Quality-preserving low-cost probabilistic 3D denoising with applications to Computed Tomography

Illia Horenko*[1], Lukas Pospisil[2], Edoardo Vecchi[1], Steffen Albrecht[1,3], Alexander Gerber[4], Beate Rehbock[5], Albrecht Stroh[6], and Susanne Gerber*[7]

[1]USI Lugano, Faculty of Informatics, Via G. Buffi 13, TI-6900 Lugano, Switzerland

[2]VSB Ostrava, Department of Mathematics, Ludvika Podeste 1875/17 708 33 Ostrava, Czech Republic

[3]Institute of Physiology, University Medical Center of the Johannes Gutenberg-University Mainz, Hanns-Dieter-Huesch-Weg 19, 55128 Mainz, Germany

[4]Institute of Occupational Medicine, Faculty of Medicine, GU Frankfurt, Theodor-Stern-Kai 7, Haus 9b, 60590 Frankfurt am Main, Germany

[5]Lung Radiology Center Berlin, Bismarckstr. 45-47, 10627 Berlin, Germany

[6]Institute of Pathophysiology, University Medical Center of the Johannes Gutenberg-University Mainz, Hanns-Dieter-Huesch-Weg 19, 55128 Mainz, Germany

[7]Institute for Human Genetics, University Medical Center of the Johannes Gutenberg-University Mainz, 55128 Mainz, Germany

August 10, 2021

*To whom correspondence should be addressed: horenkoi@usi.ch and sugerber@uni-mainz.de

Abstract

We propose a pipeline for a synthetic generation of personalized Computer Tomography (CT) images, with a radiation exposure evaluation and a lifetime attributable risk (LAR) assessment. We perform a patient-specific performance evaluation for a broad range of denoising algorithms (including the most popular Deep Learning denoising approaches, wavelets-based methods, methods based on Mumford-Shah denoising etc.), focusing both on accessing the capability to reduce the patient-specific CT-induced LAR and on computational cost scalability. We introduce a parallel probabilistic Mumford-Shah denoising model (PMS), showing that it markedly-outperforms the compared common denoising methods in denoising quality and cost scaling. In particular, we show that it allows an approximately 22-fold robust patient-specific LAR reduction for infants and a 10-fold LAR reduction for adults. Using a normal laptop the proposed algorithm for PMS allows a cheap and robust (with the Multiscale Structural Similartity index $> 90\%$) denoising of very large 2D videos and 3D images (with over 10^7 voxels) that are subject to ultra-strong Gaussian and various non-Gaussian noises, also for Signal-to-Noise Ratios much below 1.0. The code is provided for open access.

One-sentence summary: Probabilisitc formulation of Mumford-Shah principle (PMS) allows a cheap quality-preserving denoising of ultra-noisy 3D images and 2D videos.

Introduction

Computed tomography (CT) is one of the most frequently used medical imaging techniques, with over 100 million CT scans performed yearly worldwide [1]. An additional increase in the total number of CT examinations could be observed in the recent COVID-19 epidemics [2, 3]. However, distinguishing subtle CT image features relevant for diagnostics purposes typically requires significant radiation exposure and thus increases the patient's radiation-imposed lifetime attributable risk (LAR). This, in turn, leads to an additional chance of developing a radiation-exposure attributable cancer type [1]. The quantification of the LAR is a complex challenge and requires modeling the multifactorial interplay of DNA damage and repair mechanisms, as well as incorporating random/stochastic effects

that accumulate in the low-radiation regime. *In silico* simulations and analytical estimates for net effects of such stochastic radiation-triggered reactions imply a linear model for the dependence of LAR from the accumulated radiation exposure [4, 5, 6, 7] – with linear model coefficients being dependent on the patient's age and sex, as well as on the particular type of the CT. Despite some controversy regarding the possible existence of low-radiation thresholds in the LAR models suggested by some studies [8], the linear no-threshold models (LNT) are currently recommended for LAR assessment by the committee for Biologic Effects of Ionizing Radiation (BEIR VII) of the National Academy of Sciences of the USA [5] and by the World Health Organisation [1]. Several recent epidemiological and methodological studies support the statement that a safe radiation dose does not exist [9, 10, 11, 12] and that the LAR of CT is exceptionally high for infants and children [10, 11, 12]. The approximately 14 million pediatric CT scans of head, abdomen, pelvis, chest, or spine performed each year worldwide [10, 1] would therefore lead to approximately 12.000 fatal cases of cancer, of which 4800 are attributable only to the USA.

The prognosis that the reduction of the highest 25% of doses to the median could prevent 43% of these cancers [10] naturally suggests the increased use of low- and ultra-low radiation CT (for radiation exposures down to 0.5 mGy). However, a reduction of radiation exposure results in increased image noise and thus necessitates the application of reliable image denoising and feature extraction tools. Facilitated by the rapid development of emergent machine learning (ML) and deep learning (DL) algorithms, this research on the boundary between medical radiology and informatics is attracting an increasing amount of attention over the past years [13]. The currently available CT image denoising tools can be roughly subdivided into unsupervised and supervised methods. The unsupervised approaches search for a hidden pattern without prior learning, whereas the supervised techniques aim to identify features previously learned from the training data. Unsupervised methods do not require previous training, allow high-speed computations, and belong to the most frequently-used image denoising instruments [14, 15]. They include methods based on local averaging of the data (like Gaussian, weighted Gaussian, bilateral and mean average filtering) [16, 17, 18, 14] and spectral methods (like Fourier-, wavelet- and PCA-denoising) [19, 20, 21, 22, 15]. Recent years have also seen an active development of very successful CT denoising approaches based on semi-

supervised ML ideas (for example, methods based on generative adversarial networks) [23, 24] and Deep-learning algorithms for denoising- and image segmentation [25, 26, 13, 27]. The deep learning methods have been shown to be very successful for denoising and the current convention says that DL performs much better than traditional unsupervised regularized denoising algorithms.

However, recent evidences in the literature indicate that ML and DL tools can struggle when dealing with the denoising of real images, either due to the lack of an adequate training sample or to the increasing complexity (and computational cost) of the required network [28]. This is particularly true in medical imaging, where the approaches based on ML can sometimes lack accuracy [29], while DL tools tend to rely too heavily on labeled datasets and on sufficiently large training sets [30, 31, 32]. The size of the training set plays a very central role also in the denoising of CT images, where the number of instances in the training set T is significantly smaller than the feature space dimension D, corresponding to the number of voxels. A problem characterised by $D \gg T$ pertains to the so-called "small-data learning challenge" [33, 34, 35, 36, 37], and represents a scenario in which ML and DL approaches are prone to quickly overfit the small training set and to achieve an unsatisfactory performance on the validation set [38, 39, 40, 41]. To tackle this issue, several alternative approaches have been proposed [42, 43], with transfer learning representing one of the most powerful alternatives [44]. Even the latter approach presents, however, some limitations that are particularly relevant in the denoising of CT images: due to the individual variation of small-scale anatomical features and of CT operation regimes, the structural similarity assumption between the source domain and the target domain is usually not fulfilled, while remains unclear the amount and type of information that needs to be transferred if we want to avoid potential drawbacks – like, e.g., negative transfer – that could actually lead to a performance worse than the starting deep learning model [45, 46]. Thus, while a combination of transfer learning and deep learning is being widely used to attempt the solution of small data problems in the denoising of medical images [47, 48, 49], the reported results can still be dissatisfactory due to the lack of efficient strategies to systematically tackle these limitations [50].

The issues described above are not the only ones arising in the small data regime characteristic for CT: a statistically-significant systematic comparison and benchmarking of the supervised learning approaches can be strongly biased by the so-called "concept drift", i.e., a scenario in which the non-

stationarity of the learning problem leads to a mismatch between the training data and the actual application data [51, 52, 53, 54, 55]. In CT imaging, such context-dependence of supervised ML and DL tools becomes particularly problematic when there is a discrepancy between the type of patient (age, sex, body size) and noise model tackled in the training set and those tested in the validation. This context-dependence and "concept drift" can quickly lead to unfair comparisons and unsatisfying performances of supervised learning methods. Last but not least, robustness of the learning methods can be strongly confined by the existence of structural constraints inherent for the ML and DL tools in the "small data challenge" regime: for example, while spectral filtering methods tend to outperform other unsupervised denoising algorithms [14], they also have a fundamental difficulty in dealing with high noise levels in the data [19, 20]. Recently, the existence of statistically-significant overfitting boundaries has been shown empirically by employing high-performance facilities: e.g., in [56], long short-term memory (LSTM) deep neural networks [57] have been shown to systematically overfit the data and to produce results which are not statistically-significant if the condition $T \geq 13.6D + 3.8$ is not satisfied (where T is the size of data statistics and D is the number of features).

While regularised time series clustering approaches were recently demonstrated to operate in these "small data, large noise" regimes, even when the noise is an order of magnitude larger than the true signal [58, 59, 60, 61], these studies were only confined to one-dimensional denoising problems. A systematic comparison with a broad range of supervised and unsupervised methods is still lacking. Due to the stochastic nature of the noise in CT, a statistically-significant evaluation and comparison of different CT image denoising methods has to rely on sufficiently large amounts of CT images taken from the same patient under the same combination of controls (e.g., with the same tube current and the same tube voltage). However, obtaining such an extensive set of reference-imaging data for a particular patient without a medical necessity would be unethical. A systematic comparison of methods would additionally require combining such data for multiple patients in a sampled range of patient-specific parameters (age, sex, body size, etc.) as well as for a large number of practically-relevant combinations of CT controls. Furthermore, the standard quality measures like the Mean-Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNE), and the Multi-Scale Structural Similarity Index (MS-SSIM) also rely on the availability of the reference image without noise but generated with the

same set of underlying features [62, 63, 64]. Finally, combining existing CT data from different sources in a metastudy is problematic as well, due to a very high level of the individuality of the more subtle anatomic features of the human body on a small scale [65, 66] and would thus introduce a strong bias into such a comparative study, which would also lack the reference images. Furthermore, very few datasets containing CT projection data covering the low-radiation regime are currently available in open access, mainly due to the proprietary nature of this data and the (hidden) manufacturer-specific processing of the raw data [67, 68, 69, 70]. Even, when this information is available, like in the low-dose CT image and projection dataset described in [70], a systematic statistically-significant comparison is problematic since for each of the patients only a couple of images (with and without noise) are available, from overall $T = 299$ clinically performed patient CT exams - and with the radiation exposures practically not going below 3 mGy. As we will show below, this ultra-low radiation regime with radiation exposure down to 0.5 mGy and with SNR<1.0 imposes critical challenges for the bulk of currently-available denoising methods and will receive a particular attention in the tests performed below. To address these issues, we will lay two foundations in this manuscript.

First, we propose a pipeline for the automated patient-specific generation of synthetic CT images, radiation exposure estimatition and LAR computation, herewith following the strengthening movement in radiological research and using the synthetic images (e.g., like in the software tool CatSim) [67]. The created images are based on a data-driven estimation of CT image noise intensities and their relationship to CT control parameters[71, 66, 72, 73]. For this purpose, we combine the LNT model for the CT-induced lifetime attributable risk [5, 9, 10, 11, 12] with the data-driven models that relate CT noise variance to the CT voltage, current and the amount of radiation exposure [71, 74]. **Second**, we introduce the Probabilistic formulation of the Mumford-Shah formalism (PMS) and propose a regularized Scalable Probabilistic Approximation algorithm (rSPA) and its parallel extension DD-rSPA as new methods for denoising of 3D images, comparing their computational cost and denoising performances with the state-of-the-art methods in this field. Particular focus thereby is to investigate the possibility of reducing personalized LAR through improving the denoising performance in the ultra-low radiation regime (down to 0.1-0.5 mGy, with Signal-to-Noise Ratios below 1.0).

Results

Patient-specific generation of synthetic CT images, radiation exposure estimation and LAR computations

In the first step of the proposed pipeline we provide algorithms for generating synthetic noisy CT images for every relevant combination of CT control parameters, image parameters and patient dependent variables. Regarding the CT control parameters, we focus on the two most relevant ones that can be adjusted on the computer tomograph, which are the tube voltage, **kVp** and the tube current, **mA** . The CT image parameters are the standard deviation of the CT quantum noise, σ, and the CT feature contrast in Hounsfield Units (HU). The patient-dependent variables for computing the overal CT-quantum noise as well as the CT-induced additional cancer risk, **r**, are the patient's **age**, **sex**, the subject's size, **d**, in *cm*, as well as the absorbed radiation dose density $\mathbf{CTDI_{vol}}$ in milligray (mGy). The initial reference data for the automated generation of a battery of synthetic test-images can be either a set of real CT-data generated using high-dose radiation (Fig. 1A), or artificially simulated data, respectively. These reference data have to be characterized by high image quality and low quantum noise (visualized in Fig. 1B), as compared to the (ultra) low-dose CT images (Fig. 1C) that naturally contain a massive amount of noise and thus result in low CT-image quality. Figure 1D gives a graphical abstract of the workflow from image generation to the subsequent comparison of the various ML/DL-denoising methods based on the accuracy of the denoised image data. Starting with high-quality reference data, a broad range of typical CT image noises is imposed in a multitude of combinations from patient-specific and CT control variables. The obtained noisy CT images are subsequently denoised using various state-of-the-art methods. Processed and denoised images are compared to the original reference data. To model the effect of noise in CT images, we deploy and compare three different alternatives: (i) an additive Gaussian noise model that was shown to provide an adequate description of quantum noise effects in real CT images on a small scale of several centimeters [71, 75]; (ii) the non-Gaussian multiplicative noise model where the quantum noise variances change with the underlying feature color; and (iii) the empirical CT noise model sampled from the real patient data.

Computation of the noise variance σ is performed for given CT control parameters (tube current $\mathbf{mA}$, tube voltage $\mathbf{kVp}$) and patient-specific parameter (water-equivalent patient diameter $\mathbf{d}$) using the non-linear regression model introduced in [71] (see equation (1). Equation (2) of the workflow computes the effective absorbed radiation dose density $\mathbf{CTDI_{vol}}$ for a volume unit from the tube control parameters $\mathbf{mA}$ and $\mathbf{kVp}$ using the data-driven regression model established in [74]. Equation (3) of the image generation workflow computes the resulting lifetime attributable risk for a patient (LAR) utilizing the linear no-threshold model (LNT) proposed by the committee for Biologic Effects of Ionizing Radiation (BEIR VII) of the National Academy of Sciences of the USA [5, 76].

$$\begin{aligned} ln(\sigma) =& \alpha_0(kVp) + \alpha_1(kVp)d + \alpha_2(kVp)ln(mA) + \alpha_3(kVp)d^2 + \alpha_4(kVp)ln^2(mA) + , \\ & \alpha_5(kVp)dln(mA), \end{aligned} \tag{1}$$

$$CTDI_{vol} = \gamma_0(kVp, CT\,type) + \gamma_1(kVp, CT\,type)mA, \tag{2}$$

$$LAR = \beta_0(age, sex, organ) + \beta_1(age, sex, organ)CTDI_{vol}. \tag{3}$$

3D regularized Scalable Probabilistic Approximation algorithm:

In the following, we introduce the 3D regularized Scalable Probabilistic Approximation algorithm (rSPA). More algorithmic details and a complete derivation with mathematical proofs can be found in the paper supplement. rSPA (see Fig. 2 for a graphical representation) seeks a simultaneous solution of image segmentation and noise elimination problems and aims to find the spatially most-persistent decomposition of the image in terms of K latent features. Direct application of popular segmentation and clustering methods from ML to the denoising problem results in computatinally-tractable tools with a favourable linear scaling of computational cost - but resulting in suboptimal irregular segmentations that disregard the spatial ordering of the data [77, 78, 79]. Application of regularized clustering and segmentation tools that take into account the spatial ordering and regularity of the data and features (e.g., methods based on Mumford-Shah functional optimization) have unfavourable polynomial cost scaling, limiting their application to very small images or requiring very extensive computational

