Information Geometric Approaches for Patient-Specific Test-Time Adaptation of Deep Learning Models for Semantic Segmentation

Hariharan Ravishankar, Naveen Paluru, Prasad Sudhakar,
and Phaneendra K. Yalavarthy, *Senior Member, IEEE*

***Abstract*— The test-time adaptation (TTA) of deep-learning-based semantic segmentation models, specific to individual patient data, was addressed in this study. The existing TTA methods in medical imaging are often unconstrained, require anatomical prior information or additional neural networks built during training phase, making them less practical, and prone to performance deterioration. In this study, a novel framework based on information geometric principles was proposed to achieve generic, off-the-shelf, regularized patient-specific adaptation of models during test-time. By considering the pre-trained model and the adapted models as part of statistical neuromanifolds, test-time adaptation was treated as constrained functional regularization using information geometric measures, leading to improved generalization and patient optimality. The efficacy of the proposed approach was shown on three challenging problems: a) improving generalization of state-of-the-art models for segmenting COVID-19 anomalies in Computed Tomography (CT) images b) cross-institutional brain tumor segmentation from magnetic resonance (MR) images, c) segmentation of retinal layers in Optical Coherence Tomography (OCT) images. Further, it was demonstrated that robust patient-specific adaptation can be achieved without adding significant computational burden, making it first of its kind based on information geometric principles.**

***Index Terms*— Deep learning, Information geometry, Medical imaging, Semantic segmentation, Patient-specific adaptation, Generalization.**

I. Introduction

DEEP learning models have achieved state-of-the-art (SOTA) results in medical imaging applications and resulted in the adoption of artificial intelligence (AI) models for many radiological workflows [1]. Despite their success, these models often demonstrate a considerable drop in performance after deployment in real-world applications [2]. The generalization of deep-learning models, that is, *performance on unseen data*, remains one of the biggest challenges to overcome [3]. Especially in medical imaging, this is compounded by potential "distribution-shift" on unseen data due to changes in demography, subject variability, acquisition hardware, and protocols [4]. Another potential impediment towards utilizing AI models is the lack of "patient-optimality." Despite the high average performance, deep learning models are prone to failures on "individual" cases with minor input modulations than the training data [4]. This was also consistently observed across the experiments conducted in this study. Solving this problem is critical in healthcare because this variance in performance will increase the burden on care-giving experts and reduce their trust in AI-based applications.

Sustaining performance on every patient data is a challenge for the current AI methods. AI models that provide "personalised" healthcare, tailored to individual patients, are seen as the next step in medical imaging evolution [5]. The need and potential of such patient-specific models have been identified in the National Science and Technology Council's medical imaging roadmap [6]. This study proposes ways to improve the generalization of deep-learning-based medical imaging segmentation models by personalizing them to individual patients during testing. It develops a mathematical framework based on information geometry that simultaneously addresses *generalization* and *personalization*, advancing the increased use of AI models in radiological workflows.

II. Related Work

A. Model Adaptation paradigms

Consider the class of semantic segmentation problems that map N-D medical images ($X_i \in \mathcal{X}$) into multiclass pixel-wise maps ($Y_i \in \mathcal{Y}$). Feed-forward neural networks f parameterized by $\boldsymbol{\theta}$ are utilized to obtain the mappings $f_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow \mathcal{Y}$. These parameters are learned via an empirical risk minimization (ERM) process that involves optimizing the risk $J_{\text{ERM}}(\boldsymbol{\theta})$ over the source domain data $\mathcal{D}_{\mathcal{S}} = \{\mathcal{X}_{\mathcal{S}}, \mathcal{Y}_{\mathcal{S}}\}$. Common choices for loss $J_{\text{ERM}}(\boldsymbol{\theta})$ include binary cross-entropy, Dice loss, or variants. The set of neural network weights obtained at the end of this training procedure is denoted as $\boldsymbol{\theta}_{\text{ERM}}$. The target-domain input data is denoted by a collection of N subject data

This work was supported by S. Ramachandran-National Bioscience Award for Career Development awarded by Department of Biotechnology, Govt. of India, and in part by an ARG grant# ARG01-0524-230330 from the Qatar National Research Fund (a member of the Qatar Foundation).

Hariharan Ravishankar, Naveen Paluru, and Phaneendra K. Yalavarthy are with the Department of Computational and Data sciences, Indian Institute of Science, Bangalore - 560 012 India (e-mail: hariharanrav@iisc.ac.in, naveenp@iisc.ac.in, and yalavarthy@iisc.ac.in).

Prasad Sudhakar is with the Wipro-GE HealthCare, John F. Welch Technology Centre, Odyssey Building, 122, EPIP, Phase II, Whitefield, Bangalore - 560066, India (e-mail: prasad.sudhakar@gehealthcare.com).

TABLE I: Model adaptation paradigms and problem of interest.

Type	Data required	Comments
Fine-tuning	$\{X_T, Y_T\}$	Ground-truth required
Unsupervised domain adaptation	$\{X_S, Y_S\}$, $\{X_T\}$	Source domain data required
Source free domain adaptation	$\{X_T\}$	Entire target domain data required
Patient-specific adaptation	X^*	**Only subject data required**

as $\mathcal{X}_\mathcal{T} = \{X_1^*, X_2^*, ...X_N^*\}$. In a traditional machine learning setting, inference on a new sample $X^* \in \mathcal{D}_\mathcal{T}$ is obtained using $\boldsymbol{\theta}_{\text{ERM}}$. To address the loss of performance in $\mathcal{D}_\mathcal{T}$, $\boldsymbol{\theta}_{\text{ERM}}$ is adapted to the data from $\mathcal{D}_\mathcal{T}$. Table I summarizes the various model-adaptation paradigms available in the literature.

Fine-tuning methods [7] address post-deployment performance drops by collecting ground-truth annotations $\mathcal{Y}_\mathcal{T} = \{Y_1^*, Y_2^*, \cdots, Y_N^*\}$ along with $\mathcal{X}_\mathcal{T}$ and modifying the weights of the pre-trained models using supervised loss (Table I). While fine-tuning promises the best adaptation performance, these methods are impractical because expert annotation is expensive and model adaptation cycles take a longer time.

To overcome the dependence on expert annotation for model fine-tuning, "unsupervised domain-adaptation" techniques [8] have been proposed. Owing to privacy issues, it is impractical to make source data available at every target site. This constraint has given rise to the field of "source-free domain adaptation" [9], where source data is not required. Although the dependence on source data and target domain annotation has been relaxed, these adaptation methods require access to the entirety of the target domain data (Table I). Similar to fine-tuning methods, the accumulation of a sufficient amount of target domain data leads to longer model update cycles.

In this study, the problem of generalization was addressed by the *"patient-specific adaptation"* of pre-trained models. As depicted in Fig. 1, the performance of a model gets tailored to individual test subject (X^*) and not the entire target-domain data, giving rise to *patient-specific models* $\boldsymbol{\theta}^*$. Although this setting is more challenging than full-domain adaptation [10]–[12], patient-wise adaptation is more practical and serves the goal of precision medicine and patient-optimality [5].

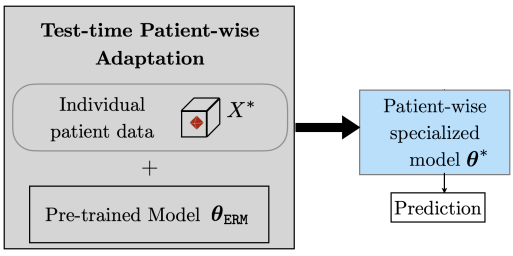


Fig. 1: Patient-specific test-time adaptation framework.

B. Test-time adaptation (TTA) methods

Test time adaptation (TTA) methods have recently demonstrated a significant impact on computer vision problems. Significant works include test-time training [13], where an auxiliary self-supervised task learned during training is replicated on test data to drive weight updates. In reference [14], a full test-time adaptation setting was proposed, where entropy minimization (Tent) was shown to effectively improve classification performance. Following Tent [14], efficient TTA without forgetting [15] and robust TTA using sharpness-aware-minimizers [16] achieved SOTA results on distribution-shifted datasets. A comprehensive survey of TTA methods was presented in reference [17]. Owing to their promise, TTA methods have recently been used for semantic segmentation in medical imaging, which fall into following three major subcategories.

1) Learning auxiliary networks during training: Autoencoders (AE-SDA) were used in [11] for patient-wise self-domain adaptation for retinal layers segmentation from OCT images. In reference [18] (TTA-DAE), the authors used a denoising auto-encoder (DAE) and adapted the input normalization sub-network during test time. The reference [12] proposed on-the-fly test-time adaptation (OF-DPG) without weight modification. A domain prior generator (DPG) is pre-trained from multiple domains (requires access to data from all domains) to generate "domain code" for the segmentation network. The downside in this class of approaches used in AE-SDA [11], TTA-DAE [18], OF-DPG [12] and their variants is that they modify the training procedure, making them unsuitable for off-the-shelf adaptation.

2) Utilizing anatomy priors: Another collection of works [9], [10] utilize invariance of anatomical priors (relative distribution of sub-anatomies in prostrate, heart, etc) across domains to drive model adaptation. In the reference [9], class-ratio priors from standard anatomical references were utilized for model adaptation for full target domain adaptation (SF-UDA). Patient-specific adaptation with the same anatomical priors along with additional shape moments (TTA-SM) has been presented in [10]. Anatomical priors are limited to anatomy segmentation and not pathology segmentation. For instance, the ratio of healthy and non-healthy regions in CT images varies with lung infection severity. Additionally, anatomical priors may differ across subjects, which can affect these methods, making these methods have limited utility in real-time.

3) Regularization: The unconstrained model adaptation on target domain patient data can lead to performance deterioration. To address this, additional constraints have been proposed to control weight perturbations. A group of methods attempt regularization on batch-normalization statistics [19], [20] using KL divergence and mean square error (MSE), respectively. The work in [21] proposed constraining batch-normalization scaling and shifting parameters (OSUDA) to be consistent between the pre-trained and adapted models. Major issues with these methods are that they attempt "weight regularization" and some of them are not patient-specific methods.

Table II presents the salient features of these studies. In short, gaps in existing methods include one of - 1) requirement of entire target domain data – not being patient-specific, 2) dependence on anatomical prior making them unsuitable for pathology/varying anatomy, 3) modifying training procedures by learning auxiliary networks, making them unsuitable for off-the-shelf adaptation, and 4) lack of regularization leading to unconstrained updates. In this study, information geometric approaches were presented to specifically address these gaps.

Information geometry enables the application of differential

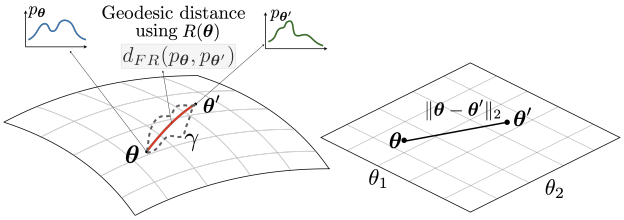


Fig. 2: Curvilinear geometry of test-time neuromanifold: functional regularization versus Euclidean weight regularization.

geometry tools for analyzing probability distributions [22]. An early application of information geometry in DL led to the popular natural gradient method [23], which used the Fisher information metric on statistical neuromanifolds parameterized by network weights [23]. Recently, information geometric approaches have been proposed for shape clustering [24], image segmentation [25], learning under noisy labels [26] and out-of-domain detection [27]. To the best of our knowledge, this work is the first attempt to apply information geometric approaches for model adaptation.

The major contributions of this study are as follows.

