A Field of Experts Prior for Adapting Neural Networks at Test Time

Neerav Karani, Georg Brunner, Ertunc Erdil, Simin Fei,
Kerem Tezcan, Krishna Chaitanya, Ender Konukoglu

Biomedical Image Computing Group, Computer Vision Laboratory, ETH Zürich *

February 14, 2022

Abstract

Supervised learning methods based on convolutional neural networks (CNNs) show promising performance in several medical image analysis tasks. Such performance, however, is marred in the presence of acquisition-related distribution shifts between training and test images. Recently, it has been proposed to tackle this problem by fine-tuning trained CNNs for each test image. Such *test-time-adaptation* (TTA) is a promising and practical strategy for improving robustness to distribution shifts as it requires neither data sharing between institutions nor annotating additional data. Previous TTA methods use a *helper* model to increase similarity between outputs and/or features extracted from a test image with those of the training images. Such helpers, which are typically modeled using CNNs and trained in a self-supervised manner, can be task-specific and themselves vulnerable to distribution shifts in their inputs. To overcome these problems, we propose to carry out TTA by matching the feature distributions of test and training images, as modelled by a field-of-experts (FoE) prior. FoEs model complicated probability distributions as products of several simpler *expert* distributions. We use the 1D marginal distributions of a trained task CNN's features as the experts in the FoE model. Further, we carry out principal component analysis (PCA) of patches of the task CNN's features, and consider the distributions of the PCA loadings as additional experts. We extensively validate the method's efficacy on 5 MRI segmentation tasks (healthy tissues in 4 anatomical regions and lesion segmentation in 1 one anatomy), using data from 17 institutions, and on a MRI registration task, using data from 3 institutions. We find that the proposed FoE-based TTA is generically applicable in multiple tasks, and outperforms all previous TTA methods for lesion segmentation. For healthy tissue segmentation, the proposed method outperforms other task-agnostic TTA methods, but a previous TTA method which is specifically designed for segmentation performs the best for most of the tested datasets. Our implementation is publicly available here.

1 Introduction

1.1 The distribution shift problem

Performance of convolutional neural networks (CNNs) trained using supervised learning degrades when the distributions of training and test samples differ. This is known as the distribution shift (DS) problem[1]. Several types of DS are pertinent in medical imaging [1]. We consider acquisition-related DS - that is, DS caused by variations in scanners and acquisition protocol parameters. Such shifts are pervasive in clinical practice; thus, tackling them suitably is crucial for large-scale adoption of deep learning methods.

*Manuscript under review.

All authors are with the Biomedical Image Computing Group at ETH Zurich, Switzerland (https://bmic.ee.ethz.ch/). Corresponding author: Neerav Karani (nkarani@vision.ee.ethz.ch).

[1]We use the acronym DS to refer to the singular 'distribution shift' as well as the plural 'distribution shifts', and call on the reader to infer the form based on the context.

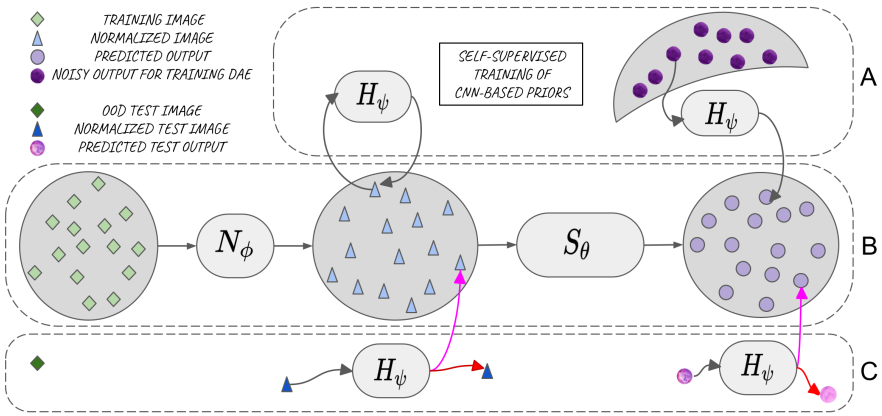

Figure 1: An illustrative schematic of the DS problem in CNN-based helper models. The figure is divided into 3 horizontal slabs. Slab B shows the mapping of the inputs (green) to the outputs (purple), via the normalized features (blue). Slab A shows the training of prior models (autoencoder (center) [12], denoising autoencoder (right) [11]) to be used for TTA: the AE is trained to auto-encode features of training images and the DAE is trained to denoise corrupted outputs (from a specific corruption distribution indicated by the crescent). Finally, slab C shows the desirable behaviour (pink arrows) and potential failure cases (red arrows) when the trained prior models are used to guide TTA.

1.2 Categories of methods to tackle the DS problem

Due to its high practical relevance, the DS problem has attracted substantial attention in the research community. In decreasing order of dependence on data from the test distribution, methods in the DS literature can be broadly categorized into the following groups: transfer learning (TL) [2], [3], [4], unsupervised domain adaptation (UDA) [5], [6], and domain generalization (DG) [7], [8], [9]. Among these three settings, DG is the most appealing - contrary to other settings, it does not require sharing data between institutions or annotating additional images. Recently, two new settings have been proposed - source-free domain adaptation (SFDA) [10] and test-time adaptation (TTA) [11], [12]. Here, the performance of CNNs trained with DG techniques is further improved by adapting them using unlabelled image(s) from the test distribution. Importantly, the adaptation in TTA or SFDA is done without access to data from the training distribution. Due to these advantages, we pose our work in the TTA setting.

1.3 Test-time adaptation

In TTA, the parameters of a previously trained CNN are adapted for each test image. The subset of the parameters that get adapted per test image is a design choice. Noting that acquisition-related DS manifest as contrast variations, one approach is to design the CNN as a concatenation of a *shallow, image-specific contrast normalization CNN*, $z = N_\phi(x)$, followed by a deep task CNN that is shared by all training and test images, $y = S_\theta(z)$. Here, x is the input image, z is the *normalized* image, and y is the output (e.g. segmentation, deformation field, enhanced image). The image-specific parameters, ϕ, are adapted by requiring adherence to a prior model, H_ψ, either in the output space [11] or in the feature space [12]. H_ψ encourages similarity between outputs or features of the test image with those of the training images. It is itself modelled using a CNN and trained in a self-supervised manner - as a denoising autoencoder (DAE) in [11] and as an autoencoder (AE) in [12].

1.4 The DS problem in H_ψ

In this work, we scrutinize the prior model, H_ψ, which is a key component in tackling the DS problem via TTA. Consider what happens when TTA is used to improve a CNN's prediction accuracy for an out-of-distribution (OOD) test image. (In general, OOD images can differ from training images in terms of acquisition settings, imaging modality (e.g. CT v/s MRI),

anatomy, etc. Here, we consider OOD test images pertaining to acquisition-related DS.) At the beginning of TTA iterations, the test features (outputs) are likely to be dissimilar to the features (outputs) corresponding to the training images. Indeed, this is symptomatic of the CNN's poor performance on OOD images. The main assumption of TTA methods like [11], [12] is that H_ψ is capable of mapping such features (outputs) to ones that are similar to features (outputs) observed during training. However, if H_ψ is modelled with a CNN, it is likely to be vulnerable to a DS problem of its own - that is, the outputs of H_ψ may be unreliable when its test inputs are from a different distribution as compared to its training inputs. An illustrative schematic of this problem is shown in Fig. 1. AEs in [12], which are trained to auto-encode features of training images, are not guaranteed to transform the features of test images to be like the features of training images. Similarly, DAEs in [11], trained to denoise corrupted outputs corresponding to a particular corruption distribution, may be unable to denoise outputs with different corruption patterns.

Although DAEs (for arbitrary corruption distributions) and AEs lack a strict probabilistic underpinning, the aforementioned TTA approaches can be roughly thought of as learning a probabilistic model of the training features (outputs), and then increasing the likelihood of the test features (outputs) under the trained model. We argue that even if CNN-based unsupervised density estimation models are used as the prior, they too are likely to suffer from the DS problem [13], [14]. For instance, one approach for TTA might be to train variational autoencoders (VAEs) to model the distribution of features of the training images, and to modify the test image's features such that their likelihood under the trained VAE increases. VAEs may even assign higher likelihood values to OOD samples than samples from their training distribution [13]. Such behaviour may render them unsuitable for TTA.

1.5 Overview of the proposed method

In this work, we propose two main changes as compared to recent TTA works. First, instead of driving TTA by minimizing the reconstruction loss of a prior model, H_ψ, we propose to match the distribution of 2D slices of a volumetric test image with the distribution of slices of training images. The distribution matching is done in the space of normalized images, z.

Second, noting the lack of DS robustness in CNN-based prior models for driving TTA, we posit that *simpler* prior models may (a) suffice to improve task performance under the considered acquisition-related DS, while (b) themselves being more robust to DS as compared to CNN-based priors. With this motivation, we model the distribution of the normalized training images, z, using a Field of Experts (FoEs) [15] formulation. FoEs (described in more detail in Sec. 3.1) combine ideas of Markov random fields (MRFs) [16] and Product of Experts (PoEs) [17]. FoEs enable modeling of complex distributions as a product of several simpler distributions. The simple distributions are those of the outputs of so-called *expert* functions, which are typically formulated as scalar functions of image patches. We propose to use the task-specific filters learned in S_θ as the FoE experts (Sec. 3.2.1). Further, we augment the FoE model with additional experts - projections onto principal components of patches in the last layer of S_θ (Sec. 3.2.3).

For TTA, we adapt the normalization module N_ϕ, so as to match the individual expert distributions of the test and training images, for all experts in the FoE model.

1.6 Summary of contributions

To summarize, we consider the acquisition-related DS problem in CNN-based medical image analyses and make the following contributions in this work: (1) we propose distribution matching for TTA, (2) we model the distribution of normalized images, z, using a FoE model, with the task-specific CNN filters acting as the expert functions, and (3) we augment the FoE model with PCA-based expert functions.

We support these technical contributions with an extensive validation on 5 image segmentation tasks, using data from 17 centers, and an image registration task, using data from 3 centers. To the best of our knowledge, this is the first work in the literature that evaluates the TTA

setting on such a large variety of anatomies and tasks for medical image analysis. The results of these experiments help us organize the current TTA literature, including the proposed method, along three axes. (1) Applicability to multiple tasks: some of the existing TTA methods are task-dependent. The proposed method relieves this constraint, and provides a general approach that can used in multiple tasks. As compared to existing task-agnostic methods, the proposed method provides similar performance for image registration and superior performance for image segmentation. (2) Performance in segmentation of anomalies: we find that DS robustness issue is particularly difficult for lesion datasets. Here, all of the existing TTA methods either fail to improve performance, and several methods even lead to performance degradation as compared to the baseline. The proposed method provides substantial performance gains in this challenging scenario. (3) Performance in segmentation of healthy tissues: in this scenario, our experiments indicate that methods specifically designed for handling distribution shifts in image segmentation outperform more general TTA methods, including the proposed method.

