Dynamic Sub-domain Modeling for Robust Medical Image Segmentation

Kyungsu Lee[1], Seo-Yeon Choi[1], Jae Youn Hwang[2], and Jonghye Woo[3]

[1] Jeonbuk National University, South Korea
[2] Daegu Gyeonbuk Institute of Science & Technology, South Korea
[3] Massachusetts General Hospital and Harvard Medical School, MA USA

Abstract. Intra-domain distribution shifts are common in medical image segmentation due to variations in acquisition protocols, reconstruction settings, and anatomical morphology, even when data originate from the same nominal domain. We propose a fully automatic segmentation framework that explicitly models latent sub-domains without gradient-based adaptation at inference. Our method dynamically discovers sub-domains via a nonparametric prototype mechanism, expanding only when novel distributional modes are detected. Each prototype conditions a specialized expert via a hypernetwork, forming a prototype-conditioned expert field over the latent manifold. Soft routing with uncertainty-based temperature scaling enables parameter-free test-time adaptation by calibrating sub-domain posteriors without updating model weights. Experiments on 2D medical benchmarks demonstrate improved robustness and granular segmentation accuracy under intra-domain shifts.[1]

Keywords: Segmentation · Domain Adaptation · Foundation Model.

1 Introduction

Recent advances in medical image segmentation have reduced large inter-domain discrepancies through domain adaptation (DA) and domain generalization (DG) techniques [10, 32, 35]. However, even when training and testing data come from the same nominal dataset, performance often saturates due to unresolved intra-domain variability. Medical images frequently exhibit multimodal distributions driven by acquisition protocols, scanner vendors, reconstruction parameters, and anatomical or pathological appearance, forming latent sub-domains within a single dataset [25]. Training one shared model over these heterogeneous sub-distributions encourages representation averaging, which can blur boundaries and weaken subtle structural cues [27]. In boundary-sensitive tasks, this averaging creates an intrinsic ceiling: gains increasingly depend on contour-level reliability rather than coarse region accuracy [18, 19]. We argue that breaking this ceiling requires explicitly modeling latent sub-domain structure rather than further homogenizing it.

[1] Project link: https://github.com/Macs-Laboratory/lsd-seg

State-of-the-art and Limitations Fine-grained segmentation methods improve boundary precision via boundary-aware objectives [18], multi-scale context aggregation [34], atrous convolution [7], and point-based refinement [19], enhancing local structural fidelity. However, these approaches typically assume a unified data distribution and therefore do not explicitly account for intra-domain multimodality, where samples within a single dataset form latent sub-domains due to acquisition and appearance variations [25]. Consequently, they are limited when boundary ambiguity stems from unresolved sub-domain shifts rather than insufficient architectural capacity. To mitigate distribution shifts more broadly, domain adaptation and generalization methods emphasize alignment and invariance: adversarial feature alignment and output-space matching reduce source–target discrepancies [10, 32], while invariant risk minimization and style-based normalization promote domain-agnostic representations [2, 36]. Despite their effectiveness for addressing inter-domain gaps, these paradigms suppress distribution-specific variation to achieve invariance, which can attenuate subtle high-frequency or contrast-dependent cues needed for precise delineation. Although recent work recognizes intra-domain variability, it still largely relies on alignment or augmentation that implicitly homogenizes sub-distributions rather than explicitly modeling their structure. Conditional computation provides an alternative direction: Mixture-of-Experts (MoE) decomposes prediction into multiple experts coordinated by input-dependent gating [16, 28], sparse routing enables scalable capacity expansion [9], and vision adaptations suggest benefits from increased representation diversity. However, existing MoE segmentation frameworks typically fix the number of experts and focus on capacity scaling or task decomposition rather than modeling intra-domain heterogeneity as a latent structural variable, and they do not address the practical setting where sub-domains are unknown or new modes may emerge during training.

Our Contribution In this work, we reinterpret single-domain segmentation as inference over a latent sub-domain variable and propose a dynamic sub-domain modeling framework to explicitly capture intra-domain heterogeneity. Instead of enforcing invariance, we treat distributional variability as a structured signal. We introduce a nonparametric prototype mechanism that adaptively discovers latent sub-domains without predefining their number, expanding capacity only when novel modes are detected. Each prototype conditions a specialized expert via a hypernetwork, forming an expert field over the latent sub-domain manifold. Soft routing with uncertainty-tempered calibration enables parameter-free test-time adaptation without gradient updates. By explicitly modeling latent sub-domains and enabling conditional specialization, the proposed framework alleviates the ceiling effect induced by intra-domain heterogeneity and improves boundary-level reliability in fully automatic medical image segmentation.

