2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230764

SAGTTA: SALIENCY GUIDED TEST TIME AUGMENTATION FOR MEDICAL IMAGE SEGMENTATION ACROSS VENDOR DOMAIN SHIFT

Suhang You[⋆] *Devavrat Tomar*[†] *Behzad Bozorgtabar*[†] *Mauricio Reyes*[⋆]

[⋆] ARTORG Center for Biomedical Engineering Research, Univ. of Bern, Switzerland
[†]Signal Processing Laboratory 5 (LTS5), EPFL, Switzerland

ABSTRACT

Test time augmentation has been shown to be an effective approach to combat domain shifts in deep learning. Despite their promising performance levels, the interpretability of the underlying used models is however low. Saliency maps have been widely used in medical image analysis as a post-hoc interpretability method for deep learning models. Beyond explainability, in this study, we propose SaGTTA (Saliency Guided Test Time Augmentation), the first learnable framework that introduces saliency information to guide test time augmentations via a novel self-supervised loss term. During test time augmentation, the proposed self-supervised saliency-guided loss aims at promoting augmentation policies that enhance the distinctiveness among class-specific saliency maps. By promoting saliency distinctiveness among different labels of the test image during test time augmentation, the data distribution discrepancy between the test image and training dataset is alleviated. We compared the proposed method with a state-of-the-art method, using a publicly available dataset, showing improvements in terms of performance, model calibration, and robustness. The code will be made publicly available at https://github.com/yousuhang/SaGTTA.

***Index Terms*—** Saliency map, Intepretability, Test time augmentation, Medical image segmentation, Domain shift

1. INTRODUCTION

Deep neural networks (DNN) have been actively proposed to tackle clinical problems and have been highly impactful in achieving state-of-the-art performance levels in medical image segmentation [1]. One common challenge in clinical practice is known as domain shift [2] which leads to a drastic decrease in performance for unseen datasets presenting large discrepancies in terms of acquisition protocols, vendors, etc. [3]. Many approaches have been proposed to mitigate the issue of domain shift. However, these methods typically require labeled multi-center/-vendor data during training to learn domain-agnostic representations [4], or leverage unlabeled data across vendors to improve the model generalisation [5], which is challenging in clinical practice due to the scarcity of test data, or the need to retrain/fine-tune trained models upon deployment. Therefore, a model that can self-adapt without these limitations has raised much interest. Recently, methods for test time adaptation and test time augmentation (TTA) have been proposed to tackle this issue. Test time adaptation aims at modifying the model parameters using unlabeled data and self-supervision [6], while TTA promotes the test data distribution to be close to the distribution of the training dataset [7]. In this study, we focus on TTA due to its flexibility and no requirements to change a model once it has been deployed (e.g., no need for re-certification of modified models).

Specifically, OptTTA [7] is a state-of-the-art approach that takes advantage of batch norm statistics learned from the training dataset to optimize and select the best augmentation policy yielding the highest similarity between the batch norm statistics of the training and augmented test set. Despite the success of OptTTA, it requires batch norm statistics which does not apply to all DNN segmentation architectures. Moreover, batch norm statistics might also not well approximate the training data population statistics in all cases, as reported in [8]. Besides, the interpretability of the underlying neural network method and associated batch-norm-based policy optimization remains low, which is of importance in clinical practice when deploying and auditing these models. Motivated by SIBNet [9], which uses saliency information within an inductive bias for improved training of classification models, we investigated the possibility of employing saliency information to guide test time augmentations via a novel self-supervised loss term. During test time augmentation, the proposed self-supervised saliency-guided loss aims at promoting augmentation policies that enhance the distinctiveness among class-specific saliency maps. By promoting saliency distinctiveness among different labels of the test image during test time augmentation, the data distribution discrepancy between the test image and training dataset is alleviated.

Our **contributions** are: 1) The first learnable framework that introduces interpretability to guide test time augmentation under domain shift. 2) The injected interpretability information not only guides the optimization of augmentation sub-policies but also provides saliency maps implicitly informing about the effect of the applied data augmentation policy. 3)

We compare our method with OptTTA in terms of model performance, model calibration, and robustness against scarcity of training data and additive Gaussian noise.

2. METHODS

Our proposed method consists of two processes as shown in Fig. 1. The first process is **optimization & selection of sub-policies** and the second one is **fine-tuning and applying** the selected sub-policies to the test dataset in an ensemble manner. In the first process, all possible augmentation sub-policies are included and each one is optimized using a few samples from the testing dataset. After optimization, the best sub-policies will be selected and further fine-tuned with the rest samples from the test dataset and applied to generate image segmentations. In the following sections, we detail each process, followed by descriptions of the proposed saliency-guided test-time self-supervised loss term.

