This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Hierarchical Cautious Optimization for Semi-Supervised Medical Image Segmentation

Aaron Olender[1], Aviad Rabinowich[2,3,4], Jayan Khawaja[2,3], Dafna Ben Bashat[3,4,5], and Leo Joskowicz[1]

[1] School of Computer Science & Engineering, Hebrew University of Jerusalem
[2] Department of Radiology, Tel-Aviv Sourasky Medical Center
[3] Gray Faculty of Medical & Health Sciences, Tel-Aviv University
[4] Sagol Brain Institute, Tel-Aviv Sourasky Medical Center
[5] Sagol School of Neuroscience, Tel-Aviv University
aaron.olender@mail.huji.ac.il

Abstract. In volumetric medical image segmentation, semi-supervised learning (SSL) leverages unlabeled data to reduce the reliance on scarce expert annotations. However, standard optimizers treat gradients from labeled and unlabeled objectives uniformly, despite their differing reliability. We propose Hierarchical Cautious Optimization (HCO), an optimizer level method that enforces a trust hierarchy between gradient sources. HCO computes momentum using gradients from labeled data and incorporates unlabeled gradients only when their update direction is aligned with this labeled-driven momentum. The method integrates seamlessly into existing momentum-based optimizers with negligible computational overhead. Across three datasets, HCO consistently improves performance, with notable gains on a challenging fetal MRI dataset where Dice scores for fetal lungs and liver increased from 0.68 to 0.84 (+24%) and from 0.71 to 0.82 (+16%), respectively. These gains across optimizers and datasets show that HCO is a practical enhancement for SSL segmentation.

Keywords: Medical image segmentation · Semi-supervised learning

1 Introduction

Volumetric medical image segmentation typically requires extensive expert annotations, which are expensive and impractical to obtain [6,13]. Semi-supervised learning (SSL) mitigates this need by leveraging small labeled and larger unlabeled datasets [17]. Most modern SSL methods use pseudo-labeling or consistency regularization to extract supervision from unlabeled samples [23,24]. While substantial effort has focused on improving these signals, the optimization process itself has received far less attention.

SSL methods typically minimize a joint objective over labeled and unlabeled data. Standard optimizers treat gradients from labeled supervision and unlabeled signals symmetrically, despite their different reliability. In practice,

unlabeled gradients are often noisy and high-variance [1]. Treating them equally may amplify confirmation bias or induce co-training collapse, where incorrect predictions reinforce one another across views [5]. This suggests that optimization strategies accounting for gradient reliability may offer a complementary path to improve SSL performance.

Most SSL segmentation methods obtain supervision from unlabeled images via pseudo-labeling [10] and consistency regularization [9,20]. The widely used Mean Teacher (MT) [24] method, in which a teacher network is updated with an exponential moving average (EMA) of the student, provides temporally smoothed targets. Subsequent work focuses on improving pseudo-label reliability through uncertainty filtering [27,30], structural priors [11,15], strong augmentations [2,28], and multi-teacher ensembling [7,21,29].

Research on robust optimization and training acceleration addresses gradient noise via sign-based updates [3,4], limiting outlier gradients by using only the sign of the update. More recently, Cautious Optimizers (CO) [12] modify momentum-based optimizers to prevent backtracking by masking updates when momentum opposes the current gradient, thereby promoting monotonic loss reduction. However, both sign-based and cautious strategies are *objective-symmetric*: they modulate the update based on the total gradient without distinguishing the signal source. In SSL, reliability is asymmetric: labeled gradients are ground-truth, while unlabeled gradients are uncertain.

In this paper, we study how labeled and unlabeled signals can be treated by the optimizer given a fixed SSL objective. We propose the *Hierarchical Cautious Optimization* (HCO), a framework for momentum-based optimizers that enforces a trust hierarchy between labeled and unlabeled gradients. HCO adapts the cautious principle of CO to the SSL setting by anchoring momentum in labeled supervision and admitting only aligned unlabeled components. It computes momentum updates exclusively from labeled gradients, thereby defining a reliable descent direction that allows unlabeled gradients to contribute only when they align with the labeled momentum. As a result, unlabeled signals can reinforce, but never contradict, trusted supervision, preventing noisy updates from corrupting the optimization trajectory.

