Attention-Based Prototype Calibration for Multi-Rater Few-Shot Medical Image Segmentation

Truong Vu, Minh Khoi Ho, and Yutong Xie*

Mohamed bin Zayed University of Artificial Intelligence
{truong.vu, khoi.ho, yutong.xie}@mbzuai.ac.ae

Abstract. Few-shot medical image segmentation methods typically assume a single ground-truth annotation, overlooking systematic variability across expert raters commonly observed in clinical datasets. We propose an attention-based prototype calibration framework for few-shot multi-rater segmentation that models rater-specific deviations from a consensus representation in prototype space. A lightweight yet principled attention operator directly refines rater prototypes without modifying the backbone feature extractor, making the approach fully compatible with existing prototype-based few-shot segmentation methods. This design preserves semantic consistency while enabling personalized segmentation outputs with minimal computational overhead. Experiments on multi-rater medical imaging datasets demonstrate consistent improvements over baseline prototype approaches, highlighting the effectiveness of structured prototype calibration for modeling annotation variability. Our code is available at `https://github.com/truong2710-cyber/JAPC`.

Keywords: Few-shot Learning · Multi-rater · Medical Image Segmentation

1 Introduction

Medical image segmentation underpins many clinical applications, including tumor assessment and organ volumetry [3,19,28]. Although deep neural networks achieve strong performance [2,7,15], they rely on densely annotated data, which is expensive and time-consuming to obtain. Few-shot medical image segmentation (FSMIS) mitigates this issue by learning transferable representations that adapt to unseen classes from only a few labeled support images. Prototype-based approaches [6,8,10,12,13,18,23,24,29], such as PANet [25] and SSL-ALPNet [20], compute class prototypes from support masks and perform dense feature matching for query segmentation, demonstrating strong generalization under limited supervision.

However, most FSMIS methods assume a single reliable annotation per image. In realistic clinical practice, images are often labeled by multiple experts

* Corresponding author: yutong.xie@mbzuai.ac.ae

whose delineations differ systematically due to boundary tolerance and clinical interpretation. Such inter-rater variability, observed in datasets such as CURVAS [21], LIDC-IDRI [1], and QUBIQ [16], reflects structured stylistic differences rather than random noise. While multi-rater segmentation has been studied in fully supervised settings through consensus estimation [26], uncertainty modeling [4,14], and personalized prediction [11,17,27], these approaches assume abundant annotations and are not tailored to few-shot learning.

This reveals a key gap: in practice, supervision is both scarce and heterogeneous. To address this, we introduce *few-shot multi-rater segmentation*, where each support image contains multiple annotations and the model produces R personalized predictions per query, one corresponding to each rater style observed in the support set, thereby avoiding consensus collapse and explicitly modeling multiple plausible segmentation modes.

Collapsing multiple annotations into a consensus discards rater-specific information, whereas training raters independently can not exploit shared semantic structure across annotators nor explicitly model the structured differences between rater styles. Hence, we propose a prototype-centric personalization framework for this setting. An attention-based prototype calibration module models structured deviations between rater-specific and consensus prototypes directly in prototype space. A calibration loss regularizes prototype updates to stabilize personalization under limited support data. To enable multi-style supervision from pseudo labels, we further introduce pseudo-style generation, synthesizing multiple rater-style variants from superpixel-based pseudo labels [9]. Finally, a two-stage training strategy improves optimization stability.

We evaluate our method on two multi-rater benchmarks across modalities: CURVAS (Abdominal CT) and QUBIQ brain-growth (Brain MRI). Using per-rater Dice as the metric, our approach consistently outperforms FSMIS and multi-rater baselines while maintaining robustness in few-shot scenarios.

Our contributions are as follows:

- We formalize the novel problem of *few-shot multi-rater segmentation*, bridging realistic multi-annotator supervision with few-shot learning.
- We propose a prototype-based personalization framework incorporating attention-based prototype calibration, calibration regularization, pseudo-style generation, and two-stage training for solving few-shot multi-rater segmentation by modeling structured inter-rater variability in prototype space.
- We demonstrate consistent empirical improvements in per-rater segmentation accuracy across multiple multi-rater medical benchmarks.