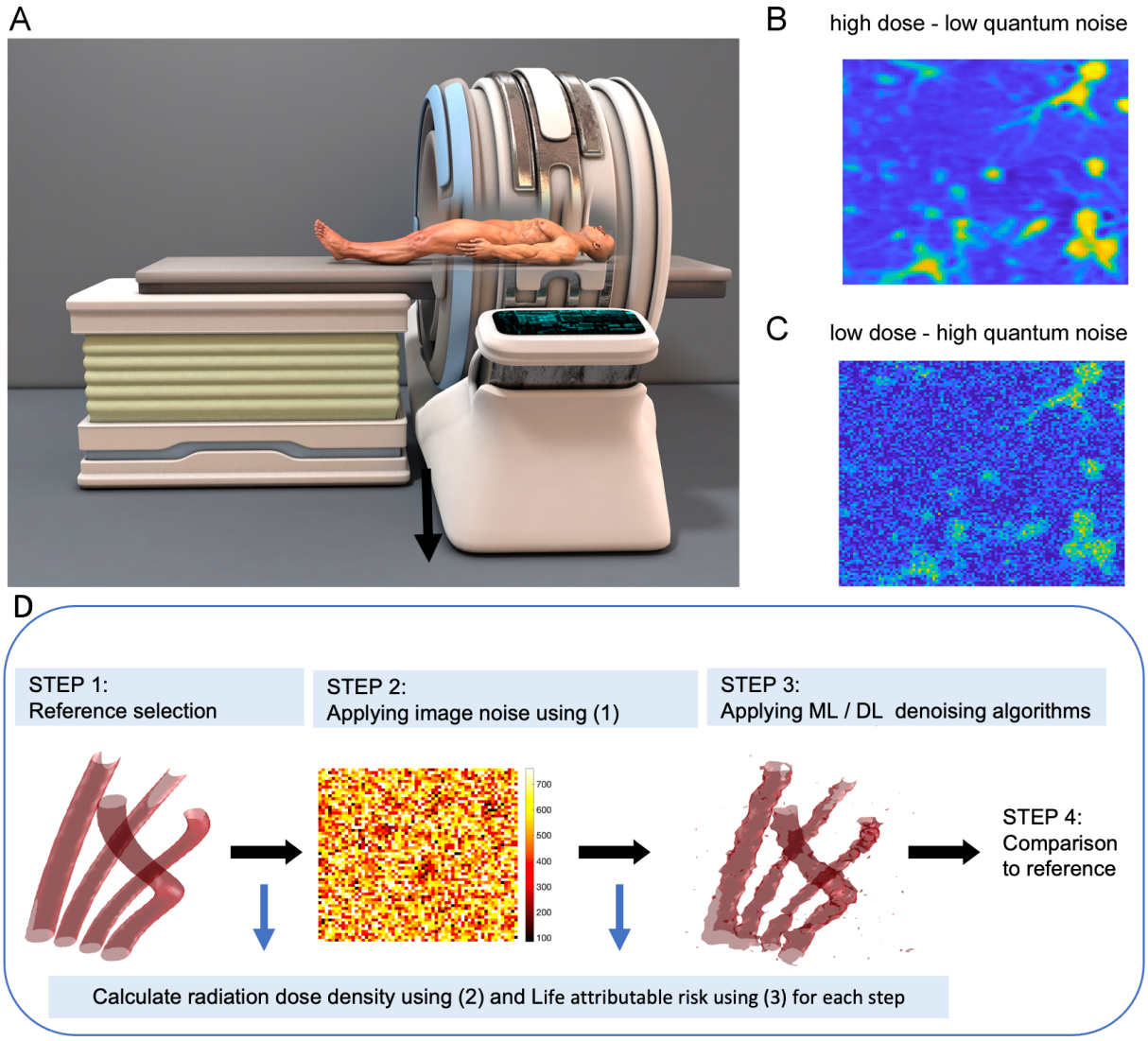

Figure 1: **Graphical representation of our proposed pipeline's workflow for automated generation and risk assessment of CT images** A: Initial reference data can be either a set of real CT-data generated using high-dose radiation or artificially simulated data. B: Exemplary high-quality and low-quantum noise image of lung vessels. C: exemplary low-dose CT images with high quantum noise. D: Workflow from image generation to subsequent benchmarking of ML/DL-denoising methods. Starting with high-quality data or artificially generated reference data, respectively, a spectrum of image noise $\boldsymbol{\sigma}$ is added for a multitude of combinations from patient-specific and CT control variables, as suggested in equation (1). The noisy images were then denoised using various state-of-the-art methods, and the processed images are compared to the original reference data.

resources [80, 58, 59, 60, 61]. In the following, the key challenge we will address with the proposed rSPA algorithm simultaneously achieving a qualitative(in terms of low error and sufficient spatial regularity of latent features) and computationally-tractable (linearly scalable) solution of the underlying optimization problem.

We consider a 3D image to be provided as an array $V = \{V(1), V(2), \ldots, V(T)\}$ of D-dimensional patch value vectors for all T of three-dimensional CT voxels, with patch values $V(t) \in \mathcal{R}^D$ being, for example, the grey-color intensities $V_d(t), d = 1, \ldots, D$ of the D-dimensional voxel patch with an index t. Without a loss of generality, in the following applications, we will consider the common grayscale CT images with one-voxel patches ($D = 1$) and T being of the order $10^5 - 10^7$. The problem of denoising can then be considered as a numerical problem of searching for K D-dimensional latent features characterized by K D-dimensional distinct feature vectors $\{C_{1,k}, \ldots, C_{D,k}\}$, with k taking values between 1 and K. Spatial characteristics of these K latent features we will be searching for will be provided by (a priori unknown) latent feature probabilities $\Gamma_k(t)$, representing the probabilities of an actual (noisy) voxel $V(t)$ to belong to a particular latent (noiseless) feature with an index k. Such numerical procedure can be performed by a broad range of clustering and segmentation algorithms from ML (e.g., K-means, Scalable Probabilistic Approximation and others) [77, 78, 79, 81]. For example, the Scalable Probabilistic Approximation algorithm [81] would minimize the sum of the errors $L_t(C, \Gamma(t))$ when approximating every vector $V(t)$ with its probabilistic representation $\tilde{V}_{C,\Gamma}(t) = \sum_{k=1}^{K} \Gamma_k(t) C_k$:

$$[C^*, \Gamma^*] \quad = \quad \underset{C,\Gamma}{\arg\min} \frac{1}{T} \sum_{t=1}^{T} L_t(C, \Gamma(t)), \tag{4}$$

where $L_t(C, \Gamma(t)) = \|V(t) - \tilde{V}_{C,\Gamma}(t))\|_2^2$. It is straightforward to see that, when C is fixed, the solution of the minimization problem (4) is equivalent to T independent minimizations of individual errors L_t with respect to their particular $\Gamma(t)$ - and can be performed independently for each t. This allows a very efficient - independent and parallel - numerical treatment of problem (4) and results in a favourable linear scaling of the computational cost with growing size and dimension of the data [81]. The downside of this nice independent and additive structure of optimization problem (4) is that it

results in solutions that are independent of any spatial permutation of the original data V, since the right-hand side of expression (4) is clearly invariant with respect to any arbitrary re-ordering of the summation indices t. This indicates that the solutions of such an optimisation problem will not change if we arbitrarily change the spatial ordering of the voxels in the original image. This invariance of the clustering outcomes with respect to the data ordering is a common characteristics of a broad class of ML methods, including, for example, Kmeans- and Fuzzy-Kmeans-clustering-methods that belong to the most popular ML algorithms, with over 3 Mio. citations according to Google-Scholar [81]. While analysing spatially-ordered data, in addition to a simple segmentation (4) of the image into K latent probabilistic features, we would like to enforce a spatial persistence of underlying features. To achieve this, we can enforce any two voxel points $V(t)$ and $V(t')$ to have similar latent probabilities of belonging to the same features if their positions are close enough to each other. In order to deal with the relative position of the voxels, we can use the kernel function, a very popular concept in ML. The simplest alternative to measure the "closeness" of two different voxels would be provided by the Euclidean kernel, defined as a distance function $\alpha_{t,t'}$ between two distinct points with indices t and t':

$$\alpha_{t,t'} = \begin{cases} 1 & \text{if } \text{dist}^{\text{Eucl}}(t,t') \leq \alpha_0, \\ 0 & \text{if } \text{dist}^{\text{Eucl}}(t,t') > \alpha_0, \end{cases} \tag{5}$$

where α_0 is some user-defined threshold (e.g., $\alpha_0 = 1$ in this paper's applications).

Then, following the idea behind the Mumford-Shah functional formulation [80], spatially-persistent optimal probabilistic approximation $\tilde{V}_{C^*,\Gamma^*}$ of the original image data V can be computed via the numerical minimization of the regularized form of the original clustering problem (4):

$$[C^*,\Gamma^*] \;=\; \underset{C,\Gamma}{\arg\,\min}\, \frac{1}{T}\left[\sum_{t=1}^{T} L_t + \frac{\bar{\varepsilon}}{\sum_{t,t'=1}^{T}\alpha_{t,t'}} \sum_{t,t'=1}^{T} \alpha_{t,t'} \|\tilde{V}_{C,\Gamma}(t) - \tilde{V}_{C,\Gamma}(t'))\|_2^2\right]. \tag{6}$$

The second term in the right-hand side of this functional controls the spatial regularity and smoothness of the obtained solutions. Please note, that, in contrast to the original clustering problem (4), problem (6) is not invariant with respect to permutations of V, and allows to optain spatially-regular solutions $[C^*,\Gamma^*]$, with the persistence that grows when increasing the scalar control parameter $\bar{\varepsilon}$. However,

these nice features of the regularized problem come at a price of losing the very-favourable linear scalability of the computational cost of problem (4): optimization with respect to different $\Gamma(t)$ can not be performed independently when C is fixed - as it is the case for the clustering problem like SPA (4), where one solves T independent K-dimensional optimization problems for $\Gamma(t)$ with fixed C. The second term in (6) - that aimed at enforcing spatial regularity and persistence - at the same time introduces the global coupling between different $\Gamma(t)$ and requires the solution of very large coupled KT-dimensional nonlinear optimization problems [80, 58, 59]. This confines the applicability of the image analysis methods based on (6) when working on common hardware (e.g., workstations) to relatively-small images, with KT not larger then 50'000-100'000 [58, 59]. Direct solution of (6) - as well as indirect Bayesian solutions of (6) based on Markov Chain Monte Carlo sampling (MCMC) - are costly beyond 1D and would require extensive use of High-Performance Computing facilities (HPC) for large realistic 3D images with $KT \approx 10^5 - 10^7$ [82, 61].

One of the key methodological insights of this work is that one can systematically derive an exact upper bound approximation of the regularized problem (6) that can be solved with a linearly scalable and parallelizable numerical algorithm for realistic 3D images (with $10^6 - 10^7$ voxels), while requiring few minutes on a common laptop:

$$
\begin{aligned}
[C^*, \Gamma^*] \;\; = \;\; & \underset{C,\Gamma}{\arg\,\min} \sum_{k=1}^{K} \left[\frac{1}{T} \sum_{t=1}^{T} \Gamma_k(t) \|V(t) - C_k\|^2 + \frac{\varepsilon \|C_k\|^2}{\sum_{t,t'=1}^{T} \alpha_{t,t'}} \sum_{t,t'=1}^{T} \alpha_{t,t'} (\Gamma_k(t) - \Gamma_k(t'))^2 \right] \\
& \text{such that } \min(V) \leq C_k \leq \max(V), \;\; \sum_{k=1}^{T} \Gamma_k(t) = 1 \text{ and } \Gamma_k(t) \geq 0 \text{ for all } t, k. \qquad (7)
\end{aligned}
$$

As proven in Lemma 1 of the paper supplement, solutions of problem (7) are also exact solutions of the original regularized problem (6) if the segmentations are discrete (i.e., if $\Gamma_k(t)$ take only discrete values 0 or 1). These solutions provide upper bound approximate minimizers of the problem (6) if $\Gamma_k(t)$ take fuzzy values between 0 and 1. In contrast to the original clustering SPA-functional (4), problem (7) has $\Gamma(t)$ outside of the norm in the first (clustering) term - and the analytical structure of the second (regularizing) term is very different from the structure that one would obtain by directly deploying common regularization tools (like Ridge, Lasso and elastic net regularizations) to

the original clustering problem (4)[1]

The numerical solution of the obtained optimization problem (7) can be computed with the monotonically-convergent rSPA algorithm: starting with some arbitrarily chosen K feature vectors C, one iterates solving the above problem for Γ (with fixed C) and minimizing of (7) for C (with fixed Γ). As proven in Lemma 2, 3 and in Theorem 1 of the paper supplement, rSPA always results in the monotonic minimization of (7), with a linear iteration cost scaling $\mathcal{O}(KDT)$. rSPA algorithm can be efficiently parallelized deploying the Domain Decomposition idea (DD) widely used in various areas. Graphical representation of the idea underlying the resulting parallel DD-rSPA algorithm is shown in the Fig. 2, a detailed description of the DD-rSPA algorithm is provided in the Section 2 of the paper supplement. Commented computer code implementing both algorithms is provided for open access at `https://www.dropbox.com/sh/rw4t6ydkpi64w8y/AAA9katysG09w7ljsvUqPwwna?dl=0` and can be run on a laptop with MATLAB installation. Numerical tests on noisy images with different sizes and noise levels reveal that the overall computational cost of both the sequential rSPA and parallel DD-rSPA algorithms grows linearly withthe image size and with decreasing Signal-to-Noise ratios (corresponding to increasing noise levels), as we can see in the panel A of the Fig. 5.

Relation of Probabilistic Mumford-Shah and rSPA algorithm to Regularized Mumford-Shah framework (MS) and Rudin-Osher-Fatemi (ROF) Total Variation model:

Mumford-Shah formalism originally introduced in [80] is one of the most well-understood and elaborated theoretical and algorithmic frameworks for edge-preserving image denoising. It aims at finding the optimally-denoised image V^d that is simultaneously *smooth* and *close enough* to the original noisy image V. Then, keeping the previously introduced notation, in the most common discrete Mumford-Shah formulation such a denoised image V^d can be found as a solution to the following optimization

[1]Applying Ridge, Lasso and elastic net regularizations with respect to both variables C and Γ in problem (4) would result in regularization terms of the form $+\epsilon_C\|C_k\| + \epsilon_\Gamma\|\Gamma_k\|$ and would require tuning at least the two regularization parameters ϵ_C and ϵ_Γ.

problem:

$$[V^d] \;=\; \underset{V^d}{\arg\,\min}\, \frac{1}{T}\sum_{t=1}^{T}\left[\|V(t)-V^d(t)\|^2+\epsilon^2\|\nabla V^d(t)\|\right], \tag{8}$$

where the first term measures the "closeness" of the original and the denoised images, the second term regularizes the "smoothness" of the denoised image by penalizing the norm of its average gradient. One of the key theoretical insights to this problem (8) was provided in the work by Rudin, Osher and Fatemi [83]: deploying the Euler-Lagrange principle they shown that the solution to the minimization problem (8) is equivalent to solving a parabolic Partial Differential Equation (PDE). This opened a way of deploying the very efficient PDE solvers and the so-called *level-set methods* to the image denoising problem. The numerical solution of both the original MS-formulation (8) and of the PDE-based ROF-formalism is commonly achieved by deploying the Galerkin ansatz:

$$V^d(t) \;=\; \sum_{k=1}^{K} C_k\Gamma_k(t), \tag{9}$$

where $\Gamma_k(t)$ is a fixed set of known basis functions (e.g., mesh functions, finite element functions, wavelet basis functions, Fourier basis functions, etc.) and C_k are the unknown coefficients that are found numerically [83, 84, 85, 86, 87, 88].

The most important difference between the Probablistic Mumford-Shah (PMS) problem formulation (7) and the common MS- and ROF-methods is the form of the Galerkin expansion (9): (i) PMS problem (7) deploys the probabilistic expansion (9), with unknown C and $\Gamma(t)$ being a priori unknown *non-parametric* probability density vectors - whereas common MS and ROF-tools dwell on a priori fixed *parametric* sets of non-probabilistic basis functions Γ. Hence, in contrast to the parametric optimization problem (8) that allows a straightforward Euler-Lagrange reformulation in form of the parabolic PDE, the introduced PMS-formulation deals with a non-parametric variational problem (7) subject to both equality and inequality constraints, that do not allow a straightforward Euler-Lagrange reformulation - and do not allow deploying the very efficient algorithms from PDE numerics for its solution. One of the central methodological developments of this manuscript was

showing that despite of this presumed limitation, it is possible to efficiently solve the PMS problem numerically (7) , with an iterative algorithm that has a linear scalability of computational cost. Direct numerical comparison of PMS and the common MS- and ROF-tools [83, 88] reveals very significant differences in denoising performance, cost and parallel scalability (see Fig. 5 A-C).