2 Related Work

2.1 Domain Generalization (DG)

From a practical point-of-view, DG is arguably most attractive among all strategies for tackling the DS problem; after training, it allows a CNN to be used directly (without any adaptation) for analyzing images from unseen test distributions. Several strategies have been proposed for DG - (i) meta learning [7], (ii) domain invariant representation learning [18], (iii) shape-appearance disentanglement [19], (iv) regularization using task-specific priors [20], (v) data augmentation [8], (vi) training with a fully-synthetic dataset of images representing a large degree of morphological, resolution and acquisition parameter variation [9], among others. These DG methods substantially improve CNN robustness with respect to DS; however, there still remains a gap to the performance that can be achieved if supervised learning were to be done using labelled images from the test distribution. Further, some of these methods rely on design choices that may be applicable only for certain anatomical regions (for instance, the procedure to generate synthetic images in [9] requires dense segmentation labels as inputs). Overall, we argue that the settings of DG and TTA are complementary in nature - the former can provide a fairly robust trained model, and the latter can further improve performance by fine-tuning the model to specifically suit the test image at hand.

2.2 Test-Time Adaptation

A relatively new approach for tackling DS is to adapt a trained model using unlabelled test image(s), but without access to the training dataset. At a broad level, works in this category vary along two axes - (a) which parameters are adapted at test time and (b) the loss function that is used to drive the adaptation. Common choices along axis (a) are (i) a normalization module in the task CNN's initial layers [11], [21], (ii) batch normalization parameters throughout the task CNN [22] and (iii) a combination of shallow adaptable modules at different layers in the task CNN [12]. Along axis (b), proposed works either minimize (i) the loss of a pre-trained self-supervised network [11], [12], [21], (ii) the entropy of predictions for the test image(s) [22], or (iii) task-specific self-supervised losses such as (1) k-space data consistency in MRI reconstruction CNNs [23], (2) cycle-consistency-based estimation of a *correction filter* to transform low-resolution (LR) test images to resemble LR images seen during training of super-resolution CNNs [24] or (3) an estimator (Stein's unbiased risk estimator) of the true loss for known noise distributions in denoising CNNs [25].

Test-image-specific adaptation has also been considered in the context of generative models. For instance, [26] proposed to fine-tune density estimation models (e.g. generative adversarial networks) for each test image, when used in the Bayesian image enhancement framework. As well, [27] observed that CNNs trained from scratch to generate a given corrupted test image from a random vector have a tendency to first generate the corresponding clean image. This has been recently leveraged for dynamic cardiac MRI reconstruction in [28].

A closely related setting to TTA is that of source-free domain adaptation (SFDA), where multiple images from the test distribution are used simultaneously for model adaptation [29], [30], [31], [10]. While SFDA has the advantage that multiple images from the test distribution may provide a regularization effect on one another during adaptation, TTA may benefit from adapting parameters to get the best performance for each test image. It may be interesting to empirically compare the performance of SFDA with the proposed TTA approach; we defer this analysis to future work.

2.3 Matching Marginal Feature Distributions for Tackling DS

Another TTA strategy is to use the statistics of the test image(s) in the batch normalization [32] layers of the task CNN. Here, no learnable parameters of the task CNN are adapted; rather the mean and variance stored in each batch normalization layer are replaced with those of the given test image(s). Effectively, at each layer, this amounts to matching the 1D Gaussian approximation of the marginal feature distribution of the test image(s) with that of the entire training dataset. Indeed, with this motivation, [33] explicitly minimize the KL-divergence between Gaussian approximations of the marginal distribution of features at a particular layer in the task CNN. These strategies has been shown to improve DS robustness in natural imaging datasets [34], [35], [33]. On the other hand, [36] recently point out that this method matches only the first two moments of the 1D distributions, and is thus prone to inaccuracies when the distributions are substantially non-Gaussian. To match higher-order moments, concurrent work [37] matches non-parametric approximations of marginal feature distributions between test and training images. The formulation presented in this work can be used with both parametric or non-parametric approximations.

We believe that an important contribution of the work in this manuscript is the interpretation of marginal feature distribution matching idea in the field-of-experts (FoE) formulation. This formulation allows us to view individual features as experts of a FoE model for the *full* probability distribution of upstream features (specifically, in our case, of the normalized images, $N_\phi(X)$). Thus, the proposed work generalizes the marginal distribution matching framework, and several previous works [34], [35], [33], [37] can be seen as instances of the proposed general framework. Furthermore, the proposed framework naturally extends to include 1D distribution matching in the space of PCA loadings of patches of CNN features.

2.4 Frequency of summary statistics for out-of-distribution (OOD) detection

Noting that density estimation models may assign higher likelihood values to OOD samples than samples of the training distribution [13], [38] instead constructed 1D PDFs of several summary statistics for the training data and evaluated the likelihood of the same statistics of test data under the constructed PDFs. In a similar vein, [39] estimate 1D marginal distributions of CNN features (using kernel density estimation) for OOD detection of MRIs.

3 Method

3.1 Background

3.1.1 Markov Random Fields (MRFs)

MRFs [16] express a probability density function of an image, z, as an energy-based model:

$$p(z) = \frac{1}{\mathcal{C}} \exp(-E(z)) \tag{1}$$

where $\mathcal{C}$ is a normalization constant. The energy of the image is defined as the sum of energies (potential functions) of all constituent $\mathcal{R}^{k\times k}$ patches (cliques), z_k:

$$E(z) = \sum_{k \in \mathcal{K}} E(z_k) \tag{2}$$

where $\mathcal{K}$ denotes the set of all $k \times k$ patches. Typically, the energy function $E(z_k)$ is defined over relatively small patches and is hand-crafted - for instance, to encode smoothness.

3.1.2 Field of Experts (FoEs)

FoEs [15] extend the MRF idea by learning the energy function from data. Specifically, the energy of image patches, z_k, is written in the Product-of-Experts (PoE) framework [17], [40]:

$$E(z_k) = -\sum_{j=1}^{J} \log p(f_j(z_k); \alpha_j) \tag{3}$$

Substituting this into Eqn. 2, the energy of the total image, z, becomes

$$E(z) = -\sum_{k \in \mathcal{K}} \sum_{j=1}^{J} \log p(f_j(z_k); \alpha_j) \tag{4}$$

The corresponding probability density function of the image, z, becomes

$$p(z) = \frac{1}{\mathcal{C}} \prod_{k \in \mathcal{K}} \prod_{j \in \text{experts}} p(f_j(z_k); \alpha_j) \tag{5}$$

Here, $f_j : \mathcal{R}^{k \times k} \rightarrow \mathcal{R}$ are *expert* functions, and α_j are parameters of the 1D distributions of experts' scalar outputs. The key idea in PoE and thus, FoE models is that each expert models a particular low-dimensional aspect of the high-dimensional data. Due to the product formulation, only data points that are assigned high probability by *all* experts are likely under the model. In [17], [40], [15], f_j and α_j are learned using an algorithm known as contrastive divergence, such that images in a training dataset are assigned low energy values, and all other points in the image space are assigned high energy values.

3.2 Field-of-Expert (FoE) Priors for TTA

We now describe the proposed TTA method for acquisition-related DS in medical imaging. Fig. 2 shows a representative CNN architecture in this framework. An image, x, is passed through a shallow normalization module, N_ϕ, which outputs a normalized image, z. N_ϕ consists of a few (2-4) convolutional layers with relatively small kernel size (1-3) and stride 1, and outputs z, which is a feature with the same spatial dimensionality and the same number of channels as x. z is passed through a deep CNN, S_θ, which produces the output y. y is formulated as per the task at hand - for instance, it can be a segmentation mask, a deformation field, a super-resolved image, etc. We consider 2D CNNs, but in principle, the method may be extended to 3D architectures as well. S_θ and N_ϕ are trained using labelled input-output pairs from the training distribution. At test-time, S_θ is fixed, while N_ϕ is adapted for each test volumetric image.

A representative architecture for S_θ is shown in the lower part of Fig. 2. Let f_l: $\mathcal{R}^{N_x \times N_y} \rightarrow \mathcal{R}^{N_{xl} \times N_{yl} \times C_l}$ denote the function that takes as input z, and outputs the features of the l^{th} convolutional layer of S_θ. Further, let f_{cl}: $\mathcal{R}^{N_x \times N_y} \rightarrow \mathcal{R}^{N_{xl} \times N_{yl}}$ denote the function that takes as input z, and outputs the c^{th} channel of the l^{th} convolutional layer of S_θ. If k_l is the receptive field at f_{cl} with respect to z, each pixel in the output of f_{cl} can be seen as a 1D projection of a $k_l \times k_l$ patch of z, i.e., an expert function.

3.2.1 The FoE-CNN model

We model the distribution of normalized images, z, using the FoE formulation (Sec. 3.1) with the 3 modifications.