2 Method

2.1 Problem Formulation

Few-shot medical image segmentation (FSMIS) aims to learn a model on a training set $\mathcal{D}_{tr}$ that can segment *unseen* semantic classes in a test set $\mathcal{D}_{te}$ given only a few labeled examples of those unseen classes, *without retraining*. The dataset

is split into $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$ with disjoint class sets $\mathcal{C}_{tr}$ and $\mathcal{C}_{te}$, where $\mathcal{C}_{tr} \cap \mathcal{C}_{te} = \varnothing$. Following the standard meta-learning protocol [20], $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$ are sampled into episodes $\{(\mathcal{S}_i, \mathcal{Q}_i)\}$ for training and evaluation, respectively. Each episode defines an N-way K-shot segmentation task with episode classes $\{c_j\}_{j=1}^{N}$ (typically N=1 in medical FSS): the support set is $\mathcal{S}_i = \{(x_s^l, y_s^l(c_j))\}_{j=1,l=1}^{N,K}$ and the query set $\mathcal{Q}_i$ contains V image-mask pairs from the same class as the support set, where $y_s^l(c_j) \in \{0,1\}^{H\times W}$ is the binary mask for class c_j in an image sized $H \times W$(background c_0 is implicit). A conventional single-label FSMIS model learns to predict one mask $\hat{y}_q^v(c_j)$ for each query image x_q^v $(v = 1, \ldots, V)$ conditioned on $\mathcal{S}_i$. After a series of episodes, we obtain the final model, which is evaluated on unseen $\mathcal{C}_{te}$ in the same N-way K-shot segmentation manner.

In realistic multi-rater datasets, each image may be annotated by multiple experts with systematic stylistic differences. We therefore extend to *few-shot multi-rater segmentation*, where each support image has R masks for each class c_j $\{y_s^{l,(r)}(c_j)\}_{r=1}^{R}$, i.e., $\mathcal{S}_i = \{(x_s^l, \{y_s^{l,(r)}(c_j)\}_{r=1}^{R})\}_{j=1,l=1}^{N,K}$. The class split and episodic protocol remain unchanged, but the goal differs: given x_q^v, the model must output R personalized predictions $\{\hat{y}_q^{v,(r)}(c_j)\}_{r=1}^{R}$, one per rater style observed in the support set, rather than a single consensus segmentation. Following previous works [18,20,24,29], we consider $N = K = 1$.

2.2 Prototype-Based Few-Shot Segmentation

In this section, we present a simple baseline approach to the *few-shot multi-rater segmentation* problem. This approach is a direct extension of existing prototype-based FSMIS methods [18,20,24,29], simply processing each rater independently.

Let $f_\theta(\cdot)$ denote a shared encoder that extracts support and query features $F_s = f_\theta(x_s)$ and $F_q = f_\theta(x_q)$, where x_s and x_q are support and query images from the same episode. Following modern prototype-based few-shot segmentation methods [6,8,10,12,13,18,23,24,29], we abstract prototype generation as $\{p_c^{(r)}\} = \Phi(F_s, y_s^{(r)})$, where $c \in \{\text{fg}, \text{bg}\}$ denotes foreground/background classes and $\Phi(\cdot)$ may generate one or multiple prototypes per class. Each prototype set $p_c^{(r)} \in \mathbb{R}^{N\times D}$ contains N class-specific prototypes of feature dimension D.