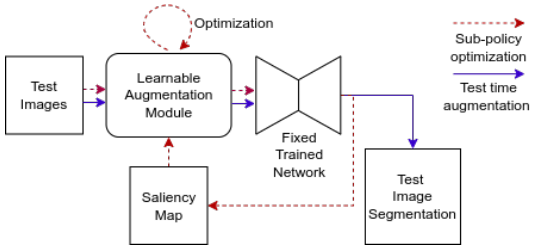

Fig. 1. SaGTTA workflow. First, each sub-policy is optimized with a few samples from the test dataset and the best sub-policies are selected. Second, the selected sub-policies are further fine-tuned with the rest of the samples from the test dataset. Finally, the selected sub-policies are applied to the test dataset to generate image segmentations.

2.1. Augmentation Module

Policy and Sub-policy. A sub-policy is defined as an image transformation or a combination of image transformations. We adopted commonly used image transformation combinations used in policy optimization, from *Identity*, *Gamma Correction*, *Gaussian Blur*, *Contrast modification*, *Brightness modification*, *Resize Crop*, *Horizontal Flip*, *Vertical Flip* and *Random Rotation*. Examples of sub-policies could be [*Gamma Correction*, *Identity*, *Random Rotation*]. A policy consists of one or more sub-policies. To apply the optimized policy, we first optimize each sub-policy in the augmentation module and combine all selected best sub-policies predictions to create segmentation maps for the test dataset.

Learnable sub-policy and Reparameterization. Among considered sub-policies, some image transformations are not learnable (e.g. *Vertical Flip*) and some are learnable (e.g. *Gamma Correction*). For sub-policies with learnable parameters, we define the process as:

$$I_p = \mathcal{T}(I_{in}; \Theta) \tag{1}$$

where I_{in} and I_p are input and augmented image volumes. $\mathcal{T}(\cdot, \Theta)$ is a sub-policy with k step transformation process parameterized by $\Theta = [\theta_1, ...\theta_i, ..., \theta_k]^T$ where θ_i is an n-dimensional vector $\theta_i \in \mathbf{R}^n$ sampled from a uniform distribution:

$$\theta_i \sim \mathcal{U}(a, b) = \mu_i + \sigma_i \cdot \mathcal{U}(-1, 1) \tag{2}$$

where a and b are the lower and upper bound of the parameter space, respectively. During optimization, the distribution of each θ_i is learned through gradient descent in the reparameterized form where $\mu_i \in \mathbf{R}^n$ and $\sigma_i \in \mathbf{R}^n$ parameters to define the sampling mean and standard deviation, respectively. $\mathcal{U}(-1, 1)$ is a n-dimensional uniform distribution where each dimension ranges in [-1, 1].

2.2. Saliency guided optimization & Sub-policy selection

2.2.1. Sub-policy optimization

Saliency maps reflect DNN's response to the input image where pixel intensities reflect pixel attribution or level of importance to the task [10]. For a segmentation model classifying each pixel into one out of K classes, a total of K saliency maps can be calculated, each yielding a class-specific saliency map. We exploit this property of saliency map methods to optimize sub-policies, by promoting class-specific saliency maps to be as distinctive as possible among all K class-specific saliency maps.

For a policy augmented image volume $I_p \in \mathbf{R^N}$, the total loss function of a sub-policy optimization is defined as the linear combination of an entropy loss ($\mathcal{L}_{ent}$), a Nuclear Norm loss $\mathcal{L}_{nn}$, and the proposed saliency-guided loss $\mathcal{L}_{sal}$:

$$\mathcal{L}_{total}(I_p) = \mathcal{L}_{ent} + \alpha \cdot \mathcal{L}_{nn}(I_p) + \beta \cdot \mathcal{L}_{sal}(I_p), \tag{3}$$

where α and β are hyper-parameters. Below we describe each loss term.

Conditional Entropy Loss ($\mathcal{L}_{ent}$): Aggregates all pixel entropy conditioned on the augmented image I_p.

$$\mathcal{L}_{ent} = -\frac{1}{N}\Sigma_{i=1}^{N}\Sigma_{c=1}^{K} p(c|I_p) \log p(c|I_p) \tag{4}$$

where $p(c|I_p)$ is the prediction probability of label c on the augmented image I_p. This loss term encourages confidence in the predictions of the model.

Nuclear Norm Regularization ($\mathcal{L}_{nn}$): Regularizes the effect of entropy minimization that reduces prediction diversity. The nuclear norm is a convex approximation of the matrix rank. Through maximization, prediction diversity and label discrimination are maintained [11].

$$\mathcal{L}_{nn} = trace(\sqrt{p(c|I_p)^* p(c|I_p)}) \tag{5}$$

where $p(c|I_p)^*$ is the conjugate transpose of $p(c|I_p)$.

