Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PG-HSAM: Prototype-Guided Hierarchical Segment Anything Model for Medical Image Segmentation with Limited Annotations

Guangzhe Zhao

zhaoguangzhe@bucea.edu.cn

Beijing University of Civil Engineering and Architecture

Yongqi Wu

Beijing University of Civil Engineering and Architecture

Xueping Wang

Beijing University of Civil Engineering and Architecture

Teng Cao

Beijing University of Civil Engineering and Architecture

Research Article

Keywords: Medical image segmentation, Segment Anything Model, Prototype learning, Contrastive learning, Limited annotations

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11404752/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PG-HSAM: Prototype-Guided Hierarchical Segment Anything Model for Medical Image Segmentation with Limited Annotations

Guangzhe Zhao[1,2*], Yongqi Wu[1,2], Xueping Wang[1,2], Teng Cao[1,2]

[1]School of Intelligence Science and Technology, Beijing University of Civil Engineering and Architecture, Beijing, 102616, China.

[2]Beijing Key Laboratory of Super Intelligent Technology for Urban Architecture, Beijing University of Civil Engineering and Architecture, Beijing, 102616, China.

*Corresponding author(s). E-mail(s): zhaoguangzhe@bucea.edu.cn;

Abstract

Adapting segmentation foundation models to medical images with limited annotations requires effective use of semantic information without substantially increasing model complexity. We propose PG-HSAM, an extension of H-SAM that introduces prototype prior fusion (PPF) and balanced pixel-to-prototype contrastive regularization (BPC). PPF maintains exponential-moving-average class prototypes, converts pixel-to-prototype similarities into a semantic probability prior, and fuses this prior with the first-stage prediction before hierarchical refinement. BPC regularizes the same feature space with class-balanced pixel sampling. For binary segmentation, foreground and background prototypes are jointly normalized to construct the semantic prior. The prototype memory is non-trainable, and BPC is used only during training. Under the limited-label protocols of H-SAM, PG-HSAM achieves mean Dice scores of 80.49%, 85.20%, and 83.51% on Synapse, LA, and PROMISE12, respectively. Compared with H-SAM, the gains are 0.81, 1.26, and 6.56 percentage points, while Synapse HD95 decreases from 18.35 to 13.70. Ablation and hyperparameter studies further verify the contributions of PPF and BPC and show that the improvement is obtained without increasing the model parameter count.

Keywords: Medical image segmentation, Segment Anything Model, Prototype learning, Contrastive learning, Limited annotations

1 Introduction

Medical image segmentation is fundamental to computer-aided diagnosis, treatment planning, and quantitative image analysis. Convolutional and Transformer-based models, including U-Net [1], nnU-Net [2], TransUNet [3], and Swin-Unet [4], have achieved strong segmentation performance. Their success, however, commonly depends on task-specific training with substantial pixel-level annotations, which are expensive to acquire in many medical applications.

The Segment Anything Model (SAM) [5] offers a large-scale pretrained segmentation backbone and has stimulated extensive research on transferring foundation-model knowledge to medical images. Direct adaptation remains challenging

because natural and medical images exhibit a substantial domain gap and because the original SAM formulation relies on external prompts. Efficient medical adaptations such as SAM-Adapter [6], AutoSAM [7], SAMed [8], and H-SAM [9] reduce the amount of task-specific training while retaining much of the pretrained model. Among them, H-SAM is particularly relevant to limited-label segmentation: its first-stage prediction is reused as a probabilistic prior that guides a second hierarchical decoder. This design provides an explicit prior interface, but the prior itself is generated from the current image and therefore does not directly exploit class-level semantic information accumulated across labeled samples.

Prototype learning provides a compact way to summarize such dataset-level semantics. A class prototype aggregates feature representations from pixels belonging to the same semantic category, while pixel-level contrastive learning can encourage features to approach their corresponding class anchors and separate from competing classes [10, 11]. Motivated by this complementarity, we introduce a shared prototype space into H-SAM and use it for two purposes: constructing a semantic prior for hierarchical refinement and regularizing the feature space from which that prior is derived.