A pixel-wise cosine similarity: $\ell_{q,c}^{(r)}(i) = \text{sim}\big(F_q(i), p_c^{(r)}\big)$ is used to compute query logits, where i is the spatial location on the feature map. The foreground logit $\ell_{q,\text{fg}}^{(r)}(i)$ and background logit $\ell_{q,\text{bg}}^{(r)}(i)$ are stacked to form $\ell_q^{(r)}(i)$. Class probabilities are obtained via softmax: $\pi_q^{(r)}(i) = \text{softmax}(\ell_q^{(r)}(i))$, where $\pi_q^{(r)} \in \mathbb{R}^{H\times W\times C}$ and $\pi_{q,c}^{(r)}$ denotes its c-channel ($c \in \{\text{fg,bg}\}$). The final predicted mask for rater r is defined as $\hat{y}_q^{(r)}(i) = \underset{c\in\{\text{fg,bg}\}}{\arg\max}\ \pi_{q,c}^{(r)}(i)$.

2.3 JAPC: Joint Attention-based Prototype Calibration

As mentioned previously, a straightforward baseline to adapt a conventional single-label few-shot model to our novel setting is to treat each rater independently: for each episode, one can process a single rater mask at a time and repeat

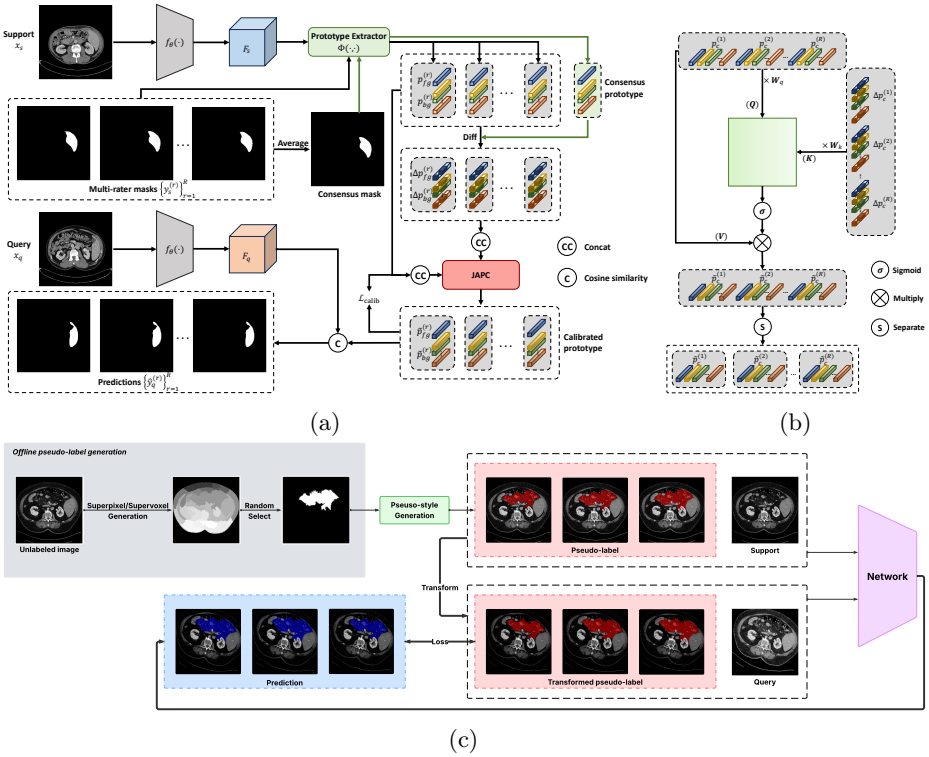


Fig. 1: Overview of the proposed framework. (a) Few-shot multi-rater segmentation pipeline. Rater-specific and consensus prototypes are extracted from support features, and calibrated by the **J**oint **A**ttention-based **P**rototype **C**alibration (**JAPC**) module to generate R personalized predictions via cosine similarity with query features. (b) Internal structure of **JAPC**, modeling structured interactions between rater prototypes and their deviations. (c) Pseudo-style generation. Pseudo labels derived from superpixel/supervoxel segmentation are transformed to synthesize multiple rater-style variants, which are used as support labels and query supervision during training. The "Network" block in (c) corresponds to the full pipeline shown in (a).

this for all R raters. While simple, this strategy ignores structured relationships among raters. In particular, it fails to exploit shared semantic information and does not explicitly model systematic boundary deviations across raters, leading to inefficient learning of stylistic differences, especially in ambiguous regions.