1) proposed a novel framework of IGTTA: based on *information geometric* **(IG)** principles to achieve patient-specific, generic, off-the-shelf, test-time adaptation **(TTA)** of semantic segmentation models in medical imaging.
2) developed a novel composite loss function for joint confidence maximization and *functional regularization* on statistical neuromanifolds using Fisher-Rao distance and other divergences to avoid performance deterioration.
3) established theoretical connections between Fisher-Rao distance and KL divergence within information-geometric framework and also study impact on performance and adaptation times.
4) demonstration of efficacy of proposed IGTTA on three challenging medical imaging segmentation problems: i) Chest CT anomaly segmentation in COVID-19 subjects (at least 4% improvement over SOTA methods across CNN and transformer-based architectures); ii) multi-site tumor segmentation in Brain MR images ($> 3-7\%$ improvement over other TTA methods across architectures), and iii) OCT retinal layers segmentation (2.8% improvement over baseline).
5) carefully curated ablation studies on design choices of the approach and comparisons with SOTA TTA methods.

TABLE II: Summary of related methods.

Method	Patient-specific	Anatomical prior reqd.	Regularization	Off-the-shelf adaptation
Tent [14]	✔	**No**	None	✔
AE-SDA [11]	✔	**No**	None	✘
TTA-DAE [18]	✔	**No**	None	✘
OF-DPG [12]	✔	**No**	None	✘
SF-UDA [9]	✘	Yes	Class Ratio	✘
TTA-SM [10]	✔	Yes	Class Ratio	✘
OSUDA [21]	✘	**No**	Weight	✔
IGTTA (Proposed)	✔	**No**	**Functional**	✔

III. Formulation and Motivation

A. Label-free surrogate objective for specialization

To obtain the patient-specific adapted model $\boldsymbol{\theta}^*$ from $\boldsymbol{\theta}_{\text{ERM}}$ for individual subject data X^*, as true segmentation map Y^* is not available during test time, a surrogate two-part composite loss function $\mathcal{G}_{\boldsymbol{\theta}}(X^*, \boldsymbol{\theta}_{\text{ERM}})$ is proposed, which requires only the pretrained model's weights and the new sample.

$$\mathcal{G}_{\boldsymbol{\theta}}(X^*, \boldsymbol{\theta}_{\text{ERM}}) = \mathcal{L}_1(X^*, \boldsymbol{\theta}) + \lambda \mathcal{L}_2(X^*, \boldsymbol{\theta}, \boldsymbol{\theta}_{\text{ERM}}). \quad (1)$$

$\mathcal{L}_1(X^*, \boldsymbol{\theta})$ is a term for improving pixel-wise confidence of adapted model's predictions, and $\mathcal{L}_2(X^*, \boldsymbol{\theta}, \boldsymbol{\theta}_{\text{ERM}})$ is a term that inhibits model collapse, ensuring divergence between pretrained model and adapted model is limited.

The intuition is to drive model weights to produce more confident voxel-level predictions while ensuring that the adapted model does not deviate from the pretrained model. The patient-specific model $\boldsymbol{\theta}^*$ can be obtained by solving,

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{G}_{\boldsymbol{\theta}}(X^*, \boldsymbol{\theta}_{\text{ERM}}). \quad (2)$$

B. Motivation for Information Geometric approaches

The major challenge in specifying $\mathcal{G}_{\boldsymbol{\theta}}(X^*, \boldsymbol{\theta}_{\text{ERM}})$ lies in the design of constraining objective $\mathcal{L}_2$. However, the constraints

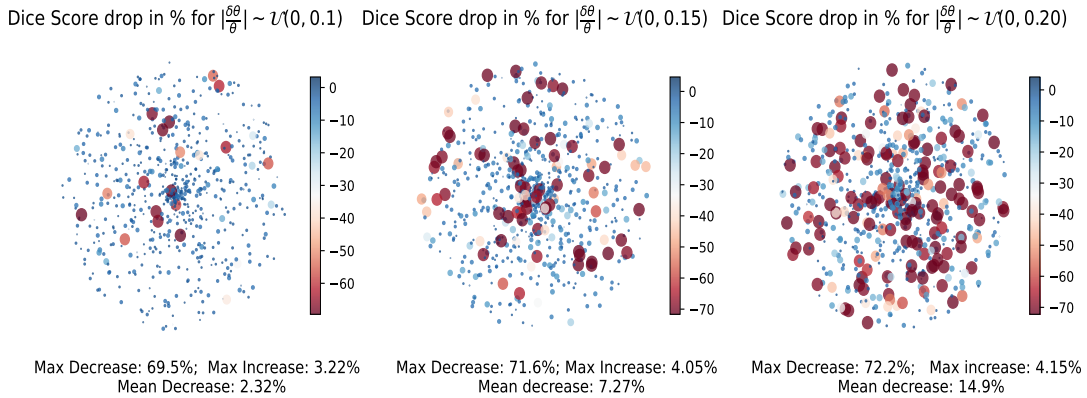


Fig. 3: Performance change from baseline in segmenting anomalies from target-domain Chest CT images for various degrees of uniformly sampled random weight perturbations of the state-of-the-art model [28]. Each simulation contained 1000 random perturbations to the learnable affine parameters of all the batch normalization layers. Color (blue to brown) and size (small to large) capture the range of performance reduction in increasing order. Negative: Dice decreases; Positive: Dice increases.

on the Euclidean geometry (Fig. 2) of weights space of the form $\|\boldsymbol{\theta}_{\text{ERM}} - \boldsymbol{\theta}^*\|_p \leq \beta$ might seem natural [19]–[21], the loss surface for target domain patient data may not be stable for even small β. To understand the robustness of the weight regularizations, a simple experiment of perturbing weights of the SOTA method [28] for segmenting anomalous regions from Chest CT images was conducted. The performance drop from the baseline for uniformly sampled random weight perturbations for various values of $\boldsymbol{\beta}$ is shown in Fig. 3. It is interesting to note that even for smaller weight perturbations of relative magnitude ≤ 0.1, there is a mean performance drop of 2.32%, with a maximum decrease of up to 69.5% dice-overlap. Indeed, there are a few perturbations that increase the performance, which are useful models, and locating them is the goal of TTA. For slightly higher perturbations sampled from $|\frac{\delta\boldsymbol{\theta}}{\boldsymbol{\theta}}| \sim \mathcal{U}(0, 0.15)$ and $|\frac{\delta\boldsymbol{\theta}}{\boldsymbol{\theta}}| \sim \mathcal{U}(0, 0.20)$, the problem is acute with more frequent failures (more brown in Fig. 3) as well larger mean decrease.

In summary, the loss-surface around $\boldsymbol{\theta}_{\text{ERM}}$ may not be flat enough for target domain subject data, potentially rendering weight regularization ineffective. Simply, *weight regularization may not guarantee functional regularization*, where the updated model can produce diverging outputs from the pre-trained model. This is observed in the experiments reported in the paper (Section IV) where the OSUDA approach [21] deteriorates the performance of many subjects.

In this study, an information geometric approach for functional regularization is proposed as the choice for $\mathcal{L}_2$. This design does not require additional priors regarding the target domain or learning auxiliary networks from the source domain, making the approach more practical, robust, and generic.

IV. Information Geometric Test-Time Adaptation

A. Geometry of test-time neuromanifold

Consider a statistical manifold $\mathcal{M}$ of the probability distributions parameterized by weights of the neural network. When these probability distributions are derived as predictions of different neural networks on test-time subject data X^* corresponding to varying weights $\boldsymbol{\theta}$, $\mathcal{M}$ is defined as the *test-time neuromanifold*. Note that the pre-trained model $\boldsymbol{\theta}_{\text{ERM}}$ and the optimal patient-specific model $\boldsymbol{\theta}^*$ are points on $\mathcal{M}$ among many other plausible neural networks.

$$\mathcal{M} \stackrel{\text{def}}{=} \{p_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(X^*) \mid \boldsymbol{\theta} = \{\theta_1, \theta_2, ...\theta_n\} \in \Theta \subset \mathcal{R}_n\}. \quad (3)$$

Here $\boldsymbol{\theta}$ is the vector of all the weights of the neural network which is made up of scalar weights θ_1, θ_2, ... θ_n. Fig. 2 depicts test-time neuromanifold for current problem of adaptation, which is a curvilinear manifold with network weights as coordinates. If $\mathcal{M}$ satisfies certain regularity constraints and is smoothly parameterized by $\boldsymbol{\theta}$ (shown in [22], [23], [26]), one can obtain the Fisher information matrix (FIM): $R(\boldsymbol{\theta})$ which can be used to define a Riemannian metric on the statistical neuromanifold. Proceeding with $R(\boldsymbol{\theta})$ as metric tensor, the Fisher-Rao distance between two neural networks on $\mathcal{M}$, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ is defined as,

$$d_{FR}(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'}) \stackrel{\text{def}}{=} \inf_{\gamma} \int_0^1 \sqrt{\frac{d\boldsymbol{\theta}(t)}{dt}^\top R(\boldsymbol{\theta}) \frac{d\boldsymbol{\theta}(t)}{dt}}\, dt. \quad (4)$$

In Eq. (4), $\gamma : [0, 1] \to \boldsymbol{\theta}$ represents a curve or a path between arbitrary pair of weights $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, such that $\gamma(0) = \boldsymbol{\theta}$ and $\gamma(1) = \boldsymbol{\theta}'$. Hence, the Fisher-Rao distance (FRD) between $p_{\boldsymbol{\theta}}$ and $p_{\boldsymbol{\theta}'}$ is the geodesic distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ using the FIM metric tensor $R(\boldsymbol{\theta})$ [22], obtained as the infimum across different γ values (Fig. 2). While the Fisher–Rao distance for arbitrary probability distributions is difficult to obtain, closed-form expressions exist for common distributions, such as the exponential family and discrete distributions [26].

Without loss of generality, let us consider mapping one pixel of X^* to one of K segmentation classes. Here, $f_{\boldsymbol{\theta}}(X^*)[j]$ corresponds to the soft-max probabilities of j^{th} pixel of interest; hence, $p_{\boldsymbol{\theta}}$ is essentially a probability simplex in $K-1$ dimensions. For such probability simplexes Δ^{K-1}, Fisher-Rao distance (FRD) between two models $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ has been derived using spherical re-parameterizations [26], [27] as

$$d_{FR}(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'}) = 2 \arccos\left(\sum_{i=1}^{K} \sqrt{f_{\boldsymbol{\theta}}(X^*)[j]_i \cdot f_{\boldsymbol{\theta}'}(X^*)[j]_i}\right), \quad (5)$$

Note that this definition of a test-time neuromanifold is unique and differs from other applications of information geometry in deep learning. In the natural gradient method [29], the neuromanifold is defined based on the distribution of predictions of training data for varying weights. In out-of-distribution detection [27], the neuromanifold corresponds to the predictions on samples from in-domain and out-of-domain data, but for a fixed pretrained model.

Relating FRD to other divergences: Recently, there have been efforts linking Fisher Rao distance (FRD) to other popular divergences. In [29], authors relate KL divergence and FRD on a neuromanifold described by softmax probabilities on two different inputs but same model. It is reiterated that neuromanifolds described in this paper are fundamentally different, where two points on the manifold are obtained via two distributions of softmax probabilities on same input X^* but two different models $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. A theorem establishing the relationship between two divergences for these manifolds is given below, modifying the result in [29].

