathrm{iso}}_{\underline{\underline{\underline{\underline{\tilde{\theta}}}}}^{2(1)}} \circ \mathscr{P}_{\underline{\Xi}^{(0)}_{\mathrm{aug}}},$$

where $\mathcal{Z}_1^{-1}$ is scale-delayed 1 step back of the 1st argument and a concatenate layer is defined as an operation $\mathscr{P}_{\underline{\Xi}^{(i)}_{\mathrm{aug}}}\left(\underline{\underline{\underline{\underline{c}}}}^{(i)}_{\mathrm{aug}}\right) = \left(\underline{\underline{\underline{\underline{c}}}}^{(i)} \quad \mathscr{B} \circ \underline{\tilde{\Xi}^{(i)}}\, \underline{\underline{\underline{\underline{\widehat{s}}}}}^{(i+1)}\right)$, i.e. input signal at concatenation layer includes bypass signal $c^{(i)}$ and unpooled-batchnormed lowpass signal $\mathscr{B} \circ \underline{\tilde{\Xi}^{(i)}}\, \underline{\underline{\underline{\underline{\widehat{s}}}}}^{(i+1)}$.

Next, we define an iterated mapping for a skip-connecting signal $\underline{\underline{\underline{\underline{c}}}}$ and a lowpass signal $\underline{\underline{\underline{\underline{s^{(I)}}}}}$ at scale I as:

$$\tilde{\mathscr{T}}_I\left(\underline{\underline{\underline{\underline{c}}}}, \underline{\underline{\underline{\underline{s^{(I)}}}}}\right) = \tilde{\mathscr{T}}^{(0)} \circ \tilde{\mathscr{T}}^{(1)} \circ \ldots \circ \tilde{\mathscr{T}}^{(I-1)}\left(\underline{\underline{\underline{\underline{c^{(I-1)}}}}} \quad \underline{\underline{\underline{\underline{s^{(I)}}}}}\right).$$

Note that $\underline{\underline{\underline{\underline{s^{(I)}}}}}$ is reconstructed from a latent variable z. Denote unknown parameters in a decoder as:

$$\alpha := \left\{ \underline{W^s}, b^s, \underline{\underline{\underline{\underline{\theta^0}}}}, \left\{ \left(\underline{\underline{\underline{\underline{\tilde{\theta}^{1(i)}}}}}, \underline{\underline{\underline{\underline{\tilde{\theta}^{2(i)}}}}} \right) \right\}_{i=1}^{I} \right\}.$$

Then, we have the following proposition:

Proposition II.6. *Given a skip-connecting signal $\underline{\underline{\underline{\underline{c}}}}$ and a latent variable y from an encoder, a decoder mapping in RQUNet-VAE is defined as*

$$\mathscr{D}_{\alpha}\left(\underline{\underline{\underline{\underline{c}}}}, z\right) := prox_{ReLU} \circ \underline{\underline{\underline{\underline{\mathfrak{C}}}}}^{iso}_{\underline{\underline{\underline{\underline{\theta^0}}}}} \circ \tilde{\mathscr{T}}_I\left(\underline{\underline{\underline{\underline{c}}}}, uvec\left(\underline{W^s} z + b^s\right)\right). \quad (19)$$

Proof. We provide a proof of Proposition II.6 in Section 5.6 in SM. □

b) RQUNet-VAE expansion: An auto-encoder is recast as a latent factor model (a decoder) whose latent variable is computed from the observed data as an encoder:

$$z = \mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right) + \mathscr{S}_{\gamma_s}^{\frac{1}{2}}\left(\underline{\underline{f}}\right) \odot \epsilon, \ \epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(\mathbf{0}_d, \mathrm{Id}_d), \quad (20)$$

$$\underline{\underline{\underline{\underline{c}}}} = \mathscr{C}_{\gamma_c}\left(\underline{\underline{f}}\right), \quad (21)$$

$$\underline{\underline{\tilde{f}}} = \mathscr{D}_{\alpha}\left(\underline{\underline{\underline{\underline{c}}}}, z\right) + \sigma \underline{\underline{e}}, \ \underline{\underline{e}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{|\Omega| \times P}(0, \mathrm{Id}), \quad (22)$$

$$\underline{\underline{\tilde{f}}} = \underline{\underline{f}} \in \mathfrak{F}, \quad (23)$$

with a known standard deviation $\sigma > 0$ and the data set $\mathfrak{F} = \left\{\underline{\underline{f_i}}\right\}_{i=1}^{T} \subset \mathbb{R}^{|\Omega| \times P}$. Combining Equation 22, Equation 20, Equation 21, and Equation 23, we obtain a variational auto-encoder:

$$\underline{\underline{f}} = \mathscr{D}_{\alpha}\left(\mathscr{C}_{\gamma_c}\left(\underline{\underline{f}}\right), \mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right) + \mathscr{S}_{\gamma_s}^{\frac{1}{2}}\left(\underline{\underline{f}}\right) \odot \epsilon\right) + \sigma \underline{\underline{e}}, \quad (24)$$

with standard normal random variables $\epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(\mathbf{0}_d, \mathrm{Id}_d)$ and $\underline{\underline{e}} = \left[e_{l,c}\right]_{l \in \Omega}^{c=1,\ldots,p}$, $e_{l,c} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.

The auto-encoder system (Equations 22-23) is recast as Bayesian inference:

$$\underline{\underline{\tilde{f}}} \mid z \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{n_1 \times n_2 \times P}\left(\mathscr{D}_{\alpha}\left(\underline{\underline{\underline{\underline{c}}}}, z\right), \sigma^2 \mathrm{Id}\right)$$

$$= \mathbb{H}_{\alpha}\left(\underline{\underline{\tilde{f}}} \mid z\right) = \mathtt{h}_{\alpha}\left(\underline{\underline{\tilde{f}}} \mid z\right) \mathrm{d}\underline{\underline{\tilde{f}}}, \quad (25)$$

$$z \mid \underline{\underline{f}} \overset{\text{i.i.d.}}{\sim} \mathbb{K}_{\alpha}\left(z \mid \underline{\underline{f}}\right) = \mathtt{k}_{\alpha}\left(z \mid \underline{\underline{f}}\right) \mathrm{d}y$$

$$\approx \mathcal{N}_d\left(\mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right), \mathrm{diag}\left\{\mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}\right)\right\}\right). \quad (26)$$

An explanation for the above hierarchical model is: an observed signal $\underline{\underline{\tilde{f}}}$ is assumed to be sampled from a normal distribution whose latent variable z is sampled from an unknown distribution $\mathbb{K}_{\alpha}$ which is approximated by a normal distribution (Equation 26) parameterized by (γ_m, γ_s). The key idea for $z \mid \underline{\underline{f}} \sim \mathbb{K}_{\alpha}$ to depend on a parameter α is because of the auto-encoder (Equations 20-23), i.e. $\underline{\underline{\tilde{f}}} = \underline{\underline{f}}$. To see this, Bayes' rule for a conditional density of random variable $z \mid \underline{\underline{f}}$ is:

$$\mathtt{p}\left(z \mid \underline{\underline{f}}\right) = \frac{\mathtt{p}\left(\underline{\underline{f}} \mid z\right) \mathtt{p}(z)}{\mathtt{p}\left(\underline{\underline{f}}\right)} = \frac{\mathtt{p}\left(\underline{\underline{\tilde{f}}} \mid z\right) \mathtt{p}(z)}{\mathtt{p}\left(\underline{\underline{f}}\right)}$$

$$\overset{(25)}{=} \frac{\mathtt{h}_{\alpha}\left(\underline{\underline{\tilde{f}}} \mid z\right) \mathtt{p}(z)}{\mathtt{p}\left(\underline{\underline{f}}\right)} := \mathtt{k}_{\alpha}\left(z \mid \underline{\underline{f}}\right),$$

which depends on a decoder's parameter α. Since the above density is intractible because of the incomputable integral

in a marginal distribution $\mathtt{p}\left(\underline{\underline{f}}\right) = \int_{\mathbb{R}^d} \mathtt{h}_\alpha\left(\underline{\underline{\tilde{f}}} \mid z\right)\mathtt{p}(z)\mathrm{d}z$, a distribution $\mathbb{K}_\alpha$ is approximated by a normal distribution:

$$\mathtt{k}_\alpha\left(z \mid \underline{\underline{f}}\right) = \frac{\mathtt{h}_\alpha\left(\underline{\underline{f}} \mid z\right)\mathtt{p}(z)}{\mathtt{p}\left(\underline{\underline{f}}\right)}$$

$$\approx \mathcal{N}_d\left(z\,; \mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right), \mathrm{diag}\left\{\mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}\right)\right\}\right). \quad (27)$$

Denote variance vector as

$$\mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}\right) = \mathscr{S}^{\frac{1}{2}}_{\gamma_s}\left(\underline{\underline{f}}\right) \odot \mathscr{S}^{\frac{1}{2}}_{\gamma_s}\left(\underline{\underline{f}}\right)$$

$$= \left(\mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}\right)_1 \quad \dots \quad \mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}\right)_d\right)^{\mathsf{T}} \in \mathbb{R}^d.$$

Choose a standard normal prior distribution $\mathbb{P}(z) = \mathtt{p}(z)\mathrm{d}z = \mathcal{N}_d\left(\mathbf{0}_d, \mathrm{Id}_d\right)$, we have the following proposition for finding model's parameters:

Proposition II.7. *Unknown parameters in RQUnet-VAE are obtained from the following minimization problem:*

$$\left(\gamma_c^\dagger, \gamma_m^\dagger, \gamma_s^\dagger, \alpha^\dagger\right) = \mathrm{argmin}\,\mathscr{L}\left(\gamma_c, \gamma_m, \gamma_s, \alpha\right) \quad (28)$$

where:

$$\mathscr{L}(\cdot) := \frac{1}{2\sigma^2}\sum_{i=1}^{T}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,Id)}\Bigg[$$

$$\left\|\underline{\underline{f_i}} - \mathscr{D}_\alpha\left(\mathscr{C}_{\gamma_c}\left(\underline{\underline{f_i}}\right), \mathscr{M}_{\gamma_m}\left(\underline{\underline{f_i}}\right) + \mathscr{S}^{\frac{1}{2}}_{\gamma_s}\left(\underline{\underline{f_i}}\right)\odot\epsilon\right)\right\|^2_{\ell_2}\Bigg]$$

$$+\sum_{i=1}^{T} KL\left[\mathcal{N}_d\left(\mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right), diag\left\{\mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}\right)\right\}\right) \,||\, \mathcal{N}_d\left(\mathbf{0}_d, Id_d\right)\right]$$

and KL-divergence is defined as:

$$KL\left[\cdot \,||\, \cdot\right]$$

$$= \frac{1}{2}\left(\left\|\mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right)\right\|^2_{\ell_2} - d + \sum_{i=1}^{d}\left(\mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}\right)_i - \log\mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}\right)_i\right)\right). \quad (29)$$

Proof. We provide a proof of Proposition II.7 in Section 5.7 in SM. □

It is clear that the only one random variable in the above minimization is ϵ, so the gradient descent method can be applied for model parameters $(\gamma_c, \gamma_m, \gamma_s, \alpha)$. Since high order Riesz-Quincunx wavelet expansion is an identity operator, for training these model parameters, we remove this layer in the training procedure. But, later we use it in its smoothing version with the trained parameters $\left(\gamma_c^\dagger, \gamma_m^\dagger, \gamma_s^\dagger, \alpha^\dagger\right)$ to truncate small wavelet coefficients of signals in skip-connection. A summary of RQUNet-VAE with a training procedure (without Riesz-Quincunx wavelet) is shown in the SM for encoder and decoder, respectively.

c) Generalized Besov space by RQUNet-VAE and proximal operators: We note that Equation 28 is a non-convex minimization problem with potentially many local minimas. Assume an existence of minima $\left(\alpha^\dagger, \gamma_c^\dagger, \gamma_m^\dagger, \gamma_s^\dagger\right)$ such that a condition of perfect reconstruction occurs, i.e. $\underline{\underline{\tilde{f}}} = \underline{\underline{f}}$; then, the auto-encoder (Equations 22-23) plays as a generalized wavelet expansion with learnable parameters:

$$\lim_{\sigma\to 0}\left\|\underline{\underline{f}} - \mathscr{D}_{\alpha^\dagger}\left(\mathscr{C}_{\gamma_c^\dagger}\left(\underline{\underline{f}}\right), \mathscr{M}_{\gamma_m^\dagger}\left(\underline{\underline{f}}\right) + \mathscr{S}^{\frac{1}{2}}_{\gamma_s^\dagger}\left(\underline{\underline{f}}\right)\odot\epsilon\right) + \sigma\underline{\underline{e}}\right\|^2_{\ell_2}$$

$$= 0,$$

having its deterministic version:

$$\underline{\underline{f}} = \mathscr{D}_{\alpha^\dagger}\left(\mathscr{C}_{\gamma_c^\dagger}\left(\underline{\underline{f}}\right), \mathscr{M}_{\gamma_m^\dagger}\left(\underline{\underline{f}}\right) + \mathscr{S}^{\frac{1}{2}}_{\gamma_s^\dagger}\left(\underline{\underline{f}}\right)\odot\epsilon\right),$$

where encoder and decoder are forward and backward wavelet transform. Then, it induces a generalized Besov space:

$$\mathfrak{B} := \left\{\underline{\underline{f}} \in \mathbb{R}^{|\Omega|\times P} : \left\|\underline{\underline{f}}\right\|_{\mathfrak{B}} := \mathscr{P}\circ\mathscr{C}_{\gamma_c}\left(\underline{\underline{f}}\right) < \infty\right\}, \quad (30)$$

where function $\mathscr{P}(\cdot)$ acts on the skip-connecting signal $\underline{\underline{c}} = \mathscr{C}_{\gamma_c}\left(\underline{\underline{f}}\right)$ for proximal mapping as described in Equation 13 in the supplemental material.

Then, given an image $\underline{\underline{f}} \in \mathbb{R}^{|\Omega|\times P}$ with an expansion in a space $\mathfrak{B}$ (30), we have a regularization in that space $\mathfrak{B}$ with the optimally trained parameters as generalized wavelet smoothing via a proximal operator:

$$\underline{\underline{\tilde{c}}} = \underset{\underline{\underline{w}}}{\mathrm{argmin}}\left\{\mathscr{P}\left(\underline{\underline{w}}\right) + \frac{1}{2\mu}\left\|\mathscr{C}_{\gamma_c}\left(\underline{\underline{f}}\right) - \underline{\underline{w}}\right\|^2_{\mathrm{F}}\right\}$$

$$= \mathrm{prox}_{\mu\mathscr{P}}\circ\mathscr{C}_{\gamma_c}\left(\underline{\underline{f}}\right)$$

$$= \underline{\underline{\mathfrak{C}}}_{\underline{\phi}}\circ\mathscr{C}_{\gamma_c}\left(\underline{\underline{f}}\right) + \underline{\underline{\mathfrak{C}}}^*_{\underline{\psi}}\circ\mathrm{prox}_{\mu\mathscr{P}}\circ\underline{\underline{\mathfrak{C}}}_{\underline{\tilde{\psi}}}\circ\mathscr{C}_{\gamma_c}\left(\underline{\underline{f}}\right), \quad (31)$$

where its element form is:

$$\underline{\underline{\tilde{c}}} = \left\{\underline{\underline{\tilde{c}^{(i)}}}\right\}_{i=0}^{I-1},\ \underline{\underline{\tilde{c}^{(i)}}} = \left\{\underline{\tilde{c}_l^{(i)}}\right\}_{l=0}^{2^{i-1}L-1},$$

$$\underline{\tilde{c}_l^{(i)}} = \underline{\mathfrak{C}}_{\underline{\phi}}\left(\underline{c_l^{(i)}}\right) + \underline{\underline{\mathfrak{C}}}^*_{\underline{\psi}}\circ\mathrm{prox}_{\mu\mathscr{P}}\circ\underline{\underline{\mathfrak{C}}}_{\underline{\tilde{\psi}}}\left(\underline{c_l^{(i)}}\right)$$

for $l = 0, \dots, 2^{i-1}L-1$, $i = 0, \dots, I-1$. Note that $\mathrm{prox}_{\mu\mathscr{P}}(\cdot)$ in Equation 31 acts on wavelet coefficients only. A reason is scaling coefficients contain main energy of a signal whose values are large. These scaling coefficients should be preserved during a shrinking process while wavelet coefficients mainly contain oscillating signals in different scales, including noise which should be removed.

Then, we have a smoothed image as:

$$\underline{\underline{\tilde{f}}} = \mathscr{D}_{\alpha^\dagger}\left(\underline{\underline{\tilde{c}}},\ \mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right) + \mathscr{S}^{\frac{1}{2}}_{\gamma_s}\left(\underline{\underline{f}}\right)\odot\epsilon\right) \in \mathbb{R}^{|\Omega|\times P}.$$

In summary, we have the following proposition:

Proposition II.8. *Given a learned parameter* $\left(\alpha^{\dagger}, \gamma_c^{\dagger}, \gamma_m^{\dagger}, \gamma_s^{\dagger}\right)$ *obtain by training RQUNet-VAE with a dataset, we have RQUNet-VAE smoothing (with a parameter* $\mu > 0$*) for an image* $\underline{\underline{\tilde{f}}}$ *as a solution of a regularization in a generalized Besov space* $\mathfrak{B}$ *(30):*

$$\underline{\underline{\tilde{f}}} = \mathscr{D}_{\alpha^{\dagger}}\left(\underline{\underline{\mathfrak{C}}}_{\underline{\phi}} \circ \mathscr{C}_{\gamma_c^{\dagger}}\left(\underline{\underline{f}}\right) + \underline{\underline{\mathfrak{C}}}^*_{\underline{\underline{\psi}}} \circ prox_{\mu\mathscr{P}} \circ \underline{\underline{\mathfrak{C}}}_{\underline{\tilde{\psi}}} \circ \mathscr{C}_{\gamma_c^{\dagger}}\left(\underline{\underline{f}}\right), \mathscr{M}_{\gamma_m^{\dagger}}\left(\underline{\underline{f}}\right) + \mathscr{S}^{\frac{1}{2}}_{\gamma_s^{\dagger}}\left(\underline{\underline{f}}\right) \odot \epsilon\right). \tag{32}$$

Proof. A proof of Proposition II.8 is directly obtained from the above paragraphs. □

E. RQUNet-VAE iterative shrinkage Lagrangian system

a) Decomposition for multi-band image: RQUNet-VAE iterative shrinkage algorithm for multi-band image $\underline{\underline{f}} \in \mathbb{R}^{|\Omega| \times P}$ is described in Algorithm 1 in the SM. This is equivalent to a nonlinear mapping with unknown model's parameters Γ: $\underline{\underline{u}}^{(0)} = \underline{\underline{f}}, \underline{\underline{\lambda}}^{(0)} = \mathbf{0}$ for $\tau = 1, \ldots, N$ and

$$\left(\underline{\underline{u}}^{(\tau)}, \underline{\underline{\lambda}}^{(\tau+1)}\right) = \mathscr{K}_{\mu}\left(\underline{\underline{f}}, \underline{\underline{u}}^{(\tau-1)}, \underline{\underline{\lambda}}^{(\tau)}; \Gamma\right). \tag{33}$$

b) Diffusion process and spectral decomposition for multiband image: For a multiband image $\underline{\underline{f}} \in \mathbb{R}^{|\Omega| \times P}$, a diffusion process by a Lagrangian system (34) is: $\underline{\underline{u}}^{(0)} = \underline{\underline{f}}, \tau = 1, \ldots, N$,

$$\left(\underline{\underline{u}}^{(\tau)}, \underline{\underline{\lambda}}^{(\tau+1)}\right) = \mathscr{K}_{\mu}\left(\underline{\underline{u}}^{(\tau-1)}, \underline{\underline{u}}^{(\tau-1)}, \underline{\underline{\lambda}}^{(\tau)}; \Gamma\right). \tag{34}$$

Given $\left\{\underline{\underline{u}}^{(\tau)}\right\}_{\tau=1}^{N}$ generated by diffusion process (34) in Algorithm 2, we have a discrete TV-like transform as in [9]:

$$\underline{\underline{\phi}}^{(\tau)} := \frac{\tau}{\beta}\left(\underline{\underline{u}}^{(\tau+1)} - 2\underline{\underline{u}}^{(\tau)} + \underline{\underline{u}}^{(\tau-1)}\right),$$

whose inverse transform is: $\underline{\underline{\tilde{f}}}^{(N)} := (1+N)\,\underline{\underline{u}}^{(N)} - N\underline{\underline{u}}^{(N+1)}$,

$$\underline{\underline{f}} = \underline{\underline{\tilde{f}}}^{(N)} + \beta \sum_{\tau=1}^{N} \underline{\underline{\phi}}^{(\tau)}.$$

Its filtered version

$$\underline{\underline{f_{H_N}}} = H^{(N)}\,\underline{\underline{\tilde{f}}}^{(N)} + \beta \sum_{\tau=1}^{N} H^{(\tau)}\,\underline{\underline{\phi}}^{(\tau)}$$

is defined, e.g. by the ideal spectral-filter $H^{(\tau)}$:

$$H^{(\tau)}_{\text{lowpass}} = \begin{cases} 0, & \tau \in \{0, \ldots, \tau_1\}, \\ 1, & \tau \in \{\tau_1, \ldots, N\} \end{cases},$$

$$H^{(\tau)}_{\text{highpass}} = \begin{cases} 1, & \tau \in \{0, \ldots, t_1\}, \\ 0, & \tau \in \{\tau_1, \ldots, N\} \end{cases},$$

$$H^{(\tau)}_{\text{bandpass}} = \begin{cases} 0, & \tau \in \{0, \ldots, \tau_1\}, \\ 1, & \tau \in \{\tau_1, \ldots, \tau_2\}, \\ 0, & \tau \in \{\tau_2, \ldots, N\} \end{cases}, ,$$

$$H^{(\tau)}_{\text{bandstop}} = \begin{cases} 1, & \tau \in \{0, \ldots, \tau_1\}, \\ 0, & \tau \in \{\tau_1, \ldots, \tau_2\}, \\ 1, & \tau \in \{\tau_2, \ldots, N\}, \end{cases}$$

where the time threshold τ_i are selected by the TV-spectrum $S^{(\tau)} := \left\|\underline{\underline{\phi}}^{(\tau)}\right\|_{\ell_1}$.

c) Decomposition for multiband time series: Next, we show how our RQUNet-VAE can be applied to satellite image time series, that is, a sequence of images of the same area.