To explicitly model structured inter-rater relationships, we introduce **JAPC**: a **J**oint **A**ttention-based **P**rototype **C**alibration module. This module aims to calibrate each rater's prototypes, highlighting the differences among raters. The module is lightweight and can be plugged into any existing prototype-based

FSMIS method. An overview of the overall framework and the JAPC module is illustrated in Fig. 1a and Fig. 1b, respectively.

First, we compute a consensus mask $y_s^{(0)} = \frac{1}{R}\sum_{r=1}^{R} y_s^{(r)}$, and derive a consensus prototype $p_c^{(0)} = \Phi(F_s, y_s^{(0)})$ where $c \in \{\text{fg}, \text{bg}\}$. For each rater and class, we define deviation $\Delta_c^{(r)} = p_c^{(r)} - p_c^{(0)}$. All rater prototypes are concatenated: $P_c = \text{concat}([p_c^{(1)}, \ldots, p_c^{(R)}])$. The shape of P_c is $N_{\text{total}} \times D$ with N_{total} being the total number of prototype vectors of all raters for class c and D being the prototype dimension.

We apply a shared self-attention module for both background prototypes P_{bg} and foreground prototype P_{fg}: $\mathbf{Q} = P_c W_q, \quad \mathbf{K} = \Delta_c W_k, \quad \mathbf{V} = P_c,$

$$\mathbf{A} = \sigma\Big(\mathbf{Q}\mathbf{K}^\top/\sqrt{D}\Big), \quad \tilde{P}_c = \mathbf{A}\mathbf{V}.$$

This joint formulation allows each rater prototype to attend to the deviations of all other raters, enabling structured modeling of shared and distinctive boundary tendencies. Calibrated prototypes of each rater $\tilde{p}_c^{(r)}$ are separated from $\tilde{P}_c$ and then used for query prediction: $\ell_{q,c}^{(r)}(i) = \text{sim}(F_q(i), \tilde{p}_c^{(r)})$.

To prevent excessive deformation from semantic anchors, we introduce calibration loss: $\mathcal{L}_{\text{calib}} = \frac{1}{R}\sum_{r=1}^{R}\sum_c \|p_c^{(r)} - \tilde{p}_c^{(r)}\|_2^2$.

2.4 Pseudo-Style Generation from Pseudo Labels

Prior work shows that superpixel-derived pseudo labels yield more diverse and generalizable representations than raw annotations [20]; thus, we use them as primary supervision. However, they lack rater-specific variations. To introduce the variations, we generate R style-conditioned variants $\tilde{y}^{(r)} = \mathcal{T}_r(y^{\text{pseudo}})$ from each pseudo label, where $\mathcal{T}_r$ is a random structured boundary transformation. Varying across episodes, these transformations preserve semantics while simulating plausible inter-rater deviations. During training only, we randomly apply 1–2 operations from erosion, dilation, opening/closing, hole filling, small-component pruning, directional boundary shifts, or Gaussian blur with small kernels to support labels and query supervision (Fig. 1c), encouraging the model to learn rater variability as structured boundary deviations around a shared object.

2.5 Two-Stage Training Strategy

Directly training the calibration module jointly with a randomly initialized encoder leads to unstable prototype dynamics under few-shot supervision. We therefore adopt a two-stage training strategy within a total budget of T iterations: the base prototype model is trained without JAPC and pseudo-style for the first $T/2$ iterations, after which they are enabled and training continues for the remaining $T/2$ iterations.

This strategy improves stability because intermediate encoder features already encode meaningful semantic structure, reducing early prototype noise. It also prevents catastrophic drift of prototypes during calibration, enabling effective modeling of subtle stylistic differences at object boundaries.

