NeuroImage 225 (2021) 117366

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion MRI

Ryutaro Tanno[a,g,*], Daniel E. Worrall[b], Enrico Kaden[a], Aurobrata Ghosh[a], Francesco Grussu[a,c], Alberto Bizzi[d], Stamatios N. Sotiropoulos[e,f], Antonio Criminisi[g], Daniel C. Alexander[a]

[a] Centre for Medical Image Computing and Department of Computer Science, UCL, Gower Street, London WC1E 6BT, UK
[b] Machine Learning Lab, University of Amsterdam, the Netherlands
[c] Institute of Neurology, Faculty of Brain Sciences, UCL, UK
[d] Neuroradiology Unit, Foundation IRCCS Carlo Besta Neurological Institute, Milan, Italy
[e] Sir Peter Mansfield Imaging Centre, School of Medicine and NIHR Biomedical Research Centre, University of Nottingham, UK
[f] Wellcome Centre for Integrative Neuroimaging, University of Oxford, UK
[g] Healthcare Intelligence, Microsoft Research Cambridge, UK

ARTICLE INFO

Keywords:
Uncertainty quantification
Deep learning
Safety
Robustness
Interpretability
Super-resolution
Image enhancement
Image synthesis
Neuroimaging
Diffusion MRI
Tractography

ABSTRACT

Deep learning (DL) has shown great potential in medical image enhancement problems, such as super-resolution or image synthesis. However, to date, most existing approaches are based on deterministic models, neglecting the presence of different sources of uncertainty in such problems. Here we introduce methods to characterise different components of uncertainty, and demonstrate the ideas using diffusion MRI super-resolution. Specifically, we propose to account for *intrinsic uncertainty* through a heteroscedastic noise model and for *parameter uncertainty* through approximate Bayesian inference, and integrate the two to quantify *predictive uncertainty* over the output image. Moreover, we introduce a method to propagate the predictive uncertainty on a multi-channelled image to derived scalar parameters, and separately quantify the effects of intrinsic and parameter uncertainty therein. The methods are evaluated for super-resolution of two different signal representations of diffusion MR images—Diffusion Tensor images and Mean Apparent Propagator MRI—and their derived quantities such as mean diffusivity and fractional anisotropy, on multiple datasets of both healthy and pathological human brains. Results highlight three key potential benefits of modelling uncertainty for improving the safety of DL-based image enhancement systems. Firstly, modelling uncertainty improves the predictive performance even when test data departs from training data ("out-of-distribution" datasets). Secondly, the predictive uncertainty highly correlates with reconstruction errors, and is therefore capable of detecting predictive "failures". Results on both healthy subjects and patients with brain glioma or multiple sclerosis demonstrate that such an uncertainty measure enables subject-specific and voxel-wise risk assessment of the super-resolved images that can be accounted for in subsequent analysis. Thirdly, we show that the method for decomposing predictive uncertainty into its independent sources provides high-level "explanations" for the model performance by separately quantifying how much uncertainty arises from the inherent difficulty of the task or the limited training examples. The introduced concepts of uncertainty modelling extend naturally to many other imaging modalities and data enhancement applications.

1. Introduction

In the last few years, deep learning techniques have permeated the field of medical image processing (Litjens et al., 2017; Shen et al., 2017). Beyond the automation of existing radiological tasks—e.g. segmentation (Kamnitsas et al., 2017b), detection (Roth et al., 2014), disease grading and classification (Araújo et al., 2017)—deep learning has been applied to a diverse set of "data enhancement" problems. Data enhancement aims to improve the quality, the information content[1], or the quantity of medical images available for research and clinics by transforming images from one domain to another (Isola et al., 2017). Previous research has shown the efficacy of data enhancement in different forms such as super-resolution (Chen et al., 2018; Oktay et al., 2016; Ravì et al., 2019), image synthesis (Kang et al., 2017; Nie et al., 2016), denoising (Benou et al., 2017; Chen et al., 2017), data harmonisation (Karayumak et al., 2018; Tax et al., 2019) across scanners and protocols, reconstruction

* Corresponding author at: Healthcare Intelligence, Microsoft Research Cambridge, UK.
E-mail address: r.tanno@cs.ucl.ac.uk (R. Tanno).

[1] Typically done by transferring information from an external source (e.g., training data).

https://doi.org/10.1016/j.neuroimage.2020.117366
Received 12 July 2019; Received in revised form 28 August 2020; Accepted 5 September 2020
Available online 9 October 2020

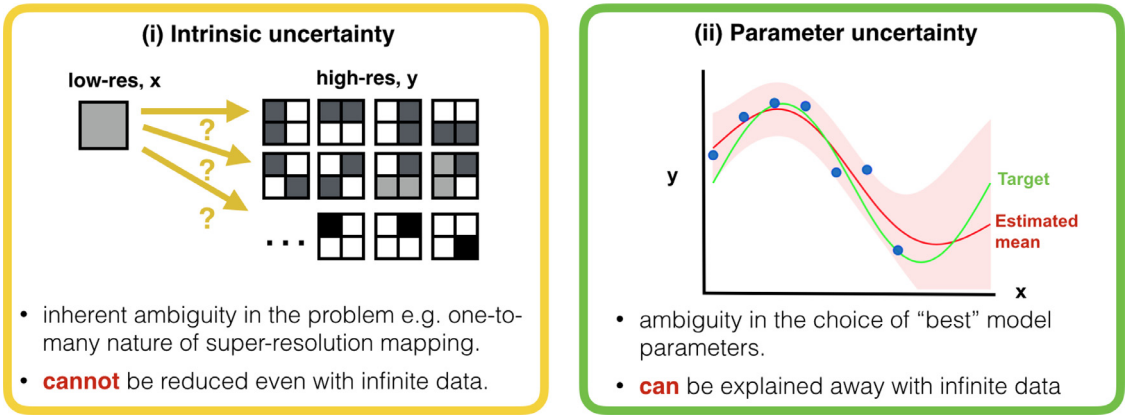

Fig. 1. Illustration of two different types of uncertainty (Hora, 1996). Intrinsic uncertainty (Wang et al., 1996) quantifies the degree of inherent ambiguity in the underlying problem. For example, in the case of super-resolution, there exist many possible high-resolution images **y** that would get mapped onto the same low-resolution input **x**. Intrinsic uncertainty is irreducible with training data. On the other hand, the parameter uncertainty (Draper, 1995) (a subtype of model uncertainty) arises from the finite training set. There exist more than one model that can explain the given training data equally well, and the parameter uncertainty quantifies the ambiguity in selecting the model parameters that best captures the target data-generating process. As illustrated in the figure on the right, parameter uncertainty decreases with more data; the green line shows the target function, the red line is the estimated mean, and the shaded region signifies the associated parameter uncertainty (standard deviation), which is higher in regions where we have fewer observations.

(Hammernik et al., 2018; Jin et al., 2017; Schlemper et al., 2018a; Sun et al., 2016; Yang et al., 2018; Yoon et al., 2019; Zhu et al., 2018), registration (Balakrishnan et al., 2018; Sokooti et al., 2017) and quality control (Esses et al., 2018; Wu et al., 2017). These advances have the potential not only to enhance the quality and efficiency of radiological care, but also facilitate scientific discoveries in medical research through increased volume and content of usable data.

However, most efforts in the development of data enhancement techniques have focused on improving the accuracy of deep learning algorithms, with little consideration of risk management. Blindly trusting the output of a given machine learning tool risks undetected failures e.g. spurious features and removal of structures (Cohen et al., 2018a). In medical applications, images inform scientific conclusions in research, and diagnostic, prognostic and interventional decisions in clinics. Therefore, translation of current proofs of principle to such safety-critical applications demands mechanisms for quantifying the risks of failures i.e. deriving uncertainty/confidence measures and explanation of their sources (Begoli et al., 2019).

Predictive failures of deep learning systems, by and large, occur due to two reasons: (i) the task itself is inherently ambiguous or (ii) the learned model is not adequate to describe the data (Der Kiureghian and Ditlevsen, 2009; Hora, 1996; Kendall and Gal, 2017; Tanno et al., 2017), as illustrated in Fig. 1. The former stems from *intrinsic uncertainty* (Wang et al., 1996), which describes ambiguity in the underlying data generating process (e.g. presence of stochasticity such as measurement noise and intrinsic ill-posed nature of the problem), and cannot be alleviated by increasing available training data or model complexity[2]. The latter is characterised by *model uncertainty* (Draper, 1995), which describes ambiguity in model specification[3]. Model uncertainty arises from (a) *parameter uncertainty*: ambiguity in fitting the model to the target mapping due to limited training data, or (b) *model bias*: errors due to insufficient flexibility of the model class (e.g. fitting a linear model to a sinusoidal process). These types of uncertainty can be reduced by collecting more data or specifying a different class of models. With the expressivity of deep neural networks, which are known to be universal approximators (Cybenko, 1989) if sufficiently large, one might reasonably assume that the model bias is small enough to be discounted. Under this assumption, intrinsic and parameter uncertainty (Fig. 1) fully characterise the predictive failures of deep learning models. Therefore, accurate estimation of these uncertainties are needed and would potentially allow practioners to understand better the limits of the models, flag doubtful predictions, and highlight test cases that are not well represented in the training data.

In this work, we introduce methods for modelling components of uncertainty in medical image enhancement systems based on deep learning. We propose to model intrinsic uncertainty through a input-dependent (heteroscedastic) noise model (Nix and Weigend, 1994) and parameter uncertainty through variational dropout (Kingma et al., 2015). We then combine and propagate these two "source" uncertainties into a spatial map of *predictive uncertainty* over the output image, which can be used to assess the output reliability on subject-specific and voxel-wise basis. Lastly, we propose a method to propagate the predictive uncertainty to arbitrary derived quantities of the output images, such as scalar indices that are commonly used for subsequent analysis, and decompose it into distinct components which separately quantify the contributions of intrinsic and parameter uncertainty.

The primary goal of this work is to evaluate the practical utility of the proposed methods for modelling uncertainty in terms of three aspects; (i) performance on unseen datasets, including generalisation to out-of-distribution data and robustness to noise/outliers; (ii) safety assessment of system output; (iii) explainability of failures. We note here that validating the "correctness" of the derived uncertainty estimates is an important fundamental problem, that is not the main focus of this work—a very challenging task as the ground truth is typically unknown. Here we take a pragmatic approach and focus our study on how useful, rather than how accurate, uncertainty modelling is in the context of medical image enhancement applications. To this end, we use Image Quality Transfer (IQT) (Alexander et al., 2014; 2017; Blumberg et al., 2018; Tanno et al., 2016) as the core test ground[4], focusing on its application to *super-resolution* of diffusion magnetic resonance imaging (dMRI) scans. For two different types of diffusion signal representations, we evaluate the effects of uncertainty modelling on generalisation by measuring the predictive accuracy on unseen test subjects in the Human Connectome Project (HCP) dataset (Sotiropoulos et al., 2013) and the

[2] Intrinsic uncertainty is also known as *aleaotoric* or statistical uncertainty.

[3] Model uncertainty is a subclass of *epistemic uncertainty* (Hora, 1996) which encompasses types of uncertainties that arise from lack of knowledge.

[4] IQT is a data-enhancement framework for propagating information from rare or expensive high quality images to lower quality but more readily available images.

Lifespan dataset (Harms et al., 2018). We additionally test the value of improved predictive performance in a downstream tractography application. We then test the capability of the predictive uncertainty map to indicate predictive errors and thus to detect potential failures on images of both healthy subjects and those in which pathologies unseen in the training data arise, specifically from glioma and multiple-sclerosis (MS) patients. Lastly, we perform the decomposition of predictive uncertainty on HCP subjects with benign abnormalities, and assess its potential value in gaining high-level interpretations of predictive performance.

2. Related works

This section provides a review of related works under several different themes. We first review the development of learning-based image enhancement methods in medical imaging applications. We then discuss the recent advances made to model and quantify uncertainty in such image enhancement problems. Lastly, we describe the existing strands of research in uncertainty modelling for other medical imaging problems and fields of applications.

Various forms of image enhancement can be cast as image transformation problems where the input image from one domain is mapped to an output image from another domain. Numerous recent methods have proposed to perform image transformation tasks as supervised regression of low quality against high quality image content. Alexander et al. (2014) proposed Image Quality Transfer (IQT), a general framework for supervised quality enhancement of medical images. They demonstrated the efficacy of their method through a random forest (RF) implementation of super-resolution (SR) of brain diffusion tensor images and estimation of advanced microstructure parameter maps from sparse measurements. More recently, deep learning, typically in the form of convolutional neural networks (CNNs), has shown additional promise in this kind of task. For example, Oktay et al. (2016) proposed a CNN model to upsample a stack of 2D MRI cardiac volumes in the through-plane direction, where the SR mapping is learnt from 3D cardiac volumes of nearly isotropic voxels. This work was later extended by Oktay et al. (2018) with the addition of global anatomical prior based on auto-encoder. Zhao et al. (2018) proposed a solution to the same SR problem for brains that utilises the high frequency information in in-plane slices to super-resolve in the through-plane direction without requiring external training data. In addition, a range of different architectures of CNNs have been considered for SR of other modalities and anatomical structures such as structural MRI (Chen et al., 2018) of brains, retinal fundus images (Mahapatra et al., 2017) and computer tomography (CT) scans of chest (Yu et al., 2017). Another problem of growing interest is image synthesis, which aims to synthesise an image of a different modality given the input image. Nie et al. (2018) employed a conditional generative adversarial network to synthesise CT from MRI with fine texture details whilst (Wolterink et al., 2017) extended this idea using a CycleGAN (Zhu et al., 2017) to leverage the abundance of unpaired training sets of CT and MR scans. In Bahrami et al. (2016), a variant of CNN was applied to predict 7T images from 3T MRI, where both contrast and resolution are enhanced. Another notable application is the harmonisation of diffusion MRIs (Blumberg et al., 2018, 2019; Karayumak et al., 2018; Tax et al., 2019) where images acquired at different scanners or magnetic field strengths are mapped to the common reference image space to allow for joint analysis.

Despite this advancement, all of these methods commit to a single prediction and lack a mechanism to communicate uncertainty in the output image. In medical applications where images can ultimately inform life-and-death decisions, quantifying reliability of output is crucial. Tanno et al. (2016) aimed to address this problem for supervised image enhancement for the first time by proposing a Bayesian variant of random forests to quantify uncertainty over predicted high-resolution MRI. They showed that the uncertainty measure correlates well with the accuracy and can highlight abnormality not represented in the training data. In our preliminary work (Tanno et al., 2017), we made an initial attempt to extend this approach with probabilistic deep-learning formulation, and showed that modelling different components of uncertainty—intrinsic and parameter uncertainty—allows one to build a more generalisable model and quantify predictive confidence. Kendall and Gal (2017) concurrently investigated the same problem in computer vision, suggesting its utility for safety-critical applications such as self-driving cars. More recently, Shi et al. (2019) extended these works in the context of medical image segmentation and proposed a mechanism to learn the intrinsic uncertainty in a supervised manner, when multiple labels are available. Dalca et al. (2018) proposed a CNN-based probabilistic model for diffeomorphic image registration with a learning algorithm based on variational inference, and demonstrated the state-of-the-art registration accuracy on established benchmarks while providing estimates of registration uncertainty. An alternative approach is ensembling where the variance of the predictions of multiple networks is used to quantify the predictive uncertainty (Lakshminarayanan et al., 2017). Schlemper et al. (2018b) proposed a novel combination of the cascaded CNN architecture and compressive sensing, equipped with a variant of ensemble techniques, which enabled robust reconstruction of highly undersampled cardiovascular diffusion MR images, and quantification of reconstruction uncertainty. Bragman et al. (2018) studied the value of uncertainty modelling for multi-task learning in the context of MR-only radiotherapy treatment planning where the synthetic CT image and the segmentation of organs at risk are simultaneously predicted from the input MRI image.

We should also note that, although not the focus of this work, research on uncertainty modelling in deep learning techniques extend to other medical image processing tasks beyond data enhancement, such as segmentation, detection and classification. For example, Nair et al. (2018, 2020) demonstrated for lesion segmentation of multiple sclerosis that the voxel-wise uncertainty metrics can be used for quality control; by filtering out predictions with high uncertainty, the model could achieve higher lesion detection accuracy. A concurrent work by Eaton-Rosen et al. (2018) showed for the task of brain tumour segmentation that the Monte Carlo (MC) sample variance from dropout (Gal and Ghahramani, 2015) can be calibrated to provide meaningful error bars over estimates of tumour volumes. Similarly, Roy et al. (2019) introduced ways to turn voxel-wise uncertainty score into structure-wise uncertainty metrics for brain parcellation task, and showed their values in performing more reliable group analysis. The uncertainty metric based on MC dropout has also shown promise in disease grading of retinal fundal images (Leibig et al., 2017; Worrall et al., 2016), and more recently an extension based on test-time augmentation was introduced by Ayhan and Berens (2018). An alternative approach to these works is to train a model that predicts the uncertainty score directly; Raghu et al. (2019) showed that this approach is more effective when opinions from multiple experts are available for each image. In a similar vein, Eaton-Rosen et al. (2019) proposed a means to estimate confidence intervals of any desired percentiles based on quantile regression, and tailored it to the task of counting objects in an given image with successful demonstration in estimating uncertainty over the measurements of different biomarkers such as histopathological cell counting and white matter hyperintensity counting. Kohl et al. (2018) and Baumgartner et al. (2019) proposed methods to generate a set of diverse and plausible segmentation proposals on a given image, capturing more realistically the high inter-reader annotation variability, which is commonly observed in medical image segmentation tasks. Lastly, Raykar et al. (2010) and Tanno et al. (2019) demonstrated for the classification of mammograms and cardiac ultra-sound images, respectively that modelling uncertainty of human annotators enables robust learning from noisy labels in the presence of large disagreement.