We evaluate HCO on three volumetric segmentation benchmarks: Left Atrium -MRI[26], Pancreas-CT [19], and a private fetal MRI dataset. In all settings, HCO is a drop-in replacement for standard optimizers. It consistently improves performance without architecture modifications or significant computational overhead.

The contributions of this paper are:

1. HCO, a general optimizer-level approach for semi-supervised learning that serves as a drop-in replacement for standard momentum-based optimizers.
2. A convergence analysis showing that HCO preserves the convergence rate of the underlying optimizer assuming a mild alignment condition, which is verified empirically.
3. Empirical and mechanistic evidence that HCO sustains unlabeled signal and improves performance across three volumetric medical image segmentation benchmarks.

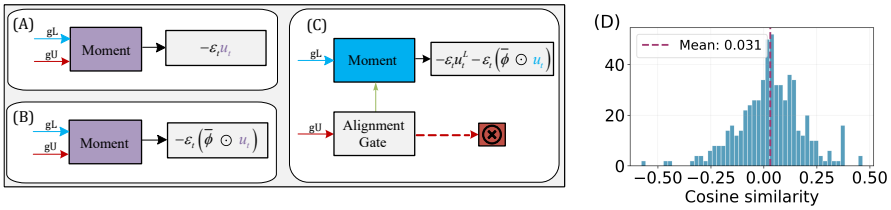


Fig. 1. Optimization mechanisms and source-level gradient conflicts. (A) Standard momentum mixes labeled (g_L) and unlabeled (g_U) gradients. (B) CO applies a sign-agreement filter with momentum. (C) HCO anchors momentum in labeled supervision and gates conflicting unlabeled components. (D) Distribution of cosine similarity between g_L and g_U illustrating frequent anti-alignment ($\cos < 0$). This motivates the HCO's alignment-gated use of unlabeled updates (see details in Study 1).

2 Method

HCO is a general optimization principle for semi-supervised learning that applies to *any* momentum-based optimizer. At each iteration, the optimization momentum is first updated using labeled gradients to establish a trusted direction; unlabeled gradients are incorporated only through components aligned in sign with this direction. Consequently, the momentum aggregates reliable supervision and reinforces unlabeled signals while suppressing conflicting updates. This hierarchical approach is motivated by observed source-level gradient conflicts: the cosine similarity between labeled and unlabeled gradients, while positive on average, exhibits a substantial negative tail that standard momentum-based optimizers do not account for (Fig. 1D).

Hierarchical Cautious AdamW. Algorithm 1 presents the AdamW instantiation of HCO used in this paper. The same principle applies directly to SGD with momentum and to related optimizers. Only the momentum computation and the resulting parameter update are modified; all other components are identical to standard AdamW. The algorithm is described line by line:

1. **Labeled Step (lines 4–10):** a bias-corrected update u_t^L is computed by accumulating the labeled gradient g_t^L into the moment buffers. It is derived solely from ground truth, u_t^L acts as the *trusted direction*.
2. **Unlabeled Momentum Refinement (lines 11–16):** the unlabeled gradient g_t^U is filtered with $\phi_t = \mathbb{I}(u_t^L \odot g_t^U > 0)$ to retain only components aligned with the trusted direction. These filtered gradients update the moments to yield a refined direction u_t.
3. **Cautious Update (lines 17–20):** The binary gate ϕ_t is normalized by its mean to $\bar{\phi}_t$ (Line 17). This maintains the update magnitude even when the mask is sparse. The update (Line 18) applies this gated correction alongside the trusted direction: $w_t \leftarrow w_{t-1} - \frac{\varepsilon_t}{2} u_t^L - \frac{\varepsilon_t}{2}(\bar{\phi}_t \odot u_t)$, followed by standard weight decay (Line 19).