***Theorem** 1:* The FRD between two soft-max predictions $p_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(X^*)$ and $p_{\boldsymbol{\theta}'} = f_{\boldsymbol{\theta}'}(X^*)$ given in eqn. (5) is related to KL divergence by,

$$1 - cos(d_{FR}(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'})/2) <= \frac{1}{2} d_{KL}(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'}), \quad (6)$$

Proof: Starting from Hellinger distance [29] between $p_{\boldsymbol{\theta}}$ and $p'_{\boldsymbol{\theta}}$ defined as,

$$H(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'}) \stackrel{\text{def}}{=} \sqrt{2}\left(1 - \sum_{i=1}^{K} \sqrt{f_{\boldsymbol{\theta}}(X^*)[j]_i \cdot f_{\boldsymbol{\theta}'}(X^*)[j]_i)}\right)^{1/2} \quad (7)$$

Using eqn. (5) in eqn. (7),

$$H(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'}) = \sqrt{2}\left(1 - cos\left(\frac{d_{FR}(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'}}{2}\right)\right)^{1/2} \quad (8)$$

Denoting KL divergence as $d_{KL}(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'})$, and noting that $H(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'})^2 <= d_{KL}(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}'})$, theorem 1 is readily obtained. ■

This result establishes that KL divergence is an approximation and an upper bound for exact geodesic distance given

by FRD. As noted in [29], despite their relationship, FRD and KL divergences can behave differently in optimization procedures. Empirical results and adaptation dynamics using these two divergences are discussed in Section. VII.

B. Design of $\mathcal{L}_2$ in IGTTA: Functional regularization

The aim of this work is to define divergences that ensure that patient-specific model $\boldsymbol{\theta}^*$ does not deviate much from $\boldsymbol{\theta}_{ERM}$ and utilize them in an adaptation procedure for semantic segmentation problems. Thus, the single-pixel definition of FRD in Eq. (4) is expanded to cover the entire volume of predictions, using two possible aggregation strategies.

1) Voxel Manifold Regularizer - IGTTA-VM: The first aggregation strategy is to accumulate $d_{FR}(p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}_{ERM}})$ over all voxels in X^* in the newly obtained patient data volume X^*.

$$\mathcal{L}_2(X^*, \boldsymbol{\theta}, \boldsymbol{\theta}_{\text{ERM}}) = \frac{1}{|X^*|} \sum_{j=1}^{|X^*|} d_{FR}(p_{\boldsymbol{\theta}}[j], p_{\boldsymbol{\theta}_{ERM}}[j]). \quad (9)$$

2) Class-ratio manifold Regularizer - IGTTA-CRM: In this aggregation strategy, the class ratio estimates for K classes are computed per slice and aggregated over all slices in the volume. Denoting class ratio estimate of s^{th} slice as $CR[s]$ and for total number of S slices, $\mathcal{L}_2$ is defined as,

$$\mathcal{L}_2(X^*, \boldsymbol{\theta}, \boldsymbol{\theta}_{\text{ERM}}) = \frac{1}{S} \sum_{s=1}^{S} d_{FR}(CR_{\boldsymbol{\theta}}[s], CR_{\boldsymbol{\theta}_{ERM}}[s]) \quad (10)$$

Note that $CR[s]$ is again a Δ^{K-1}-simplex. Therefore, the definition of Eq. (4) holds. While the Fisher–Rao distance is a divergence measure, other common divergences, such as KL divergences, are also explored and compared in experiments.

Intuitively, the voxel manifold regularizer constrains the per-pixel functional divergence of $\boldsymbol{\theta}^*$ from $\boldsymbol{\theta}_{ERM}$, whereas the class-ratio manifold regularizer ensures that the distribution of the class types per slice does not diverge. These two manifolds represent examples of the proposed approach. More advanced regularizers can be derived by using the same framework to impose spatial constraints.

C. Design of $\mathcal{L}_1$: Confidence Maximization

Recently, several studies established empirical correlations between classification accuracy and confidence in predictions [10], [14], [21]. Inspired by these, this study proposes minimizing the average pixel-wise uncertainty during the specialization procedure as a choice for $\mathcal{L}_1$. The average pixel-wise prediction confidence $\mathcal{L}_1$ is defined as the average of ℓ_{conf} over all pixels of X^*.

$$\mathcal{L}_1(X^*, \boldsymbol{\theta}) = \frac{1}{|X^*|} \sum_{j=1}^{|X^*|} \ell_{conf}(f_{\boldsymbol{\theta}}(X^*)[j]), \quad (11)$$

Table III lists the choices of (ℓ_{conf}). In addition, c denotes the winning class $c = \arg\max \{f_{\boldsymbol{\theta}}(X^*)[j]\}$.

- Soft-max entropy (ℓ_{ent}) [14] - measures uncertainty in the prediction vector. Entropy will be zero if the prediction for the winning class c is 1. If the predictions for all classes are equal, the entropy will be maximum.
- Generalized cross entropy (ℓ_{gce}) [30] - another way of measuring uncertainty. As $q \to 0$, ℓ_{gce} approaches ℓ_{ent} and as $q \to 1$, it approaches mean absolute error loss. This study is the first attempt to utilize ℓ_{gce} in medical imaging applications.
- Margin loss (ℓ_{marg}) - Proposed in this study is a simple loss function that measures difference between winning class prediction value and next best prediction.

TABLE III: List of objective functions that measure uncertainty or confidence in a vector of k-way predictions.

Confidence Loss Functions : ℓ_{conf}	**Formulae**
Soft-max entropy [14]: ℓ_{ent}	$-\sum_{r=0}^{k-1} p^r(j,\boldsymbol{\theta}) \log(p^r(j,\boldsymbol{\theta}))$
Generalized cross-entropy [30]: ℓ_{gce}	$q^{-1}(1-p^c(j,\boldsymbol{\theta})),\ q \in [0,1]$
Margin (proposed): ℓ_{marg}	$-(p^c(j,\boldsymbol{\theta}) - \max_{r \neq c} p^r(j,\boldsymbol{\theta}))$

Recently, there have been works that explored margin-loss based training paradigms for improved calibration and generalization [31], [32]. These works are inherently training-time methodologies where the aim is to achieve a balance between performance and calibration. In this work, margin-based confidence maximization is proposed for test-time adaptation, which has not been explored in prior works in general or for medical image segmentation.

D. Algorithmic procedure for IGTTA

By using the definitions of $\mathcal{L}_1$ and $\mathcal{L}_2$, the specification of the label-free surrogate loss $\mathcal{G}_{\boldsymbol{\theta}}(.)$ in Eq. (1) is complete. To obtain $\boldsymbol{\theta}^*$ from $\boldsymbol{\theta}_{ERM}$ customized for X^*, an iterative optimization procedure was proposed, similar to the standard learning procedures for training neural networks. The gradients computed using $\mathcal{G}_{\boldsymbol{\theta}}(X^*, \boldsymbol{\theta}_{\text{ERM}})$ are backpropagated to update the model weights for a predetermined number of steps (typically 10, with $\lambda = 1$ in Eq. (1)). To establish architecture-agnostic effectiveness of proposed framework, we demonstrate results using SOTA medical imaging segmentation architectures that are a) CNN-based : AnamNet [28], residual U-Net [11], etc and b) vision-transformer based: MissFormer [33]. In all the adaptations performed in this work starting from CNN-based architectures, learnable affine parameters of BatchNorm layers (shift and scale) were updated along with the usage of subject-level statistics (mean and standard deviation). This is similar to the results of references [10], [14], [21], where updating these limited the number of parameters has been demonstrated to handle co-variate shift. In experiments involving adaptations of vision transformer based architecture: MissFormer [33], we update the affine parameters of LayerNorm layers.

V. Example Problems and Data-sets

In this study, three segmentation problems were investigated to show the generic applicability of proposed IGTTA to multiple modalities (CT, MR, and OCT), showing utility for both anatomy and pathology segmentation, and robustness to

the number of classes (3 to 11). Furthermore, they cover the common domain adaptation scenarios in medical imaging because of 1) covariate shift across multiple sites and demography, 2) Anatomy/Pathology sub-type differences, and 3) variability due to scanner-types.

A. Segmentation of anomalies in COVID-19 CT images

1) Problem set-up and data-sets: The first problem is segmenting Chest CT images obtained from COVID-19 patients into three classes: background, healthy lung regions, and anomalies. Common anomalies observed in chest CT scans of COVID-19 patients include ground-glass opacities (GGO), consolidation, and pleural effusion [28].

The axial chest CT images utilized in this work were obtained from two publicly available open-source datasets:1) Dataset I - Italian Society of Medical and Interventional Radiology [34] and 2) Dataset II - radiopedia [35]. Dataset I consists of 100 slices from approximately 40 patient CT scans, and Dataset II contains 829 slices obtained from 9 subjects.

2) Baseline Models: To study the benefits of the proposed adaptation scheme, models built using Dataset I were evaluated on Dataset II. The light-weight CNN architecture called "AnamNet" proposed in [28] has reported state-of-the-art (SOTA) results for this problem, while training on 270 augmented images from Dataset I and testing on 704 slices from Dataset II. Firstly, 90 slices with significant lung regions were selected out of 100 training slices and then augmentations included horizontal and vertical flip leading to a total of 270 slices. In this study, the same experimental setup was used. In addition to AnamNet, seven other architectures reported in [28] were also utilized.

B. Multi-Site Brain Tumor Segmentation on MR images

1) Problem set-up and data-sets: The Federated Tumor Segmentation challenge (FETS) [36] involves cross-institutional, multi-contrast preoperative MRI scans of the brain, containing heterogeneous tumors, namely gliomas. Aggregated from 23 different sites, FETS data contained a total of 1251 patient scans, with each scan containing four contrasts: T1-weighted (T1), T1-contrast enhanced (T1ce), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR). The goal is to map voxels into four classes: enhancing tumor (ET), peritumoral edema (ED), necrotic and non-enhancing tumor core (CoreT), and background. According to the FETS challenge, these 1251 scans were aggregated from sites with varying demography, device, and disease stage, making it a suitable large-scale database for experimentation, besides the clinically relevant goal of segmenting gliomas. To understand the efficacy of IGTTA for cross-institutional domain adaptation, patient-specific adaptation of models built on Site 1 as the source domain (511 subjects) was attempted using individual patient scans from the remaining 22 sites (740 subjects).

2) Baseline Models: The two-dimensional (2D) segmentation model was built using the AnamNet [28] architecture utilizing all four contrasts and 90 middle slices from the 3-D MRI of the training subjects. The models were built for 100 epochs with batch size $= 24$ and Adam optimizer (learning rate $= 0.01$). The results were compared for semantically meaningful tumor classes: Whole Tumor (WT): ED + EnhT + CoreT, Enhancing tumor (ET), and Tumor Core (TC): ED + CoreT, as proposed in [36].

C. Segmentation of retinal layers from OCT images

1) Problem set-up and data-sets: Finally, vendor-agnostic segmentation of retinal layers from 2D OCT images is considered. The 11 retinal layers under consideration and their acronyms [11] are as follows: 1) vitreous background, 2) retinal nerve fiber layer (RNFL), 3) ganglion cell/inner plexiform layer (GCIP), 4) Inner nuclear layer (INL), 5) Outer plexiform layer (OPL), 6) Outer nuclear layer (ONL), 7) Inner segment (IS), 8) Outer segment (OS), 9) Retinal pigment epithelium (RPE), 10) Choroid background 11) other. Two publicly available OCT datasets from two different scanners 1) Heidelberg Spectralis scanner [37] and 2) Cirrus scanner [11], were analyzed. The Spectralis dataset consists of 35 3-D subject volumes with 49 slices (2D) of size 496×1024 per volume, whereas the Cirrus dataset consists of six subjects with eight slices of size 1024×512. The pre-processed and standardized images along with manual annotations were made available in the AE-SDA [11].