2.6 Overall Objective

The rater-specific segmentation loss is defined as pixel-wise cross-entropy: $\mathcal{L}_{\text{seg}}^{(r)} = \frac{1}{|\Omega|}\sum_{i\in\Omega}\text{CE}\big(\text{softmax}(\ell_q^{(r)}(i)), y_q^{(r)}(i)\big)$, where Ω denotes spatial locations. The overall segmentation loss is averaged across raters: $\mathcal{L}_{\text{seg}} = \frac{1}{R}\sum_{r=1}^{R}\mathcal{L}_{\text{seg}}^{(r)}$.

Following previous works [20,24,25], we also employ the alignment loss, where the query images act as the support set to predict labels of the support images. Please note that the prototypes in this reverse learning process are also calibrated as in the forward process:

$$\mathcal{L}_{\text{align}} = \frac{1}{R}\sum_{r=1}^{R}\mathcal{L}_{\text{align}}^{(r)} = \frac{1}{R|\Omega|}\sum_{r=1}^{R}\sum_{i\in\Omega}\text{CE}\big(\text{softmax}(\ell_s^{(r)}(i)), y_s^{(r)}(i)\big).$$

The final objective is $\mathcal{L} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{calib}}$.

3 Experiments

3.1 Datasets

We evaluate on two multi-rater medical segmentation datasets spanning different modalities and numbers of raters: **(i) CURVAS** [21] (abdominal CT), termed Abd-CT, contains three organs (kidney (K), pancreas (P), liver (L)), each annotated by three experts. We follow the official split with 20 training and 65 test scans. **(ii) QUBIQ Brain-Growth** [16] (Brain-MRI) includes one class (brain-growth (B)) annotated by seven raters. As the official test set is unavailable, we use the validation set for evaluation, resulting in 34 training and 5 test scans.

3.2 Evaluation Metrics

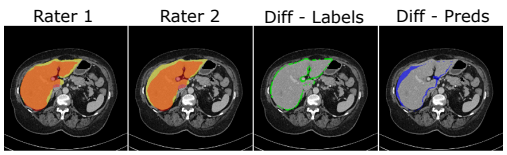
Following previous works[20,24,27], we evaluate segmentation performance using the Dice score (%). Suppose $\text{Dice}_{R_i}^{c}$ is the Dice score (%) of rater i with respect to class c. For each dataset, we report Dice_{m}^{c} (the average Dice across raters for class c), $\text{Dice}_{m}^{R_i}$ (the average Dice across classes for rater R_i), and $\text{Dice}_{m} = \frac{1}{C}\sum_{c}\text{Dice}_{m}^{c}$ (the average Dice across all classes and raters).

3.3 Few-shot Settings

We follow two common few-shot evaluation protocols [20,22,24]. In **Setting 1 (standard)**, the model is trained and tested on all classes without explicit class partitioning; test classes may appear as background during training but are not supervised as foreground. In **Setting 2 (strict unseen-class)**, test classes are entirely excluded from training, and any image containing test-class pixels is removed to ensure true class disjointness. Label partitioning differs accordingly. In Setting 1, pseudo-label training requires no partitioning. In Abd-CT under Setting 2, pancreas and liver often co-occur in 2D slices, so they are grouped (upper abdomen); when testing one group, all its classes are excluded from training. Since Brain-MRI contains only one class, it is evaluated only under Setting 1.

Table 1: Setting 1 results. We report per-class mean Dice on Abd-CT and Brain-MRI. Voxel-based (supervoxel) methods are not applicable to Brain-MRI as images are single-sliced. Best and second best values are **bold** and underlined.

Settings	Method	Abd-CT				Brain-MRI
		Dice_m^K	Dice_m^P	Dice_m^L	Dice_m	Dice_m^B
Full-sup	Backbone	85.02	57.46	91.40	77.96	87.42
Multi-rater	MRNet [11]	56.19	15.20	68.29	46.56	62.35
	D-Persona [27]	64.72	26.67	67.16	53.05	57.20
Few-shot supervoxel	Q-Net [23]	61.06	39.56	57.44	52.69	–
	RPT [29]	63.82	**41.87**	71.66	59.12	–
Few-shot superpixel	SSL-ALPNet [20]	69.33	37.68	77.42	61.48	61.37
	SSL-ALPNet [20] + Ours	**71.04**	39.06	78.26	**62.79**	65.78
	DSPNet [24]	68.03	37.96	78.58	61.52	54.02
	DSPNet [24] + Ours	68.63	39.09	**79.03**	62.25	**66.62**