To summarize, existing approaches fail to explicitly model latent intra-domain variability within a single dataset, leading to performance saturation. We reinterpret single-domain segmentation as inference over latent sub-domains and propose a dynamic expert framework that explicitly models this structure.

2 Problem Definition

In this research, we study single-domain medical image segmentation under latent intra-domain heterogeneity, where samples from the same nominal dataset arise from multiple hidden modes (e.g., acquisition protocol, scanner vendor, and patient-specific anatomy). We model this setting via an unobserved sub-domain variable $z \in \mathcal{Z}$ such that $(x, y) \sim \sum_{z\in\mathcal{Z}} p(z)p(x, y|z)$, inducing a mixture conditional $p(y|x) = \sum_{z\in\mathcal{Z}} p(z|x)p(y|x, z)$. The objective is to learn a predictor $f_\theta : \mathcal{X} \to \mathcal{Y}$ that remains accurate under shifts in the test-time sub-domain prior $p(z)$, including rare or previously unseen modes, while strictly avoiding test-time parameter updates (i.e., no gradient-based adaptation) and relying solely on input-dependent inference-time specialization.

3 Method

Overview We propose a fully automatic segmentation framework that explicitly models latent intra-domain sub-domains while performing parameter-free inference (i.e., no test-time gradient updates). Given a 2D medical image $x \in \mathbb{R}^{H\times W}$ with annotation y, we extract frozen foundation features $F = E(x)$ from a SAM2/3-style encoder $E(\cdot)$ and obtain an automatic prompt $q = P(x)$ (e.g., box, sparse points, or a coarse mask). We approximate a mixture-of-functions

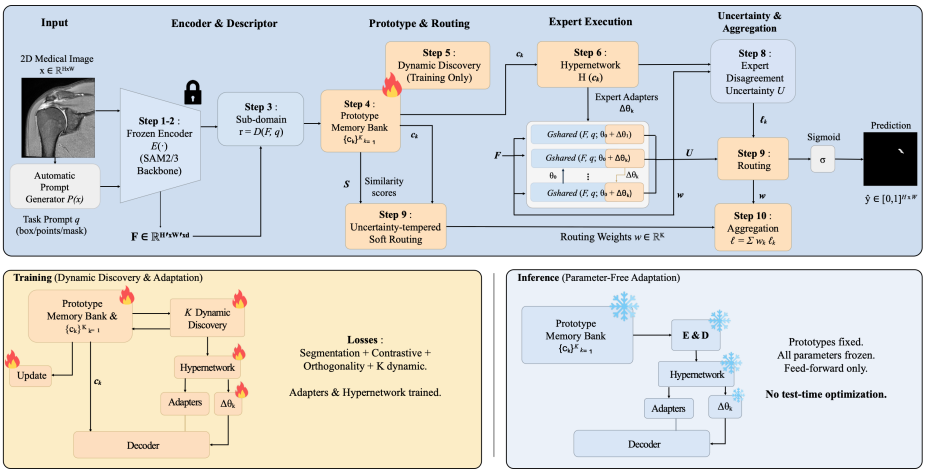

Fig. 1. Overview of our proposed framework. Frozen modules (**blue**) extract features F and prompt q from image x. A learnable descriptor block (**orange**) computes r to dynamically manage prototypes $\{c_k\}$ within a memory bank (**green**) during training (dashed **gray** area). These prototypes drive a hypernetwork (**orange**) to generate expert adaptations $\Delta\theta_k$ for a shared decoder, producing logits $\{\ell_k\}$. The final prediction $\hat{y}$ is obtained by uncertainty-tempered soft routing across these experts. Inference is completely parameter-free.

by (i) learning a sub-domain descriptor, (ii) dynamically discovering a set of prototypes without predefining K, and (iii) instantiating a prototype-conditioned expert field via a hypernetwork that generates lightweight adaptations for a shared mask decoder. For the resulting K experts, we compute logits $\{\ell_k\}_{k=1}^{K}$ and routing weights $\{w_k\}_{k=1}^{K}$ and compose the prediction as

$$\hat{y} = \sigma\left(\sum_{k=1}^{K} w_k \, \ell_k\right), \tag{1}$$

where $\sigma(\cdot)$ is sigmoid (or softmax for multi-class segmentation) and K is determined during training by the proposed discovery rule.