Given a multi-band video $\underline{\underline{f}} = \left\{\underline{\underline{f_1}}, \ldots, \underline{\underline{f_T}}\right\} \in \mathbb{R}^{|\Omega| \times P \times T}$, $\underline{\underline{f_t}} = \left\{\underline{f_{t,1}}, \ldots, \underline{f_{t,P}}\right\}$, $\underline{f_{t,p}} \in \mathbb{R}^{|\Omega|}$ and scalling and wavelet bases, e.g. 1D Haar bases:

$$\phi_{I,t,m} = \begin{cases} 1, & t \geq 2^I m \;\&\; t < 2^I(1+m) \\ 0, & \text{else} \end{cases},$$

$$\xi_{i,t,m} = \begin{cases} 1, & t \geq 2^i m \;\&\; t < 2^i\left(\frac{1}{2}+m\right) \\ -1, & t \geq 2^i\left(\frac{1}{2}+m\right) \;\&\; t < 2^i(1+m) \\ 0, & \text{else} \end{cases}, \tag{35}$$

a time-wavelet smoothing expansion for a video $\underline{\underline{f}} \in \mathbb{R}^{|\Omega| \times P \times T}$ by proximal operator $\mu\mathscr{P}$ is defined as: $t = 1, \ldots, T$,

$$\underline{\underline{\tilde{f}_t}} = \frac{1}{2^{\frac{I}{2}}} \sum_{s=0}^{2^{-I}T-1} \frac{1}{2^{\frac{I}{2}}} \sum_{t'=1}^{T} \underline{\underline{f_{t'}}}\,\phi_{I,t',s}\phi_{I,t,s} + \sum_{i=1}^{I} \frac{1}{2^{\frac{i}{2}}} \sum_{s=0}^{2^{-i}T-1} \text{prox}_{\mu\mathscr{P}}\Big\{ \underbrace{\frac{1}{2^{\frac{i}{2}}} \sum_{t'=1}^{T} \underline{\underline{f_{t'}}}\,\xi_{i,t',s}}_{:=\underline{\underline{w_{i,m}}}} \Big\} \xi_{i,t,s},$$

which is equivalent to:

$$\underline{\underline{\tilde{f}}} := \mathscr{G}_{\phi}\left(\underline{\underline{f}}\right) + \mathscr{W}_{\xi}^* \circ \text{prox}_{\mu\mathscr{P}} \circ \mathscr{W}_{\xi}\left(\underline{\underline{f}}\right) = \left\{\underline{\underline{\tilde{f}_1}}, \ldots, \underline{\underline{\tilde{f}_T}}\right\}. \tag{36}$$

Wavelet coefficient in (36) is $\underline{\underline{w_i}} = \left\{\underline{\underline{w_{i,1}}}, \ldots, \underline{\underline{w_{i,T}}}\right\} := \mathscr{W}_{\xi}\left\{\underline{\underline{f}}\right\} \in \mathbb{R}^{|\Omega| \times P \times 2^{-i}T}$, $\underline{\underline{w_{i,m}}} = \left\{\underline{w_{i,m,1}}, \ldots, \underline{w_{i,m,P}}\right\}$

and $w_{i,m,p} \in \mathbb{R}^{|\Omega|}$. Combined with Equation (34), RQUnet-VAE's iterative shrinkage algorithm for multi-band video $\underline{\underline{f}} \in \mathbb{R}^{n_1 \times n_2 \times P \times T}$ is described in Algorithm 2 in SM, called RQUNet-VAE scheme 2.

F. RQUNet-VAE based segmentation

In this section, we propose a mathematical framework for a segmentation problem with our RQUNet-VAE which serves as a smoothing term. Given a dataset of original images (without artificially added noise) with ground-truth masks, we first train the UNet-VAE to obtain the optimal model parameters. Next we predict a segmented image from a noisy image using the RQUNet-VAE with a smoothing term as a truncation scheme of the N-th order Riesz-Quincunx wavelet expansion on the skip-connecting signals. If the training data is $\left(\mathfrak{F}, \mathfrak{F}^{\text{gt}}\right) := \left\{\underline{\underline{f}}_i, \underline{\underline{f}}_i^{\text{gt}}\right\}_{i=1}^{n} \subset \mathbb{R}^{|\Omega| \times P} \times \mathbb{R}^{|\Omega| \times K}$ and K-dimensional hot key tensors of the ground-truth masks, i.e. pixel intensity and its allocation are:

$$\underline{\underline{f}}_i := \left[f_{i,l}\right]_{l \in \Omega}\,,\ f_{i,l} \in \mathbb{R}^P\,,\ \underline{\underline{f}}_i^{\text{gt}} := \left[f_{i,l}^{\text{gt}}\right]_{l \in \Omega}\,,$$

where $\underline{f}_{i,l}^{\text{gt}} = \left[f_{i,l,k}^{\text{gt}}\right]_{k=1}^{K} \in \mathbb{R}^K$ is a one-hot-key vector. Given the loss function: $\underline{\underline{x}}, \underline{\underline{y}} \in \mathbb{R}^{|\Omega| \times K}$,

$$\mathscr{H}\left(\underline{\underline{x}}, \underline{\underline{y}}\right) = \sum_{l \in \Omega} \sum_{k=1}^{K} x_{l,k} \log \frac{\exp\left(y_{l,k}\right)}{\sum_{h=1}^{K} \exp\left(y_{l,h}\right)}\,;$$

similar to proposition II.7, a loss function of our RQUNet-VAE based segmentation problem is as follows:

Proposition II.9. *Unknown parameters in RQUNet-VAE based segmentation are obtained from the following minimization problem:*

$$\left(\underline{\underline{\theta}}^{\dagger}, \gamma_c^{\dagger}, \gamma_m^{\dagger}, \gamma_s^{\dagger}, \alpha^{\dagger}\right) = \text{argmin} \sum_{i=1}^{n} \mathscr{L}\left(\underline{\underline{\theta}}, \gamma_c, \gamma_m, \gamma_s, \alpha\right) \tag{37}$$

where:

$$\mathscr{L}(\cdot) := KL\left[\mathcal{N}_d\left(\mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}_i\right), diag\left\{\mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}_i\right)\right\}\right) || \mathcal{N}_d\left(\mathbf{0}_d, Id_d\right)\right]$$

$$- \frac{1}{2n\sigma^2} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, Id)}\Big[\mathscr{H}\Big(\underline{\underline{f}}_i^{gt}, \underline{\underline{\mathfrak{C}}}_{\underline{\underline{\theta}}}^{iso} \circ \mathscr{D}_{\alpha}\Big(\mathscr{C}_{\gamma_c}\left(\underline{\underline{f}}_i\right), \mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}_i\right)$$

$$+ \mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}_i\right)^{\frac{1}{2}} \odot \epsilon\Big)\Big)\Big]\,,$$

and the K-L divergence is defined in (29).

Proof. We provide a proof of Proposition II.9 in Section 5.8 in SM. □

After training the model parameters by minimizing the loss function (37), the segmentation of a new noisy image is predicted:

$$f^{\text{new}} = f_0^{\text{new}} + \sigma\epsilon\,,$$

with standard Gaussian noise $\epsilon = \left[\epsilon_{l,c}\right]_{l \in \Omega}^{c=1,\ldots,p} \in \mathbb{R}^{|\Omega| \times P}$ and $\epsilon_{l,c} \sim \mathcal{N}(0,1)$. From proposition II.8 by adding N-th order Riesz Quincunx wavelet truncation with proximal operator $\text{prox}_{\mu\mathscr{P}}(\cdot)$ parameterized by a smoothing parameter μ as:

$$u\left(\underline{\underline{f}}^{\text{new}}\right) = \left\{u_1\left(\underline{\underline{f}}^{\text{new}}\right), \ldots, u_K\left(\underline{\underline{f}}^{\text{new}}\right)\right\} \in \mathbb{R}^{|\Omega| \times K}$$

$$= \underline{\underline{\underline{\mathfrak{C}}}}_{\underline{\underline{\theta}}}^{\text{iso}} \circ \mathscr{D}_{\alpha^{\dagger}}\Big(\underline{\underline{\mathfrak{C}}}_{\underline{\phi}} \circ \mathscr{C}_{\gamma_c^{\dagger}}\left(\underline{\underline{f}}^{\text{new}}\right) + \underline{\underline{\mathfrak{C}}}_{\underline{\underline{\psi}}}^{*} \circ \text{prox}_{\mu\mathscr{P}} \circ \underline{\underline{\mathfrak{C}}}_{\underline{\underline{\tilde{\psi}}}} \circ \mathscr{C}_{\gamma_c^{\dagger}}\left(\underline{\underline{f}}^{\text{new}}\right),$$

$$\mathscr{M}_{\gamma_m^{\dagger}}\left(\underline{\underline{f}}_i^{\text{new}}\right) + \mathscr{S}_{\gamma_s^{\dagger}}^{\frac{1}{2}}\left(\underline{\underline{f}}_i^{\text{new}}\right) \odot \epsilon\Big)\,,$$

$$\underline{\underline{y}}^{\text{new}} = \text{argmax}_{k=1,\ldots,K} u_k\left(\underline{\underline{f}}^{\text{new}}\right) \in \mathbb{R}^{|\Omega| \times K}\,,\ u_k\left(\underline{\underline{f}}^{\text{new}}\right) \in \mathbb{R}^{|\Omega|}\,.$$

III. Experimental Results

A. Experimental Setup

This section describes the dataset used for our experiments, introduces competitor solutions, and defines evaluation metrics to measure the ability of an algorithm to reduce the noise of an image.

1) Datasets: We used Sentinel-2 satellite images (S30:MSI harmonized, V1.5) processed as part of the Harmonized Landsat Sentinel-2 (HLS) dataset obtained from USGS DAAC (https://lpdaac.usgs.gov/data/get-started-data/collection-overview/missions/harmonized-landsat-sentinel-2-hls-overview/) [5]. The S30:MSI harmonized surface reflectance product is resampled from the original 10m to 30m resolution and adjusted to Landsat8 spectral response function in order to ultimately create a harmonized time series with a 2-3 day revisit. Radiometric and geometric corrections are applied to convert data to surface reflectance, adjust for BRDF differences, and spectral bandpass differences. The study was focused on the HLS tile 18STH covering a large part of Northern Virginia in the US, for 2016-2020. From the main HLS tile's time series, 510 images were randomly created with a size of 256×256 pixels and including only the three visible bands (4R, 3G, 2B). Quality Assessment (QA) layers were used to exclude images with more than 30 percent cloud shadow, adjacent cloud, cloud and cirrus clouds. The 500 images were used for training of our RQUNet-VAE and competitor approaches, while ten images were used for testing in the denoising and decomposition experiments. The ten test images are visualized in SM.

For the time series decomposition experiments the above Sentinel-2 data were used to create 80, randomly located time series of length 99 images and image size 40×40 pixels. Images were normalized from 0 to 1 using image minimum and maximum. The RQUNet-VAE was trained on each image in the dataset to obtain all unknown model parameters. All the Sentinel-2 datasets are available on Github [1].

[1] https://github.com/trile83/RQUNetVAE

For our image segmentation experiments in Section III-D we used National Agriculture Imagery Program (NAIP) images coinciding with high resolution ground truth land cover data of each pixel in an image [18]. We acquired NAIP Nature Color imagery of northern Virginia consisting of RGB bands which is similar to some commercial satellite imagery (e.g. BlackSky, Planet). For training and validation (ground truth) we used the 1m resolution land cover dataset for the Chesapeake Bay watershed [1]. The dataset contains six classes - Water, Tree and shrubs, Herbaceous, Barren, Impervious (roads) and Impervious (other). We combined Impervious (other) and Impervious (roads) together into an integrated Impervious class. This provided us with four classes Water, Tree and Shrubs, Grass, and Impervious. For the segmentation experiments these classes were grouped into three classes, Vegetated (tree, grass, shrubs), Water and Impervious.

2) Evaluated Algorithms: Since RQUNet-VAE Scheme 1 is based on harmonic analysis, we compared it to state-of-the-art approaches including wavelet CDF 9/7 [33], [30], [6], curvelet [3] and Riesz Dyadic wavelet kernel [23], while the iterative Scheme 2 method was compared to state-of-the-art iterative methods such as directional TV-L2 [27], [10], [31] and GIAF [23].

Note that since the competitor methods were designed for gray-scale images, the methods were applied independently to each band of every multi-band image. The RQUNet-VAE, on the other hand, was designed for multi-channels images.

3) Evaluation Metrics:

a) Image reconstruction: To evaluate the ability of an algorithm to reduce the noise of an image, this section defines two commonly used measures. Given a clean multi-channel image $\underline{\underline{f}} \in \mathbb{R}^{|\Omega| \times P}$ as ground-truth and a denoised image $\underline{\underline{f}}^{\dagger} \in \mathbb{R}^{|\Omega| \times P}$, we use

- peak-signal-to-noise-ratio (PSNR):

$$\text{PSNR} = 10 \log \frac{\max\left(\underline{\underline{f}}\right)}{\text{MSE}}, \text{MSE} = \frac{1}{n_1 n_2 P} \left\| \underline{\underline{f}} - \underline{\underline{f}}^{\dagger} \right\|_{\ell_2}^2$$

- Structural similarity index (SSIM) [41]:

$$\text{SSIM} = \frac{\left(2\mu_f \mu_{f^\dagger} + c_1\right)\left(2\sigma_{ff^\dagger}\right)}{\left(\mu_f^2 + \mu_{f^\dagger}^2 + c_1\right)\left(\sigma_f^2 + \sigma_{f^\dagger}^2 + c_2\right)}$$

where (μ_f, σ_f^2) and $(\mu_{f^\dagger}, \sigma_{f^\dagger}^2)$ are mean and variance of ground-truth $\underline{\underline{f}}$ and its denoised image $\underline{\underline{f}}^{\dagger}$ and $\sigma_{ff^\dagger}$ are their covariance. $c_1 = (0.01r)^2, c_2 = (0.03r)^2$ where r is the dynamic range of pixel-values.

b) Image segmentation: To evaluate our RQUNet-VAE for segmentation of noisy test data, given that our statistical method provides uncertainty quantification, we propose the following evaluation framework:

Given a noisy test image $\underline{\underline{f}} \in \mathbb{R}^{|\Omega| \times P}$ as an input of the trained RQUNet-VAE segmentation and its ground-truth $\underline{\underline{f}}^{\text{gt}} = \left[f_l^{\text{gt}}\right]_{l \in \Omega} \in \mathbb{R}^{n_1 \times n_2 \times K}, f_l \in \mathbb{R}^K$ and the trained RQUNet-VAE by minimizing the loss function (37) from the clean dataset, we run prediction n time to obtain segmented masks $u^{(i)}\left(\underline{\underline{f}}\right) = \left[u_l^{(i)}\left(\underline{\underline{f}}\right)\right]_{l \in \Omega} \in \mathbb{R}^{|\Omega| \times K}, u_l^{(i)}\left(\underline{\underline{f}}\right) \in \mathbb{R}^K$ for $i = 1, \ldots, n$. For segmentation problem, class-balanced accuracy is defined for every pixel $l \in \Omega$:

$$\widehat{p}_l^n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\{u_l^{(i)}(\underline{\underline{f}}), f_l^{\text{gt}}\}}, \ \delta_{\{u_l^{(i)}(\underline{\underline{f}}), f_l^{\text{gt}}\}} = \begin{cases} 1, & u_l^{(i)}(\underline{\underline{f}}) = f_l^{\text{gt}} \\ 0, & \text{else} \end{cases} \tag{38}$$

which is modeled by a Binomial random variable approximated by Normal distribution (for large n):

$$Y_i := \delta_{\{u_l^{(i)}(\underline{\underline{f}}), f_l^{\text{gt}}\}} \sim \text{Bernoulli}\left(\widehat{p}_l^n, 1\right), \tag{39}$$

$$p_l^n = \frac{1}{n} \sum_{i=1}^{n} Y_i \sim \frac{1}{n} \text{Binomial}\left(\widehat{p}_l^n, n\right) \approx \mathcal{N}\left(\widehat{p}_l^n, \frac{\widehat{p}_l^n (1 - \widehat{p}_l^n)}{n}\right). \tag{40}$$

4) Training Procedure: For comparision with other state-of-the-art-methods, all parameters have been optimized for minimal mean-square-error (MSE) via heuristic search for each individual training image.

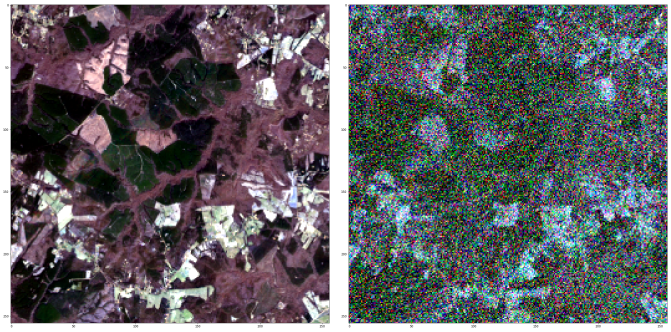
Since RQUNet-VAE is a hybrid model of deterministic high-order Riesz-Quincunx wavelet, we apply a training procedure for UNet-VAE with 100 epochs and a batch size of 16 using the Adam optimization method [12]. The source code of our training in PyTorch can be found at https://github.com/trile83/RQUNetVAE.

Image Set		Sentinel-2, std = 0.04
Number of Images		10
RQUNet-VAE scheme 1	PSNR	**38.693**
	SSIM	**0.969**
Riesz Dyadic	PSNR	37.433
	SSIM	0.959
curvelet	PSNR	36.314
	SSIM	0.942
wavelet CDF 9/7	PSNR	36.061
	SSIM	0.945
RQUNet-VAE scheme 2	PSNR	**39.087**
	SSIM	**0.971**
GIAF-Riesz Dyadic	PSNR	38.987
	SSIM	0.969
TV-L2 ($L = 2$)	PSNR	38.522
	SSIM	0.968
TV-L2 ($L = 9$)	PSNR	38.53
	SSIM	0.968

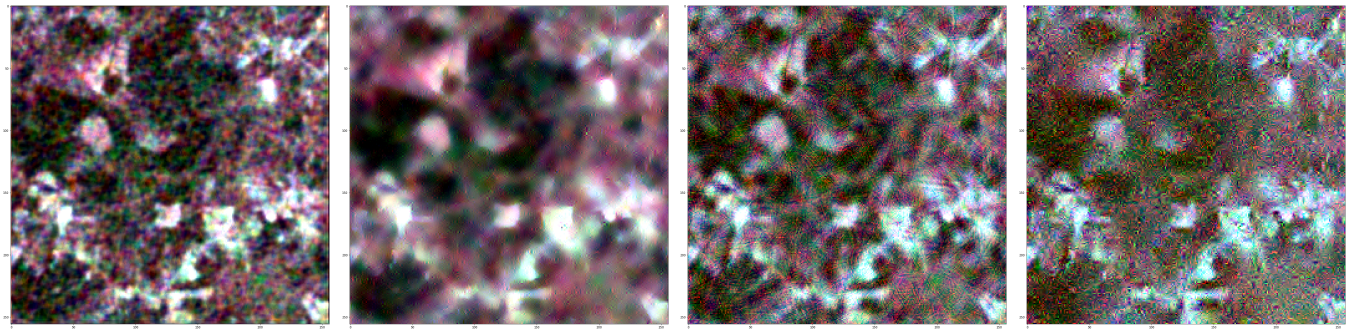
TABLE I
PSNR AND SSIM: MEAN OVER THE THREE IMAGE SETS (1ST ND 2ND BEST IN BOLD).

B. Denoising Experiments

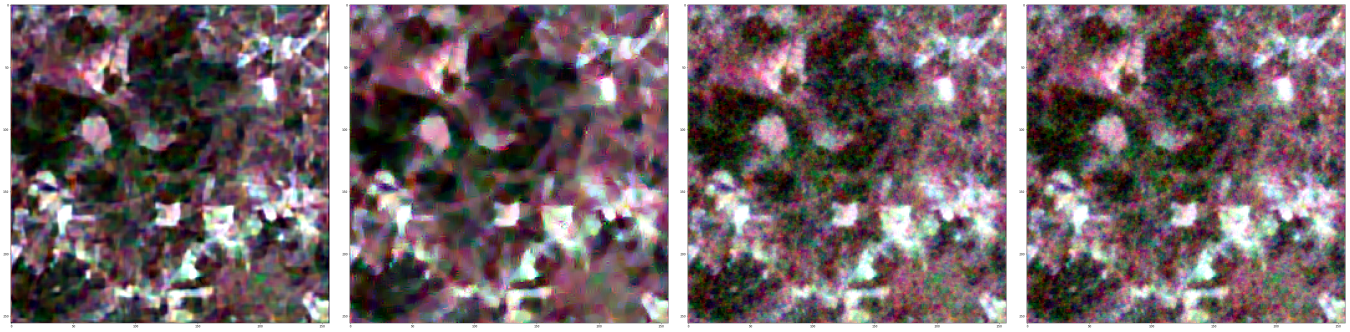
To impose artificial noise on images for evaluation, Gaussian noise was added to each band of the Sentinel-2 images with a standard deviation of std $= 0.04$. All images were normalized to interval $[0, 1]$ using the image minimum and maximum. To give some examples, original image shown in Figure 2(a) and the corresponding image with added noise in Figure 2(b). (All ten test images with added noise are visualized in SM).