Application and comparison of the rSPA method with standard methods:

Next, we compare the denoising performance for a broad selection of supervised and unsupervised algorithms using the synthetic CT images generated with the above-introduced pipeline. As a noise-less CT reference we first use the patient data exemplified in Fig. 3A. It has 274'625 voxels and represents a cubic CT area of around $5cm \times 5cm \times 5cm$. The data came from a high-radiation CT (180 $\mathbf{mA}$ tube current, $\mathbf{CTDI_{vol}}$ 15.4 mGy, section from a thorax CT of a 19 year old female patient). For each particular combination of tube-specific and patient-specific parameters, we used this reference image to create statistics of 100 different independent noisy synthetic CT images for every parameter combination.Fig. 3B shows the increase in noise when reducing radiation exposure. To illustrate the performance of DL on these data we first apply one of the most widely-used DL denoising networks: the Convolutional Neuronal Network DnCNN-3 from [25], with over 3264 citations according to Google Scholar. It was trained on a comprehensive collection of imaging datasets (including the Berkeley segmentation dataset, with over a million of image pairs for training) in a very broad range of Signal-to-Noise ratios and noise types (both Gaussian and non-Gaussian). Figures 3C and 3D show the effects of denoising by DL DnCNN-3 from [25] and rSPA, respectively in low- and ultralow-radiation CT. Fig. 3E) shows a 3-dimensional segmentation obtained from a stack of such high-radiation CT data whereas Figures 3F and 3G give the segmentation based on the images denoised using DnCNN-3 and rSPA, respectively. Figures 3E-F are all obtained from two feature isosurfaces at 625 and 200 Hounsfield Units (HU), respectively, representing the interior of blood vessels in the lung volume segment.

Apparently, rSPA provides denoised images and segmentations that are much closer to the high-radiation reference images. Particularly, we observe that - as the noise increases - DL denoising methods start recognizing features from noise artifacts that were not part of the true reference images.

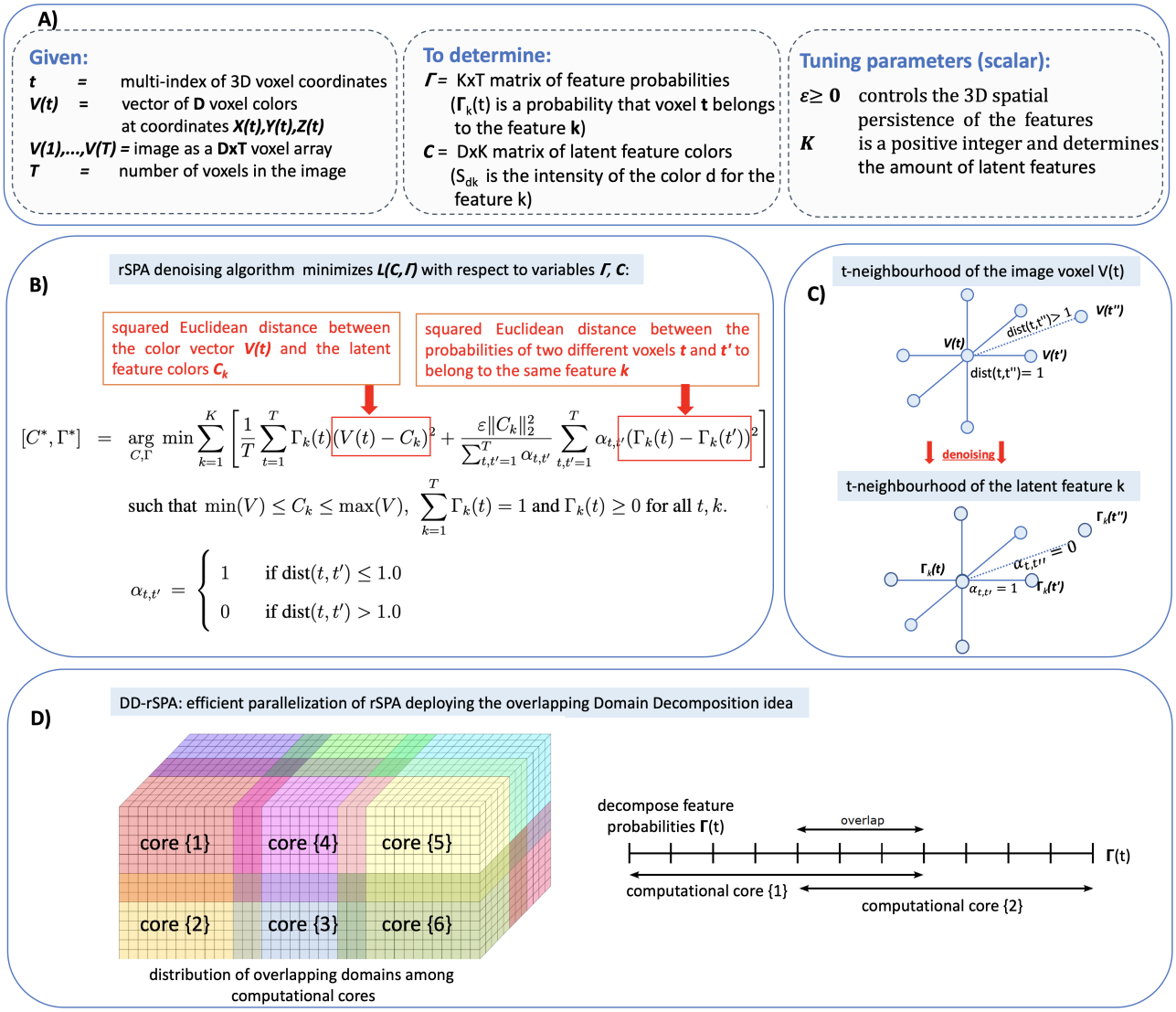

Figure 2: **Graphical overview of the regularized Scalable Probablistic Approximation (rSPA) and its parallel extension DD-rSPA: A)** Summary of the parameters and variables. **b)** Core rSPA algorithm idea: 3D-denoising with the regularized Scalable Probabilistic Approximation algorithm (rSPA). Given the (noisy) CT voxel data V, rSPA minimizes the function $L(C, \Gamma)$ and seeks for the optimal segmentation of V in terms of the K spatially-persistent latent features characterized by the latent feature probabilities in K rows of the matrix Γ as well as by the latent colors as K columns of the latent color matrix C. Persistency of the feature segmentation is imposed by the second term of the right-hand side of the function $L(C, \Gamma)$, that penalizes the differences in the feature probability values in the spatially-neighboring points. **C)**: Denoising idea: latent feature probabilities are persistent (slowly-changing) 3D functions. **D)**: graphical representation of the overlapping Domain Decomposition used in the parallel DD-rSPA algorithm.

As already mentioned above, such deterioration of the performance of ML and DL methods can be attributed to various reasons, including, on one hand, the insufficient training data set and a "small data challenge" [38, 39, 40, 41] and, on the other hand, induced by the "concept drift", stemming from the mismatch between the type of image features and the noise model used in model training and the noise model in the validation data [51, 52, 53].

To discern the potential impact of "concept drift" - and to rule-out the possibility that the "hallucinations" observed for DL CNN in Fig. 3 in the ultra-low radiation regime are induced by the insufficient training dataset - we additionally train the DnCNN-3 from [25] first with 10'000 image pairs (with and without noise) of spheres and circles of various sizes - and then with further 40'000 image pairs. We performed this two-stage training procedure to evaluate the performance improvement induced by providing more training data. The complete additional training took around 8 days on a machine with 28 CPUs (Intel Xeon Gold 6240R 2.4G, 14C/28T) and 384 GB RAM (DDR4-2933) using up to 90% of the physical cores and 120GB of memory. The resulting denoising network is provided for open access at `https://www.dropbox.com/s/ia69h9fhgud2vpt/additionallytrained_DnCNN-3_network.mat?dl=0`. We found that using a larger training dataset (with further 40'000 image pairs) can only bring negligible improvements, confirming the earlier finding reported in [25]. Noisy images in every pair were created using the empirically sampled non-Gaussiann CT noise at various levels, covering low and ultra-low radiation regimes (down to 0.2 mGy, corresponding to the Signal-to-Noise ratios between 5 and 0.1). In the Figure 4 we show some of the results obtained from the application of additionally trained DnCNN to the noisy images of circles and spheres that were not used in the training, deploying the same empirically-sampled non-Gaussian CT noise model as used in the training at the medium noise level (SNR=5, corresponding to the low-radiation CT) and at the high noise level (SNR=0.5, corresponding to the ultra-low radiation CT). Complete comparisons are provided as movie files and are available at `https://www.dropbox.com/sh/n2dbl4h9p4o0p92/AABRkAalhXoaiKFO7ixsSzKga?dl=0`. In the Fig. 4, we observe the same effect of a quick deterioration of DL denoising quality with the increasing noise as in the Fig. 3: at the medium noise level DL provides high-quality denoising, outperforming a very popular unsupervised 3D wavelets denoising tool [19, 20, 21, 22, 15]. However, at the high noise levels

DL is getting outperformed by the 3D wavelet denoising. Interestingly, the best performance, in both cases, is achieved when applying the DL denoising to the data that has been previously denoised by rSPA.

Making an interim assessment of these results, we can conclude that the deteriorating performance of DL denoising is neither a result of a "concept drift" (since the type of features and the noise model deployed in the training and in validation were the same) - neither a consequence of the training data set insufficiency (since we observed only negligible performance improvements of DL when expanding the additional training data from 10K to 50K image pairs). A possible explanation can be given by the fact that here we observe a fundamental robustness boundary of DL denoising in the high noise regime, similar to the Donoho-boundary for wavelets methods [19, 20]. As we will see in the following, further numerical results provided below give additional support to this hypothesis.

In the next step, we compare the computational cost scaling, denoising performance scaling and parallelisability scalings for DL, TV-regularized Mumford-Shah denoising from [88], sequential rSPA, parallel DD-rSPA and parallel DD-rSPA followed by DL. We are particularly interested in analysing the dependence of these characterstics from the image size and noise intensity. For every combination of image size and noise level, we create 10 randomly-generated images of spheres and circles with the non-Gaussian noise - matching the characteristics of the additionally trained DnCNN-3 to avoid the bias through "concept drift". The code reproducing these results is available at `https://www.dropbox.com/sh/6p3q62zaelcyugz/AACkEjggyKcIAdgtoHGWClWPa?dl=0`. The results are summarized in Fig. 5, and their computation of results required around 30 hours on a laptop with a MacBook Pro 3,1 GHz Quad-Core Intel Core i7 (4 cores) with 16 GB RAM. The measurement of the computational cost for DL considered only the pure time of applying the fully-trained DL network to a noisy image - and did not include the time needed for the additional training (that was around 8 days on the workstation as mentioned above). As it can be seen from Fig. 5, the overall costs of all considered methods scale linearly with the image size - and parallel DD-rSPA demonstrating the weak scaling of parallel computation cost (see Fig. 5C). DD-rSPA allows the denoising of a 3D image with $10^7 - 10^8$ voxels in the ultra-low radiation regime (SNR $= 0.5$) at around 3-10 minutes on a MacBook Pro laptop with 4 cores. Interestingly, the costs of DL and common MS denoisings

practically do not depend on the noise level, whereas the cost of rSPA and DD-rSPA grows linearly with the decreasing SNR. According to the Theorem 1 of the paper supplement, the iteration cost of rSPA and DD-rSPA does not depend on the noise intensity - and this linear dependence of the overall cost on noise is solely explained by the linear increase in the number of rSPA and DD-rSPA iterations required to achieve the solution of the minimization problem (7) with the linearly reducing SNR. In another words, these results show that DL and common MS-denoising invest the same amount of work at different noise levels, whereas rSPA and DD-rSPA invest work linearly-proportional to the SNR - and increasing with the relative increase of the noise. A comparison of the denoising quality scalings in Fig. 5 provides additional evidence towards the hypothesis formulated above: deterioration of the denoising performance of DL in the area of large noise (small SNR) and smaller image sizes - where DL is getting outperformed by the 3D Wavelet-Denoising - is not the result of an insufficient training dataset or the "concept drift". It can be explained with the existence of a fundamental robustness boundary of DL denoising in the high noise regimes, with SNR<1.0. This finding is also confirmed by inspecting the performance of the DL when it is applied to the images that were previously denoised by DD-rSPA (light blue surface in Fig. 5B): this combination of unsupervised DD-rSPA followed by the supervised DL exhibits the best performance among all the considered methods in this high noise regime.

Next, from synthetic CT images generated from circles and spheres, we come back to the analysis of CT images generated from real anatomical features. Using the CT image generation and LAR-assessment pipeline, we compare the performance of denoising methods in a broad range of absorbed radiation dose densities. This comparison is made for two synthetic noise models (Fig. 6A and Fig. 6B, with Gaussian and non-Gaussian noise) and for the empirical nonparametric CT noise model obtained from the real patient data (Fig. 6C). The results of this comparison are shown in terms of three major image quality measures. As expected, the Gaussian 3D filtering exhibits the best performance among the common tools for all three additive Gaussian noise scenarios from Fig. 6A. On the other hand, the non-Gaussian deep learning DnCNN denoising outperforms the other tools (except rSPA) in the non-Gaussian and empirical noise situations, as can be seen in Figs. 6B and 6C. However, in the overall comparison, the rSPA method markedly outperforms all considered denoising tools in

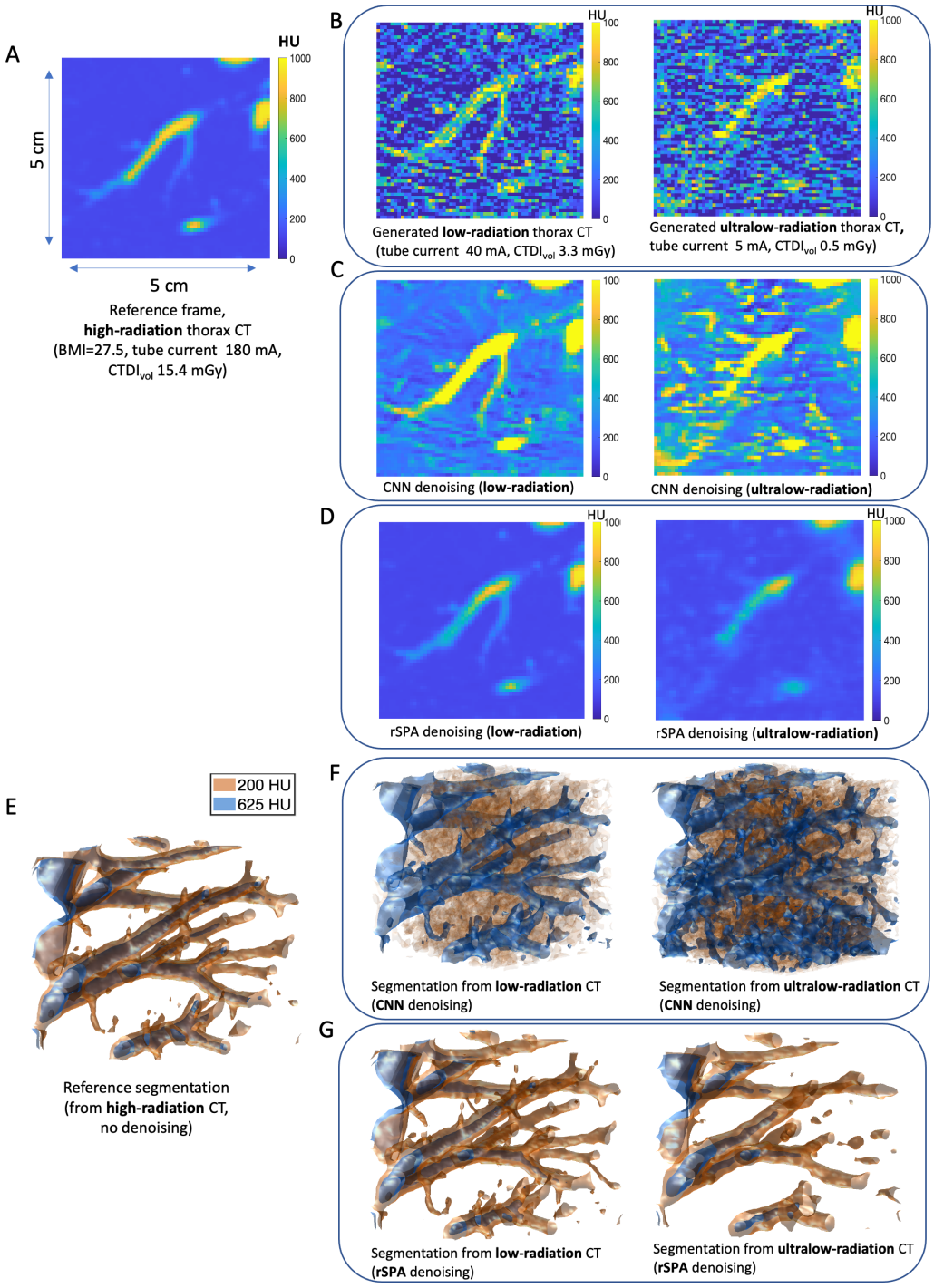

Figure 3: **Radiation exposure, quantum noise and denoising performance of CNNs and rSPA in low-radiation and ultralow-radiation thorax CT regimes** A: Reference Data of a thorax CT voxel fragment (approx. $5cm^3$) of a 19 y.o. female with the BMI 27.5, acquired with the Somatum Emotion 16 2007 (Siemens) at 130 kV tube voltage. B: Simulated decreasing of the radiation exposure $\mathbf{CTDI_{vol}}$ from 15.6 mGy (reference frame) to 3.3mGy (for low-radiation simulations) and 0.5 mGy (ultra-low-radiation) results in a significant increase of quantum noise. C: Reconstructed images using CNNs. D: Reconstructed images using rSPA. E: 3D segmentation of the original reference frame. F: 3D segmentation based on the images denoised using CNNs. G: 3D-segmentation of the images denoised by rSPA.