Based on this idea, PG-HSAM introduces two components. Prototype prior fusion (PPF) converts exponential-moving-average class prototypes into a pixel-wise semantic probability map and combines it with the first-stage prediction before the second-stage mask attention. Balanced pixel-to-prototype contrastive regularization (BPC) constrains the same feature space during training with class-balanced pixel sampling. Rather than generating additional prompts or adding a separate prototype prediction branch, PG-HSAM uses the existing prior pathway of H-SAM to inject prototype information into the second-stage decoder. The prototype statistics are stored as non-trainable buffers, and BPC is disabled during inference.

The main contributions are threefold. First, we construct a prototype-derived semantic prior and fuse it with the first-stage prediction through the existing hierarchical refinement pathway of H-SAM. Second, we introduce a class-balanced pixel-to-prototype objective that regularizes the feature space used to build the prototype prior. Third, we extend the formulation to both multi-class CT and binary MRI segmentation and evaluate it on Synapse, LA, and PROMISE12 under limited-label settings, together with ablation, hyperparameter, complexity, and qualitative analyses.

2 Related work

2.1 Medical segmentation and SAM adaptation

Transformer-based segmentation methods such as TransUNet [3], Swin-Unet [4], MISSFormer [12], TransDeepLab [13], and DAE-Former [14] improve global contextual modeling, but they still rely on task-specific supervision. SAM [5] motivates a different direction by transferring large-scale pretrained representations to downstream segmentation tasks. MedSAM [15] adapts SAM with large-scale medical data, whereas SAM-Adapter [6], SAMed [8], and H-SAM [9] focus on more parameter-efficient adaptation. PG-HSAM follows the prompt-free hierarchical design of H-SAM and specifically targets the quality of the prior supplied to its second-stage refinement.

2.2 Prototype-guided representation learning

Prototype learning represents semantic categories with compact feature anchors and has been widely used in few-shot segmentation [16]. Recent SAM-based methods also employ prototypes in different parts of the segmentation pipeline. SurgicalSAM [17] uses prototype-based class prompts, PGP-SAM [18] uses prototypes for automatic prompt generation, and CCF-SAM [19] introduces prototype guidance in a coarse-to-fine framework. In PG-HSAM, the prototype memory is instead converted into a semantic probability prior for the second-stage decoder of H-SAM. BPC uses the same prototype feature space as a training-time constraint. Thus, prototype guidance is incorporated without adding a separate prompt generator or prototype-specific prediction branch.

3 Method

3.1 Overview

PG-HSAM is built on the two-stage hierarchical decoder of H-SAM [9]. The SAM image encoder is adapted with LoRA [20], while the hierarchical decoder produces and refines segmentation probabilities in two stages. Let $\mathbf{P}^{(1)}$ denote the first-stage probability map. H-SAM uses this prediction as a prior for the second decoder; PG-HSAM retains this interface but enriches the prior with class-level semantics accumulated across labeled samples. As illustrated in Fig. 1, a shared prototype memory is updated from normalized pixel features. PPF converts the stored prototypes into a semantic probability map and fuses it with $\mathbf{P}^{(1)}$, whereas BPC regularizes the corresponding feature space during training.

3.2 Prototype memory and prior fusion

Let $\mathbf{f}_i \in \mathbb{R}^C$ denote the pixel feature at spatial position i, and let $\mathbf{z}_i = \mathbf{f}_i/(\|\mathbf{f}_i\|_2 + \epsilon)$ be its ℓ_2-normalized representation. After resizing the ground-truth mask to the feature resolution, the batch prototype of class k is computed as

$$\begin{aligned} \mathbf{q}_k &= \frac{1}{|\Omega_k|} \sum_{i \in \Omega_k} \mathrm{sg}(\mathbf{z}_i), \\ \mathbf{c}_k &\leftarrow m\mathbf{c}_k + (1-m)\mathbf{q}_k. \end{aligned} \tag{1}$$

where Ω_k is the set of pixels belonging to class k, $0 \le m < 1$ is the EMA momentum, and $\mathrm{sg}(\cdot)$ denotes stop-gradient. The update is performed only for classes represented in the current batch. The running prototypes $\mathbf{c}_k$ are stored as non-trainable buffers and therefore do not enlarge the set of optimized model parameters.