(a) Original image (b) Noisy image, $\sigma = 0.04$



(c) RQUnet-VAE, $(\text{PSNR}, \text{SSIM}) = (38.502, 0.966)$ (d) Riesz Dyadic, $(\text{PSNR}, \text{SSIM}) = (37.555, 0.96)$ (e) curvelet, $(\text{PSNR}, \text{SSIM}) = (36.494, 0.944)$ (f) wavelet CDF 9/7, $(\text{PSNR}, \text{SSIM}) = (35.646, 0.936)$

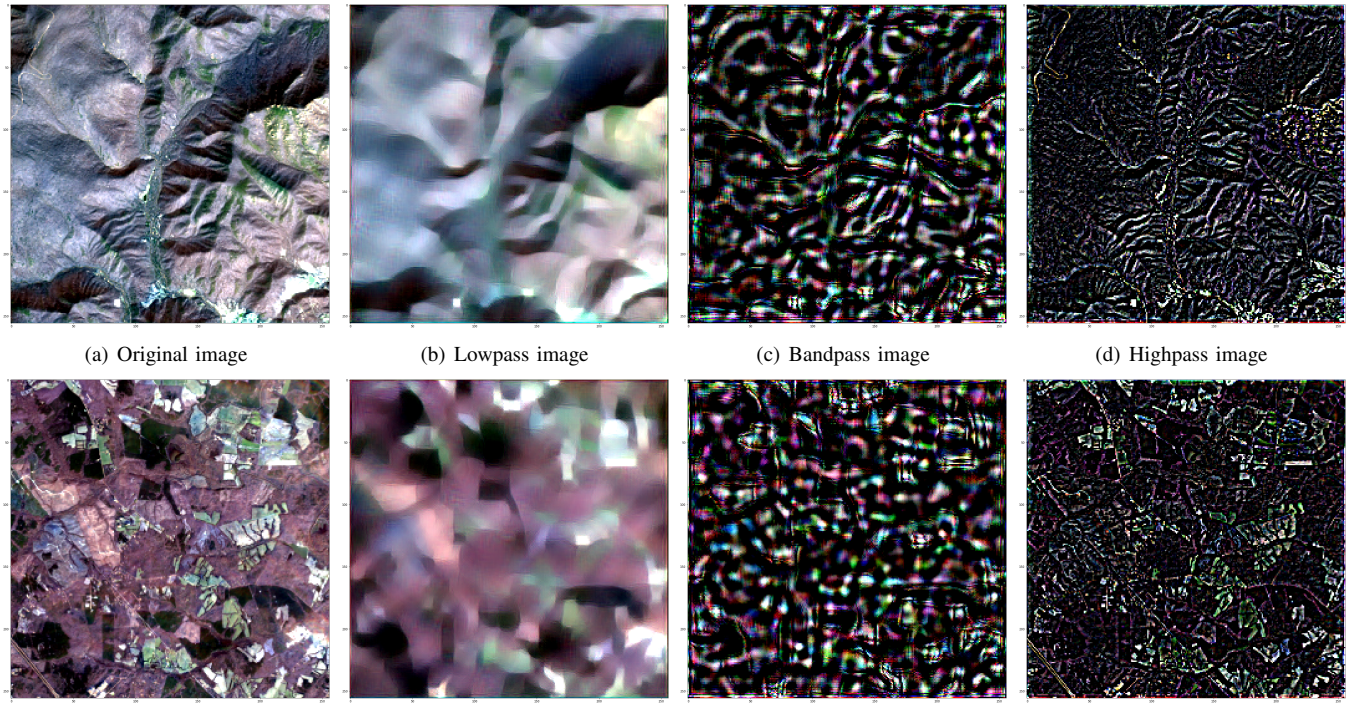


(g) RQUnet-VAE scheme 3, $(\text{PSNR}, \text{SSIM}) = (38.822, 0.97)$ (h) GIAF-Riesz Dyadic, $(\text{PSNR}, \text{SSIM}) = (38.936, 0.97)$ (i) TV-L2 $L = 2$, $(\text{PSNR}, \text{SSIM}) = (38.513, 0.967)$ (j) TV-L2 $L = 9$, $(\text{PSNR}, \text{SSIM}) = (38.535, 0.966)$

Fig. 2. Qualitative image denoising results for RQUnet-VAE and competitor approaches. Image denoising by the RQUnet-VAE with standard deviation $\sigma = 0.04$ and optimal α^*.

1) Quantitative Results: Table I provides the results of the comparison of the RQUNet-VAE Schemes 1 and 2 with the state-of-the-art methods described in Section III-A2 using the evaluation metrics described in Section III-A3. For the ten test images described in Section III-A1 the proposed RQUNet-VAE Scheme 2 yields the highest Peak-Signal-to-Noise-Ratio (PSNR) and the highest Structural Similarity Index (SSIM) of all competitors (Table 1). Among all the approaches based on harmonic analysis, RQUNet-VAE Scheme 1 yields the best results. Note GIAF also iteratively computes scaling and wavelet coefficients, similar to our RQUNet-VAE Scheme 2. The main difference, however, is that GIAF employs Riesz Dyadic wavelet kernel for scaling and wavelet functions whereas RQUNet-VAE Scheme 2 uses our adapted RQUNet-VAE. Using this scheme, allows more redundant parameters to better learn specific features from the data and thus, better adapt to specific images. (Denoising results for all other test images are visualized in SM).

2) Qualitative Results: Figures 1, 2 and 3 in SM are original, noisy and denoised images with the 2nd scheme of our RQUNet-VAE. Figures 2(c)-2(j) provide qualitative results to assess how well the RQUNet-VAE schemes are able to reduce artificially added noise individual Sentinel2 images. Figure 2(c) shows the result of denoising the noisy image of Figure 2(b) using our RQUNet-VAE Scheme 1. Since Scheme 1 is based on harmonic analysis, we compare it to noise reduction using a Riesz Dyadic wavelet kernel [23] in Figure 2(d), using curvelets [3] in Figure 2(e), and using wavelet CDF 9/7 [33], [30], [6] in Figure 2(f). The RQUNet-VAE Scheme 1 best preserves the edges of objects of the original image. The Riez Dyadic wavelet kernel in Figure 2(d) yields a blurred denoised image. In contrast, wavelets (Figure 2(f)) better retain edges, but much of the noise remains. Across all these figures, it is clearly discernible



(a) Original image (b) Lowpass image (c) Bandpass image (d) Highpass image

(e) Original image (f) Lowpass image (g) Bandpass image (h) Highpass image

Fig. 3. Image decomposition by RQUNet-VAE by spectral analysis with a threshold $T = 3$ for spectral histogram. Parameters: number of iteration Iter $= 30$, a smoothing parameter $\alpha = 0.4$.

that the proposed RQUNet-VAE Scheme 1 provides the best combination of noise reduction and delineation of edges and objects.

Figure 2(g) shows the denoised image using the RQUNet-VAE Scheme 2, compared to other iterative methods including GIAF [23] in Figure 2(h) and directional TV-L2 [27], [10], [31] using a number of directions $L = 2$ in Figure 2(i) and $L = 9$ in Figure 2(j). Figure 2 (g) illustrates that RQUNet-VAE significantly reduces artefacts while preserving texture pattern, contrast and sharp edges of objects in the reconstructed image, such that it is most similar to the original image. Note that the GIAF-Riesz dyadic method in Figure 2 (h) also performs very well, but a reconstructed images also contained some artefacts. The RQUNet-VAE Scheme 2 again provides the best trade-off between reduction of noise and delineation of edges and objects, where GIAF (Figure 2(h)) oversmoothes edges between objects while TV-L2 (Figure 2(i) and Figure 2(j)) still retain clearly visible noise. This demonstrates that the RQUNet-VAE with learned and deterministic frames increases sparsity in a generalized Besov space. Section III-D will demonstrate how this balance of noise reduction and discrimination of edges improves machine learning results when applying the RQUNet-VAE to segmentation of high resolution images for land cover mapping.

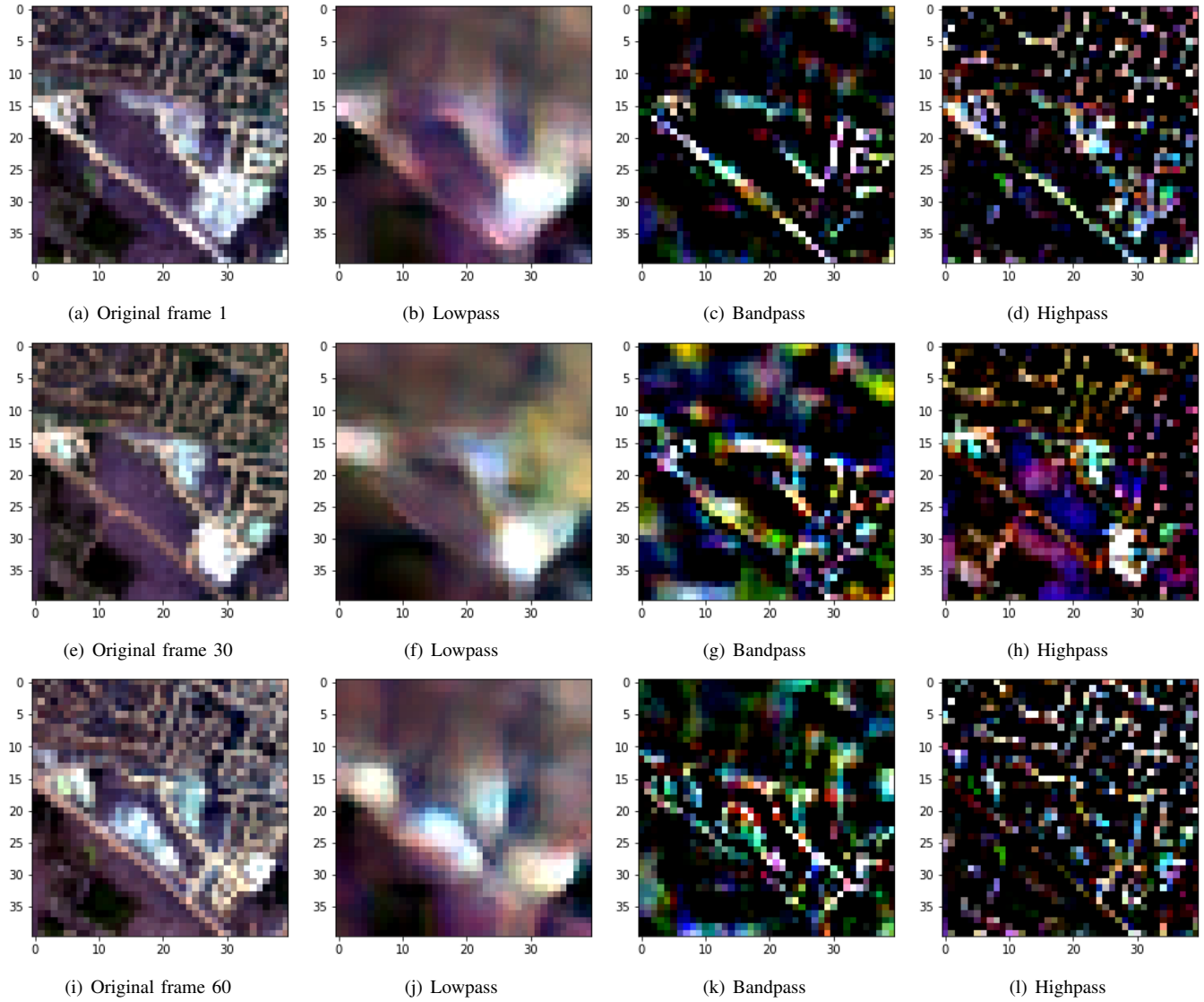
C. Image Decomposition Experiments

This section evaluates how RQUNet-VAE Scheme 2 decomposes and subsequently denoises images. RQUNet-VAE Scheme 2 uses an iterative scheme to decompose an image into (i) a lowpass image, (ii) a bandpass image, and (iii) a highpass image. Denoising can be achieved by truncating highpass features. The RQUNet-VAE was applied to the Sentinel-2 image dataset described in Section III-A1. Figure 3 shows examples of the decomposition for four images. Figures 3(a,e) shows the original Sentinel2 images of a typical rural landscape with forest cover, cultivated fields and small, distributed structures. The corresponding lowpass images in Figures 3(b,f) captures the larger land cover parcels, without the fine-scale texture which has been removed. The corresponding bandpass in Figures 3(c,g) captures most of the signal of texture, whereas the highpass image in Figure Figures 3(d,h) captures very fine-scale texture, oscillating patterns, along with noise.

a) Time series decomposition: The RQUNet-VAE was applied to a Sentinel2 time series decomposition by using Haar wavelet smoothing in time as a diffusion process, (Section II-E). The proposed smoothing technique is therefore simultaneously applied in spatial domain (by RQUNet-VAE) and temporal domain (by Haar wavelet in time), following Algorithm 2 in SM. The smoothing parameter was set at $\alpha = 0.03$ for all decompositions.

The Sentinel2 time series was comprised as follows: $\left\{\underline{\underline{f_i}}\right\}_{i=1}^{80}$, $\underline{\underline{f_i}} \in \mathbb{R}^{40\times40\times3\times99}$, where each of the 99 images have three channels of image size 40×40. The time series is then padded to $\underline{\underline{f_i}} \in \mathbb{R}^{64\times64\times3\times120}$. The RQUNet-VAE is trained on each image in the padded dataset to obtain all unknown model parameters. An Adam optimizer was used to train RQUNet-VAE with 200 epochs and batch-size of 16.

The trained parameters are then applied to Algorithm 2 with 10 iterations for spectral decomposition. Note that Algorithm 2 is an extension of generalized intersection algorithms with



(a) Original frame 1 (b) Lowpass (c) Bandpass (d) Highpass

(e) Original frame 30 (f) Lowpass (g) Bandpass (h) Highpass

(i) Original frame 60 (j) Lowpass (k) Bandpass (l) Highpass

Fig. 4. Time series decomposition by RQUNet-VAE with GIAF based spectral decomposition for site 296.

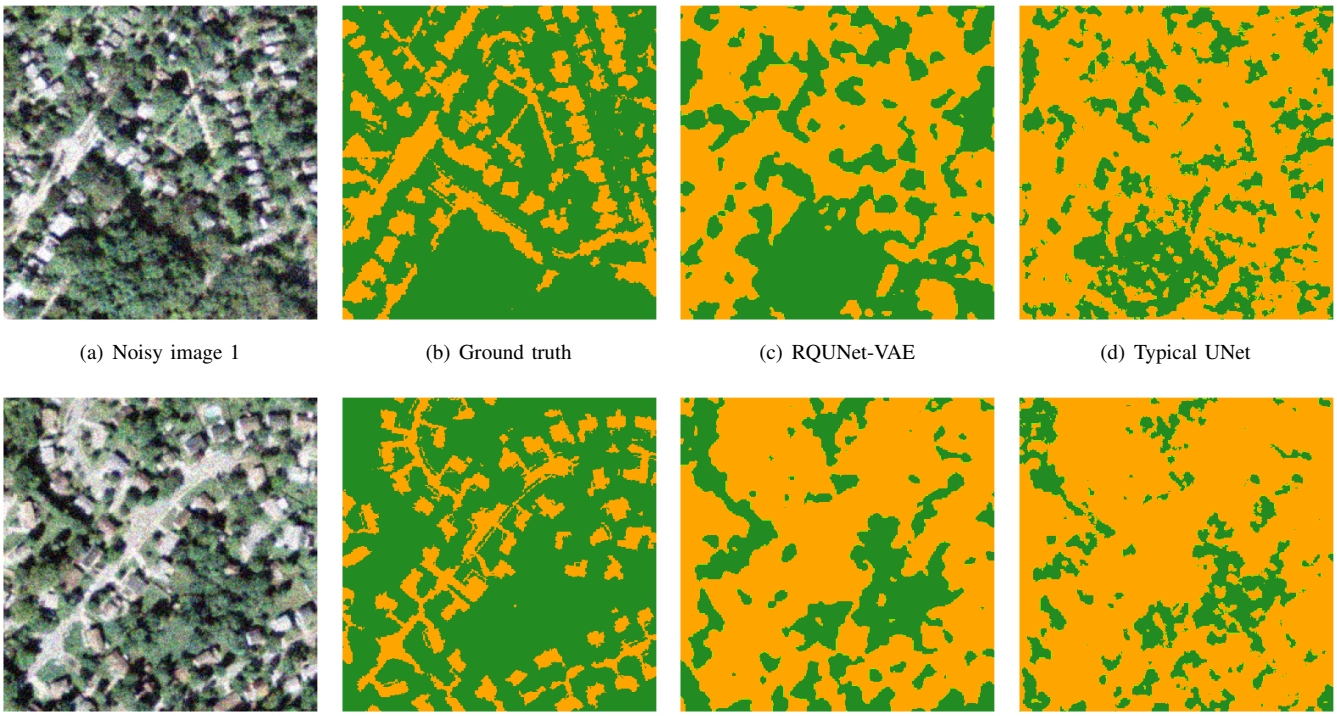
fixpoints [23] for video processing via spectral decomposition into lowpass, bandpass and highpass videos.

Figure 4 illustrates the time series decomposition with the RQUNet-VAE and diffusion process. The smoothing parameter was set at $\alpha = 0.03$ for all decompositions. Similar to image decomposition in the previous sections, this time series decomposition extracts a homogeneous time series and a residual time series which contains small objects (e.g. roads), noise, and texture.

D. Image Segmentation by RQUNet-VAE

The RQUNet-VAE was applied to the task of high resolution aerial image segmentation. The goal is to demonstrate that the RQUNet-VAE makes the image segmentation more robust to noise compared to the existing U-Net architecture. We first provide a brief background to the problem of image segmentation in Section III-D1. Then, in Section III-D2 we apply 1) the traditional U-Net architecture and 2) the RQUNet-VAE to the problem of image segmentation and show that RQUNet-VAE yields better segmentation results where artificial noise is added to images. This shows that our proposed RQUNet-VAE makes the existing U-Net architecture more robust to noise.

1) Background on Image Segmentation: The conventional image segmentation procedure has the following steps, (i) pre-processing to remove noise and unwanted small objects, (ii) segmentation, (iii) post-processing with morphological operators. This 3-step process requires a large number of parameters that need to be selected by an expert. Therefore, attempts have been made to incorporate a smoothing term into various models. The Mumford-Shah model [17] was proposed to introduce a smoothing term in a minimization approach for segmentation. However, computation of this smoothing term is an NP-hard problem and therefore not feasible [17]. Later, the Chan-Vese segmentation model [4] was proposed that was solvable by relying on the level set method [19]. By adding a smoothing term in a variational formulae, this model successfully segments objects in a noisy background. Inspired by Chan-Vese model [4], we use RQUNet-VAE for segmentation by introducing a smoothing term, i.e. Riesz-Quincunx wavelet truncation, directly into the UNet-VAE. The



(a) Noisy image 1 (b) Ground truth (c) RQUNet-VAE (d) Typical UNet

(e) Noisy image 2 (f) Ground truth (g) RQUNet-VAE (h) Typical UNet

Fig. 5. Segmentation result with RQUNet-VAE as a smoothing term for noisy images. a,e) NAIP images with artificial noise added (std $= 0.08$) as input to segmentation; b,f) ground-truth segmentation masks; (c,g) segmentation masks returned by our RQUNet-VAE with a smoothing parameter $\alpha = 0.5$; (d,h) segmentation mask returned by the typical UNet architecture [24].

advantage is that the RQUNet-VAE includes the smoothing term inside the Unet and thus combines (1) the ability of a Unet to learn features from an image, with (2) denoising capabilities enabled by the Riesz-Quincunx wavelet truncation. This smoothing term should eliminate small scale objects in a segmented image, e.g. texture and noise. Then, there is no need for separate pre- or post-processing steps, as these are all performed by the RQUNet-VAE segmentation.

2) Experimental Evaluation of RQUNet-VAE for Image Segmentation: We apply both 1) the traditional U-Net [24] and 2) RQUNet-VAE to the problem of segmentation using the 20 NAIP images described in Section III-A1 using two classes of land cover, impervious and vegetated. Figure 5 shows the segmentation results for two sample images after noise has been added (Section III-B, standard deviation of 0.08). The corresponding ground truth land cover masks are shown in Figures 5(b) and 5(f). Table II provides quantitative segmentation results, including overall accuracy, precision, recall, and F1-score for both of the two classes. Due to the added noise, the segmentation accuracy of the U-Net is very low at 0.6640 and 0.6170 for the two images, respectively Table II. The precision and recall for the two classes (non-impervious and impervious), indicate that this low accuracy is attributed to a bias towards the impervious class, as suggested by the low recall for the non-impervious class (0.3780 for the traditional UNet). This implies that more than two out of five impervious pixels are incorrectly classified as vegetated.

The performance of the traditional UNet is much lower with the added noise than on the original, which indicates that the addition of the noise substantially confuses the U-Net. In contrast, the RQUNet-VAE coped much better with the noise, yielding higher accuracy values of 0.7057 ± 0.0013 and 0.6634 ± 0.0012 for the two images, respectively. The negative impact of noise on the segmentation accuracy is substantially reduced when using the RQUNet-VAE which makes the U-Net more robust to noise. Furthermore, the RQUNet-VAE is better able to discriminate the vegetated class, shown by the higher F1-scores Table II.