Algorithm 1 AdamW HCO

Require: Initial parameters w_0, step sizes $\{\varepsilon_t\}$, $\beta_1, \beta_2 \in [0, 1)$, weight decay $\gamma \geq 0$, stability constant $\epsilon > 0$.

1: Initialize $t \leftarrow 0$, $m_0 \leftarrow 0$, $v_0 \leftarrow 0$
2: **while** not converged **do**
3: $t \leftarrow t + 1$
4: Compute $g_t^L \leftarrow \nabla_w \mathcal{L}_{\mathrm{L}}(w_{t-1})$
5: $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t^L$
6: $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2)(g_t^L)^2$
7: $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$
8: $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$
9: $u_t^L \leftarrow \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
10: $g_t^U \leftarrow \nabla_w \mathcal{L}_{\mathrm{U}}(w_{t-1})$
11: $\phi_t \leftarrow \mathbb{I}(u_t^L \odot g_t^U > 0)$
12: $m_t \leftarrow m_t + (1 - \beta_1)(g_t^U \odot \phi_t)$
13: $v_t \leftarrow v_t + (1 - \beta_2)(g_t^U \odot \phi_t)^2$
14: $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$
15: $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$
16: $u_t \leftarrow \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
17: $\overline{\phi}_t \leftarrow \phi_t / \max(\mathrm{mean}(\phi_t), \epsilon)$
18: $w_t \leftarrow w_{t-1} - \frac{\varepsilon_t}{2} u_t^L - \frac{\varepsilon_t}{2}(\overline{\phi}_t \odot u_t)$
19: $w_t \leftarrow w_t - \varepsilon_t \gamma w_t$
20: **end while**
21: **return** w_t

Notation. At iteration t, the network parameters are $w_t \in \mathbb{R}^d$. The gradients and moving averages have the same dimension. The operations are element-wise.

- $\boldsymbol{w}_t$ Network parameters after t updates.
- $\boldsymbol{g}_t^L, \boldsymbol{g}_t^U$ Labeled / unlabeled gradients.
- $\boldsymbol{m}_t, \boldsymbol{v}_t$ Exponential moving average buffers of the first and second moments.
- $\boldsymbol{u}_t^L$ Trusted direction derived from labeled gradients g_t^L.
- $\boldsymbol{u}_t$ Final cautious momentum after incorporating aligned components of g_t^U.
- $\mathbb{I}(\cdot)$ Element-wise indicator function returning 1 for true conditions, 0 for false ones.
- $\boldsymbol{\phi}_t$ Binary alignment masks that identify components of g_t^U aligned with u_t^L and u_t.
- $\overline{\boldsymbol{\phi}}_t$ Scaled binary masks that preserve update magnitude despite selective component filtering.

Parameters: ε_t (learning rate), $\beta_{1,2}$ (decay factors), ϵ (stability const).

Design rationale. Binary gating is the minimal mechanism that enforces the trust hierarchy, removing only unlabeled components that would reverse the trusted labeled direction. Since only the entry of g_t^U into the momentum buffer is gated, the momentum u_t and the parameter update stay continuous-valued and update diversity is preserved.

Computational cost. Alignment checks are element-wise and reuse existing quantities (u_t^L, u_t, g_t^U). This adds a wall-clock time overhead of $< 1\%$ in our PyTorch implementation.

Convergence Analysis. We analyze the convergence of HCO by adapting standard non-convex analyses of adaptive gradient methods (e.g., Adam/AdamW) [8,14,18], with one additional *alignment* condition to account for the HCO gating mechanism. We assume that the SSL objective $\mathcal{L}(w) = \mathcal{L}_L(w) + \lambda \mathcal{L}_U(w)$ is L-smooth and bounded below, and that the stochastic gradients of $\mathcal{L}_L$ and $\mathcal{L}_U$ are conditionally unbiased with bounded second moments.