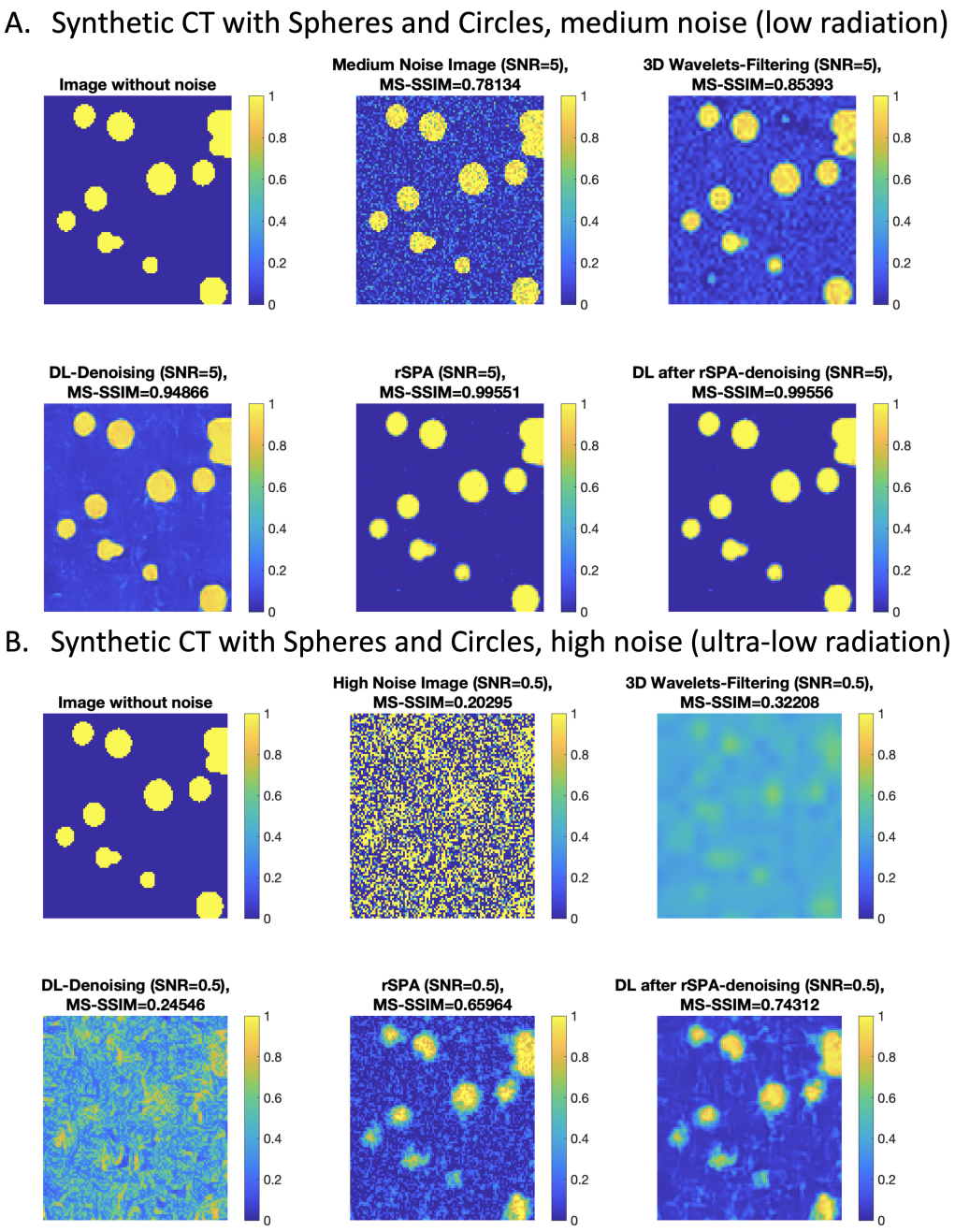

Figure 4: **Comparing denoising performance on synthetic CT images of noisy circles, with DL from Fig. 3 additionally trained to recognize circles for non-Gaussian noise model**: A: medium noise scenario, corresponding to low-radiation regime with around 3.3 mGy; B: high noise scenario, corresponding to ultra-low radiation regime with 0.5 mGy.

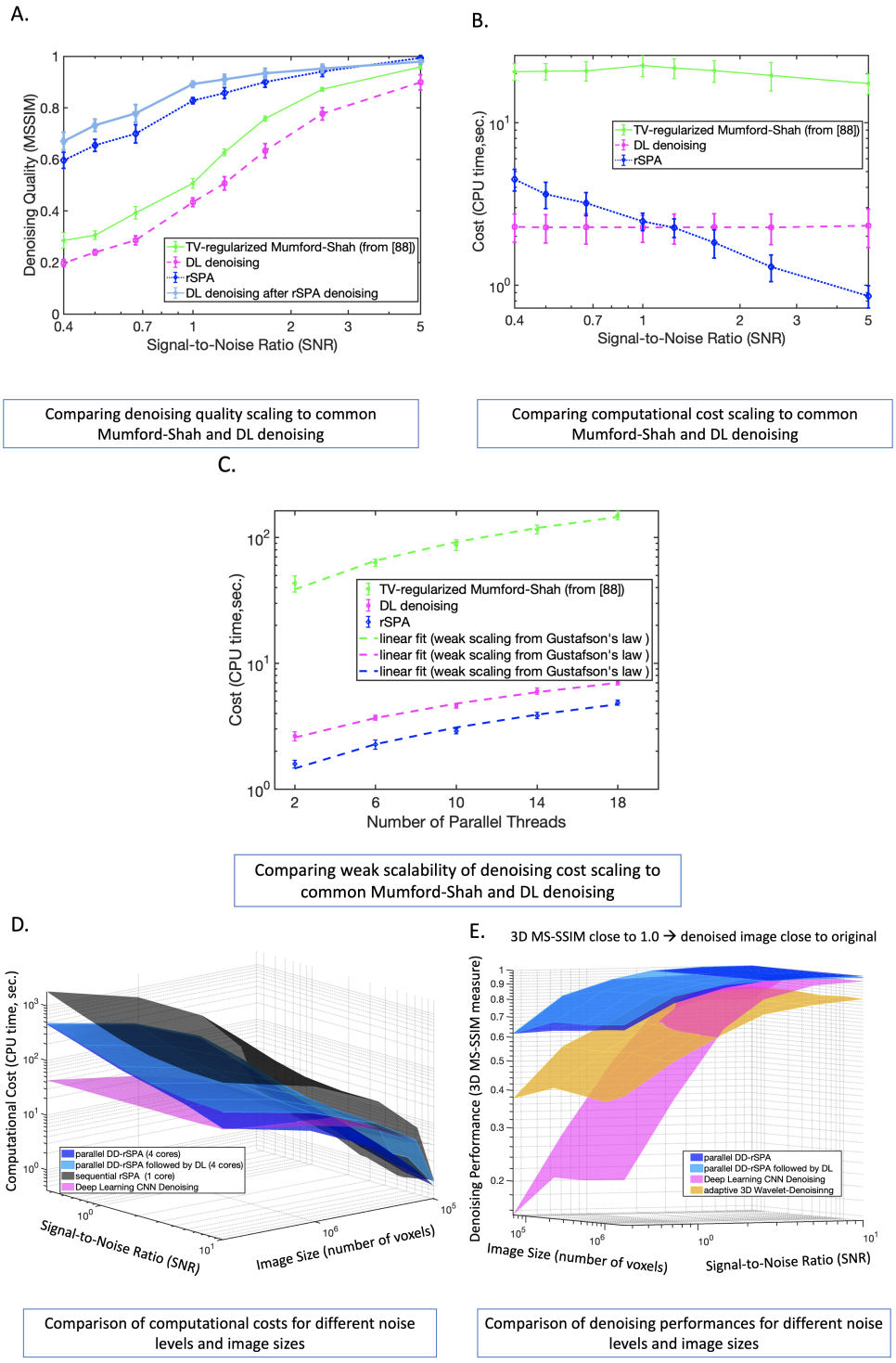

Figure 5: **Comparing denoising quality, cost and parallelizability :** A-C: comparison of PMS rSPA algorithm to the regularized Mumford-Shah denoising tool introduced in [88] and to the additionally trained DL denoising algorithm from Figs. 3 and 4; D-E: computational cost scaling and performance for DL (without taking into account time for additional training), sequential rSPA, parallel DD-rSPA and DD-rSPA followed by DL. Each point of each method's curve and surface is obtained from statistical averaging of the respective values obtained analyzing 10 randomly-generated images with these particular combinations of image size and noise level.

all image quality measures for all three noise models. As can be seen from Fig. 6, rSPA allows to achieve the same quality of the denoised image obtained with DnCNN ($\mathbf{3D\ MS-SSIM}$ around 0.9) with around 15-fold smaller absorbed radiation dose density ($\mathbf{CTDI_{vol} = 0.95mGy}$ for rSPA vs. $\mathbf{CTDI_{vol} = 15mGy}$ for DnCNN).

In Fig. 7, we compared the average denoising performances measured with the $\mathbf{3D\ MS-SSIM}$ image quality measure for a range of practically-relevant CT feature color intensity differences, lifetime attributable risks (LAR), and absorbed radiation dose densities. The results again demonstrate, that rSPA is superior to all other considered tools in all analyzed regimes. $\mathbf{3D\ MS-SSIM}$ of the blue surfaces corresponding to rSPA is close to 1.0 almost everywhere, indicating that the denoised images are very close to the reference CT images without noise. The powerful effect of image quality-preserving LAR reduction by denoising, especially in the female infants, is visible in Fig. Fig. 7B. Denoising with rSPA allows achieving the same imaging quality as using DnCNN ($\mathbf{3D\ MS-SSIM}$ around 0.97 for feature color differences around 50-100 HU) with a 22.6-fold smaller LAR ($\mathbf{LAR = 0.015\%}$ for rSPA vs. $\mathbf{LAR = 0.34\%}$ for DnCNN).

Finally, in Fig. 8 we evaluate the performance of DL with and without preliminary DD-rSPA denoising, comparing it to the denoising performance of DD-rSPA for the synthetic noisy CT images generated with real anatomic features from thorax CT. The noiseless thorax CT image used as reference in this performance comparison is available at `https://www.dropbox.com/s/29x0xivg8l80q10/female_lung_thorax_CT_image_section_v2.mat?dl=0`. The dotted lines show a 95% nonparametric confidence intervals (c.i.) obtained for every value of $\mathbf{CTDI_{vol}}$ from 100 different independently-generated noisy synthetic CT images, using the MATLAB-function *quantile()*. These results support our previous findings: applying DL to the image previously denoised with DD-rSPA provides a statistically-significant improvement of DL denoising performance.

Discussion

We introduced an algorithmic pipeline for the generation of synthetic patient-specific CT images and radiation-induced risk assessment. We used it to compare various CT image denoising approaches in

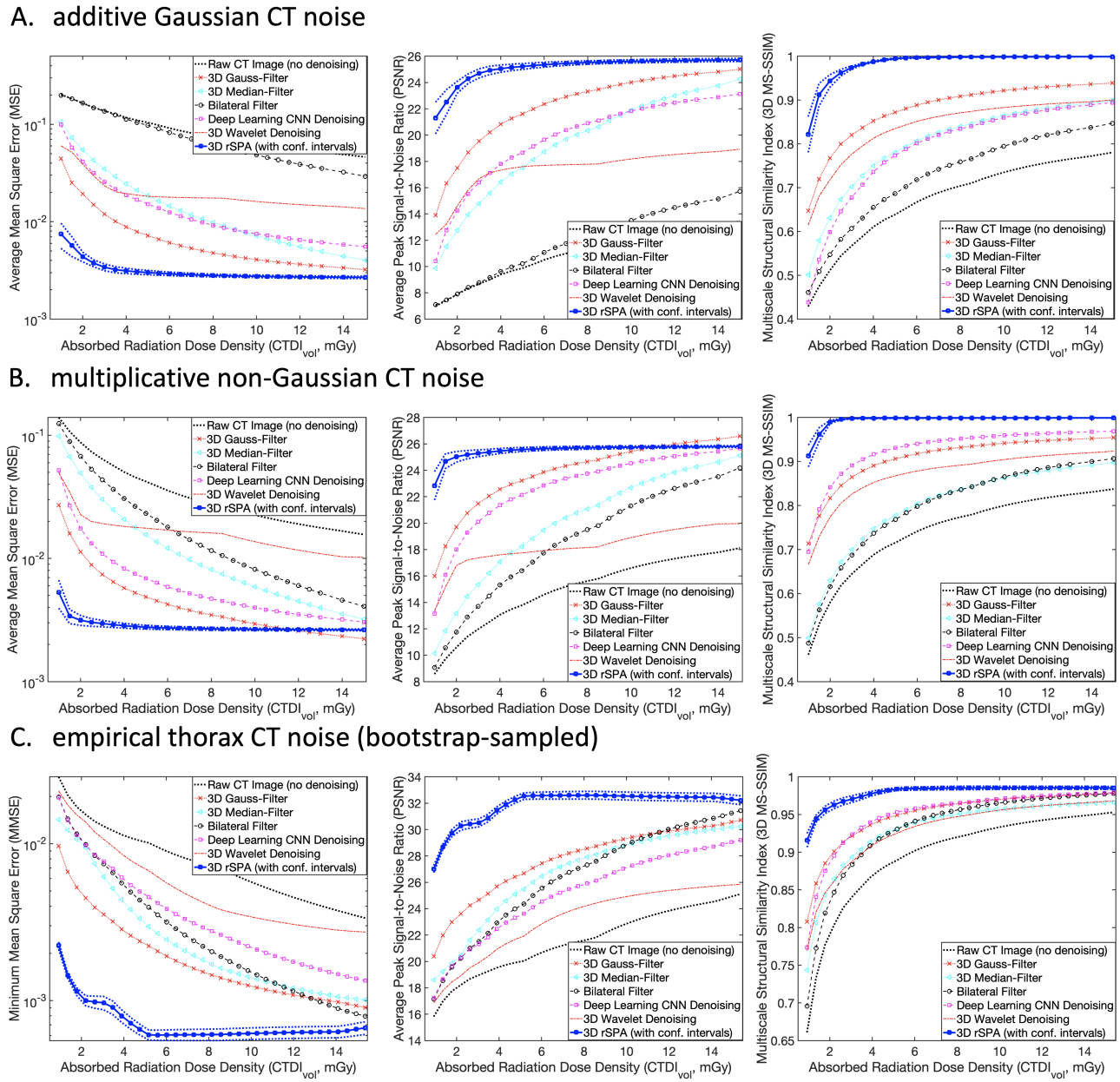

Figure 6: **Comparing CT image denoising performances for three CT noise models**: (A) additive Gaussian noise model (CT noise variance is independent of the feature color); (B) multiplicative non-Gaussian noise model (CT noise variance changes with the amplitude of the underlying color signal); (C) empirical noise obtained from the thorax CT patient data. In (A) and (B), generation of synthetic images was performed for a patient with a water-equivalent diameter of 30 cm, which is subject to a Thorax CT with a typical tube voltage of 120 kV in the range of tube currents between 5 mA-180 mA and a set of artificial anatomic features from Fig.2A (with a feature contrast of 200 HU). In (C), real patient data were used. Comparison is performed with three primary image quality criteria: (left panels) with the mean squared error; (middle panels) with the Peak Signal-to-Noise Ratio; (right panels) with the 3D Multiscale Structural Similarity Index.