2) Baseline Models: The experimental set-up provided in AE-SDA [11] was reproduced from their publicly available code-base. The 2D segmentation backbone using residual U-Net [38] was built on subject data from Spectralis data-set and evaluated on 6 subject volumes from Cirrus data-sets (8 slices per volume). Multiple data augmentation strategies were utilized during training - including gamma adjustment, flips, Gaussian noise addition, etc [38]. The models were trained for 20 epochs, with batch size $= 2$ the Adam optimizer (learning rate $= 0.01$).

VI. Experiments and Results

A. SOTA methods comparison and implementation details

The proposed IGTTA method was compared with following six test-time adaptation methods, as discussed in Section II.

- Classic TTA method: a) Tent [14].
- TTA with weight regularizer: OSUDA [21]. Tent (PS) and OSUDA (PS) adaptations are patient-specific to be consistent with the proposed IGTTA.
- TTA with auxiliary networks a) TTA-DAE [18] : with auxiliary denoising auto-encoders b) AE-SDA: self-domain adaptation with auto-encoders [11].
- TTA using anatomical priors a) SFDA: source-free domain adaptation [9] b) TTA-SM: patient-specific adaptation with shape moments [10].

For all the three problems, patient-specific models $\{\boldsymbol{\theta}_i^*\}$ for individual subject data X_i^* were built using the proposed IGTTA method and SOTA methods. To ensure fair comparison, same baseline models $\boldsymbol{\theta}_{ERM}$ were used as starting weights of adaptation by respective methods. These base models are explicitly stated in title of respective quantitative comparison tables. For each of the SOTA method, best set

of adaptation parameters from respective works were used. For Tent [14], number of adaptation steps were chosen to be 10. For OSUDA [21], SFDA [9] and TTA-SM [10], number of adaptation steps were set to 200 following the respective works, whereas for TTA-DAE [18] it was set to 150.

Implementation details: The design choices of proposed approach in IGTTA include: 1) choice of neuromanifold for $\mathcal{L}_2$: voxel versus class-ratio manifold, and 2) choice of divergence function (Fisher-Rao distance versus KL-divergence) 3) number of adaptation steps. Variants of proposed approach based on combinations of neuromanifold and divergence functions are referred to as 1) IGTTA-VM$_{FRD}$: FRD on voxel manifold 2) IGTTA-VM$_{KL}$: KL divergence on voxel manifold 3) IGTTA-CRM$_{FRD}$: FRD on class-ratio manifold 4) IGTTA-CRM$_{KL}$: KL divergence on class-ratio manifold. Across all experiments, we have chosen adaptation steps as 10 for voxel-manifold based variants of proposed method (IGTTA-VM$_{FRD}$ and IGTTA-VM$_{KL}$). This is to have lesser adaptation times and compare with Tent which uses 10 as the number of adaptation steps in their experiments. For variants of proposed approach based on class-ratio manifold (IGTTA-CRM$_{FRD}$ and IGTTA-CRM$_{KL}$), we have chosen number of adaptation steps as 50 to still have 4 times lesser adaptation steps compared to other methods.

For each of the three problems, best results of proposed approach among the 4 IGTTA variants are shared. Detailed discussion on impact of choice of neuromanifold, choice of divergence function, comparisons of results across 4 variants,

TABLE IV: COVID-19 anomaly segmentation from Chest CT images: Mean and standard deviations of Dice score and ASSD for normal and abnormal lung regions, evaluated on 704 images from target-domain. Reported results are obtained from 3 independent runs.

(a) Comparisons of the proposed IGTTA with other test-time adaptation methods in literature with base model as AnamNet [28].

TTA Method	Dice Score ↑		% ↑ in dice of abnorm. class	ASSD ↓		↓ in ASSD of abnorm. class
	Normal	Abnormal		Normal	Abnormal	
AnamNet [28] (Unadapted)	0.968 ± 0.005	0.730 ± 0.025	NA	1.104 ± 0.217	13.432 ± 2.458	NA
Tent(PS) [14]	0.976 ± 0.001	0.694 ± 0.031	-3.6	0.931 ± 0.204	21.107 ± 1.12	-7.675
OSUDA(PS) [21]	0.977 ± 0.001	0.638 ± 0.033	-9.2	1.167 ± 0.204	19.611 ± 1.12	-6.179
SFDA [39]	0.951 ± 0.019	0.608 ± 0.072	-12.2	2.031 ± 0.964	25.337 ± 12.01	-11.905
TTA-SM [10]	0.950 ± 0.011	0.627 ± 0.067	-10.3	2.576 ± 0.628	33.753 ± 9.105	-20.321
TTA-DAE [18]	0.973 ± 0.006	0.692 ± 0.007	-3.8	1.508 ± 0.232	19.976 ± 2.067	-6.544
IGTTA-VM$_{FRD}$ (Proposed)	**0.978 ± 0.007**	**0.758 ± 0.038**	**2.8**	**0.833 ± 0.313**	**8.512 ± 3.585**	**4.92**

(b) Comparisons of the proposed IGTTA with other test-time adaptation methods in literature with base Model as MissFormer [33].

TTA Method	Dice Score ↑		% ↑ in dice of abnorm. class	ASSD ↓		↓ in ASSD of abnorm. class
	Normal	Abnormal		Normal	Abnormal	
MissFormer [33] (Unadapted)	0.930 ± 0.002	0.473 ± 0.006	NA	2.795 ± 0.368	39.683 ± 2.813	NA
Tent(PS) [14]	0.953 ± 0.013	0.532 ± 0.065	5.8	1.298 ± 0.289	32.319 ± 0.804	7.365
OSUDA(PS) [21]	0.954 ± 0.013	0.534 ± 0.065	5.9	1.301 ± 0.287	32.313 ± 0.802	7.371
SFDA [39]	0.951 ± 0.006	0.564 ± 0.012	9.1	1.259 ± 0.026	31.324 ± 0.415	8.359
TTA-SM [10]	0.945 ± 0.003	0.297 ± 0.004	-17.6	1.216 ± 0.121	34.799 ± 0.700	4.885
TTA-DAE [18]	0.953 ± 0.011	0.408 ± 0.108	-6.6	1.222 ± 0.218	33.198 ± 0.288	6.486
IGTTA-VM$_{FRD}$ (Proposed)	**0.956 ± 0.013**	**0.575 ± 0.002**	**10.2**	**1.162 ± 0.026**	**30.857 ± 0.274**	**8.826**

(c) Effect of the proposed IGTTA on 8 different architectures compared in [28]. In each row, bold signifies the best results obtained before and after the proposed IGTTA-based adaptation. Best results are highlighted using box.

Architecture	Base model - Dice Score ↑		IGTTA-VM$_{FRD}$ - Dice Score ↑		Base model - ASSD ↓		IGTTA-VM$_{FRD}$ - ASSD ↓	
	Normal	Abnormal	Normal	AbNormal	Normal	Abnormal	Normal	AbNormal
UNet	0.949 ± 0.006	0.636 ± 0.025	**0.955 ± 0.005**	**0.663 ± 0.024**	1.529 ± 0.156	21.135± 2.277	**1.385 ±0.076**	**18.308 ±1.903**
ENet	0.95 ± 0.005	0.675 ± 0.016	**0.953 ± 0.002**	**0.688 ± 0.019**	1.609 ±0.377	19.735 ±5.094	**1.339 ±0.082**	**17.141 ±2.83**
UNet++	0.957 ± 0.006	0.687 ± 0.032	**0.962 ± 0.005**	**0.71 ± 0.025**	1.177 ±0.121	15.688 ±2.747	**1.007 ±0.067**	**12.37± 1.619**
SegNet	0.943 ± 0.008	0.609 ± 0.022	**0.953 ± 0.001**	**0.658 ± 0.005**	1.976 ±0.101	27.624 ±1.026	**1.912 ±0.236**	**21.334 ±3.816**
AttUNet	0.952 ± 0.08	0.654 ± 0.041	**0.955 ± 0.008**	**0.669 ± 0.041**	1.229± 0.148	16.166± 3.278	**1.139 ±0.147**	**13.961± 2.703**
LEDNet	0.93 ± 0.007	0.615 ± 0.032	**0.932 ± 0.007**	**0.63 ± 0.035**	2.994 ± 1.877	22.3 ± 1.194	**2.994± 0.351**	**21.994 ±1.745**
MissFormer [33]	0.930 ± 0.002	0.473 ± 0.0006	**0.956 ± 0.013**	**0.575 ± 0.002**	2.795± 0.368	39.683 ±2.813	**1.162± 0.026**	**30.857 ±0.174**
AnamNet [28]	0.968 ± 0.005	0.73 ± 0.025	**0.978 ± 0.007**	**0.758 ± 0.038**	1.104 ± 0.217	13.432 ± 2.458	**0.833 ±0.313**	**8.512± 3.585**

adaptation dynamics are presented in ablation studies (Section. VII). Additionally, a prescribed recipe to choose for new set of problems is presented. As stated earlier, BatchNorm parameters were adapted for CNN based arhcitectures and LayerNorm parameters were adapted for MissFormer models. These parameters were adapted using Adam optimizer with learning rate of 5e-4 and initial decay rates for first & second moments of gradient set to 0.99.

Comparison Metrics: The methods are compared quantitatively using variety of metrics starting with Dice Similarity score of adapted model's prediction and ground truth segmentation. Comparisons are also reported for surface distance-based metrics 1) ASSD - Average symmetric surface distance 2) HD - Hausdorff distance. It should be noted that in MR brain tumor segmentation problem, Tumor core (TC) and enhancing tumor (ET) regions are often smaller, have inconsistent boundaries and hence calculating surface distance based metrics can be erroneous and misleading. Hence, Sensitivity as an additional metric is compared for these classes, since we are interested in True Positives (TP) performance of methods for successive evaluation by clinical experts. It is pointed that dice score is nothing but f1-score, hence all four parameters - TP, TN, FP, FN are covered by these set of metrics. One of the biggest risks of adaptation, is deterioration of adapted model compared to base model's performance, which indicates failure of adaptation procedure. To capture these events, we also report number of patient-wise adaptation failures that each of the methods attain across problems. It is noted that for every experiment, we report mean and standard deviations of results obtained from 3 independent runs.

B. Summary of Results

1. IGTTA outperforms other test-time adaptation methods:

a) Segmentation of anomalies in COVID-19 CT images: The proposed method achieved SOTA performance on all metrics of interest for both CNN-based and transformer-based segmentation models. Quantitative Comparisons with the other methods are presented in Table. IVa and Table. IVb with AnamNet and MissFormer as base models respectively.

Starting with AnamNet based pre-trained models, all the comparison methods lead to deterioration in performance (Table. IVa). Reasons for this deterioration include 1) high variability in the degree of infection: methods which depend on pre-determined class-ratios from training distribution like SFDA [9] and TTA-SM [10] suffer when fraction of anomalous region changes across subjects. 2) Variation in the imaging properties of the nine subjects - TTA-DAE [18] which learns auxiliary denoising auto-encoder on training distribution suffers with shift in imaging distribution, thereby leading to reduction in post-adaptation performance. 3) insufficient regularization: Tent [14] suffers from over-eager adaptation clearly highlighting the need for stronger regularization. Though OSUDA [21] does have weight regularization, it is clearly insufficient in individual patient-wise adaptation setting, where full test-domain data is unavailable. The benefits of functional regularization is clearly visible in proposed method of IGTTA, which improves the SOTA baseline model of AnamNet [28] by a significant margin of 2.8% in Dice score of abnormal

TABLE V: Multi-Site Brain Tumor Segmentation on MR images: Mean and standard deviations of various evaluation metrics for WT, ET, TC classes evaluated on of 740 target domain subjects from 22 target sites. Reported results are obtained from 3 independent runs.