HCO-Specific Descent Condition. The core deviation of HCO from standard Adam is the effective update direction Δ_t (labeled momentum plus a masked unlabeled correction; Algorithm 1, line 18). Substituting Δ_t into the standard descent inequality used in non-convex Adam analyses [18] yields a single HCO-specific requirement: that gating preserves descent on the *combined* objective *in*

expectation – not per-step or per-coordinate. Formally, there exists $c > 0$ such that:

$$\mathbb{E}[\langle \nabla \mathcal{L}(w_{t-1}), \Delta_t \rangle \,|\, \mathcal{F}_{t-1}] \geq c \, \|\nabla \mathcal{L}(w_{t-1})\|^2. \tag{1}$$

Empirical verification. At each self-training step of a complete BCP run on Left-Atrium MRI (8/72 split), we compute the cosine similarity between the flattened full-parameter labeled and unlabeled gradient vectors g_L and g_U. Fig. 1D aggregates these per-step values over the entire run into a histogram. The distribution is positive in expectation even without gating, supporting the plausibility of the descent condition in (1), while its negative tail motivates HCO's filtering of conflicting unlabeled components.

Result. When (1) is satisfied, substituting the gated update direction Δ_t into the L-smooth descent inequality and telescoping over t with step sizes $\varepsilon_t \propto 1/\sqrt{t}$ [18] cancels the higher-order terms and yields the usual non-convex rate:

$$\min_{0 \leq t \leq T-1} \mathbb{E}\big[\|\nabla \mathcal{L}(w_t)\|^2\big] = \mathcal{O}(1/\sqrt{T}). \tag{2}$$

3 Experiments

Datasets. We validate HCO on three datasets: (1) **Left-Atrium** [26]: 100 annotated cardiac 3D MRIs for atrium segmentation (preprocessing follows [2,25]); (2) **Pancreas-CT** [19]: 82 annotated abdominal CTs for pancreas segmentation (preprocessing follows [22]); and (3) **Fetal-MRI** (in-house): 92 labeled and 600 unlabeled 3D MRIs (gestational age 28–39 weeks; 3T Siemens TruFISP; $0.78 \times 0.78 \times 2$ mm), collected with IRB approval. A set of 50 labeled cases is used for testing.

Implementation. Experiments were conducted in PyTorch on NVIDIA RTX 4090 GPUs. Hyperparameters follow the original SSL frameworks unless specified otherwise. HCO introduces no additional hyperparameters and requires no tuning beyond the baseline optimizer settings. We report Dice, Jaccard, Average Surface Distance (ASD), and 95% Hausdorff Distance (HD95).

Experimental Design. We conduct four studies in which the *only* change is the optimizer (baseline vs. HCO). HCO is evaluated across two representative 3D SSL paradigms (BCP, CMT) and three optimizer families (SGD, Adam, AdamW), isolating optimizer-level effects and demonstrating that HCO yields gains independent of the underlying SSL method.

Study 1 (BCP): Generalization across datasets and optimizers.

Within the BCP (bidirectional copy-paste) framework [2], we evaluate HCO on the Left-Atrium MRI using SGD and on Pancreas-CT using Adam. We quantify robustness across anatomy, modality, and optimizer choice. We keep the BCP pre-training checkpoint fixed and swap only the optimizer during self-training (2K pre-train + 15K self-train), using the same 3D V-Net backbone [16] and pipeline on both datasets (Table 1). We use a paired Wilcoxon signed-rank test across test volumes. For the Left Atrium MRI dataset, we average per-scan metrics over two seeds prior to testing and report performance under two labeled

Table 1. Results of the BCP and VNet segmentation models, the optimizers SGD and Adam, and their optimization variations (none, CO, HCO) on two datasets. Left-Atrium MRI: 20 test cases, mean (std) of two seeds. Pancreas-CT: 18 test cases, single run (std shown in paretheses where relevant). L/U indicates the number of Labeled/Unlabeled cases. Metrics: Dice and Jaccard (%), ASD and HD95 (voxels). Best per dataset/method results are shown in bold.