(i) **Multiple patch sizes**: Firstly, note that in the original FoE model, the energy function is defined in terms of input patches of a *single* patch size. We consider multiple patch sizes to

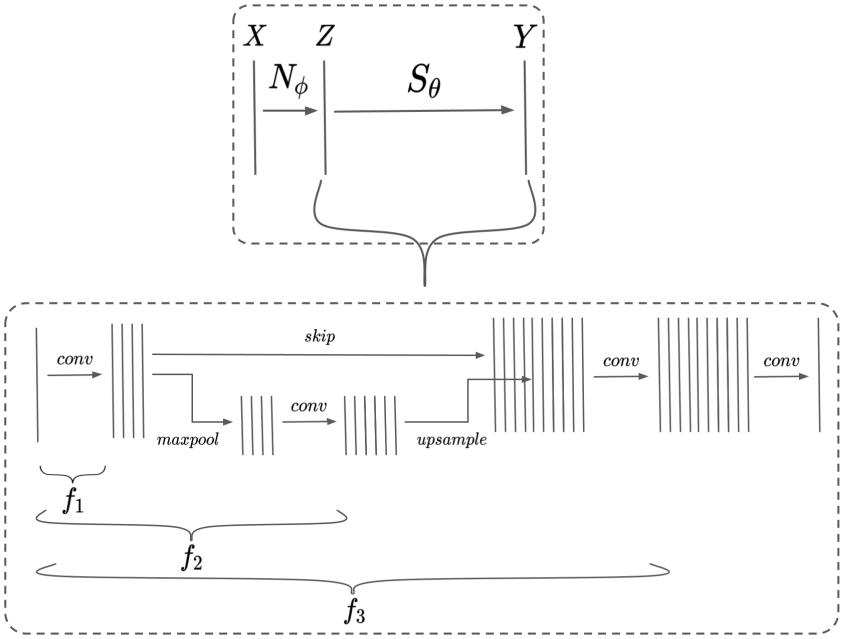

Figure 2: Representative schematic of a test-time adaptable CNN.

define the energy. Specifically, if S_θ consists of L convolutional layers, we consider L patch sizes - namely, the receptive fields of all the convolutional layers of S_θ.

$$E(z) = \sum_{k=k_1}^{k_L} \sum_{k\times k \text{ patches}} E((z_k)) \tag{6}$$

(ii) **Task-specific experts**: Secondly, we define the energy function for each patch size, using a separate PoE model. However, unlike [15], we do not learn the expert functions using contrastive divergence. Instead, we construct a task-specific FoE model by using the functions f_{cl} of S_θ as C_l experts to describe the energy of patches of z of size $k_l * k_l$:

$$E(z_{k_l}) = -\sum_{c=1}^{C_l} \log\, p(f_{cl}(z_{k_l}; \alpha_{cl})) \tag{7}$$

As previously noted, $f_{cl}(z_{k_l})$ are individual pixels of the c^{th} channel of the l^{th} convolutional layer of S_θ. Thus, $p(f_{cl}(z_{k_l}; \alpha_{cl}))$ is the 1D distribution of these pixel values, and α_{cl} are its parameters. Combining Eqns 7 and 6, and inserting the resulting energy function into the FoE formulation (Sec. 3.1), the corresponding PDF of the normalized images can be written as:

$$p(z) = \frac{1}{\mathcal{C}} \prod_{l=1}^{L} \prod_{k_l * k_l \; patches} \prod_{c=1}^{C_l} p(f_{cl}(z_{k_l})) \tag{8}$$

Change of notation: For ease of reading, let us denote expert outputs, $f_{cl}(z_{k_l})$, by u and their distribution, $p(f_{cl}(z_{k_l}); \alpha_{cl})$, by $p_{cl}(u; \alpha_{cl})$. Also, note that the product over $k_l * k_l$ patches of Z is the product over the pixels of f_{cl}. Thus, we have:

$$p(z) = \frac{1}{\mathcal{C}} \prod_{l=1}^{L} \prod_{c=1}^{C_l} \prod_{i=1}^{N_{xl} * N_{yl}} p_{cl}(u_i; \alpha_{cl}) \tag{9}$$

The functions learned in S_θ act as *task-specific* experts. We hypothesize that matching the distributions of the outputs of such experts during TTA is likely to be beneficial for improving the task performance for the test images.

(iii) **Estimation of experts' distributions**: We approximate the expert distributions, $p_{cl}(u;\alpha_{cl})$, as 1D Gaussian distributions, with $\alpha_{cl} = \{\mu_{cl}, \sigma_{cl}\}$:

$$p_{cl}(u;\alpha_{cl}) = \mathcal{N}(\mu_{cl}, \sigma_{cl}),\ \mu_{cl} = \frac{1}{N_z}\sum_z \frac{1}{N_{xl}*N_{yl}}\sum_i u_i,\ \sigma^2_{cl} = \frac{1}{N_z}\sum_z \frac{1}{N_{xl}*N_{yl}}\sum_i (u_i - \mu_{cl})^2 \tag{10}$$

Here, the outer sum, $\sum_z$, is over all samples of normalized images z, and the inner sum, $\sum_i$, is over all pixels of the feature at the c^{th} channel of the l^{th} layer.

Eqn. 9 defines the complete field of CNN experts probability model (FoE-CNN) of the normalized images, z, with the individual expert PDFs given either by Eqn. 10.

We further analyze the effect on TTA of modelling $p_{cl}(u)$ using kernel density estimation (KDE) (see 'analysis experiments' in Sec. 4.1.3). While this approach can capture higher-order moments of the distributions, we observed that the resulting PDFs were relatively similar to their Gaussian approximations. Thus, for simplicity, we propose to use the Gaussian approximation in the method, and show the effect of using KDE in the appendix.

3.2.2 TTA using FoE-CNN

We propose to use to the FoE-CNN model for TTA in the following setting: at the training site (e.g. hospital), multiple labelled volumetric images are available from the training distribution, but at the test site, we would like to adapt the model for each volumetric test image separately. Therefore, we consider subject-specific distributions $p^s(z)$ (Eqn. 9), consisting of subject-specific 1D PDFs, $p^s_{cl}(u)$ (Eqn. 10). That is, after training N_ϕ and S_θ using data from the training distribution, we compute and save the 1D PDFs, $p^s_{cl}(u)$, for all channels of all layers, for all training subjects. These are transferred to the test site. A practical advantage here is that only summary statistics of the 1D distributions are transferred - this provides benefits in terms of privacy and memory requirements, as compared to transferring large CNN models or the training distribution images themselves. Now, for TTA, we have to make the following two design choices.

(i) **Log-likelihood maximization v/s Distribution matching**: Given a test subject t, there are two possible ways to carry out TTA. One option is to maximize the log-likelihood of the normalized image corresponding to the test image, under the FoE-CNN model computed for the training images. Further, since the distribution of the training subjects are also modelled subject-wise, we additionally take an expectation over the training subjects:

$$max_\phi\ E_{p(s)}\left[E_{p^t(z)}\ \log\ p^s(z)\right] \rightarrow max_\phi\ E_{p(s)}\left[E_{p^t(z)}\ \sum_{l=1}^{L}\sum_{c=1}^{C_l}\sum_{i=1}^{N_{xl}*N_{yl}}\ \log\ p^s_{cl}(u_i)\right] \tag{11}$$

We approximate the expectation with respect to $p^t(z)$ using randomly chosen 2D slices of the test subject's volumetric image. A potential problem with this TTA formulation may be that it attract all pixels u_i towards the modes of p^s_{cl}. To circumvent this issue, we propose to model the distribution of the normalized images corresponding to the 2D slices of the test subject, $p^t(z)$, also using the FoE-CNN model (Eqn. 9). Now, a suitable divergence measure, D, between this and the distributions of the training subjects can be minimized.

$$min_\phi\ E_{p(s)}\ D(p^s(z), p^t(z)) \tag{12}$$

However, the normalization constant $\mathcal{C}$ in Eqn. 9 is intractable to compute and may be different for the two distributions. As well, commonly used divergence measures (such as f-divergences) require integration over the entire space over which the distributions are defined. Clearly, this is not possible for the high-dimensional normalized images, z. Therefore, for TTA, we match all the 1D expert distributions, p_{cl}, for all channels of all layers. That is, we minimize $L_{FoE-CNN}$ with respect to ϕ, where

$$L_{FoE-CNN} = E_{p(s)}\left[\ \frac{1}{L}\sum_{l=1}^{L}\frac{1}{C_l}\sum_{c=1}^{C_l} D(p^s_{cl}(u), p^t_{cl}(u))\ \right] \tag{13}$$

In particular, we minimize the KL-divergence between the individual 1D distributions for the training and test images. As the 1D PDFs are approximated as Gaussians, the KL-divergence can be computed in closed form. Further, for this choice of divergence measure, we show (Appendix Sec. 6.1) that minimizing the objective in Eqn. 12 is equal to minimizing the one in Eqn. 13 plus log of the normalization constant $\mathcal{C}$ for the test image.

(ii) **Incorporating information from multiple training subjects**: As mentioned previously, we consider subject-specific distributions of the normalized images. This provides us with two options for carrying out distribution matching for TTA: (a) minimize the *divergence* of the test subject's distribution with the expected distribution over all training subjects: $min\ D(E_{p(s)}[p^s_{cl}(u)], p^t_{cl}(u))$. (b) minimize the *expected divergence* of the test subject's distribution with the distribution of each training subject: $min\ E_{p(s)}[D(p^s_{cl}(u), p^t_{cl}(u))]$. For KL-divergence, we show in Appendix 6.2 that two objectives are related as follows:

$$D_{KL}\big(E_{p(s)}[p^s_{cl}(u)], p^t_{cl}(u)\big) = -\,E_{p(s)}[D_{KL}(p^s_{cl}(u), E_{p(s)}[p^s_{cl}(u)])] + E_{p(s)}[D_{KL}(p^s_{cl}(u), p^t_{cl}(u))] \tag{14}$$

As the first term on the right-hand side of Eqn 14 does not depend on the test image, TTA should, in principle, be equivalent for both ways of incorporating information from multiple training subjects. However, computing (b) in practice requires only one monte-carlo (MC) approximation, while computing (a) requires three MC approximations over the training subjects. Thus, the variance of (b) will be less than that of (a) [41]. With this reasoning, we choose (b) over (a) in the proposed TTA objective (Eqn 13).

3.2.3 Additional experts using principal component analysis (PCA)

We note that the task-specific experts, f_{cl}, in the proposed probability model (Eqn. 9) take as inputs patches of increasing patch sizes. The experts f_{cL} have the largest receptive field, k_L, - thus, they model spatial correlations in $k_L \times k_L$ patches. Depending on the architecture of S_θ, this may or may not cover the entire spatial dimensionality of the normalized image z. We hypothesize that considering spatial correlations in even larger image patches may further improve the proposed TTA. Furthermore, even within the already considered patch sizes, the task-specific experts derived from S_θ may not necessarily capture all spatial correlations that are relevant for distinguishing and improving the task performance when faced with acquisition-related DS.

(i) **The FoE-CNN-PCA model**: We consider additional expert functions that encode spatial correlations at the layer with the largest receptive field. To do so, we use PCA [42], [43]. For all the training images, we extract the last layer features, $f_{cL}(z)$. Next, for each channel of $f_{cL}(z)$, we extract $r \times r$ patches with stride d. We carry out PCA of these patches and save the first G principal components. Now, for each channel c, we compute the PCA coefficients, v, for all extracted patches of all training images. The functions that output the PCA coefficients are considered the additional experts. We compute subject-wise 1D PDFs in each principal dimension, $p^s_{cg}(v)$, where $c = 1, 2, ...C_L$, $g = 1, 2, ...G$, $s = 1, 2, ...n_{tr}$.

(ii) **PCA of *active* patches**: For the task of image segmentation, we noticed that the marginal distributions of the features f_{cL} have two distinct modes - one corresponding to the regions of interest, and one to "background" regions, which are not relevant for the task at hand. In several segmentation applications, the background consists of many more pixels than the foreground classes combined. In such cases, PCA may be unable to find directions of variance within the foreground regions, matching marginal distributions of which may be more useful for TTA. To tackle this problem, we consider only *active* patches while doing PCA. Active patches are defined as those whose central pixel's predicted foreground segmentation probability is greater than a threshold τ.