A more comprehensive evaluation of RQUNet-VAE vs U-Net on 20 images with various levels of noise is given in Figure 6. This figure shows the mean accuracy (solid line) of the U-Net and the RQUNet-VAE across all 20 images. When using the original image with no noise ($\sigma = 0$), the RQUNet-VAE only provides a marginal improvement over U-Net. As the noise level σ is increased, both RQUNet-VAE and U-Net exhibit a drop in accuracy. However, that the drop in accuracy is substantially slower for RQUNet-VAE . In particular, for noise values around $\sigma = 0.1$, the RQUNet-VAE has up to 10% higher accuracy than U-Net. This improvement stems from the variational auto-encoder approach, which, through variational changes to the latent representation of an image, allows the identification of pixels with a high probability of belonging to one class while being assigned to another class by the deterministic U-Net. Figure 6 gives the results using two parameters setting of RQUNet-VAE , $\alpha = 0.5$ and $\alpha = 1.0$. In both cases the resulting accuracy is nearly identical (the two lines are almost perfectly on top of each other), showing that the the performance is not very sensitive to the choice of α and thus that RQUNet-VAE is robust to non-optimal

TABLE II
QUANTITATIVE COMPARISON BETWEEN RQUNET-VAE AND THE TRADITIONAL UNET ARCHITECTURE. SINCE RQUNET-VAE USES A VARIATIONAL (NON-DETERMINISTIC) APPROACH, WE PROVIDE MEAN AND STANDARD DEVIATION AGGREGATED OVER 20 RUNS. PRECISION, RECALL, AND F1-SCORE ARE PROVIDED FOR BOTH CLASSES [VEGETATED, IMPERVIOUS].

Algorithm	Accuracy (overall)	Precision (per class)	Recall (per class)	F1-Score (per class)
RQUNet-VAE Mean (Image 1)	0.7057	[0.9047 0.4498]	[0.5457 0.8660]	[0.6807 0.5921]
RQUNet-VAE Stdev (Image 1)	±0.0013	±[0.0008 0.0014]	±[0.0030 0.0016]	±[0.0024 0.0012]
Traditional UNet (Image 1)	0.6640	[0.9460 0.3960]	[0.3780 0.9500]	[0.5400 0.5590]
RQUNet-VAE Mean (Image 2)	0.6734	[0.9837 0.3258]	[0.3662 0.9806]	[0.5337 0.4890]
RQUNet-VAE Stdev (Image 2)	±0.0012	±[0.0006 0.0009]	±[0.0027 0.0007]	±[0.0028 0.0010]
Traditional UNet (Image 2)	0.6170	[0.9860 0.2900]	[0.2440 0.9890]	[0.3920 0.4490]

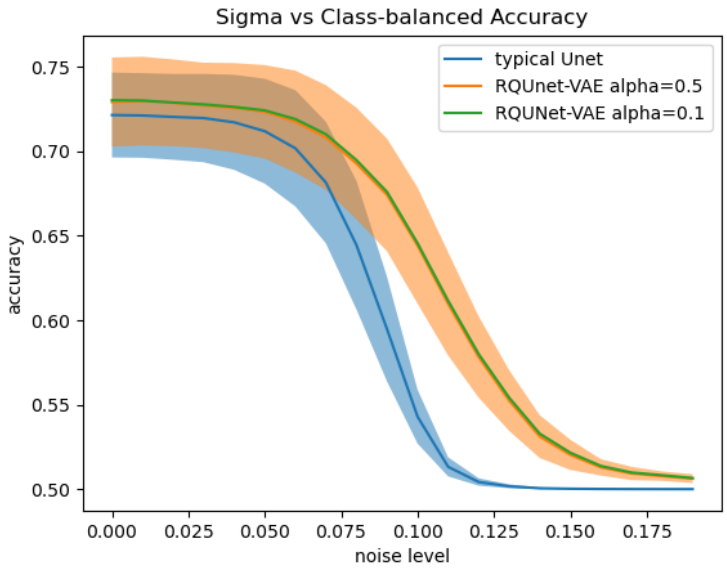


Fig. 6. Outcome of segmentation experiments using the RQUNet-VAE and traditional UNet applied to 20 NAIP images with increasing added noise (0-0.20). The confidence intervals (blue and orange bands) are calculated from the mean and standard deviation of accuracy of 50 prediction iterations for each noise level and alpha value (0.5, 0.1) in the RQ scheme. This illustrates the results of variational terms of RQUNet-VAE.

choices of α. In summary, Figure 6 shows that our proposed RQUNet-VAE is much more robust to Gaussian noise added to an image, as the reduction of segmentation accuracy with higher noise levels, is less pronounced.

IV. CONCLUSION

In this work, we introduced the RQUNet-VAE which augments the existing UNet architecture with a generalized wavelet expansion approach that we extended to a diffusion process to enable spectral decomposition. To the best of our knowledge this is the first approach that enables image decomposition and image denoising using a variational variant of the UNet architecture. An important application of this decomposition is denoising, achieved by truncating highpass features, that is, discarding information of decomposed images having the highest variance. We apply our proposed RQUNet-VAE to image denoising and segmentation of multi-band satellite images and their time-series which often contain noise due to multiple causes. During quantitative comparisons of noise reduction the RQUNet-VAE yields the highest PSNR and SSIM of all competitor methods. Among all the approaches based on harmonic analysis, RQUNet-VAE Scheme 2 yields the best results. For the application of satellite image denoising our proposed RQUNet-VAE provides superior qualitative performance compared to other competitors. The denoising by RQUNet-VAE Scheme 1 was visually compared to the noise reduction using a Riesz Dyadic wavelet and curvelets and using wavelet CDF and it provided the best combination of noise reduction and delineation of edges and objects. Furthermore the propose RQUNet-VAE Scheme 2, was compared to other iterative methods including GIAF and directional TV-L2 and it resulted in the best trade-off between reduction of noise and delineation of edges and objects, whereas the GIAF oversmoothed edges between objects while TV-L2 clearly retained visible noise.

To quantitatively measure the improvement of the RQUNet-VAE against the traditional UNet, we applied it to high resolution aerial image segmentation. Our experiments show only a slight improvement over the traditional UNet for segmantation of the original images with little noise. However, as artificial noise is added to images, we observe that the segmentation quality of the UNet decreases more rapidly than when using the RQUNet-VAE . The superior performance of RQUNet-VAE is due to a neural network in UNet-VAE with a learned dictionary obtained from the training procedure to increase the level of sparsity for an input image in a generalized Besov space. This property is due to a fundamental concept in signal processing, that the signal is sparse in some transformed domain. This demonstrates that the RQUNet-VAE architecture is

able to substantially increase the robustness to noise compared to the traditional UNet for the application of satellite image segmentation.

V. Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), Space-based Machine Automated Recognition Technique (SMART) program, via contract 2021-20111000003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. The research was conducted in collaboration with BlackSky. We thank Diego Torrejon and Isaac Corley of BlackSky for valuable comments on the manuscript.

References

[1] J. Allenby and C. Phelan. Implementing technology and precision conservation in the Chesapeake bay. *Chesapeake Conservancy*, 2013.

[2] J.-F. Cai, R. H. Chan, and Z. Shen. A framelet-based image inpainting algorithm. *Applied and Computational Harmonic Analysis*, 24(2):131–149, 2008.

[3] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise c2 singularities. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(2):219–266, 2004.

[4] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001.

[5] M. Claverie, J. Ju, J. G. Masek, J. L. Dungan, E. F. Vermote, J.-C. Roger, S. V. Skakun, and C. Justice. The harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote sensing of environment*, 219:145–161, 2018.

[6] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *Journal of Fourier analysis and applications*, 4(3):247–269, 1998.

[7] P. Esser, E. Sutter, and B. Ommer. A variational U-Net for conditional appearance and shape generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8857–8866, 2018.

[8] M. Feilner, D. Van De Ville, and M. Unser. An orthogonal family of Quincunx wavelets with continuously adjustable order. *IEEE Transactions on Image Processing*, 14(4):499–510, 2005.

[9] G. Gilboa. A total variation spectral framework for scale and texture analysis. *SIAM journal on Imaging Sciences*, 7(4):1937–1961, 2014.

[10] T. Goldstein and S. Osher. The split Bregman method for L_1-regularized problems. *SIAM journal on imaging sciences*, 2(2):323–343, 2009.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[14] W. Kleynhans, B. P. Salmon, J. C. Olivier, F. Van den Bergh, K. J. Wessels, T. L. Grobler, and K. C. Steenkamp. Land cover change detection using autocorrelation analysis on MODIS time-series data: Detection of new human settlements in the gauteng province of south africa. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 5(3):777–783, 2012.

[15] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger. A probabilistic U-Net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018.

[16] J. Li and D. P. Roy. A global analysis of Sentinel-2A, Sentinel-2B and Landsat-8 data revisit intervals and implications for terrestrial monitoring. *Remote Sensing*, 9(9):902, 2017.

[17] D. B. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 1989.

[18] NAIP. National agriculture imagery program (NAIP) information sheet, 2015. https://www.fsa.usda.gov/Internet/FSA_File/naip_info_sheet_2015.pdf.

[19] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of computational physics*, 79(1):12–49, 1988.

[20] J. R. Partington, J. R. Partington, et al. *An introduction to Hankel operators*. Cambridge University Press, 1988.

[21] N. G. Polson, J. G. Scott, and B. T. Willard. Proximal algorithms in statistics and machine learning. *Statistical Science*, 30(4):559–581, 2015.

[22] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin. Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems*, 29, 2016.

[23] R. Richter, D. H. Thai, and S. F. Huckemann. Generalized intersection algorithms with fixpoints for image decomposition learning. *CoRR*, abs/2010.08661, 2020.

[24] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[25] D. P. Roy, H. Huang, L. Boschetti, L. Giglio, L. Yan, H. H. Zhang, and Z. Li. Landsat-8 and Sentinel-2 burned area mapping-A combined sensor multi-temporal change detection approach. *Remote Sensing of Environment*, 231:111254, 2019.

[26] D. P. Roy, Y. Jin, P. Lewis, and C. Justice. Prototyping a global algorithm for systematic fire-affected area mapping using MODIS time series data. *Remote sensing of environment*, 97(2):137–162, 2005.

[27] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[28] B. P. Salmon, W. Kleynhans, F. Van den Bergh, J. C. Olivier, T. L. Grobler, and K. J. Wessels. Land cover change detection using the internal covariance matrix of the extended Kalman filter over multiple spectral bands. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 6(3):1079–1085, 2013.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[30] W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM journal on mathematical analysis*, 29(2):511–546, 1998.

[31] D. H. Thai and C. Gottschlich. Directional global three-part image decomposition. *EURASIP Journal on Image and Video Processing*, 2016(1):1–20, 2016.

[32] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, 2020.

[33] M. Unser and T. Blu. Mathematical properties of the jpeg2000 wavelet filters. *IEEE transactions on image processing*, 12(9):1080–1090, 2003.

[34] M. Unser, N. Chenouard, and D. Van De Ville. Steerable pyramids and tight wavelet frames in $L_2(\mathbb{R}^d)$. *IEEE Transactions on Image Processing*, 20(10):2705–2721, Oct. 2011.

[35] M. Unser, D. Sage, and D. Van De Ville. Multiresolution monogenic signal analysis using the Riesz-Laplace wavelet transform. *IEEE Transactions on Image Processing*, 18(11):2402–2418, Nov. 2009.

[36] M. Unser and D. Van De Ville. Wavelet steerability and the higher-order Riesz transform. *IEEE Transactions on Image Processing*, 19(3):636–652, Mar. 2010.

[37] F. Van Den Bergh, K. J. Wessels, S. Miteff, T. L. Van Zyl, A. D. Gazendam, and A. K. Bachoo. Hitempo: A platform for time-series analysis of remote-sensing satellite data in a high-performance computing environment. *International Journal of Remote Sensing*, 33(15):4720–4740, 2012.

[38] T. Van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[39] E. Vermote, C. Justice, M. Claverie, and B. Franch. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sensing of Environment*, 185:46–56, 2016.

[40] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE transactions on Signal Processing*, 50(6):1417–1428, 2002.

[41] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

[42] C. E. Woodcock, T. R. Loveland, M. Herold, and M. E. Bauer. Transitioning from change detection to monitoring with remote sensing: A paradigm shift. *Remote Sensing of Environment*, 238:111558, 2020.

[43] J. C. Ye, Y. Han, and E. Cha. Deep convolutional framelets: A general deep learning framework for inverse problems. *SIAM Journal on Imaging Sciences*, 11(2):991–1048, 2018.

[44] R. Yin, T. Gao, Y. M. Lu, and I. Daubechies. A tale of two bases: Local-nonlocal regularization on image patches with convolution framelets. *SIAM Journal on Imaging Sciences*, 10(2):711–750, 2017.

[45] H. K. Zhang, D. P. Roy, L. Yan, Z. Li, H. Huang, E. Vermote, S. Skakun, and J.-C. Roger. Characterization of Sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences. *Remote sensing of environment*, 215:482–494, 2018.

[46] Z. Zhu. Change detection using Landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:370–384, 2017.

Riesz-Quincunx-Unet Variational Auto-Encoder for Satellite Image Denoising: Supplemental Material

Duy H. Thai[1], Xiqi Fei[1], Tri M. Le[1], Andreas Züfle[2], Konrad Wessels[1]

[1]George Mason University, Department of Geography and Geoinformation Science, USA

[2]Emory University, Department of Computer Science, USA

1 Additional Images

This section provides details on the ten satellite images used for testing of our proposed RQUNet-VAE approach. Figure 1 shows the original images, Figure 2 shows the same images with artificial noise added, and Figure 3 shows the denoised images using our proposed RQUNet-VAE.

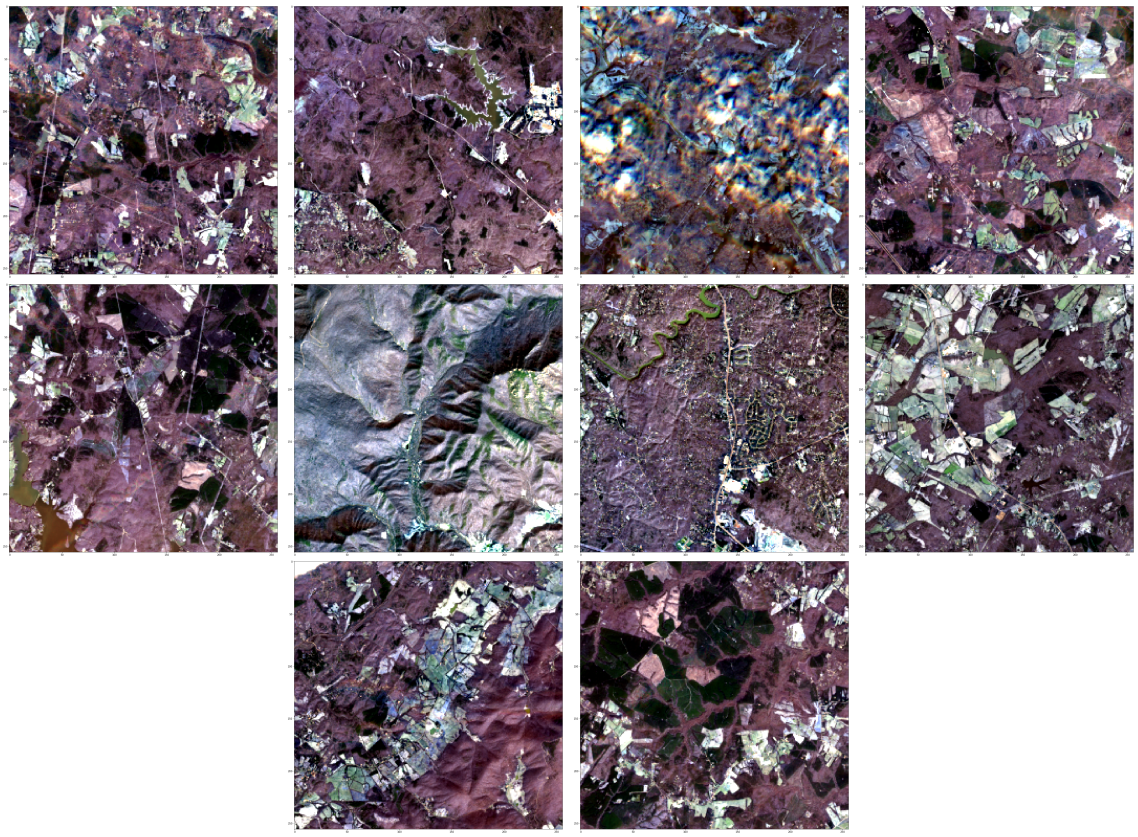


Figure 1: Original images.

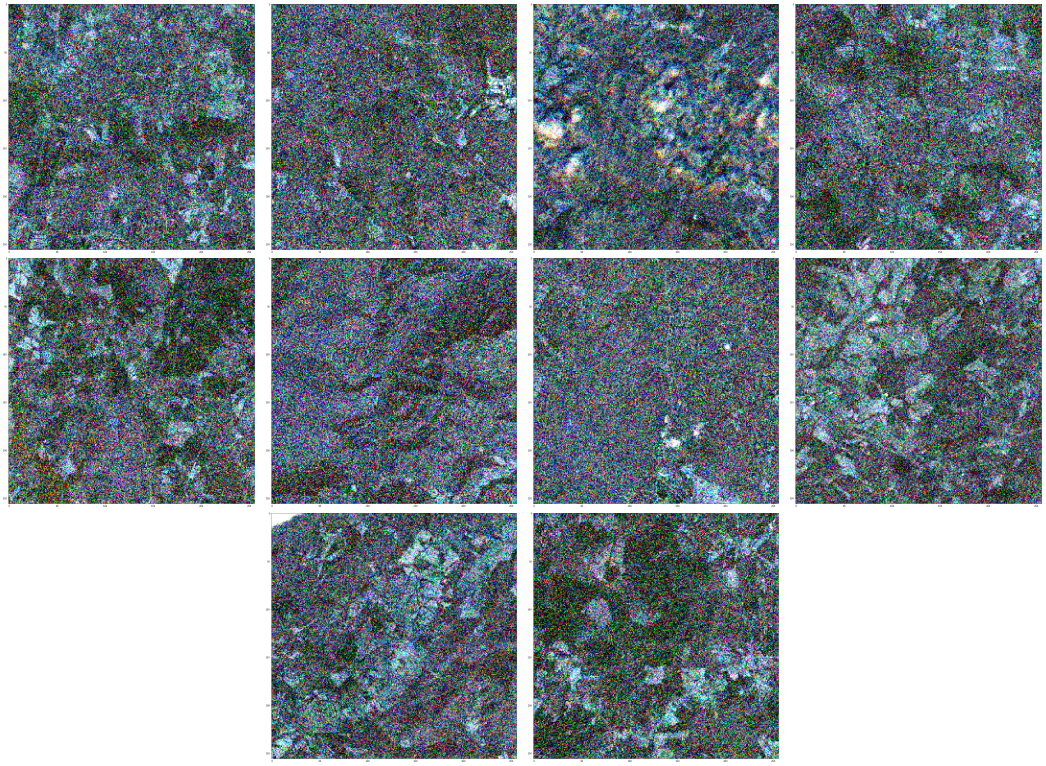


Figure 2: Noisy images with variance $\sigma = 0.04$.