Sub-domain Descriptor and Dynamic Prototype Discovery To represent intra-domain variability in a task-relevant manner, we learn a compact descriptor $r = D(F, q) \in \mathbb{R}^m$ using a lightweight head on top of frozen features. Concretely, we aggregate global appearance cues $g = \mathrm{Pool}(F)$ and prompt-aligned local cues $\tilde{\ell} = \mathrm{ROIPool}(F; q)$ and form $r = \phi([g; \tilde{\ell}])$, where $\phi(\cdot)$ is a small MLP and $[\cdot;\cdot]$ denotes concatenation. Latent sub-domains are represented by prototypes $\{c_k \in \mathbb{R}^m\}_{k=1}^{K}$ maintained nonparametrically; given r, we compute cosine similarities $s_k = \frac{r^\top c_k}{\|r\|\|c_k\|}$ and declare novelty if $\max_k s_k < \tau$, in which case we spawn $c_{K+1} \leftarrow r$ and increment K; otherwise we assign $k^\star = \arg\max_k s_k$ and update by EMA $c_{k^\star} \leftarrow (1-\eta)c_{k^\star} + \eta r$. Prototype maintenance is performed only during training for stability: low-support prototypes are pruned by assignment counts and highly similar prototypes are merged when $\frac{c_i^\top c_j}{\|c_i\|\|c_j\|} > \rho$ to prevent uncontrolled growth and fragmentation.

Prototype-Conditioned Expert Field via Hypernetwork Each prototype defines a specialized expert, but storing a separate decoder per prototype is inefficient when K is dynamic. We instead generate expert-specific adaptation parameters through a hypernetwork $H(\cdot)$ as $\Delta\theta_k = H(c_k)$ and apply them to a shared SAM-style mask decoder G_{shared} with base parameters θ_0. Each expert logit is computed as $\ell_k = G_{\mathrm{shared}}(F, q; \theta_0 + \Delta\theta_k)$, where $\Delta\theta_k$ is implemented as a lightweight adaptation (e.g., low-rank/adapters). This treats expert parameters as a smooth function over the prototype manifold, enabling immediate instantiation of experts for newly discovered sub-domains while preserving parameter sharing across nearby modes.

Uncertainty-Tempered Soft Routing We route each input to the expert field using similarity-based soft composition, with weights $w_k = \mathrm{softmax}\left(\frac{s_k}{T_r}\right)$, where T_r controls routing sharpness. To avoid overconfident routing under ambiguous inputs or boundary cases, we modulate T_r using an uncertainty proxy derived from expert disagreement: with $p_k = \sigma(\ell_k)$, we estimate $U = \mathrm{mean}(\mathrm{Var}_k(p_k))$ and set $T_r = T_0 + \alpha U$, so that higher disagreement yields smoother mixtures

(larger T_r) and lower disagreement yields sharper specialization. This calibration is fully feed-forward and requires no test-time optimization.

Optimization and Inference Training minimizes a segmentation objective augmented with regularizers that structure the descriptor–prototype space and prevent expert collapse:

$$\mathcal{L} = \mathcal{L}_{seg} + \alpha\,\mathcal{L}_{proto} + \beta\,\mathcal{L}_{ortho}, \tag{2}$$

where $\mathcal{L}_{seg} = \mathcal{L}_{Dice}(\hat{y}, y) + \lambda\,\mathcal{L}_{CE}(\hat{y}, y)$, the prototype contrastive term is $\mathcal{L}_{proto} = -\log \frac{\exp(s_{k^\star}/t)}{\sum_j \exp(s_j/t)}$, and the diversity regularizer is $\mathcal{L}_{ortho} = \sum_{i \neq j} \left\| \Delta\theta_i^\top \Delta\theta_j \right\|_F^2$. At inference, all parameters (including E, P, D, H, and G_{shared}) remain frozen and the prototype set discovered during training is fixed; for each test image we compute (F, q, r), obtain $\{s_k\}$, generate $\{\Delta\theta_k\}$, compute uncertainty-tempered $\{w_k\}$, and output $\hat{y}$ via Eq. (1), achieving robustness to latent intra-domain shifts without any test-time updates.