However, within the context of medical image enhancement, these lines of research performed only limited validation of the quality and utility of uncertainty modelling. In this work, we formalise and extend the preliminary ideas in Tanno et al. (2017) and provide a comprehensive set of experiments to evaluate the proposed uncertainty modelling

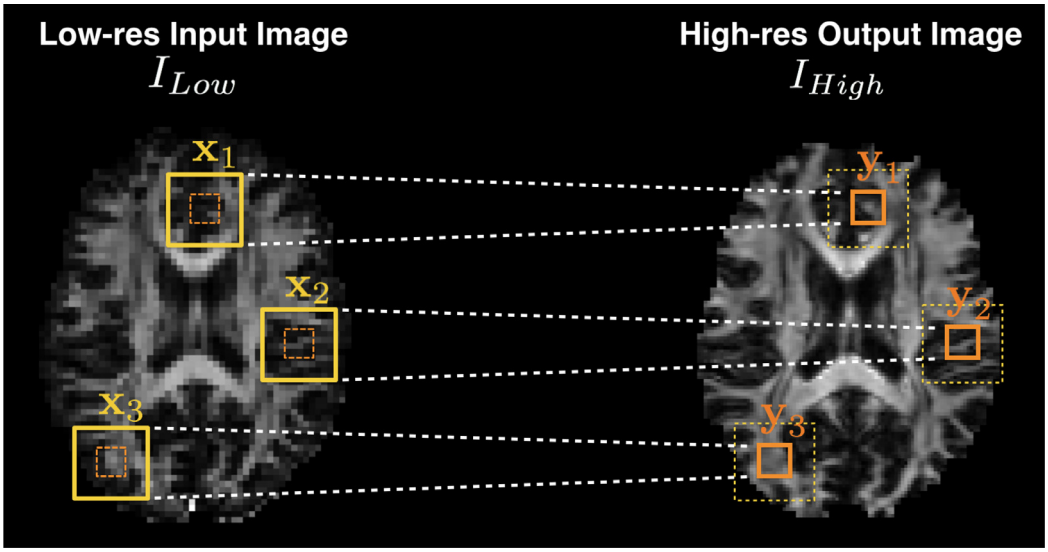

Fig. 2. Illustration of the patch-wise regression in super-resolution application. The conditional distribution over the high quality image $p(I_{High}|I_{Low})$ is assumed to factorise over local neighbourhoods $\{(\mathbf{x}_i, \mathbf{y}_i)\}_i$. In this case, for each input subvolume $\mathbf{x}_i$ (in yellow), the high resolution version of the smaller centrally located neighbourhood, $\mathbf{y}_i$ (in orange) is regressed.

techniques in a diverse set of datasets, which vary in demographics, scanner types, acquisition protocols or pathology. Our proposed framework models different components of uncertainty, namely intrinsic and parameter uncertainty, and provides conclusive evidence that this improves performance thanks to different regularisation effects. In addition, we propose a method to decompose predictive uncertainty over an arbitrary function of the output image (e.g. morphological measurements) into its sources, in order to provide a high-level explanation of model performance on the given input.

3. Methods

This section describes the methods for modelling different components of uncertainty that arise in data enhancement. Firstly, we provide an overview of Image Quality Transfer (IQT) which formulates data enhancement as a supervised learning problem. Secondly, using the IQT framework, we introduce methods to model *intrinsic* and *parameter uncertainty*, separately, focusing on the application of super-resolution. We then combine the two approaches and estimate the overall uncertainty over prediction (*predictive uncertainty*) by approximating the variance of the predictive distribution (Eq. (9)). Lastly, we propose a method for decomposing predictive uncertainty into its sources—intrinsic and parameter uncertainty—in an attempt to provide quantifiable explanations for the confidence on model output (Eq. (13)).

3.1. Background: image quality transfer

Alexander et al. (2014) proposed Image Quality Transfer (IQT), the first supervised learning based framework for data enhancement of medical images, and here we survey its general formulation which forms the testing ground of this work. IQT performs data enhancement via regression of low quality against high quality image content. In order to overcome the memory demands of processing 3-dimensional medical images, along with other subsequent work such as (Bahrami et al., 2016; Oktay et al., 2016; 2018; Yang et al., 2016), IQT assumes factorisability over local neighbourhoods (also called patches) and models the conditional distribution of high-quality image I_{High} given the corresponding low-quality input I_{Low} as:

$$p(I_{High}|I_{Low}) = \prod_{i\in S} p(\mathbf{y}_i|\mathbf{x}_i) \quad (1)$$

where $\{\mathbf{y}_i\}_{i\in S}$ is a set of disjoint high-quality subvolumes with S denoting the set of their indices, which together constitute the whole image I_{High}, while $\{\mathbf{x}_i\}_{i\in S}$ is a set of potentially overlapping low-quality subvolumes, each of which contains and is spatially larger than the corresponding $\mathbf{y}_i$, as illustrated in Fig. 2. In other words, here we assume that the high-resolution neighbourhoods are statistically independent given the corresponding low-resolution versions. We define each local neighbourhood as a cubic sub-volume. The locality assumption reduces the problem of learning $p(I_{High}|I_{Low})$ to the much less memory intensive problem of learning $p(\mathbf{y}|\mathbf{x})$. In other words, IQT formulates the data enhancement task as a patch-wise regression where an input low-quality image I_{Low} is split into smaller overlapping sub-volumes $\{\mathbf{x}_i\}_{i\in S}$ and the corresponding non-overlapping high-quality sub-volumes $\{\mathbf{y}_i\}_{i\in S}$ are independently predicted according to the patch regressor $p(\mathbf{y}|\mathbf{x})$. The final prediction for the 3D high-quality volume I_{high} is constructed by tesellating the output patches $\{\mathbf{y}_i\}_{i\in S}$.

The original implementation of IQT (Alexander et al., 2014; 2017; Tanno et al., 2016) employed a variant of random forests (RFs) to model $p(\mathbf{y}|\mathbf{x})$ while more recent (Bahrami et al., 2016; Oktay et al., 2016; 2018; Yang et al., 2016) approaches use variants of convolutional neural networks (CNNs). Either way, the machine learning algorithm is trained on pairs of high-quality and low-quality patches $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ extracted from a set of image volumes, and is used to perform the data-enhancement task of interest. Typically, such patch pairs D are synthesised by down-sampling a collection of high quality images to approximate their counterparts in a particular low-quality scenario (Alexander et al., 2014; Oktay et al., 2016). In this work, we focus on the task of super-resolution (SR) where the spatial resolution of I_{high} is higher than the input image I_{low}.

3.2. Baseline super-resolution model: 3D-ESPCN

As the baseline architecture for modelling $p(\mathbf{y}|\mathbf{x})$, we adapt efficient subpixel-shifted convolutional network (ESPCN) (Shi et al., 2016) to 3D data. ESPCN is a recently proposed method with the capacity to perform real-time per-frame SR of videos while retaining high accuracy on 2D natural images. We have chosen to base on this architecture for its simplicity and computational performance. Most CNN-based SR techniques first up-sample a low-resolution input image (e.g. through bilinear interpolation, Dong et al., 2016; deconvolution, McDonagh et al., 2017; Oktay et al., 2016; fractional-strided convolution, Johnson et al., 2016,

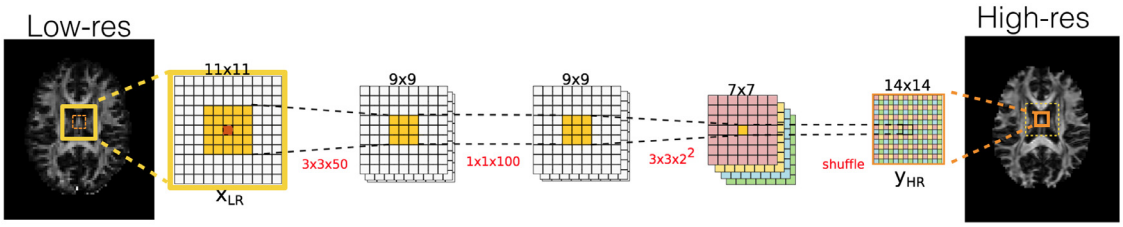

Fig. 3. 2D illustration of an example baseline network (ESPCN Shi et al., 2016) with upsampling rate, $r = 2$. The receptive field of the central 2^2 pixels in the output patch is 5^2 pixels in the input patch and is shown in yellow. The shuffling operation at the end periodically rearranges the final feature maps from the low-resolution space into the high-resolution space.

etc.) and then refine the high-resolution estimate through a series of convolutions. These methods suffer from the fact that (1) the up-sampling can be a lossy process and (2) refinement in the high-resolution space has a higher computational cost than in the low-resolution space. By contrast, ESPCN performs convolutions in the low-resolution-space, up-sampling afterwards. The reduced resolution of feature maps dramatically decreases the computational and memory costs, which is more pronounced in processing 3D data.

More specifically the ESPCN is a fully convolutional network, with a special *shuffling operation* on the output, which identifies individual feature channel dimensions with spatial locations in the high-resolution output. Fig. 3 shows a 2D illustration of an example ESPCN when the fully convolutional part of the network consists of 3 convolutional layers, each followed by a ReLU, and the final layer has cr^2 feature maps where r is the upsampling rate and c is the number of channels in the output image (e.g. 6 in the case of DT images). The shuffling operation takes the feature maps of shape $h \times w \times cr^2$ and remaps pixels from different channels into different spatial locations in the high-resolution output, producing a $rh \times rw \times c$ image, where h and w denote height and width of the pre-shuffling feature maps. This shuffling operation in 3D is given by $S(F)_{i,j,k,c} = F_{\lfloor i/r \rfloor, \lfloor j/r \rfloor, \lfloor k/r \rfloor, (r^3-1)c + \mathrm{mod}(i,r) + r \cdot \mathrm{mod}(j,r) + r^3 \cdot \mathrm{mod}(k,r)}$ where F is the pre-shuffled feature maps. The combined effects of the last convolution and shuffling is effectively a learned interpolation, and an efficient implementation of deconvolution layer (Zeiler et al., 2011) where the kernel size is divisible by the size of the stride (Shi et al., 2016). Therefore, it is less susceptible to checker-board like artifacts commonly observed with deconvolution operations (Odena et al., 2016).

At test time, the prediction of higher resolution volume is performed through *shift-and-stitch* operation. The network takes each sub-volume $\mathbf{x}$ in a low-resolution image, and predicts the corresponding high-resolution sub-volume $\mathbf{y}$. By tessellating the predictions from appropriately shifted inputs $\mathbf{x}$, the whole high-resolution volume is reconstructed. With convolutions being local operations, each output voxel is only inferred from a local region in the input volume, and the spatial extent of this local connectivity is referred to as the *receptive field*. For a given input subvolume, the network increases the resolution of the central voxel of each receptive field e.g. the central 2^3 output voxels are estimated from the corresponding 5^3 receptive field in the input volume, as coloured yellow in Fig. 3.

Given training pairs of high-resolution and low-resolution patches $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, we optimise the network parameters by minimising the sum of per-pixel mean-squared-error (MSE) between the ground truth $\mathbf{y}$ and the predicted high-resolution patch $\mu_\theta(\mathbf{x})$ over the training set. Here θ denotes all network parameters. This is equivalent to minimising the negative log likelihood (NLL) under the Gaussian noise model $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}; \mu_\theta(\mathbf{x}), \sigma^2 I)$ with fixed isotropic variance σ^2.

3.3. Intrinsic uncertainty and heteroscedastic noise model

Intrinsic uncertainty quantifies the inherent ambiguity of the underlying problem that is irreducible with data as illustrated in Fig. 1(i). Here we capture intrinsic uncertainty by estimating the variance of the target conditional distribution $p(\mathbf{y}|\mathbf{x}, \theta)$. In medical images, intrinsic uncertainty is often spatially and channel-wise varying. For example, super-resolution could be fundamentally harder on some anatomical structures than others due to signal variability as shown in Tanno et al. (2016). It may also be the case that some channels of the image volume might contain more complex, non-linear and noisy signals than other channels e.g. higher order terms in diffusion signal representations. To capture such potential variation of intrinsic uncertainty, we model $p(\mathbf{y}|\mathbf{x}, \theta)$ as a Gaussian distribution with input-dependent varying variance:

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathcal{N}\big(\mathbf{y}; \mu(\mathbf{x}; \boldsymbol{\theta}_1), \Sigma(\mathbf{x}; \theta_2)\big) = \frac{\exp\Big(\big(\mathbf{y} - \mu(\mathbf{x}; \theta_1)\big)^{\mathrm{T}} \Sigma^{-1}(\mathbf{x}; \theta_2)\big(\mathbf{y} - \mu(\mathbf{x}; \theta_1)\big)\Big)}{\sqrt{(2\pi)^{\dim(\mathbf{y})} \cdot \det \Sigma(\mathbf{x}; \theta_2)}} \tag{2}$$

where the mean $\mu(\mathbf{x}; \theta_1)$ and the covariance $\Sigma(\mathbf{x}; \theta_2)$ are functions of input $\mathbf{x}$ and modelled by two separate 3D-ESPCNs (as shown in Fig. 4), which we refer to as "mean network" and "covariance network", and are parametrised by θ_1 and θ_2, respectively. Here $\dim(\mathbf{y})$ denotes the dimension of the output patch $\mathbf{y}$. We note that the input patch $\mathbf{x}$ varies spatially, which makes the estimated variance spatially varying and different for respective channels. Fig. 4 shows a 2D illustration of our 3D architecture. For each low-resolution input patch $\mathbf{x}$, we use the output of the mean network $\mu(\mathbf{x}; \theta_1)$ at the top as the final estimate of the high-resolution ground truth $\mathbf{y}$ whilst the diagonal elements of the covariance $\Sigma(\mathbf{x}; \theta_2)$ quantify the corresponding intrinsic uncertainty over individual components in $\mu(\mathbf{x}; \theta_1)$ and over different channels. Lastly, we note that this is a specifc instance of a broad class of models, called *heteroscedastic noise models* (Nix and Weigend, 1994; Rao, 1970) where the variance is a function of the value of the input. In contrast, the baseline 3D-ESPCN can be viewed as an example of *homoscedastic noise models* with $\mathbf{y} = \mu_\theta(\mathbf{x}) + \sigma\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$ with constant variance σ^2 across all spatial locations and image channels, which is highly unrealistic in most medical images.

We jointly optimise the parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ of the mean network and the covariance network by minimising the negative loglikelihood (NLL)[5]:

$$\mathcal{L}_\theta(\mathcal{D}) = \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}} -\log\, p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \tag{3}$$

$$= \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}} -\log\, \mathcal{N}\big(\mathbf{y}_i; \mu(\mathbf{x}_i; \boldsymbol{\theta}_1), \Sigma(\mathbf{x}_i; \boldsymbol{\theta}_2)\big) \tag{4}$$

$$= \mathcal{M}_\theta(\mathcal{D}) + \mathcal{H}_\theta(\mathcal{D}) + c \tag{5}$$

[5] One may wonder how it is possible to estimate the heteroscedastic variance when you observe only one possible output patch $\mathbf{y}$ for each input $\mathbf{x}$ in the training data. Here we are assuming that the conditional distribution $p(\mathbf{y}|\mathbf{x}, \theta)$ is locally "smooth"; if two input low-res patches are similar $\mathbf{x}_1 \approx \mathbf{x}_2$, then we should also have $p(\mathbf{y}|\mathbf{x}_1, \theta) \approx p(\mathbf{y}|\mathbf{x}_2, \theta)$. Therefore, intuitively speaking, by minimising the negative log-likelihood Eq. (5), the model estimates the mean and the variance of the conditional distribution for each $\mathbf{x}$ based on the output labels of other similar input samples.

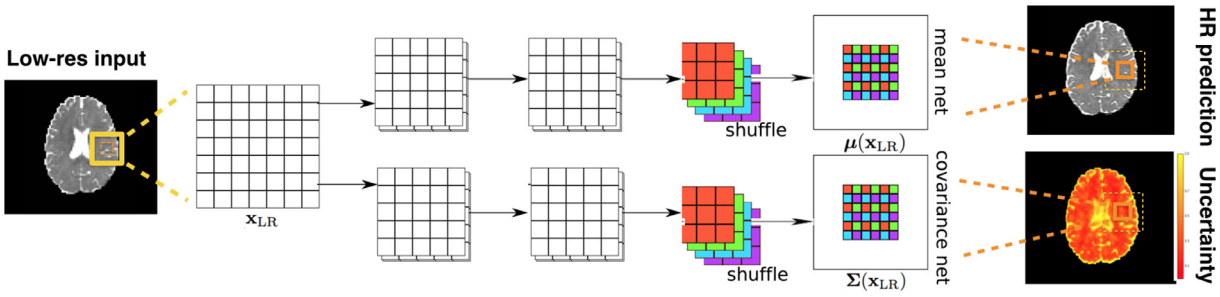

Fig. 4. 2D illustration of the proposed dual-path architecture which estimates the mean and diagonal covariance of the Gaussian conditional distributions as functions of the input low-resolution subvolume **x**. The "mean network" $\mu(\cdot)$ at the top generates the high-resolution prediction, while the "covariance network" $\Sigma(\cdot)$ at the bottom estimates the corresponding covariance matrix at the selected location in the volume. The diagonal entries of the covariance are used to quantify the intrinsic uncertainty. The parameters of both networks are learned by minimising the common loss function (Eq. (5)).

where c is a constant and the remaining terms are given by

$$\mathcal{M}_\theta(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\big(\mathbf{y}_i - \mu(\mathbf{x}_i;\theta_1)\big)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i;\theta_2)\big(\mathbf{y}_i - \mu(\mathbf{x}_i;\theta_1)\big),$$

$$\mathcal{H}_\theta(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\log\det\Sigma(\mathbf{x}_i;\theta_2).$$

Here $\mathcal{M}_\theta(\mathcal{D})$ denotes the mean squared Mahalanobis distance with respect to the predictive distribution $p(\mathbf{y}|\mathbf{x},\theta)$. For simplicity, in this work we assume diagonality of the covariance matrix $\Sigma(\mathbf{x};\theta_2)$. This means that the Mahalanobis distance term $\mathcal{M}_\theta(\mathcal{D})$ equates to the sum of MSEs across all pixels and channels in the output, weighted by the inverse of the corresponding variance (estimated intrinsic uncertainty)[6]. This term naturally encourages assigning high uncertainty to regions with higher MSEs, robustifying the training to noisy labels and outliers. On other other hand, $\mathcal{H}_\theta(\mathcal{D})$ represents the mean differential entropy and discourages the spread of $\Sigma_{\theta_2}(\mathbf{x})$ from growing too large. We note that the covariance network is used to modulate the training of the mean network and quantify intrinsic uncertainty during inference while only the mean network generates the final prediction, requiring a single 3D-ESPCN to perform super-resolution.