Method	L/U	Dice↑	Jaccard↑	ASD↓	HD95↓
Left-Atrium MRI (3D V-Net backbone; BCP)					
VNet-SGD	4/0	52.55	39.60	4.91	47.05
VNet-SGD	8/0	82.74	71.72	3.26	13.35
VNet-SGD	80/0	91.47	84.36	1.51	5.18
BCP (4 labeled samples)					
BCP-SGD (Repr)	4/76	88.27 (3.26)	79.15 (5.01)	2.26 (1.03)	8.21 (4.18)
BCP-SGD_{CO}	4/76	88.37 (3.11)	79.29 (4.81)	**2.19 (0.88)**	8.07 (3.97)
BCP-SGD_{HCO}	4/76	**88.83 (2.64)**	**79.99 (4.17)**	**2.19 (0.75)**	**6.55 (2.73)**
BCP (8 labeled samples)					
BCP-SGD (Repr)	8/72	89.64 (3.55)	81.41 (5.52)	1.78 (0.86)	6.93 (3.93)
BCP-SGD_{CO}	8/72	89.76 (2.93)	81.54 (4.72)	1.72 (0.70)	6.84 (3.65)
BCP-SGD_{HCO}	8/72	**90.75 (2.17)**	**83.13 (3.63)**	**1.64 (0.58)**	**5.52 (2.26)**
Pancreas-CT (3D V-Net backbone; BCP)					
VNet-Adam	12/0	69.96	55.55	1.64	14.27
VNet-Adam	62/0	82.60	70.81	1.33	5.61
BCP-Adam	12/50	82.91	70.97	2.25	6.43
BCP-$Adam_{CO}$	12/50	82.58 (5.37)	70.65 (7.57)	3.11 (2.38)	8.84 (9.24)
BCP-$Adam_{HCO}$	12/50	**83.71 (4.94)**	**72.26 (7.00)**	**2.05 (1.38)**	**5.46 (3.99)**

data regimes following the BCP protocol: **8 labeled / 72 unlabeled** and **4 labeled / 76 unlabeled** samples (Table 1).

Results. On the Left-Atrium MRI dataset, SGD_{HCO} consistently improves overlap and boundary metrics over SGD. For eight labeled samples, SGD_{HCO} improves Dice 89.64→90.75 (p=0.0016) and Jaccard 81.41→83.13 (p=0.0016), while reducing HD95 6.93→5.52 (p=0.0063) and ASD 1.78→1.64 (p=0.2729). For four labeled samples, the Dice improves 88.27→88.83 (p=0.0379) with a corresponding HD95 reduction 8.21→6.55 (p=0.0086).

Pancreas-CT (Adam). Using the standard BCP split (**12 labeled / 50 unlabeled** scans; Table 1), replacing $Adam_{CO}$ with $Adam_{HCO}$ improves Dice 82.58→83.71 (p=0.0152) and Jaccard 70.65→72.26 (p=0.0134), while reducing HD95 8.84→5.46 (p=0.0149) and ASD 3.11→2.05 (p=0.0152).

Study 2 (CMT): Generalization across SSL frameworks and anatomies. We further test transferability by integrating $AdamW_{HCO}$ into the consistency-based CMT framework [21] for fetal lung and liver segmentation. Following CMT, we use a 3D V-Net [16], random flips/crops, and minibatches of eight

Table 2. Fetal MRI segmentation results (mean (std)) for liver and lungs. **AdamW$_{\text{sup}}$**: supervised baseline with 5 labeled scans. **AdamW**: standard semi-supervised baseline. **AdamW$_{\text{CO}}$**: Cautious Optimizer variant [12]. **AdamW$_{\text{HCO}}$**: full Hierarchical Cautious Optimizer (ours). Metrics: Dice/Jaccard (%), Average Surface Distance (ASD, voxels), and 95th percentile Hausdorff Distance (HD95, voxels). Bold highlights the best result for each organ and metric.