4 Experiments

Settings We evaluate on 20 public segmentation datasets spanning diverse modalities: endoscopy (PolypGen [1], Kvasir-SEG [17], CVC-ClinicDB [4]), dermoscopy (ISIC 2018) [8], fundus imaging (REFUGE [24], DRIVE [31]), histopathology (MoNuSeg [21], GlaS [30]), ultrasound (BUSI [14], CAMUS [22]), MRI (ACDC [5], PROMISE12 [23], BraTS [33], ISLES'22 [13], WMH [20], MSD Hippocampus [29]), CT (KiTS21 [12], BTCV [11], LiTS [6], LIDC-IDRI [3]), and PET/CT (autoPET [15], HECKTOR [26]). 3D volumes are processed slice-wise into 2D samples. Official splits are used when available; otherwise, patient-level splits are used to prevent leakage. For intra-domain robustness evaluation, subdomains are defined via metadata (e.g., center, pathology) when available, or constructed via clustering of training image descriptors. Additionally, we compare with strong medical segmentation models, including nnU-Net and MedNeXt, and state-space/Mamba architectures (U-Mamba, SegMamba, VM-UNet) in 2D or slice-wise settings. Promptable foundation models are evaluated using SAM2, MedSAM, SAM-Med2D, and MedSAM2, all using the same automatic prompt generator for fairness. We additionally perform ablations on dynamic prototype discovery (fixed-K experts), hypernetwork-based expert generation (static adapters), and uncertainty-tempered routing. All methods are trained under a unified 2D setup with identical preprocessing, augmentation, and optimization per dataset. Our model uses a frozen SAM2/3 encoder with fully automatic prompting, applied consistently to promptable baselines. We report Dice and IoU, HD95 (or ASSD), and Boundary-F1, as well as worst-subdomain Dice and inter-sub-domain variance for robustness. All evaluations are performed without test-time adaptation, and runtime and peak GPU memory are reported.

Table 1. Main results averaged over 20 datasets (mean ± 95% CI).

Method	Dice ↑	HD95 ↓	Boundary-F1 ↑	Worst Dice ↑
(1) nnU-Net (2D)	80.4 ± 1.7	12.9 ± 0.3	74.8 ± 1.7	75.3 ± 6.5
(2) MedNeXt (2D)	82.3 ± 1.8	12.7 ± 0.3	76.7 ± 1.8	77.3 ± 1.3
(3) VM-UNet	81.3 ± 1.8	12.8 ± 0.3	75.7 ± 1.8	75.8 ± 1.2
(4) U-Mamba (2D)	82.4 ± 1.9	12.6 ± 0.3	76.8 ± 1.9	75.9 ± 3.8
(5) SegMamba (slice-wise 2D)	84.0 ± 1.8	12.4 ± 0.3	78.4 ± 1.8	78.6 ± 4.0
(6) SAM2 + APG	82.5 ± 1.8	12.6 ± 0.3	76.9 ± 1.8	76.4 ± 9.1
(7) MedSAM + APG	84.2 ± 1.6	12.4 ± 0.2	78.6 ± 1.6	78.2 ± 1.7
(8) SAM-Med2D + APG	84.4 ± 2.0	12.3 ± 0.3	78.8 ± 2.0	78.6 ± 8.9
(9) MedSAM2 + APG	84.8 ± 1.8	12.3 ± 0.3	79.2 ± 1.8	78.8 ± 9.2
(10) Ours (Fixed-K experts)	84.6 ± 1.8	12.3 ± 0.3	79.0 ± 1.8	78.9 ± 8.4
(11) Ours (w/o hypernetwork; static adapters)	82.8 ± 1.8	12.6 ± 0.3	77.2 ± 1.8	76.4 ± 9.3
(12) Ours (w/o uncertainty-tempered routing)	83.1 ± 1.8	12.5 ± 0.3	77.5 ± 1.8	77.6 ± 6.0
(13) Ours (global-only descriptor; w/o ROI)	83.9 ± 1.8	12.4 ± 0.3	78.3 ± 1.8	77.5 ± 5.0
(14) Ours (Full)	**87.9 ± 1.7**	**11.8 ± 0.3**	**82.3 ± 1.7**	**82.2 ± 7.3**