3.4. Parameter uncertainty and variational dropout

Parameter uncertainty signifies the ambiguity in selecting the parameters of the model that best describes the training data as illustrated in Fig. 1.(ii). The limitation of the previously introduced 3D-ESPCN baseline (Section 3.2) and its heteroscedastic extension (Section 3.3) is their reliance on a single estimate of network parameters. In many medical imaging problems, the amount of training data is modest; in such cases, this point estimate approach increases the risk of overfitting (Gal and Ghahramani, 2015).

We combat this problem with a Bayesian approach. Specifically, instead of resorting to a single network of fixed parameters, we consider the (posterior) distribution over all the possible settings of network parameters given training data $p(\theta|\mathcal{D})$. This probability density encapsulates the parameter uncertainty, with its spread of mass describing the ambiguity in selecting most appropriate models to explain the training data $\mathcal{D}$. However, in practice, the posterior $p(\theta|\mathcal{D})$ is intractable due to the difficulty in computing the normalisation constant. We, therefore, propose to approximate $p(\theta|\mathcal{D})$ with a simpler distribution $q_\phi(\theta)$ (Blei et al., 2017). Specifically, we adapt a technique called *variational dropout* (Kingma et al., 2015) to convolution operations from its original version introduced for feedforward NNs.

[6] In the case of full covariance, $\mathcal{M}_\theta(\mathcal{D})$ becomes the MSE in the basis of principle components, weighted by the corresponding eigenvalues.

Binary dropout (Srivastava et al., 2014) is a popular choice of method for approximating posterior distributions (Gal and Ghahramani, 2015) with demonstrated utility in medical imaging applications (Bragman et al., 2018; Eaton-Rosen et al., 2018; Leibig et al., 2017; Nair et al., 2018; Roy et al., 2019; Worrall et al., 2016; Yang et al., 2016). However, typically hyper-parameters (dropout rates) need to be pre-set before the training, requiring inefficient cross-validation and thus substantially constraining the flexibility of approximate distribution family $q_\phi(\cdot)$ (often a fixed dropout rate per layer). This limitation motivates us to use variational dropout (Kingma et al., 2015) that extends such approach with a way to learn the dropout rate from data for every single weight in the network and theoretically enables a more effective approximation of the posterior distribution. Another established class of methods is stochastic gradient Markov chain Monte Carlo (SG-MCMC) method (Chen et al., 2014; Ma et al., 2015; Neal, 1993; Welling and Teh, 2011). However, in this work, we do not not consider SG-MCMC methods because they remain, although unbiased, computationally inefficient due to the requirement of evaluating an ensemble of models for posterior computation, and are slow to converge for high-dimensional problems.

Variational dropout (Kingma et al., 2015) employs a form of variational inference to approximate the posterior $p(\theta|\mathcal{D})$ by a member of tractable family of distributions $q_\phi(\theta) = \prod_{ij}\mathcal{N}(\theta_{ij};\eta_{ij},\alpha_{ij}\eta_{ij}^2)$ parametrised by $\phi = \{\eta_{ij},\alpha_{ij}\}_{ij}$, such that Kullback-Leibler (KL) divergence $\mathrm{KL}\big(q_\phi(\theta)||p(\theta|\mathcal{D})\big)$ is minimised. Here, θ_{ij} denotes an individual element in the convolution filters of CNNs as a random variable with parameters α_{ij}(dropout rate) and η_{ij} (mean), and the posterior over the set of all weights is effectively approximated with a product of univariate Gaussian distributions. In practice, introducing a prior $p(\theta)$ and applying Bayes' rule allow us to rewrite the minimisation of the KL divergence as maximisation of the quantity known as the evidence lower bound (ELBO) (Blei et al., 2017). Here during training, we learn the variational parameters $\phi = \{\eta_{ij},\alpha_{ij}\}_{ij}$ by minimising the negative ELBO (to be consistent with the NLL cost function in Eq. (3)):

$$\mathcal{L}_\phi(\mathcal{D}) = \sum_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{D}}\Big(\mathbb{E}_{q_\phi(\theta)}[-\log p(\mathbf{y}_i|\mathbf{x}_i,\theta)] + \mathrm{KL}(q_\phi(\theta)||p(\theta))\Big) \tag{6}$$

An accurate approximation for the KL term for log-uniform prior $p(\theta)$ is proposed in Molchanov et al. (2017), which is employed here. On the other hand, the first term (referred to as the reconstruction term) cannot be computed exactly, thus we employ the following MC approximation by sampling S samples of network parameters from the posterior:

$$\mathbb{E}_{q_\phi(\theta)}[-\log p(\mathbf{y}|\mathbf{x},\theta)] \approx \frac{1}{S}\sum_{s=1}^{S}-\log p(\mathbf{y}|\mathbf{x},\theta^{(s)}),\quad \theta^{(s)}\sim q_\phi(\theta) \tag{7}$$

Adapting the local reparametrisation trick presented in Kingma et al. (2015) to a convolution operation, we derive the implementation of posterior sampling $\theta^{(s)}\sim q_\phi(\theta)$ such that the vari-

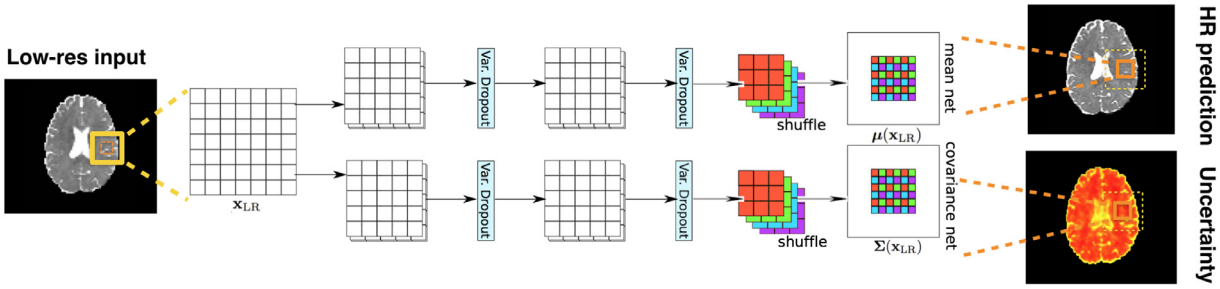

Fig. 5. 2D illustration of a heteroscedastic network with variational dropout. Diagonal covariance is again assumed. The top 3D-ESPCN estimates the mean and the bottom one estimates the covariance matrix of the likelihood. Variational dropout is applied to feature maps after every convolution where Gaussian noise is injected into feature maps $F_{\mathrm{out}} = \mu_Y + \sigma_Y \odot \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ (see Eq. (8)).

ance of gradients over each mini-batch is low[7]. In practice, this amounts to replacing each standard convolution kernel with a "Bayesian" convolution, which proceeds as follows. Firstly, we define two separate convolution kernels: $\boldsymbol{\eta} \in \mathbb{R}^{c \times k^2}$ ("mean" kernels) and $\alpha \odot \eta^2 \in \mathbb{R}^{c \times k^2}$ ("variance" kernels) where $\odot$ denotes the element-wise multiplication, c is the number of input channel and k is the kernel width. Input feature maps F_{in} and its elementwise squared values are convolved by respective kernels to compute the "mean" and "variance" of the output feature maps $\mu_Y \triangleq {F_{\mathrm{in}}}^* \boldsymbol{\eta}$ and $\sigma_Y^2 \triangleq F_{\mathrm{in}}^2 * (\alpha \odot \eta^2)$ (Fig. 5). Lastly, the final output feature maps F_{out} are computed by drawing a sample from $\mathcal{N}(\mu_Y, \sigma_Y^2)$ i.e. computing the following quantity:

$$F_{\mathrm{out}} \triangleq \mu_Y + \sigma_Y \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I). \tag{8}$$

Every forward pass (i.e. computation of each $p(\mathbf{y}|\mathbf{x}, \theta^{(s)})$) with variational dropout is thus performed via a sequence of Bayesian convolutions. Since the injected Gaussian noise ϵ is independent of the variational parameters $\phi = \{\eta_{ij}, \alpha_{ij}\}_{ij}$, the approximate reconstruction term in Eq. (7) is differentiable with respect to them (Kingma and Welling, 2014).

3.5. Joint modelling of intrinsic and parameter uncertainty

We now describe how to combine the methods for modelling intrinsic and parameter uncertainty. Operationally, we take the dual architecture (Fig. 4) used to model intrinsic uncertainty, and apply variational dropout to every convolution layer in it. The intrinsic uncertainty is modelled in the heteroscedastic Gaussian model $p(\mathbf{y}|\mathbf{x}, \theta_1, \theta_2) = \mathcal{N}\big(\mathbf{y}; \mu(\mathbf{x}; \theta_1), \Sigma(\mathbf{x}; \theta_2)\big)$ while the parameter uncertainty is captured in the approximate posterior $q_\phi(\theta_1, \theta_2) \approx p(\theta_1, \theta_2|D)$ obtained from variational dropout.

At test time, for each low-resolution input subvolume $\mathbf{x}$, we would like to compute the predictive distribution $p(\mathbf{y}|\mathbf{x}, D)$ over the high-resolution output $\mathbf{y}$. We approximate this quantity by $q_\phi^*(\mathbf{y}|\mathbf{x})$ by taking the "average" of all possible network predictions $p(\mathbf{y}|\mathbf{x}, \theta) = \mathcal{N}\big(\mathbf{y}; \mu(\mathbf{x}; \theta_1), \Sigma(\mathbf{x}; \theta_2)\big)$ from all settings of the parameters θ_1, θ_2, weighted by the associated approximate posterior distribution $q_\phi(\theta_1, \theta_2)$. More formally, we need to compute the integral below:

$$q_\phi^*(\mathbf{y}|\mathbf{x}) \triangleq \int \underbrace{\mathcal{N}(\mathbf{y}; \mu(\mathbf{x}; \theta_1), \Sigma(\mathbf{x}; \theta_2)))}_{\text{Network prediction}} \cdot \underbrace{q_\phi(\theta_1, \theta_2)}_{\text{Approx. posterior}} \, \mathrm{d}\theta_1 \mathrm{d}\theta_2 \tag{9}$$

$$\approx \int p(\mathbf{y}|\mathbf{x}, \theta_1, \theta_2) \cdot p(\theta_1, \theta_2|D) \mathrm{d}\theta_1 \mathrm{d}\theta_2 = p(\mathbf{y}|\mathbf{x}, D) \tag{10}$$

where the last line represents the true predictive distribution $p(\mathbf{y}|\mathbf{x}, D)$ which is estimated by our model $q_\phi^*(\mathbf{y}|\mathbf{x})$. However, in practice, the integral $q_\phi^*(\mathbf{y}|\mathbf{x})$ cannot be evaluated in closed form because the likelihood $\mathcal{N}\big(\mathbf{y}; \mu(\mathbf{x}; \theta_1), \Sigma(\mathbf{x}; \theta_2)\big)$ is a highly non-linear function of input $\mathbf{x}$ as given in Eq. (2). At test time, we therefore estimate, for each input $\mathbf{x}$, the mean and covariance of the approximate predictive distribution $q_\phi^*(\mathbf{y}|\mathbf{x})$ with the unbiased Monte Carlo estimators:

$$\hat{\mu}_{\mathbf{y}|\mathbf{x}} \triangleq \frac{1}{T} \sum_{t=1}^{T} \mu(\mathbf{x}; \theta_1^t) \xrightarrow[T \to \infty]{} \mathbb{E}_{q_\phi^*(\mathbf{y}|\mathbf{x})}[\mathbf{y}] \tag{11}$$

$$\hat{\Sigma}_{\mathbf{y}|\mathbf{x}} \triangleq \frac{1}{T} \sum_{t=1}^{T} \Big(\Sigma(\mathbf{x}; \theta_2^t) + \mu(\mathbf{x}; \theta_1^t) \mu(\mathbf{x}; \theta_1^t)^{\mathrm{T}} \Big) - \hat{\mu}_{\mathbf{y}|\mathbf{x}} \hat{\mu}_{\mathbf{y}|\mathbf{x}}^{\mathrm{T}} \xrightarrow[T \to \infty]{} \mathrm{cov}_{q_\phi^*(\mathbf{y}|\mathbf{x})}[\mathbf{y}, \mathbf{y}] \tag{12}$$

where $\{(\theta_1^t, \theta_2^t)\}_{t=1}^T$ are samples of the network parameters (i.e. convolution kernels) drawn from the approximate posterior $q_\phi(\theta_1, \theta_2)$. In other words, the inference performs T stochastic forward passes at test time by injecting noise into features according to Eq. (8), and amalgamates the corresponding network outputs to compute the sample mean $\hat{\mu}_{\mathbf{y}|\mathbf{x}}$ and sample covariance $\hat{\Sigma}_{\mathbf{y}|\mathbf{x}}$. We use the sample mean $\hat{\mu}_{\mathbf{y}|\mathbf{x}}$ as the final prediction of an high-resolution output patch $\mathbf{y}$ and use the diagonal elements of the sample covariance $\hat{\Sigma}_{\mathbf{y}|\mathbf{x}}$ to quantify the corresponding uncertainty, which we refer to as *predictive mean* and *predictive uncertainty*, respectively.

3.6. Uncertainty decomposition and propagation

Predictive uncertainty arises from the combination of two source effects, namely intrinsic and parameter uncertainty, for which we have previously introduced methods for estimation. Lastly, we introduce a method based on variance decomposition for disentangling these effects and quantifying their contributions separately in predictive uncertainty. We consider such decomposition problem in the presence of an arbitrary transformation of the output variable $\mathbf{y}$.

The users of super-resolution algorithms are often interested in the quantities that are derived from the predicted high-resolution images, rather than the images themselves. For example, quantities such as the principal direction (first eigenvalue of the DT), mean diffusivity (MD) and fractional anisotropy (FA) are typically calculated from diffusion tensor images (DTIs) and used in the downstream analysis. We therefore consider an generic function[8] $g : \mathcal{Y} \to \mathbb{R}^m$ which transforms the high-resolution multi-channel data $\mathbf{y}$ to a quantity of interest of dimension m e.g. MD and FA maps, and propose a way to propagate the predictive uncertainty over $\mathbf{y}$ to the transformed domain (i.e. compute the variance of $p(g(\mathbf{y})|D, \mathbf{x})$) and decompose it into the "intrinsic" and "parameter" components. Specifically, by using the law of total variance[9]

[7] See the proof for feedforward networks given in Kingma et al. (2015) which generalises to convolutions.

[8] We assume here that the transform g is a measurable function with well-defined expectation and variance.

[9] The total law of variance (also known as Eve's Law) states that if random variables if A and B are random variables on the same probability space, and the mean and the variance of A and B are well-defined, then $\mathbb{V}[A] = \mathbb{V}[\mathbb{E}[A|B]] + \mathbb{E}[\mathbb{V}[A|B]]$.

(Weiss, 2006), we perform the following decomposition:

$$\mathbb{V}_{p(\mathbf{y}|\mathbf{x},D)}[g(\mathbf{y})] = \Delta_p(g(\mathbf{y})) + \Delta_i(g(\mathbf{y})) \tag{13}$$

where the respective component terms are defined as:

$$\Delta_p(g(\mathbf{y})) \triangleq \mathbb{E}_{p(\theta|D)}\left[\mathbb{V}_{p(g(\mathbf{y})|\theta,\mathbf{x},D)}[g(\mathbf{y})] - \mathbb{V}_{p(g(\mathbf{y})|\theta,\mathbf{x},D)}[g(\mathbf{y})|\theta]\right] \tag{14}$$

$$= \underbrace{\mathbb{V}_{p(\theta|D)}[\mathbb{E}_{p(g(\mathbf{y})|\theta,\mathbf{x},D)}[g(\mathbf{y})|\theta]]}_{\text{propagated parameter uncertainty}} \tag{15}$$

$$\Delta_i(g(\mathbf{y})) \triangleq \underbrace{\mathbb{E}_{p(\theta|D)}[\mathbb{V}_{p(g(\mathbf{y})|\theta,\mathbf{x},D)}[g(\mathbf{y})|\theta]]}_{\text{propagated intrinsic uncertainty}} \tag{16}$$

We refer to the components $\Delta_p(g(\mathbf{y}))$ and $\Delta_i(g(\mathbf{y}))$ as "propagated" parameter and intrinsic uncertainty. Intuitively, the first term quantifies the difference in variance between the cases where we have variable parameters and fixed parameters. In other words, this quantifies how much predictive uncertainty on the derived quantity arises, on average, from the variability in parameters. The second term on the other hand quantifies the average variance of the model prediction when the parameters are fixed, which signifies the model-independent uncertainty due to data i.e. intrinsic uncertainty. Assuming that the considered neural network is identifiable[10] and sufficiently complex to capture the underlying data generating process, as the amount of training data increases, the posterior $p(\theta|D)$ tends to a Dirac delta function and thus the first term diminishes to zero while the second term remains. A similar variance decomposition technique was employed in Bowsher and Swain (2012) to understand how the variation in cell signals of interest (e.g. gene expression) in a bio-chemical network is caused by the fluctuations of other environmental variables (e.g. transcription rate and biological noise). In our case, we employ the variance decomposition technique to separate the effects of network parameters from the intrinsic uncertainty in the prediction of $g(\mathbf{y})$.