Organ	Optimizer	Metrics			
		Dice ↑	Jaccard ↑	ASD ↓	HD95 ↓
Liver	AdamW$_{\text{sup}}$	61.09 (3.66)	47.82 (3.88)	7.15 (3.39)	21.79 (6.73)
	AdamW	70.99 (8.57)	56.10 (9.88)	10.95 (4.15)	31.80 (10.54)
	AdamW$_{\text{CO}}$	63.36 (9.60)	47.48 (9.98)	13.53 (4.15)	37.02 (9.45)
	AdamW$_{\text{HCO}}$	**82.15 (5.40)**	**70.32 (7.49)**	**4.94 (2.27)**	**15.21 (6.91)**
Lungs	AdamW$_{\text{sup}}$	51.06 (8.94)	40.75 (8.73)	9.11 (0.67)	26.22 (2.95)
	AdamW	67.54 (10.24)	52.80 (11.93)	8.12 (3.83)	25.09 (10.25)
	AdamW$_{\text{CO}}$	56.96 (10.71)	41.38 (10.80)	12.29 (4.49)	34.52 (10.00)
	AdamW$_{\text{HCO}}$	**84.17 (6.70)**	**73.54 (9.73)**	**3.56 (2.68)**	**13.30 (10.01)**

patches (4 labeled/4 unlabeled, 144×144×64), training for 200 epochs with LR 1×10^{-4}. To reduce split bias, results are averaged over 16 label/unlabel folds (4 labeled × 4 unlabeled), training from scratch for each.

Results. As shown in Table 2, AdamW$_{\text{HCO}}$ consistently improves both overlap and boundary metrics for lungs and liver over AdamW and AdamW$_{\text{CO}}$, with all gains statistically significant (paired one-sided t-test, $p<10^{-18}$). Qualitatively, AdamW$_{\text{HCO}}$ tends to produce sharper boundaries and fewer false negatives in low-contrast regions (Fig. 2).

Study 3: Hierarchy ablation. We isolate the contribution of HCO's labeled–unlabeled trust hierarchy by comparing base optimizers, cautious optimization (CO) [12], and HCO across all experimental settings. Unlike HCO, CO applies elementwise cautious masking but treats labeled and unlabeled gradients symmetrically, without a source-aware trust hierarchy. Across Left-Atrium MRI, Pancreas-CT, and Fetal-MRI (Tables 1–2), CO provides little or negative benefit over standard optimization, while HCO consistently improves both overlap and surface metrics. On Left-Atrium MRI, CO closely matches SGD (89.76% vs. 89.64% Dice), whereas HCO reaches 90.75%. On Pancreas-CT, CO underperforms both Adam and HCO (82.57% vs. 82.91% and 83.71% Dice), with substantially worse ASD accuracy. On Fetal-MRI, CO performs worst across organs, e.g., lung Dice 56.96% vs. 67.54% for AdamW and 84.17% for AdamW$_{\text{HCO}}$. This demonstrates that cautious masking alone is insufficient and that hierarchical trust is critical to HCO's effectiveness.

Study 4: Loss-ratio dynamics and interpretation. We track the ratio $\mathcal{L}_{\text{U}}/\mathcal{L}_{\text{L}}$ throughout training in Studies 1 and 2 to assess whether unlabeled data exerts sustained optimization influence. In mean-teacher SSL, a rapidly collapsing ratio does not indicate success—it often reflects *trivial agreement*, where the