Main Results Table 1 summarizes performance averaged over 20 datasets (mean ± 95% CI). Our full model achieves the best Dice (87.9) and the lowest HD95 (11.8), outperforming strong baselines, including nnU-Net, MedNeXt, and recent Mamba-based models, under a unified 2D evaluation protocol. Compared to the strongest non-foundation baseline, SegMamba (84.0 Dice), our method improves Dice by +3.9, and it surpasses promptable foundation variants such as MedSAM2 + APG (84.8 Dice) by +3.1, demonstrating that dynamic sub-domain modeling provides benefits beyond backbone scaling or prompt engineering. The largest gains are observed in Boundary-F1 and Worst Dice (82.2 vs. 78.8), indicating improved contour precision and reduced intra-domain performance variance. Notably, the improvement in Worst Dice suggests enhanced robustness to challenging or underrepresented subdomains rather than uniform average-case gains. Ablation results further confirm that dynamic prototype discovery, hypernetwork-based expert generation, and uncertainty-tempered routing each contribute to the final robustness improvement, thereby validating the necessity of explicitly modeling latent sub-domain structure.

Latent Structure and Robustness Analysis Figure 2 provides an empirical analysis of the learned representation geometry, cross-dataset robustness, and expert-level uncertainty behavior. As shown in Fig. 2(a), the UMAP projection reveals that Dataset A, although treated as a single domain during training, forms five compact clusters (S1–S5), indicating that the descriptor uncovers latent intra-domain substructures related to acquisition or anatomical variability. In contrast, Dataset B forms a clearly separated cluster, demonstrating that the representation captures both intra- and inter-domain structure within a unified embedding space. This dual-scale organization supports our formulation of segmentation as inference over a latent mixture rather than a unimodal mapping.

This structured latent space translates into consistent robustness gains. The win–loss heatmap in Fig. 2(b) shows predominantly positive Dice improvements (Ours − Baseline) across 20 datasets and 13 competitors, with gains up to +0.4 Dice in challenging cases. Improvements are especially pronounced in worst-case

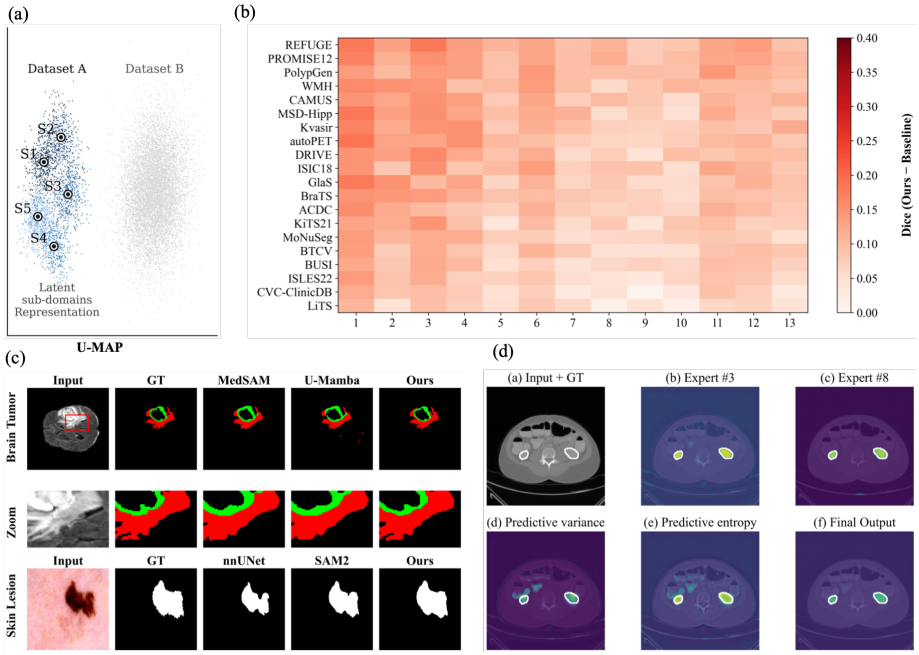

Fig. 2. Experimental analysis of latent sub-domain modeling. (a) UMAP visualization of learned descriptors from BraTS and WMH. (b) Win–loss heatmap showing Dice improvement (Ours − Baseline) across 20 datasets and 13 comparison methods. (c) Qualitative segmentation results on BraTS and ISIC. (d) Uncertainty analysis of expert predictions, including predictive variance and entropy maps.