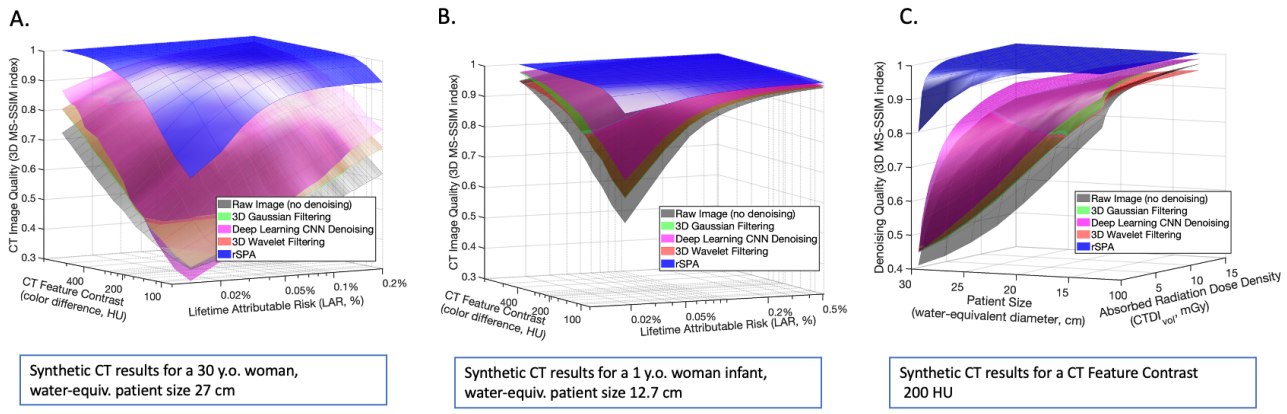

Figure 7: **Comparing denoising methods with the average Multiscale Structural Similarity Index (3D MS-SSIM):** (A) varying the true underlying feature contrast and LAR for a synthetic 30 y.o. female patient with a water-equiv. cross-section of 27cm; (B) varying the true underlying feature contrast and LAR for a synthetic 1 year old female infant patient with a water-equiv. cross-section of 12.7cm; (C) denoising performance comparison when varying the patient size and the effective absorbed radiation dose density, with the 200 Hounsfield Units (HU) feature contrast differences.

a range of practically-relevant CT regimes. The ultra-low radiation CT regime represents a three-fold challenge for all of the standard denoising methods: (i) reduction of the radiation exposure leads to a substantial increase of the noise, eventually making it impossible for standard unsupervised and spectral denoising tools (e.g., based on wavelets) to separate the noise from the underlying true image signals; (ii) heterogeneity and a high level of the individuality of anatomic features (e.g., of blood vessel networks) on a small scale - as well as the variability of patient sizes, CT conditions and a "small data challenge"- can lead to a problem of "concept drift" common for supervised methods, making the identification of some pre-trained features and patterns in the noisy CT images particularly difficult; (iii) as was shown in Fig. 3, in the ultra-low radiation regime performance of one of the most popular supervised denoising CNNs trained in a wide range of noise regimes [25] quickly deteriorates. To discern the potential impact of "concept drift" - and to rule-out the possible insufficiency of the training dataset - we additionally trained the DnCNN-3 from [25] first with 10'000 image pairs (with and without noise) - and then with 40'000 image pairs. We found that using a larger training dataset (with further 40'000 additional image pairs) only brought negligible improvements, thus confirming the earlier findings reported in [25].

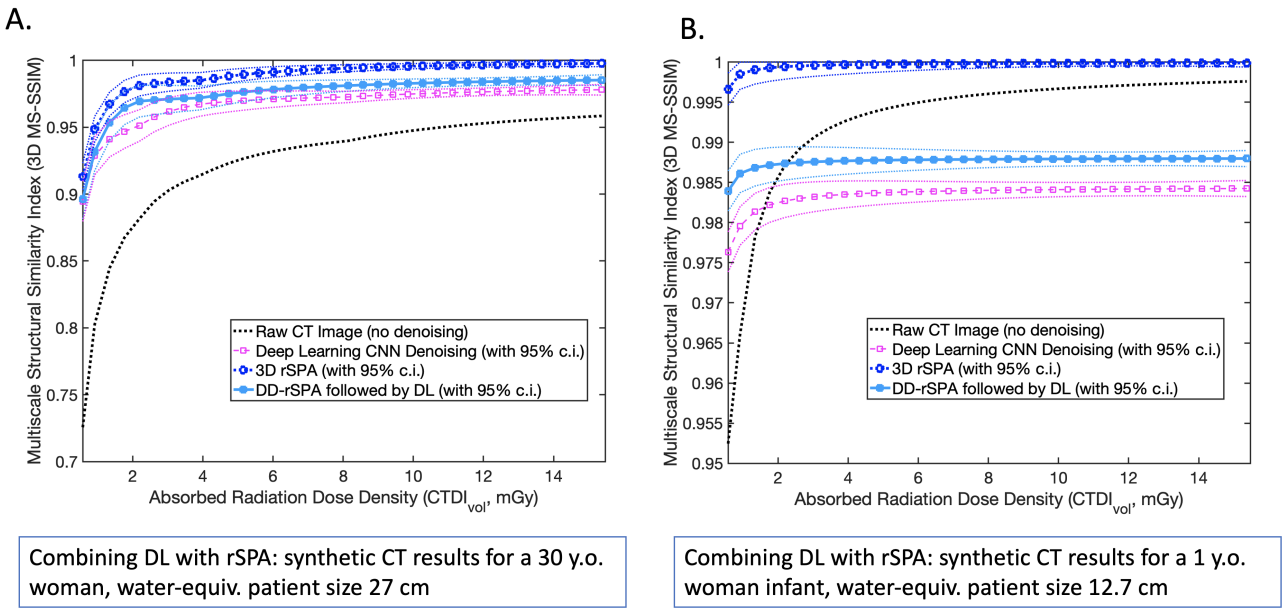

Figure 8: **Comparing denoising methods with the average Multiscale Structural Similarity Index (3D MS-SSIM) for simulated thorax CT:** (A) varying the absorbed radiation dose for a synthetic 30 y.o. female patient with a water-equiv. cross-section of 27cm; (B) varying the absorbed radiation dose for a synthetic 1 year old female infant patient with a water-equiv. cross-section of 12.7cm. Noiseless thorax CT image used as reference in this performance comparison is available at `https://www.dropbox.com/s/29x0xivg8l80q10/female_lung_thorax_CT_image_section_v2.mat?dl=0`. Dotted lines show 95% nonparametric confidence intervals (c.i.) obtained for every value of $\mathbf{CTDI_{vol}}$ from 100 different independently-generated noisy synthetic CT images, using the MATLAB-function *quantile()*.

To tackle those challenges, we introduced the Probabilistic Mumford-Shah formalism (PMS) (7) and shown that it can be efficiently solved numerically, by means of the an unsupervised regularized Scalable Probabilistic Approximation method (rSPA) that seeks a simultaneous solution of image segmentation and noise elimination problems. We could prove that it provides a computationally-cheap (with a linear cost scaling, see Fig. 5, Lemma 1-3 and Theorem 1 of the paper supplement) exact upper bound approximation of the numerically much more expensive regularized probabilistic segmentation problem (6). We also introduced DD-rSPA, a parallel extension of the rSPA algorithm based on the decomposition of the 3D-domain in overlapping subdomains (see Fig. 2 for a graphical overview, while a detailed description of the DD-rSPA algorithm is given in the Section 2 of the paper supplement). Commented code for both algorithms was provided for open access. Numerical tests with noisy images of different sizes and noise levels were summarized in the Fig. 5, revealing that : (i) the overall computational cost of both the sequential rSPA and the parallel DD-rSPA algorithms grows linearly with the image size and with the decreasing Signal-to-Noise ratios (corresponding to increasing noise levels), whereas the common Mumford-Shah and DL-denoising tools (as well as the methods like 3D wavelets denoising) "invest" the same amount of computational work independently of image SNR; (ii) the deteriorating performance of the DL denoising observed in Figs. 3,4 and 5 is neither a result of a "concept drift" (since the type of features and the noise model deployed in the training and in validation were the same) - nor a consequence of the training dataset insufficiency (since we observed only negligible performance improvements of DL when expanding the additional training data from 10K to 50K image pairs). The scaling of DL performance decay observed in Fig. 5 exhibits a much steeper robustness boundary than the Donoho-boundary [19, 20] of the wavelets denoising robustness (compare magenta and orange surfaces in Fig. 5E).

We deployed further tests that included artificial and real data, Gaussian and non-Gaussian, additive (Fig. 6A), multiplicative (Fig. 6B) and nonparametric empirical CT noise scenarios (Fig. 6C) as well as continuous and discontinuous feature boundaries. These results show that rSPA outperformed all of the other considered denoising methods in all evaluated performance measures.

The favourable linear scaling and parallelization opportunity provided by the DD-rSPA algorithm allow using a normal laptop for the tasks that would otherwise required extensive hardware (e.g.,

workstations and HPC facilities): as can be seen from the Fig. 5, DD-rSPA allows qualitative (with $\mathbf{3DMS-SSIM}$ around 0.9) denoising of a 3D image with 10^7 voxels in ultra-low radiation regime (SNR=0.5) at around 3 minutes on a MacBook Pro Laptop with 4 cores. Non of the other denoising methods tested was able to come even close to this performance.

Results summarized in Figs. 6, 7 and 8 show that using rSPA and DD-rSPA opens a possibility to gain a significant patient-specific reduction of the radiation-imposed risks, allowing an around 20-fold estimated reduction of LAR for infants and an around 10-fold LAR reduction for adults. Based on the risk assessment protocol introduced in [10], the results from Fig.7 B indicate that adopting this personalized denoising methodology for ultra-low radiation CT in the pediatric praxis might be the key to prevent around 90% of the deadly cancers induced by pediatric CTs. This could be up to 11'000 cases yearly worldwide that can be potentially prevented.

As can be seen from Figs. 4, 5 and 8, applying DL to the images previously denoised with DD-rSPA provides a statistically-significant improvement of DL denoising. This opens a possibility to boost the performance of the supervised DL and ML methods recently developed in CT imaging. Many of the existing tools were trained in the regimes with moderate and low noise levels - and preliminary unsupervised denoising with DD-rSPA can extend their applicability to the ultra-low radiation regimes with very high noise levels.

The sequential rSPA and the parallel DD-rSPA algorithms can also be directly applied to the denoising and segmentation of ultra-noisy 2D and 3D movie data from different areas. In a case of 2D movies, time axis of a movie can be considered as the a third image dimension in rSPA. Another possible application area - the 3D movies - emerge for example in fMRI applications in various biomedical areas (e.g., in cardiology), where the main challenge is detecting the moving boundary of the inner organ and distinguishing it from other eventual shapes in a time-resolved noisy dynamics [89]. Some examples of such DD-rSPA movie denoising are available at `https://www.dropbox.com/sh/n2dbl4h9p4o0p92/AABRkAalhXoaiKFO7ixsSzKga?dl=0`. Finally, beyond CT data denoising and segmentation, we also see direct application possibilities for other imaging techniques, such as: fiber-optic fluorescence imaging, diffusion tensor imaging and for large-scale 3D segmentation tasks from electron microscopy images.

Methods

Synthetic CT image generation model

To create the additive Gaussian CT noise, we used the parameter value *'gaussian'*, non-Gaussian multiplicative noise images were created using the function $imnoise()$ with the parameter value 'speckle' The variants parameter σ is in both cases selected according to the description below. MATLAB code implementing this CT image generation workflow is available at https://www.dropbox.com/sh/rr0no9vdo8osx44/AAAHQxXJnxT8P0LPs7wTRBv7a?dl=0. Generation of the nonparametric empirical CT noise was implemented in the function $create\_CT\_image\_noise()$ available at `https://www.dropbox.com/s/xbwwrk9y2napgpy/create_CT_image_noise.m?dl=0`.

Common CT image denoising and image quality assessment methods

We used the same software platform (MATLAB) and the same hardware (Mac workstation with 28 CPU cores) for all calculations to guarantee a fair comparison of the denoising methods and to rule-out the software- and platform-induced differences that can bias this comparison. All deployed common denoising and image quality assessment tools are available in the MATLAB functions from the "Image Processing", "Deep Learning", "Machine Learning" and "Wavelets" toolboxes of MathWorks. We used denoising methods based on local window filtering of the data (3D Gaussian filtering with the MATLAB-function $imgaussfilt3()$, 3D local median filtering with the MATLAB-function $medfilt3()$ and bilateral filtering with the MATLAB-function $imbilatfilt()$) [16, 17, 18, 14], spectral denoising methods (the 3D wavelets denoising with the MATLAB-function $wavedec3()$) [19, 20, 21, 22, 15] and a deep learning denoising method based on pre-trained feed-forward denoising convolutional neural networks (DnCNNs, with the MATLAB-functions $denoiseImage()$ and $denoisingNetwork()$) [25, 26, 13, 27]. For each of the considered images, the standard deviation of the local Gaussian smoothing kernel σ was changed in the range $\sigma = [0.2, 0.4, 0.6, \ldots, 2]$. The value leading to the least MSE deviation between the denoised and the original CT image was taken to com-

pute the curves in Fig. 4 and Fig.5. Similarly, for the optimal 3D wavelet filtering all of the wavelet bases available in MATLAB were checked for all of the possible depths of level decompositions - and the wavelet decomposition with the minimal MSE error was taken. Pre-training of DnCNN was done with over 20 Mio images and was provided in the "Deep Learning Toolbox". Image quality measures plotted in Fig.4 and Fig.5 were computed using the MATLAB-functions from the "Image Processing Toolbox": 3D mean-squared error (MSE) [62] was computed as the average over the 2D MSE errors obtained with the MATLAB-function $immse()$; 3D Peak Signal-to-Noise Ration (PSNR) was obtained as an average over the 2D PSNR image error measures [63] implemented in the MATLAB-function $psnr()$; 3D Multi-Scale Structural Similarity Index Measure (3D MS-SSIM) [64] with the 3D image volume measure MATLAB-function $multissim3()$.

Statistical post-processing

The curves in Figures 6-8 show averages over individual denoising results obtained for 100 different independently-generated noisy synthetic CT images that were obtained for every particular combination of tube-specific and patient specific parameters. In Figure 5, the surfaces represent averages over 10 randomly-realized noisy CT images. To provide a fair comparison, same random CT image realizations were used with every denoising method. Dotted lines in Figures 6 and 8 show 95% nonparametric confidence intervals (c.i.) computed with the MATLAB-function *quantile()*.

Data and code availability

Code is available for open access at `https://www.dropbox.com/sh/rw4t6ydkpi64w8y/AAA9katysG09w7ljsvUqPwwna?dl=0` under the BSD 3-Clause License.

Acknowledgement The authors would like to thank Piotr Didyk (USI Lugano) for helpful discussions and comments. This work was supported by the "Emergent AI Center" of the JGU Mainz (financed by the Carl-Zeiss-Stiftung) and by the Mercator Fellowship in the DFG Collaborative Research Center 1114 "Scaling Cascades in Complex Systems".

Competing Interests The authors declare that they have no competing interests.

References and Notes

[1] *Communicating radiation risks in paediatric imaging: information to support health care discussions about benefit and risk* (World Health Organization, 2016).

[2] Ai, T. *et al.* Correlation of chest CT and RT-PCR testing in Coronavirus disease 2019 (covid-19) in china: A report of 1014 cases. *Radiology* **0**, 200642 (2020). URL https://doi.org/10.1148/radiol.2020200642. PMID: 32101510, https://doi.org/10.1148/radiol.2020200642.

[3] Bernheim, A. *et al.* Chest CT findings in coronavirus disease-19 (covid-19): Relationship to duration of infection. *Radiology* **295**, 200463 (2020). URL https://doi.org/10.1148/radiol.2020200463. PMID: 32077789, https://doi.org/10.1148/radiol.2020200463.