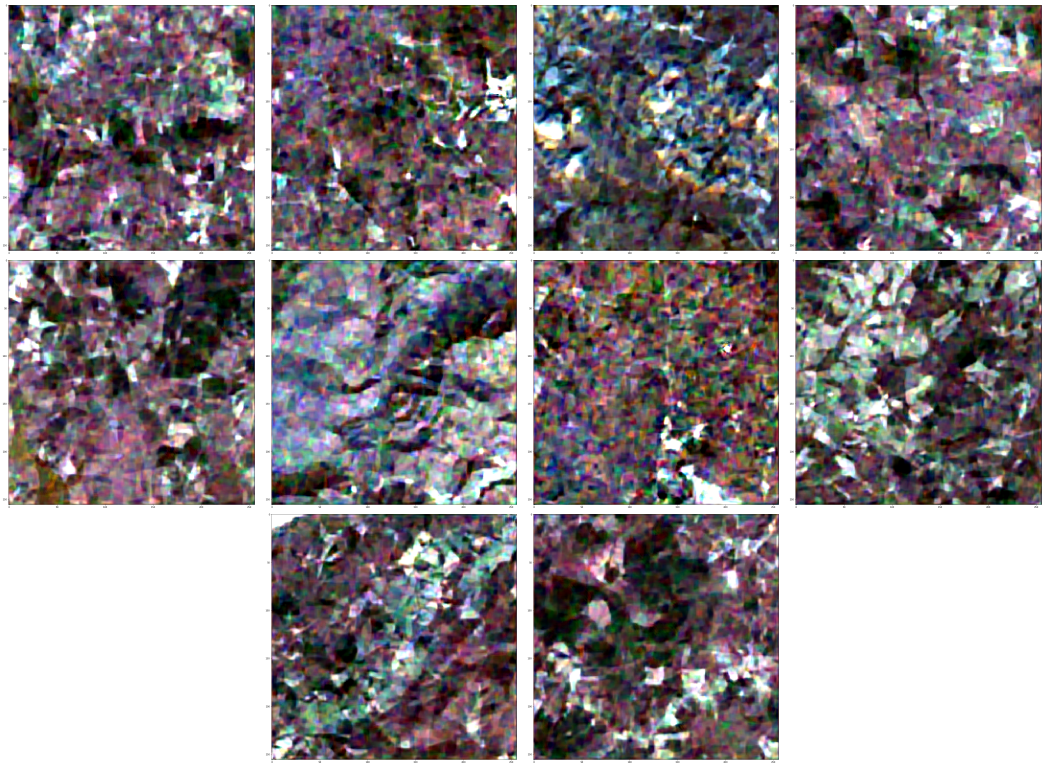
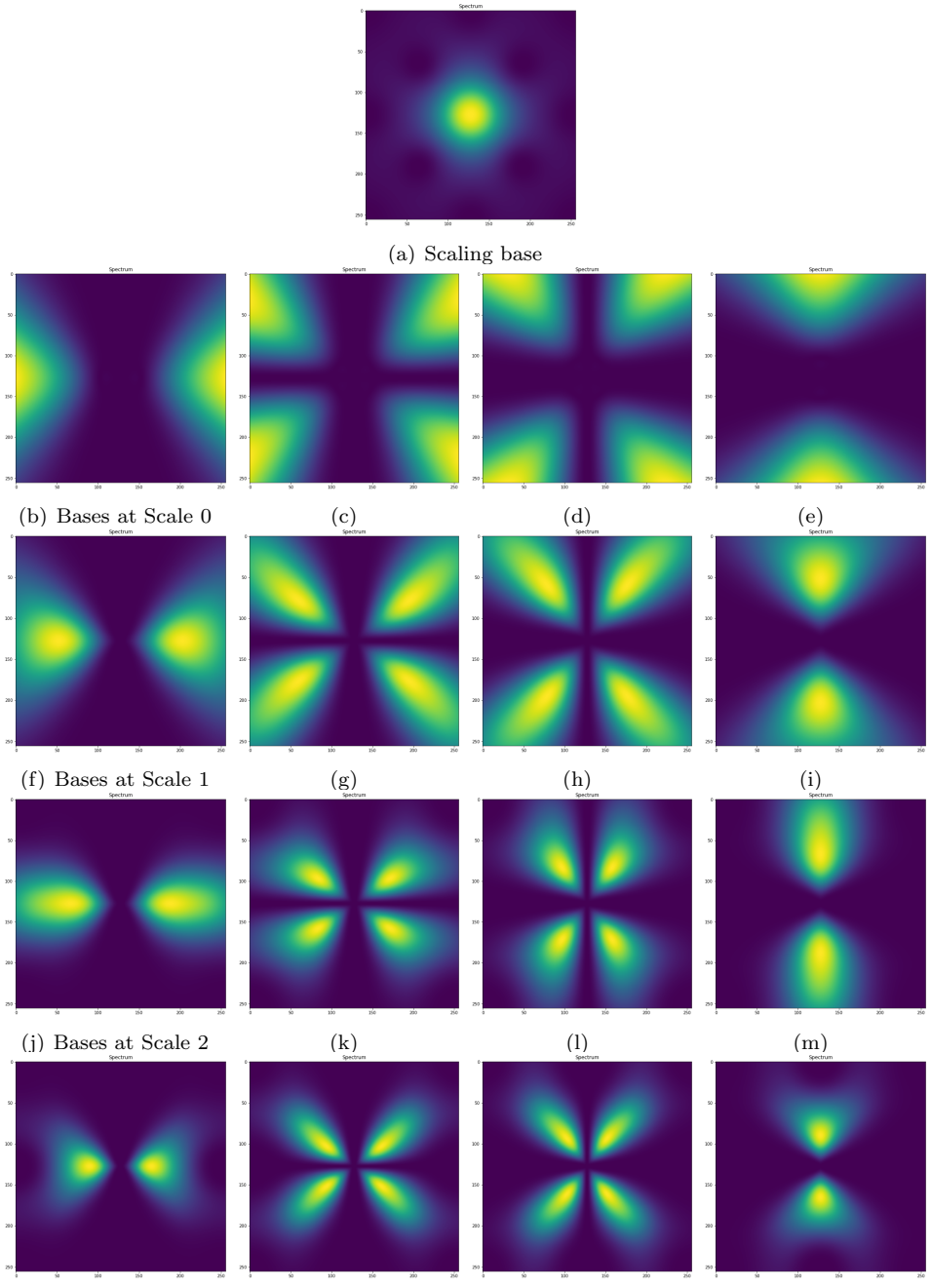


Figure 3: Denoised images for $\sigma = 0.04$ by RQUnet-VAE smoothing.

2 Riesz Quincunx Filter Banks



(a) Scaling base

(b) Bases at Scale 0 (c) (d) (e)

(f) Bases at Scale 1 (g) (h) (i)

(j) Bases at Scale 2 (k) (l) (m)

(n) Bases at Scale 3 (o) (p) (q)

Figure 4: This figure illustrates N-th order Riesz Quincunx Filter banks in the Fourier domain with 3 scales, a fractional order of Bspline $\gamma = 1.2$ and order of Riesz transform $N = 3$.

Figure 4 illustrates the N-th order Riesz Quincunx filter banks in the Fourier domain: rows are for 4 wavelet scales and columns are 4 directions for each scale corresponding to $N = 0, ..., 3$ orders in Riesz transform.

3 Detailed Encoder and Decoder of RQUNet-VAE

Figure 5 shows the detailed architecture of the RQUNet-VAE encoder (left) and decoder (right).

4 Additional Background and definitions

4.1 Hankel matrices and convolutional operations

The following list provides definition of Hankel matrix and its inverse and a set of convolution operations used in this work.

- Extended Hankel matrix of an image $\underline{f} \in \mathbb{R}^{|\Omega|}$:

$$\mathscr{H}_{d_1|n_2}(\underline{f}) = \begin{pmatrix} \mathscr{H}_{d_1}(f_1) & \dots & \mathscr{H}_{d_1}(f_{n_2}) \end{pmatrix} \in \mathbb{R}^{n_1 \times n_2 d_1},$$

  with a Hankel matrix:

$$\mathscr{H}_{d_1}(f_i) = \begin{pmatrix} f_i[1] & \dots & f_i[d_1] \\ \vdots & \ddots & \vdots \\ f_i[n_1] & \dots & f_i[n_1 + d_1 - 1] \end{pmatrix} \in \mathbb{R}^{n_1 \times d_1}.$$

- The extended Hankel matrix of a multi-channel image $\underline{\underline{f}} \in \mathbb{R}^{|\Omega| \times P}$:

$$\mathscr{H}_{n_1|d_2|P}\left(\underline{\underline{f}}\right) = \begin{pmatrix} \mathscr{H}_{n_1|d_2}(\underline{f_1}) & \dots & \mathscr{H}_{n_1|d_2}(\underline{f_P}) \end{pmatrix}.$$

  Given $\{e_k\}_{k=1}^{n_1}$ as an orthonormal basis of $\mathbb{R}^{n_1}$, $\left\{\underline{\tilde{e}_k} = \frac{1}{\sqrt{d_1}}\mathscr{H}_{d_1}(e_k)\right\}_{k=1}^{n_1} \subset \mathbb{R}^{n_1 \times d_1}$ is the orthonormal basis of $\mathcal{H}(n_1, d_1) = \left\{h : h \in \mathbb{R}^{n_1 \times d_1}\right\}$, i.e. $\langle \underline{\tilde{e}_k}, \underline{\tilde{e}_l} \rangle_{\mathrm{F}} = \delta_{kl}$ and an orthonormal expansion:

$$\mathscr{H}_{d_1}(f_i) = \sum_{k=1}^{n_1} \underbrace{\left\langle \underline{\tilde{e}_k}, \mathscr{H}_{d_1}(f_i) \right\rangle_{\mathrm{F}}}_{=\sqrt{d_1} f_i[k]} \underline{\tilde{e}_k}.$$

- The inverse of an extended Hankel matrices for a matrix $\underline{g} = \begin{pmatrix} \underline{g_1} & \dots & \underline{g_{n_2}} \end{pmatrix} \in \mathbb{R}^{n_1 \times d_1 n_2}, \underline{g_i} \in \mathbb{R}^{n_1 \times d_1}$ as:

$$\mathscr{H}^{\dagger}_{d_1|n_2}\left(\underline{g}\right) = \begin{pmatrix} \mathscr{H}^{\dagger}_{d_1}\left(\underline{g_1}\right) & \dots & \mathscr{H}^{\dagger}_{d_1}\left(\underline{g_{n_2}}\right) \end{pmatrix} \in \mathbb{R}^{|\Omega|}, \tag{1}$$

  having an inverse Hankel matrix:

$$\mathscr{H}^{\dagger}_{d_1}\left(\underline{g_i}\right) = \frac{1}{\sqrt{d_1}} \begin{pmatrix} \left\langle \underline{\tilde{e}_1}, \underline{g_i} \right\rangle_{\mathrm{F}} \\ \vdots \\ \left\langle \underline{\tilde{e}_{n_1}}, \underline{g_i} \right\rangle_{\mathrm{F}} \end{pmatrix} \in \mathbb{R}^{n_1}. \tag{2}$$

- Filter banks and their family of matrices for the Riesz-Quincunx wavelet:
  - scaling filter bank $\underline{\phi} \in \mathbb{R}^{d_1 \times d_2}$,
  - primal wavelet filter bank $\underline{\underline{\psi}} = \left\{\underline{\psi_1}, \dots, \underline{\psi_P}\right\} \in \mathbb{R}^{d_1 \times d_2 \times P}, \underline{\psi_i} \in \mathbb{R}^{d_1 \times d_2}$,

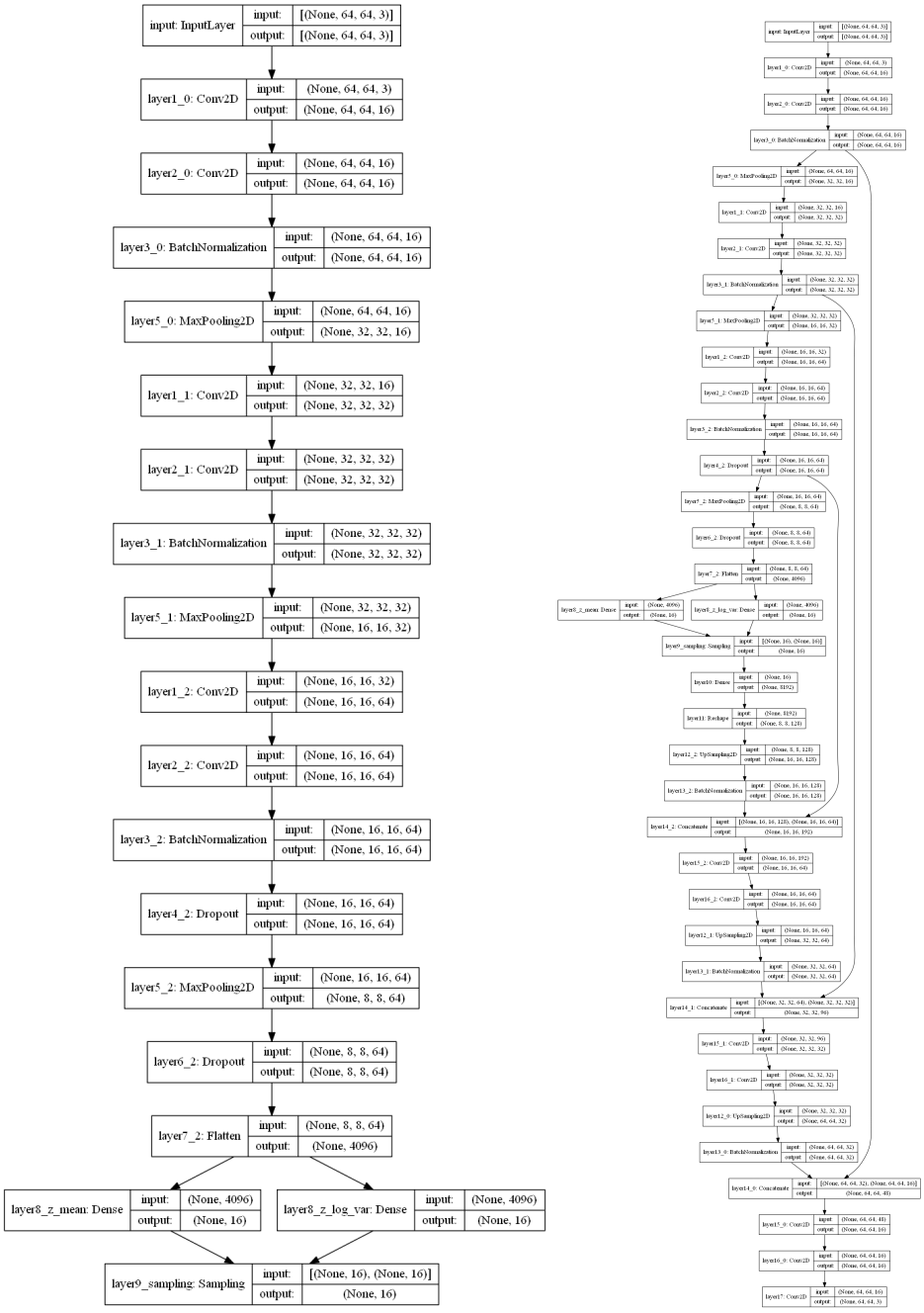


Figure 5: RQUnet-VAE Encoder (left) and Decoder (right)

– dual wavelet filter bank $\underline{\underline{\tilde{\psi}}} = \left\{\underline{\tilde{\psi}_1}, \dots, \underline{\tilde{\psi}_P}\right\} \in \mathbb{R}^{d_1 \times d_2 \times P}, \underline{\tilde{\psi}_i} \in \mathbb{R}^{d_1 \times d_2}$,

– having matrix forms

$\underline{\underline{\Psi}} = \left\{\underline{\Psi_1}, \dots, \underline{\Psi_P}\right\}$, $\underline{\underline{\tilde{\Psi}}} = \left\{\underline{\tilde{\Psi}_1}, \dots, \underline{\tilde{\Psi}_P}\right\}$ such that for $p \in 1, ..., P$:

$$\underline{\Psi_p} = \begin{pmatrix} \underline{\Psi_p^1} \\ \vdots \\ \underline{\Psi_p^{d_2}} \end{pmatrix} = \begin{pmatrix} \psi_{p,1}^1 & \dots & \psi_{p,d_2}^1 \\ \vdots & \ddots & \vdots \\ \psi_{p,1}^{d_2} & \dots & \psi_{p,d_2}^{d_2} \end{pmatrix}, \underline{\tilde{\Psi}_p} = \begin{pmatrix} \underline{\tilde{\Psi}_p^1} \\ \vdots \\ \underline{\tilde{\Psi}_p^{d_2}} \end{pmatrix} = \begin{pmatrix} \tilde{\psi}_{p,1}^1 & \dots & \tilde{\psi}_{p,d_2}^1 \\ \vdots & \ddots & \vdots \\ \tilde{\psi}_{p,1}^{d_2} & \dots & \tilde{\psi}_{p,d_2}^{d_2} \end{pmatrix} \in \mathbb{R}^{d_2 d_1 \times d_2} \tag{3}$$

and

$$\underline{\Phi} = \begin{pmatrix} \underline{\Phi^1} \\ \vdots \\ \underline{\Phi^{d_2}} \end{pmatrix} = \begin{pmatrix} \phi_1^1 & \dots & \phi_{d_2}^1 \\ \vdots & \ddots & \vdots \\ \phi_1^{d_2} & \dots & \phi_{d_2}^{d_2} \end{pmatrix}, \underline{\tilde{\Phi}} = \begin{pmatrix} \underline{\tilde{\Phi}^1} \\ \vdots \\ \underline{\tilde{\Phi}^{d_2}} \end{pmatrix} = \begin{pmatrix} \tilde{\phi}_1^1 & \dots & \tilde{\phi}_{d_2}^1 \\ \vdots & \ddots & \vdots \\ \tilde{\phi}_1^{d_2} & \dots & \tilde{\phi}_{d_2}^{d_2} \end{pmatrix} \in \mathbb{R}^{d_2 d_1 \times d_2} \tag{4}$$

with $\phi_{i_2}^{i_1}[i_3] := \phi[i_3, i_2 - i_1], \psi_{p,i_2}^{i_1}[l] := \psi_p[i_3, i_2 - i_1], \tilde{\psi}_{p,i_2}^{i_1}[i_3] := \tilde{\psi}_p[i_3, i_2 - i_1]$ and $\phi_{i_2}^{i_1}, \psi_{p,i_2}^{i_1}, \tilde{\psi}_{p,i_2}^{i_1} \in \mathbb{R}^{d_1}, i_1 = 1, \dots, d_2$.

– Local bases are defined as: $\xi_i, \tilde{\xi}_i \in \mathbb{R}^{n_1}$,

$$\underline{\Xi} = \begin{pmatrix} \xi_1 & \dots & \xi_d \end{pmatrix}, \underline{\tilde{\Xi}} = \begin{pmatrix} \tilde{\xi}_1 & \dots & \tilde{\xi}_d \end{pmatrix} \in \mathbb{R}^{n_1 \times d}, \tag{5}$$

- Filter banks and their family of matrices for Unet-VAE:

$$\underline{\underline{\theta}} = \left\{\underline{\underline{\theta_1}}, \dots, \underline{\underline{\theta_Q}}\right\} \in \mathbb{R}^{d_1 \times d_2 \times P \times Q}, \underline{\underline{\theta_q}} = \left\{\underline{\theta_{q,1}}, \dots, \underline{\theta_{q,P}}\right\} \in \mathbb{R}^{d_1 \times d_2 \times P}, \ \underline{\theta_{q,p}} \in \mathbb{R}^{d_1 \times d_2},$$

whose matrix form is:

$$\underline{\underline{\Theta}} = \left\{\underline{\Theta_1}, \dots, \underline{\Theta_Q}\right\} \in \mathbb{R}^{d_1 d_2 P \times d_2 \times Q}, \underline{\Theta_q} = \begin{pmatrix} \underline{\Theta_{q,1}} \\ \vdots \\ \underline{\Theta_{q,P}} \end{pmatrix} \in \mathbb{R}^{d_1 d_2 P \times d_2}, \tag{6}$$

$$\underline{\Theta_{q,p}} = \begin{pmatrix} \underline{\Theta_{q,p}^1} \\ \vdots \\ \underline{\Theta_{q,p}^{d_2}} \end{pmatrix} = \begin{pmatrix} \theta_{q,p,1}^1 & \dots & \theta_{q,p,d_2}^1 \\ \vdots & \ddots & \vdots \\ \theta_{q,p,1}^{n_2} & \dots & \theta_{q,p,d_2}^{d_2} \end{pmatrix} = \begin{pmatrix} \theta_{q,p,1} & \dots & \theta_{q,p,d_2} \end{pmatrix} \in \mathbb{R}^{d_1 d_2 \times d_2},$$

$\theta_{q,p,i_2}^{i_1}[i_3] := \theta_{q,p}[i_3, i_2 - i_1], \ i_1 = 1, \dots, d_2, \ \theta_{q,p,i_2}^{i_1} \in \mathbb{R}^{d_1}$.

- Convolution Operations: Given filter bank matrices $\left(\underline{\phi}, \underline{\underline{\psi}}, \underline{\underline{\theta}}\right)$ with their matrix forms $\left(\underline{\Phi}, \underline{\underline{\Psi}}, \underline{\underline{\Theta}}\right)$ in (Equations 3-6) respectively, we have convolution operations for a matrix $\underline{f} \in \mathbb{R}^{n_1 \times n_2}$ and a multi-band image $\underline{\underline{f}} = \{\underline{f_1}, \dots, \underline{f_P}\} \in \mathbb{R}^{|\Omega| \times P}$:

– **1D convolution:**

$$\mathfrak{C}_{\phi_{i_1}^{i_2}} : \mathbb{R}^{n_1} \to \mathbb{R}^{n_1}; \ \mathfrak{C}_{\phi_{i_1}^{i_2}}(f_{i_2}) = \left(\sum_{k_1=1}^{d_1} f_{i_2}[k_1] \check{\phi}_{i_1}^{i_2}[k_2 - k_1] \right)_{k_2=1}^{n_1} = \mathscr{H}_{d_1}(f_{i_2}) \phi_{i_1}^{i_2}, \tag{7}$$

– **2D convolution:**

$$\underline{\mathfrak{C}}_{\underline{\phi}} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times d_2} ; \ \underline{\mathfrak{C}}_{\underline{\phi}}(\underline{f}) = \left(\sum_{k_1=1}^{d_1} \sum_{k_2=1}^{d_2} f[k_1, k_2] \check{\phi}[r_1 - k_1, r_2 - k_2] \right)_{r_1=1,\dots,n_1}^{r_2=1,\dots,d_2}, \tag{8}$$

– **Matrix-family convolution:**

$$\underline{\underline{\mathfrak{C}}}_{\underline{\underline{\psi}}} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times d_2 \times P} ; \ \underline{\underline{\mathfrak{C}}}_{\underline{\underline{\psi}}}(\underline{f}) = \left\{ \underline{\mathfrak{C}}_{\underline{\psi}_1}(\underline{f}), \dots, \underline{\mathfrak{C}}_{\underline{\psi}_P}(\underline{f}) \right\} = \mathscr{H}_{n_1|d_2}(\underline{f}) \underline{\underline{\Psi}}, \tag{9}$$

– **Anisotropic matrix-family convolution:**

$$\underline{\underline{\mathfrak{C}}}^{\text{ani}}_{\underline{\underline{\psi}}} : \mathbb{R}^{n_1 \times n_2 \times P} \to \mathbb{R}^{n_1 \times d_2 \times P} ;$$

$$\underline{\underline{\mathfrak{C}}}^{\text{ani}}_{\underline{\underline{\psi}}} \left(\underline{\underline{f}} \right) = \left\{ \underline{\mathfrak{C}}_{\underline{\psi}_1}(\underline{f}_1), \dots, \underline{\mathfrak{C}}_{\underline{\psi}_P}(\underline{f}_P) \right\} = \left\{ \mathscr{H}_{n_1|d_2}(\underline{f}_1) \underline{\Psi}_1, \dots, \mathscr{H}_{n_1|d_2}(\underline{f}_P) \underline{\Psi}_P \right\} .$$