We first consider a special case where the transform g is an identity map i.e. $g(\mathbf{y}) = \mathbf{y}$. Since the likelihood is modelled by a Gaussian distribution with heteroscedastic noise i.e. $p(\mathbf{y}|\theta_1, \theta_2, \mathbf{x}, D) = \mathcal{N}(\mathbf{y}; \mu(\mathbf{x}; \theta_1), \Sigma(\mathbf{x}; \theta_2))$, we see that the parameter and intrinsic uncertainty are given by

$$\Delta_p(\mathbf{y}) = \mathbb{V}_{p(\theta_1|D)}[\mu_{\theta_1}(\mathbf{x})], \quad \Delta_i(\mathbf{y}) = \mathbb{E}_{p(\theta_2|D)}[\Sigma_{\theta_2}(\mathbf{x})] \tag{17}$$

which can be approximated by the components of the MC variance estimator in Eq. (12) :

$$\widehat{\Delta}_p(\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} \mu(\mathbf{x}; \theta_1^t)\mu(\mathbf{x}; \theta_1^t)^{\mathrm{T}} - \hat{\mu}_{\mathbf{y}|\mathbf{x}}\hat{\mu}_{\mathbf{y}|\mathbf{x}}^{\mathrm{T}} \tag{18}$$

$$\widehat{\Delta}_i(\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} \Sigma(\mathbf{x}; \theta_2^t) \tag{19}$$

where $\{(\theta_1^t, \theta_2^t)\}_{t=1}^T$ are drawn from the approximate posterior $q_\phi(\theta_1, \theta_2)$.

More generally, when the transform g is complicated, MC sampling provides an alternative implementation. Given samples of model parameters $\{\theta_t\}_{t=1}^T \sim q(\theta|D)$ and $\{g_j^t\}_{j=1}^J \sim p(g(\mathbf{y})|\theta_t, \mathbf{x}, D)$ for $t = 1, \ldots, T$, we estimate both the progapated parameter and intrinsic uncertainty by simply using sample mean and sample variance:

$$\widehat{\Delta}_p(g(\mathbf{y})) \triangleq \frac{1}{T}\sum_t (\hat{\mu}^t)^2 - \Big(\frac{1}{(J-1)T}\sum_{j,t}(g_j^t)\Big)^2 \tag{20}$$

$$\widehat{\Delta}_i(g(\mathbf{y})) \triangleq \frac{1}{(J-1)T}\sum_{j,t}(g_j^t)^2 - \frac{1}{T}\sum_t (\hat{\mu}^t)^2 \tag{21}$$

$$\hat{\mu}^t = \frac{1}{J}\sum_j g_j^t. \tag{22}$$

These estimators are, although unbiased, higher in variance than the case where g is the identity (Eqs. (18) and (19)), due to two sources of sampling, thus requiring more samples for reliable estimation of respective uncertainty components.

4. Experiments and results

In this section, we evaluate the proposed uncertainty modelling techniques for super-resolution of diffusion MR images. First, we quantitatively study the effects of modelling uncertainty on the super-resolution performance by comparing our probabilistic CNN models against the relevant baselines in two different types of diffusion signal representations. Secondly, we evaluate the value of predictive uncertainty as a realiability metric of output images on multiple datasets of both healthy subjects and those with unseen pathological structures such as brain tumour (Glioma) and multiple sclerosis (MS).

4.1. Datasets

We make use of the following four diffusion MRI datasets to evaluate different benefits of the proposed technique:

- *Human connectome project dataset:* we use the diffusion MRI data from the WU-Minn HCP (release Q3) (Van Essen et al., 2013) as the source of the training datasets. The dataset enjoys very high image resolution, signal levels and coverage of the measurement space, enabled by the combination of custom imaging, reconstruction innovations and a lengthy acquisition protocol (circa 59 min) (Sotiropoulos et al., 2013). Each subject's data set contains 288 diffusion weighted images (DWIs) of voxel size 1.25^3 mm^3 of which 18 have nominal $b = 0$ and the three high-angular-resolution-diffusion-imaging (HARDI) shells of 90 directions have nominal b-values of 1000, 2000, and 3000 smm^{-2} (see Sotiropoulos et al., 2013 for the full acquisition details). The data are preprocessed by correcting distortions including susceptibility-induced, eddy currents and motion as outlined in Glasser et al. (2013).
- *Lifespan dataset:* this dataset (available online at http://lifespan.humanconnectome.org contains 26 subjects of much wider age range (8–75 years) than the main HCP cohorts (22–36 years), and is acquired with a shortened version of the main HCP protocol (circa 36 min) with lower resolution (1.5 mm isotropic voxels) and only two HARDI shells, with $b = 1000$ and 2500 smm^{-2}. Due to the differences in sequence timing and voxel resolution, Lifespan dataset has a different signal-to-noise ratio from HCP dataset. However, we also note that the protocol still leverages the special features of the HCP scanners, providing images of substantially better quality than standard sequences. In this work, we focus the elderly subjects (45–75 years old) and utilise this out-of-training-distribution dataset to assess the robustness of our techniques to domain shifts. It is well known and widely accepted that there are significant differences in numerous diffusion metrices between young adults and elderly subjects. Most notably, the DTI fractional anisotropy (FA) is significantly lower and the DTI mean diffusivity (MD) is significantly higher in elderly subjects compared to young adults. We refer the interested readers to references Westlye et al. (2009), Lebel et al. (2012), and Salat (2014) for more details.
- *Prisma dataset:* two healthy male adults (29 and 33 years old respectively) were scanned twice at different image resolutions using the clinical 3T Siemens Prisma scanner in FMRIB, Oxford. Both datasets contain diffusion MRI data with 21 $b = 0$ images and three 90-direction HARDI shells, b-values of 1000, 2000, and 3000 smm^{-2},

[10] We note that a neural network is, in general, not identifiable i.e. there exist more than a single set of parameters that capture the same target distribution $p(g(\mathbf{y})|\mathbf{x})$. In such cases, the posterior distribution $p(\theta|D)$ does not collapse to a single Dirac Delta function with infinite amount of observations—it rather converges to a mixture of all sets of network parameters Θ such that $p(g(\mathbf{y})|\theta^*, \mathbf{x}) = p(g(\mathbf{y})|\mathbf{x}) \forall \theta^* \in \Theta$. However, the expectation $\mathbb{E}_{p(g(\mathbf{y})|\theta,\mathbf{x},D)}[g(\mathbf{y})|\theta]$ is the same for all $\theta \in \Theta$ and thus the propagated parameter uncertainty $\Delta_p(g(\mathbf{y}))$ converges to zero.

Table 1
Details of training data for two diffusion MR signal representations, DTIs and MAP-MRIs. The first two columns from the right denote the size of the input **x** and output patches **y** of dimension [width, height, depth, channels] while the third and the fourth columns show the number of patch pairs (**x, y**) extracted from each subject, and the total number of training subjects used, respectively.

Data	Size of input **x**	Size of output **y**	No. pairs (**x, y**) per subject	No. subjects
DTIs	$11 \times 11 \times 11 \times 6$	$14 \times 14 \times 14 \times 6$	8000	16
MAP-MRIs	$21 \times 21 \times 21 \times 22$	$14 \times 14 \times 14 \times 22$	4000	16

each for two resolutions, 2.50 mm and 1.35 mm isotropic voxels (see Alexander et al., 2017 for full acquisition details). In addition, each of these datasets also includes a standard 3D T1-weighted MPRAGE (1 mm isotropic resolution). The Prisma scanner is less powerful than the bespoke HCP scanner and cannot achieve sufficient signal at 1.25 mm resolution, but the 1.35 mm data provides a pseudo ground-truth for IQT resolution enhancement of the 2.5 mm data.

- *Pathology dataset:* we use two separate datasets which consist of images of brain tumour (Glioma) (Figini et al., 2018) and multiple sclerosis (MS) patients, respectively. The data of each wubject with glioma contains DWIs with $b = 700$ s/mm^2 while the measurement of each MS patient is of $b = 1200$ s/mm^2. Both datasets have isotropic voxel size 2^3 mm^3, which is closer to the image resolution of commonplace clinical scanners. We use these datasets to assess the behaviour of predictive uncertainty on images with pathological features that are not represented in the training data set.

In all the experiments, super-resolution is performed on diffusion parameter maps derived from the DWIs in the above datasets. In particular, we consider two diffusion MRI models, namely the diffusion tensor (DT) model (Basser et al., 1994) and Mean Apparent Propagator (MAP) MRI (Özarslan et al., 2013), where the former is the simplest and most standard diffusion parameter map, and the latter is a high-order generalisation of the former with the capacity to characterise signals from more complex tissue structures (e.g. fibre crossing regions), a requirement for successful tractography applications. We compute both of these diffusion parameter maps using the implementation from Alexander et al. (2017), which is available at https://github.com/ucl-mig/iqt.

We fit the DT model to the combination of $b = 0$ images and $b = 1000$ s/mm^2 HARDI shell for the HCP and Lifespan datasets, and $b = 700$ s/mm^2 shell for the brain tumour dataset. In all cases, weighted linear least squares are employed for the fitting, taking into account the spatially varying b-values and gradient directions in the HCP dataset. On the other hand, in the case of MAP-MRI, 22 coefficients of basis functions up to order 4 are estimated via (unweighted) least squares to all three shells of the HCP, Lifespan and Prisma datasets. As noted in Alexander et al. (2017), the choice of scale parameters (see Özarslan et al., 2013) $\mu_x = \mu_y = \mu_z = 1.2 \times 10^{-3}$ mm empirically minimises the fitting error in the HCP dataset, and is used for all datasets.

Training datasets in all experiments are constructed by artificially downsampling very high-resolution images in the HCP dataset. In particular, we employ the following downsampling procedure: (i) the raw DWIs of selected subjects are blurred by applying the mean filter of size $r \times r \times r$ independently over channels with r denoting the upsampling rate; (ii) the DT or MAP parameters are computed for every voxel; (iii) the spatial resolution of the resultant parameter maps are reduced by taking every r pixels. A coupled library of low-resolution and high-resolution patches is then constructed by associating each patch in the downsampled DTI/MAP-MRI with the corresponding patch in the ground truth DTI or MAP-MRI. In this case, we ensure the low-resolution patch to be centrally and entirely contained within the corresponding high-resolution patch (as illustrated by the yellow and orange squares in Fig. 3). We then randomly select a pre-set number of patches from each subject in the training pool to create a training dataset as detailed in Table 1. In addition to the 8 subjects used in the prior work (Alexander et al., 2014; Tanno et al., 2016; 2017), we randomly select additional 8 subjects from the HCP cohort and include them in the training subject pool. Patches are standardised channel-wise by subtracting the mean of foreground pixel intensities of the corresponding subject and dividing by its standard deviation. Moreover, since MAP-MRI datasets contain outliers due to model fitting, in large enough quantity to influence the training of the baseline 3D-ESPCN model, we remove them by clipping the voxel intensity values of the respective 22 channels separately at 0.1% and 99.9% percentiles computed over all the foreground voxels in the whole training dataset.

4.2. Network architectures and training

For the training of all CNN models, we minimised the associated loss function using Adam (Kingma and Ba, 2014) for 200 epochs with initial learning rate of 10^{-3} and $\beta = [0.9, 0.999]$, with minibatches of size 12. We hold out 50% of training patch pairs as a validation set. The best performing model was selected based on the mean-squared-error (MSE) on the validation set.

For the super-resolution of DTIs, as in Shi et al. (2016), we use a minimal architecture for the baseline 3D-ESPCN, consisting of three $3D$ convolutional layers with filters $(3^3, 50) \rightarrow (1^3, 100) \rightarrow (3^3, 6r^3)$ where r is upsampling rate and 6 is the number of channels in DTIs. As illustrated in Fig. 3, the dimensions of convolution filters are chosen, so each $5^3 \cdot 6$ low-resolution receptive field patch maps to a $r^3 \cdot 6$ high-resolution patch, which mirrors competing random forest based methods (Alexander et al., 2014; Tanno et al., 2016) for a fair comparison. On the other hand, for MAP-MRI, which is a more complex image modality with 21 channels, we employ a deeper model with 6 convolution layers $(5^3, 256) \rightarrow (3^3, 256) \rightarrow (3^3, 128) \rightarrow (3^3, 128) \rightarrow (3^3, 64) \rightarrow (3^3, 21r^3)$ prior to the shuffling operation, which expands the receptive field on each $r^3 \cdot 21$ high-resolution patch to $15^3 \cdot 21$ input low-resolution patch. Every convolution layer is followed by a ReLU non-linearity except the last one in the architecture, and batch-normalisation (Ioffe and Szegedy, 2015) is additionally employed for MAP-MRI super-resolution between convolution layer and ReLU non-linearity.

The mean and variance networks in the heteroscedastic noise model introduced in Section 3.3 are implemented as two separate baseline 3D-ESPCNs of the architectures, specified above for DTIs and MAP-MRIs. Positivity of the variance is enforced by passing the output through a softplus function $f(x) = \ln(1 + e^x)$ as in Lakshminarayanan et al. (2017).

For variational dropout, we considered two flavours: Var.(I) optimises per-weight dropout rates, and Var.(II) optimises per-filter dropout rates. More formally, the "drop-out rate" α_{ij} in the approximate posterior $q_\phi(\theta_{ij}) = \mathcal{N}(\theta_{ij}; \eta_{ij}, \alpha_{ij}\eta_{ij}^2)$ is different for every element in each convolution kernel in the former while the latter has common α_{ij} shared across each kernel. In preliminary analysis, we found that the number of samples per data point for estimating reconstruction term (Eq. (7)) can be set to $S = 1$ so long as the batch size is sensibly large ($M = 12$).

We also note the default training with binary and Gaussian dropout also employs $S = 1$ (Srivastava et al., 2014) along with other MC variational inference methods for neural networks such as Kingma and

Welling (2014), Kingma et al. (2015), and Gal et al. (2017a). Variational dropout is applied to both the baseline and heteroscedastic models without changing the architectures. For both binary and Gaussian dropout modes, we incorporate the dropout operations of fixed rate p in every convolution layer of the baseline 3D-ESPCN architecture.

All models are trained on simulated datasets generated from 16 HCP subjects as detailed in Section 4.1. We also retrained the random forest models employed in Tanno et al. (2016), Alexander et al. (2017) on equivalent datasets. It takes under 60/360 min to train a single network on DTI/MAP-MRI data on a single TITAN X GPU. All models are implemented in the TensorFlow framework (Abadi et al., 2016) and the codes will be released at https://github.com/rtanno21609/UncertaintyNeuroimageEnhancement.

4.3. Benefits on super-resolution performance

We evaluate the effects of modelling different components of uncertainty on the prediction performance of our models for super-resolution of DTI and MAP-MRI on two datasets—HCP and Lifespan as detailed in Section 4.1. The first dataset contains 16 unseen subjects from the same HCP cohort used for training, while the second one consists of 10 subjects from the HCP Lifespan dataset. The latter tests generalisability, as they are acquired with a different protocol at lower resolution (1.5 mm isotropic), and contain subjects of a different age range (45–75 years) to the original HCP data (22–36 years). We perform $\times 2$ upsampling in all spatial directions. The reconstruction quality is measured with root-mean-squared-error (RMSE), peak-signal-to-noise-ratio (PSNR) and mean-structural-similarity (MSSIM) (Wang et al., 2004) on two separate regions: (i) "interior"; the set of pixels whose the 5^3 cubic neighbourhood is entirely contrained within the brain mask; (ii) "exterior"; the remaining set of pixels in the brain mask, as shown in Fig. 6. This is because the current state-of-the-art methods based on random forests (RFs) such IQT-RF (Alexander et al., 2017) and BIQT-RF (Tanno et al., 2016) are only trained on patches from the interior region and requires a separate procedure on the brain boundary. In addition, the estimation problem is quite different in boundary regions, but remains valuable particularly for applications such as tractography where seed or target regions are often in the cortical surface of the brain. We only present the RMSE results, but the derived conclusions remain the same for the other two metrics (see Section C in the Supplementary materials). Aside from the interpolation techniques, for each method an ensemble of 10 models are trained on different trainings set (generated by randomly extracting patch pairs from the common 16 HCP training subjects) and for each model, the average error metric over the test subjects are first calculated. The mean and standard deviations of such average errors are computed across the model ensemble and reported in Tables 2 and 3.

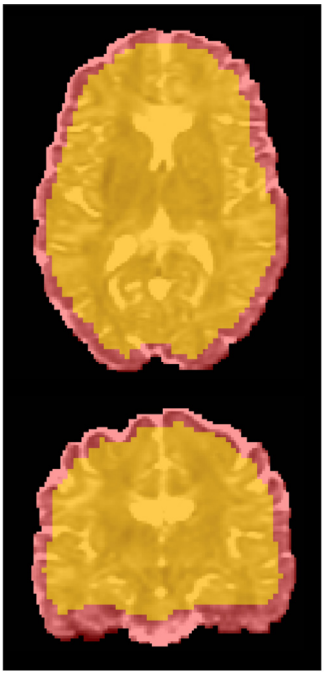

Fig. 6. Visualisation of "interior" (yellow) and "exterior" regions (red). The interior region consists of a set of patches contained entirely within the brain while the exterior region consists of partial patches that contain mixtures of brain and background voxels.

Table 2 shows that our baseline achieves 8.5%/39.8% reduction in RMSE for the super-resolution of DTIs on the HCP dataset on the interior/exterior regions with respect to the best published method, BIQT-RF (Tanno et al., 2016). While the standard deviations are higher, the improvements are more pronounced in MAP-MRI super-resolution, reducing the average RMSEs by 49.6% and 63.5% on the interior and exterior regions. We note that that IQT-RF and BIQT-RF are only trained on interior patches, and super-resolution on boundary patches requires a separate *ad hoc* procedure. Despite including exterior patches in training our model, which complicates the learning task, the baseline CNN out-performs the RF methods on both regions. We see similar improvements in the out-of-distribution Lifespan dataset.