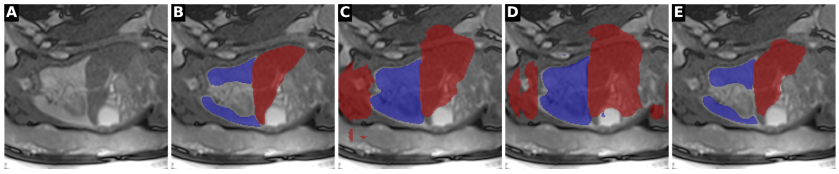
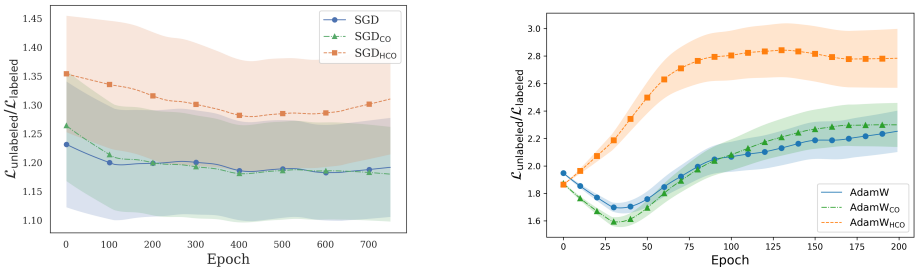


Fig. 2. A. Input MRI, **B.** Ground truth, **C.** AdamW, **D.** AdamW_{CO}, **E.** $\text{AdamW}_{\text{HCO}}$(Ours). Lungs are shown in blue and liver in red.



(a) Left Atrium / BCP (*SGD* family) (b) Fetal MRI / CMT (*AdamW* family)

Fig. 3. Evolution of the unlabeled-to-labeled loss ratio. Each curve is a LOWESS trend line of $\mathcal{L}_{\text{U}}/\mathcal{L}_{\text{L}}$ across epochs; shaded bands show ±SEM computed across steps within each epoch. **(a)** On the LA dataset, SGD_{HCO} (orange) maintains a higher $\mathcal{L}_{\text{U}}/\mathcal{L}_{\text{L}}$ throughout training than standard SGD (blue). **(b)** On fetal MRI, $\text{AdamW}_{\text{HCO}}$ (orange) exhibits a higher ratio than vanilla AdamW (blue) and the caution-only variant (green).

student quickly matches the EMA teacher, making $\mathcal{L}_{\text{U}}$ small but uninformative for generalization.

Ratios are aggregated within each epoch and summarized with SEM across runs (different seeds with a fixed split for BCP; different splits with a fixed seed for CMT), and visualized as a LOWESS trendline of the step-wise ratios (Fig. 3).

Fig. 3 shows that HCO consistently maintains a higher unlabeled-to-labeled loss ratio throughout training compared to baseline optimizers. We attribute this to HCO's gating of conflicting or low-trust unlabeled updates, which reduces shortcut minimization of $\mathcal{L}_{\text{U}}$ via rapid agreement and helps prevent premature collapse of the unsupervised objective. On Left-Atrium MRI with BCP, SGD_{HCO} increases the ratio by about 10% relative to SGD, consistent with sustained engagement of unlabeled supervision. A similar trend is observed on Fetal-MRI under CMT, where $\text{AdamW}_{\text{HCO}}$ preserves a higher and more stable ratio than both AdamW and AdamW_{CO}. The sustained higher ratio under HCO, combined with consistent performance gains across datasets and frameworks, suggests that HCO preserves useful unlabeled supervision rather than allowing it to collapse.

4 Conclusion

We introduced Hierarchical Cautious Optimization (HCO), an optimizer-level approach that enforces an explicit trust hierarchy between labeled and unlabeled gradient signals in semi-supervised learning. HCO forms the momentum direction from labeled gradients and admits unlabeled gradients only when they align with this direction. This principle can be applied to any momentum-based optimizer by modifying only how unlabeled gradients enter the momentum buffer. Across three volumetric medical segmentation datasets and two SSL frameworks, HCO consistently improves performance under low-label regimes, and remains effective with Adam, AdamW, and SGD with momentum. Together, these robust gains and negligible implementation overhead make HCO a strong default optimizer choice when labeled data is scarce. HCO assumes the labeled gradient is a trustworthy reference, and its binary sign-based gate ignores gradient magnitude; its convergence guarantee is likewise conditional on an in-expectation alignment assumption. These assumptions may not hold in every setting.