(iii) **TTA using FoE-CNN-PCA**: The principal components computed on the training images, as well as the expert PDFs of the principal coefficients are transferred to the test site.

When a test image t arrives, patches of its features, $f_{cL}(z)$, are extracted, active patches are retained and the saved principal components are used to compute the corresponding expert PDFs, $p^t_{cg}(v)$. The matching of the additional PCA coefficient PDFs is included in the TTA optimization. That is, we minimize $L_{FoE-CNN-PCA}$ with respect to ϕ, where

$$L_{FoE-CNN-PCA} = E_{p(s)}\Big[\frac{1}{L}\sum_{l=1}^{L}\frac{1}{C_l}\sum_{c=1}^{C_l} D_{KL}(p^s_{cl}(u), p^t_{cl}(u)) + \lambda\frac{1}{C_L}\sum_{c=1}^{C_L}\frac{1}{G}\sum_{g=1}^{G} D_{KL}(p^s_{cg}(v), p^t_{cg}(v))\Big] \quad (15)$$

A hyperparameter, λ, is used to weigh the contribution of the PCA experts with respect to the CNN ones.

4 Experiments

We validated the proposed method for tackling the DS problem on two medical image analysis tasks - segmentation (Sec. 4.1) and atlas registration (Sec. 4.2).

4.1 Segmentation

4.1.1 Datasets

We considered MRI segmentation for 5 anatomies (names of the segmented foreground classes are shown brackets) - (i) T2w prostate (whole organ), (ii) Cine cardiac (myocardium, left and right ventricles), (iii) T1w spine (spinal cord grey matter), (iv) healthy T1w brain (cerebellum gray matter, cerebellum white matter, cerebral gray matter, cerebral white matter, thalamus, hippocampus, amygdala, ventricles, caudate, putamen, pallidum, ventral DC, CSF and brain stem) and (v) diseased FLAIR brain (cerebral white matter hyper-intensities). In total, we used data from 17 centers. We used FreeSurfer [44] generated ground truth segmentations for the healthy brain images, while expert manual ground truth annotations were available for all other datasets. Table 1 shows details of all datasets; Fig. 3 shows example images.

Dataset	Center	Vendor	Field	N_I	$N_{tr}\|N_{vl}\|N_{ts}$
Prostate [45], [46], [47]					
NCI-13	RUNMC, Nijmegen	S	3	30	15|5|10
NCI-13	BMC, Boston	P	1.5	30	15|5|10
Promise12	UCL, London	S	1.5, 3	13	(6|2|5)x2
Promise12	HK, Bergen	S	1.5	12	(5|2|5)x2
Promise12	BIDMC, Boston	G	3	12	(5|2|5)x2
Private	USZ, Zurich	S	3	68	48|10|10
Cardiac [48]					
M&Ms	CSF, Barcelona	P	1.5	50	30|10|10
M&Ms	UHE, Hamburg	P	1.5	25	10|5|10
M&Ms	HVDH, Barcelona	S	1.5	75	55|10|10
Spinal Cord Grey Matter [49]					
SCGM	PM, Montreal	S	3	10	(5|2|3)x3
SCGM	USZ, Zurich	S	3	10	(5|2|3)x3
SCGM	VU, Nashville	P	3	10	(5|2|3)x3
SCGM	UCL, London	P	3	10	(5|2|3)x3
Brain [50], [51]					
HCP	HCP, Missouri	S	3	35	20|5|10
ABIDE	AC, Caltech	S	3	25	10|5|10
White Matter Hyperintensities [52]					
WMH-17	UMC, Utrecht	P	3	20	(10|5|5)x2
WMH-17	NUHS, Singapore	S	3	20	(10|5|5)x2

Table 1: Details of segmentation datasets for 5 anatomies. The vendors S, P and G refer to Siemens, Philips and GE, respectively. N_I refers to the total number of 3D images, and the last column refers to the training, validation and test split. For some datasets, the split is followed by x2 or x3. This refers to the number of dataset splits that were done to get a reasonable number of test images in datasets with a low N_I. Among the prostate datasets, RUNMC and UCL were acquired with surface coil, while the rest of the datasets were acquired with endo-rectal coil.

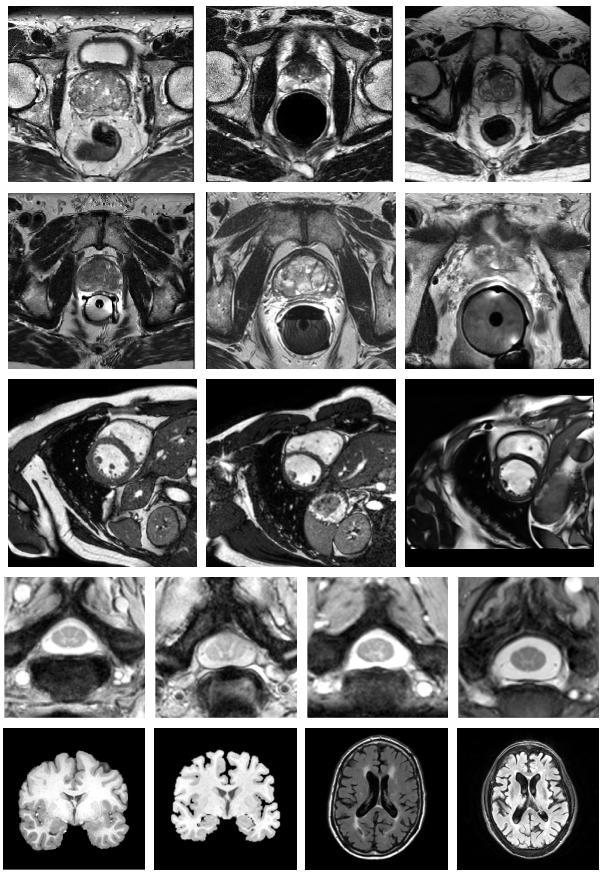

Figure 3: Example images from the different datasets used for the segmentation experiments. Rows 1, 2: prostate T2w MRIs from different centers (RUNMC, BMC, UCL, HK, BIDMC and USZ), row 3: cardiac T1w images (CSF (l), UHE (c), HVHD (r)), row 4: spine MRIs, row 5 (first two): brain T1w MRIs of healthy subjects (HCP (l), ABIDE-CALTECH (r)), row 5 (last two): brain FLAIR MRIs of subjects with white matter hyperintensities (UMC (l), NUHS (r)). Please refer to Table 1 for details about the imaging protocol differences.

4.1.2 Pre-processing

We pre-processed all images by (a) removing bias fields with the N4 algorithm [53], (b) linearly normalizing the intensities to 0-1 range using the 1^{st} and 99^{th} percentile values, and clipping the values to 0 and 1, (c) re-scaling all images of a particular anatomy to the same in-plane isotropic resolution: $0.625\,mm^2$, $1.33\,mm^2$, $0.25\,mm^2$, $0.7\,mm^2$ and $1.0\,mm^2$ for prostate, cardiac, spine, brain and WMH respectively, and (d) cropping / padding zeros to have the same in-plane image size: 200x200 for the spine images and 256x256 for other anatomies. The evaluation for each test image was done in its original resolution and size. For the brain datasets, we additionally set the intensities of the skull voxels to 0.

4.1.3 Experiments

We used the same architecture for N_ϕ and S_θ as in [11]. N_ϕ consisted of 3 convolutional layers of kernel size 3, number of output channels 16, 16 and 1, and an expressive activation function ($act(x) = \exp -(x^2/\sigma^2)$) with a learnable scale σ for each channel. S_θ followed a U-Net [54] like encoder-decoder structure with skip connections, and batch normalization layers following each convolutional layer. The ReLU activation function was used in S_θ.

For each anatomy, we used the institution in the first row in Table 1 as the training distribution, and the remaining institutions as separate test distributions. In this setup, we carried out the following experiments:

(i) **Baseline**: We trained a CNN ($N_\phi + S_\theta$) using labelled images from the training distribution. The training was done by minimizing the Dice loss [55] using an Adam optimizer with a learning rate of 0.001 and a batch size of 16. The optimization was run for 30000 iterations, and the model selection criterion was the average Dice score on the validation dataset. For datasets where the total number of images was very small, splits were created as indicated in Table 1, and average test scores are reported. The dataset splits were designed in such a way that we had 10 test volumes from each test distribution (except for the spine images, where the number of test volumes was 9).

(ii) **Strong baseline**: As described in Sec. 2.1, several domain generalization methods have been proposed to tackle acquisition-related DS in medical image analysis. We implemented stacked data augmentations [8], which present an effective and general DG approach. The implementation details were the same as in [11]: for every image in a training batch, each transformation (translation, rotation, scaling, elastic deformations, gamma contrast modification, additive brightness and additive Gaussian noise) was applied with probability 0.25. This functioned as a *strong baseline*, the performance of which we sought to improve with the proposed TTA approach.

(iii) **Benchmark**: The best performance on images from a test distribution can be achieved by training a new model in a supervised manner, using a separate set of labelled images from the test distribution. As some of the datasets contained only a small number of images to start with, we instead used a *transfer learning* benchmark - that is, the model trained on the training distribution was fine-tuned using labelled images from the test distribution. The fine-tuning was done with the Adam optimizer for 5000 iterations, with a learning rate of 0.0001 and batch size of 16. This model served as the benchmark.

(iv) **Test-Time Adaptation**: We compared the proposed approach (TTA-FoE-CNN-PCA) with four existing TTA works: TTA-Entropy-Min [22], TTA-DAE [11], TTA-AE [12] and TTA-FoE-CNN [37].

Using the *strong baseline* model as the starting point, TTA was run for N_{tta} epochs for each test subject. In each epoch, averaged gradients over batches of size b_{tta} were used to update the network parameters with a learning rate of lr_{tta}. N_{tta} was set to 200 for the healthy brain dataset (due to its high through-plane size) and to 1000 for all other datasets. b_{tta} was set to 8 for all datasets except SCGM, where it was set to 2 as some images had less than 8 slices. Specific implementation details for each TTA method are provided below.

(a) **TTA-Entropy-Min** [22]: The normalization module, N_ϕ, was adapted for each test subject, with lr_{tta} as 0.0001.

(b) **TTA-DAE** [11]: A 3D denoising autoencoder was trained in the space of segmentation labels, using the same corruption distribution as proposed in the original paper. Similar to the original implementation, healthy brain segmentations were downsampled in the through-plane direction by a factor of 4, to overcome memory issues. lr_{tta} was set to 0.001.

