Supplementary Material for "Enhancing Consistency Regularization in Semi-Supervised Medical Image Segmentation: A Posterior Perturbation Perspective"

Hongyu Zhang[1], Haipeng Chen[1], Yingda Lyu[1],
Zenan Shi[1], Yongping Yang[2], Yu Wang[1*]

[1] *College of Computer Science and Technology, Jilin University, Changchun, Jilin, China*
[2] *Department of Colorectal Anal Surgery, the Second Hospital of Jilin University, Changchun, Jilin, China*

1 Derivation of the Evidence Lower Bound (ELBO)

Consider the marginal likelihood,

$$\log p_\theta(x, y) = \log \int p_\theta(x, y, r)\, \mathrm{d}r, \tag{1}$$

where the joint distribution factorizes as

$$p_\theta(x, y, r) = p_\theta(y \mid x, r)\, p_\theta(x \mid r)\, p(r). \tag{2}$$

By introducing a variational distribution $q_\phi(r \mid x)$ and inserting the identity $\frac{q_\phi(r|x)}{q_\phi(r|x)}$ under the integral, the expression becomes

$$\log p_\theta(x, y) = \log \int p_\theta(x, y, r)\, \frac{q_\phi(r \mid x)}{q_\phi(r \mid x)}\, \mathrm{d}r = \log\, \mathbb{E}_{r\sim q_\phi(r|x)} \Big[\tfrac{p_\theta(x,y,r)}{q_\phi(r|x)} \Big]. \tag{3}$$

Implies Jensen's inequality $\log \mathbb{E}[\cdot] \geq \mathbb{E}[\log(\cdot)]$

$$\log p_\theta(x, y) \;\geq\; \mathbb{E}_{r\sim q_\phi(r|x)} \Big[\log \tfrac{p_\theta(x,y,r)}{q_\phi(r|x)} \Big]. \tag{4}$$

Substituting $p_\theta(x, y, r) = p_\theta(y \mid x, r)\, p_\theta(x \mid r)\, p(r)$ into the logarithm and rearranging terms yields

$$\log p_\theta(x, y) \;\geq\; \mathbb{E}_{r\sim q_\phi(r|x)} \Big[\log p_\theta(y \mid x, r) + \log p_\theta(x \mid r) + \log p(r) - \log q_\phi(r \mid x) \Big]. \tag{5}$$

*Corresponding author: wangyu001@jlu.edu.cn

Rewriting $\mathbb{E}_{q_\phi(r|x)}[\log q_\phi(r \mid x) - \log p(r)]$ as the KL divergence

$$D_{\mathrm{KL}}\big(q_\phi(r \mid x) \,\big\|\, p(r)\big) = \mathbb{E}_{r\sim q_\phi(r|x)}\big[\log q_\phi(r \mid x) - \log p(r)\big] \tag{6}$$

leads to

$$\log p_\theta(x, y) \;\geq\; \mathbb{E}_{r\sim q_\phi(r|x)}\big[\log p_\theta(y \mid x, r) + \log p_\theta(x \mid r)\big] \;-\; D_{\mathrm{KL}}\big(q_\phi(r \mid x) \,\big\|\, p(r)\big). \tag{7}$$

Hence, defining

$$\mathrm{ELBO}(\theta, \phi; x, y) = \mathbb{E}_{r\sim q_\phi(r|x)}\big[\log p_\theta(y \mid x, r) + \log p_\theta(x \mid r)\big] \;-\; D_{\mathrm{KL}}\big(q_\phi(r \mid x) \,\big\|\, p(r)\big), \tag{8}$$

we see that maximizing this ELBO with respect to both θ and ϕ provides a tight lower bound on $\log p_\theta(x, y)$.

2 Derivation of the KL Divergence $\mathcal{L}_{KL}$

Consider the prior

$$p(r) = \mathcal{N}\big(\mathbf{0},\, \varepsilon\, \mathbf{I}\big)$$

with $\varepsilon > 0$ and the posterior

$$q_\phi(r \mid x) = \mathcal{N}\big(\mu_\phi(x),\, \Sigma_\phi(x)\big).$$

Suppose $r \in \mathbb{R}^d$. The Kullback–Leibler divergence between two multivariate Gaussians $q = \mathcal{N}(\mu_q, \Sigma_q)$ and $p = \mathcal{N}(\mu_p, \Sigma_p)$ is given by

$$D_{KL}[q \parallel p] = \frac{1}{2}\Big[\mathrm{tr}\big(\Sigma_p^{-1}\Sigma_q\big) \;+\; (\mu_p - \mu_q)^\top \Sigma_p^{-1}(\mu_p - \mu_q) \;-\; d \;+\; \log \tfrac{\det(\Sigma_p)}{\det(\Sigma_q)}\Big]. \tag{9}$$