For multi-class segmentation, foreground class prototypes are used to construct the prototype prior. For a valid foreground-class set $\mathcal{V}_{\mathrm{fg}}$, the probability assigned to class k at location i is

$$P_{ik}^{p} = \frac{\exp(\mathbf{z}_i^\top \mathbf{c}_k/\tau)}{\sum_{j \in \mathcal{V}_{\mathrm{fg}}} \exp(\mathbf{z}_i^\top \mathbf{c}_j/\tau)}, \qquad k \in \mathcal{V}_{\mathrm{fg}}, \tag{2}$$

where τ is the temperature. Classes whose prototypes have not yet been initialized are excluded from the normalization and receive zero probability. Binary segmentation requires an explicit competition between foreground and background. We therefore maintain both foreground and background prototypes and normalize them jointly:

$$P_{ik}^{p} = \frac{\exp(\mathbf{z}_i^\top \mathbf{c}_k/\tau)}{\sum_{j \in \{0,1\}} \exp(\mathbf{z}_i^\top \mathbf{c}_j/\tau)}, \qquad k \in \{0,1\}. \tag{3}$$

This foreground–background formulation allows the same prototype-guided refinement mechanism to be applied to the LA and PROMISE12 binary segmentation tasks without changing the hierarchical decoder.

The prototype distribution $\mathbf{P}^p$ complements the image-specific first-stage prediction with class-level semantic information. PPF combines the two sources by a convex fusion:

$$\widetilde{\mathbf{P}} = \alpha\mathbf{P}^{(1)} + (1-\alpha)\mathbf{P}^{p}, \tag{4}$$

where $\alpha \in [0,1]$ controls the relative contribution of the original H-SAM prior and the prototype-derived prior. The fused map is renormalized and supplied to the second-stage mask cross-attention inherited from H-SAM. Before all required prototypes have been initialized, the detached first-stage probability map is used directly, preventing unstable early prototype estimates from affecting refinement.

3.3 Balanced pixel-to-prototype regularization

PPF uses prototype similarity at the refinement interface, but the quality of this prior depends on the discriminability of the underlying feature space. BPC therefore regularizes the same normalized features used to construct the prototype prior. Foreground anchors are the detached EMA prototypes. For BPC, a background anchor is computed from the normalized mean of background features in the current batch so that foreground features are contrasted against an explicit competing class during optimization.

Medical segmentation masks are often highly imbalanced: large organs can contribute far more pixels than small structures. Directly averaging the contrastive loss over all pixels would therefore bias optimization toward dominant classes. To limit this effect and keep the additional training