– **Isotropic matrix-family convolution:**

$$\underline{\underline{\underline{\mathfrak{C}}}}^{\text{iso}}_{\underline{\underline{\underline{\theta}}}} : \mathbb{R}^{n_1 \times n_2 \times P} \to \mathbb{R}^{n_1 \times d_2 \times Q}, \underline{\underline{\tilde{f}}} = \left\{ \underline{\tilde{f}}_1, \dots, \underline{\tilde{f}}_Q \right\} = \underline{\underline{\underline{\mathfrak{C}}}}^{\text{iso}}_{\underline{\underline{\underline{\theta}}}} \left(\underline{\underline{f}} \right) = \mathscr{H}_{n_1|d_2|P} \left(\underline{\underline{f}} \right) \underline{\underline{\Theta}}$$

with:

$$\underline{\tilde{f}_q} = \sum_{p=1}^{P} \underline{\mathfrak{C}}_{\underline{\theta_{q,p}}}(\underline{f_p}) = \sum_{p=1}^{P} \mathscr{H}_{n_1|d_2}(\underline{f_p}) \underline{\Theta_{q,p}} = \mathscr{H}_{n_1|d_2|P} \left(\underline{\underline{f}} \right) \underline{\Theta_q} .$$

Then, we have the following relation for 2D convolution operation:

Proposition 4.1. *A 2D convolution is defined by Hankel matrix by describing it via a 1D convolution:*

$$\underline{\mathfrak{C}}_{\underline{\phi}}(\underline{f})[k_1, k_2] = \sum_{i=1}^{d_2} \mathfrak{C}_{\phi^i_{k_2}}(f_i)[k_1] = \sum_{i=1}^{d_2} \left(\mathscr{H}_{d_1}(f_i) \phi^i_{k_2} \right)[k_1], k_1 = 1, \dots, n_1, k_2 = 1, \dots, d_2$$

which is written in a matrix form as:

$$\underline{\mathfrak{C}}_{\underline{\phi}} \left(\underline{f} \right) = \mathscr{H}_{n_1|d_2} \left(\underline{f} \right) \underline{\Phi} . \tag{10}$$

Proof. We provide a proof of Proposition 4.1 in Section 5.9. □

Moreover, note that we have an adjoint operator $\underline{\underline{\mathfrak{C}}}^*_{\underline{\underline{\psi}}} = \sum_{p=1}^{P} \mathfrak{C}^*_{\psi_p} : \mathbb{R}^{n_1 \times d_2 \times P} \to \mathbb{R}^{n_1 \times n_2}$. And, for $d_1 = n_1, d_2 = n_2$, we have $\underline{\mathfrak{C}}_{\underline{\phi}}(\underline{f}) = \mathcal{F}^{-1} \left(\underline{\widehat{P}} \odot \underline{\widehat{F}} \right)$, $\underline{\widehat{P}} = \left[\widehat{P} \left(e^{j\omega} \right) \right]_{\omega \in [-\pi, \pi]^d}$ and $\underline{\widehat{F}} = \left[\widehat{F} \left(e^{j\omega} \right) \right]_{\omega \in [-\pi, \pi]^d}$. Its adjoint operator $\underline{\mathfrak{C}}^*_{\underline{\phi}} = \underline{\mathfrak{C}}_{\underline{\check{\phi}}} : \mathbb{R}^{n \times m} \to \mathbb{R}^{n \times m}$ is defined with a discrete time-reversed kernel

$$\check{\phi}(x_1, x_2) = \phi(-x_1, -x_2) \overset{\mathcal{F}}{\longleftrightarrow} \widehat{\check{\phi}}(\omega_1, \omega_2) = \widehat{\phi}^*(\omega_1, \omega_2)$$

whose discretized version is:

$$\check{\phi}[k_1, k_2] = \phi[-k_1, -k_2] \overset{\mathcal{F}}{\longleftrightarrow} \widehat{P}^* \left(e^{j\omega_1}, e^{j\omega_2} \right) = \widehat{P} \left(e^{-j\omega_1}, e^{-j\omega_2} \right) .$$

4.2 Proximal Operators and Moreau-Yosida envelope

Given a non-smooth function $\mathscr{P}(\cdot)$ which can be convex or non-convex, its Moreau-Yosida envelope ($\frac{1}{\mu} > 0$-Lipschitz differentiable) is

$$\mathscr{P}_\mu(\cdot) := \inf_{\underline{u}\in\mathbb{R}^{n_1\times n_2}} \left\{ \mathscr{P}(\underline{u}) + \frac{1}{2\mu}\|\underline{u} - \cdot\|_{\mathrm{F}}^2 \right\}, \tag{11}$$

with $\lim_{\mu\to 0} \mathscr{P}_\mu(\theta) = \mathscr{P}(\theta)$ and $\|\cdot\|_{\mathrm{F}}$ is Frobenius norm. Its gradient

$$\nabla\mathscr{P}_\mu(\cdot) = \frac{1}{\mu}\left(\cdot - \mathrm{prox}_{\mu\mathscr{P}}(\cdot)\right) \,:\, \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2} \tag{12}$$

is defined by a proximity operator

$$\mathrm{prox}_{\mu\mathscr{P}}(\cdot) := \operatorname*{argmin}_{\underline{u}\in\mathbb{R}^{n_1\times n_2}} \left\{ \mathscr{P}(\underline{u}) + \frac{1}{2\mu}\|\underline{u} - \cdot\|_{\mathrm{F}}^2 \right\}. \tag{13}$$

We also note that some activation functions in neural network can be well written as a proximal operator associated with its abstract function, e.g. Rectified Linear Unit (ReLU) activation function:

$$\begin{aligned}\mathrm{prox}_{\mathrm{ReLU}}\left(\underline{f}\right) &= \operatorname*{argmin}_{\underline{u}\in\mathbb{R}^{n_1\times n_2}} \left\{ \mathscr{P}_{\mathrm{ReLU}}(\underline{u}) + \frac{1}{2}\left\|\underline{u} - \underline{f}\right\|_{\mathrm{F}}^2 \right\} \\ &= \left[\max\left(0\,, f_{i_1,i_2}\right)\right]_{i_1,i_2=0}^{n_1-1,n_2-1}, \ \underline{f} = \left[f_{i_1,i_2}\right]_{i_1,i_2=0}^{n_1-1,n_2-1},\end{aligned}$$

where $\mathscr{P}_{\mathrm{ReLU}}(\cdot)$ is a non-defined function.

4.3 Kullback Leibler divergence:

Given distributions $\mathtt{F}(\mathrm{d}z) = \mathtt{f}(z)\mathrm{d}z$ and $\mathtt{G}(\mathrm{d}z) = \mathtt{g}(z)\mathrm{d}z$ on a domain $\mathbb{R}^d$, their KL-divergence is:

$$\begin{aligned}\mathrm{KL}\left(\mathtt{F} \,||\, \mathtt{G}\right) &= \mathbb{E}_{Z\sim\mathtt{F}}\left[\log\frac{\mathtt{f}(Z)}{\mathtt{g}(Z)}\right] = \int_{\mathbb{R}^d} \mathtt{f}(z)\log\frac{\mathtt{f}(z)}{\mathtt{g}(z)}\mathrm{d}z \\ &\geq 0\,.\end{aligned} \tag{14}$$

5 Proofs

5.1 Proof of Proposition II.1

Due to the unity condition (Equation 1) and $\mathscr{H}^{\dagger}_{d_1|n_2} \circ \mathscr{H}_{d_1|n_2} = \mathrm{Id}$, convolutional framelet decomposition is:

$$\underline{f} = \mathscr{H}^{\dagger}_{d_1|n_2}\left(\underline{\tilde{\Xi}\Xi}^{\mathrm{T}} \mathscr{H}_{d_1|n_2}\left(\underline{f}\right) \underline{\Phi\tilde{\Phi}}^{\mathrm{T}}\right)$$

where:

$$\underline{c_f} = \underline{\Xi}^{\mathrm{T}} \mathscr{H}_{d_1|n_2}\left(\underline{f}\right) \underline{\Phi} = \begin{pmatrix} \xi_1^{\mathrm{T}} \\ \vdots \\ \xi_d^{\mathrm{T}} \end{pmatrix} \mathscr{H}_{d_1|n_2}\left(\underline{f}\right) \begin{pmatrix} \phi_1 & \dots & \phi_{d_2} \end{pmatrix} = \begin{pmatrix} \xi_1^{\mathrm{T}} \mathscr{H}_{d_1|n_2}\left(\underline{f}\right) \phi_1 & \dots & \xi_1^{\mathrm{T}} \mathscr{H}_{d_1|n_2}\left(\underline{f}\right) \phi_{d_2} \\ \vdots & \ddots & \vdots \\ \xi_d^{\mathrm{T}} \mathscr{H}_{d_1|n_2}\left(\underline{f}\right) \phi_1 & \dots & \xi_d^{\mathrm{T}} \mathscr{H}_{d_1|n_2}\left(\underline{f}\right) \phi_{d_2} \end{pmatrix}$$

$$= \begin{pmatrix} c_{f,1} & \dots & c_{f,d_2} \end{pmatrix} = \left[c_f^{l,s}\right]_{l=1,\dots,d}^{s=1,\dots,d_2}$$

where:

$$c_f^{l,s} = \xi_l^{\mathrm{T}} \mathscr{H}_{d_1|n_2}\left(\underline{f}\right) \phi_s = \xi_l^{\mathrm{T}} \begin{pmatrix} \mathscr{H}_{d_1}(f_1) & \dots & \mathscr{H}_{d_1}(f_{n_2}) \end{pmatrix} \begin{pmatrix} \phi_s^1 \\ \vdots \\ \phi_s^{n_2} \end{pmatrix} = \sum_{i=1}^{n_2} \xi_l^{\mathrm{T}} \mathscr{H}_{d_1}(f_i) \phi_s^i$$

$$= \sum_{i=1}^{n_2} \left\langle f_i \,, \mathfrak{C}_{\phi_s^i}(\xi_l) \right\rangle_{\ell_2} .$$

The last equality is due to $u^{\mathrm{T}} \mathscr{H}_{d_1}(a) v = u^{\mathrm{T}} \mathfrak{C}_v(a) = \langle a \,, \mathfrak{C}_v(u) \rangle_{\ell_2}$ for $u \,, v \,, a \in \mathbb{R}^d$.

Now, we expand $\underline{f}$:

$$\underline{f} = \mathscr{H}^{\dagger}_{d_1|n_2}\left(\underline{\tilde{\Xi} c_f \tilde{\Phi}^{\mathrm{T}}}\right) = \left(\mathscr{H}^{\dagger}_{d_1}\left(\underline{\tilde{\Xi} c_f \tilde{\Phi}^{1,\mathrm{T}}}\right) \quad \dots \quad \mathscr{H}^{\dagger}_{d_1}\left(\underline{\tilde{\Xi} c_f \tilde{\Phi}^{n_2,\mathrm{T}}}\right)\right)$$

$$= \left(\mathscr{H}^{\dagger}_{d_1}\left(\sum_{s=1}^{d_2} \sum_{l=1}^{d} c_f^{l,s} \tilde{\xi}_l \tilde{\phi}_s^{1,\mathrm{T}}\right) \quad \dots \quad \mathscr{H}^{\dagger}_{d_1}\left(\sum_{s=1}^{d_2} \sum_{l=1}^{d} c_f^{l,s} \tilde{\xi}_l \tilde{\phi}_s^{n_2,\mathrm{T}}\right)\right)$$

$$= \left(\frac{1}{d_1} \sum_{s=1}^{d_2} \sum_{l=1}^{d} c_f^{l,s} \mathfrak{C}_{\tilde{\phi}_s^1}\left(\tilde{\xi}_l\right) \quad \dots \quad \frac{1}{d_1} \sum_{s=1}^{d_2} \sum_{l=1}^{d} c_f^{l,s} \mathfrak{C}_{\tilde{\phi}_s^{n_2}}\left(\tilde{\xi}_l\right)\right)$$

$$= \frac{1}{d_1} \sum_{s=1}^{d_2} \left(\mathfrak{C}_{\tilde{\phi}_s^1}\left(\underline{\tilde{\Xi}} c_{f,s}\right) \quad \dots \quad \mathfrak{C}_{\tilde{\phi}_s^{n_2}}\left(\underline{\tilde{\Xi}} c_{f,s}\right)\right) .$$

This concludes the proof. Note that the 3rd equality is due to:

$$\underline{\tilde{\Xi} c_f \tilde{\Phi}^{i,\mathrm{T}}} = \begin{pmatrix} \tilde{\xi}_1 & \dots & \tilde{\xi}_m \end{pmatrix} \begin{pmatrix} c_f^{1,1} & \dots & c_f^{1,d_2} \\ \vdots & \ddots & \vdots \\ c_f^{d,1} & \dots & c_f^{d,d_2} \end{pmatrix} \begin{pmatrix} \tilde{\phi}_{1,\mathrm{T}}^i \\ \vdots \\ \tilde{\phi}_{d_2}^{i,\mathrm{T}} \end{pmatrix} = \sum_{s=1}^{d_2} \sum_{l=1}^{d} \tilde{\xi}_l c_f^{l,s} \tilde{\phi}_s^{i,\mathrm{T}} ;$$

and, the 4th equality is due to:

$$\mathscr{H}_{d_1}^{\dagger}\left(\sum_{s=1}^{d_2}\sum_{l=1}^{d} c_f^{l,s}\tilde{\xi}_l\tilde{\phi}_s^{n_2,\mathrm{T}}\right) = \frac{1}{\sqrt{d_1}}\sum_{s=1}^{d_2}\sum_{l=1}^{d} c_f^{l,s}\begin{pmatrix} \left\langle \frac{1}{\sqrt{d_1}}\mathscr{H}_{d_1}\left(\underline{e_1}\right), \tilde{\xi}_l\tilde{\phi}_s^{n_2,\mathrm{T}}\right\rangle_{\mathrm{F}} \\ \vdots \\ \left\langle \frac{1}{\sqrt{d_1}}\mathscr{H}_{d_1}\left(\underline{e_{n_1}}\right), \tilde{\xi}_l\tilde{\phi}_s^{n_2,\mathrm{T}}\right\rangle_{\mathrm{F}} \end{pmatrix}$$

$$= \frac{1}{d_1}\sum_{s=1}^{d_2}\sum_{l=1}^{d} c_f^{l,s}\begin{pmatrix} \left\langle \underline{e_1}, \mathfrak{C}_{\tilde{\phi}_s^{n_2}}\left(\tilde{\xi}_l\right)\right\rangle_{\mathrm{F}} \\ \vdots \\ \left\langle \underline{e_{n_1}}, \mathfrak{C}_{\tilde{\phi}_s^{n_2}}\left(\tilde{\xi}_l\right)\right\rangle_{\mathrm{F}} \end{pmatrix} = \frac{1}{d_1}\sum_{s=1}^{d_2}\sum_{l=1}^{d} c_f^{l,s}\mathfrak{C}_{\tilde{\phi}_s^{n_2}}\left(\tilde{\xi}_l\right).$$

5.2 Proof of Proposition II.2

Given a wavelet expansion acting on a discrete function $f \in \ell_2\left(\mathbb{Z}^2\right)$:

$$f[k] = \sum_{m\in\mathbb{Z}^2}\left\langle f, \tilde{\varphi}_I(\cdot - m)\right\rangle_{\ell_2}\varphi_I(k-m) + \sum_{i=0}^{I}\sum_{l=0}^{L}\sum_{m\in\mathbb{Z}^2}\left\langle f, \tilde{\psi}_{il}(\cdot - m)\right\rangle_{\ell_2}\psi_{il}(k-m),$$

we compute its scaling and wavelet coefficients in the Fourier domain:

$$c_I[m] = \left\langle f, \tilde{\varphi}_I\left(\cdot - m\right)\right\rangle_{\ell_2} \overset{\mathcal{F}}{\longleftrightarrow} \widehat{C}_I\left(e^{j\omega}\right) = \sum_{m\in\mathbb{Z}^2} c_I[m]e^{-j\langle m,\omega\rangle_{\ell_2}} = \sum_{k\in\mathbb{Z}^2} f[k]\sum_{m\in\mathbb{Z}^2}\check{\tilde{\varphi}}_I\left(m-k\right)e^{-j\langle m,\omega\rangle_{\ell_2}}. \tag{15}$$

To compute a Fourier transform of $\check{\tilde{\varphi}}_I\left(m-k\right)$, we find a Fourier transform of its continuous version for $x \in \mathbb{R}^d$:

$$\check{\tilde{\varphi}}_I\left(x-k\right) \overset{\mathcal{F}}{\longleftrightarrow} e^{-j\langle k,\omega\rangle_{\ell_2}}\int_{\mathbb{R}^2}\check{\tilde{\varphi}}_I(x)e^{-j\langle x,\omega\rangle_{\ell_2}}\mathrm{d}x = \widehat{\tilde{\varphi}}_I^{*}(\omega)e^{-j\langle k,\omega\rangle_{\ell_2}}.$$

By Poisson summation formulae, we have the following identity:

$$\sum_{m\in\mathbb{Z}^d}\check{\tilde{\varphi}}_I\left(m-k\right)e^{-j\langle m,\omega\rangle_{\ell_2}} = \sum_{m\in\mathbb{Z}^d}\widehat{\tilde{\varphi}}_I^{*}\left(2\pi m+\omega\right)e^{-j\langle k,2\pi m+\omega\rangle_{\ell_2}}$$

Note $e^{j2\pi m} = 1, \forall m \in \mathbb{Z}^d$; then, scaling coefficient $c_I[m] \overset{\mathcal{F}}{\longleftrightarrow} \widehat{C}_I\left(e^{j\omega}\right)$ in (15) is

$$\widehat{C}_I\left(e^{j\omega}\right) = \sum_{k\in\mathbb{Z}^2} f[k]\sum_{m\in\mathbb{Z}^2}\widehat{\tilde{\varphi}}_I^{*}\left(2\pi m+\omega\right)e^{-j\langle k,2\pi m+\omega\rangle_{\ell_2}} = \sum_{m\in\mathbb{Z}^2}\widehat{\tilde{\varphi}}_I^{*}\left(2\pi m+\omega\right)\widehat{F}\left(e^{j(2\pi m+\omega)}\right)$$

$$= \widehat{F}\left(e^{j\omega}\right)\left[\widehat{\tilde{\varphi}}_I^{*}(\omega) + \sum_{m\in\mathbb{Z}^2\setminus\{0\}}\widehat{\tilde{\varphi}}_I^{*}\left(2\pi m+\omega\right)\right], \tag{16}$$

where $f[k] \overset{\mathcal{F}}{\longleftrightarrow} \widehat{F}\left(e^{j\omega}\right)$. Similarly, wavelet coefficient $d_{il}[m] \overset{\mathcal{F}}{\longleftrightarrow} \widehat{D}_{il}\left(e^{j\omega}\right)$ is:

$$\widehat{D}_{il}\left(e^{j\omega}\right) = \widehat{F}\left(e^{j\omega}\right)\left[\widehat{\tilde{\psi}}_{il}^{*}(\omega) + \sum_{m\in\mathbb{Z}^2\setminus\{0\}}\widehat{\tilde{\psi}}_{il}^{*}\left(2\pi m+\omega\right)\right]. \tag{17}$$

Now, we compute wavelet expansion in the Fourier domain as:

$$\widehat{F}\left(e^{j\omega}\right) = \sum_{k\in\mathbb{Z}^2} f[k]e^{-j\langle k\,,\omega\rangle_{\ell_2}}$$

$$= \sum_{m\in\mathbb{Z}^2} c_I[m] \sum_{k\in\mathbb{Z}^2} \varphi_I\left(k-m\right) e^{-j\langle k\,,\omega\rangle_{\ell_2}} + \sum_{i=0}^{I}\sum_{l=0}^{L}\sum_{m\in\mathbb{Z}^2} d_{il}[m] \sum_{k\in\mathbb{Z}^2} \psi_{il}\left(k-m\right) e^{-j\langle k\,,\omega\rangle_{\ell_2}}$$

$$= \sum_{k\in\mathbb{Z}^2} \widehat{\varphi}_I\left(2\pi k+\omega\right) \sum_{m\in\mathbb{Z}^2} c_I[m]e^{-j\langle m\,,2\pi k+\omega\rangle_{\ell_2}} + \sum_{i=0}^{I}\sum_{l=0}^{L}\sum_{k\in\mathbb{Z}^2} \widehat{\psi}_{il}\left(2\pi k+\omega\right) \sum_{m\in\mathbb{Z}^2} d_{il}[m]e^{-j\langle m\,,2\pi k+\omega\rangle_{\ell_2}}$$

$$= \widehat{C}_I\left(e^{j\omega}\right)\left(\widehat{\varphi}_I(\omega) + \sum_{k\in\mathbb{Z}^2\setminus\{0\}} \widehat{\varphi}_I\left(2\pi k+\omega\right)\right) + \sum_{i=0}^{I}\sum_{l=0}^{L} \widehat{D}_{il}\left(e^{j\omega}\right)\left(\widehat{\psi}_{il}(\omega) + \sum_{k\in\mathbb{Z}^2\setminus\{0\}} \widehat{\psi}_{il}\left(2\pi k+\omega\right)\right).$$

The 3rd equality is from Poisson summation formulae. We obtain the unity condition (Equation 6) from Equation 16 and Equation 17. To compensate the error in Equation 6, we modify a primal wavelet frame at scale 0 as:

$$1 = \widehat{\tilde{\varphi}}_I^*(\omega)\widehat{\varphi}_I(\omega) + \sum_{i=1}^{I}\sum_{l=0}^{L} \widehat{\tilde{\psi}}_{il}^*(\omega)\widehat{\psi}_{il}(\omega) + \sum_{l=0}^{L} \widehat{\tilde{\psi}}_{0l}^*(\omega)\widehat{\psi}_{0l}(\omega) + \widehat{e}(\omega)$$

$$\Leftrightarrow\ \widehat{\tilde{\psi}}_0(\omega) := \widehat{\psi}_0(\omega) + \frac{\widehat{e}(\omega)}{\widehat{\tilde{\psi}}_0^*(\omega)}\,.$$

This is due to a unity property of L-th Riesz transform $\sum_{l=0}^{L}\left|\widehat{\mathscr{R}^l}(\omega)\right|^2 = 1$. Finally, the unity condition (Equation 6) becomes:

$$\widehat{\tilde{\varphi}}_I^*(\omega)\widehat{\varphi}_I(\omega) + \widehat{\tilde{\psi}}_0^*(\omega)\widehat{\tilde{\psi}}_0(\omega) + \sum_{i=1}^{I}\sum_{l=0}^{L} \widehat{\tilde{\psi}}_{il}^*(\omega)\widehat{\psi}_{il}(\omega) = 1\,.$$

Then, multiply 2 sides with $F\left(e^{j\omega}\right)$ and take an inverse Fourier transform, we obtain a wavelet expansion:

$$\widehat{F}\left(e^{j\omega}\right) = \widehat{F}\left(e^{j\omega}\right)\widehat{\tilde{\varphi}}_I^*(\omega)\widehat{\varphi}_I(\omega) + \sum_{i=0}^{I}\sum_{l=0}^{L} \widehat{F}\left(e^{j\omega}\right)\widehat{\tilde{\psi}}_{il}^*(\omega)\widehat{\psi}_{il}(\omega)$$

$$\overset{\mathcal{F}}{\longleftrightarrow}\ f[k] = \sum_{m\in\mathbb{Z}^2} \left\langle f\,,\tilde{\varphi}_I(\cdot-m)\right\rangle_{\ell_2} \varphi_I(k-m) + \sum_{i=0}^{I}\sum_{l=0}^{L}\sum_{m\in\mathbb{Z}^2} \left\langle f\,,\tilde{\psi}_{il}(\cdot-m)\right\rangle_{\ell_2} \psi_{il}(k-m)$$

with $\widehat{\psi}_{0n}(\omega) = \widehat{\tilde{\psi}}_{0n}(\omega)$.