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

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)
2. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11514–11524 (2023)
3. Bernstein, J., Wang, Y., Azizzadenesheli, K., Anandkumar, A.: SIGNSGD: compressed optimisation for non-convex problems. In: ICML. Proceedings of Machine Learning Research, vol. 80, pp. 559–568. PMLR (2018)
4. Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Pham, H., Dong, X., Luong, T., Hsieh, C., Lu, Y., Le, Q.V.: Symbolic discovery of optimization algorithms. In: NeurIPS (2023)
5. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: ICCV. pp. 9620–9629. IEEE (2021)
6. Cheplygina, V., de Bruijne, M., Pluim, J.P.W.: Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Anal. **54**, 280–296 (2019)
7. Gao, N., Zhou, S., Wang, L., Zheng, N.: PMT: progressive mean teacher via exploring temporal consistency for semi-supervised medical image segmentation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXVIII. Lecture Notes in Computer Science, vol. 15126, pp. 144–160. Springer (2024). https://doi.org/10.1007/978-3-031-73113-6_9, https://doi.org/10.1007/978-3-031-73113-6_9
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (Poster) (2015)

9. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: ICLR (Poster). OpenReview.net (2017)
10. Lee, D.H.: Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. ICML 2013 Workshop : Challenges in Representation Learning (WREPL) (07 2013)
11. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: MICCAI 2020. pp. 552–561. Springer (2020)
12. Liang, K., Chen, L., Liu, B., Liu, Q.: Cautious optimizers: Improving training with one line of code. In: International Conference on Learning Representations (ICLR) (2026), https://openreview.net/forum?id=zBPZeRjfgu, accepted at ICLR 2026; arXiv:2411.16085
13. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (Dec 2017). https://doi.org/10.1016/j.media.2017.07.005, http://dx.doi.org/10.1016/j.media.2017.07.005
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (Poster). OpenReview.net (2019)
15. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8801–8809 (2021)
16. Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571. IEEE Computer Society (2016)
17. Ouali, Y., Hudelot, C., Tami, M.: An overview of deep semi-supervised learning (2020), https://arxiv.org/abs/2006.05278
18. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. CoRR **abs/1904.09237** (2019)
19. Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: MICCAI 2015. pp. 556–564. Springer (2015)
20. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: NIPS. pp. 1163–1171 (2016)
21. Shen, Z., Cao, P., Yang, H., Liu, X., Yang, J., Zaïane, O.R.: Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. In: IJCAI. pp. 4199–4207. ijcai.org (2023)
22. Shi, Y., Zhang, J., Ling, T., Lu, J., Zheng, Y., Yu, Q., Qi, L., Gao, Y.: Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. IEEE Trans. Medical Imaging **41**(3), 608–620 (2022)
23. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E.D., Kurakin, A., Li, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: NeurIPS (2020)
24. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
25. Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J.: Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: MICCAI (5). Lecture Notes in Computer Science, vol. 13435, pp. 34–43. Springer (2022)

26. Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A.K., Yang, X., Heng, P., Ni, D., Li, C., Tong, Q., Si, W., Puybareau, É., Khoudli, Y., Géraud, T., Zhao, J.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical Image Anal. **67**, 101832 (2021)
27. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: MICCAI 2019. pp. 605–613. Springer (2019)
28. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV. pp. 6022–6031. IEEE (2019)
29. Zhao, Z., Wang, Z., Wang, L., Yu, D., Yuan, Y., Zhou, L.: Alternate diverse teaching for semi-supervised medical image segmentation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part V. Lecture Notes in Computer Science, vol. 15063, pp. 227–243. Springer (2024). https://doi.org/10.1007/978-3-031-72652-1_14, https://doi.org/10.1007/978-3-031-72652-1_14
30. Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. Int. J. Comput. Vis. **129**(4), 1106–1120 (2021)