(c) **TTA-AE** [12]: Instead of adapting N_ϕ, *adaptor* modules A^x, A^1, A^2 and A^3 were introduced and adapted for each test subject as was done in the original article. We experimented with different settings of [12] so as to get the best results for the datasets used in our experiments (Appendix 6.3). The architectures of the adaptors were kept the same as proposed in [12], with one change: the instance normalization layers in A_X were discarded as they lead to instability during TTA. Two other changes were done to further improve the performance and stability: (a) average gradients over all batches in a single TTA epoch were used for the TTA updates (as described in Sec 3.5 in [11]) and (b) the lr_{tta} was set to 0.00001. Five 2D autoencoders (AEs) (with the same architectures as in [12]) were trained and the weight of the orthogonality loss, λ_{orth}, was set to 1.0, as done in [12]. We observed that driving the TTA using losses from two AEs (at the input and output layers) provided better performance than using all 5 AEs.

With these modifications, TTA-AE worked in a stable manner, without having to resort to early stopping as done in [12].

(d) **TTA-FoE-CNN**: At the end of the *strong baseline* training, the FoE-CNN model was constructed by computing 1D PDFs for all channels of all layers of S_θ, for each training subject. For the chosen architecture of S_θ, this amounted to 704 channels. As the PDFs were approximated as Gaussians, two parameters were stored per PDF. In principle, this method resembles the approach proposed in [37].

(e) **TTA-FoE-CNN-PCA** For computing the additional expert PDFs of the FoE-CNN-PCA model, the following steps were followed: (a) For all training images, features from the last layer of S_θ were extracted (from here, a 1x1 convolutional layer provided the segmentation logits). In the chosen architecture, these features were of the same spatial dimensions as the images and had $C_L = 16$ channels. (b) For each channel in these features, patches of size $r \times r = 16 \times 16$ were extracted with stride $d = 8$. (c) From these, only *active* patches (that is, patches whose central pixel's predicted foreground probability was greater than $\tau = 0.8$) were retained. As CNNs typically make high confidence predictions, this step is likely to be insensitive to the exact value of τ. To obtain a comparable number of *active* patches to other anatomies, the stride d was set to 2 for the WMH images, where the foreground size was particularly small. (d) PCA was done using the active patches of all training images, and the first $G = 10$ principal directions were identified. (e) Finally, 1D PCA expert PDFs were computed similar to the 1D CNN expert PDFs: for all channels of the last layer of S_θ, for all principal directions, for each training subject. In total, we had $C_L \times G = 160$ PCA expert PDFs for each training subject. The hyperparameter, λ, was empirically set to 0.1, and lr_{tta} to 0.0001.

(v) **Analysis Experiments**

(a) **Approximating Expert Distributions with KDEs rather than as Gaussians**: In the experiments above, we approximated the individual expert distributions of the FoE model (Eqn. 9) as Gaussian distributions. As the expert distributions are in 1D, we also considered non-parametric estimation methods, such as kernel density estimation (KDE) [56], [57], [58]. This approach is provides an alternative to the *soft-binning*-based non-parametric approximation in [37]. In general, KDEs have the two important downsides. Firstly, the number of data points required to get a reliable density estimate grows exponentially with dimensionality. This is not a concern in low dimensions. Secondly, KDEs require access to the training samples to evaluate the PDF at a given test sample. Again, in low dimensions (e.g 1D), it may be feasible to evaluate and save the KDE over the entire domain of interest when one has access to the training samples. Thus, the training samples are no longer required at test time. Accordingly, we compute

$$p_{cl}(u) = \frac{1}{N_z} \sum_z \frac{1}{N_{xl} * N_{yl}} \sum_i \frac{1}{\sqrt{2\pi}} \exp\left(-\alpha \, ||u - u_i||_2^2\right)) \tag{16}$$

Being more expressive than Gaussians, KDEs can potentially capture higher-order moments of the expert distributions - thus leading to more accurate distribution matching and better TTA performance. Implementation-wise, when the 1D PDFs were estimated as Gaussians, the KL-divergence could be computed in closed form. When KDEs were used, we numerically computed the integral in the KL-divergences using Riemann sums.

(b) **Effect of the weighting between the CNN and the PCA experts**: The effect of the weighting parameter, λ, in Eqn. 15, was empirically analyzed for the 5 test distributions of the prostate segmentation experiment.

4.1.4 Results

The following points can be inferred from the quantitative results of our segmentation experiments (Table 2).

(i) The baseline demonstrates that the DS problem exists for all the 5 anatomies. The difference between the Dice scores on the training and test distributions is sometimes as high as 60 Dice

Method \ Test	UCL	HK	BIDMC	BMC	USZ	UHE	HVHD	USZ	VU	UCL	AC	NUHS
	Prostate					Cardiac		Spine			Brain	WMH
	Supervised Learning on Training Distribution											
Baseline	0.50	0.68	0.29	0.28	0.67	0.86	0.38	0.61	0.82	0.79	0.69	0.00
	Domain Generalization											
Strong baseline [8]	0.77	0.82	0.62	0.77	0.76	0.85	0.80	0.67	0.84	0.88	0.76	0.37
	Test Time Adaptation											
Entropy Min. [22]	0.77	0.81	$0.68^{\triangle}$	0.77	$0.80^{\blacktriangle}$	0.85	$0.80^{\triangle}$	0.67	0.84	0.88	$0.81^{\blacktriangle}$	$0.36^{\blacktriangledown}$
DAE [11]	$\mathbf{0.84}^{\blacktriangle}$	$\mathbf{0.84}^{\triangle}$	$\mathbf{0.75}^{\blacktriangle}$	$\mathbf{0.81}^{\triangle}$	$\mathbf{0.82}^{\blacktriangle}$	$\mathbf{0.87}^{\blacktriangle}$	0.81	**0.69**	$0.80^{\triangledown}$	0.80	$\mathbf{0.82}^{\blacktriangle}$	0.37
AE [12]	0.78	0.83	$0.51^{\triangledown}$	**0.79**	0.79	$0.86^{\blacktriangle}$	0.80	$\mathbf{0.69}^{\triangle}$	$0.84^{\triangle}$	$0.88^{\triangle}$	$0.78^{\blacktriangle}$	$0.24^{\blacktriangledown}$
FoE-CNN [37]	0.78	$0.77^{\triangledown}$	0.64	0.76	0.76	0.86	$\mathbf{0.82}^{\blacktriangle}$	0.68	$\mathbf{0.85}^{\triangle}$	$\mathbf{0.89}^{\triangle}$	$0.79^{\blacktriangle}$	$0.24^{\blacktriangledown}$
FoE-CNN-PCA (Ours)	0.79	0.81	$0.73^{\triangle}$	0.75	0.78	0.85	$\mathbf{0.82}^{\blacktriangle}$	0.68	$0.83^{\triangledown}$	0.88	$0.79^{\blacktriangle}$	$\mathbf{0.42}^{\blacktriangle}$
	Transfer Learning											
Benchmark	0.80	0.85	0.82	0.83	0.84	0.88	0.83	0.78	0.85	0.90	0.88	0.77

Table 2: Dice scores (averaged over all foreground labels and all test subjects) for the segmentation test-distribution datasets. In each column, the highest Dice score among the TTA methods has been highlighted. The Dice scores for test images from the training distribution are: (a) for the baseline: RUNMC 0.86, CSF: 0.82, PM: 0.88, HCP: 0.87, UMC: 0.71, (b) for the strong baseline: RUNMC 0.91, CSF: 0.83, PM: 0.89, HCP: 0.87, UMC: 0.72. Results for the NUHS dataset are mean values over 4 runs. Paired permutation tests were done to measure the statistical significance of the improvement or degradation caused by each TTA method over the strong baseline. $^{\triangle}$ ($^{\triangledown}$) and $^{\blacktriangle}$ ($^{\blacktriangledown}$) indicate improvement (degradation) with p-value less than 0.05 and 0.01, respectively. The stricter significance test (p-value 0.01) was done to counter the multiple comparison problem [59].

points; a model that provides almost perfect segmentations on test images from the training distribution can potentially provide completely un-usable segmentations on test images from a shifted distribution (e.g. test images from a different hospital).

(ii) Data augmentation [8] helps vastly. This *strong baseline* is much more robust to DS than the baseline - in some cases, providing a performance jump as high as 50 Dice points. These results corroborate numerous similar findings in the current literature. Given the generality and effectiveness of the approach, we believe it is imperative that works studying DS robustness in CNN-based medical image segmentation should include stacked data augmentation during training.

(iii) A gap to the benchmark still remains - in most cases, heuristic data augmentation falls short of rivalling the performance of supervised fine-tuning.

(iv) Results of the TTA methods are described below. When making statements about statistical significance of results in the text below, we follow a strict threshold based on Benferroni correction to account for the multiple comparison problem [59]. For each dataset, permutation tests ($n = 100000$) were used to compute statistical significance of the performance improvement or degradation provided by each TTA method with respect to the strong baseline. Thus, 5 comparisons were made for each dataset. So, the p-value threshold was divided by 5.

(a) Entropy minimization-based TTA [22] does not require construction of additional models to capture training distribution traits; yet, it provides performance improvement in some cases. Also, unlike other works [10], we largely do not observe the problem that the entropy minimization leads to all pixels being predicted as the same class. This might have been due to the limited adaptation ability provided by N_ϕ.

However, the performance gains are statistically significant for only 2 out of the 12 test datasets. As well, the performance degrades the strong baseline significantly (although marginally) for the lesion test dataset. Another downside of this method is that it can only be applied for tasks with categorical outputs.

(b) TTA-DAE [11] provides the best performance for the most number of test datasets for healthy tissue segmentations. For 5 out of 11 healthy test datasets, the improvements provided

by this method over the strong baseline are statistically significant. It also leads to a drop of 5 and 8 Dice points in the mean results for the two spine datasets; however, permutation tests show that the drops may be due to large degradation for a small number of test subjects within those datasets. Even so, the large drops in performance for particular subjects may be indicative of the DAE's DS problem - that is, the DAE's outputs may be unreliable when it is fed with segmentations that do not match the heuristically designed noise distribution used for its training.

Furthermore, the DAE fails to improve performance for the lesion dataset. We believe that this reflects its inapplicability to tackle the DS problem in anatomies where reliable shape priors cannot be learned.

In terms of applicability, the DAE-based TTA is also restricted in terms of the tasks that it can applied to. For segmentation, the DAE could be trained by heuristically designing a suitable corruption distribution. It is unclear how to achieve this for other tasks.