Reconstruction is faster than the RF baselines; the 3D-ESPCN is capable of estimating the whole high-resolution DTI/MAP-MRI under 10/60 seconds on a CPU and 1/10 second(s) on a GPU. On the other hand, BIQT-RF takes $\sim$ 10 min with 8 trees on both DTIs and MAP-MRIs. The fully convolutional architecture of the model enables to process input patches of different size from that of training inputs, and we achieve faster reconstruction by using larger input patches of dimension $25^3 \cdot c$ where c is the number of channels. We also note that the reconstruction time of the variational dropout based models increases by a factor of the number of MC samples used at test time, although it is possible, with more memory, to leverage GPU parallelisation by making multiple copies of each input patch and treating them as a mini-batch. On the other hand, the heteroscedastic CNN enjoys the same inference speed of the baseline since only the mean network is used for reconstruction (the covariance network is only employed to quantify the estimated intrinsic uncertainty).

Table 2 shows that, on both HCP and Lifespan data, modelling both intrinsic and parameter uncertainty (i.e. Hetero. + Variational Dropout (I), (II)) achieves the best reconstruction accuracy in DTI super-resolution. We observe that modelling intrinsic uncertainty with the heteroscedastic network on its own further reduces the average RMSE of the baseline 3D-ESPCN on the interior region with high statistical significance ($p < 10^{-3}$). However, poorer performance is observed on the exterior than the baseline. On the other hand, using 200 MC weight samples[11], we see modelling parameter uncertainty with variational dropout (see Variational Dropout.(I)-CNN) performs best on both datasets on the exterior region. Combination of heteroscedastic model and variational dropout (i.e. Hetero. + Variational Dropout (I) or (II)) leads to the top 2 performance on both datasets on the interior region and reduces errors on the exterior to the level comparable or better than the baseline.

Similarly, Table 3 shows that the best performance in MAP-MRI super-resolution comes from the combined models (i.e. Hetero. + Variational Dropout.(I) and (II)). We observe that as with the DTI case, modelling intrinsic uncertainty through the heteroscedastic network improves the reconstruction accuracy on the interior region, whilst the errors on the exterior are increased with respect to the baseline 3D-ESPCN. Moreover, the improvement is pronounced when the outliers

[11] We observed that drawing more than 200 MC samples at inference time was sufficient for the average RMSE and its standard deviation to converge.

Table 2
Super-resolution results on diffusion tensor images (DTIs) of HCP and Lifespan datasets for different upsampling methods. For each method, an ensemble of 10 models are trained on different training sets generated by randomly extracting a set of patch pairs from the common 16 HCP subjects. For each model, the average RMSE ($\times 10^{-4}$mm^2/s) over subjects in respective datasets is first computed and the mean/std of such average RMSE over the ensemble are then reported. Best results in bold red, and the second best in blue.

Models	HCP (interior)	HCP (exterior)	Life (interior)	Life (exterior)
CSpline-interpolation	10.069 ± n/a	31.738 ± n/a	32.483 ± n/a	49.066 ± n/a
β-Spline interpolation	9.578 ± n/a	98.169 ± n/a	33.429 ± n/a	186.049 ± n/a
IQT-RF	6.974 ± 0.024	23.139 ± 0.351	10.038 ± 0.019	25.166 ± 0.328
BIQT-RF	6.972 ± 0.069	23.110 ± 0.362	9.926 ± 0.055	25.208 ± 0.290
3D-ESPCN(baseline)	6.212 ± 0.017	13.609 ± 0.084	8.902 ± 0.020	16.389 ± 0.114
+ Binary Dropout ($p = 0.1$)	6.319 ± 0.015	13.738 ± 0.048	9.093 ± 0.024	16.489 ± 0.099
+ Gaussian Dropout ($p = 0.05$)	6.463 ± 0.034	14.168 ± 0.051	9.184 ± 0.048	16.653 ± 0.092
+ Variational Dropout (I)	6.194 ± 0.013	**13.412 ± 0.041**	8.874 ± 0.027	**16.147 ± 0.051**
+ Variational Dropout (II)	6.201 ± 0.015	13.479 ± 0.047	8.878 ± 0.031	16.230 ± 0.075
+ Hetero.	6.135 ± 0.029	15.469 ± 0.231	8.885 ± 0.041	17.208 ± 0.211
+ Hetero. + Variational Dropout (I)	6.121 ± 0.015	13.591 ± 0.051	**8.837 ± 0.043**	16.261 ± 0.053
+ Hetero. + Variational Dropout (II)	**6.116 ± 0.013**	13.622 ± 0.099	8.861 ± 0.031	16.387 ± 0.098

Table 3
Super-resolution results on MAP-MRIs of HCP and Lifespan datasets for different upsampling methods. For each method, an ensemble of 5 models are trained on different training sets generated by randomly extracting a set of patch pairs from the common 16 HCP subjects. For each model, the average RMSE over subjects in respective datasets is first computed and the mean/std of such average RMSEs ($\times 10^{-2}$) over the ensemble are then reported. Best results in bold red, and the second best in blue. In addition, the performance of 3D-ESPCN and its probabilistic variants trained on data without outlier removal are also included.

Models	HCP (interior)	HCP (exterior)	Life (interior)	Life (exterior)
CSpline interpolation	5.234 ± n/a	30.362 ± n/a	7.135 ± n/a	29.232 ± n/a
β-Spline interpolation	4.852 ± n/a	63.446 ± n/a	6.523 ± n/a	56.937 ± n/a
IQT-RF (Alexander et al., 2017)	4.538 ± 0.113	25.541 ± 0.131	5.882 ± 0.121	26.137 ± 0.279
BIQT-RF (Tanno et al., 2016)	4.838 ± 0.129	25.523 ± 0.175	5.949 ± 0.131	27.509 ± 0.233
3D-ESPCN(baseline)	2.285 ± 0.126	9.316 ± 0.127	4.195 ± 0.163	11.922 ± 0.192
+ Binary Dropout ($p = 0.1$)	2.283 ± 0.154	9.272 ± 0.132	4.120 ± 0.178	11.652 ± 0.204
+ Gaussian Dropout ($p = 0.1$)	2.370 ± 0.155	9.335 ± 0.144	4.327 ± 0.157	11.907 ± 0.211
+ Variational Dropout (I)	2.155 ± 0.122	9.205 ± 0.193	3.997 ± 0.153	11.547 ± 0.177
+ Variational Dropout (II)	2.172 ± 0.128	9.112 ± 0.173	3.972 ± 0.132	11.511 ± 0.172
+ Hetero.	1.998 ± 0.132	11.294 ± 0.216	3.872 ± 0.140	12.084 ± 0.129
+ Hetero + Variational Dropout (I)	**1.951 ± 0.122**	9.102 ± 0.181	**3.572 ± 0.171**	**11.037 ± 0.192**
+ Hetero + Variational Dropout (II)	1.969 ± 0.119	**9.052 ± 0.162**	3.606 ± 0.141	11.311 ± 0.195
3D-ESPCN(without outlier removal)	3.425 ± 0.163	13.284 ± 0.239	6.032 ± 0.229	15.513 ± 0.273
+ Hetero.	2.264 ± 0.153	11.306 ± 0.172	3.919 ± 0.140	12.821 ± 0.150
+ Hetero + Variational Dropout (I)	2.138 ± 0.159	10.022 ± 0.187	3.681 ± 0.193	12.133 ± 0.205
+ Hetero + Variational Dropout (II)	2.133 ± 0.188	9.988 ± 0.209	3.690 ± 0.184	12.052 ± 0.212

due to model fitting errors are not removed in the training data. In this case, we see that the reconstruction accuracy of 3D-ESPCN dramatically decreases, whilst in contrast it is only marginally compromised when equipped with the heteroscedastic noise model, displaying robustness to outliers. Lastly, we note that the top-2 accuracy are consistently achieved by the joint modelling of intrinsic and parameter uncertainty (i.e. Hetero. + Variational Dropout.(I) and (II)) on both the interior and exterior regions on both HCP and Lifespan datasets.

The performance difference of heteroscedastic network between the interior and the exterior region roots from the loss function. The Mahalanobis term $\mathcal{M}_\theta(D)$ in Eq. (5) imposes a larger penalty on the regions with smaller intrinsic uncertainty. The network therefore allocates less of its resources towards the regions with higher uncertainty (e.g. boundary regions) where the statistical mapping from the low-resolution to high-resolution space is more ambiguous, and biases the model to fit the regions with lower uncertainty. However, we note that the performance of the heteroscedastic network is still considerably better than the standard interpolation and RF-based methods. By augmenting the model with variational dropout, the exterior error of the heteroscedastic model is dramatically reduced, indicating the "smoothing" regularisation effect of dropout against overfitting to low-uncertainty areas (i.e., regions with high data frequency). We also observe concomitant performance improvement on the interior regions on both datasets, which additionally shows the benefits of such regularisation even in low-uncertainty areas.

Both Table 2 and Table 3 show that the use of variational dropout attains lower errors than the models with fixed dropout probabilities p, namely, Binary and Gaussian dropout (Srivastava et al., 2014). Different instances of both dropout models are trained for a range of p by linearly increasing on the interval [0.05,0.3] with increment 0.05, and the test errors for the configurations with smallest RMSE on the validation set are reported in Tables 2 and 3. As with variational dropout models, 200 MC samples are used for inference. In all cases, two variants of variational dropout (I) and (II) outperform the networks with the best binary or Gaussian dropout models, showing the benefits of learning dropout probabilities p rather than fixing them in advance. We should also note that the dropout operation is commonly turned off during test time. Such point estimate approach based on mean approximation marginally reduces the average reconstruction accuracy, as shown in a wide range of applications in Gal and Ghahramani (2015), and does not give an estimate of the predictive variance.

Lastly, to test the real world utility of the observed improvement in reconstruction performance, we further assessed the benefits of super-resolution with a tractography experiment on the Prisma dataset, which contains two DWIs of the same subject at two different image resolutions—1.35 mm and 2.5 mm isotropic voxels, as detailed in Section 4.1. Fig. 13 in the Supplementary materials shows that IQT via our best performing CNN (3D-ESPCN + Hetero. + Variational Dropout (I)) makes a tangible difference in downstream tractography. For more details, we refer the readers to Section E in the Supplementary material.

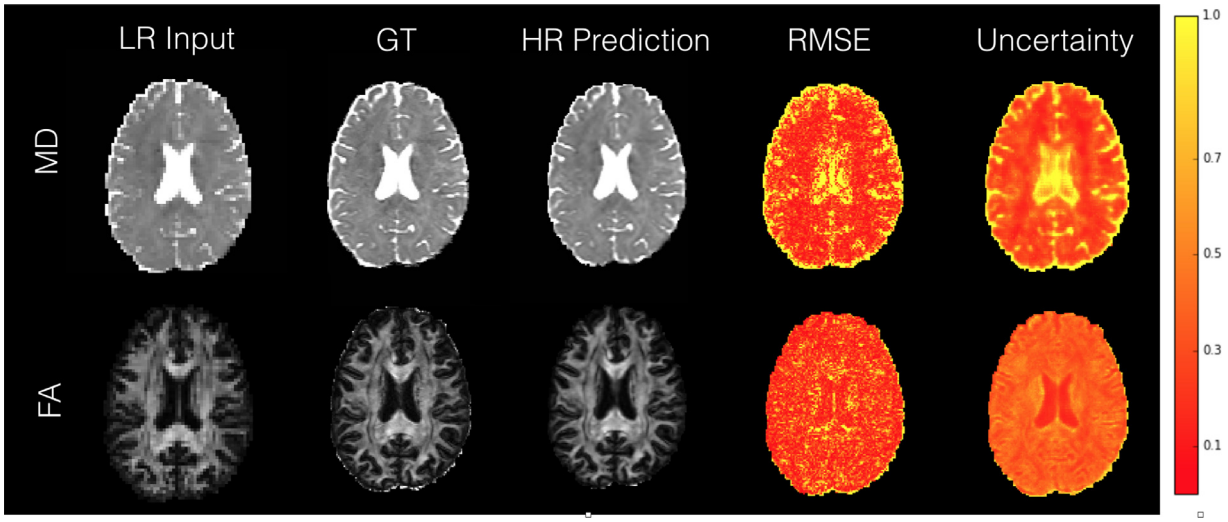

Fig. 7. Comparison between voxel-wise RMSE and predictive uncertainty maps for FA and MD computed on a HCP test subject (min-max normalised for MD and FA separately). Low-res input, ground truth and the mean of high-resolution predictions are also shown.

4.4. Reliability assessment of model predictions

In this section, we investigate the utility of uncertainty modelling in quantifying and understanding the reliability of model predictions. Firstly, in Section 4.4.1, we investigate the utility of the derived predictive uncertainty map as a proxy measure of reconstruction accuracy on healthy test subjects from both HCP and Lifespan datasets. Secondly, in Section 4.4.2, we study the behaviours of uncertainty maps in the presence of abnormal features that are not present in the training data.

4.4.1. Healthy test subjects

We employ the most performant CNN model (3D-ESPCN + Hetero. + Variational Dropout(I)) to generate the high-resolution predictions of *mean diffusivity* (MD) and *fractional anisotropy* (FA), and their associated predictive uncertainty maps. Here we draw 200 samples of high-resolution DTI predictions for each subject from the predictive distribution $q^*_\phi(\mathbf{y}|\mathbf{x})$, and then the FA and MD maps of each prediction are computed. The sample mean and standard deviation are then calculated from these samples to generate the final estimates of high-resolution MD/FA maps and their corresponding predictive uncertainty.

Fig. 7 displays high correspondence between the error (RMSE) maps and the predictive uncertainty on both FA and MD of a HCP test subject. This demonstrates the potential utility of uncertainty map as a surrogate measure of prediction accuracy. In particular, the MD uncertainty map captures subtle variations within the white matter and the cerebrospinal fluid (CSF) at the centre. Also, in accordance with the low reconstruction accuracy, high predictive uncertainty is observed in the CSF in MD. This is expected since the CSF is essentially free water with low signal-to-noise-ratio (SNR) and is also affected by biological noise such as cardiac pulsations. The reconstruction errors are high in FA prediction on the bottom-right quarter of the brain boundary, close to the skull, which is also reflected in the uncertainty map.

Fig. 7 also shows strong correlation between the intensity value of the prediction and the predictive uncertainty. This is expected since the error map itself correlates strongly with the intensity values. However, we note this is not always the case and we would like to point to some example cases. For instance, we observe in the top row of Fig. 7 that uncertainty is lower in a region of grey matter with higher MD intensity, and also captures the subtle variations in accuracy within the central CSF, which has approximately uniform intensity. Similarly, the FA map in the bottom row shows that the uncertainty on the bottom right brain boundary is particularly high in accordance with high RMSE, while the FA values there are very low. We also observe similar behaviours even in the presence of abnormalities (as described in greater detail in the subsequent section): Fig. 11 (b) shows that the propagated parameter uncertainty assigns higher or comparable degree of uncertainty to the MS lesions than the central CSF which has significantly higher MD values.

Fig. 8 tests the utility of predictive uncertainty map in discriminating potential predictive failures in the predicted high-resolution MD map. We define ground truth "safe" voxels as the ones with reconstruction error (RMSE) smaller than a fixed value, and the task is to separate them from the remaining ground-truth "risky" voxels by thresholding on their predictive uncertainty values. The threshold for defining safe voxels is set to 1.5×10^{-4} s/mm^2, such that the risky voxels mostly concentrate on the outer-boundary and the CSF regions (which account for 17.5% of all voxels under consideration). Here the positive class is defined as "safe" while the negative class is defined as "risky". Fig. 8 (a) shows the corresponding receiver operating characteristic (ROC) curve of such binary classification task, which plots the true-positive-rate (TPR) against the false-positive-rate (FSR) computed based on all the voxels in the 16 HCP training subjects. In this case, TPR decribes the percentage of correctly detected safe voxels out of all the safe ones, while FPR is defined as the percentage of risky voxels that are wrongly classified as safe out of all the risky voxels. We then select the best threshold by maximising the F1 score, and use this to classify the voxels in each predicted high-resolution MD into "safe" and "risky" ones for all subjects in the test HCP dataset and the Lifespan dataset. Fig. 8 (b) shows the inter-subject average of the TPR and FPR on both datasets. While on average TPR slightly worsens compared to the results on the training subjects, FPR improves in both cases—notably, this uncertainty-based classification is able to correctly identify 96% of risky predictions on unseen subjects from out-of-training-distribution dataset, namely Lifespan, which differs in demographics and underlying acquisition. Fig. 8 (c) visualises the classification results to the pre-defined "ground truth" on one of the Lifespan subjects, which illustrates that the generated "warning" aggressively flags potentially risky voxels at the cost of thresholding out the safe ones.