subdomain performance, confirming that explicitly modeling latent sub-domains reduces intra-domain variance rather than merely increasing mean Dice. Qualitative results in Fig. 2(c) further illustrate sharper and more coherent boundaries in ambiguous peripheral regions compared to strong baselines, consistent with the Boundary-F1 improvements. Finally, Fig. 2(d) shows that predictive variance and entropy concentrate near uncertain boundaries, while uncertainty-tempered routing adaptively interpolates across experts to produce calibrated and stable predictions. These observations indicate that robustness arises from structured expert specialization and adaptive mixture, rather than enforced invariance.

Collectively, these findings empirically support our central hypothesis that intra-domain heterogeneity is better modeled as a mixture of specialized functions over a latent sub-domain manifold, leading to reduced worst-case degradation, improved boundary fidelity, and interpretable internal structure without test-time adaptation.

Mechanism Analysis and Stability. Figure 3 analyzes the robustness mechanism and stability properties of the proposed dynamic sub-domain modeling

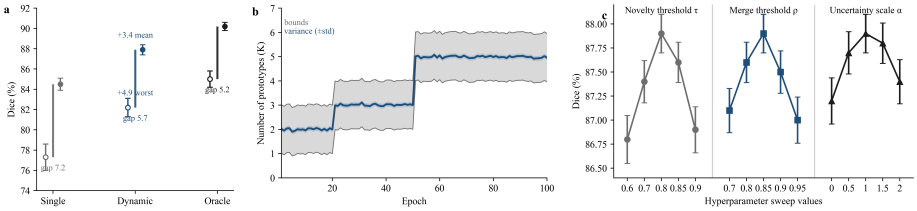

Fig. 3. Mechanism and stability analysis of dynamic sub-domain modeling. (a) Oracle sub-domain evaluation comparing mean and worst-sub-domain Dice. (b) Prototype evolution over 100 training epochs with bounded growth and variance bands. (c) Hyperparameter sensitivity to τ, merge threshold ρ, and uncertainty scale α.

framework. Under oracle sub-domain labels (Fig. 3(a)), the dynamic model improves not only mean Dice but also worst-subdomain Dice, thereby reducing the robustness gap between average and worst-case performance. This result indicates that explicitly modeling latent sub-domains mitigates performance variance induced by hidden mixture shifts rather than merely increasing overall accuracy. Figure 3(b) presents prototype dynamics over 100 training epochs. Although the number of prototypes expands adaptively, the growth remains bounded, and the variance band across runs stays well within predefined limits. The coexistence of bounded expansion and limited variance suggests stable nonparametric discovery without uncontrolled fragmentation of prototypes. Finally, Fig. 3(c) evaluates sensitivity to the novelty threshold τ, merge threshold ρ, and uncertainty scaling parameter α. Performance remains stable across a broad range of values, with no sharp degradation near the selected operating points. This robustness to hyperparameter variation confirms that the gains arise from the modeling principle itself rather than from delicate threshold tuning.

Additional experiments, including per-dataset breakdowns, statistical significance tests, runtime and memory analysis, capacity-matched controls, prompt-sensitivity analysis, and extended prototype/routing visualizations, are provided in the project link for reproducibility and transparency.
(Project link: https://github.com/Macs-Laboratory/lsd-seg)

5 Conclusion

We presented a dynamic sub-domain modeling framework for robust single-domain medical image segmentation. Our approach explicitly captures latent intra-domain heterogeneity and enables fully feed-forward, parameter-free inference without requiring test-time gradient updates. The proposed pipeline learns a task-relevant sub-domain descriptor, discovers prototypes nonparametrically to allocate capacity as novel modes emerge, and instantiates a prototype-conditioned expert field via a hypernetwork. Furthermore, uncertainty-tempered soft routing mitigates overconfident specialization under ambiguous inputs. Experimental results demonstrate that this design improves robustness to shifts

in hidden-mode mixtures while providing interpretability of the latent structure through learned prototypes and routing weights. Despite its strengths, the approach involves trade-offs between expert specialization and prototype fragmentation. Future work will focus on enhancing prompt generation, exploring richer sub-domain representations, and developing prototype dynamics that better model continuous manifolds while maintaining strict no-update deployment constraints.