[4] Brenner, D. J. *et al.* Cancer risks attributable to low doses of ionizing radiation: assessing what we really know. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 13761–13766 (2003). URL https://pubmed.ncbi.nlm.nih.gov/14610281.

[5] Radiation, committee and research, board and studies, division and council, national. *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2* 1–406 (2006).

[6] Brenner, D. J. & Hall, E. J. Computed tomography —an increasing source of radiation exposure. *New England Journal of Medicine* **357**, 2277–2284 (2007). URL https://doi.org/10.1056/NEJMra072149.

[7] Gillespie, D. T. A diffusional bimolecular propensity function. *The Journal of chemical physics* **131**, 164109–164109 (2009). URL https://pubmed.ncbi.nlm.nih.gov/19894929.

[8] Siegel, J. A. *et al.* The BEIR VII estimates of low-dose radiation health risks are based on faulty assumptions and data analyses: A call for reassessment. *Journal of Nuclear Medicine* **59**, 1017–1019 (2018). URL http://jnm.snmjournals.org/content/59/7/1017.abstract. http://jnm.snmjournals.org/content/59/7/1017.full.pdf+html.

[9] Berrington de González, A. *et al.* Projected Cancer Risks From Computed Tomographic Scans Performed in the United States in 2007. *Archives of Internal Medicine* **169**, 2071–2077 (2009). URL https://doi.org/10.1001/archinternmed.2009.440. https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/415368/ioi90109_2071_2077.pdf.

[10] Miglioretti, D. L. *et al.* The Use of Computed Tomography in Pediatrics and the Associated Radiation Exposure and Estimated Cancer Risk. *JAMA Pediatrics* **167**, 700–707 (2013). URL https://doi.org/10.1001/jamapediatrics.2013.311. https://jamanetwork.com/journals/jamapediatrics/articlepdf/1696279/poi130056.pdf.

[11] Duncan, J. R., Lieber, M. R., Adachi, N. & Wahl, R. L. Radiation dose does matter: Mechanistic insights into dna damage and repair support the linear no-threshold model of low-dose radiation health risks. *Journal of Nuclear Medicine* **59**, 1014–1016 (2018). URL http://jnm.snmjournals.org/content/59/7/1014.short. http://jnm.snmjournals.org/content/59/7/1014.full.pdf+html.

[12] Huang, R., Liu, X., He, L. & Zhou, P.-K. Radiation exposure associated with computed tomography in childhood and the subsequent risk of cancer: A meta-analysis of cohort studies. *Dose-response : a publication of International Hormesis Society* **18**, 1559325820923828–1559325820923828 (2020). URL https://pubmed.ncbi.nlm.nih.gov/32425727.

[13] Choy, G. *et al.* Current applications and future impact of machine learning in radiology. *Radiology* **288**, 318–328 (2018). URL https://doi.org/10.1148/radiol.2018171820. PMID: 29944078, https://doi.org/10.1148/radiol.2018171820.

[14] Koziol, P. *et al.* Comparison of spectral and spatial denoising techniques in the context of high definition ft-ir imaging hyperspectral data. *Scientific Reports* **8**, 14351 (2018). URL https://doi.org/10.1038/s41598-018-32713-7.

[15] Roels, J. *et al.* An interactive imagej plugin for semi-automated image denoising in electron microscopy. *Nature Communications* **11**, 771 (2020). URL https://doi.org/10.1038/s41467-020-14529-0.

[16] Wirjadi, O. & Breuel, T. Approximate separable 3d anisotropic gauss filter. vol. 2, 11 – 149 (2005).

[17] Tomasi, C. & Manduchi, R. Bilateral filtering for gray and color images. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV ?98, 839 (IEEE Computer Society, USA, 1998).

[18] Harms, J., Wang, T., Petrongolo, M., Niu, T. & Zhu, L. Noise suppression for dual-energy CT via penalized weighted least-square optimization with similarity-based regularization. *Medical Physics* **43**, 2676–2686 (2016).

[19] Donoho, D. L. De-noising by soft-thresholding. *IEEE Transactions on Information Theory* **41**, 613–627 (1995).

[20] Arias-Castro, E. & Donoho, D. L. Does median filtering truly preserve edges better than linear filtering? *Ann. Statist.* **37**, 1172–1206 (2009). URL https://doi.org/10.1214/08-AOS604.

[21] Wang, Y., Che, X. & Ma, S. Nonlinear filtering based on 3d wavelet transform for mri denoising. *EURASIP Journal on Advances in Signal Processing* **2012**, 40 (2012). URL https://doi.org/10.1186/1687-6180-2012-40.

[22] Tang, S. & Tang, X. Statistical CT noise reduction with multiscale decomposition and penalized weighted least squares in the projection domain. *Medical physics* **39**, 5498–5512 (2012). URL https://pubmed.ncbi.nlm.nih.gov/22957617.

[23] Yang, Q. *et al.* Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging* **37**, 1348–1357 (2018). URL https://pubmed.ncbi.nlm.nih.gov/29870364.

[24] Konefal, A. *et al.* Unpaired low-dose CT denoising network based on cycle-consistent generative adversarial network with prior image information. *Computational and Mathematical Methods in Medicine* **2019**, 8639825 (2019). URL https://doi.org/10.1155/2019/8639825.

[25] Zhang, K., Zuo, W., Chen, Y., Meng, D. & Zhang, L. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *Trans. Img. Proc.* **26**, 3142–3155 (2017). URL https://doi.org/10.1109/TIP.2017.2662206.

[26] Chen, H. *et al.* Low-dose CT via convolutional neural network. *Biomedical optics express* **8**, 679–694 (2017). URL https://pubmed.ncbi.nlm.nih.gov/28270976.

[27] Topal, E., Löffler, M. & Zschech, E. Deep learning-based inaccuracy compensation in reconstruction of high resolution xCT data. *Scientific reports* **10**, 7682–7682 (2020). URL https://pubmed.ncbi.nlm.nih.gov/32376852.

[28] Tian, C. *et al.* Deep learning on image denoising: An overview. *Neural Networks* (2020).

[29] Kaur, P., Singh, G. & Kaur, P. A review of denoising medical images using machine learning approaches. *Current medical imaging* **14**, 675–685 (2018).

[30] Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical image analysis* **42**, 60–88 (2017).

[31] Lundervold, A. S. & Lundervold, A. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik* **29**, 102–127 (2019).

[32] Razzak, M. I., Naz, S. & Zaib, A. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps* 323–350 (2018).

[33] Liu, S. & Deng, W. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, 730–734 (IEEE, 2015).

[34] Gaonkar, B., Hovda, D., Martin, N. & Macyszyn, L. Deep learning in the small sample size setting: cascaded feed forward neural networks for medical image segmentation. In *Medical Imaging 2016: Computer-Aided Diagnosis*, vol. 9785, 97852I (International Society for Optics and Photonics, 2016).

[35] Zhao, W. Research on the deep learning of the small sample data based on transfer learning. In *AIP Conference Proceedings*, vol. 1864, 020018 (AIP Publishing LLC, 2017).

[36] Keshari, R., Ghosh, S., Chhabra, S., Vatsa, M. & Singh, R. Unravelling small sample size problems in the deep learning world. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 134–143 (IEEE, 2020).

[37] D'souza, R. N., Huang, P.-Y. & Yeh, F.-C. Structural analysis and optimization of convolutional neural networks with a small sample size. *Scientific reports* **10**, 1–13 (2020).

[38] Dietterich, T. Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)* **27**, 326–327 (1995).

[39] Zhang, C., Vinyals, O., Munos, R. & Bengio, S. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893* (2018).

[40] Rice, L., Wong, E. & Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, 8093–8104 (PMLR, 2020).

[41] Hosseini, M. *et al.* I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews* (2020).

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

[43] Ying, X. An overview of overfitting and its solutions. In *Journal of Physics: Conference Series*, vol. 1168, 022022 (IOP Publishing, 2019).

[44] Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**, 1345–1359 (2009).

[45] Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *Journal of Big data* **3**, 1–40 (2016).

[46] Jang, Y., Lee, H., Hwang, S. J. & Shin, J. Learning what and where to transfer. In *International Conference on Machine Learning*, 3030–3039 (PMLR, 2019).

[47] Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: Understanding transfer learning for medical imaging. *arXiv preprint arXiv:1902.07208* (2019).

[48] Alzubaidi, L. *et al.* Towards a better understanding of transfer learning for medical imaging: a case study. *Applied Sciences* **10**, 4523 (2020).

[49] Alzubaidi, L. *et al.* Novel transfer learning approach for medical imaging with limited labeled data. *Cancers* **13**, 1590 (2021).

[50] Tartaglione, E., Barbano, C. A., Berzovini, C., Calandri, M. & Grangetto, M. Unveiling covid-19 from chest x-ray with deep learning: a hurdles race with small data. *International Journal of Environmental Research and Public Health* **17**, 6933 (2020).

[51] Tsymbal, A. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin* **106**, 58 (2004).

[52] Žliobaitė, I. Learning under concept drift: an overview. *arXiv preprint arXiv:1010.4784* (2010).

[53] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. & Bouchachia, A. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* **46**, 1–37 (2014).

[54] Alippi, C. Learning in non-stationary environments. In Rosa, A. C., Guervós, J. J. M. & Filipe, J. (eds.) *ECTA 2014 - Proceedings of the International Conference on Evolutionary Computation Theory and Applications, part of IJCCI 2014, Rome, Italy, 22-24 October, 2014*, IS–11 (SciTePress, 2014).

[55] Souza, V. M. A., dos Reis, D. M., Maletzke, A. & Batista, G. E. A. P. A. Challenges in benchmarking stream learning algorithms with real-world data. *Data Mining and Knowledge Discovery* **34**, 1805–1858 (2020). URL https://doi.org/10.1007/s10618-020-00698-5.

[56] Horenko, I. On a scalable entropic breaching of the overfitting barrier for small data problems in machine learning. *Neural Computation* **0**, 1–17 (0). URL https://doi.org/10.1162/neco_a_01296. PMID: 32521216, https://doi.org/10.1162/neco_a_01296.

[57] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–80 (1997).

[58] Horenko, I. Finite element approach to clustering of multidimensional time series. *SIAM Journal on Scientific Computing* **32**, 62–83 (2010).

[59] Metzner, P., Putzig, L. & Horenko, I. Analysis of persistent nonstationary time series and applications. *Commun. Appl. Math. Comput. Sci.* **7**, 175–229 (2012). URL https://doi.org/10.2140/camcos.2012.7.175.

[60] Gerber, S. & Horenko, I. Improving clustering by imposing network information. *Science Advances* **1** (2015). URL https://advances.sciencemag.org/content/1/7/e1500163. https://advances.sciencemag.org/content/1/7/e1500163.full.pdf.

[61] Pospisil, L., Gagliardini, P., Sawyer, W. & Horenko, I. On a scalable nonparametric denoising of time series signals. *Commun. Appl. Math. Comput. Sci.* **13**, 107–138 (2018). URL https://doi.org/10.2140/camcos.2018.13.107.

[62] Wackerly, D. D., III, W. M. & Scheaffer, R. L. *Mathematical Statistics with Applications* (Duxbury Advanced Series, 2002), sixth edition edn.

[63] Huynh-Thu, Q. & Ghanbari, M. The accuracy of psnr in predicting video quality for different video scenes and frame rates. *Telecommunication Systems* **49**, 35–48 (2012). URL https://doi.org/10.1007/s11235-010-9351-x.

[64] Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *Trans. Img. Proc.* **13**, 600?612 (2004). URL https://doi.org/10.1109/TIP.2003.819861.

[65] Hallgrímsson, B. & Hall, B. *Variation: A Central Concept in Biology* (Elsevier Science, 2011). URL https://books.google.de/books?id=3b4pToaZ3JEC.

[66] Meyer, M. *et al.* Reproducibility of ct radiomic features within the same patient: Influence of radiation dose and ct reconstruction settings. *Radiology* **293**, 583–591 (2019). URL https://doi.org/10.1148/radiol.2019190928. PMID: 31573400, https://doi.org/10.1148/radiol.2019190928.

[67] De Man, B. *et al.* Catsim: a new computer assisted tomography simulation environment. In *Medical Imaging 2007: Physics of Medical Imaging*, vol. 6510, 65102G (International Society for Optics and Photonics, 2007).

[68] Yu, L., Shiung, M., Jondal, D. & McCollough, C. H. Development and validation of a practical lower-dose-simulation tool for optimizing computed tomography scan protocols. *Journal of computer assisted tomography* **36**, 477–487 (2012).

[69] McCollough, C. H. *et al.* Low-dose ct for the detection and classification of metastatic liver lesions: results of the 2016 low dose ct grand challenge. *Medical physics* **44**, e339–e352 (2017).

[70] Moen, T. R. *et al.* Low-dose ct image and projection dataset. *Medical physics* **48**, 902–911 (2021).

[71] Li, X. & Samei, E. Comparison of patient size-based methods for estimating quantum noise in CT images of the lung. *Medical physics* **36**, 541–6 (2009).

[72] Solomon, J., Lyu, P., Marin, D. & Samei, E. Noise and spatial resolution properties of a commercially available deep?learning based ct reconstruction algorithm. *Medical Physics* (2020).

[73] Samei, E., Kinahan, P., Nishikawa, R. M. M. & Maidment, A. Virtual Clinical Trials: Why and What (Special Section Guest Editorial). *Journal of Medical Imaging* **7**, 1 – 5 (2020). URL https://doi.org/10.1117/1.JMI.7.4.042801.

[74] Anam, C. *et al.* Volume computed tomography dose index (CTdivol) and size-specific dose estimate (ssde) for tube current modulation (tcm) in CT scanning. *International Journal of Radiation Research* **16**, 289–297 (2018). URL https://iranjournals.nlai.ir/1271/article_335823.html.

[75] Koyuncu, H. & Ceylan, R. Elimination of white Gaussian noise in arterial phase CT images to bring adrenal tumours into the forefront. *Computerized Medical Imaging and Graphics* **65**, 46 – 57 (2018). URL http://www.sciencedirect.com/science/article/pii/S0895611117300472. Advances in Biomedical Image Processing.

[76] Council, N. R. *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2* (The National Academies Press, Washington, DC, 2006). URL https://www.nap.edu/catalog/11340/health-risks-from-exposure-to-low-levels-of-ionizing-radiation.

[77] Bezdek, J. C., Ehrlich, R. & Full, W. FCM: The fuzzy c-means clustering algorithm. *Computers and Geosciences* **10**, 191–203 (1984).

[78] Höppner, F., Klawonn, F., Kruse, R. & Runkler, T. *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition* (John Wiley & Sons, 1999).

[79] Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* **31**, 651 – 666 (2010). URL http://www.sciencedirect.com/science/article/pii/S0167865509002323. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[80] Mumford, D. & Shah, J. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. on Pure and Applied Mathematics* **42**, 577–685 (1989). URL http://doi.wiley.com/10.1002/cpa.3160420503.

[81] Gerber, S., Pospisil, L., Navandar, M. & Horenko, I. Low-cost scalable discretization, prediction, and feature selection for complex systems. *Science Advances* **6** (2020). URL https://advances.sciencemag.org/content/6/5/eaaw0961. https://advances.sciencemag.org/content/6/5/eaaw0961.full.pdf.

[82] de Wiljes, J., Majda, A. & Horenko, I. An adaptive Markov chain Monte Carlo approach to time series clustering of processes with regime transition behavior. *SIAM Multiscale Model. Simul.* **11**, 415–441 (2013).

[83] Rudin, L. I., Osher, S. & Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena* **60**, 259–268 (1992). URL https://www.sciencedirect.com/science/article/pii/016727899290242F.

[84] Chambolle, A. Finite-differences discretizations of the mumford-shah functional. *ESAIM: Mathematical Modelling and Numerical Analysis* **33**, 261–288 (1999).

[85] Lysaker, O. M., Lundervold, A. & Tai, X. Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* **12 12**, 1579–90 (2003).

[86] Chan, T. F. & Shen, J. *Image processing and analysis: variational, PDE, wavelet, and stochastic methods* (SIAM, 2005).

[87] Pock, T., Cremers, D., Bischof, H. & Chambolle, A. An algorithm for minimizing the mumford-shah functional. In *2009 IEEE 12th International Conference on Computer Vision*, 1133–1140 (2009).

[88] Hohm, K., Storath, M. & Weinmann, A. An algorithmic framework for mumford–shah regularization of inverse problems in imaging. *Inverse Problems* **31**, 115011 (2015).