Lastly, we have also performed the same quantitative analysis on the FA map (see Fig. 12 in Section B in the Supplementary material). We observe similar results with the optimal threshold on predictive uncertainty, achieving the average detection rate of 10% on both "within-distribution" HCP test set and 8% on the "out-of-distribution" Lifespan data set. The slight increase in the TRP can be ascribed to the more noisy

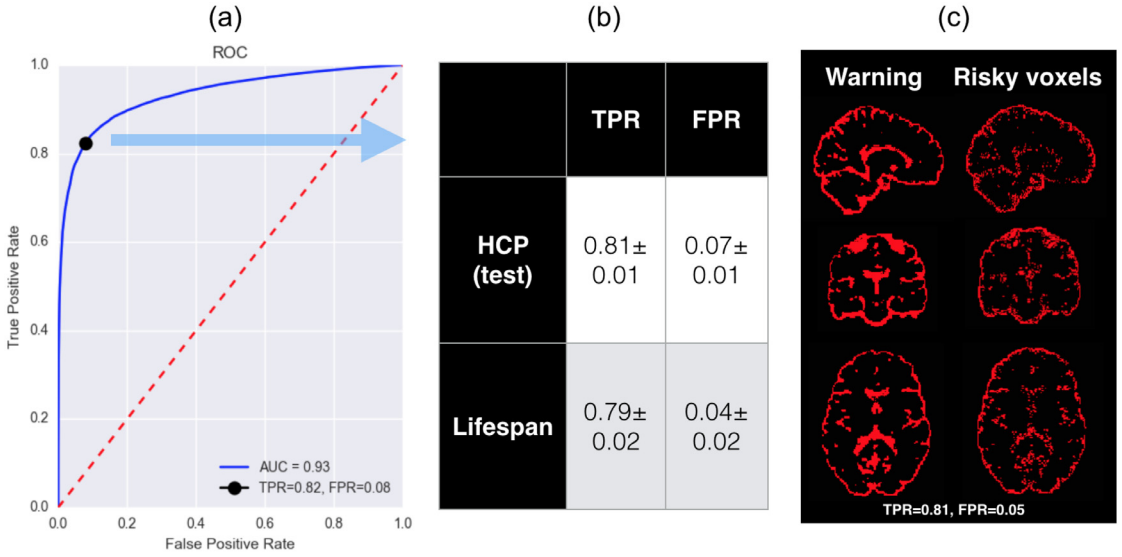

Fig. 8. Discrimination of "safe" voxels in the predicted high-resolution MD map by thresholding on predictive uncertainty. Here a single 3D-ESPCN + Hetero. + Variational Dropout (I) model is used to quantify the predictive uncertainty over each image volume. (a) The ROC curve plots the true positive rate (TPR) against false positive rate (FPR) computed for a range of threshold values on the foreground voxels in the training subjects. Best threshold (black dot) was selected such that F1 score is maximised and is employed to separate "safe" voxels from "risky" ones; (b) the average TPR and FPR over the 16 test HCP subjects and the 16 Lifespan subjects are shown; (c) an example visualisation of the "ground truth" safe (black) and risky (red) voxels on a Lifespan subject along with the corresponding classification results denoted as "warning".

distribution of the high error voxels. In this case, a large proportion of "risky" voxels concentrate on the brain parenchyma and structural boundary, which can also be detected with reasonable accuracy.

4.4.2. *Unseen abnormalities and uncertainty decomposition*

We separately visualise the propagated intrinsic and parameter uncertainty over the predicted high-resolution MD map on images of subjects with a variety of different unseen abnormal structures, such as benign cysts, tumours (Glioma) and focal lesions caused by multiple sclerosis (MS). We emphasise here that all these images have been acquired with different protocols. Specifically, benign cysts in the HCP datasets represent abnormalities in images acquired with the same protocol as the training data, while tumours and MS lesions are examples of pathologies present in out-of-distribution imaging protocols. In all cases, we use the SR network, Hetero.+Variational Dropout (I), trained on healthy subjects from HCP dataset. For each of 200 different sets of parameters $\{\theta_t\}_{t=1}^{200}$ sampled from the posterior distribution $q(\theta|D)$, we draw 10 samples of high-resolution DTIs from the likelihood, $\{\mathbf{y}_j^t\}_{j=1}^{10} \sim p(\mathbf{y}|\theta_t, \mathbf{x}, D)$, compute the corresponding MD, and approximate the two constituents of predictive uncertainty with the MC estimators given in Eqs. (20) and (21).

Fig. 9 shows the reconstruction accuracy along with the components of predictive uncertainty over the high-resolution MD map of a HCP test subject, which contains a benign abnormality (a small posterior midline arachnoid cyst). The error (RMSE) and propagated intrinsic uncertainty are plotted on the same scale whereas the propagated model uncertainty is plotted on 1/5 of the scale for clear visualisation. In this case, the predictive uncertainty is dominated by the intrinsic component. In particular, low propagated intrinsic uncertainty is observed in the interior of the cyst relative to its boundary in accordance with the high accuracy in the region. This is expected as the interior structure of a cyst is highly homogeneous with low variance in signals and the super-resolution task should therefore be relatively straightforward. On the other hand, the component of parameter uncertainty is high on the interior structure which also makes sense as such homogeneous features are underrepresented in the training data of healthy subjects. This example illustrates how decoupling the effects of intrinsic and parameter uncertainty potentially allows one to make sense of the predictive performance.

Fig. 10 visualises the uncertainty components generated by the same CNN model trained on datasets of varying size. We see that the propagated parameter uncertainty diminishes as the training set size increases, while the propagated intrinsic uncertainty stays more or less constant. This result is indeed what is expected as described in Fig. 1; the specification of network weights becomes more confident i.e. the variance of the posterior distribution decreases as the amount of training data increases, while the effect of intrinsic uncertainty is only determined by the underlying problem and is irreducible with the amount of data. On the other hand, when the standard binary or Gaussian dropout was employed instead of variational dropout, we observed that the effect of parameter uncertainty stayed more or less constant with the size of training data. This may be a consequence of the posterior variance largely determined by the prespecified drop-out rates, which in turn results in more static variance of predictive distribution.

We further validate our method on clinical images with previously unseen pathologies. We note that the pathology data contain images acquired with standard clinical protocols with voxel size slightly smaller than that of the training low-resolution images and lower signal-to-noise ratio.

Fig. 11 shows that pathological areas not represented in the training set are flagged as highly uncertain. Although the ground truth is not available in this case, the uncertainty can be quantified instead to flag potential low accuracy areas. Fig. 11 (a) shows that the propagated parameter uncertainty highlights the tumour core, and speckly artefacts in the input image, which are not represented in the training data. On the other hand, the intrinsic uncertainty component is high on the whole region of pathology covering both the tumour core and its surrounding edema. Fig. 11 (b) shows that high parameter uncertainty is assigned to a large part of focal lesions in MS, while the intrinsic uncertainty is mostly prevalent around the boundaries between anatomical structures and CSF. We also observe that the super-resolution sharpens the original image without introducing noticeable artifacts; in particular, for the brain tumour image, some of the partial volume effects are cleared.

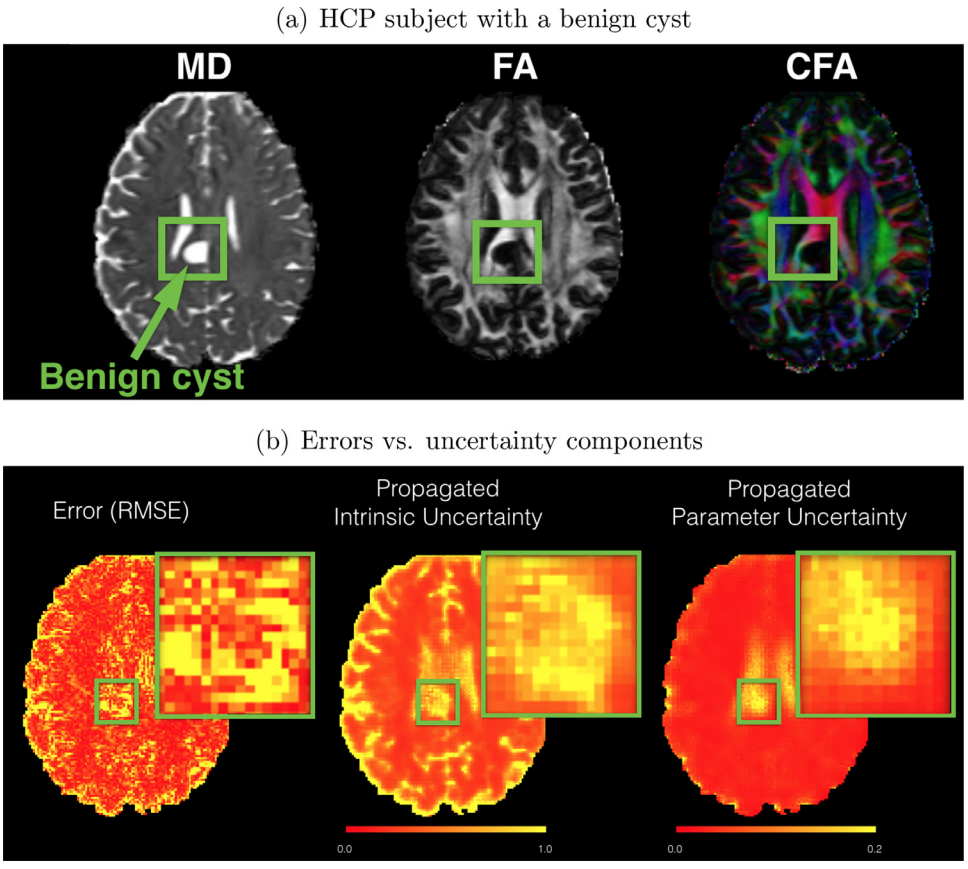

Fig. 9. Visualisation of (a) MD, FA and colour FA maps computed from the DTI of a HCP subject with a small posterior midline arachnoid cyst in the central part of the brain. (b) the corresponding reconstruction accuracy (RMSE) in MD and the corresponding components of predicted uncertainty.

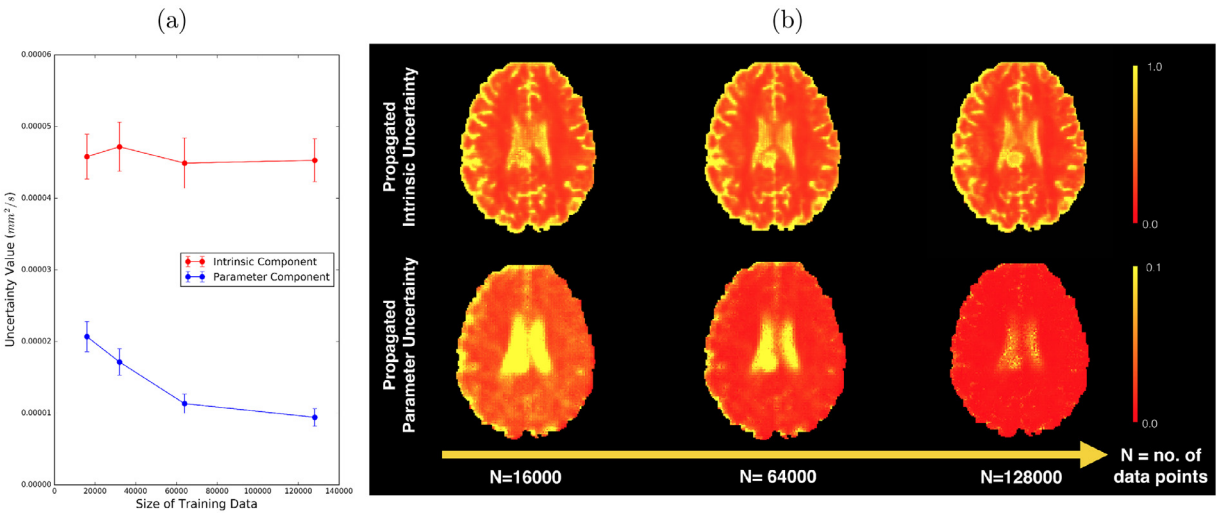

Fig. 10. Training set size vs propagated intrinsic/parameter uncertainty. (a) shows the quantitative results on the whole HCP test population. For a fixed training data size, an ensemble of 10 each 3D-ESPCN + Hetro. + Variational Dropout (I) models are trained on different training sets generated by randomly extracting a set of patch pairs from the common 16 HCP training subjects. The average uncertainty components from each model are first computed over the HCP test subjects, and the mean/std of such average uncertainty values over the model ensemble are then reported. (b) visualises the respective uncertainty components from a single model on the MD map of an unseen HCP subject with a benign cyst. The uncertainty maps are normalised across all the figures in each row.

5. Discussion and conclusion

We introduce a probabilistic deep learning (DL) framework for quantifying three types of uncertainties that arise in data-enhancement applications, and demonstrate its potential benefits in improving the safety of such systems towards practical deployment. The framework models *intrinsic uncertainty* through heteroscedastic noise model and *parameter uncertainty* through approximate Bayesian inference in the form of variational dropout, and finally integrates the two to quantify *predictive uncertainty* over the system output. Experiments focus on the super-resolution application of image quality transfer (IQT) (Alexander et al., 2017) and study several desirable properties of such framework, which lack in the existing body of data enhancement methods based on deterministic DL models.

Firstly, results on a range of applications and datasets show that modelling uncertainty improves overall prediction performance.

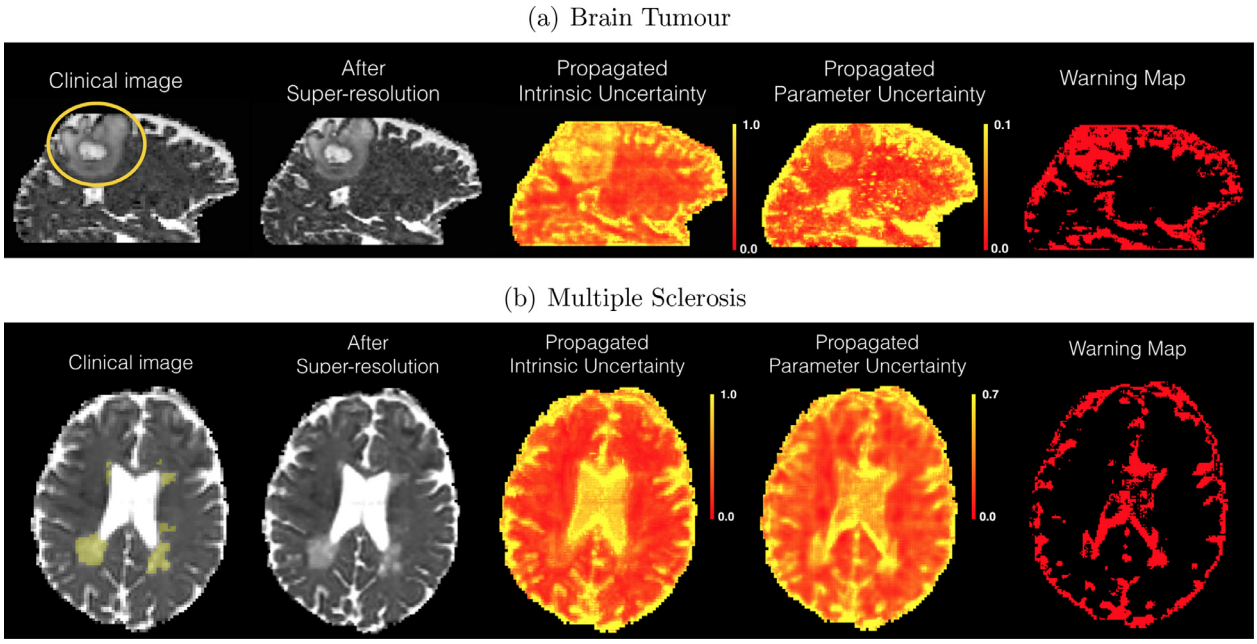

Fig. 11. Visualisation of propagated uncertainty components on clinical images with pathology that was not present in the training data. The super-resolution is performed on the clinical images due to low-resolution, and thus the ground truths are not available in both cases. (a) shows the results on the data of a Glioma patient, and the yellow circle indicates the region of tumour. (b) shows the same set of results on a MS patient with labels of focal lesions obtained from a neurologist indicated in yellow. Each row shows from left to right: (i) MD map computed from the original DTI; (ii) MD map computed from the output of super-resolution; (iii), (iv) maps of the estimated propagated intrinsic and parameter uncertainty; (v) "warning map" obtained from the same threshold value used in Section 4.4.1, which flag large parts of the pathological features in both cases.

Tables 2 and 3 show that modelling the combination of both *intrinsic* and *parameter* uncertainty achieves the state-of-the-art accuracy on super-resolution of DTIs and MAP-MRI coefficients in both of the HCP test dataset and the Lifespan dataset, improving on the present best methods based on random-forests (RF-IQT Alexander et al., 2017 and RF-BIQT Tanno et al., 2016) and interpolation—the standard method to estimate sub-voxel information used in clinical visualisation software. In particular, results on the Lifespan dataset, which differs from the training data in age range and acquisition protocol, indicates the better generalisability of our method. In addition, Fig. 13 shows that such combined model also benefits downstream tractography in comparison with the previous methods, illustrating the potential utility of the method for downstream connectivity analysis. Such improvement in the predictive performance arises from the regularisation effects imparted by the modelling of respective uncertainty components. Specifically, modelling intrinsic uncertainty through the heteroscedastic network improves robustness to outliers, while modelling parameter uncertainty via variational dropout defends against overfitting. For example, Table 3 shows that the predictive performance of the 3D-ESPCN + Hetero. model is only marginally compromised even when the outliers are not removed from training data, while the baseline 3D-ESPCN results in much poorer performance. This can be ascribed to the ability of the variance network $\Sigma_{\theta_2}(\cdot)$ in the 3D-ESPCN + Hetero. architecture to attenuate the effects of outliers by assigning small weights (i.e. high uncertainty) in the weighted MSE loss function as shown in Eq. (21). However, this loss attenuation mechanism can also encourage the network to overfit to low-uncertainty regions, potentially focusing less on ambiguous yet important parts of the data—we indeed observe in Table 3 that the heteroscedastic network performs considerably worse than the baseline 3D-ESPCN on the exterior regions while the reverse is observed on the interior part. Such overfitting to low-uncertainty interior regions is alleviated by modelling parameter uncertainty with variational dropout (Kingma et al., 2015), as evidenced by the dramatic error reduction in the exterior region on both HCP and Lifespan datasets.