Saliency Guided Loss ($\mathcal{L}_{sal}$). The proposed $\mathcal{L}_{sal}$ aims at promoting distinctiveness of class-specific saliency maps.

$$\mathcal{L}_{sal}(I_p) = \binom{K-1}{2}^{-1} \Sigma_{c_i=1}^{K-2} \Sigma_{c_j=c_i+1}^{K-1} CosSim(S_{c_i}, S_{c_j}) \tag{6}$$

where $CosSim(S_{c_i}, S_{c_j})$ is the cosine similarity between the saliency of label c_i and c_j. $\binom{K-1}{2}^{-1}$ takes the mean of all paired labels where $\binom{}{}$ is the combination operator that calculates all combinations of $K-1$ non-background labels. Each saliency map is calculated using Integrated Gradient [12] simply adapted for segmentation.

2.2.2. *Best sub-policies selection and test time augmentation*

After optimization of all sub-policies using gradient descent, the top-m sub-policies yielding the lowest loss values $\mathcal{L}_{total}$ (Eq.(3)) are selected to be further fine-tuned. During test time augmentation, we randomly draw M samples of Θs for each sub-policy and apply them to the test dataset. The ensemble augmented segmentation probability $\hat{p}(c|I_{test})$ is finally calculated as:

$$\hat{p}(c|I_{test}) = \frac{1}{mM} \Sigma_{i=1}^{m} \Sigma_{j=1}^{M} p(c|{I_p}^{i,j}) \tag{7}$$

where m is the number of selected sub-policies, and ${I_p}^{i,j}$ is the augmented image from sub-policy indexed i and jth sampled parameter set Θ.

3. EXPERIMENTS AND RESULTS

3.1. Dataset & Implementation

Dataset. We apply SaGTTA to the publicly available Spinal Cord Gray Matter Segmentation dataset [13], which includes four medical centers collected by different vendors. The annotated class labels are Gray Matter and White Matter. In our experiments, and following [7], we specifically focus on adapting test data from site #3 to the model trained with site #1 since this corresponds to the most challenging domain shift scenario for the dataset [7].

Implementation Details. Images were re-sampled to isotropic 1mm resolution with bi-cubic interpolation. For the test data, we only applied re-sampling to the Axial plane. We trained the segmentation model using 2D U-net due to volume inconsistency in the cranio-caudal direction. During training, we used batch-wise weighted cross entropy loss and applied RMSprop optimizer with a learning rate of 10^{-5} for 250K iterations. During optimization of sub-policies, we set $\alpha = -0.005$ and $\beta = -0.01$ empirically and selected top-3 (i.e, m=3) sub-policies for fine-tuning using AdamW optimizer with a learning rate of 10^{-3}. All experiments were implemented in PyTorch with NVIDIA GeForce GTX 1080ti GPU. Segmentation performance was measured by dice similarity coefficient (DSC).

Table 1. DSC(%) comparison among compared methods. The model is trained with site 1 image volumes and tested on site 3 image volumes. The paired t-test was applied to 20 reruns of OptTTA and SaGTTA, $P < 0.005$.

Method	Direct Test	OptTTA	SaGTTA	Grid Search
DSC Mean%	57.09	79.43	80.16	80.36
DSC Std%	19.99	4.16	2.75	2.45

3.2. Results

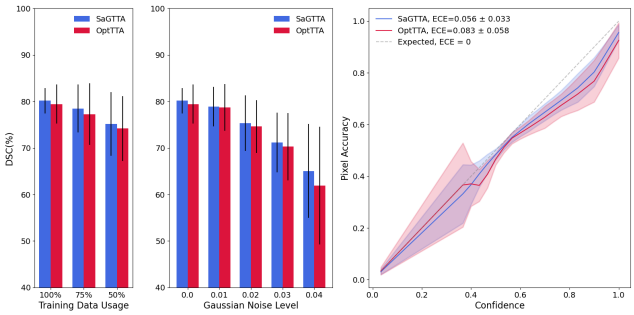

Fig. 2. SaGTTA DSC compared to OptTTA. Left: Mean and std DSC of augmented segmentation using different percentages of training data used for the trained model. Middle: Mean and std DSC of augmented segmentation compared at different levels of additive Gaussian noise added to the training data. Right: Reliability plot of augmented segmentation results with expected calibration error (ECE) for SaGTTA and OptTTA. The dashed line corresponds to perfect calibration.