[89] Paragios, N., Duncan, J. & Ayache, N. *Handbook of Biomedical Imaging: Methodologies and Clinical Research* (Springer, 2015).

Supplementary material for the manuscript entitled: Quality-preserving low-cost probabilistic 3D denoising with applications to Computed Tomography

Illia Horenko, Lukas Pospisil, Edoardo Vecchi, Steffen Albrecht, Alexander Gerber, Beate Rehbock, Albrecht Stroh, and Susanne Gerber

This document provides supplementary material for the manuscript entitled: Order of magnitude risk reduction in Computed Tomography with the unsupervised machine learning denoising. In particular, we provide the complete mathematical formulation of 3D regularized Scalable Probabilistic Approximation (rSPA) optimization problem and present:

- **Lemma 1** - derivation of the rSPA problem formulation,
- **Algorithm 1** - a subspace algorithm for solving rSPA optimization problem in pseudo-code. The algorithm consists of two consequent inner optimization problems; namely so-called C-problem and Γ-problem.
- **Lemma 2** - the solvability and the computational cost of solving the C-problem,
- **Lemma 3** - the solvability and the computational cost of solving the Γ-problem,
- **Theorem 1** - the properties and the computational cost of solving rSPA problem. The proof is based on Lemma 2 and Lemma 3.

1 Regularized Scalable Probabilistic Approximation Algorithm (rSPA)

Formulation: Let $t \in \mathbb{N}^3$ be a multi-index of 3D voxel coordinates and

$$V := [V(1), \ldots, V(T)] \in \mathbb{R}^{D,T}$$

be a 3D CT image represented as a matrix of given D-dimensional voxel colours at 3D coordinates $X := [X(1), \ldots, X(T)] \in \mathbb{R}^{3,T}$.

We will be searching for a probabilistic approximation $\tilde{V}_{C,\Gamma}$ of the image in terms of K latent features characterized by K distinct color vectors $\{C_{1,k}, \ldots, C_{D,k}\}$, with k taking values between 1 and K. Spatial characteristics of these K latent features that we will be searching for will be provided by (a priori unknown) latent feature probabilities $\Gamma_k(t)$, being the probabilities of an actual (noisy) voxel $V(t)$ to belong to a particular latent (noiseless) feature with an index k:

$$\tilde{V}_{C,\Gamma} := [\sum_{k=1}^{K} \Gamma_k(1) C_k, \ldots, \sum_{k=1}^{K} \Gamma_k(T) C_k] \in \mathbb{R}^{D,T}$$

Then, following the idea behind the Mumford-Shah functional formulation [6], spatially-persistent optimal probabilistic approximation $\tilde{V}_{C^*,\Gamma^*}$ of the original image data V can be computed via the numerical minimization of the function:

$$\begin{aligned}
[C^*, \Gamma^*] &= \arg\min_{C,\Gamma} \min_{\Gamma \in \Omega_\Gamma, C \in \Omega_C} \mathcal{L}(C, \Gamma) = \\
&= \arg\min_{C,\Gamma} \min_{\Gamma \in \Omega_\Gamma, C \in \Omega_C} \left[\frac{1}{T} \sum_{t=1}^{T} \mathrm{dist}^2(V(t), \tilde{V}_{C,\Gamma}(t)) + \frac{\bar{\varepsilon}}{\sum_{t,t'=1}^{T} \alpha_{t,t'}} \sum_{t,t'=1}^{T} \alpha_{t,t'} \mathrm{dist}^2(\tilde{V}_{C,\Gamma}(t), \tilde{V}_{C,\Gamma}(t')) \right],
\end{aligned} \tag{1}$$

where $\mathrm{dist}^2(\cdot,\cdot)$ is a square of some distance (e.g., Euclidean distance, l_1-distance, etc.), feasible sets are given as

$$\Omega_\Gamma = \{\Gamma = [\Gamma(1), \ldots, \Gamma(T)] \in \mathbb{R}^{K,T} : \quad \Gamma \geq 0 \text{ and } \forall t : \sum_{k=1}^{K} \Gamma_k(t) = 1\}, \tag{2}$$

$$\Omega_C = \{C = [C_1, \ldots, C_K] \in \mathbb{R}^{D,K} : \quad \min(V) \leq C \leq \max(V)\}, \tag{3}$$

and function $\alpha_{t,t'}$ is the indicator function of the voxel neighborhood defined as

$$\alpha_{t,t'} = \begin{cases} 1 & \text{if } \mathrm{dist}(X(t), X(t')) \leq \alpha_0, \\ 0 & \text{if } \mathrm{dist}(X(t), X(t')) > \alpha_0, \end{cases}$$

with $\bar{\varepsilon} \geq 0$ and $\alpha_0 > 0$ being some user-defined parameters.

Lemma 1. *(rSPA as an approximate upper bound formulation for probabilistic segmentations with Euclidean distance) Approximate solutions of the problem (1,2,3) with Eulidean distance*

$$\mathrm{dist}(x, y) = \|x - y\|_2, \tag{4}$$

can be found minimizing its upper bound

$$\begin{aligned}
[C^*, \Gamma^*] &= \arg\min_{C,\Gamma} \min_{\Gamma \in \Omega_\Gamma, C \in \Omega_C} L(C, \Gamma) = \\
&= \arg\min_{C,\Gamma} \min_{\Gamma \in \Omega_\Gamma, C \in \Omega_C} \sum_{k=1}^{K} \left[\frac{1}{T} \sum_{t=1}^{T} \Gamma_k(t) \|V(t) - C_k\|_2^2 + \frac{\bar{\varepsilon} \|C_k\|_2^2}{\sum_{t,t'=1}^{T} \alpha_{t,t'}} \sum_{t,t'=1}^{T} \alpha_{t,t'} (\Gamma_k(t) - \Gamma_k(t'))^2 \right].
\end{aligned} \tag{5}$$

Moreover, $L(C, \Gamma) \geq \mathcal{L}(C, \Gamma)$ *(for all* C, Γ *).*

Proof. Since the square of any norm is a convex function and for any $t = 1, \ldots, T$ coefficients $\Gamma_k(t)$ forms the coefficients of convex combination, we can apply Jensens inequality to the first term of (1) and obtain

$$\mathrm{dist}^2(V(t), \tilde{V}_{C,\Gamma}(t)) = \left\| V(t) - \sum_{k=1}^{K} \Gamma_k(t) C_k \right\|_2^2 = \left\| \sum_{k=1}^{K} \Gamma_k(t) \left(V(t) - C_k \right) \right\|_2^2 \leq \sum_{k=1}^{K} \Gamma_k(t) \left\| V(t) - C_k \right\|_2^2 .$$

In the case of the second term, we use the properties of the norm

$$\begin{aligned} \forall u, v \in \mathcal{V}: \quad & \|u + v\| \leq \|u\| + \|v\|, \\ \forall u \in \mathcal{V}, \forall \alpha \in \mathbb{R}: \quad & \|\alpha u\| = |\alpha| \cdot \|u\|, \end{aligned}$$

and we get

$$\begin{aligned} \mathrm{dist}^2(\tilde{V}_{C,\Gamma}(t), \tilde{V}_{C,\Gamma}(t')) &= \left\| \left(\sum_{k=1}^{K} \Gamma_k(t) C_k \right) - \left(\sum_{k=1}^{K} \Gamma_k(t') C_k \right) \right\|_2^2 = \left\| \sum_{k=1}^{K} \left(\Gamma_k(t) - \Gamma_k(t') \right) C_k \right\|_2^2 \\ &\leq \sum_{k=1}^{K} \left\| \left(\Gamma_k(t) - \Gamma_k(t') \right) C_k \right\|_2^2 \leq \sum_{k=1}^{K} \left(\Gamma_k(t) - \Gamma_k(t') \right)^2 \left\| C_k \right\|_2^2 . \end{aligned}$$

□

Algorithm: Approximate solutions of the optimization problem (5,2,3) can be found using the iterative subspace algorithm, i.e., it is solved as a sequence of split optimization problems, see Algorithm 1.

Let V be given voxel data, K be a fixed number of latent features, and $\bar{\varepsilon} \geq 0$ be a priori chosen regularization parameter.
Choose a feasible initial approximation $\Gamma^0 \in \Omega_\Gamma$ and set iteration counter $\mathrm{it} = 0$.

while $\|L(C, \Gamma^{\mathrm{it}}) - L(C^{\mathrm{it}-1}, \Gamma^{\mathrm{it}-1})\|$ *is not sufficiently small*

*solve the problem with fixed $\Gamma^{\mathrm{it}-1}$ (**C-problem**)*

$$C^{\mathrm{it}} = \arg\min_{C} \min_{C \in \Omega_C} L(C, \Gamma^{\mathrm{it}-1}) \tag{6}$$

*solve the problem with fixed C^{it} (**Γ-problem**)*

$$\Gamma^{\mathrm{it}} = \arg\min_{\Gamma} \min_{\Gamma \in \Omega_\Gamma} L(C^{\mathrm{it}}, \Gamma) \tag{7}$$

$\mathrm{it} = \mathrm{it} + 1$
endwhile

Return an approximation of the latent features color C^{it} and an approximation of latent feature affiliation probability vectors Γ^{it}.

Algorithm 1: Regularized Scalable Probabilistic Approximation algorithm (rSPA).

Lemma 2. *(The properties of C-problem* (6)*)*

1. *the optimization problem* (6) *has always solution,*
2. (6) *is a box-constrained convex Quadratic Programming problem (QP) with diagonal Hessian matrix and it has analytical solution,*
3. *evaluation of analytical solution for solving problem* (6) *is* $\mathcal{O}(TKD)$.

Proof.

1. Let $\Gamma = \hat{\Gamma}$ be fixed. We are dealing with minimization problem with continuous convex objective function on closed set, therefore by Weierstrass Extreme value theorem [3], the problem has always solution.

2. The objective function of problem (5) with Euclidean measure (4) is given by

$$L(C,\hat{\Gamma}) = \sum_{k=1}^{K}\left[\left(\frac{1}{T}\sum_{t=1}^{T}\hat{\Gamma}_k(t)\|V(t)-C_k\|_2^2\right)+\kappa_k\|C_k\|^2\right] \tag{8}$$

where we denoted constant

$$\kappa_k = \frac{\bar{\varepsilon}}{\sum_{t,t'=1}^{T}\alpha_{t,t'}}\sum_{t,t'=1}^{T}\alpha_{t,t'}(\hat{\Gamma}_k(t)-\hat{\Gamma}_k(t'))^2, \quad k=1,\dots,K. \tag{9}$$

Since the minimization of (8) with respect to (3) is separable in $k=1,\dots,K$, we have

$$C_k^* = \arg\min_{C_k}\min_{C_k\in\Omega_{C_k}}\left(\frac{1}{T}\sum_{t=1}^{T}\hat{\Gamma}_k(t)\|V(t)-C_k\|_2^2\right)+\kappa_k\|C_k\|^2 = \arg\min_{C_k}\min_{C_k\in\Omega_{C_k}}\underbrace{\frac{1}{2}C_k^TA_kC_k - C_k^Tb_k}_{=f_k(C_k)} \tag{10}$$

with

$$\Omega_{C_k} = \{C_k\in\mathbb{R}^D: \quad \min(V)\le C_k\le\max(V)\}$$

and

$$A_k = \sigma_kI_D, \qquad \sigma_k = \left(\frac{1}{T}\sum_{t=1}^{T}\hat{\Gamma}_k(t)\right)+\kappa_k, \qquad b_k = \sum_{t=1}^{T}\hat{\Gamma}_k(t)V(t), \tag{11}$$

where $I_D\in\mathbb{R}^D$ is identity matrix. Please, notice that for any non-empty cluster (i.e., $\sum_{t=1}^{T}\Gamma_k(t)>0$) $\sigma_k>0$ and therefore (10) is stricly convex optimization problem on closed convex set and consequently (10) has unique solution. If the cluster is empty, then (10) can be simplified to

$$C_k^* = \arg\min_{C_k}\min_{C_k\in\Omega_{C_k}}0,$$

which has infinite number of solutions, i.e., any $C_k^*\in\Omega_{C_k}$ solves the problem.

The problem (10) is (again) separable in $d=1,\dots,D$ and we can write

$$\begin{aligned} C_{d,k}^* &= \arg\min_{C_{d,k}}\min_{C_{d,k}\in\Omega_{C_{d,k}}}\tfrac{1}{2}\sigma_kC_{d,k}^2-b_{d,k}C_{d_k} = \arg\min_{C_{d,k}}\min_{C_{d,k}\in\Omega_{C_{d,k}}}\tfrac{1}{2}C_{d,k}^2-\tfrac{b_{d,k}}{\sigma_k}C_{d_k} \\ &= \arg\min_{C_{d,k}}\min_{C_{d,k}\in\Omega_{C_{d,k}}}\|C_{d,k}-\tfrac{b_{d,k}}{\sigma_k}\|_2 = P_{\Omega_{C_{d,k}}}\left(\tfrac{b_{d,k}}{\sigma_k}\right) \end{aligned}$$

with interval

$$\Omega_{C_{d,k}} = [\min(V),\max(V)]\subset\mathbb{R}$$

and projection onto this interval $P_{\Omega_{C_{d,k}}}:\mathbb{R}\to\Omega_{C_{d,k}}$ given by

$$P_{\Omega_{C_{d,k}}}(\tau) := \min\{\max(V),\max\{\min(V),\tau\}\}. \tag{12}$$

3. The computation of K sums of T vectors of dimension D in (11) and (9) followed by the computation of projection (12) is $\mathcal{O}(TKD)$.

□

Lemma 3. *(The properties of Γ-problem (7))*

1. *Problem (7) is a convex QP on separable simplexes (i.e., with bound inequality and linear equality constraints).*
2. *Assembling the objects of problem (7) is $\mathcal{O}(TKD)$.*
3. *One iteration of Spectral Projected Gradient method for QP (SPG-QP, [7]) for solving problem (7) is $\mathcal{O}(TK)$.*

Proof.

1. Let $C = \hat{C}$ be fixed. The objective function of problem (5) with Euclidean measure (4) is given by

$$L(\hat{C}, \Gamma) = \sum_{k=1}^{K} \left[-\left(\sum_{t=1}^{T} \hat{\Gamma}_k(t) w_{k,t} \right) + \xi_k \sum_{t,t'=1}^{T} \alpha_{t,t'} \left(\Gamma_k(t) - \Gamma_k(t') \right)^2 \right] \tag{13}$$

with constants

$$w_{k,t} = -\frac{1}{T} \|V(t) - \hat{C}_k\|_2^2, \qquad \xi_k = \frac{\bar{\varepsilon} \operatorname{dist}^2(C_k, 0)}{\sum_{t,t'=1}^{T} \alpha_{t,t'}}. \tag{14}$$

In the following, we simplify the objective function (13) into the standard QP form.

Let us denote the diagonalization of matrix into vector

$$\gamma = \operatorname{vec}(\Gamma^T),$$

and introduce multi-index (t, k) of vector γ by

$$\gamma_{(t,k)} = \gamma_{(k-1)T+t} = \Gamma_k(t),$$

where γ_j is j-th component of $\gamma \in \mathbb{R}^{KT}$. At first, notice that quadratic term

$$\alpha_{t,t'} (\Gamma_k(t) - \Gamma_k(t'))^2 = \alpha_{t,t'} [\Gamma_k(t), \Gamma_k(t')] \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \Gamma_k(t) \\ \Gamma_k(t') \end{bmatrix} = \gamma_{(:,k)}^T H(t, t') \gamma_{(:,k)},$$

where $\gamma_{(:,k)} = [\gamma_{(1,k)}, \ldots, \gamma_{(T,k)}]^T \in \mathbb{R}^T$ and the components of matrix $H(t, t') \in \mathbb{R}^{T,T}$ are given by

$$H_{t_1,t_2}(t, t') = \begin{cases} \alpha_{t,t'} & \text{if } t_1 = t \text{ and } t_2 = t, \\ \alpha_{t,t'} & \text{if } t_1 = t' \text{ and } t_2 = t', \\ -\alpha_{t,t'} & \text{if } t_1 = t \text{ and } t_2 = t', \\ -\alpha_{t,t'} & \text{if } t_1 = t' \text{ and } t_2 = t, \\ 0 & \text{elsewhere.} \end{cases}$$

Using this notation, we are able to simplify the quadratic term of (13) into

$$\sum_{k=1}^{K} \xi_k \sum_{t=1}^{T} \sum_{t'=1}^{T} \alpha_{t,t'} (\Gamma_k(t) - \Gamma_k(t'))^2 = \sum_{k=1}^{K} \xi_k \gamma_{(:,k)}^T \underbrace{\left(\sum_{t=1}^{T} \sum_{t'=1}^{T} H(t, t') \right)}_{=\hat{A}} \gamma_{(:,k)} = \frac{1}{2} \gamma^T \underbrace{\left(\Xi \otimes \left(2\hat{A} \right) \right)}_{=A} \gamma,$$

where $\Xi = \operatorname{diag}(\xi_1, \ldots, \xi_K) \in \mathbb{R}^{K,K}$ is diagonal matrix and $\otimes$ denotes matrix Kronecker product. Matrix $A \in \mathbb{R}^{KT,KT}$ is a block-diagonal matrix of K diagonal blocks $2\xi_k \hat{A} \in \mathbb{R}^{T,T}$. Let us remark that the matrix

$$\sum_{t'=1}^{T} H(t, t')$$

forms the Laplace matrix corresponding to graph of neighborhood of vortex t (stencil). Consequently, matrix $\hat{A}$ is composed from contributions from all stencils constructed in vortexes in the system. Such a matrix is symmetric positive semidefinite.