(a) Comparisons of the proposed IGTTA with other test-time adaptation methods in literature with base model as AnamNet [28].

TTA Method	Dice Score ↑			# Failures on WT	Sensitivity ↑		HD ↓
	Enhancing Tumor (ET)	Tumor Core (TC)	Whole Tumor (WT)		Enhancing Tumor (ET)	Tumor Core (TC)	Whole Tumor (WT)
AnamNet [28] (Unadapted)	0.339 ± 0.015	0.353 ± 0.010	0.357 ± 0.087	NA	0.571 ± 0.014	0.516 ± 0.005	23.866 ± 10.561
Tent(PS) [14]	0.458 ± 0.098	0.448 ± 0.090	0.496 ± 0.115	267	0.663 ± 0.140	0.620 ± 0.131	12.384 ± 0.347
OSUDA(PS) [21]	0.526 ± 0.053	0.512 ± 0.053	0.600 ± 0.066	174	0.752 ± 0.095	0.697 ± 0.085	11.539 ± 0.592
SFDA [39]	0.505 ± 0.023	0.501 ± 0.035	0.595 ± 0.046	140	0.718 ± 0.073	0.661 ± 0.104	11.938 ± 0.451
TTA-SM [10]	0.635 ± 0.018	0.632 ± 0.020	0.717 ± 0.010	73	0.810 ± 0.018	0.704 ± 0.023	12.101 ± 0.075
TTA-DAE [18]	0.621 ± 0.015	0.620 ± 0.010	0.701 ± 0.002	118	0.770 ± 0.016	0.686 ± 0.039	12.235 ± 0.098
IGTTA-CRM$_{KL}$ (Proposed)	**0.644 ± 0.014**	**0.642 ± 0.022**	**0.754 ± 0.009**	**26**	**0.830 ± 0.007**	**0.719 ± 0.002**	**10.1 ± 2.436**

(b) Comparisons of the proposed IGTTA with other test-time adaptation methods in literature with base model as MissFormer [33].

TTA Method	Dice Score ↑			# Failures on WT	Sensitivity ↑		HD ↓
	Enhancing Tumor (ET)	Tumor Core (TC)	Whole Tumor (WT)		Enhancing Tumor (ET)	Tumor Core (TC)	Whole Tumor (WT)
MissFormer [33] (Unadapted)	0.586 ± 0.009	0.581 ± 0.013	0.757 ± 0.017	NA	0.751 ± 0.014	0.658 ± 0.012	14.375 ± 1.958
Tent(PS) [14]	0.427 ± 0.024	0.354 ± 0.020	0.457 ± 0.023	168	0.631 ± 0.002	0.593 ± 0.013	13.435 ± 1.029
OSUDA(PS) [21]	0.429 ± 0.021	0.358 ± 0.018	0.459 ± 0.022	167	0.64 ± 0.011	0.601 ± 0.022	13.419 ± 1.095
SFDA [39]	0.535 ± 0.011	0.529 ± 0.01	0.706 ± 0.008	151	0.719 ± 0.011	0.629 ± 0.005	15.797 ± 0.606
TTA-SM [10]	0.540 ± 0.002	0.530 ± 0.001	0.679 ± 0.005	116	0.767 ± 0.016	0.673 ± 0.009	13.965 ± 0.198
TTA-DAE [18]	0.561 ± 0.009	0.557 ± 0.031	0.710 ± 0.006	103	0.793 ± 0.020	0.699 ± 0.031	12.478 ± 0.243
IGTTA-CRM$_{KL}$ (Proposed)	**0.605 ± 0.016**	**0.613 ± 0.003**	**0.783 ± 0.021**	**52**	**0.818 ± 0.013**	**0.737 ± 0.004**	**9.261 ± 0.434**

class. Additionally, it achieves lowest ASDD metrics amongst all the metrics for both the classes.

Similarly, improved performance of the proposed method is observed when using MissFormer [33] model as baseline (Table. IVb). Firstly, baseline performance of unadpated model in itself is lower compared to AnamNet [28] due to limited training data and imbalance amongst segmentation classes. This behaviour is also observed with other larger models in Table IVc. Hence majority of adaptation methods lead to improvement in performance over the baseline, except TTA-SM [10] and TTA-DAE [18] for similar reasons stated above. Tent [14], OSUDA [40] SFDA [9] achieve improved performance compared to baseline, but proposed approach of IGTAA outperform all of them by a significant margin. IGTTA achieves 10.2% increase in dice score, while decreasing the ASDD metric by almost 25% for the abnormal class compared to the baseline.

b) Multi-Site MR Brain Tumor Segmentation: Results are shared in Table. Va and Table. Vb for AnamNet and MissFormer base models respectively. It is apparent that the proposed IGTTA method significantly outperforms other TTA methods on the task of adapting models built on site 1 data (511 subjects) to other 22 target domain sites (740 subjects) on all metrics of interest: 1) Dice Score 2) Hausdorff distance (HD) on whole tumor 3) Sensitivity on other two classes 4) Total number of failures post adaptation for both base-models. In Table. Va, Compared to the unadapted AnamNet model, IGTTA improves the dice score for all subtypes of tumor tissues (WT, ET, TC) by atleast 29%.

Similarly, improved performance of the proposed method is observed when using MissFormer [33] model as well in Table. Vb. The proposed method again achieves best performance on all metrics of interest across all tumor tissue types. It is pointed that, MissFormer model has a higher baseline performance compared to AnamNet, as this problem has higher training data, which can be exploited by a larger model. However, this poses a challenge for all the comparison methods as they have more parameters to adapt during test-time which can lead to reduced performance without careful adaptation, This is indeed the cases, where all the methods starting from MissFormer model as baseline, suffer from performance deterioration owing to reasons described earlier. Another thing to note is that, irrespective of base model's performance, IGTTA achieves similar post-adaptation performance in Tables. Va and Vb. This is a highly encouraging result, which showcases power of IGTTA to bridge the gap between base models of varying performances.

The site-wise adaptation performance is shown in Fig. 4 for one run of AnamNet, where IGTTA improves performance on all the target sites. It should be noted that both unregularized TTA (Tent [14]) and weight-regularized TTA (OSUDA [21]) deteriorate the performance on site 13.

c) **Retinal layers segmentation on OCT images**: As shown in Fig. 6(a), compared to the no-adaptation case, IGTTA achieves an overall Dice increase of 2.8% on the Cirrus dataset with models trained on the Spectralis dataset). Improvements over the AE-SDA method [11] are 0.8%. All methods based on entropy minimization (IGTTA, Tent, and OSUDA) perform satisfactorily because of the higher number of classes (11). The proposed approach achieved marginally higher class-wise performance for seven out of eight types of retinal layers. A slight drop in performance was observed for the RPE labels.

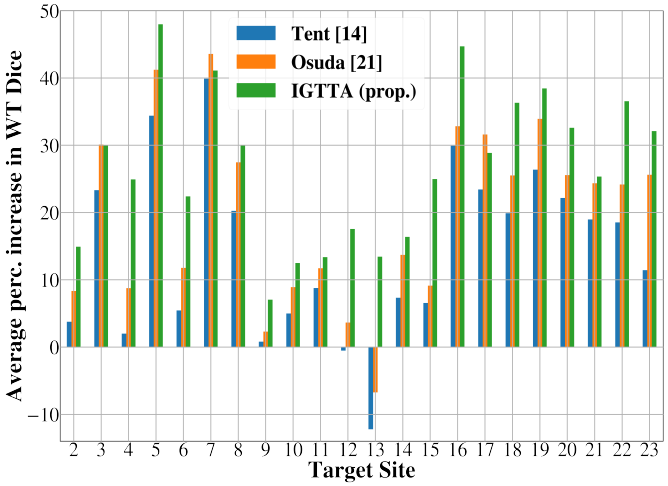


Fig. 4: Multi-site brain tumor segmentation from MR Images: Percentage increase in WT Dice across different target Sites compared for proposed IGTTA, Tent [14] and OSUDA [21], for one run of AnamNet model trained on Site 1 and tested on 22 other sites as target for a total of 740 subjects.

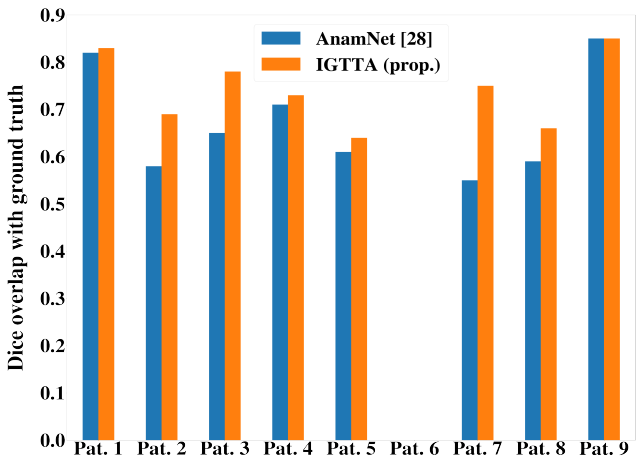
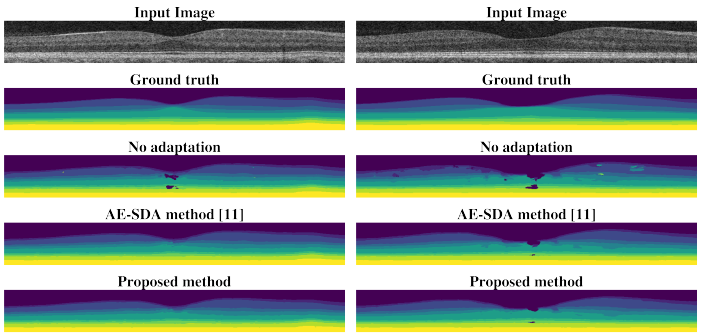


Fig. 5: Chest CT anomaly segmentation: Patient-wise comparisons of Dice overlap for abnormal classes between the proposed approach and AnamNet [28] on the test dataset.

2. IGTTA achieves patient optimality: Patient optimality refers to successful adaptation of a neural network that has been trained on population data, to patient's data during test-time. This can be demonstrated by showing how proposed method improves on every individual patient post-adaptation. **a) For the chest CT segmentation problem**, IGTTA consistently improves anomaly segmentation across all nine subjects across all runs and for all architectures. Fig. 5 shows that for one run of adaptation with AnamNet as basemodel, IGTTA obtains mean Dice score increase of 6.34%, with maximum increase of 19.58% for Patient 7, followed by 13.86% and 11.21% increase on Patient 3 and Patient 2, respectively. Interestingly, the proposed approach improves on patients,

Retinal Layers	Target: Cirrus scanner data; Source: Spectralis scanner data				
	No adapt	Tent [14]	OSUDA [21]	AE-SDA method [11]	IGTTA-CRM$_{KL}$ Method
RNFL	0.754	0.796	0.796	0.780	**0.798**
GCIP	0.817	0.866	0.866	0.855	**0.867**
INL	0.752	0.792	0.792	0.780	**0.794**
OPL	0.644	0.679	0.679	0.672	**0.683**
ONL	0.874	0.895	0.895	0.888	**0.897**
IS	0.860	0.875	0.875	0.87	**0.875**
OS	0.873	0.891	0.891	0.889	**0.892**
RPE	**0.842**	0.838	0.838	**0.842**	0.837
Overall	0.802	0.829	0.829	0.822	**0.830**

(a) Dice score comparison.



(b) Sample OCT Image 1 results. (c) Sample OCT Image 2 results.