Substituting $\mu_q = \mu_\phi(x)$, $\Sigma_q = \Sigma_\phi(x)$, $\mu_p = \mathbf{0}$, and $\Sigma_p = \varepsilon\,\mathbf{I}$ leads to three primary terms. First, the trace term $\mathrm{tr}(\Sigma_p^{-1}\Sigma_q) = \frac{1}{\varepsilon}\,\mathrm{tr}\big(\Sigma_\phi(x)\big)$. Second, the quadratic form $(\mu_p - \mu_q)^\top \Sigma_p^{-1}(\mu_p - \mu_q) = \frac{1}{\varepsilon}\,\mu_\phi(x)^\top\,\mu_\phi(x)$. Third, the log-determinant ratio $\log \frac{\det(\Sigma_p)}{\det(\Sigma_q)} = d\,\log\varepsilon - \log\det\big(\Sigma_\phi(x)\big)$. Collecting all pieces yields

$$D_{KL}[q_\phi(r \mid x) \parallel p(r)] = \tfrac{1}{2}\Big[\tfrac{1}{\varepsilon}\,\mathrm{tr}\big(\Sigma_\phi(x)\big) + \tfrac{1}{\varepsilon}\,\mu_\phi(x)^\top\,\mu_\phi(x) - d + d\,\log\varepsilon - \log\det\big(\Sigma_\phi(x)\big)\Big]. \tag{10}$$

When ε and d are constants, the term $d(\log\varepsilon - 1)$ can be ignored for optimization purposes, so one defines

$$\mathcal{L}_{KL} = \tfrac{1}{\varepsilon}\Big[\mathrm{tr}\big(\Sigma_\phi(x)\big) + \mu_\phi(x)^\top\,\mu_\phi(x)\Big] - \log\det\big(\Sigma_\phi(x)\big), \tag{11}$$

which forms the core KL objective up to an additive constant.

3 Derivation of the Product of Multivariate Gaussians

Consider two Gaussian densities $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ over $\mathbb{R}^D$. Their PDFs can be written as

$$p_1(x) = (2\pi)^{-\frac{D}{2}}\,[\det(\Sigma_1)]^{-\frac{1}{2}}\,\exp\Big[-\tfrac{1}{2}\,(x-\mu_1)^\top\,\Sigma_1^{-1}\,(x-\mu_1)\Big], \tag{12}$$

$$p_2(x) = (2\pi)^{-\frac{D}{2}}\,[\det(\Sigma_2)]^{-\frac{1}{2}}\,\exp\Big[-\tfrac{1}{2}\,(x-\mu_2)^\top\,\Sigma_2^{-1}\,(x-\mu_2)\Big]. \tag{13}$$

Multiplying $p_1(x)$ and $p_2(x)$ yields

$$p_1(x)\,p_2(x) = (2\pi)^{-D}\,[\det(\Sigma_1)\,\det(\Sigma_2)]^{-\frac{1}{2}}\,\exp\Big[-\tfrac{1}{2}\,\big((x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)^\top\Sigma_2^{-1}(x-\mu_2)\big)\Big]. \tag{14}$$

Defining $\Lambda_1 := \Sigma_1^{-1}$, $\Lambda_2 := \Sigma_2^{-1}$, and $\Lambda_{\text{new}} := \Lambda_1 + \Lambda_2$, one expands and reorganizes the exponent to complete the square. This step identifies the center $\mu_{\text{new}} = \Lambda_{\text{new}}^{-1}(\Lambda_1\,\mu_1 + \Lambda_2\,\mu_2)$ and shows that

$$p_1(x)\,p_2(x) \;\propto\; \exp\Big[-\tfrac{1}{2}\,\big(x-\mu_{\text{new}}\big)^\top\,\Lambda_{\text{new}}\,\big(x-\mu_{\text{new}}\big)\Big], \tag{15}$$

which is an unnormalized Gaussian with mean μ_{new} and precision matrix Λ_{new}.

Extending this to a product of d_z Gaussians in the same space $\mathbb{R}^D$ leads to

$$\Sigma_{\text{joint}}^{-1} = \sum_{i=1}^{d_z}\Sigma_i^{-1}, \quad \mu_{\text{joint}} = \Sigma_{\text{joint}}\sum_{i=1}^{d_z}\Sigma_i^{-1}\,\mu_i, \tag{16}$$

which can be applied to derive, for instance,

$$\Sigma_\phi(x) = \Big(\sum_{i=1}^{d_z}\Sigma_\phi^i(z_r^i)^{-1}\Big)^{-1}, \quad \mu_\phi(x) = \Sigma_\phi(x)\sum_{i=1}^{d_z}\Sigma_\phi^i(z_r^i)^{-1}\,\mu_\phi^i(z_r^i). \tag{17}$$

Thus, multiplying multiple Gaussians in $\mathbb{R}^D$ simply sums their precision matrices and combines their means via precision weighting.