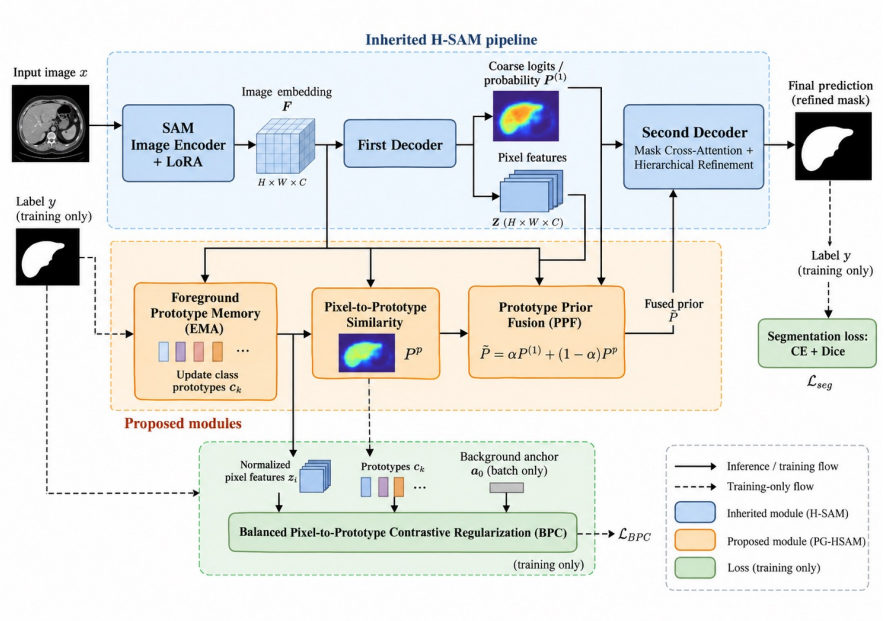


Fig. 1 Overall architecture of PG-HSAM. The shared prototype memory is updated from normalized pixel features. PPF converts the prototypes into a semantic probability prior and fuses it with the first-stage prediction before second-stage refinement, while BPC regularizes the same prototype feature space during training. Prototype statistics are stored as non-trainable buffers.

cost small, at most S pixels are sampled from each represented class. For a sampled pixel i with label y_i, the pixel-to-prototype contrastive loss and its class-balanced aggregation are

$$\ell_i = -\log \frac{\exp(\mathbf{z}_i^\top \mathbf{a}_{y_i}/\tau)}{\sum_{j\in\mathcal{A}} \exp(\mathbf{z}_i^\top \mathbf{a}_j/\tau)}, \qquad \mathcal{L}_{\mathrm{BPC}} = \frac{1}{|\mathcal{C}_B|} \sum_{k\in\mathcal{C}_B} \frac{1}{|\mathcal{S}_k|} \sum_{i\in\mathcal{S}_k} \ell_i, \tag{5}$$

where $\mathcal{A}$ denotes the set of available anchors, $\mathcal{S}_k$ is the sampled pixel set of class k, and $\mathcal{C}_B$ contains the classes represented in the current batch. Equal averaging across represented classes prevents a class from receiving a larger loss weight simply because it occupies more pixels. Gradients propagate through the feature-producing network, whereas all prototype anchors remain detached.

Both H-SAM decoders retain their original cross-entropy and Dice supervision. The BPC term is added to the hierarchical segmentation objective:

$$\mathcal{L} = \mathcal{L}_{\text{H-SAM}} + \lambda_e \mathcal{L}_{\mathrm{BPC}}, \tag{6}$$

where λ_e is linearly warmed up from zero to $\lambda_{\max}$ during the first T epochs. The warm-up reduces the influence of prototype regularization while the feature distribution and prototype memory are still being established. At inference, BPC is disabled; the stored prototypes are used only to construct the PPF prior. Consequently, the proposed regularizer adds no inference-time loss computation or trainable parameters.

4 Experiments

4.1 Experimental setup

We follow the limited-label data splits and preprocessing protocol of H-SAM [9]. For Synapse, the official 18 training and 12 test cases are used, with 221 of the 2,212 training slices retained under the 10% slice-based setting. Eight abdominal organs are evaluated: spleen, right kidney, left kidney, gallbladder, liver, stomach, aorta, and pancreas. For LA, the dataset is split into 80 training and 20 test scans, and the same four labeled training scans used by H-SAM are retained. For PROMISE12, we use the 40/10 training/test split

and the same three labeled training cases as H-SAM. Consistent with the H-SAM protocol, only labeled images are used to train our model; the remaining scans in the semi-supervised dataset formulations are not used.

PG-HSAM and H-SAM use SAM ViT-B with LoRA rank 4 and 512×512 input resolution and are trained for 300 epochs. AdamW is used with an initial learning rate of 0.0025. Unless otherwise stated, PPF uses $\alpha = 0.65$, EMA momentum $m = 0.95$, and temperature $\tau = 0.1$. BPC samples at most $S = 256$ pixels per represented class, with $\lambda_{\max} = 0.1$ and a five-epoch warm-up. Training is performed on two NVIDIA GeForce RTX 3090 GPUs with a batch size of 8. H-SAM and PG-HSAM use the same data split, preprocessing, training budget, checkpoint-selection rule, and inference procedure. Dice and HD95 are used for evaluation; higher Dice and lower HD95 indicate better performance.