Secondly, experiments on the images of healthy and pathological brains have demonstrated the utility of *predictive uncertainty* as a reliability metric of output images. Fig. 13 illustrates the strong correspondence between the maps of predictive uncertainty and the reconstruction quality (voxel-wise RMSE) in the downstream derived quantites such as FA and MD maps. In addition, Fig. 11 shows that such uncertainty measure also highlights pathological structures not observed in the training data. We have also tested the utility of predictive uncertainty in discriminating voxels with sufficiently low RMSEs in the predicted high-resolution MD maps. As shown in Fig. 8, the optimal threshold selected on the HCP training dataset is capable of detecting over 90% of non-reliable predictions—voxels with RMSE above a certain threshold—not only on the unseen subjects in the same HCP cohort but also on subjects from the out-of-sample Lifespan dataset, that are statistically disparate from the training distribution (e.g. different age range and acquisition protocol). These results combined demonstrate the utility of predictive uncertainty map as a means to quantify output safety, and provides a subject-specific alternative to standard population-group reliability metrics (e.g. mean reconstruction accuracy in a held-out cohort of subjects). Such conventional group statistics can be misleading in practice; for instance, the information that a super-resolution algorithm is reliable 99% of the time on a dataset of 1000 subjects may not accurately represent the performance on a new unseen individual if the person is not well-represented in the cohort (e.g. pathology, different scanners, etc). In contrast, predictive uncertainty provides a metric of reliability, tailored to each individual at hand.

Thirdly, our preliminary experiments show that decomposition of the effects of intrinsic and parameter uncertainty in the predictive uncertainty provides a layer of explanations into the performance of the considered deep learning methods. Fig. 9 shows that the low reconstruction error in the centre of the benign cyst can be explained by the dominant intrinsic uncertainty, which indicates the inherent simplicity of super-resolution task in such homogeneous region, whilst the unfamiliarity of such structure in the healthy training dataset is reflected in the high parameter uncertainty. Assuming that the estimates of decomposed uncertainty components are sufficiently accurate, we could act on them to further improve the overall safety of the system. Imagine a scenario where reconstruction error is consistently high on certain

image structures, if the parameter uncertainty is high but intrinsic uncertainty is low, this indicates that collecting more training data would be beneficial. On the other hand, if the parameter uncertainty is low and intrinsic uncertainty is high, this would mean that we need to regard such errors as inevitability, and abstain from predictions to ensure safety or account for them appropriately in subsequent analysis. We, however, note that in our experiments, the receptive field of the employed network is relatively small. In consequence, both the intrinsic and parameter uncertainty are estimated purely based on the statistics of local patches. Future work will study the effects of accounting for semantic information on the quality of estimated uncertainty components by comparing networks of varying receptive fields.

The proposed methods for estimating intrinsic and parameter uncertainty make several simplifying assumptions in the form of likelihood model $p(\mathbf{y}|\theta, \mathbf{x})$ and posterior distributions over network parameters $p(\theta|D)$. Firstly, the likelihood model takes the form of a Gaussian distribution with a diagonal covariance matrix. This means that the likelihood model is not able to capture multi-modality of the predictive distribution i.e. the presence of multiple different solutions. While the full predictive distribution (Eq. (9)) is not necessarily unimodal in theory due to the integration with the posterior distribution, we observe in practice that the drawn samples are not very diverse. Future work should explore the benefits of employing more complex forms of likelihood functions such as mixture models (Bishop, 1994; Kohl et al., 2018), diversity losses (Bouchacourt et al., 2016; Guzman-Rivera et al., 2012; Lee et al., 2018) and more powerful density estimators (Huang et al., 2018; Kohl et al., 2018; Odena et al., 2017; Papamakarios et al., 2017; Rezende and Mohamed, 2015). Also, the diagonality of covariance matrices means that the output pixels are assumed statistically independent given the input. Although the predicted images display high inter-pixel consistency, modelling the correlations between neighbouring pixels (Chandra and Kokkinos, 2016) may further improve the reconstruction quality. Analogous to the likelihood function, variational dropout (Kingma et al., 2015), which is used in this work, approximates the posteriors $p(\theta|D)$ by Gaussian distributions with diagonal covariance, imposing restrictive assumptions of unimodality and statistical independence between neural network weights. More recent advances in the Bayesian deep learning research (Louizos and Welling, 2016, 2017; Oh et al., 2020; Pawlowski et al., 2017; Zhang et al., 2019; Krueger et al., 2017) could be used to enhance the quality of parameter uncertainty estimation by allowing the model to capture multi-modality and statistical dependencies between parameters. We also refer the readers to a recent review paper by Zhang et al. (2018) on this topic for a balanced perspetive on possible approaches. We should note that both the mean and variance MC estimators of very high dimensional posterior distribution converge with only a few hundred samples in our case, because of this simplistic choice of the variational distributions. However, it is likely that, in order to approximate the posterior with a more complex family of distributions, a larger number of samples would be necessary.

The lack of "ground truths" renders the quantitative evaluation of the "accuracy" of the derived uncertainty estimates extremely challenging. Unfortunately, the distribution of interest $p(\mathbf{y}|\mathbf{x})$ is unknown in real-world medical imaging applications including the task of dMRI super-resolution. However, we envision the use of image simulation would provide new means to quantify the differences of various methods of modelling uncertainty. For the validation of intrinsic uncertainty estimate, we plan to create a synthetic image dataset with the known target distribution $p(\mathbf{y}|\mathbf{x})$. For example, one possibility is to pass a set of medical images through a known stochastic transformation to define the target output images[12]. This way, the "ground truth" intrinsic noise is known and the fidelity of the intrinsic uncertainty estimate can be quantified. It would also be interesting to study how the relative accuracy of intrinsic uncertainty estimates from different methods measured on a variety of such synthetic datasets translate to the measure of practical utility (e.g., detection rate of predictive failures). On the other hand, the validation of the parameter uncertainty is more challenging since the target distribution of interest (i.e., the posterior distribution over the parameters) is not available even if the underlying data distribution $p(\mathbf{y}|\mathbf{x})$ is known as is the case in synthetic datasets. However, controllable and realistic means to edit input images (Clatz et al., 2005; Park et al., 2019; Prastawa et al., 2009) (e.g., simulation of pathological structures of different controllable parameters such as size and shape) would allow ones to study in a systematic fashion what kinds of "out-of-distribution" structures can be detected through the estimate of parameter uncertainty for different Bayesian NN models.

Another important future challenge is the clinical validation of predictive uncertainty as a reliability metric of output images. To this end, we need to design a more clinically meaningful definition of success and failure of the data enhancement algorithm at hand. Despite the high accuracy in distinguishing between predictive failures and successes attained with our method (Fig. 8), our definition of reconstruction quality, namely voxel-wise RMSE, does not necessarily represent the real utility of the output image. One possible approach would be to have clinical experts to label the potential failures in the super-resolved images, be it for a targeted application (e.g. diagnosis of some neurological conditions) or for general usage in clinical practice. A more economical alternative, which does not require extra label acquisition, is to define the prediction success in downstream measurements of interest i.e. functions of the output images $g(\cdot)$, such as morphometric measurements of anatomical or pathological structures (e.g. volumes). The propagation method (Eq. (13)) introduced in Section 3.6 can be utilised to quantify uncertainty components in the space of target measurement $g(\cdot)$. Measuring the correlation between such propagated uncertainty estimates and the corresponding errors would be a useful indicator of how well the uncertainty measure reflects the accuracy of the chosen measurement $g(\cdot)$. Lastly, our initial results on the brain tumour dataset motivate a larger-scale quantitative validation of uncertainty estimates in the presence of pathology. Future work must examine the effect of including patients' dataset in the training data on the estimate of uncertainty components.

There are many ways in which uncertainty information could be utilised by radiologists or other users of data enhancement algorithms. First, predictive uncertainty can be used to decide when to abstain from predictions in high-risk regions of images (e.g. anomalies, out-of-distribution examples or inherently ambiguous features). For example, the original input low-resolution image can be augmented by overlaying the high-resolution prediction only in locations with sufficiently low uncertainty, before presenting to clinicians. As demonstrated by Fig. 8 in the context of super-resolution, such uncertainty-based quality control of predictions is potentially an effective means to maintain high accuracy of output images and also to safeguard against hallucination or removal of structures (Cohen et al., 2018a). Second, the uncertainty information could be used for active learning (Settles, 2009) to decide which images should be labelled and included in the training set to maximally improve the model performance. Prior work (Gal et al., 2017; Gorriz et al., 2017) define the acquisition function so as to select examples with high parameter uncertainty, and achieve promising results in classification and segmentation tasks. In particular, these methods are able to construct a compact and effective training dataset, and consequently improve the prediction accuracy while reducing the training time. The same idea could be naturally extended to data enhancement problems, that are typically formulated as multivariate regression tasks. For example, in the case of IQT, we could simulate a library of low-resolution and high-resolution image pairs from a large public dataset (e.g. HCP), and incrementally expand the training data by adding more examples from such a library. We should note, however, that in many data enhancement applications, obtaining a new "label" may require an extra

[12] For example, one could use a patch-wise cubic transform with diminishing noise $y = \bar{\mathbf{x}}^3 + \frac{1}{\bar{\mathbf{x}}^3}\epsilon \in \mathbb{R}$ where $\bar{\mathbf{x}}$ denotes the sum of all elements in $\mathbf{x}$ and $\epsilon \sim \mathcal{N}(0, 1)$. Repeated application of this function to neighbouring patches in the input image synthesises the target image.

acquisition possibly with a different scanner or modality, which may be logistically challenging. Third, another important application is transfer learning (Pan and Yang, 2010) where uncertainty information could be used to leverage knowledge from different but related domains or tasks. In many data enhancement applications, the test distribution can considerably deviate from the training distribution. For example, the algorithm might be trained on a synthetic dataset or images acquired from a scanner that is very different from the one used in the hospital where one plans to deploy the model. Therefore, a mechanism to adapt performance within a specific environment (e.g., based on the local patient population) (Kamnitsas et al., 2017a), possibly in an online fashion (Karani et al., 2018; Baweja et al., 2018), is in demand. Recent work have shown that the Bayesian formalism provides a natural framework to use uncertainty in order to account for the difference and commonality between distributions to guide information transfer in continual learning (Kirkpatrick et al., 2017; Nguyen et al., 2018) or few-shot learning (Finn et al., 2018; Yoon et al., 2018) settings. Exploring the benefits of these ideas in the context of medical image enhancement remains future work.

Another noteworthy limitation of the current super-resolution approach is the dependence on a fixed downsampling model, which may deviate from the test environment. To some degrees, our experiments already substantiate robustness of our approach; for example, our experiment in Fig. 13 shows that the super-resolution algorithm can improve the quality of tractography on a real low-quality image even in situations where the input resolution is different from that of the training data. Moreover, we also found that the preliminary investigations into alternative downsampling strategies (e.g., replacing the block averaging with bilinear interpolation) produce little variation in the results. However, a more thorough evaluation is needed for a clinical adoption to establish when our super-resolution method fails, and whether such safety boundary can be quantified through the estimated predictive uncertainty in a range of test environments. In addition, future work will also investigate whether a more realistic emulation of the image generation process in place of the simple downsampling (e.g., using a diffusion MRI simulator) enhances practical applications.

The proposed framework for uncertainty quantification is formulated for multivariate regression in the general form, and thus is naturally applicable to many other image enhancement challenges such as: rapid image acquisition techniques e.g., compressed sensing (Sun et al., 2016), MR fingerprinting (Cohen et al., 2018b; Ma et al., 2013) or sparse reconstruction (Hammernik et al., 2018; Schlemper et al., 2018a); denoising (Benou et al., 2017) and dealiasing (Han et al., 2018; Yang et al., 2018); image synthesis tasks e.g., estimating T2-weighted images from T1 (Jog et al., 2015; Rousseau, 2008; Ye et al., 2013), estimating CT images from MRI (Bragman et al., 2018; Burgos et al., 2015; Nie et al., 2018), and generating a high-field scan from a low-field scan (Bahrami et al., 2016); data harmonisation (Karayumak et al., 2018; Mirzaalian et al., 2016; Tax et al., 2019) which aims to learn mappings among imaging protocols to reduce confounds in multicentre studies. Our results on image quality transfer (Alexander et al., 2017) illustrate the potential of the uncertainty modelling techniques to improve the safety of these applications by not only improving the predictive accuracy, but also providing a mechanism to quantify risks and safeguard against potential malfunction.

Data and code availability statement

The Human Connectome Project dataset (release Q3) (Van Essen et al., 2013) and the Lifespan dataset (Harms et al., 2018) are publicly available. The Prisma data is available upon request. The glioma and multiple-sclerosis datasets are part of on-going studies at the Humanitas Research Hospital, Italy and Institute of Neurology at UCL, UK repectively, and we are bound by the policies of the data providers. The code will be released at https://github.com/rtanno21609/SaferNeuroimageEnhancement upon publication.

CRediT authorship contribution statement

Ryutaro Tanno: Writing - original draft, Conceptualization, Methodology, Software, Validation. **Daniel E. Worrall:** Methodology, Software, Writing - review & editing. **Enrico Kaden:** Conceptualization, Writing - review & editing, Visualization. **Aurobrata Ghosh:** Conceptualization, Writing - review & editing. **Francesco Grussu:** Resources, Writing - review & editing. **Alberto Bizzi:** Resources, Writing - review & editing. **Stamatios N. Sotiropoulos:** Resources, Writing - review & editing. **Antonio Criminisi:** Funding acquisition, Supervision, Writing - review & editing. **Daniel C. Alexander:** Writing - original draft, Funding acquisition, Conceptualization, Supervision.

Acknowledgement

We would like to thank Felix Bragmann at Babylon Health, Zach Eaton-Rosen at UCL/KCL and Stefano Blumberg at UCL for their valuable feedback. We would also like to thank Samuel Hurley whom helped with the Prisma acquisitions in FMRIB at University of Oxford. The tumour data were acquired as part of a clinical research project (Figini et al., 2018) lead by Alberto Bizzi, MD at the Humanitas Research Hospital in Milan, Italy. We are also grateful to Mark S. Graham at Visulytix and Gary Zhang at UCL for connecting us with Alberto Bizzi. The multiple sclerosis (MS) data were acquired as part of a study at UCL Institute of Neurology, funded by the MS Society UK and the UCL Hospitals Biomedical Research Centre (PIs: David Miller and Declan Chard). The HCP data were provided by the WU-Minn Consortium (PIs: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by NIH and Washington University.

EU Horizon 2020 grant CDS-QuaMRI 634541-2 and EPSRC grants R014019, R006032, N018702, and M020533 support DCA's work on this topic. FG has received funding under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 634541 and from the EPSRC (R006032/1 and M020533/1). RT was supported by Microsoft scholarship.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.neuroimage.2020.117366 .

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: a system for large-scale machine learning.. In: Proceedings of the OSDI, 16, pp. 265–283.

Alexander, D.C., et al., 2014. Image quality transfer via random forest regression: applications in diffusion MRI. In: Proceedings of the MICCAI. Springer, pp. 225–232.

Alexander, D.C., Zikic, D., Ghosh, A., Tanno, R., Wottschel, V., Zhang, J., Kaden, E., Dyrby, T.B., Sotiropoulos, S.N., Zhang, H., et al., 2017. Image quality transfer and applications in diffusion MRI. NeuroImage 152, 283–298.

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., Campilho, A., 2017. Classification of breast cancer histology images using convolutional neural networks. PloS One 12 (6), e0177544.

Ayhan, M. S., Berens, P., 2018. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. Proceedings of the Medical Imaging with Deep Learning (MIDL) Conference.

Bahrami, K., Shi, F., Rekik, I., Shen, D., 2016. Convolutional neural network for reconstruction of 7T-like images from 3T MRI using appearance and anatomical features. In: Proceedings of the MICCAI DLDLM Workshop. Springer, pp. 39–47.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2018. An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9252–9260.

Basser, P.J., Mattiello, J., LeBihan, D., 1994. Mr diffusion tensor spectroscopy and imaging. Biophys. J. 66 (1), 259–267.

Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E., 2019. PHiSeg: capturing uncertainty in medical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention.

Baweja, C., Glocker, B., Kamnitsas, K., 2018. Towards continual learning in medical imaging. Medical Imaging meets NIPS Workshop, 32nd Conference on Neural Information Processing Systems (NIPS). Montréal, Canada.

Begoli, E., Bhattacharya, T., Kusnezov, D., 2019. The need for uncertainty quantification in machine-assisted medical decision making. Nat. Mach. Intell. 1 (1), 20.

Benou, A., Veksler, R., Friedman, A., Raviv, T.R., 2017. Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced MRI sequences. Med. Image Anal. 42, 145–159.

Bishop, C.M., 1994. Mixture Density Networks. Technical Report. Citeseer.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: a review for statisticians. J. Am. Stat. Assoc. 112 (518), 859–877.

Blumberg, S.B., Palombo, M., Khoo, C.S., Tax, C., Tanno, R., Alexander, D.C., 2019. Multi-stage prediction networks for data harmonization. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer.

Blumberg, S.B., Tanno, R., Kokkinos, I., Alexander, D.C., 2018. Deeper image quality transfer: training low-memory neural networks for 3D images. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 118–125.

Bouchacourt, D., Mudigonda, P.K., Nowozin, S., 2016. Disco nets: dissimilarity coefficients networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 352–360.

Bowsher, C.G., Swain, P.S., 2012. Identifying sources of variation and the flow of information in biochemical networks. In: Proceedings of the National Academy of Sciences.

Bragman, F.J., Tanno, R., Eaton-Rosen, Z., Li, W., Hawkes, D.J., Ourselin, S., Alexander, D.C., McClelland, J.R., Cardoso, M.J., 2018. Uncertainty in multitask learning: joint representations for probabilistic mr-only radiotherapy planning. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention.

Burgos, N., Cardoso, M.J., Guerreiro, F., Veiga, C., Modat, M., McClelland, J., Knopf, A.-C., Punwani, S., Atkinson, D., Arridge, S.R., et al., 2015. Robust CT synthesis for radiotherapy planning: Application to the head and neck region. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 476–484.

Chandra, S., Kokkinos, I., 2016. Fast, exact and multi-scale inference for semantic image segmentation with deep Gaussian CRFs. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 402–418.