(c) Autoencoder-based TTA [12] provides performance improvement in several cases. However, statistically significant improvements could only be obtained for 2 test datasets; even in these cases, the improvements were marginal (1 and 2 dice points). This method also also lead to a drop of 12 and 13 Dice points for the prostate BIDMC and the WMH dataset, respectively; the latter was statistically significant.

(d) The FoE-CNN that in principle resembles the concurrent approach of [37] is overall, less performant than the PCA-based extended FoE proposed in the current work.

(e) As compared to the strong baseline, the proposed FoE-CNN-PCA based TTA improves performance for 7 and retains performance for 2 out of the 12 test distributions. In particular, the proposed method shows promising performance gains in cases where the other task-agnostic methods falter substantially (e.g. prostate BIDMC and WMH). Out of these, the improvements are statistically significant for 3 test datasets, including the lesion dataset.

The 3 test distributions where the method leads to a performance drop, the drop is relatively small: 3, 1 and 1 Dice points. We claim that this illustrates the stability of the proposed TTA method and validates our initial hypothesis - FoE-based TTA improves performance in the face of acquisition-related DS in medical imaging, while itself being substantially more robust to the DS shift problem that other priors such as the DAE [11] or the AE [12] may be vulnerable to.

Importantly, the proposed method provides the best performance for the task of WMH segmentation - indicating its superiority in cases where CNN-based helper modules such as DAEs [11] may be unable to learn appropriate shape priors. Notably, all competing methods from the literature fail to improve DS robustness for this lesion segmentation experiment; the proposed method is the only approach that shows promising results in this challenging scenario.

(v) Analysis Experiments

(a) Approximating Expert Distributions with KDEs rather than as Gaussians: Comparing the KDEs v/s Gaussian approximations (Fig. 4), we observed that the actual distributions do not differ substantially from their Gaussian approximations. This is also reflected in the TTA results in Table 3 - performance of the proposed method is very similar for both estimates of expert distributions.

(b) Effect of the weighting between the CNN and the PCA experts: Results of this hyper-parameter tuning are shown in Table 4. The introduction of PCA experts with $\lambda = 0.1$ improves TTA performance for 4 of the 5 prostate datasets. However, increasing λ to 1.0 leads to performance decrease in 3 of the 5 datasets. Based on these results, we choose $\lambda = 0.1$ for all datasets of all anatomies.

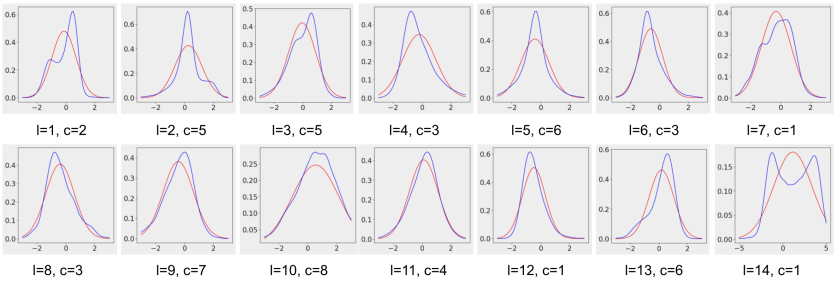

Figure 4: Comparison of KDEs v/s Gaussian approximations (corresponding to a single prostate RUNMC training subject) for modeling the channel PDFs of different layers of the trained segmentation network. $l = 14$ is the last-but-one layer of the network. From here, a 1x1 convolution gives the segmentation logits. In each layer (l), the channel (c) with the visually most-non-gaussian KDE is chosen for visualization. With this choice, some non-Gaussianity is observed in the initial and final layers, while the layers in the middle of the segmentation CNN has highly Gaussian marginal distributions.

Method \ Test	UCL	HK	BIDMC	BMC	USZ
			TTA-FoE-CNN-PCA		
Gaussian	0.79	0.81	0.73	0.75	0.78
KDE	0.79	0.81	0.74	0.76	0.78

Table 3: Effect of approximating 1D distributions of the FoE model with Gaussians v/s kernel density estimation (KDE). Both approximations lead to very similar TTA performance. Fig. 4 provides visual justification of this observation - the 1D distributions of CNN as well as the PCA experts are sufficiently well approximated with Gaussians.

Method \ Test	UCL	HK	BIDMC	BMC	USZ
			TTA-FoE-CNN		
$\lambda = 0.0$	0.78	0.77	0.64	0.76	0.76
			TTA-FoE-CNN-PCA		
$\lambda = 0.1$	**0.79**	0.81	0.73	0.75	**0.78**
$\lambda = 1.0$	0.77	**0.82**	**0.74**	0.74	0.77

Table 4: Effect of the weighting parameter between the CNN and PCA experts in TTA-FoE-CNN-PCA. Based on these results, we choose $\lambda = 0.1$ for all datasets of all anatomies.

4.2 Registration

Next, we checked if the proposed method can tackle acquisition-related DS in another task of high practical importance - registration of brain scans with an atlas. The registration CNN is set up as follows. [2] Let a be an atlas and x be the image. Let a_s and x_s be the corresponding segmentation labels. We treat a as the moving image and register it to x, the fixed image. x is first passed through the normalization module, N_ϕ, to obtain z. z and a are concatenated and passed through a deep CNN, S_θ, which outputs a velocity field v_0. v_0 is exponentiated via a *squaring-and-scaling* layer [60] to obtain a diffeomorphic deformation field, Φ. The Dice loss between the warped moving segmentation, $a_s \odot \Phi$, and x_s is used for training N_ϕ and S_θ. For each test image, N_ϕ is adapted with the proposed TTA method.

[2] Ideally, such registration would be done in 3D. However, to avoid memory issues in 3D CNNs, we conduct experiments in a 2D setup. We believe that this still serves as credible evidence of the method's applicability in this task.

4.2.1 Datasets

We used HCP [50] T1w images as those from the training distribution and ABIDE-STANFORD (AS) [51] and OASIS [61] as two test distributions. We used the atlas provided by [62]. Example images are shown in Fig. 5.

4.2.2 Implementation details

All images were re-sampled to an isotropic 1 mm^3 resolution. Upon visual inspection, the axial slices of the atlas, the HCP and OASIS datasets were roughly aligned in the through plane direction, while the AS volumes were shifted by 10 slices. After accounting for this, we extracted the central 40 axial slices from all volumes. We used 3-label (background, white matter, grey matter) Freesurfer [44] segmentations for HCP, AS and expert segmentations for the atlas and OASIS.

Among the TTA methods, we note that TTA-Entropy-Min. [22] can only be applied in cases where S_θ outputs a probability distribution over a fixed number of classes; it is unclear how to extend this for regression. Also, TTA-DAE [11] requires a denoising autoencoder to be trained with corruption patterns that are expected at test time. Designing such corruptions for the registration task is non-trivial. Thus, we compare the proposed method with TTA-AE [12] only.

Method \ Test	HCP	AS	OASIS
Baseline	0.847	0.751	0.864
Strong baseline [8]	0.843	0.786	0.873
TTA-AE [12]	-	0.795	0.868
TTA-FoE-CNN-PCA	-	$0.795^{\triangle}$	$0.870^{\triangledown}$
Benchmark	-	0.821	0.883

Table 5: Dice scores (averaged over all foreground labels and all test subjects) for the registration experiments. We measured the statistical significance of the improvement or degradation caused by each TTA method over the strong baseline, using paired permutation tests. $^{\triangle}$ ($^{\triangledown}$) and $^{\blacktriangle}$ ($^{\blacktriangledown}$) indicate improvement (degradation) with p-value less than 0.05 and 0.025, respectively. The stricter significance test (p-value 0.025) was done to counter the multiple comparison problem [59].

4.2.3 Results

The following points can be inferred from the quantitative results of our registration experiments (Table 5).

(i) In the baseline, the DS problem is quite stark for the AS dataset, but relatively mild for the OASIS dataset.

(ii) Stacked data augmentation [8] (with the same hyperparameters as in the segmentation experiments) provides substantial gains for registration as well. In this *strong baseline*, the performance of OASIS is already almost as good as in the benchmark. However, a gap between the *strong baseline* and the benchmark exists for the AS dataset. In a practical setting, TTA methods would typically be unaware of the extent to which the DS problem exists before TTA. In this scenario, TTA methods should ideally improve performance for datasets which suffer from the DS problem, and retain performance for other datasets.

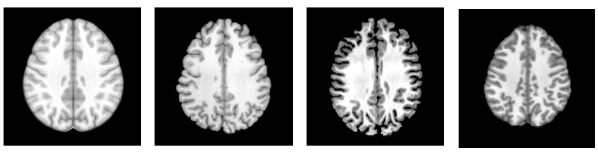

Figure 5: From left to right: a 2D slice from the atlas and example slices from three datasets: HCP, ABIDE-STANFORD (AS) and OASIS.

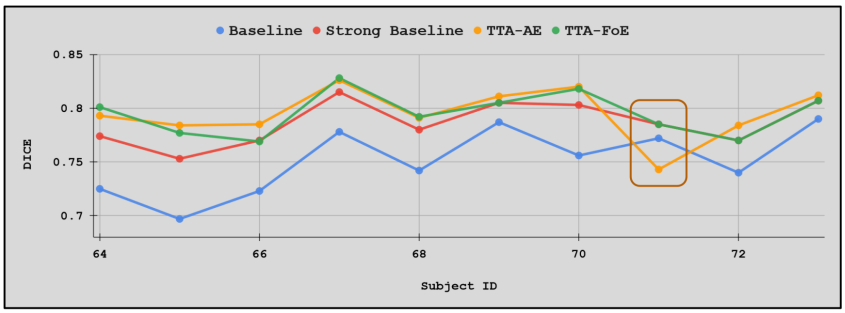

Figure 6: Dice scores for individual subjects of the AS dataset. Overall, both TTA methods perform similarly well for most subjects. However, the brown box highlights one subject, where TTA-AE leads to performance degradation as compared to both the baseline as well as the strong baseline. Such degradation does not occur with the proposed method.

(iii) On average, both the proposed method and TTA-AE [12] improve the performance of the strong baseline for the AS dataset and retain it for the OASIS dataset. However, the changes in performance brought about by TTA are not statistically significant for any dataset. Subject-wise results are shown in Fig. 6.

We believe that this set of experiments demonstrates that, in principle, the proposed method can be applied to the image registration task. Further, it achieves comparable results in this task to a previously existing task-agnostic TTA method [12]. However, the registration task seems to be particularly challenging for both methods.

5 Discussion

In this section, we discuss the strengths and limitations of the proposed methods, and outline avenues for further research.