Method	$\text{Dice}_m^{R_1}$	$\text{Dice}_m^{R_2}$	$\text{Dice}_m^{R_3}$	Dice_m
Q-Net [23]	54.73	54.80	51.85	53.79
RPT [29]	51.59	51.70	49.79	51.03
SSL-ALPNet [20]	57.01	58.08	54.65	56.58
+ **Ours**	59.23	59.95	56.88	58.69
DSPNet [24]	58.62	59.18	56.54	58.11
+ **Ours**	**60.23**	**60.82**	**57.82**	**59.62**

Fig. 2: *(left)* Qualitative example with ground truth (yellow) and predictions (red) and their rater-wise pixel difference (green & blue respectively). *(right)* Per-rater macro-averaged Dice over Kidney/Pancreas/Liver on Abd-CT Setting 2.

3.4 Implementation Details

We compare with representative FSMIS methods, including SSL-ALPNet [20] and DSPNet [24] (superpixel-based), Q-Net [23] and RPT [29] (supervoxel-based), as well as multi-rater methods MRNet [11] and D-Persona [27] trained on the support set. All few-shot methods use identical support–query splits and share the same DeepLabv3-ResNet101 [5] backbone. A fully supervised model is included as an upper bound. We adopt default hyperparameters for all benchmarks. On Abd-CT, FSMIS models are trained for 50k iterations, and ours initializes from the 25k checkpoint following the two-stage schedule. On Brain-MRI, FSMIS models are trained for 10k iterations, and ours starts from 5k. Multi-rater models and fully supervised baseline are trained for 50 epochs each.

3.5 Results

Few-shot performance. Across both settings, our method consistently improves strong few-shot baselines on Abd-CT and Brain-MRI (Tables 1 and 2). On Abd-CT, despite modest absolute gains, improvements are statistically significant: using each test scan once as support gives 65 Dice_m results, and one-tailed paired p-tests show significance in all cases ($p < 0.01$), with p-values of

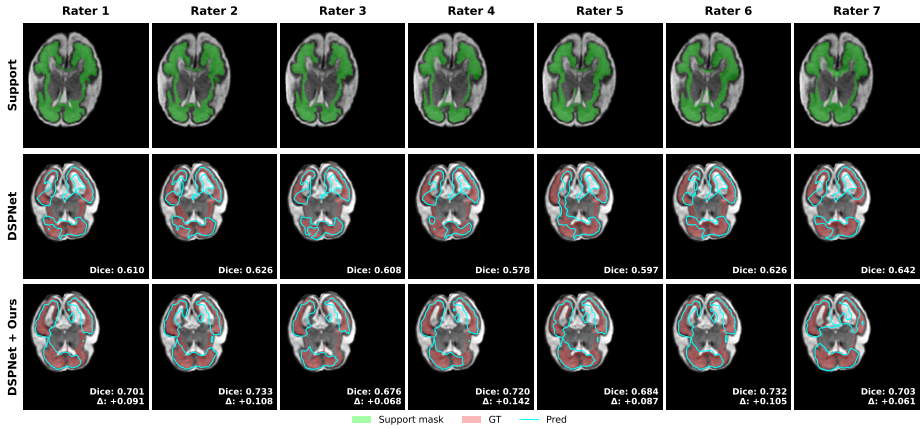


Fig. 3: Visualization of support annotations and segmentation results in Brain-MRI. Our method consistently performs better across all 7 raters.

$5.25\times10^{-19}/2.71\times10^{-24}$ for SSL-ALPNet/DSPNet in Setting 1 and 5.84×10^{-12} $/2.20\times10^{-65}$ in Setting 2. On Brain-MRI, gains are more pronounced, particularly over DSPNet, demonstrating the effectiveness of personalized prototype calibration under multi-rater supervision.