4.2 Quantitative comparison

Table 1 summarizes the Synapse results. PG-HSAM achieves 80.49% mean Dice and 13.70 HD95, compared with 79.68% Dice and 18.35 HD95 for H-SAM. This corresponds to a 0.81 percentage-point Dice improvement and a 4.65 absolute reduction in HD95, equivalent to a 25.34% relative decrease. At the organ level, PG-HSAM improves seven of the eight Dice scores, covering the spleen, both kidneys, gallbladder, stomach, aorta, and pancreas. Liver performance remains essentially unchanged (94.18% versus 94.20%), indicating that the aggregate improvement is not driven by sacrificing an already well-segmented large organ.

Table 2 reports the results on the two binary MRI datasets. PG-HSAM reaches 85.20% Dice on LA and 83.51% on PROMISE12, improving over H-SAM by 1.26 and 6.56 percentage points, respectively. No dataset-specific prototype module is introduced for these tasks; only the foreground–background normalization in Eq. (3) is used. The gains therefore indicate that prototype-guided refinement is not limited to the multi-class abdominal CT setting.

4.3 Ablation study

Contribution of PPF and BPC.

Table 3 examines the contribution of PPF and BPC on Synapse. Introducing PPF increases Dice from 79.68% to 80.45% and reduces HD95 from 18.35 to 16.84, showing that a prototype-derived semantic prior improves the second-stage refinement even without the additional contrastive objective. Adding BPC on top of PPF further reduces HD95 to 13.70 while preserving the Dice gain. The different behavior of the two metrics is informative: PPF provides most of the region-overlap improvement, whereas BPC mainly strengthens boundary-sensitive performance by shaping the same feature space from which the semantic prior is constructed.

Hyperparameter sensitivity.

We further study the PPF fusion coefficient α in Eq. (4) and the maximum BPC weight $\lambda_{\max}$ in Eq. (6). When varying one hyperparameter, the other is fixed at its default value. As shown in Table 4, $\alpha = 0.65$ and $\lambda_{\max} = 0.10$ give the best Dice and HD95 among the tested settings. Across the tested α values, Dice varies by only 0.27 percentage points; across the tested $\lambda_{\max}$ values, the range is 0.37 points. The relatively small Dice variation also indicates that the method is not highly sensitive to moderate changes in either hyperparameter, while HD95 shows a clearer preference for the selected values.

4.4 Efficiency analysis

Table 5 compares model size and segmentation performance with several SAM-based methods. The transformer depth and parameter counts of SAMed, SAM-Adapter, and H-SAM are taken from H-SAM [9]. PG-HSAM uses the same four-layer hierarchical decoder as H-SAM and therefore has the same total parameter count of 112.3M. Since the prototype memory is stored as non-trainable buffers and BPC is only used during training, neither component adds model parameters. Under the same parameter count, PG-HSAM improves the mean Dice from 79.68% to 80.49%.

Table 1 Comparison on Synapse using the 10% labeled setting. Higher Dice and lower HD95 are better.

Method	Organ-wise Dice (%)								Mean Dice↑	HD95↓
	Spl	RK	LK	Gal	Liv	Sto	Aor	Pan		
TransUNet [3]	82.63	74.68	81.13	52.97	93.52	69.53	84.07	45.57	73.01	35.21
Swin-Unet [4]	79.43	68.69	76.84	56.13	90.22	64.15	71.57	39.16	68.27	38.98
TransDeepLab [13]	77.16	68.53	77.81	52.05	90.69	63.94	70.04	42.80	67.88	38.40
MISSFormer [12]	83.19	60.98	70.09	59.07	92.57	64.83	77.69	36.19	68.08	41.41
DAE-Former [14]	78.60	66.52	68.58	57.84	91.48	64.42	73.63	40.27	67.67	36.41
AutoSAM [7]	68.70	77.44	76.53	24.87	88.06	52.70	75.19	34.58	62.26	31.67
SAM-Adapter [6]	72.42	68.38	66.77	22.38	89.69	53.15	66.74	26.12	58.28	54.22
SAMed [8]	85.82	82.25	82.62	63.15	92.72	67.20	78.72	52.12	75.57	23.02
H-SAM [9]	89.95	84.30	83.58	67.96	**94.20**	74.70	84.18	58.60	79.68	18.35
PG-HSAM (ours)	**90.66**	**84.64**	**84.33**	**70.19**	94.18	**74.90**	**85.49**	**59.60**	**80.49**	**13.70**