5.1 Strengths of the proposed method

1. **Lesion segmentation performance**: All existing TTA methods failed to tackle the DS robustness problem for lesion datasets. Furthermore, 3 out of the 4 existing methods lead to statistically significant performance degradation over the strong baseline. In particular, TTA-DAE, which shows strong performance for healthy tissue segmentation, fails to improve performance for lesions due to the difficulty in learning appropriate shape priors. The proposed method provided substantial as well as statistical significant performance improvement in this challenging scenario.

2. **Applicability to multiple tasks**: Our experiments indicate that the proposed method can, in principle, be applied to multiple tasks. Such generality is an important asset; the DS problem is likely to occur in all medical image analysis tasks.

3. **Generalization of previous works**: This work makes the novel contribution of casting the marginal distribution matching idea in a Field-of-Experts formulation. This observation allows us view several recent works [34], [35], [33], [37] as instances of our general framework, and enables us to build on these works by introducing additional expert functions in the form of principle loadings of feature patches.

5.2 Limitations of the proposed method

1. **Performance on healthy tissue segmentation is not as good as TTA-DAE**: Although the proposed method improves performance of the strong baseline in a large number of the test datasets, methods specifically designed for image segmentation often outperform the more general method developed in this work.

2. **Matching the distribution of individual experts rather than the full FoE distribution**: An important relaxation in TTA-FoE is between Eqn. 12 and Eqn. 13. Eqn. 12 seeks to match the full FoE distribution between test and training images. However, this is not possible as the computation of the normalization constant $\mathcal{C}$ is intractable. Thus, we instead carry out the relaxed optimization, as shown in Eqn. 13 - minimizing divergence between the distributions of individual experts. It is unclear if the relaxed optimization is theoretically guaranteed to converge, or if the alignment of individual experts may compete with one another. In practice, we observe the optimization to converge for all the test images, across all test distributions and anatomical regions. We believe that this behaviour could have been aided by the initial closeness of the individual expert distributions. Thus, the proposed TTA method works well for *small DS* (due to changing scanners or acquisition protocol parameters within the same imaging modality), but may not be suitable for *large DS* (for instance, across imaging modalities).

5.3 Avenues for further exploration

1. **Choice of expert functions of the FoE Model**: In initial product-of-experts [17], [40] and field-of-experts [15] works, the experts are parameterized and learned from data, such that the probability model assigns high likelihood values to the true data - for example, using algorithms such as contrastive divergence. Further, parameters of the expert PDFs are also learned from data. In contrast, in this work, we used two types of experts - (1) the task-specific convolutional filters learned in the segmentation or registration CNN and (2) projections onto principal components of patches in the last layer of the segmentation or registration CNN. Thus, we used task-specific experts, and only learned the parameters of the expert PDFs from data. In other words, we aligned the test and training normalized images, in terms of their projections that are the most relevant for the task CNN to perform the task at hand. Such a task-specific probability model could be augmented with learned experts, as proposed in earlier works [17], [40], [15]. The extended model would potentially capture further projections of the normalized images, apart from the task-specific projections considered in this work. It is unclear if alignment along such directions between test and training images would further improve TTA performance; we defer this analysis to future work.

2. **Choice of the divergence measure to be minimized for TTA**: We minimize the KL-divergence between expert distributions. It may be interesting to investigate if aligning distributions by minimizing other divergences may lead to improved TTA performance. For instance, in concurrent work, [37] minimize a symmetric version of the KL divergence. Leveraging the low-dimensionality of the expert outputs, even divergence measures that cannot be computed in closed form, may be easy to compute numerically.

Acknowledgments

This work was supported by the following grants: (a) Swiss Platform for Advanced Scientific Computing (PASC), (b) Swiss National Science Foundation Grant 205320-200877, (c) Clinical Research Priority Program (CRPP) Grant on Artificial Intelligence in Oncological Imaging Network, University of Zurich.

References

[1] Daniel Castro, Ian Walker, and Ben Glocker, "Causality matters in medical imaging," *Nature Communications*, 2020.

[2] Annegreet Van Opbroek, Arfan Ikram, Meike Vernooij, and Marleen De Bruijne, "Transfer learning improves supervised image segmentation across imaging protocols," *IEEE transactions on medical imaging*, 2014.

[3] Nima Tajbakhsh, Jae Shin, Suryakanth Gurudu, Todd Hurst, Christopher B Kendall, Michael Gotway, and Jianming Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE transactions on medical imaging*, 2016.

[4] Neerav Karani, Krishna Chaitanya, Christian Baumgartner, and Ender Konukoglu, "A lifelong learning approach to brain mr segmentation across scanners and protocols," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018.

[5] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al., "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *International conference on information processing in medical imaging*. Springer, 2017.

[6] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K Moyo, Michael R Savona, Richard G Abramson, and Bennett A Landman, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE transactions on medical imaging*, 2018.

[7] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker, "Domain generalization via model-agnostic learning of semantic features," in *Advances in Neural Information Processing Systems*, 2019.

[8] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu, and Z. Xu, "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE Transactions on Medical Imaging*, 2020.

[9] Billot Benjamin, Douglas Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian Dalca, and Juan Eugenio Iglesias, "Synthseg: Domain randomisation for segmentation of brain mri scans of any contrast and resolution," 2021.

[10] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed, "Source-relaxed domain adaptation for image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020.

[11] Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu, "Test-time adaptable neural networks for robust medical image segmentation," *Medical Image Analysis*, vol. 68, 2021.

[12] Yufan He, Aaron Carass, Lianrui Zuo, Blake E Dewey, and Jerry L Prince, "Autoencoder based self-supervised test-time adaptation for medical image analysis," *Medical Image Analysis*, p. 102136, 2021.

[13] Eric Nalisnick, Akihiro Matsukawa, Yee Teh, Dilan Gorur, and Balaji Lakshminarayanan, "Do deep generative models know what they don't know?," in *International Conference on Learning Representations*, 2019.

[14] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich, "Deep anomaly detection with outlier exposure," in *International Conference on Learning Representations*, 2019.

[15] Stefan Roth and Michael J Black, "Fields of experts: A framework for learning image priors," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005.

[16] Stuart Geman and Donald Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, 1984.

[17] Geoffrey Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, 2002.

[18] Maxime W Lafarge, Josien PW Pluim, Koen AJ Eppenhof, and Mitko Veta, "Learning domain-invariant representations of histological images," *Frontiers in medicine*, 2019.

[19] Xiao Liu, Spyridon Thermos, Alison O'Neil, and Sotirios A Tsaftaris, "Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021.

[20] Quande Liu, Qi Dou, and Pheng-Ann Heng, "Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020.

[21] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International Conference on Machine Learning*. PMLR, 2020.

[22] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *International Conference on Learning Representations*, 2021.

[23] Davis Gilton, Gregory Ongie, and Rebecca Willett, "Model adaptation for inverse problems in imaging," *IEEE Transactions on Computational Imaging*, 2021.

[24] Shady Abu Hussein, Tom Tirer, and Raja Giryes, "Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[25] Shakarim Soltanayev and Se Young Chun, "Training deep learning based denoisers without ground truth data," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. 2018, Curran Associates, Inc.

[26] Shady Abu Hussein, Tom Tirer, and Raja Giryes, "Image-adaptive gan based reconstruction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[27] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Deep image prior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[28] Jaejun Yoo, Kyong Hwan Jin, Harshit Gupta, Jerome Yerly, Matthias Stuber, and Michael Unser, "Time-dependent deep image prior for dynamic mri," *IEEE Transactions on Medical Imaging*, 2021.

[29] Vidit Jain and Erik Learned-Miller, "Online domain adaptation of a pre-trained cascade of classifiers," in *CVPR 2011*. IEEE, 2011.

[30] Boris Chidlovskii, Stephane Clinchant, and Gabriela Csurka, "Domain adaptation in the absence of source domain data," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[31] Jian Liang, Dapeng Hu, and Jiashi Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International Conference on Machine Learning*. PMLR, 2020.

[32] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015.

[33] Masato Ishii and Masashi Sugiyama, "Source-free domain adaptation via distributional alignment by matching batch normalization statistics," *arXiv preprint arXiv:2101.10842*, 2021.

[34] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu, "Adaptive batch normalization for practical domain adaptation," *Pattern Recognition*, 2018.

[35] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge, "Improving robustness against common corruptions by covariate shift adaptation," *Advances in Neural Information Processing Systems*, 2020.

[36] Collin Burns and Jacob Steinhardt, "Limitations of post-hoc feature alignment for robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[37] C. Eastwood, I. Mason, C. Williams, and B. Schölkopf, "Source-free adaptation to measurement shift via bottom-up feature restoration," in *10th International Conference on Learning Representations (ICLR)*, Apr. 2022.

[38] Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon, "Density of states estimation for out of distribution detection," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021.

[39] Ertunc Erdil, Krishna Chaitanya, Neerav Karani, and Ender Konukoglu, "Task-agnostic out-of-distribution detection using kernel density estimation," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*, Cham, 2021, Springer International Publishing.

[40] Max Welling, Simon Osindero, and Geoffrey E Hinton, "Learning sparse topographic representations with products of student-t distributions," *Advances in neural information processing systems*, 2002.

[41] Zdravko Botev and Ad Ridder, "Variance reduction," *Wiley StatsRef: Statistics Reference Online*, 2014.

[42] Karl Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 1901.

[43] Harold Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, 1933.

[44] Bruce Fischl, "Freesurfer," *Neuroimage*, 2012.

[45] N Bloch, A Madabhushi, H Huisman, J Freymann, J Kirby, M Grauer, A Enquobahrie, C Jaffe, L Clarke, and Farahani K., "Nci-isbi 2013 challenge: Automated segmentation of prostate structures.," 2015.

[46] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al., "Evaluation of prostate segmentation algorithms for mri: the promise12 challenge," *Medical image analysis*, 2014.

[47] Anton S Becker, Alexander Cornelius, Cäcilia S Reiner, Daniel Stocker, Erika J Ulbrich, Borna K Barth, Ashkan Mortezavi, Daniel Eberli, and Olivio F Donati, "Direct comparison of pi-rads version 2 and version 1 regarding interreader agreement and diagnostic accuracy for the detection of clinically significant prostate cancer," *European journal of radiology*, 2017.

[48] Víctor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martín-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al., "Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge," *IEEE Transactions on Medical Imaging*, 2021.

[49] Ferran Prados, John Ashburner, Claudia Blaiotta, Tom Brosch, Julio Carballido-Gamio, Manuel Jorge Cardoso, Benjamin N Conrad, Esha Datta, Gergely Dávid, Benjamin De Leener, et al., "Spinal cord grey matter segmentation challenge," *Neuroimage*, 2017.

[50] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al., "The wu-minn human connectome project: an overview," *Neuroimage*, 2013.