Fig. 6: Retinal layers segmentation from OCT Images: Quantitative comparisons of the proposed IGTTA method with other TTA methods. The segmentation of two sample Cirrus data slices was presented with no adaptation and AE-SDA [11] methods.

where performance is low, and maintains the same level in cases where it is already high. Patient 6 had completely healthy lungs. The compared TTA methods led to a deterioration in performance across subjects, with Tent being a strong competitor with two failures out of nine subjects data.

Another way to measure patient-optimality is to ensure model does not deteriorate in performance post-adaptation, otherwise considered as adaptation failure.

b) In brain tumor segmentation problem (Table. Va and Vb), all the comparison methods lead to higher number of adaptation failures. For instance, Tent [14] achieves lesser WT dice score post-adaptation on 267/740 cases (for AnamNet base model) and 168/740 (for MissFormer base model), capturing the issue of unconstrained adaptation. All the other methods, also have very high number of failures, owing to reasons like domain-shift, lack of sufficient regularization, varying distribution of classes, etc. The proposed method of IGTTA has least number of failures, highlighting the power of functional regularization, which ensures adapted model's prediction are constrained from diverging from base model. Note that this reduced number of failures is coupled with improvement in the overall Dice scores of tumor sub-tissues.

3. IGTTA achieves architectural-agnostic generalization improvement: As demonstrated earlier, on both CT and MR segmentation problems, the proposed method of IGTTA achieves highest adaptation performance using both AnamNet and MissFormer based architectures. These two architectures cover a reasonable spectrum of architectures observed in medical imaging problems, with AnamNet being light-weight, edge-device suitable, easy to train, suitable for less data (CT problem). MissFormer [33] being a vision-transformer based architecture has advantages of more parameters, richness of representation, higher expected performance, better suited for large data problems (MR problem). For the retinal layer segmentation with OCT, we used residual U-Net archiecture as suggested in AE-SDA [11] method which is SOTA for this problem, to allow fair comparison with other methods. To further demonstrate the architectural-agnostic genaralization of proposed IGTTA, we used 7 other popular segmentation architectures and share the adaptation results in Table. IVc segmentation in Chest CT images. It is clear that proposed method, improves on every segmentation architecture. This is a significant result showing the universal utility of IGTTA to improve any off-the-shelf pre-trained model.

4. IGTTA has competitive adaptation times: It is crucial that patient-wise adaptation techniques have acceptable space and time complexities. As noted earlier, IGTTA-VM$_{FRD}$ and IGTTA-VM$_{KL}$ are run for 10 adaptation steps and hence have similar inference times to Tent [14]. The other two variants require four times less adaptation steps, compared to SFDA [9], TTA-SM [10] and OSUDA [40]. Additionally, the proposed method is significantly faster than methods using auxiliary networks which include TTA-DAE [18] and AE-SDA [11]. In OCT segmentation problem, for every patient data in Cirrus datasets (eight slices per volume), IGTTA takes ≈ 1.5 s compared to 11.43 s for AE-SDA, giving an advantage of 7 times faster adaptation. All experiments were conducted using PyTorch libraries on an NVIDIA Quadro RTX A6000 GPU.

5. Qualitative Comparisons: a) Figure 7 shows the improvement due to proposed IGTTA across varying levels of lung size and infections on selected slices from four different test subjects. On Slice 23 (first row), Tent [14] erodes a part of the anomaly in the left lung. IGTTA reduced only the false positives at the top of the right lung. On Slice 316 (second row), the proposed method completely erases all false positives for anomalies, which is critical as this subject has fully healthy lungs, while Tent magnifies them more. On Slice 198 (third row), the proposed method has improved on false positives and false negatives - by eroding false findings in both lungs, while also correctly expanding the circular anomaly region in the right lung against the slight under-segmentation achieved by AnamNet. Slice 674 (last row) captures comparisons of lung slices affected by severe infections.

b) Figure 8 depicts the results on selected slices from varying target sites and subjects for brain tumor segmentation. In the first row, the unadapted source domain model assigns all tumor sub-tissues to the edema class. IGTTA faithfully recovered all three tissues, whereas Tent and OSUDA failed to enhance the tumor region. In the second row, perils of unregularized and weight-regularized adaptation are shown, where both Tent and OSUDA erase the predictions completely, while proposed IGTTA achieves satisfactory performance on all three classes. Similarly, the third row shows the improvement obtained with proposed IGTTA over both the unadapted

and other TTA methods.

Figure 6(b) and 6(c) shows comparison of retinal layer segmentation for baseline, proposed approach and AE-SDA method [11] on two selected slices from Cirrus data-set. In Fig. 6(b), the base-model produces segmentations with holes, loss of continuity and label mismatches compared to ground-truth. Both the proposed approach and AE-SDA method [11] show improvements over the no-adaptation output. Similarly, in a sample slice from the Cirrus dataset (Fig. 6(c)), the base model demonstrates deteriorated segmentation performance. While the proposed approach filled the holes and improved consistency across layers, the AE-based method improved small misclassified patches, which is also reflected in the quantitative comparisons (table of Fig. 6(a)).

In summary, the presented work evaluates and establishes the efficacy of the proposed framework of IGTTA in a variety of patient-wise adaptation settings. These include 1) problems containing varying degrees of infection, highly imbalanced and fewer classes - Chest CT segmentation 2) larger domain shift, multi-site adaptation problems - MR tumor segmentation 3) highly balanced and large number of classes (11) - OCT retinal layer segmentation. These problems cover the typical adaptation scenarios encountered in medical imaging segmentation studies, demonstrating the utility of proposed approach as a generic method.

The effectiveness of entropy based uncertainty minimization (Tent) is a strong function of number of classes. Larger the number of classes, more effective entropy minimization becomes. This is because, reduced entropy measure on a likelihood vector with larger number of classes, implies that there is one class among the many that has the highest likelihood mass, making it the likely/decisive winner. This effect is observed in OCT problem (number of classes is 11), where all the SOTA methods including Tent [14] benefit from entropy/uncertainty minimization and need for regularization is lessened. However, in problems like CT and MR, need for functional regularization becomes much stronger. Both un-regularized solution in Tent [14] and weight-regularizer solution in OSUDA [40] suffer from risk of harmful adaptation in patient-wise adaptation scenarios. Additionally, OSUDA is effective only when entire target domain data is available, which is not a practical scenario.

Our framework of IGTTA clearly benefits from interplay of functional regularization and confidence maximization reduc-

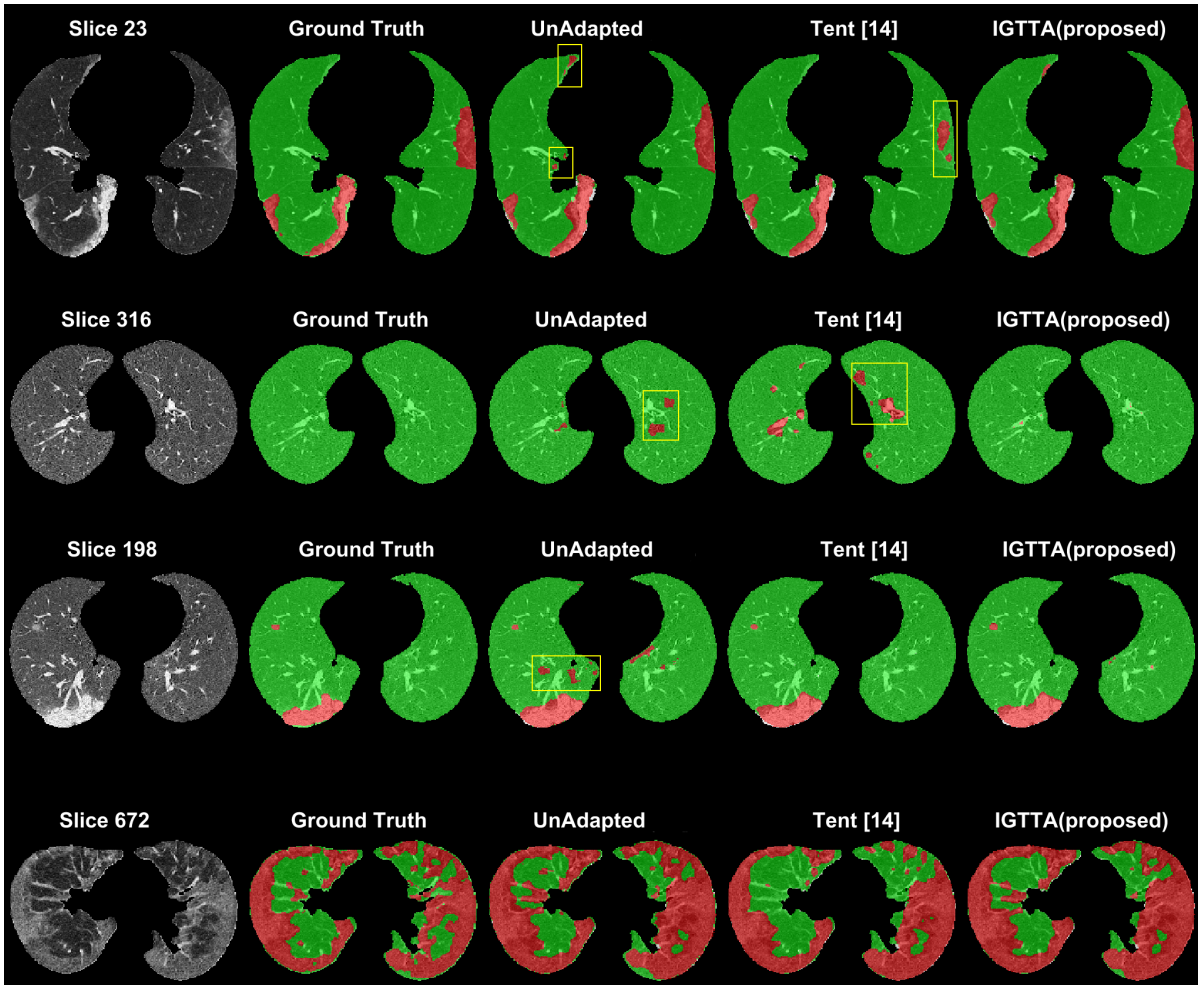


Fig. 7: Anomaly segmentation from Chest CT images: Qualitative comparisons of segmentations from the proposed method and other methods on selected slices from test dataset. Images correspond to adaptation experiments using run 1 of AnamNet [28] as base model. Green: Lung region (Normal); Red: Anomalies (Abnormal), Yellow boxes: Inaccurate segmented region.