Acknowledgments. This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HR22C1832). Also, this research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00354020).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ali, S., Jha, D., Ghatwary, N., Realdon, S.: A multi-centre polyp detection and segmentation dataset for generalisability assessment. Scientific Data **10**(1) (2023)
2. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
3. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2) (2011)
4. Bernal, J., Sánchez, Fernández-Esparrach, G., Gil, D., Rodríguez, C.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43** (2015)
5. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11) (2018)
6. Bilic, P., Christ, P., Li, H.B., Vorontsov: The liver tumor segmentation benchmark (lits). Medical image analysis **84** (2023)
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4) (2017)
8. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
9. Fedus, W., Zoph: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120) (2022)
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59) (2016)

11. Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula: Multi-organ abdominal ct reference standard segmentations (Feb 2018). https://doi.org/10.5281/zenodo.1169361
12. Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao: The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. arXiv preprint arXiv:2307.01984 (2023)
13. Hernandez Petzsche, M.R., De La Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L.: Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific data **9**(1) (2022)
14. Huang, J., Zhang, J., Zhang, Y., Li, X., Ma, X., Deng, J., Shen, H., Wang, D., Mei, L., Lei, C.: BUSI_WHU: Breast cancer ultrasound image dataset. Mendeley Data, V3 (2025). https://doi.org/10.17632/k6cpmwybk3.3
15. Ingrisch, M., Dexl, J., Jeblick, K.: Automated lesion segmentation in whole-body pet/ct-multitracer multicenter generalization. In: 27th International Conference on Medical Image Computing and Computer Assisted Intervention (2024)
16. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1) (1991)
17. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling. Springer (2020)
18. Kervadec, H., Bouchtiba, J.: Boundary loss for highly unbalanced segmentation. Medical Image Analysis **67** (Jan 2021). https://doi.org/10.1016/j.media.2020.101851
19. Kirillov, A., Wu, Y.: Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)
20. Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE transactions on medical imaging **38**(11) (2019)
21. Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., Chen, H., Heng, P.A., Li, J., Hu, Z., et al.: A multi-organ nucleus segmentation challenge. IEEE transactions on medical imaging **39**(5) (2019)
22. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE transactions on medical imaging **38**(9) (2019)
23. Litjens, G., van Ginneken, B., Huisman, H., van de Ven, W., C.: PROMISE12: Data from the MICCAI grand challenge: Prostate mr image segmentation 2012. Zenodo (2023). https://doi.org/10.5281/zenodo.8026660, medical Image Analysis, vol. 18, no. 2, pp. 359–373. Updated the zip files, fixed some issues. Dataset
24. Orlando: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical Image Analysis **59** (2020). https://doi.org/10.1016/j.media.2019.101570
25. Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S.: Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)
26. Quetin, S., Enger, S.A.: Automatic lesion and lymph node segmentation from pet and ct scans of the head and neck region: a hecktor 2025 challenge report. OpenReview preprint (2025), available at https://openreview.net/forum?id=dO98B1XEiL
27. Saeedan, F., Weber, N.: Detail-preserving pooling in deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)

28. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
29. Simpson, A.L., Antonelli, M., Bakas, S., Bilello: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
30. Sirinukunwattana, K., Pluim, J.P.W., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U.: Gland segmentation in colon histology images: The glas challenge contest (2016)
31. Staal, J., Abràmoff, M.D., Niemeijer, M.: Ridge-based vessel segmentation in color images of the retina. IEEE transactions on medical imaging **23**(4) (2004)
32. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
33. Verdier: The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. arXiv preprint arXiv:2405.18368 (2024)
34. Zhao, H., Shi, J., Qi, X., Wang, X.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017)
35. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(4) (2022)
36. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. arXiv preprint arXiv:2104.02008 (2021)