Table 2 Dice (%) on LA and PROMISE12 under the limited-label settings of H-SAM.

Method	LA	PROMISE12
SAMed [8]	82.27	78.55
H-SAM [9]	83.94	76.95
PG-HSAM (ours)	**85.20**	**83.51**

Table 3 Ablation study of PPF and BPC on Synapse.

Configuration	PPF	BPC	Dice (%)↑	HD95↓
H-SAM	×	×	79.68	18.35
H-SAM + PPF	✓	×	80.45	16.84
PG-HSAM	✓	✓	**80.49**	**13.70**

Table 4 Sensitivity to the PPF fusion coefficient α and maximum BPC weight $\lambda_{\max}$ on Synapse. When one hyperparameter is varied, the other is fixed at its default value.

Hyperparameter	Value	Dice (%)↑	HD95↓
	0.55	80.22	15.08
α	**0.65**	**80.49**	**13.70**
	0.75	80.31	14.89
	0.05	80.12	14.04
$\lambda_{\max}$	**0.10**	**80.49**	**13.70**
	0.20	80.13	14.60

Table 5 Comparison of model complexity and segmentation performance on Synapse.

Method	Transformer layers	Total parameters	Mean Dice (%)
SAMed [8]	2	108.8M	75.57
SAMed [8]	4	112.5M	76.80
SAMed [8]	6	116.2M	78.05
SAM-Adapter [6]	2	131.5M	72.80
H-SAM [9]	4	112.3M	79.68
PG-HSAM (ours)	4	112.3M	**80.49**

4.5 Qualitative results

Figure 2 compares representative Synapse predictions from PG-HSAM, H-SAM, and SAMed. PG-HSAM produces more spatially coherent masks and fewer isolated false-positive or false-negative regions in several challenging cases. The improvement is particularly meaningful for structures whose boundaries occupy a small fraction of the image, where modest spatial errors can substantially affect HD95. These observations are consistent with the quantitative results: the improvement in mean Dice is moderate, whereas the reduction in HD95 is considerably larger.

5 Discussion

The experiments show that prototype information is useful for the hierarchical refinement stage of H-SAM. On Synapse, adding PPF improves both Dice and HD95, and seven of the eight organ-wise Dice scores increase. The results on LA and PROMISE12 show that the same formulation also applies to binary MRI segmentation when foreground and background prototypes are normalized jointly.

The ablation study further separates the effects of the two proposed components. PPF accounts for most of the Dice improvement, increasing the score from 79.68% to 80.45%.