5.3 Proof of Proposition II.3

Denote $\underline{\psi_0} := \underline{\varphi_I}$ and $\underline{\tilde{\psi}_0} := \underline{\tilde{\varphi}_I}$ whose matrix forms are $\underline{\Psi_0} = \underline{\Phi_I}\,, \underline{\tilde{\Psi}_0} = \underline{\tilde{\Phi}_I} \in \mathbb{R}^{n_1n_2\times n_2}$ as in (3), respectively. Note that $\underline{\mathfrak{C}}_{\underline{\phi}} := \underline{\mathfrak{C}}^*_{\underline{\varphi_I}}\underline{\mathfrak{C}}_{\underline{\tilde{\varphi}_I}}$. A convolutional form of an expansion (**??**) is recast with

Hankel matrix as:

$$\underline{f} = \underline{\mathfrak{C}}^*_{\underline{\varphi}_I} \underline{\mathfrak{C}}_{\underline{\tilde{\varphi}}_I} \left(\underline{f}\right) + \sum_{p=1}^{P} \underline{\mathfrak{C}}^*_{\underline{\psi}_p} \underline{\mathfrak{C}}_{\underline{\tilde{\psi}}_p} \left(\underline{f}\right) = \sum_{p=0}^{P} \underline{\mathfrak{C}}^*_{\underline{\psi}_p} \underline{\mathfrak{C}}_{\underline{\tilde{\psi}}_p} \left(\underline{f}\right) = \sum_{p=0}^{P} \mathscr{H}_{n_1|n_2} \left(\underline{w_{f,p}}\right) \underline{\check{\Psi}_p}$$

$$= \sum_{p=0}^{P} \begin{pmatrix} \mathscr{H}_{n_1}\left(w_{f,p,1}\right) & \dots & \mathscr{H}_{n_1}\left(w_{f,p,n_2}\right) \end{pmatrix} \begin{pmatrix} \underline{\check{\Psi}^1_p} \\ \vdots \\ \underline{\check{\Psi}^{n_2}_p} \end{pmatrix} = \sum_{p=0}^{P} \sum_{i=1}^{n_2} \mathscr{H}_{n_1}\left(w_{f,p,i}\right) \underbrace{\begin{pmatrix} \check{\psi}^i_{p,1} & \dots & \check{\psi}^i_{p,d_2} \end{pmatrix}}_{=\underline{\check{\Psi}^i_p}}$$

$$= n_1 \sum_{p=0}^{P} \frac{1}{n_1} \sum_{i=1}^{n_2} \begin{pmatrix} \mathfrak{C}_{\check{\psi}^i_{p,1}}\left(\underline{\tilde{\Xi}} w_{f,p,i}\right) & \dots & \mathfrak{C}_{\check{\psi}^i_{p,n_2}}\left(\underline{\tilde{\Xi}} w_{f,p,i}\right) \end{pmatrix}$$

$$= n_1 \sum_{p=0}^{P} \mathscr{H}^{\dagger}_{n_1|n_2} \left(\underline{w_{f,p}}\, \underline{\check{\Psi}^{\mathrm{T}}_p}\right)$$

$$= n_1 \mathscr{H}^{\dagger}_{n_1|n_2} \left(\sum_{p=0}^{P} \underline{w_{f,p}}\, \underline{\check{\Psi}^{\mathrm{T}}_p}\right) = n_1 \mathscr{H}^{\dagger}_{n_1|n_2} \left(\underline{w_f}\, \underline{\check{\Psi}^{\mathrm{T}}}\right) ,$$

where wavelet coefficient is

$$\underline{w_{f,p}} := \underline{\mathfrak{C}}_{\underline{\tilde{\psi}}_p} \left(\underline{f}\right) = \mathscr{H}_{n_1|n_2} \left(\underline{f}\right) \underline{\tilde{\Psi}_p}$$

$$= \begin{pmatrix} w_{f,p,1} & \dots & w_{f,p,n_2} \end{pmatrix} \in \mathbb{R}^{n_1 \times n_2} ,\ p = 0, \dots, P$$

$$\Leftrightarrow\ \underline{w_f} := \begin{pmatrix} \underline{w_{f,0}} & \dots & \underline{w_{f,P}} \end{pmatrix} = \mathscr{H}_{n_1|n_2} \left(\underline{f}\right) \underline{\tilde{\Psi}} .$$

Note that the last 3rd equality is due to proposition **??** with an identity local basis $\underline{\tilde{\Xi}} = \mathrm{Id}$. The 2nd last equality is because $\mathscr{H}^{\dagger}_{n_1|n_2}$ is a linear operator: Given a matrix $\underline{g_p} = \begin{pmatrix} \underline{g_{p,1}} & \dots & \underline{g_{p,n_2}} \end{pmatrix}$ and constant $\{a_p\}_{p=0}^{P} \subset \mathbb{R}$, we have:

$$\sum_{p=0}^{P} a_p \mathscr{H}^{\dagger}_{n_1|n_2} \left(\underline{g_p}\right) = \begin{pmatrix} \sum_{p=0}^{P} a_p \mathscr{H}^{\dagger}_{n_1} \left(\underline{g_{p,1}}\right) & \dots & \sum_{p=0}^{P} a_p \mathscr{H}^{\dagger}_{n_1} \left(\underline{g_{p,n_2}}\right) \end{pmatrix} .$$

And, by a definition of an inverse Hankel matrix (2), we have:

$$\sum_{p=0}^{P} a_p \mathscr{H}^{\dagger}_{n_1} \left(\underline{g_{p,i}}\right) = \frac{1}{\sqrt{n_1}} \begin{pmatrix} \left\langle \underline{\tilde{e}_1}, \sum_{p=0}^{P} a_p \underline{g_{p,i}} \right\rangle_{\mathrm{F}} \\ \vdots \\ \left\langle \underline{\tilde{e}_{n_1}}, \sum_{p=0}^{P} a_p \underline{g_{p,i}} \right\rangle_{\mathrm{F}} \end{pmatrix} = \mathscr{H}^{\dagger}_{n_1} \left(\sum_{p=0}^{P} a_p \underline{g_{p,i}}\right) .$$

Thus, we have a linear property of an extended Hankel matrix:

$$\sum_{p=0}^{P} a_p \mathscr{H}^{\dagger}_{n_1|n_2}\left(\underline{g_p}\right) = \left(\mathscr{H}^{\dagger}_{n_1}\left(\sum_{p=0}^{P} a_p \underline{g_{p,1}}\right) \quad \dots \quad \mathscr{H}^{\dagger}_{n_1}\left(\sum_{p=0}^{P} a_p \underline{g_{p,n_2}}\right)\right)$$

$$\overset{(1)}{=} \mathscr{H}^{\dagger}_{n_1|n_2}\left(\left(\sum_{p=0}^{P} a_p \underline{g_{p,1}} \quad \dots \quad \sum_{p=0}^{P} a_p \underline{g_{p,n_2}}\right)\right)$$

$$= \mathscr{H}^{\dagger}_{n_1|n_2}\left(\sum_{p=0}^{P} a_p \underline{g_p}\right) .$$

In the end, we have the proposed non-subsampled Riesz-Quincunx wavelet is a framelet decomposition with an identity local basis:

$$\underline{f} = n_1 \mathscr{H}^{\dagger}_{n_1|n_2}\left(\mathscr{H}_{n_1|n_2}\left(\underline{f}\right)\underline{\tilde{\Psi}}\,\underline{\check{\Psi}}^{\mathrm{T}}\right) .$$

Since $\mathscr{H}^{\dagger}_{n_1|n_2} \circ \mathscr{H}_{n_1|n_2} = \mathrm{Id}$, this implies the unity condition:

$$\frac{1}{n_1}\mathrm{Id}_{n_1 n_2 \times n_1 n_2} = \underline{\tilde{\Psi}}\,\underline{\check{\Psi}}^{\mathrm{T}} = \underline{\tilde{\Phi}_I}\,\underline{\check{\Phi}_I}^{\mathrm{T}} + \sum_{p=1}^{P} \underline{\tilde{\Psi}_p}\,\underline{\check{\Psi}_p}^{\mathrm{T}} .$$

5.4 Proof of Proposition II.4

The mapping $\mathscr{T}^{(i)}$ in Equation 13 is defined with $\underline{\underline{\underline{\mathfrak{C}}}}^{\mathrm{iso}}_{\underline{\underline{\underline{\theta^{1(i)}}}}} : \mathbb{R}^{n_1 \times n_2 \times P} \to \mathbb{R}^{n_1 \times d_2 \times 2^i L}$ and $\underline{\underline{\underline{\mathfrak{C}}}}^{\mathrm{iso}}_{\underline{\underline{\underline{\theta^{2(i)}}}}} : \mathbb{R}^{n_1 \times d_2 \times 2^i L} \to \mathbb{R}^{n_1 \times n_2 \times 2^i L}$. Given a local basis $\underline{\Xi^{(i)}_{\mathrm{aug}}} = \left(\mathrm{Id} \quad \underline{\Xi^{(i)}}\right)$, an encoder is defined with a lowpass signal and skip-connecting signal as outputs:

$$\underline{\underline{c^{(i)}_{\mathrm{aug}}}} = \underline{\Xi^{(i)\mathrm{T}}_{\mathrm{aug}}}\mathscr{T}^{(i)}\left(\underline{\underline{s^{(i-1)}}}\right) = \begin{pmatrix} \mathscr{T}^{(i)}\left(\underline{\underline{s^{(i-1)}}}\right) \\ \underline{\Xi^{(i),\mathrm{T}}}\mathscr{T}^{(i)}\left(\underline{\underline{s^{(i-1)}}}\right) \end{pmatrix} = \begin{pmatrix} \underline{\underline{c^{(i)}}} \\ \underline{\underline{s^{(i)}}} \end{pmatrix} , \; i = 1, \dots, I , \; \underline{\underline{s^{(0)}}} = \underline{\underline{f}} , \; \underline{\underline{s^{(I)}}} = \mathscr{R}_p\left(\underline{\underline{s^{(I)}}}\right) \tag{18}$$

$$\Leftrightarrow \begin{cases} \underline{\underline{s^{(I)}}} = \mathscr{R}_p \circ \underline{\Xi^{(I),\mathrm{T}}}\mathscr{T}^{(I)} \circ \dots \circ \underline{\Xi^{(1),\mathrm{T}}}\mathscr{T}^{(1)}\left(\underline{\underline{f}}\right) , \\ \underline{\underline{c^{(i)}}} = \mathscr{T}^{(i)} \circ \underline{\Xi^{(i-1),\mathrm{T}}}\mathscr{T}^{(i-1)} \circ \dots \circ \underline{\Xi^{(1),\mathrm{T}}}\mathscr{T}^{(1)}\left(\underline{\underline{f}}\right) , \; i = 1, \dots, I , \end{cases} \tag{19}$$

where

$$\underline{\underline{c^{(i)}}} \in \mathbb{R}^{2^{-(i-1)} n_1 \times 2^{-(i-1)} n_2 \times 2^{(i-1)} L} , \; \underline{\underline{s^{(i)}}} \in \mathbb{R}^{2^{-(i-1)-1} n_1 \times 2^{-(i-1)-1} n_2 \times 2^{(i-1)} L} .$$

Each subband of the skip connection $\underline{\underline{c^{(i)}}} = \left\{\underline{c^{(i)}_0}, \dots, \underline{c^{(i)}_{2^i L-1}}\right\}, \underline{c^{(i)}_l} \in \mathbb{R}^{2^{-(i-1)} n_1 \times 2^{-(i-1)} n_2}$, is passed to the N-th order Riesz Quincunx wavelet expansion (Equation 11) at scale I': $i = 0, \dots, I-1, l = 0, \dots, 2^i L - 1$, $k \in \mathbb{Z}^2$:

$$\underline{c^{(i)}_l} = \underline{\mathfrak{C}}_{\underline{\phi}}\left(\underline{c^{(i)}_l}\right) + \sum_{p=1}^{P} \underline{\mathfrak{C}^*_{\psi_p}}\,\underline{\mathfrak{C}}_{\underline{\tilde{\psi}_p}}\left(\underline{c^{(i)}_l}\right) = \underline{\mathfrak{C}}_{\underline{\phi}}(\underline{c^{(i)}_l}) + \underline{\underline{\mathfrak{C}}}^*_{\underline{\psi}}\,\underline{\underline{\mathfrak{C}}}_{\underline{\tilde{\psi}}}(\underline{c^{(i)}_l})$$

$$= \left(\mathscr{H}_{2^{-(i-1)} n_1|2^{-(i-1)} n_2}(\underline{c^{(i)}_l})\underline{\Phi}\right) + \sum_{p=1}^{P}\left(\mathscr{H}_{2^{-(i-1)} n_1|2^{-(i-1)} n_2}\left(\mathscr{H}_{2^{-(i-1)} n_1|2^{-(i-1)} n_2}\left(\underline{c^{(i)}_l}\right)\underline{\tilde{\Psi}_p}\right)\underline{\check{\Psi}_p}\right)$$

$$\Leftrightarrow \underline{c^{(i)}_l} = \mathfrak{W}\,\mathfrak{W}^{-1}\left\{\underline{c^{(i)}_l}\right\} ,$$

which satisfies the condition of perfect reconstruction. Scaling and wavelet functions are defined in the previous section.

5.5 Proof of Proposition II.5

The proof is straight-forward by noting that a lowpass signal as an encoder's output in (19)

$$\underline{\underline{s^{(I)}}} = \mathscr{R}_p \circ \underline{\Xi^{(I),\mathrm{T}}} \mathscr{T}^{(I)} \circ \ldots \circ \underline{\Xi^{(1),\mathrm{T}}} \mathfrak{T}^{(1)} \left(\underline{\underline{f}} \right) .$$

And, unknown parameters are: $\gamma_m := \{\gamma_c \,, \underline{W^\mu} \,, b^\mu\}$ and $\gamma_s := \{\gamma_c \,, \underline{W^\sigma} \,, b^\sigma\}$.

5.6 Proof of Proposition II.6

Now, we define an "inverse" mapping of (Equation 15) as a linear perceptron network and a reshape operation mapping:

$$\mathscr{F}^s(\cdot) = \mathrm{uvec}\left(\underline{W^s} \cdot + b^s\right) \,:\, \mathbb{R}^d \to \mathbb{R}^{2^{-i-1} n_1 \times 2^{-i-1} n_2 \times 2^i L} \tag{20}$$

with $\underline{W^s} \in \mathbb{R}^{2^{-(I+1)} n_1 n_2 L \times d}$ and $b^s \in \mathbb{R}^{2^{-(I+1)} n_1 n_2 L}$.

Then, combining Equations 16 and 20, a reconstructed lowpass image generated from the latent variable y in Equation 16:

$$\underline{\underline{\widehat{s}^{(I)}}} = \mathscr{F}^s(y) = \mathrm{uvec}\left(\underline{W^s} \mathscr{M}_{\gamma_m} \left(\underline{\underline{f}} \right) + b^s \right) + \mathrm{uvec}\left(\left(\underline{W^s} \mathscr{S}_{\gamma_s}^{\frac{1}{2}} \left(\underline{\underline{f}} \right) \right) \odot \epsilon \right)$$

$$\in \mathbb{R}^{2^{-i-1} n_1 \times 2^{-i-1} n_2 \times 2^i L} \,,\; \epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d \left(\mathbf{0}_d \,, \mathrm{Id}_d \right) .$$

The last equality is because $\mathrm{uvec}(\cdot)$ is a linear operator.

Given an augmented local basis $\underline{\tilde{\Xi}^{(i)}_{\mathrm{aug}}} = \begin{pmatrix} \mathrm{Id}_N & \mathscr{B} \circ \underline{\tilde{\Xi}^{(i)}} \end{pmatrix}$, a decoder at scale $i \in \{I-1 \,, \ldots \,, 0\}$ is defined by an output encoder and a skip connection $\underline{\underline{\widehat{c}^{(i)}_{\mathrm{aug}}}} = \begin{pmatrix} \underline{\underline{c^{(i)}}} & \underline{\underline{\widehat{s}^{(i+1)}}} \end{pmatrix}$ as: $\underline{\underline{\widehat{s}^{(I)}}} = \mathfrak{F}^s(y)$,

$$\begin{cases} \underline{\underline{\widehat{c}^{(i-1)}_{\mathrm{aug}}}} = \tilde{\mathscr{T}}^{(i)} \left(\underline{\underline{\widehat{c}^{(i)}_{\mathrm{aug}}}} \right) = \begin{pmatrix} \underline{\underline{c^{(i-1)}}} & \underline{\underline{\widehat{s}^{(i)}}} \end{pmatrix} \,,\; i = I \,, \ldots \,, 1 \,, \\ \underline{\underline{\widehat{s}^{(0)}}} = \tilde{\mathscr{T}}^{(0)} \left(\underline{\underline{\widehat{c}^{(0)}_{\mathrm{aug}}}} \right) \,,\; i = 0 \,, \end{cases} \tag{21}$$

$$\Leftrightarrow \;\; \underline{\underline{\widehat{s}^{(0)}}} = \tilde{\mathscr{T}}^{(0)} \circ \tilde{\mathscr{T}}^{(1)} \circ \ldots \circ \tilde{\mathscr{T}}^{(I-1)} \left(\underline{\underline{c^{(I-1)}}} \quad \underline{\underline{\widehat{s}^{(I)}}} \right) := \tilde{\mathscr{T}}_I \left(\underline{\underline{c}} \,, \mathfrak{F}^s(y) \right) . \tag{22}$$

Then, we have Proposition II.6.