As shown in Table 1, compared to OptTTA, SaGTTA shows improved performance for mean and std DSC. To compare to the best or upper bound of performance in this dataset, we also ran an extensive grid search for parameters of all sub-policies and selected the 3 best sub-policies, with results showing that both OptTTA and SaGTTA converge very closely to the upper bound of performance.

We also compared SaGTTA to OptTTA in three other experiments. First, we checked for improved robustness to smaller training dataset size by reducing its size to 75% and 50%. Second, we added different levels of Gaussian noise to the training data. As shown in Fig. 2, SaGTTA performs better in all ablated studies with larger means of DSC and smaller stds of DSC. Third, we compared the calibration properties of the two methods. The expected calibration error (ECE) was calculated for the white matter and gray matter of each test volume (lower values are better). SaGTTA achieved better ECE values at 0.056 ± 0.033, indicating better cali-

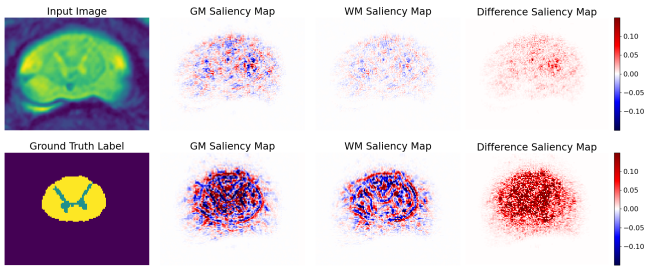

Fig. 3. Saliency map Comparison. From left to right: Original image and ground truth, Saliency maps of gray matter (GM) and white matter (WM) before and after applying SaGTTA first second row, respectively. Last column: Image differences of saliency maps before and after SaGTTA.

bration properties than OptTTA. Shown in Fig. 3, we also compared the saliency maps of each label before and after SaGTTA. The saliency guides the test data to increase attention, especially in the edge areas of tissues. Specifically, the silhouettes of the gray matter and white matter borders can be observed in each label's saliency maps after SaGTTA.

4. CONCLUSION

In our study, we propose SaGTTA, a saliency-guided test time augmentation, and the first learnable framework that introduces interpretability to guide test time augmentation. SaGTTA yielded performance and robustness improvements, along with improved calibration and interpretable information at the benefit of optimized augmentation policies.

5. COMPLIANCE WITH ETHICAL STANDARDS & ACKNOWLEDGMENTS

Compliance with ethical standards. This research study used human subject data publicly available at GRAY MATTER SPINAL CORD SEGMENTATION CHALLENGE [13]. Ethical approval was not required as confirmed by the license attached and its publication description.

Acknowledgments. This work is supported by Swiss Personalized Health Network (SPHN) initiative.

6. REFERENCES

[1] F.Isensee, P. F.Jaeger, S. A.Kohl, J.Petersen, and K. H.Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[2] W.Yan, Y.Wang, S.Gu, L.Huang, F.Yan, L.Xia, and Q.Tao, "The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 623–631.

[3] B.Glocker, R.Robinson, D. C.Castro, Q.Dou, and E.Konukoglu, "Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects," *arXiv preprint arXiv:1910.04597*, 2019.

[4] Q.Dou, D.Coelho de Castro, K.Kamnitsas, and B.Glocker, "Domain generalization via model-agnostic learning of semantic features," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[5] C. S.Perone, P.Ballester, R. C.Barros, and J.Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," *NeuroImage*, vol. 194, pp. 1–11, 2019.

[6] Y.Sun, X.Wang, Z.Liu, J.Miller, A.Efros, and M.Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International conference on machine learning*. PMLR, 2020, pp. 9229–9248.

[7] D.Tomar, G.Vray, J.-P.Thiran, and B.Bozorgtabar, "Opttta: Learnable test-time augmentation for source-free medical image segmentation under domain shift," in *Medical Imaging with Deep Learning*, 2021.

[8] Y.Wu and J.Johnson, "Rethinking" batch" in batchnorm," *arXiv preprint arXiv:2105.07576*, 2021.

[9] D.Mahapatra, A.Poellinger, and M.Reyes, "Interpretability-guided inductive bias for deep learning based medical image," *Medical image analysis*, vol. 81, pp. 102551, 2022.

[10] K.Simonyan, A.Vedaldi, and A.Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[11] S.Cui, S.Wang, J.Zhuo, L.Li, Q.Huang, and Q.Tian, "Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3941–3950.

[12] M.Sundararajan, A.Taly, and Q.Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[13] F.Prados, J.Ashburner, C.Blaiotta, T.Brosch, J.Carballido-Gamio, M. J.Cardoso, B. N.Conrad, E.Datta, G.Dávid, B.De Leener, et al., "Spinal cord grey matter segmentation challenge," *Neuroimage*, vol. 152, pp. 312–329, 2017.