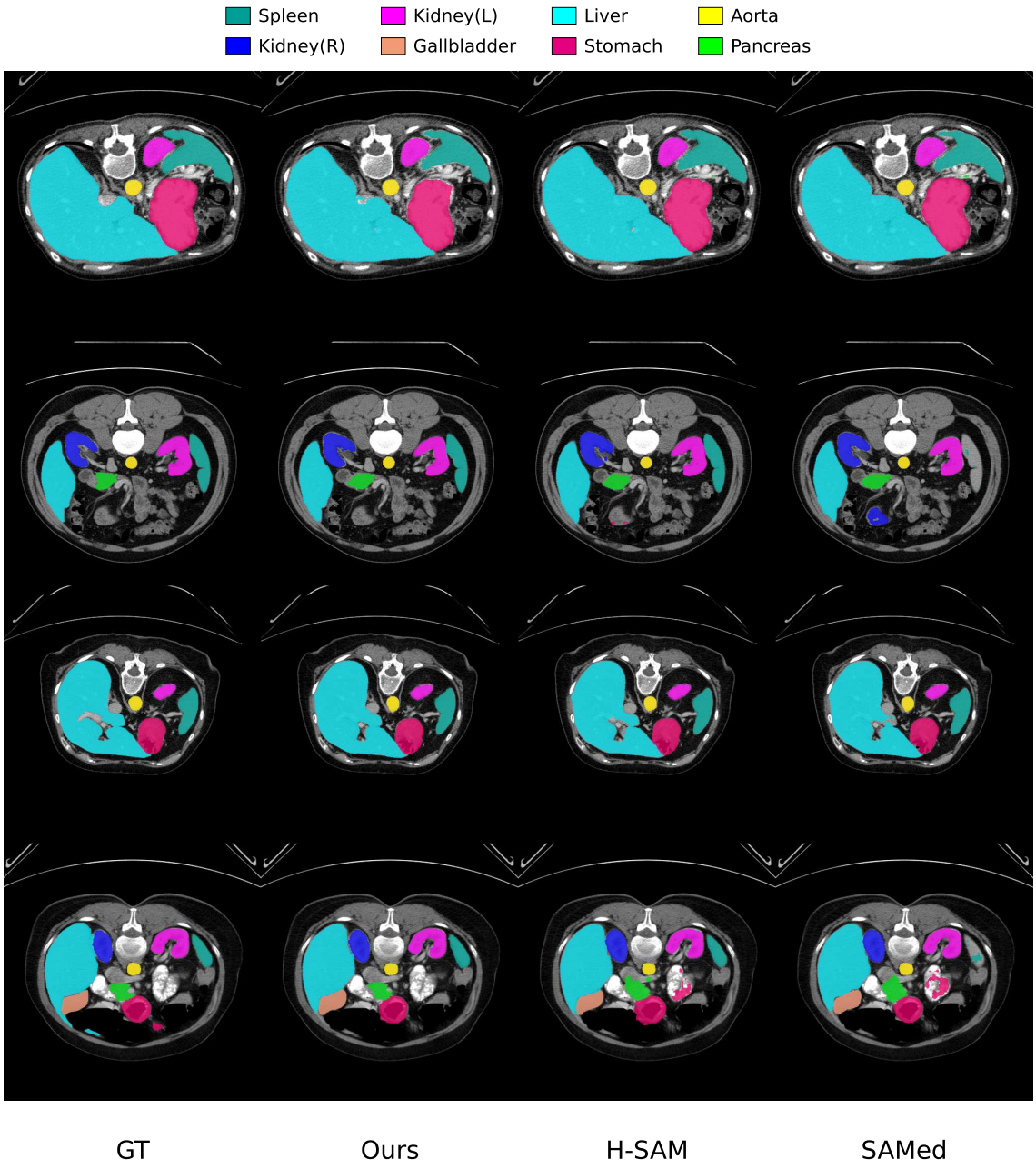


Fig. 2 Qualitative comparison on Synapse. Columns show the ground truth, PG-HSAM, H-SAM, and SAMed predictions. The same cases and class-color mapping are used across methods to facilitate direct visual comparison.

Adding BPC changes Dice only slightly but reduces HD95 from 16.84 to 13.70. This result suggests that the contrastive constraint is especially useful for improving boundary-sensitive performance while maintaining the overlap gain introduced by PPF.

The selected hyperparameters are also reasonably stable within the tested ranges. Dice remains above 80% for all values of α and $\lambda_{\max}$, with $\alpha = 0.65$ and $\lambda_{\max} = 0.10$ giving the best overall result. In addition, PG-HSAM keeps the same 112.3M parameter count as H-SAM because the

prototype memory is non-trainable and BPC does not introduce network layers.