Rater-wise robustness. Improvements are consistent across annotators (Fig. 2 and 3), indicating stable personalization rather than gains limited to specific raters. Qualitative results further show that our predictions better capture shared structures while adapting to boundary variations among raters.

Minimal overhead. Our method adds only 0.2% parameters to SSL-ALPNet and 0.3% to DSPNet, while incurring less than 5% extra computational time.

Ablation study. Table 3 verifies the contribution of each component on Abd-CT Setting 2. Pseudo-style supervision provides a modest improvement, attention and two-stage training further enhance performance, and calibration loss significantly strengthens prototype regularization. Combining all modules yields the best overall result, confirming their complementary effects.

Limitations. A limitation is that performance on pancreas may still lag behind supervoxel-based methods, which benefit from stronger region-level constraints; although pseudo-style supervision reduces fragmentation, it does not always eliminate this gap.

4 Conclusion

We introduced the problem of few-shot multi-rater medical image segmentation, bridging the gap between conventional few-shot learning and realistic multi-

Table 2: Abd-CT Setting 2 with mean Dice for each organ.

Method	Dice_m^K	Dice_m^P	Dice_m^L	Dice_m
Q-Net [23]	47.86	36.33	77.19	53.79
RPT [29]	50.29	**39.88**	62.91	51.03
SSL-ALPNet [20]	58.50	38.23	73.02	56.58
SSL-ALPNet [20] + Ours	61.49	<u>38.24</u>	76.33	<u>58.69</u>
DSPNet [24]	<u>62.24</u>	35.32	<u>76.77</u>	58.11
DSPNet [24] + Ours	**64.10**	36.78	**77.98**	**59.62**

Table 3: Ablation study on Abd-CT Setting 2 of DSPNet.

Pseudo style	Attn	2-stage	Calib loss	Dice_m^K	Dice_m^P	Dice_m^L	Dice_m
×	×	×	×	62.24	35.32	76.77	58.11
✓	×	×	×	62.43	34.21	<u>78.32</u>	58.32
✓	✓	×	✓	**66.00**	34.94	77.05	<u>59.33</u>
✓	✓	✓	×	62.86	<u>36.53</u>	**78.48**	59.29
✓	✓	✓	✓	<u>64.10</u>	**36.78**	77.98	**59.62**

annotator supervision. To address the challenges of heterogeneous annotations under limited support data, we proposed a prototype-based personalization framework with pseudo-style generation, joint attention-based prototype calibration, calibration regularization, and two-stage training. The proposed method explicitly models structured inter-rater variations while preserving shared semantic representations. Experiments on CURVAS and QUBIQ demonstrate consistent improvements in Dice scores across modalities and anatomical structures. Our results highlight the importance of modeling annotation heterogeneity in few-shot regimes and provide a practical foundation for personalized few-shot medical segmentation.

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

References

1. Armato, S.G.I.e.a.: 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), 915–931 (February 2011). https://doi.org/10.1118/1.3528204
2. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10076–10095 (2024)
3. Banerjee, T., Singh, D.P., Kour, P., Swain, D., Mahajan, S., Kadry, S., Kim, J.: A novel unified Inception-U-Net hybrid gravitational optimization model (UIGO) incorporating automated medical image segmentation and feature selection for liver tumor detection. Scientific Reports **15**(1), 29908 (2025)
4. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 119–127. Springer (2019)
5. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
6. Cheng, Z., Wang, S., Xin, T., Zhou, T., Zhang, H., Shao, L.: Few-Shot Medical Image Segmentation via Generating Multiple Representative Descriptors. IEEE Transactions on Medical Imaging **43**(6), 2202–2214 (2024). https://doi.org/10.1109/TMI.2024.3358295