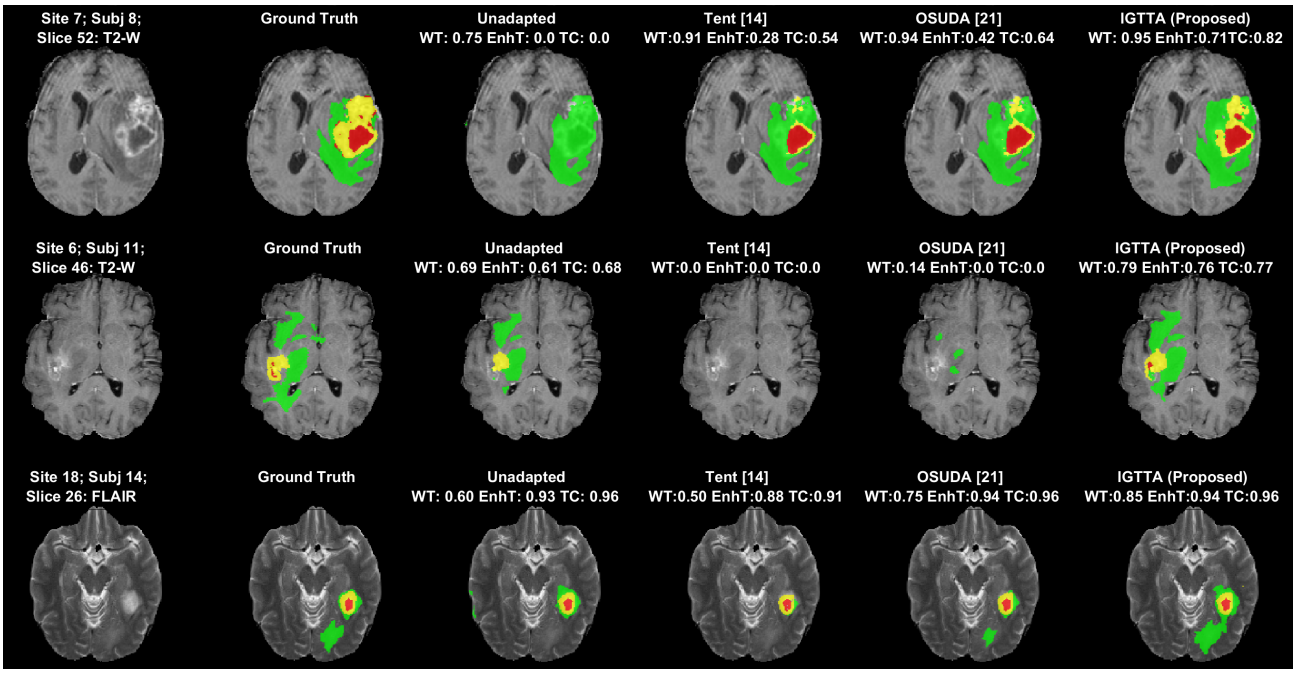


Fig. 8: Brain tumor segmentation for multimodal MR images: Qualitative comparisons of the proposed method and other methods on selected slices from different target sites for models trained on site 1. Images shown correspond to adaptation experiments using run 1 of AnamNet [28] as base model. Yellow: enhancing tumor, Green: peritumoral edema, Red: the necrotic and non-enhancing tumor core. Dice comparisons for WT, ET, and TC are also presented.

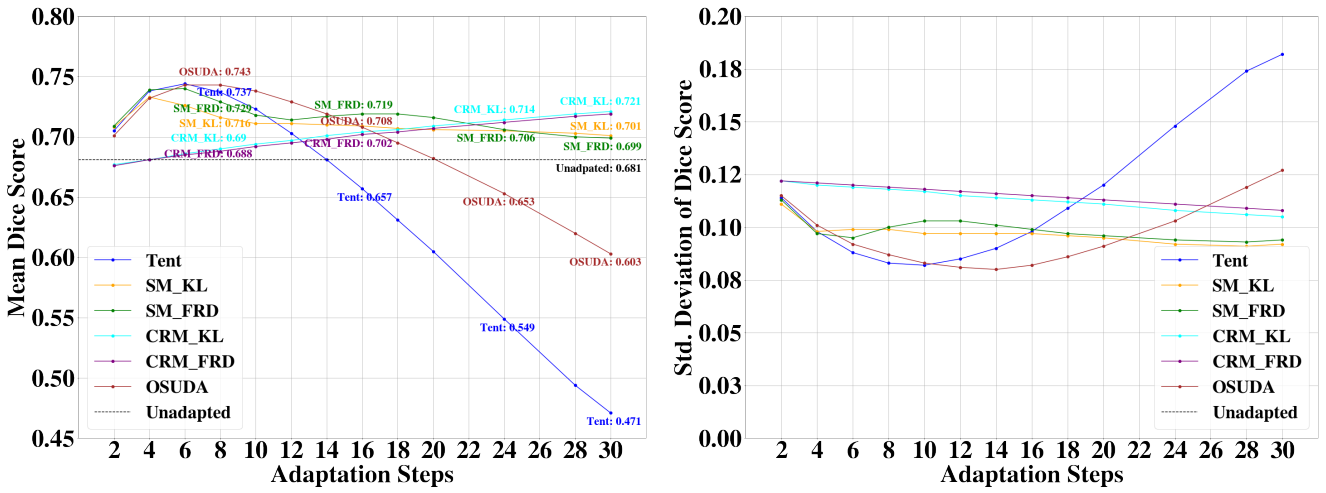
ing uncertainty. Methods that depend on class-ratio priors like SFDA [11] and TTA-SM [10] reduce in effectiveness whenever class-ratio distribution changes from source domain to target domain. This was acutely noticed in CT problem where degree of infection varied from subject to subject and across domains. These methods are more suited for anatomy segmentation than anomaly segmentation, where class-ratios can be expected to be fairly stable across domains. Finally, methods that learn additional neural networks during training like TTA-DAE [18] suffer from shift in imaging distribution, rendering the auxiliary networks less effective and often detrimental. Also, these methods do not lend themselves to off-the-shelf adaptation, as they rely on modifying training procedure by learning extra networks. As noted in the introduction section (Table. II), IGTTA addresses the major gaps in SOTA methods leading to a generic, off-the-shelf, regularized and patient-wise adaptation method suitable to a spectrum of modalities and adaptation scenarios.

VII. Ablation Studies and Analysis

In this section, we discuss complementary benefits of variants of the IGTTA framework and their suitability to different adaptation scenarios. Tables. VI and VII compares results for all variants of IGTTA for chest CT anomaly segmentation and brain tumor segmentation in MR images respectively. Firstly, it should be noted that even though the best performing variant's results were earlier shared for each problem, all the variants outperform SOTA methods on all problems. **Choice of neuromanifold:** The influencing factors are,

1) Extent of domain shift: The voxel manifold imposes stronger functional regularization as it constrains every voxel in the patient scan.This is useful when the domain shift is not drastic like in the CT segmentation problem (Table VI). The class-ratio manifold is a property derived from the voxel manifold, is less constrained, and is useful when the domain shifts are larger, as in the MR segmentation problem. Further, the number of voxels in CT volumes are much larger compared to MR or OCT volumes, which means voxel manifold gets influenced by more number of voxels in objective function.
   It is pointed out that asymptotically, voxel manifold based adaptations will converge to base model's output (eqn. 1), providing a safety net over harmful adaptation. This is reflected in Table. VII, where voxel manifold based methods have less number of adaptation failures at the expense of lesser performance improvement.
2) Number of classes and imbalance of class-ratios: Since class-ratio manifold constraints the distribution of class-ratios, it can fluctuate a lot when total number of classes is low and/or imbalance amongst class-ratios is high. Hence, we observed that for CT problem, voxel manifold was more suited and for other two problems class-ratio manifold yielded better results.
3) Adaptation times: Generally, it is observed that a more iterations are required to match the performance of the class-ratio manifold with the results of the voxel manifold (Fig. 9), roughly 2-3 times. This is because, the class-ratio manifold based functional regularization works only with class-ratio distribution and not all the voxels and hence takes longer to stabilize.

Choice of divergence: As shown earlier, FRD and KL divergence are equivalent, with KL being theoretically an upper



(a) Evolution of Mean of Dice Score

(b) Evolution of Standard deviation of Dice Score

Fig. 9: Adaptation dynamics of discussed methods with varying adaptation steps: using run 1 of AnamNet as base model on selected group of subjects from COVID-19 CT segmentation dataset.

TABLE VI: Chest CT anomaly segmentation: Comparisons for neuromanifold types and choice of divergence measures with AnamNet [28] as base model.

Variant	Neuro-Manifold	Divergence Measure	Dice Abnormal	Dice Normal
IGTTA-VM$_{\text{FRD}}$	**Voxel**	**Fisher Rao.**	**0.758 ± 0.038**	**0.978 ± 0.007**
IGTTA-VM$_{\text{KL}}$	**Voxel**	**KL Div.**	0.753 ± 0.028	0.977 ± 0.001
IGTTA-CRM$_{\text{FRD}}$	**Shape**	**Fisher Rao.**	0.751 ± 0.028	0.971 ± 0.003
IGTTA-CRM$_{\text{KL}}$	**Shape**	**KL Div.**	0.748 ± 0.028	0.973 ± 0.002

TABLE VII: Brain tumor segmentation from multi-modal MR images: Comparisons for neuromanifold types and choice of divergence measures with AnamNet [28] as base model.

Variant	Neuro-Manifold	Divergence Measure	Whole Tumor Dice Score ↑	# Failures on WT
IGTTA-VM$_{\text{FRD}}$	**Voxel**	**Fisher Rao.**	0.643 ± 0.003	**19/740**
IGTTA-VM$_{\text{KL}}$	**Voxel**	**KL Div.**	0.652 ± 0.052	21/740
IGTTA-CRM$_{\text{FRD}}$	**Shape**	**Fisher Rao.**	0.749 ± 0.012	24/740
IGTTA-CRM$_{\text{KL}}$	**Shape**	**KL Div.**	**0.754 ± 0.009**	26/740

bound for FRD. In [23], it was observed that training dynamics varied between the two divergence choices, which is what we have observed with our experiments of test-time adaptation also. Exploring theoretical underpinnings behind the differences in their training/adaptation dynamics are beyond the scope of this paper, will be taken for future research.

Fig. 9 captures adaptation dynamics of all variants of proposed approach along with Tent [14] and OSUDA [21] providing empirical reasoning for observations. Evolution of mean and standard deviation of Dice score for abnormal class as a function of number of adaptation steps has been depicted. These scores are obtained from group of subjects in CT dataset evaluated using one run of AnamNet as base model. These subjects had higher amount of infection/abnormal voxels, making them suitable for this analysis. Firstly, as shown in Fig. 9(a), Tent [14] starts deteriorating in performance after initial few adaptation steps, clearly highlighting perils of unconstrained adaptation. OSUDA's weight regularization is insufficient with performance dropping below base model, later than Tent. Note that all proposed variants of IGTTA provide better performance than unadapted model's performance. Secondly, variants based on class-ratio manifold (CRM_KL and CRM_FRD), gradually increase with steps (Fig. 9(a)), taking almost twice the adaptation steps to reach performance of voxel manifold methods (VM_KL and VM_FRD) at step 10. As noted earlier, with more steps voxel manifold based methods asymptotically converge to base model's performance. In Fig. 9(b), it is noted that, Tent [14] and OSUDA [21] have increasing standard deviation with number of steps. In other words, performance of adaptation of these methods vary highly with subjects. In contrast, all variants of proposed approach have decreasing and nearly flat standard deviation across subjects, making the proposed approach reliable - which is another benefit of functional regularization.

FRD gives exact geodesic distance on the neuromanifold, hence it should be used if adaptation times are not a constraint. Since optimizing on upper-bound can potentially lead to faster convergence, if adaptation times are a concern, it is recommended to use KL divergence. The recipe that has yielded best results in faster adaptation times is as follows: For lesser domain shifts, larger volumes and lesser number of classes use combination of voxel manifold and FRD/KL divergences (CT problem). For more number of classes, larger domain shifts use class-ratio manifold with KL divergence (MR and OCT problems), for a balance of improved performance and satisfactory adaptation times.

Choice of confidence function and λ: The proposed framework of IGTTA has many components that come together during adaptation - statistical neuro-manifold, choice of functional regularizers and confidence functions. Hence, the interplay amongst them has led to fairly robust performances across different choices and parameters. For instance, performance was noted to be fairly consistent amongst λ values in the range

TABLE VIII: Chest CT anomaly segmentation: Dice scores and ECE for IGTTA-VM$_{\mathbf{KL}}$ studying the impact of l_{conf} loss with AnamNet [28] as base model.