4 Differentiating from VAE

Our learning framework differs from a Variational Autoencoder (VAE) [1] in several key aspects: **(i) Learning Paradigm**: VAEs function as unsupervised generative models $p_\theta(x|z)$, while our framework operates as a semi-supervised conditional classification model $p_\theta(y|x,r)$. The generative pathway $p_\theta(x|r)$ in our framework is used exclusively to optimize the ELBO during training and is discarded during inference. **(ii) Modeling Objective**: Unlike conditional VAEs (cVAEs) [2], which condition the latent variable distribution on both x and y (i.e., $q_\phi(z|x,y)$), our framework models $q_\phi(r|x)$, relying only on x. This label-independent perturbation learning improves data efficiency, particularly valuable in semi-supervised settings with limited labeled data, and results in a distinct ELBO formulation. **(iii) Parameter Efficiency**: VAEs often have high parameter demands, making training on 3D tasks challenging. In contrast, our approach incorporates **AIH** to reduce parameter costs, enabling efficient segmentation of high-resolution medical volume.

5 Effects of loss weight

We optimized the weights of $\mathcal{L}_{rec}$ (α), $\mathcal{L}_{KL}$ (β), and the peak weight of $\mathcal{L}_{cc}$ (γ), with the search process detailed in Fig. 1. Experimental results on the LA dataset show that α is relatively insensitive to variations within the range of 0.5 to 2.0, but a notable performance drop occurs when $\alpha = 4.0$. Based on this observation, we selected $\alpha = 1.0$. For β, we found that larger values can lead to posterior collapse [3], while a smaller value of $\beta = 0.001$ effectively mitigates this issue. However, setting β too small (*e.g.*, $\beta = 0.0001$) may result in under-regularization, leading to slight

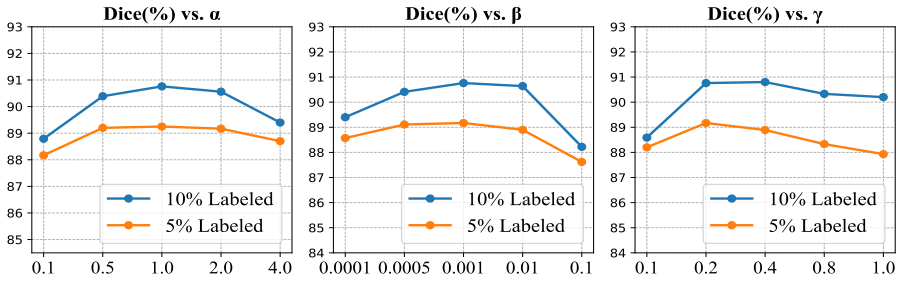

Figure 1: Impact of loss weights α ($\mathcal{L}_{rec}$), β ($\mathcal{L}_{KL}$), and γ ($\mathcal{L}_{cc}$) on the model's Dice performance on the LA dataset.

performance degradation. Regarding γ, the performance curves exhibit consistent variation across different values, with $\gamma = 0.2$ yielding the best results.

Table 1: Comparison of the complexity for various semi-supervised medical segmentation methods on the LA dataset. Metrics include total training parameters (Train Para., M), training Multiply-Accumulate Operations (MACs, G), training time (Wall-Clock Time, h), and total test parameters (Test Para., M), all measured on a single NVIDIA GeForce RTX 3090 GPU.

Methods	**Complexity**			
	Train Para.[M]	MACs.[G]	Train Time[h]	Test Para.[M]
V-Net [4]	9.45	246.5	-	9.45
MC-Net+ [5]	15.25	747.4	5.0	9.45
BCP [6]	18.92	493.7	5.5	9.45
MRP [7]	12.35	505.3	6.0	9.45
Ours	23.27	819.6	5.8	9.45

6 Overhead analysis

We evaluate the training parameter count, computational cost per iteration, training time (Wall-Clock Time), and test parameter count for our method compared to V-Net [4], MC-Net+ [5], BCP [6], and MRP [7] on the LA dataset, using a single NVIDIA GeForce RTX 3090 GPU, as summarized in Tab. 1. While our method introduces a slight increase in complexity compared to existing approaches, the design of **AIH** and the use of a single Monte Carlo sampling step ensure that the increases remain within an acceptable range. Moreover, these computational costs are offset by the performance improvements observed. Our bidirectional consistency training distills the knowledge from the complex $\mathcal{B}_p$ to the simpler $\mathcal{B}_s$. Therefore, during testing, only $\mathcal{B}_s$ (a V-Net) is used as the inference model, with all other components discarded. This approach is consistent with comparable methods, ensuring that the computational cost during inference and deployment remains unchanged.

References

[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[2] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[3] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

[4] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[5] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022.

[6] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11514–11524, 2023.

[7] Jiawei Su, Zhiming Luo, Sheng Lian, Dazhen Lin, and Shaozi Li. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. *Medical Image Analysis*, 94:103111, 2024.