[51] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al., "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry*, 2014.

[52] Hugo Kuijf, Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, Jorge Cardoso, Adria Casamitjana, et al., "Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge," *IEEE transactions on medical imaging*, 2019.

[53] Nicholas Tustison, Brian Avants, Philip Cook, Yuanjie Zheng, Alexander Egan, Paul Yushkevich, and James Gee, "N4itk: improved n3 bias correction," *IEEE transactions on medical imaging*, 2010.

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

[55] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016.

[56] David Scott, Richard Tapia, and James Thompson, "Kernel density estimation revisited," *Nonlinear Analysis: Theory, Methods & Applications*, 1977.

[57] Richard Davis, Keh-Shin Lii, and Dimitris Politis, "Remarks on some nonparametric estimates of a density function," in *Selected Works of Murray Rosenblatt*. Springer, 2011.

[58] Emanuel Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, 1962.

[59] JM Bland and DG Altman, "Multiple significance tests: the bonferroni method.," *BMJ: British Medical Journal*, vol. 310, no. 6973, pp. 170, 1995.

[60] Adrian Dalca, Guha Balakrishnan, John Guttag, and Mert Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018.

[61] Daniel Marcus, Tracy Wang, Jamie Parker, John Csernansky, John Morris, and Randy Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," *Journal of Cognitive Neuroscience*, 2007.

[62] Vladimir Fonov, Alan Evans, Kelly Botteron, Robert Almli, Robert McKinstry, Louis Collins, Brain Development Cooperative Group, et al., "Unbiased average age-appropriate atlases for pediatric studies," *Neuroimage*, 2011.

[63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015.

6 Appendix

6.1 Approximating KL-divergence minimization of the full FoE model with KL-divergence minimization of individual expert distributions

We show this analysis for Product of Experts (PoEs). It also holds for FoEs, which are a specific instance of the PoEs formulation. Consider PoE models for the source and target domain normalized images.

$$p^s(z) = \frac{\hat{p}^s(z)}{\mathcal{C}_s}, \;\; \mathcal{C}_s = \int_z \hat{p}^s(z)dz, \;\; \hat{p}^s(z) = \prod_{j=1}^{J} p_j^s(u_j), \;\; u_j = f_j(z)$$

$$p^t(z) = \frac{\hat{p}^t(z)}{\mathcal{C}_t}, \;\; \mathcal{C}_t = \int_z \hat{p}^t(z)dz, \;\; \hat{p}^t(z) = \prod_{j=1}^{J} p_j^t(u_j), \;\; u_j = f_j(z)$$

Here, we explicitly show the subscript j in variables u to indicate that different experts have different 1D co-domains. Now, consider KL-divergence minimization between these distributions:

$$min_\phi \, D_{KL}(p^s(z), p^t(z)) \rightarrow min_\phi \int_z p^s(z) \, \log \, \frac{p^s(z)}{p^t(z)} \, dz$$

$$\rightarrow min_\phi \int_z \frac{\hat{p}^s(z)}{\mathcal{C}_s} \, \log \, \frac{\mathcal{C}_t}{\mathcal{C}_s} \frac{\hat{p}^s(z)}{\hat{p}^t(z)} \, dz$$

$$\rightarrow min_\phi \int_z \frac{\hat{p}^s(z)}{\mathcal{C}_s} \, \log \, \frac{\mathcal{C}_t}{\mathcal{C}_s} \, dz + \int_z \frac{\hat{p}^s(z)}{\mathcal{C}_s} \frac{\hat{p}^s(z)}{\hat{p}^t(z)} \, dz$$

$$\rightarrow min_\phi \, \log \, \frac{\mathcal{C}_t}{\mathcal{C}_s} + \int_z \frac{\hat{p}^s(z)}{\mathcal{C}_s} \frac{\hat{p}^s(z)}{\hat{p}^t(z)} \, dz$$

Note that during TTA, ϕ is fixed for computing the source-domain distribution, while is variable for computing the target-domain distribution. Thus, ignoring the 'source-domain-only' terms, the minimization can be stated as follows:

$$\rightarrow min_\phi \, \log \, \mathcal{C}_t + \int_z \hat{p}^s(z) \, \log \, \frac{\hat{p}^s(z)}{\hat{p}^t(z)} \, dz$$

$$\rightarrow min_\phi \, \log \, \mathcal{C}_t +$$
$$\int_{u_1,u_2,..u_J} \prod_{j=1}^{J} p_j^s(u_j) \, \log \, \frac{\prod_{j=1}^{J} p_j^s(u_j)}{\prod_{j=1}^{J} p_j^t(u_j)} \, du_1 du_2 ... du_J$$

$$\rightarrow min_\phi \, \log \, \mathcal{C}_t +$$
$$\sum_{j=1}^{J} \int_{u_1,u_2,..u_J} \prod_{j=1}^{J} p_j^s(u_j) \, \log \, \frac{p_j^s(u_j)}{p_j^t(u_j)} \, du_1 du_2 ... du_J$$

$$\rightarrow min_\phi \, \log \, \mathcal{C}_t + \sum_{j=1}^{J} \int_{u_j} p_j^s(u_j) \, \log \, \frac{p_j^s(u_j)}{p_j^t(u_j)} \, du_j$$

As the normalization constant $\mathcal{C}_t$ is intractable, we ignore it in our optimization:

$$\approx min_\phi \sum_{j=1}^{J} \int_{u_j} p_j^s(u_j) \, \log \, \frac{p_j^s(u_j)}{p_j^t(u_j)} \, du_j$$

$$\rightarrow min_\phi \sum_{j=1}^{J} D_{KL}(p_j^s(u_j), p_j^t(u_j))$$

6.2 How to incorporate information from multiple training subjects?

Consider the KL-divergence between the expected distribution over all training subjects and the distribution of the test subject. For simplicity of notation, let us consider only one 1D expert's distribution.

$$
\begin{aligned}
& D_{KL}\big(E_{p(s)}[p^s(u)], p^t(u)\big) \\
&= \int_u \Big(\int_s p(s) p^s(u) ds \Big) \, log \, \frac{\int_s p(s) p^s(u) ds}{p^t(u)} \, du \\
&= \int_s p(s) \Big(\int_u p^s(u) \, log \, \frac{\int_s p(s) p^s(u) ds}{p^t(u)} \, du \Big) ds \\
&= \int_s p(s) \Big(\int_u p^s(u) \, log \, \frac{\int_s p(s) p^s(u) ds}{p^t(u)} \, \frac{p^s(u)}{p^s(u)} \, du \Big) ds \\
&= \int_s p(s) \Big(\int_u p^s(u) \, log \, \frac{\int_s p(s) p^s(u) ds}{p^s(u)} \, du + \int_u p^s(u) \, log \, \frac{p^s(u)}{p^t(u)} \, du \Big) ds \\
&= \int_s p(s) \Big(\int_u p^s(u) \, log \, \frac{E_{p(s)}[p^s(u)]}{p^s(u)} \, du + \int_u p^s(u) \, log \, \frac{p^s(u)}{p^t(u)} \, du \Big) ds \\
&= - \, \mathbb{E}_{p(s)}[D_{KL}(p^s(u), E_{p(s)}[p^s(u)])] + \mathbb{E}_{p(s)}[D_{KL}(p^s(u), p^t(u))] \\
&\leq \mathbb{E}_{p(s)}[D_{KL}(p^s(u), p^t(u))]
\end{aligned}
$$

6.3 TTA-AE variants

[12] propose a autoencoder-based method for TTA. We made some minor changes in their method to get optimal results on the datasets used in our experiments. We did this analysis for 5 prostate segmentation test distributions, and used the optimal settings for the other datasets.

Architecture: In the proposed method, the adaptable module, N_ϕ is trained on the training distribution and further adapted for each test image. In contrast, [12] introduce 4 *adaptors*, A^x, A^1, A^2, A^3, as different layers in the task CNN directly at test time. A^1, A^2, A^3 are initialized to be identity mappers, while A^x is randomly initialized. In our experiments, we found that the randomly initialized A^x (with the same architecture as in [12]) substantially altered the image intensities before any TTA iterations were done. Due to this, the Dice scores at the start of TTA iterations dropped to almost 0, and could not be recovered by the TTA. We could resolve this with the help of two changes to the architecture of A^x: (i) Instead of initializing the convolutional weights with mean 0, we initialize with mean as the inverse of number input channels and variance as proposed in [63], (ii) we removed instance normalization layers from A^x. The initial Dice scores (TTA epoch 0) were now reasonable ('Architecture' in Table 6), although much lower than the strong baseline. The TTA iterations improve the results, but are unable to cross the strong baseline.

Optimization: We observed that the Dice scores fluctuated heavily across the TTA iterations. After reducing the learning rate from 0.001 (used in [12]) to 0.00001 and using the gradient accumulation strategy proposed in [11], we observed improved performance ('Optimization' in Table 6). However, the Dice scores initially improved and then dropped after about 100 epochs, for 3 of the 5 test distributions.

Loss: Plotting the evolution of the losses of the 5 AEs: one each at the input AE^x and the output layers AE^y, and 3 at different features at different depths (AE^{F1}, AE^{F2}, AE^{F3}) in the task CNN, we observed that the accuracy of AE^x and AE^y correlated well with the Dice scores, while this was untrue for the feature-level AEs. Thus, we carried out TTA driven only by AE^x and AE^y. In this setting, TTA-AE provided performance improvement in a stable manner ('Loss' in Table 6). We used this setting for the experiments on the rest of the datasets.

Method \ Test	UCL	HK	BIDMC	BMC	USZ
Domain Generalization					
Strong baseline [8]	0.77	0.82	0.62	0.78	0.77
TTA-AE [12] Variants					
Modification in:	Details				
Architecture	Removing instance normalization in A^x				
TTA Epoch 0	0.76	0.71	0.48	0.67	0.57
TTA Epoch 10	0.56	0.73	0.51	0.50	0.76
Optimization	Lower learning rate, gradient accumulation				
TTA Epoch 0	0.76	0.71	0.48	0.67	0.57
TTA Epoch 10	0.78	0.74	0.50	0.71	0.65
TTA Epoch 100	0.77	0.83	0.56	0.78	0.78
TTA Epoch 1000	0.65	0.78	0.57	0.73	0.79
Loss	Using AEs only at input & output layers				
TTA Epoch 0	0.76	0.71	0.48	0.67	0.57
TTA Epoch 10	0.78	0.74	0.48	0.71	0.64
TTA Epoch 100	0.79	0.82	0.51	0.78	0.78
TTA Epoch 1000	0.78	0.83	0.50	0.79	0.79

Table 6: Performance of TTA-AE [12] variants.