The current study focuses on the H-SAM framework with a ViT-B backbone and limited-label training protocols. Although the method improves results on three datasets, its behavior under other SAM backbones, different annotation budgets, and cross-domain settings has not been examined here. These settings are left for future work.

6 Conclusion

We presented PG-HSAM, which introduces prototype prior fusion and balanced pixel-to-prototype regularization into the hierarchical decoding framework of H-SAM. PPF uses class prototypes to construct a semantic prior for second-stage refinement, while BPC regularizes the same feature space during training. The formulation is applied to both multi-class and binary segmentation without adding model parameters. Under the limited-label protocols of H-SAM, PG-HSAM achieves Dice scores of 80.49%, 85.20%, and 83.51% on Synapse, LA, and PROMISE12, respectively, and reduces Synapse HD95 from 18.35 to 13.70 compared with H-SAM. The ablation and hyperparameter results further confirm the contribution and stability of the proposed components.

Declarations

Funding

This research received no external funding.

References

[1] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28

[2] Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, 203–211 (2021) https://doi.org/10.1038/s41592-020-01008-z

[3] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021) https://doi.org/10.48550/arXiv.2102.04306

[4] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision Workshops, pp. 205–218 (2022). https://doi.org/10.1007/978-3-031-25066-8_9

[5] Kirillov, A., Mintun, E., Ravi, N., *et al.*: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)

[6] Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: Sam-adapter: Adapting segment anything in underperformed scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 3367–3375 (2023). https://doi.org/10.1109/ICCVW60793.2023.00361

[7] Hu, X., Xu, X., Shi, Y.: How to efficiently adapt large segmentation model (sam) to medical images. arXiv preprint (2023)

[8] Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023) https://doi.org/10.48550/arXiv.2304.13785

[9] Cheng, Z., Wei, Q., Zhu, H., Wang, Y., Qu, L., Shao, W., Zhou, Y.: Unleashing the potential of sam for medical adaptation via hierarchical decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3511–3522 (2024). https://doi.org/10.1109/CVPR52733.2024.00337

[10] Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A.C.: Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: Proceedings of the IEEE/CVF International

Conference on Computer Vision, pp. 8219–8228 (2021)

[11] Khosla, P., Teterwak, P., Wang, C., *et al.*: Supervised contrastive learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 18661–18673 (2020)

[12] Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: An effective transformer for 2d medical image segmentation. IEEE Transactions on Medical Imaging **42**(5), 1484–1494 (2023) https://doi.org/10.1109/TMI.2022.3230943

[13] Azad, R., Heidari, M., Shariatnia, M., Aghdam, E.K., Karimijafarbigloo, S., Adeli, E., Merhof, D.: Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation. In: Predictive Intelligence in Medicine, pp. 91–102 (2022). https://doi.org/10.1007/978-3-031-16919-9_9

[14] Azad, R., Arimond, R., Aghdam, E.K., Kazerouni, A., Merhof, D.: Daeformer: Dual attention-guided efficient transformer for medical image segmentation. In: Predictive Intelligence in Medicine, pp. 83–95 (2023). https://doi.org/10.1007/978-3-031-46005-0_8

[15] Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024) https://doi.org/10.1038/s41467-024-44824-z

[16] Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9197–9206 (2019)

[17] Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z.: Surgicalsam: Efficient class promptable surgical instrument segmentation. Proceedings of the AAAI Conference on Artificial Intelligence **38**(7), 6890–6898 (2024) https://doi.org/10.1609/aaai.v38i7.28514

[18] Yan, Z., Yin, Z., Lin, T., Zeng, X., Liang, K., Ma, Z.: Pgp-sam: Prototype-guided prompt learning for efficient few-shot medical image segmentation. In: IEEE International Symposium on Biomedical Imaging, pp. 1–5 (2025). https://doi.org/10.1109/ISBI60581.2025.10980911

[19] Liu, Z., Li, M., Yang, F.: Contrastive coarse-to-fine medical segmentation with prototype guidance and dual-granularity fusion. Neurocomputing **670**, 132603 (2026) https://doi.org/10.1016/j.neucom.2025.132603

[20] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)