7. Conze, P.H., Andrade-Miranda, G., Singh, V.K., Jaouen, V., Visvikis, D.: Current and emerging trends in medical image segmentation with deep learning. IEEE Transactions on Radiation and Plasma Medical Sciences **7**(6), 545–569 (2023)
8. Ding, H., Sun, C., Tang, H., Cai, D., Yan, Y.: Few-Shot Medical Image Segmentation With Cycle-Resemblance Attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2488–2497 (January 2023)
9. Felzenszwalb, P., Huttenlocher, D.: Efficient Graph-Based Image Segmentation. International Journal of Computer Vision **59**, 167–181 (09 2004). https://doi.org/10.1023/B:VISI.0000022288.19776.77
10. Hansen, S., Gautam, S., Jenssen, R., Kampffmeyer, M.: Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. Medical Image Analysis **78**, 102385 (2022)
11. Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., Zheng, Y.: Learning Calibrated Medical Image Segmentation via Multi-Rater Agreement Modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12341–12351 (June 2021)
12. Jiang, J., Zhang, H.: Concentrate on weakness: mining hard prototypes for few-shot medical image segmentation. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. IJCAI '25 (2025). https://doi.org/10.24963/ijcai.2025/139
13. Kim, H., Hansen, S., Kampffmeyer, M.: Tied Prototype Model for Few-Shot Medical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 651–661. Springer (2025)
14. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems **31** (2018)
15. Kumar, S.: Advancements in medical image segmentation: A review of transformer models. Computers and Electrical Engineering **123**, 110099 (2025)
16. Li, H.B., Navarro, F., Ezhov, I., Bayat, A., Das, D., Kofler, F., Shit, S., Waldmannstetter, D., Paetzold, J.C., Hu, X., et al.: QUBIQ: Uncertainty quantification for biomedical image segmentation challenge. arXiv preprint arXiv:2405.18435 (2024)
17. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Modeling annotator preference and stochastic annotation error for medical image segmentation. Medical Image Analysis **92**, 103028 (2024)
18. Lin, Y., Chen, Y., Cheng, K.T., Chen, H.: Few shot medical image segmentation with cross attention transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 233–243. Springer (2023)
19. Liu, Q., Liu, M., Zhu, Y., Liu, L., Zhang, Z., Wang, Y.: DAUNet: A deformable aggregation UNet for multi-organ 3D medical image segmentation. Pattern Recognition Letters **191**, 58–65 (2025)
20. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-Supervised Learning for Few-Shot Medical Image Segmentation. IEEE Transactions on Medical Imaging **41**(7), 1837–1848 (2022). https://doi.org/10.1109/TMI.2022.3150682
21. Riera-Marín, M., Kleiß, J.M., Aubanell, A., Antolín, A.: CURVAS dataset (Sep 2024). https://doi.org/10.5281/zenodo.13767408, https://doi.org/10.5281/zenodo.13767408

22. Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C.: 'Squeeze & excite'guided few-shot segmentation of volumetric images. Medical image analysis **59**, 101587 (2020)
23. Shen, Q., Li, Y., Jin, J., Liu, B.: Q-net: Query-informed few-shot medical image segmentation. In: Proceedings of SAI Intelligent Systems Conference. pp. 610–628. Springer (2023)
24. Tang, S., Yan, S., Qi, X., Gao, J., Ye, M., Zhang, J., Zhu, X.: Few-shot medical image segmentation with high-fidelity prototypes. Medical Image Analysis **100**, 103412 (2025)
25. 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)
26. Warfield, S., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23**, 903–921 (2004)
27. Wu, Y., Luo, X., Xu, Z., Guo, X., Ju, L., Ge, Z., Liao, W., Cai, J.: Diversified and Personalized Multi-rater Medical Image Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11470–11479 (June 2024)
28. Xia, Q., Zheng, H., Zou, H., Luo, D., Tang, H., Li, L., Jiang, B.: A comprehensive review of deep learning for medical image segmentation. Neurocomputing **613**, 128740 (2025)
29. Zhu, Y., Wang, S., Xin, T., Zhang, H.: Few-shot medical image segmentation via a region-enhanced prototypical transformer. In: International conference on medical image computing and computer-assisted intervention. pp. 271–280. Springer (2023)