Choice of ℓ_{conf} functions	Dice Score ↑		ECE ↓ (3 classes)
	Abnormal	Normal	
Soft-max entropy ℓ_{sme}	**0.758 ± 0.038**	**0.978 ± 0.007**	**0.475**
Generalized cross-entropy ℓ_{gce}	0.747 ± 0.029	0.973 ± 0.003	0.479
Margin ℓ_{margin}	**0.759 ± 0.012**	**0.978 ± 0.005**	**0.475**

TABLE IX: Retinal Layers segmentation from OCT images: Average Dice scores and ECE for IGTTA-CRM$_{\mathbf{KL}}$ studying the impact of l_{conf} loss with residual UNet [11] as base model.

Choice of ℓ_{conf} functions	Dice Score ↑	ECE ↓ (11 classes)
Soft-max entropy ℓ_{sme}	**0.830 ± 0.013**	**0.036**
Generalized cross-entropy ℓ_{gce}	0.829 ± 0.017	0.037
Margin ℓ_{margin}	**0.830 ± 0.025**	**0.034**

$[0.5, 1]$, while slight drop ($< 0.3\%$ in the CT problem) was observed for ($\lambda = 0.1$), which is to be expected.

To further understand the impact of different confidence functions, we compute average expected calibration error (ECE) in Tables VIII and IX. While metrics like Dice score and surface distances measure performance of the methods against ground-truth, ECE is a standard metric that measures how well calibrated the adapted models are. For our computations, we used number of calibration bins = 10, and we report average expected calibration error across all the classes. We chose CT and OCT problems, as they cover two ends of spectrum in terms of number of classes (CT =3, OCT =11). As shown in Table VIII, all the confidence functions have comparable results on both Dice score and ECE (since only 3 classes), including the proposed Margin-based confidence function. It is interesting to note that on the OCT problem (Table IX), overall ECE across confidence functions is significantly lower (since there are 11 classes) and margin based confidence function, achieves meagre decrease over other functions. The reasons for similar performances of these confidence functions and how to optimally select the best iterate of adapted model based on uncertainty will be explored in future work. In summary, the proposed approach performs consistently across different choices of confidence loss functions (including the proposed ℓ_{margin}) and λ values and produces improved results on all three problems. Robustness to hyper-parameter tuning is another salient feature of the proposed approach.

VIII. Discussion and Conclusion

Traditionally, generalization has been targeted as a training time goal, where different algorithmic strategies are utilized to ensure that the performance on data with distribution shift is acceptable. However, TTA treats generalization as an inference time goal and promises a path towards precision health. As a recent area of research, one of the bigger unsolved and less addressed problem is to ensure that the adapted model does not deteriorate in performance owing to over-eager adaptation.

The aim of this study was to ensure that sufficient algorithmic caution is exercised in patient-specific adaptation. An intuitive and practical methodology for robust TTA grounded in the theoretical framework of information geometry was presented. Extensive experiments on a variety of clinically relevant and challenging problems demonstrate the promise of IGTTA as a solution for achieving the dual goal of patient optimality and generalization. It was demonstrated that IGTTA does not require any additional information, is fast, and can be applied to any off-the-shelf model irrespective of the training procedure and architecture.

The functional regularization provided by the proposed IGTTA makes it a generic framework and other methods like TTA-SM [10] that use anatomical priors become special case of the same. Importantly, information geometric methods have been shown to provide effective representative learning [41], by exploiting structure of the data. The improved performance in adaptation observed in this study can be attributed to information geometric methods providing compact functional space based on the structure of test data for updating weights, thus making it work universally across architectures.

The max margin loss in this study contributes to the practicality of IGTTA based test-time adaptation (TTA) of deep-learning models for semantic segmentation. The max margin loss aims to improve the confidence calibration of the model, which is crucial in medical imaging applications, where accurate uncertainty estimation can significantly impact clinical decisions. This work provided the theoretical underpinnings of employing max margin loss to lower the Expected Calibration Error (ECE). The ablation studies (Table IX) provided experimental validation that max margin loss contributes to reduction in the calibration error, thus making the adapted model's predictions more reliable. Future research will be devoted to uncertainty-aware test-time adaptation, where early stopping and selection of right hyper-parameters can be attempted in addition to reducing the calibration error.

There are a few limitations of IGTTA and other TTA methods that provide scope for future work. One of the unaddressed problems in this study is determining *when* and *how* much to adapt. The proposed IGTTA is a conservative framework that intentionally inhibits functional divergence, thereby limiting the extent of possible improvement, unlike unregularized versions, such as Tent. It will be useful to identify candidates for adaptation and control the amount of adaptation in a principled manner. Another direction is to reduce the dependencies on the hyperparameters, some form of parameter-free adaptation framework is ideal. Encouraged by the success of proposed IGTTA in this study, extending it to other medical image analysis tasks, such as classification and image regression, will be taken as the next step.

The code utilized to generate the results in this study is shared here: `https://github.com/hariharanrav/IGTTA_TMI/tree/main`. Additional results are also presented in the same link.

REFERENCES

[1] M. Kim *et al.*, "Deep Learning in Medical Imaging," *Neurospine*, vol. 16, no. 4, pp. 657–668, Dec. 2019.

[2] "An epic failure: Overstated ai claims in medicine." [Online]. Available: https://mindmatters.ai/2021/08/an-epic-failure-overstated-ai-claims-in-medicine/

[3] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro, "Exploring Generalization in Deep Learning," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[4] H. Guan and M. Liu, "Domain adaptation for medical image analysis: a survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2021.

[5] K. B. Johnson *et al.*, "Precision Medicine, AI, and the Future of Personalized Health Care," *Clinical and Translational Science*, vol. 14, no. 1, pp. 86–93, Jan. 2021.

[6] National Science and Technology Council, "Roadmap for Medical Imaging Research and Development," p. 19, 2017.

[7] H. Ravishankar *et al.*, "Understanding the mechanisms of deep transfer learning for medical images," in *DLMIA 2016, MICCAI 2016, Proceedings 1*. Springer, 2016, pp. 188–196.

[8] H. Guan and M. Liu, "Domain adaptation for medical image analysis: a survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2021.

[9] M. Bateson *et al.*, "Source-free domain adaptation for image segmentation," *Medical Image Analysis*, vol. 82, p. 102617, 2022.

[10] M. Bateson and et al., "Test-time adaptation with shape moments for image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention 2022*. Springer, 2022, pp. 736–745.

[11] Y. He *et al.*, "Autoencoder based self-supervised test-time adaptation for medical image analysis," *Medical Image Analysis*, vol. 72, p. 102136, Aug. 2021.

[12] J. M. J. Valanarasu *et al.*, "On-the-fly test-time adaptation for medical image segmentation," *arXiv preprint arXiv:2203.05574*, 2022.

[13] Y. Sun *et al.*, "Test-Time Training with Self-Supervision for Generalization under Distribution Shifts," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 9229–9248.

[14] D. Wang *et al.*, "Tent: Fully Test-Time Adaptation by Entropy Minimization," in *ICLR*, Sep. 2020.

[15] S. Niu *et al.*, "Efficient test-time model adaptation without forgetting," in *ICML*. PMLR, 2022, pp. 16 888–16 905.

[16] S. Niu and et al., "Towards stable test-time adaptation in dynamic wild world," *preprint arXiv:2302.12400*, 2023.

[17] J. Liang, R. He, and T. Tan, "A comprehensive survey on test-time adaptation under distribution shifts," *arXiv preprint :2303.15361*, 2023.

[18] N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu, "Test-time adaptable neural networks for robust medical image segmentation," *Medical Image Analysis*, vol. 68, p. 101907, 2021.

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

[20] S. M. Ahmed *et al.*, "Cross-modal knowledge transfer without task-relevant source data," in *ECCV*. Springer, 2022, pp. 111–127.

[21] X. Liu *et al.*, "Adapting Off-the-Shelf Source Segmenter for Target Medical Image Segmentation," in *MICCAI 2021*. Springer, Cham, Sep. 2021, pp. 549–559.

[22] S.-I. Amari, *Information geometry and its applications*. Springer, 2016, vol. 194.

[23] S. I. Amari, "Natural gradient works efficiently in learning," *Neural computation*, vol. 10, no. 2, pp. 251–276, 1998.

[24] S. A. Gattone *et al.*, "A shape distance based on the fisher–rao metric and its application for shapes clustering," *Physica A: Statistical Mechanics and its Applications*, vol. 487, pp. 93–102, 2017.

[25] J. Pinele, J. E. Strapasson, and S. I. Costa, "The fisher–rao distance between multivariate normal distributions: Special cases, bounds and applications," *Entropy*, vol. 22, no. 4, p. 404, 2020.

[26] H. K. Miyamoto, F. C. Meneghetti, and S. I. Costa, "The fisher–rao loss for learning under label noise," *Information Geometry*, pp. 1–20, 2022.

[27] E. D. C. Gomes, F. Alberge, P. Duhamel, and P. Piantanida, "Igeood: An information geometry approach to out-of-distribution detection," *arXiv preprint arXiv:2203.07798*, 2022.

[28] N. Paluru *et al.*, "Anam-Net: Anamorphic Depth Embedding-Based Lightweight CNN for Segmentation of Anomalies in COVID-19 Chest CT Images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 932–946, Mar. 2021.

[29] M. Picot, F. Messina, M. Boudiaf, F. Labeau, I. B. Ayed, and P. Piantanida, "Adversarial robustness via fisher-rao regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2698–2710, 2022.

[30] E. Rusak *et al.*, "Adapting ImageNet-scale models to complex distribution shifts with self-learning," *arXiv:2104.12928 [cs]*, Apr. 2021, arXiv: 2104.12928. [Online]. Available: http://arxiv.org/abs/2104.12928

[31] B. Murugesan, B. Liu, A. Galdran, I. B. Ayed, and J. Dolz, "Calibrating segmentation networks with margin-based label smoothing," *Medical Image Analysis*, vol. 87, p. 102826, 2023.

[32] B. Liu, I. Ben Ayed, A. Galdran, and J. Dolz, "The devil is in the margin: Margin-based label smoothing for network calibration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 80–88.

[33] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "Missformer: An effective transformer for 2d medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1484–1494, 2022.

[34] "COVID-19 DATABASE - SIRM." [Online]. Available: https://sirm.org/category/senza-categoria/covid-19/

[35] D. J. Bell, "COVID-19 — Radiology Reference Article — Radiopaedia.org." [Online]. Available: https://radiopaedia.org/articles/covid-19-4?lang=us

[36] S. Pati, U. Baid, M. Zenk, B. Edwards, M. Sheller, G. A. Reina, P. Foley, A. Gruzdev, J. Martin, S. Albarqouni *et al.*, "The federated tumor segmentation (fets) challenge," *arXiv preprint :2105.05874*, 2021.

[37] Y. He, A. Carrass *et al.*, "Retinal layer parcellation of optical coherence tomography images: Data resource for multiple sclerosis and healthy controls," *Data in Brief*, vol. 22, pp. 601–604, Feb. 2019.

[38] Y. He *et al.*, "Segmenting retinal OCT images with inter-B-scan and longitudinal information," in *Medical Imaging 2020: Image Processing*, vol. 11313. SPIE, Mar. 2020, pp. 840–845.

[39] M. Bateson *et al.*, "Source-Free Domain Adaptation for Image Segmentation," *arXiv:2108.03152 [cs]*, Aug. 2021.

[40] X. Liu, F. Xing, C. Yang, G. El Fakhri, and J. Woo, "Adapting off-the-shelf source segmenter for target medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 549–559.

[41] S. Garg and Y. Liang, "Functional regularization for representation learning: A unified theoretical perspective," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 187–17 199, 2020.