Chen, H., Zhang, Y., Zhang, W., Liao, P., Li, K., Zhou, J., Wang, G., 2017. Low-dose CT via convolutional neural network. Biomed. Opt. Exp. 8 (2), 679–694.

Chen, T., Fox, E., Guestrin, C., 2014. Stochastic gradient hamiltonian monte carlo. In: Proceedings of the International Conference on Machine Learning, pp. 1683–1691.

Chen, Y., Shi, F., Christodoulou, A.G., Xie, Y., Zhou, Z., Li, D., 2018. Efficient and accurate MRI super-resolution using a generative adversarial network and 3d multi-level densely connected network. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 91–99.

Clatz, O., Sermesant, M., Bondiau, P.-Y., Delingette, H., Warfield, S.K., Malandain, G., Ayache, N., 2005. Realistic simulation of the 3-D growth of brain tumors in MR images coupling diffusion with biomechanical deformation. IEEE Trans. Med. Imaging 24 (10), 1334–1346.

Cohen, J. P., Luck, M., Honari, S., 2018a. Distribution Matching Losses can Hallucinate Features in Medical Image translation. Proceedings pf the. International Conference on Medical Image Computing and Computer-Assisted Intervention 529–536.

Cohen, O., Zhu, B., Rosen, M.S., 2018b. Mr fingerprinting deep reconstruction network (drone). Magn. Reson. Med. 80 (3), 885–894.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. 2 (4), 303–314.

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2018. Unsupervised learning for fast probabilistic diffeomorphic registration. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention.

Der Kiureghian, A., Ditlevsen, O., 2009. Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2), 105–112.

Dong, C., Loy, C.C., He, K., Tang, X., 2016. Image super-resolution using deep convolutional networks. IEEE PAMI 38 (2), 295–307.

Draper, D., 1995. Assessment and propagation of model uncertainty. J. R. Stat. Soc.: Ser. B (Methodol.) 57 (1), 45–70.

Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., Cardoso, M.J., 2018. Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 691–699.

Eaton-Rosen, Z., Varsavsky, T., Ourselin, S., Cardoso, M.J., 2019. As easy as 1, 2... 4? Uncertainty In counting tasks for medical imaging. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 356–364.

Esses, S.J., Lu, X., Zhao, T., Shanbhogue, K., Dane, B., Bruno, M., Chandarana, H., 2018. Automated image quality evaluation of T2-weighted liver MRI utilizing deep learning architecture. J. Magn. Reson. Imaging 47 (3), 723–728.

Figini, M., Riva, M., Graham, M., Castelli, G.M., Fernandes, B., Grimaldi, M., Baselli, G., Pessina, F., Bello, L., Zhang, H., et al., 2018. Prediction of isocitrate dehydrogenase genotype in brain gliomas with MRI: single-shell versus multishell diffusion models. Radiology 289 (3), 788–796.

Finn, C., Xu, K., Levine, S., 2018. Probabilistic model-agnostic meta-learning. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 9516–9527.

Gal, Y., Ghahramani, Z., 2015. Dropout as a Bayesian approximation: Insights and applications. In: Proceedings of the Deep Learning Workshop, ICML.

Gal, Y., Hron, J., Kendall, A., 2017. Concrete dropout. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 3581–3590.

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep Bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 1183–1192.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., et al., 2013. The minimal preprocessing pipelines for the human connectome project. NeuroImage 80, 105–124.

Gorriz, M., Carlier, A., Faure, E., Giro-i Nieto, X., 2017. Cost-Effective Active Learning for Melanoma Segmentation. NeurIPS Machine Learning for Healthcare (ML4H) Workshop.

Guzman-Rivera, A., Batra, D., Kohli, P., 2012. Multiple choice learning: learning to produce multiple structured outputs. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 1799–1807.

Hammernik, K., Klatzer, T., Kobler, E., Recht, M.P., Sodickson, D.K., Pock, T., Knoll, F., 2018. Learning a variational network for reconstruction of accelerated MRI data. Magn. Reson. Med. 79 (6), 3055–3071.

Han, Y., Yoo, J., Kim, H.H., Shin, H.J., Sung, K., Ye, J.C., 2018. Deep learning with domain adaptation for accelerated projection-reconstruction mr. Magn. Reson. Med. 80 (3), 1189–1205.

Harms, M.P., Somerville, L.H., Ances, B.M., Andersson, J., Barch, D.M., Bastiani, M., Bookheimer, S.Y., Brown, T.B., Buckner, R.L., Burgess, G.C., et al., 2018. Extending the human connectome project across ages: imaging protocols for the lifespan development and aging projects. NeuroImage 183, 972–984.

Hora, S.C., 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. Reliab. Eng. Syst. Saf. 54 (2-3), 217–223.

Huang, X., Liu, M.-Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image–to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172–189.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning, pp. 448–456.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134.

Jin, K.H., McCann, M.T., Froustey, E., Unser, M., 2017. Deep convolutional neural network for inverse problems in imaging. IEEE Trans. Image Process. 26 (9), 4509–4522.

Jog, A., Carass, A., Roy, S., Pham, D.L., Prince, J.L., 2015. MR image synthesis by contrast learning on neighborhood ensembles. Med. Image Anal. 24 (1), 63–76.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the ECCV. Springer, pp. 694–711.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017a. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Proceedings of the International Conference on Information Processing in Medical Imaging. Springer, pp. 597–609.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017b. Efficient multi-scale 3D CNN with fully connected crf for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.

Kang, E., Min, J., Ye, J.C., 2017. A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. Med. Phys. 44 (10).

Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E., 2018. A lifelong learning approach to brain mr segmentation across scanners and protocols. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 476–484.

Karayumak, S.C., Kubicki, M., Rathi, Y., 2018. Harmonizing diffusion MRI data across magnetic field strengths. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 116–124.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of the Advances in Neural Information Processing Systems, pp. 5580–5590.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. CoRR. abs/1412.6980

Kingma, D.P., Salimans, T., Welling, M., 2015. Variational dropout and the local reparameterization trick. In: Proceedings of the NIPS, pp. 2575–2583.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes. In: Proceedings of the International Conference on Learning Representations.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al., 2017. Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. 114 (13), 3521–3526.

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S.A., Rezende, D.J., Ronneberger, O., 2018. A probabilistic u-net for segmentation of ambiguous images. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 6965–6975.

Krueger, D., Huang, C.-W., Islam, R., Turner, R., Lacoste, A., Courville, A., 2017. Bayesian Hypernetworks. NeurIPS Workshop on Bayesian Deep Learning.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 6402–6413.

Lebel, C., Gee, M., Camicioli, R., Wieler, M., Martin, W., Beaulieu, C., 2012. Diffusion tensor imaging of white matter tract evolution over the lifespan. NeuroImage 60 (1), 340–352.

Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H., 2018. Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 35–51.

Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S., 2017. Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. 7 (1), 17816.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Louizos, C., Welling, M., 2016. Structured and efficient variational deep learning with matrix gaussian posteriors. In: Proceedings of the International Conference on Machine Learning, pp. 1708–1716.
Louizos, C., Welling, M., 2017. Multiplicative normalizing flows for variational bayesian neural networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 2218–2227.
Ma, D., Gulani, V., Seiberlich, N., Liu, K., Sunshine, J.L., Duerk, J.L., Griswold, M.A., 2013. Magnetic resonance fingerprinting. Nature 495 (7440), 187.
Ma, Y.-A., Chen, T., Fox, E., 2015. A complete recipe for stochastic gradient MCMC. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 2917–2925.
Mahapatra, D., Bozorgtabar, B., Hewavitharanage, S., Garnavi, R., 2017. Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 382–390.
McDonagh, S., Hou, B., Kamnitsas, K., Oktay, O., Alansary, A., Kainz, B., 2017. Context-Sensitive Super-Resolution for Fast Fetal Magnetic Resonance Imaging. Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment. Springer, Cham, pp. 116–126.
Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C., Morey, R.A., Flashman, L., et al., 2016. Inter-site and inter-scanner diffusion MRI data harmonization. NeuroImage 135, 311–323.
Molchanov, D., Ashukha, A., Vetrov, D., 2017. Variational dropout sparsifies deep neural networks. Proceedings of the 34th International Conference on Machine Learning, PMLR, pp. 2498–2507.
Nair, T., Precup, D., Arnold, D.L., Arbel, T., 2018. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 655–663.
Nair, T., Precup, D., Arnold, D.L., Arbel, T., 2020. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med. Image Anal. 59, 101557.
Neal, R.M., 1993. Bayesian learning via stochastic dynamics. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 475–482.
Nguyen, C.V., Li, Y., Bui, T.D., Turner, R.E., 2018. Variational continual learning. In: Proceedings of the International Conference on Learning Representations.
Nie, D., Cao, X., Gao, Y., Wang, L., Shen, D., 2016. Estimating ct image from MRI data using 3D fully convolutional networks. In: Proceedings of the Deep Learning and Data Labeling for Medical Applications. Springer, pp. 170–178.
Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., Wang, Q., Shen, D., 2018. Medical image synthesis with deep convolutional adversarial networks. IEEE Transactions on Biomedical Engineering 65 (12), 2720–2730.
Nix, D.A., Weigend, A.S., 1994. Estimating the mean and variance of the target probability distribution. In: Proceedings of the IEEE WCCI, 1. IEEE, pp. 55–60.
Odena, A., Dumoulin, V., Olah, C., 2016. Deconvolution and checkerboard artifacts. Distill doi:10.23915/distill.00003.
Odena, A., Olah, C., Shlens, J., 2017. Conditional image synthesis with auxiliary classifier gans. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 2642–2651.
Oh, C., Adamczewski, K., Park, M., 2020. Radial and directional posteriors for bayesian neural networks. Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence.
Oktay, O., Bai, W., Lee, M., Guerrero, R., Kamnitsas, K., Caballero, J., de Marvao, A., Cook, S., O'Regan, D., Rueckert, D., 2016. Multi-input cardiac image super-resolution using convolutional neural networks. In: Proceedings of the MICCAI. Springer.
Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T., O?Regan, D.P., et al., 2018. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE Trans. Med. Imaging 37 (2), 384–395.
Özarslan, E., Koay, C.G., Shepherd, T.M., Komlosh, M.E., İrfanoğlu, M.O., Pierpaoli, C., Basser, P.J., 2013. Mean apparent propagator (MAP) MRI: a novel diffusion imaging method for mapping tissue microstructure. NeuroImage 78, 16–32.
Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.
Papamakarios, G., Pavlakou, T., Murray, I., 2017. Masked autoregressive flow for density estimation. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 2338–2347.
Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2337–2346.
Pawlowski, N., Brock, A., Lee, M.C., Rajchl, M., Glocker, B., 2017. Implicit weight uncertainty in neural networks. Bayesian Deep Learning Workshop at NeurIPS.
Prastawa, M., Bullitt, E., Gerig, G., 2009. Simulation of brain tumors in mr images for evaluation of segmentation efficacy. Med. Image Anal. 13 (2), 297–311.
Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Mullainathan, S., Kleinberg, J.M., 2019. Direct Uncertainty Prediction for Medical Second Opinions. In: International Conference on Machine Learning. Proceedings of Machine Learning Research, pp. 5281–5290.
Rao, C.R., 1970. Estimation of heteroscedastic variances in linear models. J. Am. Stat. Assoc. 65 (329), 161–172.
Ravì, D., Szczotka, A.B., Pereira, S.P., Vercauteren, T., 2019. Adversarial training with cycle consistency for unsupervised super-resolution in endomicroscopy. Med. Image Anal. 53, 123–131.
Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L., 2010. Learning from crowds. J. Mach. Learn. Res. 11 (Apr), 1297–1322.
Rezende, D.J., Mohamed, S., 2015. Variational inference with normalizing flows. Proceedings of the 32nd International Conference on Machine Learning, PMLR, pp. 1530–1538.
Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M., 2014. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 520–527.
Rousseau, F., 2008. Brain hallucination. In: Proceedings of the ECCV. Springer, pp. 497–508.
Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N., et al., 2019. Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage 195, 11–22.
Salat, D.H., 2014. Diffusion tensor imaging in the study of aging and age-associated neural disease. In: Diffusion MRI. Elsevier, pp. 257–281.
Schlemper, J., Caballero, J., Hajnal, J.V., Price, A.N., Rueckert, D., 2018a. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. IEEE Trans. Med. Imaging 37 (2), 491–503.
Schlemper, J., Yang, G., Ferreira, P., Scott, A., McGill, L.-A., Khalique, Z., Gorodezky, M., Roehl, M., Keegan, J., Pennell, D., et al., 2018b. Stochastic deep compressive sensing for the reconstruction of diffusion tensor cardiac MRI. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention.
Settles, B., 2009. Active Learning Literature Survey. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.
Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248.
Shi, H., Worrall, D., Veeling, B., Huisman, H., Welling, M., 2019. Supervised uncertainty quantification for segmentation with multiple annotations. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention.
Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the CVPR, pp. 1874–1883.
Sokooti, H., de Vos, B., Berendsen, F., Lelieveldt, B.P., Išgum, I., Staring, M., 2017. Nonrigid image registration using multi-scale 3D convolutional neural networks. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 232–239.
Sotiropoulos, S.N., et al., 2013. Advances in diffusion MRI acquisition and processing in the human connectome project. NeuroImage 80, 125–143.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Sun, J., Li, H., Xu, Z., et al., 2016. Deep ADMM-net for compressive sensing MRI. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 10–18.
Tanno, R., Ghosh, A., Grussu, F., Kaden, E., Criminisi, A., Alexander, D.C., 2016. Bayesian image quality transfer. In: Proceedings of the MICCAI. Springer, pp. 265–273.
Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., Silberman, N., 2019. Learning from noisy labels by regularized estimation of annotator confusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Tanno, R., Worrall, D.E., Ghosh, A., Kaden, E., Sotiropoulos, S.N., Criminisi, A., Alexander, D.C., 2017. Bayesian image quality transfer with CNNs: exploring uncertainty in dMRI super-resolution. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 611–619.
Tax, C.M., Grussu, F., Kaden, E., Ning, L., Rudrapatna, U., Evans, J., St-Jean, S., Leemans, A., Koppers, S., Merhof, D., et al., 2019. Cross-scanner and cross-protocol diffusion MRI data harmonisation: a benchmark database and evaluation of algorithms. NeuroImage 195, 285–299.
Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.-M.H., et al., 2013. The WU-Minn human connectome project: an overview. NeuroImage 80, 62–79.
Wang, H., Levi, D.M., Klein, S.A., 1996. Intrinsic uncertainty and integration efficiency in bisection acuity. Vis. Res. 36 (5), 717–739.
Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Proc. 13 (4).
Weiss, N.A., 2006. A Course in Probability. Addison-Wesley.
Welling, M., Teh, Y.W., 2011. Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 681–688.
Westlye, L.T., Walhovd, K.B., Dale, A.M., Bjørnerud, A., Due-Tønnessen, P., Engvig, A., Grydeland, H., Tamnes, C.K., Østby, Y., Fjell, A.M., 2009. Life-span changes of the human brain white matter: diffusion tensor imaging (DTI) and volumetry. Cereb. Cortex 20 (9), 2055–2068.
Wolterink, J.M., Dinkla, A.M., Savenije, M.H., Seevinck, P.R., van den Berg, C.A., Išgum, I., 2017. Deep MR to CT synthesis using unpaired data. In: Proceedings of the International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 14–23.
Worrall, D.E., Wilson, C.M., Brostow, G.J., 2016. Automated retinopathy of prematurity case detection with convolutional neural networks. In: Proceedings of the MICCAI DLDLM Workshop. Springer, pp. 68–76.
Wu, L., Cheng, J.-Z., Li, S., Lei, B., Wang, T., Ni, D., 2017. Fuiqa: Fetal ultrasound image quality assessment with deep convolutional networks. IEEE Trans. Cybern. 47 (5), 1336–1349.
Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P.L., Ye, X., Liu, F., Arridge, S., Keegan, J., Guo, Y., et al., 2018. Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. IEEE Trans. Med. Imaging 37 (6), 1310–1321.

Yang, X., Kwitt, R., Niethammer, M., 2016. Fast predictive image registration. In: Proceedings of the MICCAI DLDLM Workshop. Springer, pp. 48–57.
Ye, D.H., Zikic, D., Glocker, B., Criminisi, A., Konukoglu, E., 2013. Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 606–613.
Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., Ahn, S., 2018. Bayesian model-agnostic meta-learning. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 7332–7342.
Yoon, Y.H., Khan, S., Huh, J., Ye, J.C., 2019. Efficient b-mode ultrasound image reconstruction from sub-sampled RF data using deep learning. IEEE Trans. Med. Imaging 38 (2), 325–336.
Yu, H., Liu, D., Shi, H., Yu, H., Wang, Z., Wang, X., Cross, B., Bramler, M., Huang, T.S., 2017. Computed tomography super-resolution using convolutional neural networks. In: Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 3944–3948.
Zeiler, M.D., Taylor, G.W., Fergus, R., 2011. Adaptive deconvolutional networks for mid and high level feature learning. In: Proceedings of the ICCV. IEEE, pp. 2018–2025.
Zhang, C., Bütepage, J., Kjellström, H., Mandt, S., 2018. Advances in variational inference. IEEE transactions on pattern analysis and machine intelligence 41 (8), 2008–2026.
Zhang, R., Li, C., Zhang, J., Chen, C., Wilson, A.G., 2019. Cyclical stochastic gradient MCMC for bayesian deep learning. Proceedings of the International Conference on Learning Representations.
Zhao, C., Carass, A., Dewey, B.E., Woo, J., Oh, J., Calabresi, P.A., Reich, D.S., Sati, P., Pham, D.L., Prince, J.L., 2018. A deep learning based anti-aliasing self super-resolution algorithm for MRI. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 100–108.
Zhu, B., Liu, J.Z., Cauley, S.F., Rosen, B.R., Rosen, M.S., 2018. Image reconstruction by domain-transform manifold learning. Nature 555 (7697), 487.
Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2380–7504.