The objective function can be written in the form of convex quadratic function

$$L(\hat{S}, \Gamma) = \frac{1}{2} \gamma^T A \gamma - w^T \gamma. \tag{15}$$

The feasible set (2) defines the lower bound constraints and equality constraints of the optimization problem. This feasible set is closed and convex, the objective function (15) is continuous, therefore using the Weierstrass Extreme value theorem [3], the optimization problem has always solution.

2. Before solving the QP problem (15), we assemble the Hessian matrix A, linear term b, and constraints Ω_Γ (2). The assembly of the linear term (14), where one has to sum TK values of distance functions between vectors of dimension D. If the complexity of chosen distance function evaluation is $\mathcal{O}(D)$, then the overall complexity is $\mathcal{O}(TKD)$.

3. Spectral Projected Gradient method for QP (SPG-QP, [7]) is an iterative algorithm for solving minimization problems of convex quadratic function $f(x) := \frac{1}{2}x^T Ax - b^T x, f : \mathbb{R}^n \to \mathbb{R}$ on closed convex feasible set $\Omega \subset \mathbb{R}^n$ defined by separable constraints with simple projections

$$P_\Omega(x) = \arg\min_{y\in\Omega} \|x - y\|. \tag{16}$$

From the initial approximation $x^0 \in \Omega$, the process is generating the approximations x^{it} by

$$x^{\mathrm{it}+1} = x^{\mathrm{it}} + \beta_{\mathrm{it}} d^{\mathrm{it}}, \tag{17}$$

where $d^{\mathrm{it}} \in \mathbb{R}^n$ is *projected gradient* computed as

$$d^{\mathrm{it}} = x^{\mathrm{it}} - P_\Omega(x - \alpha_{\mathrm{it}} \nabla f(x^{\mathrm{it}})). \tag{18}$$

The step-size α_{it} is computed by Barzilai-Borwein rule [1]

$$\alpha_{\mathrm{it}} = \frac{\langle x^{\mathrm{it}} - x^{\mathrm{it}-1}, x^{\mathrm{it}} - x^{\mathrm{it}-1}\rangle}{\langle \nabla f(x^{\mathrm{it}}) - \nabla f(x^{\mathrm{it}-1}), x^{\mathrm{it}} - x^{\mathrm{it}-1}\rangle} \tag{19}$$

and the step-size β_{it} is a result of Grippo, Lampariello, and Lucidi line-search method [5] for satisfying so-called *generalized Armijo criteria*

$$f(x^{\mathrm{it}} + \beta_{\mathrm{it}} d^{\mathrm{it}}) < f_{\max} \tau \beta_{\mathrm{it}} \langle \nabla f(x^{\mathrm{it}}), d^{\mathrm{it}}\rangle \tag{20}$$

with safeguarding parameter $\tau \in (0, 1)$ and $f_{\max}$ is a maximum function value in previous $m \geq 1$ iterations. The original SPG algorithm has been proposed by [2] for solving general optimization problems and the convergence is based on satisfaction of condition (20). Recently, [7] show that in the case of quadratic objective function, the line-search algorithm can be replaced by direct formula which satysfies (20)

$$\beta^{\mathrm{it}} = \min\left\{1, (1-\tau)\xi + \sqrt{(1-\tau)^2\xi^2 + \frac{2(f_{\max} - f(x^{\mathrm{it}}))}{\langle Ad^{\mathrm{it}}, d^{\mathrm{it}}\rangle}}\right\} \quad \text{with} \quad \xi = -\frac{\langle \nabla f(x^{\mathrm{it}}), d^{\mathrm{it}}\rangle}{\langle Ad^{\mathrm{it}}, d^{\mathrm{it}}\rangle}. \tag{21}$$

The algorithm non-monotonically decreases the norm of projected gradient and the function value until the stopping criteria is satisfied.

In the general case, the most time-consuming operation is the multiplication by Hessian matrix A, all other computations includes the evaluation of scalar products. In our case, the matrix has a special pattern; it is a block-diagonal matrix of K diagonal blocks of band matrices. Computational complexity of multiplication with such a matrix is $\mathcal{O}(KT)$. Please notice that the feasible set (2) is separable in T and the projection onto the set can be computed independently for each column of matrix Γ

$$\Gamma_{:,t} \in \{\gamma \in \mathbb{R}^K : \quad \gamma \geq 0 \text{ and } \sum_{k=1}^{K} \gamma_k = 1\}.$$

The projection onto each individual *simplex* is $\mathcal{O}(K)$, [4].

Summing up all the operations performed during one iteration of SPG-QP algorithm, the overall computation complexity is $\mathcal{O}(TK)$.

□

Theorem 1. *(The computational complexity of Algorithm 1)*

1. *Algorithm 1 generates the approximations with monotonically non-increasing objective function.*
2. *Let* dist *be Eulidean distance* (4)*. One iteration of Algorithm 1 is* $\mathcal{O}(TKD)$*.*

Proof.

1. Since the iterations solves the optimization problems (6) and (7), we have

$$\forall C \in \mathbb{R}^{D,K} : L(C, \Gamma^{\mathrm{it}-1}) \geq L(C^{\mathrm{it}}, \Gamma^{\mathrm{it}-1}) \quad \text{and} \quad \forall \Gamma \in \Omega_\Gamma : L(C^{\mathrm{it}}, \Gamma) \geq L(C^{\mathrm{it}}, \Gamma^{\mathrm{it}}).$$

Choosing $C = C^{\mathrm{it}-1}$ and $\Gamma = \Gamma^{\mathrm{it}-1}$, we get

$$L(C^{\mathrm{it}-1}, \Gamma^{\mathrm{it}-1}) \geq L(C^{\mathrm{it}}, \Gamma^{\mathrm{it}-1}) \geq L(C^{\mathrm{it}}, \Gamma^{\mathrm{it}}).$$

2. The statement is the consequence of Lemma 2 and Lemma 3.

□

2 Parallel Regularized Scalable Probabilistic Approximation Algorithm based on overlapping Domain Decomposition (DD-rSPA)

In the case of the computation in real-world applications, we are dealing with two main challenges: the computational demand (the number of operations that have to be performed to obtain the solution) and the memory limitation (the amount of information which can be processed by given machine). Both of these issues can be solved by High-Performance Computing (HPC). In this case, the algorithm runs on the machine which consists of several computational units (cores, processors, graphics processing unit) which are operating with distributed memory. The computational capacity of the largest supercomputers in the world can achieve more than 10^{17} FLOPS (floating-point operations per second) and can operate with several petabytes of memory. However, the massively parallel architectures cannot be efficiently utilized without appropriate massively parallel algorithms. For example in the case of discretized solution of partial differential equations with a huge number of variables, the original problem can be decomposed into smaller independent subproblems using so-called Domain Decomposition methods (DD). The idea is to solve subproblems in parallel using the individual computational units of the machine (i.e., nodes, cores, GPUs) and the only limitations arise in the case of global communication for the satisfaction of the continuity of global solution through domains. In practice, two different approaches are commonly used: overlapping DD, where the subdomains overlap by more than the interface (e.g., Schwarz alternating method or additive Schwarz method), and non-overlapping methods, where the subdomains intersect only on their interface (e.g., Balancing domain decomposition (BDDC), or Finite Element Tearing and Interconnecting (FETI)).

To analyze the problem of the global continuity and non-separability, suppose that we decompose the solution into two disjoint parts $\Gamma_{\{1\}}$ and $\Gamma_{\{2\}}$. Then the objective function of corresponding quadratic optimization Γ-problem (15) can be written as (after appropriate permutation of indexes)

$$\begin{aligned} f(\gamma) = f(\gamma_{\{1\}}, \gamma_{\{2\}}) &= \tfrac{1}{2}[\gamma_{\{1\}}^T, \gamma_{\{2\}}^T] \begin{bmatrix} A_{\{1,1\}} & A_{\{1,2\}} \\ A_{\{1,2\}}^T & A_{\{2,2\}} \end{bmatrix} \begin{bmatrix} \gamma_{\{1\}} \\ \gamma_{\{2\}} \end{bmatrix} - [w_{\{1\}}^T, w_{\{2\}}^T] \begin{bmatrix} \gamma_{\{1\}} \\ \gamma_{\{2\}} \end{bmatrix} \\ &= \underbrace{\frac{1}{2}\gamma_{\{1\}}^T A_{\{1,1\}}\gamma_{\{1\}} - w_{\{1\}}^T\gamma_{\{1\}}}_{=f_{\{1\}}(\gamma_{\{1\}})} + \underbrace{\frac{1}{2}\gamma_{\{2\}}^T A_{\{2,2\}}\gamma_{\{2\}} - w_{\{2\}}^T\gamma_{\{2\}}}_{=f_{\{2\}}(\gamma_{\{2\}})} + \gamma_{\{1\}}^T A_{\{1,2\}}\gamma_{\{2\}}. \end{aligned} \tag{22}$$

Using this equality, we can observe that the original minimization problem is separable into two disjoint minimization problems except the coupling term $\gamma_{\{1\}}^T A_{\{1,2\}}\gamma_{\{2\}}$. If we solve the problem separably to obtain $\gamma_{\{1\}}$ and $\gamma_{\{2\}}$ on separated computational units, we have to additionally handle with this term.

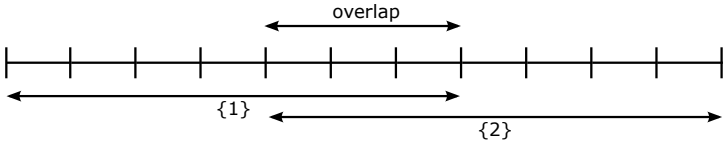

Figure 1: The overlapping Domain Decomposition: *the domain is separated into continuous overlapping parts. Each computational unit computes the corresponding local solution, however, the information in overlap have to be communicated to satisfy the continuity of global solution through domains.*

In our case, we implement the Schwartz Domain Decomposition method and separate the domain into overlapping domains, see Figure 1. For the demonstration, the figure presents the DD into two domains in 1D, but the approach is easily extendable to 3D and multiple domains in each direction, see Figure 2.

In the first step of algorithm, we solve the problem in each domain separately, i.e., we solve corresponding QP problem with appropriate block of the Hessian matrix and the block of linear term. Each domain sends the solution in overlap to the neighbouring domains and this vector is used for the computation of coupling term in local objective function. This operation can be written in terms of (22) - suppose that the local unknown part of the solution is $\gamma_{\{d\}}$ and the rest of the solution is $\gamma_{\{.\}}$. Then the objective function can be decomposed into (using the appropriate permutation of indexes)

$$\begin{aligned} f(\gamma_{\{d\}}, \gamma_{\{.\}}) &= \tfrac{1}{2}\gamma_{\{d\}}^T A_{\{d,d\}}\gamma_{\{d\}} - w_{\{d\}}^T\gamma_{\{d\}} + f_{\{.\}}(\gamma_{\{.\}}) + \gamma_{\{d\}}^T A_{\{d,.\}}\gamma_{\{.\}} \\ &= \tfrac{1}{2}\gamma_{\{d\}}^T A_{\{d,d\}}\gamma_{\{d\}} - (w_{\{d\}} - A_{\{d,.\}}\gamma_{\{.\}})^T\gamma_{\{d\}} + f_{\{.\}}(\gamma_{\{.\}}). \end{aligned}$$

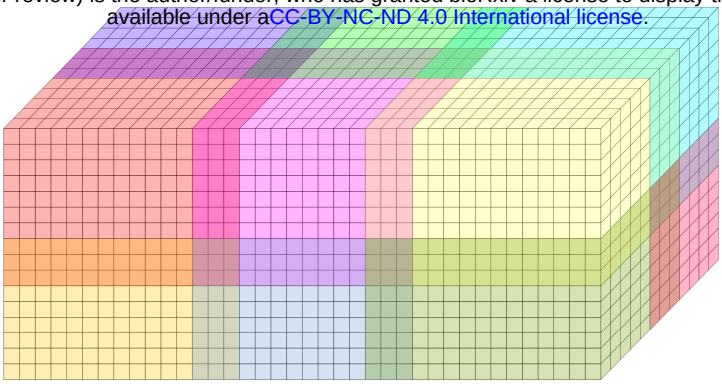

Figure 2: The overlapping Domain Decomposition in 3D: *the simplest way how to decompose the 3D domain into domains is to introduce overlapping rectangular cuboids. Such a decomposition simlifies the implementation and follows the sparsity of Hessian matrix in* Γ *problem.*

In the local domain, the unknown of the problem is the local $\gamma_{\{d\}}$, therefore the term $f_{\{.\}}(\gamma_{\{.\}})$ is constant, does not have any impact on the optimizer, and can be ignored. Please notice that the coupling matrix $A_{\{d,.\}}$ is a block of a sparse matrix and if the size of the overlap is sufficiently larger than the radius of the indicator function of the voxel neighborhood α_0, then the overlap information from the neighboring domains is sufficient information for assembling the correct overall objective function. In the iterative process, we update the linear term using the information from neighbours, solve a new QP problem, and communicate the update of overlap to neighbours, see Algorithm 2. In each iteration, we compare the local solution in overlap with the solution obtained from neighbours and if the difference is sufficiently small, we stop the algorithm.

Let A, w *be data of optimization problem* (15) *and let* $\{1\}, \ldots, \{N_D\}$ *denote the domains.*
Let $\forall d : \hat{w}_{\{d\}} = w_{\{d\}}$.

repeat

in parallel: solve the local problem

$$\gamma_d^* = \underset{\gamma_{\{d\}}}{\arg} \min_{\gamma_{\{d\}} \in \Omega_\Gamma} \frac{1}{2} \gamma_{\{d\}}^T A_{\{d,d\}} \gamma_{\{d\}} - \hat{w}_{\{d\}}^T \gamma_{\{d\}} \tag{23}$$

communication: send and receive the solution in overlap from neighbouring domains $\gamma_{\{.\}}$
in parallel: update linear term

$$\hat{w}_{\{d\}} = w_{\{d\}} - A_{\{d,.\}} \gamma_{\{.\}} \tag{24}$$

until $\|\gamma_{\{d\},overlap} - \gamma_{\{.\}}\|$ *is not sufficiently small*

Return an globally continuous parallel solution $\gamma_{\{1\}}, \ldots, \gamma_{\{N_D\}}$.

Algorithm 2: Schwartz Domain Decomposition method for solving Γ-problem in parallel.

References

[1] J. Barzilai and J. M. Borwein. Two point step size gradient methods. *IMA Journal of Numerical Analysis*, 8:141–148, 1988.

[2] E. G. Birgin, J. M. Martínez, and M. M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10:1196–1211, 2000.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, 1st edition, 2004.

[4] Y. Chen and X. Ye. Projection onto a simplex. *Unpublished manuscript, arXiv:1101.6081*, 2011.

[5] L. Grippo, F. Lampariello, and S. Lucidi. A nonmonotone line search technique for Newton's method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, 1986.

[6] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. on Pure and Applied Mathematics*, 42(5):577–685, 1989.

[7] L. Pospíšil, P. Gagliardini, W. Sawyer, and I. Horenko. On a scalable nonparametric denoising of time series signals. *Communications in Applied Mathematics and Computational Science*, 13:107–138, 2018.