5.7 Proof of Proposition II.7

Since $\mathtt{p}\left(y \mid \underline{\underline{f}} \right) = \mathtt{k}_\alpha \left(y \mid \underline{\underline{f}} \right)$, a likelihood of the latent variable in an encoder (Equation 26) is computed by a marginal likelihood:

$$\mathtt{p}(z) = \int_{\mathcal{F}} \mathtt{p}\left(z \mid \underline{\underline{f}} \right) \mathtt{p}\left(\underline{\underline{f}} \right) \mathrm{d}\underline{\underline{f}} = \mathbb{E}_{\underline{\underline{f}} \sim \mathfrak{F}} \left[\mathtt{k}_\alpha \left(z \mid \underline{\underline{f}} \right) \right] .$$

Assume we have a bunch of realization of latent variable z as $\mathfrak{Z} = \{z_i\}_{i=1}^{n_z} \subset \mathbb{R}^d$; then, we have a maximum of an expected log-marginal-likelihood $\mathtt{p}(z)$ as:

$$\alpha^\dagger = \underset{\alpha}{\operatorname{argmax}}\, \mathbb{E}_{\underline{\underline{f}} \sim \mathfrak{F}} \left[\log \mathtt{k}_\alpha \left(\mathfrak{Z} \mid \underline{\underline{f}} \right) \right] = \underset{\alpha}{\operatorname{argmax}}\, \mathbb{E}_{\underline{\underline{f}} \sim \mathfrak{F}} \left[\frac{1}{n_y} \sum_{i=1}^{n_y} \log \mathtt{k}_\alpha \left(z_i \mid \underline{\underline{f}} \right) \right] \;,\; z_i \overset{\text{i.i.d.}}{\sim} \mathbb{K}_\alpha \left(z \mid \underline{\underline{f}} \right)$$

which is

$$\begin{aligned}
\left(\gamma_m^\dagger \,, \gamma_s^\dagger \,, \alpha^\dagger \right) &= \underset{\gamma_m, \gamma_s, \alpha}{\operatorname{argmax}}\, \mathbb{E}_{\underline{\underline{f}} \sim \mathfrak{F}} \left[\frac{1}{n_z} \sum_{i=1}^{n_z} \log \mathtt{k}_\alpha (z_i \mid \underline{\underline{f}}) \right] \;,\; z_i \overset{\text{i.i.d.}}{\sim} \mathbb{G}_{\gamma_m, \gamma_s}(z \mid \underline{\underline{f}}) = \mathtt{g}_{\gamma_m, \gamma_s} \left(z \mid \underline{\underline{f}} \right) \mathrm{d}z \\
&\approx \underset{\gamma_m, \gamma_s, \alpha}{\operatorname{argmax}}\, \mathbb{E}_{\underline{\underline{f}} \sim \mathfrak{F}} \left[\mathbb{E}_{z \sim \mathbb{G}_{\gamma_m, \gamma_s}(\cdot \mid \underline{\underline{f}})} \left[\log \mathtt{k}_\alpha \left(z \mid \underline{\underline{f}} \right) \right] \right] \\
&= \underset{\gamma_m, \gamma_s, \alpha}{\operatorname{argmax}} \left\{ \mathbb{E}_{\underline{\underline{f}} \sim \mathfrak{F}} \left[\mathbb{E}_{z \sim \mathbb{G}_{\gamma_m, \gamma_s} \left(\cdot \mid \underline{\underline{f}} \right)} \left[\log \mathtt{k}_\alpha (z \mid \underline{\underline{f}}) \right] - \mathbb{E}_{z \sim \mathbb{G}_{\gamma_m, \gamma_s}(\cdot \mid \underline{\underline{f}})} \left[\log \mathtt{g}_{\gamma_m, \gamma_s} (z \mid \underline{\underline{f}}) \right] \right] \right\} \\
&= \underset{\gamma_m, \gamma_s, \alpha}{\operatorname{argmax}} \left\{ -\mathbb{E}_{\underline{\underline{f}} \sim \mathfrak{F}} \left[\mathbb{E}_{z \sim \mathbb{G}_{\gamma_m, \gamma_s} \left(\cdot \mid \underline{\underline{f}} \right)} \left[\log \frac{\mathtt{g}_{\gamma_m, \gamma_s} \left(z \mid \underline{\underline{f}} \right)}{\mathtt{k}_\alpha \left(z \mid \underline{\underline{f}} \right)} \right] \right] \right\} \\
&= \underset{\gamma_m, \gamma_s, \alpha}{\operatorname{argmin}}\, \mathbb{E}_{\underline{\underline{f}} \sim \mathfrak{F}} \left[\mathrm{KL} \left(\mathbb{G}_{\gamma_m, \gamma_s} \left(z \mid \underline{\underline{f}} \right) \,||\, \mathbb{K}_\alpha \left(z \mid \underline{\underline{f}} \right) \right) \right] \,. \qquad (23)
\end{aligned}$$

In the end, we need to find $(\gamma_m \,, \gamma_s \,, \theta)$ that minimize the KL-distance between the true distribution and its approximated version in Equation 27 over all data $\mathfrak{F}$ as:

$$\min_{\gamma_m, \gamma_s, \alpha} \mathbb{E}_{\underline{\underline{f}} \sim \mathfrak{F}} \left[\mathrm{KL} \left(\mathbb{G}_{\gamma_m, \gamma_s}(z \mid \underline{\underline{f}}) \,||\, \mathbb{K}_\alpha(z \mid \underline{\underline{f}}) \right) \right] \,. \qquad (24)$$

From Bayes' rule (Equation 27), we derive an evidence lower bound of the KL-divergence in (24):

$$\begin{aligned}
&\mathrm{KL} \left(\mathbb{G}_{\gamma_m, \gamma_s}(z \mid \underline{\underline{f}}) \,||\, \mathbb{K}_\alpha(z \mid \underline{\underline{f}}) \right) \\
=&\mathbb{E}_{z \sim \mathbb{G}_{\gamma_m, \gamma_s}(\cdot \mid \underline{\underline{f}})} \left[\log \frac{\mathtt{g}_{\gamma_m, \gamma_s}(z \mid \underline{\underline{f}})}{\mathtt{k}_\alpha(z \mid \underline{\underline{f}})} \right] \overset{(27)}{=} \mathbb{E}_{z \sim \mathbb{G}_{\gamma_m, \gamma_s}(\cdot \mid \underline{\underline{f}})} \left[\log \frac{\mathtt{p}(\underline{\underline{f}}) \mathtt{g}_{\gamma_m, \gamma_s}(z \mid \underline{\underline{f}})}{\mathtt{h}_\alpha \left(\underline{\underline{f}} \mid z \right) \mathtt{p}(z)} \right] \\
=& - \mathbb{E}_{z \sim \mathbb{G}_{\gamma_m, \gamma_s} \left(\cdot \mid \underline{\underline{f}} \right)} \left[\log \mathtt{h}_\alpha \left(\underline{\underline{f}} \mid z \right) \right] + \mathbb{E}_{z \sim \mathbb{G}_{\gamma_m, \gamma_s} \left(\cdot \mid \underline{\underline{f}} \right)} \left[\log \frac{\mathtt{g}_{\gamma_m, \gamma_s} \left(z \mid \underline{\underline{f}} \right)}{\mathtt{p}(z)} \right] + \log \mathtt{p} \left(\underline{\underline{f}} \right)
\end{aligned}$$

which is equivalent to:

$$\begin{aligned}
\mathcal{L} \left(\mathbb{G}_{\gamma_m, \gamma_s} \,, \alpha \,; \underline{\underline{f}} \right) &:= \mathbb{E}_{z \sim \mathbb{G}_{\gamma_m, \gamma_s} \left(\cdot \mid \underline{\underline{f}} \right)} \left[\log \mathtt{h}_\alpha \left(\underline{\underline{f}} \mid z \right) \right] - \mathrm{KL} \left[\mathbb{G}_{\gamma_m, \gamma_s} \left(z \mid \underline{\underline{f}} \right) \,||\, \mathbb{P}(z) \right] \\
&= \log \mathtt{p} \left(\underline{\underline{f}} \right) - \mathrm{KL} \left(\mathbb{G}_{\gamma_m, \gamma_s} \left(z \mid \underline{\underline{f}} \right) \,||\, \mathbb{K}_\alpha \left(z \mid \underline{\underline{f}} \right) \right) \\
&\leq \log \mathtt{p} \left(\underline{\underline{f}} \right) \,.
\end{aligned}$$

The last inequality is due to $\mathrm{KL}(\cdot \,||\, \cdot) \geq 0$. Note that, via a reparamization trick from Equation 20

$$z \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d\left(\mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right), \mathrm{diag}\left\{\mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}\right)\right\}\right) = \mathbb{G}_{\gamma_m,\gamma_s}\left(z \mid \underline{\underline{f}}\right),$$

$$\Leftrightarrow \; z = \mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right) + \mathscr{S}^{\frac{1}{2}}_{\gamma_s}\left(\underline{\underline{f}}\right) \odot \epsilon, \; \epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{Id})$$

an evidence bound is recast as:

$$\mathcal{L}\left(\mathbb{G}_{\gamma_m,\gamma_s}, \alpha; \underline{\underline{f}}\right) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathrm{Id})}\left[\log \mathtt{h}_\alpha\left(\underline{\underline{f}} \mid z = \mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right) + \mathscr{S}^{\frac{1}{2}}_{\gamma_s}\left(\underline{\underline{f}}\right) \odot \epsilon\right)\right] - \mathrm{KL}\left[\mathbb{G}_{\gamma_m,\gamma_s}\left(z \mid \underline{\underline{f}}\right) \,||\, \mathbb{P}(z)\right].$$

Then, a minimization (24) is equivalent to maximize an expected evidence bound over all observation $\mathfrak{F}$:

$$\begin{aligned}
\left(\gamma_m^\dagger, \gamma_s^\dagger, \alpha^\dagger\right) &= \underset{\gamma_m,\gamma_s,\alpha}{\mathrm{argmax}}\, \mathbb{E}_{\underline{\underline{f}}\sim\mathfrak{F}}\left[\log \mathtt{p}\left(\underline{\underline{f}}\right) - \mathrm{KL}\left(\mathbb{G}_{\gamma_m,\gamma_s}\left(z \mid \underline{\underline{f}}\right) \,||\, \mathbb{K}_\alpha\left(z \mid \underline{\underline{f}}\right)\right)\right] \\
&= \underset{\gamma_m,\gamma_s,\alpha}{\mathrm{argmax}}\, \mathbb{E}_{\underline{\underline{f}}\sim\mathfrak{F}}\left[\mathcal{L}\left(\mathbb{G}_{\gamma_m,\gamma_s}, \alpha; \underline{\underline{f}}\right)\right] \\
&= \underset{\gamma_m,\gamma_s,\alpha}{\mathrm{argmax}}\, \frac{1}{T}\sum_{i=1}^{T} \mathcal{L}\left(\mathbb{G}_{\gamma_m,\gamma_s}, \alpha; \underline{\underline{f_i}}\right). \qquad (25)
\end{aligned}$$

This is because $\mathtt{p}\left(\underline{\underline{f}}\right)$ is independent to $(\gamma_m, \gamma_s, \alpha)$.

For choosing $\mathbb{P}(z) = \mathcal{N}_d(\mathbf{0}_d, \mathrm{Id}_d)$ and $\mathbb{G}_{\gamma_m,\gamma_s}\left(z \mid \underline{\underline{f}}\right) = \mathcal{N}_d\left(\mathscr{M}_{\gamma_m}\left(\underline{\underline{f}}\right), \mathrm{diag}\left\{\mathscr{S}_{\gamma_s}\left(\underline{\underline{f}}\right)\right\}\right)$, the KL-divergence term is defined above. From Equation 25, we have a likelihood:

$$\mathtt{h}_\alpha\left(\underline{\underline{f}} \mid z\right) = \mathcal{N}_D\left(\underline{\underline{f}}; \mathscr{D}_\alpha\left(\underline{\underline{c}}, z\right), \sigma^2\mathrm{Id}\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left\|\underline{\underline{f}} - \mathscr{D}_\alpha\left(\underline{\underline{c}}, z\right)\right\|^2_{\ell_2}\right). \qquad (26)$$

Then, a minimization (25) becomes the following minimization problem:

$$\left(\gamma_m^\dagger, \gamma_s^\dagger, \alpha^\dagger\right) = \underset{\gamma_m,\gamma_s,\alpha}{\mathrm{argmax}}\left\{\mathscr{L}(\gamma_m, \gamma_s, \alpha) = \sum_{i=1}^{T} \mathcal{L}\left(\mathbb{G}_{\gamma_m,\gamma_s}, \alpha; \underline{\underline{f_i}}\right)\right\}. \qquad (27)$$

5.8 Proof of Proposition II.9

We firstly define a vector-valued function for K-clusters:

$$u : \mathbb{R}^P \to \mathbb{R}^K; \; u\left(f_{i,l}\right) = \begin{pmatrix} u_1\left(f_{i,l}\right) \\ \vdots \\ u_K\left(f_{i,l}\right) \end{pmatrix}, \; u_k\left(f_{i,l}\right) \in \mathbb{R},$$

whose definition is defined in a matrix form by our proposed RQUnet-VAE as:

$$
\begin{aligned}
u\left(\underline{\underline{f_i}}\right) &:= \left[u\left(f_{i,l}\right)\right]_{l\in\Omega} \\
&= \left\{u_1\left(\underline{\underline{f_i}}\right), \ldots, u_K\left(\underline{\underline{f_i}}\right)\right\} \in \mathbb{R}^{n_1\times n_2\times K},\ u_k\left(\underline{\underline{f_i}}\right) \in \mathbb{R}^{n_1\times n_2} \\
&= \underline{\underline{\underline{\mathfrak{C}^{\mathrm{iso}}}}}_{\underline{\underline{\theta}}} \circ \mathscr{D}_\alpha\left(\mathscr{C}_{\gamma_c}\left(\underline{\underline{f_i}}\right), \mathscr{M}_{\gamma_m}\left(\underline{\underline{f_i}}\right) + \mathscr{S}^{\frac{1}{2}}_{\gamma_s}\left(\underline{\underline{f_i}}\right) \odot \epsilon\right), \qquad (28)
\end{aligned}
$$

where an isotropic matrix-family convolution is for the K-clusters:

$$
\underline{\underline{\underline{\mathfrak{C}^{\mathrm{iso}}}}}_{\underline{\underline{\theta}}} : \mathbb{R}^{|\Omega|\times P} \to \mathbb{R}^{|\Omega|\times K}.
$$

We define distributions for the set $\left(\mathfrak{f}, \mathfrak{f}^{\mathrm{gt}}\right)$ as $\mathbb{Q}\left(f_{i,l}\right) = q\left(f_{i,l}\right) \mathrm{d}f_{i,l}, \mathbb{P}\left(f_{i,l}\right) = p\left(f_{i,l}\right) \mathrm{d}f_{i,l}$ and their densities are:

$$
q\left(f_{i,l}\right) = \left[q_k\left(f_{i,l}\right)\right]_{k=1}^K \in \mathbb{R}^K,\ q_k\left(f_{i,l}\right) = \mathrm{softmax}\left(u\left(f_{i,l}\right)\right)_k = \frac{\exp\left(u_k\left(f_{i,l}\right)\right)}{\sum_{h=1}^K \exp\left(u_h\left(f_{i,l}\right)\right)},
$$

$$
p\left(f_{i,l}\right) = \left[p_k\left(f_{i,l}\right)\right]_{k=1}^K \in \mathbb{R}^K,\ p_k\left(f_{i,l}\right) = f^{\mathrm{gt}}_{i,l,k} \in \{0,1\},\ \sum_{k=1}^K p_k\left(f_{i,l}\right) = 1;
$$

then, cross-entropy loss between distributions $\mathbb{P}$ and $\mathbb{Q}$ is:

$$
\begin{aligned}
\mathcal{H}(\mathbb{P}, \mathbb{Q}) = -\mathbb{E}_{\mathbb{P}}\left[\log \mathbb{Q}\right] &:= -\sum_{k=1}^K \mathbb{E}_{f\sim\mathfrak{f}}\left[p_k(f) \log q_k(f)\right] \\
&= \lim_{n\to\infty} -\frac{1}{n}\sum_{i=1}^n \mathscr{H}\left(\underline{\underline{f^{\mathrm{gt}}_i}}, u\left(\underline{\underline{f_i}}\right)\right),
\end{aligned}
$$

$$
\mathscr{H}\left(\underline{\underline{f^{\mathrm{gt}}_i}}, u\left(\underline{\underline{f_i}}\right)\right) = \sum_{l\in\Omega}\sum_{k=1}^K f^{\mathrm{gt}}_{i,l,k} \log \mathrm{softmax}\left(u\left(f_{i,l}\right)\right)_k.
$$

The last equality is due to Monte Carlo approximation method.

5.9 Proof of proposition 4.1

We start with a 2D convolution (8)

$$
\underline{\tilde{f}} = \underline{\mathfrak{C}}_{\underline{\phi}}(\underline{f}) = \left(\tilde{f}_1 \quad \ldots \quad \tilde{f}_{d_2}\right) \in \mathbb{R}^{n_1\times d_2},\ \tilde{f}_i \in \mathbb{R}^{n_1}.
$$

Then, we rewrite an element of the above 2D convolution as a 1D convolution for a 2D image $\underline{f} = \begin{pmatrix} f_1 & \dots & f_{n_2} \end{pmatrix} \in \mathbb{R}^{|\Omega|}$, $f_i \in \mathbb{R}^{n_1}$: $r_1 = 1, \dots, n_1$, $r_2 = 1, \dots, d_2$,

$$\begin{aligned}
\tilde{f}[r_1, r_2] = \underline{\mathfrak{C}}_{\underline{\phi}}(\underline{f})[r_1, r_2] &= \sum_{k_1=1}^{d_1} \sum_{k_2=1}^{d_2} f[k_1, k_2] \check{\phi}[r_1 - k_1, r_2 - k_2] \\
&= \sum_{k_2=1}^{d_2} \left(\sum_{k_1=1}^{d_1} f_{k_2}[k_1] \check{\phi}_{r_2}^{k_2}[r_1 - k_1] \right) \\
&= \sum_{k_2=1}^{d_2} \mathfrak{C}_{\phi_{r_2}^{k_2}}(f_{k_2})[r_1] := \tilde{f}_{r_2}[r_1] \,,
\end{aligned}$$

where column vector notations are

$$f_{k_2}[k_1] := f[k_1, k_2] \,, \; \phi_{r_2}^{k_2}[r_1 - k_1] := \phi_{k_2}[r_1 - k_1, r_2] := \phi[r_1 - k_1, r_2 - k_2] \,.$$

In conclusion, we have a useful relation between 2D and 1D convolution operations: $k_1 = 1, \dots, n_1$, $k_2 = 1, \dots, d_2$,

$$\tilde{f}_{k_2}[k_1] = \underline{\mathfrak{C}}_{\underline{\phi}}(\underline{f})[k_1, k_2] = \sum_{i=1}^{d_2} \mathfrak{C}_{\phi_{k_2}^i}(f_i)[k_1] = \sum_{i=1}^{d_2} \left(\mathscr{H}_{d_1}(f_i) \phi_{k_2}^i \right)[k_1] \,, \tag{29}$$

where $f_{k_2}[k_1] := f[k_1, k_2]$.

Then, a multi-input-numlti-output convolution is:

$$\begin{aligned}
\underline{\tilde{f}} &= \left(\tilde{f}_1 \,, \quad \dots \,, \quad \tilde{f}_{d_2} \right) \in \mathbb{R}^{n_1 \times d_2} \,, \; \tilde{f}_i \in \mathbb{R}^{n_1} \,, \\
&= \sum_{i=1}^{d_2} \left(\mathfrak{C}_{\phi_1^i}(f_i) \,, \quad \dots \,, \quad \mathfrak{C}_{\phi_{d_2}^i}(f_i) \right) \overset{(7)}{=} \sum_{i=1}^{d_2} \mathscr{H}_{n_1}(f_i) \underbrace{\left(\phi_1^i \,, \quad \dots \,, \quad \phi_{d_2}^i \right)}_{= \underline{\Phi}^i} \\
&= \mathscr{H}_{n_1 | d_2}(\underline{f}) \underline{\Phi} \,.
\end{aligned}$$

This concludes a relation (8).

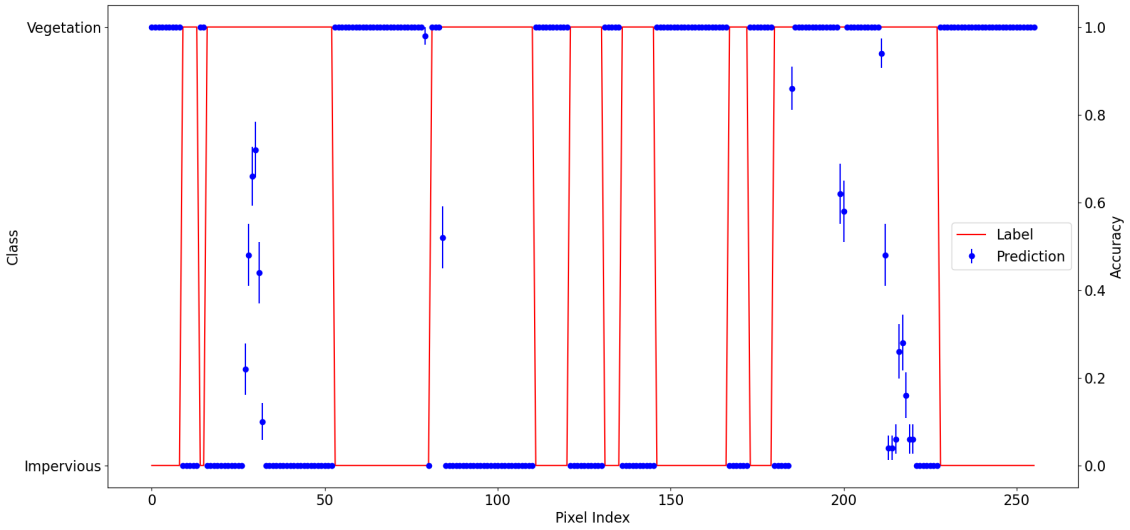


Figure 6: The probability that individual pixels are assigned to the impervious class along single horizontal line of pixels in an image during 50 RQUNet-VAE predictions compared to ground truth. Blue dots correspond to predicted accuracy (right y-axis) whereas the red line corresponds to the ground truth land cover class (left y-axis). This illustrates the results of variational terms of RQUNet-VAE.

6 Additional Experiments

While the main paper describes precision, recall, and F1-scores for entire images, this experiment looks at the accuracy of individual pixels an image to understand where our RQUNet-VAE incurs errors and where the variational approach is able to identify pixels having low confidence. For this purpose, we look at one horizontal line of pixels taken from the vertical center of the first noisy image shown in Figure 2. For each pixel, from the left to the right of the image, Figure 5.9 shows the ground truth label and the prediction accuracy. We see that for this line of pixels, the ground truth changes frequently between the impervious and the vegetation (non-impervious) class. We observe that there are many pixels where RQUNet-VAE provides a prediction accuracy of 1, implying that the class was predicted correctly in each of 50 runs of RQUNet-VAE (which is not deterministic due to the variational auto-encoder module). For example, on the very left of Figure 5.9 we observe that numerous vegetation pixels were classified correctly in all cases. Then however, the ground truth has a few pixels classified as impervious, which our RQUNet-VAE is not able to capture. For these pixels, we have an accuracy of 0, meaning that even the repeated iterations of our variational approach are not able to correctly classify these pixels. Across this horizontal line of pixels we also have numerous pixel where the variational approach provides a non-binary classification. For these cases, we provide the accuracy (as the fraction of correct classifications among the 50 runs) as well as the standard deviation of this accuracy denoted by the whiskers around the point estimate in the figure. We observe that in these cases, the variational approach allows our RQUNet-VAE to possibly make better decisions by allowing to base a decision on multiple runs and take the consensus of all runs.

While these results appear weak due to many miss-classifications, we again reiterate that the underlying image has been heavily obfuscated with noise (compare the original images in Figure 1 with the noised images in Figure 2) and that the experiments in the main paper show that the classic U-Net yields worse results, thus showing that our proposed RQUNet-VAE augments the traditional U-Net architecture by making it more robust to noise.