RefineSeg: Dual Coarse-to-Fine Learning for Medical Image Segmentation

Anghong Du[1], Nay Aung[4,5], Theodoros N. Arvanitis[1], Stefan K. Piechnik[2], Joao A C Lima[3], Steffen E. Petersen[4,5], and Le Zhang[1,4(✉)]

[1] School of Engineering, College of Engineering and Physical Sciences, University of Birmingham, Birmingham, UK
[2] Oxford Center for Clinical Magnetic Resonance Research (OCMR), Division of Cardiovascular Medicine, John Radcliffe Hospital, University of Oxford, Oxford, UK
[3] Division of Cardiology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
[4] William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University London, London, UK
[5] Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, West Smithfield, London, UK
axd1038@student.bham.ac.uk; l.zhang.16@bham.ac.uk

Abstract. High-quality pixel-level annotations of medical images are essential for supervised segmentation tasks, but obtaining such annotations is costly and requires medical expertise. To address this challenge, we propose a novel coarse-to-fine segmentation framework that relies entirely on coarse-level annotations, encompassing both target and complementary drawings, despite their inherent noise. The framework works by introducing transition matrices in order to model the inaccurate and incomplete regions in the coarse annotations. By jointly training on multiple sets of coarse annotations, it progressively refines the network's outputs and infers the true segmentation distribution, achieving a robust approximation of precise labels through matrix-based modeling. To validate the flexibility and effectiveness of the proposed method, we demonstrate the results on two public cardiac imaging datasets, ACDC and MSCMRseg, and further evaluate its performance on the UK Biobank dataset. Experimental results indicate that our approach surpasses the state-of-the-art weakly supervised methods and closely matches the fully supervised approach. *Our code is available at* `https://github.com/AnghongDu/RefineSeg-MICCAI2025`.

Keywords: Segmentation · Coarse Label · Weakly-Supervise Learning

1 Introduction

The success of deep supervised learning in image segmentation has been largely attributed to the availability of large-scale datasets with accurate pixel-level annotations [23] [3]. However, such annotations are especially costly and time-consuming to acquire in the medical domain, where expert-level annotation is

required and often affected by ambiguous boundaries and inter-observer variability [5] [22]. These challenges are further compounded by strict privacy regulations, making large-scale, high-quality annotation even more difficult. For instance, even for experienced experts, precisely delineating cardiac structures such as the left ventricle (LV), right ventricle (RV), and myocardium (MYO) is highly challenging due to ambiguous boundaries. These inherent annotation uncertainties introduce label noise, making the dataset prone to inconsistencies [22]. As a result, despite large imaging repositories like UK Biobank [13], curating high-quality labels remains a significant challenge, motivating the development of robust learning approaches capable of handling coarse annotations, which is particularly crucial in the medical domain.

To address this challenge, semi-supervised learning (SSL) and weakly supervised learning (WSL) have been widely explored. SSL leverages a small number of labeled samples alongside a large pool of unlabeled data for joint training [9]. While SSL approaches have demonstrated effectiveness in improving model performance, they still require a considerable amount of fully labeled images as supervision. WSL exploit annotations that are easier to obtain than pixel-wise labels, such as bounding boxes [17] [14], scribbles [10] [21] and point labels [20]. Despite their lower annotation cost, WSL suffers from annotation noise, as weak labels often fail to provide precise object boundaries, leading to increased uncertainty during training. For example, in datasets like ACDC, the MYO is embedded within the LV, making it difficult for bounding boxes to isolate the classes precisely. A more effective weak annotation strategy is scribble-based annotation, where annotators simply draw lines and circles within the object of interest (OOI) region to provide guidance. However, such methods often rely on post-processing (e.g., ScribFormer [10] uses random walk propagation that assumes closed-loop strokes) to generate full segmentation masks. In this work, we adopt the coarse annotations that offer more information than scribble labels while avoiding non-target pixels being grabbed into the bounding box. Meanwhile, creating coarse annotations, such as rough OOI and non-OOI boundaries, has a cost similar to that of scribble annotations and can be performed by non-experts. Given these advantages, leveraging computational methods to refine noisy pixels from coarse annotations provides a highly efficient and low-cost approach to enriching large-scale dataset annotations.

Our contribution: We propose a novel weakly supervised segmentation framework that enables end-to-end joint training using both *Positive* (target) and *Negative* (complementary) coarse annotations. Unlike the previous WSL approaches, our method models and disentangles the complex mappings from the input images to the coarse annotations and to the true segmentation distribution simultaneously by introducing transition matrices regularizing. For evaluation, we conduct comprehensive experiments on the ACDC, MSCMRseg, and UK Biobank datasets, achieving segmentation performance that surpasses state-of-the-art weakly supervised methods and closely matches fully supervised approach. Moreover, our method offers a promising pathway for making it feasible

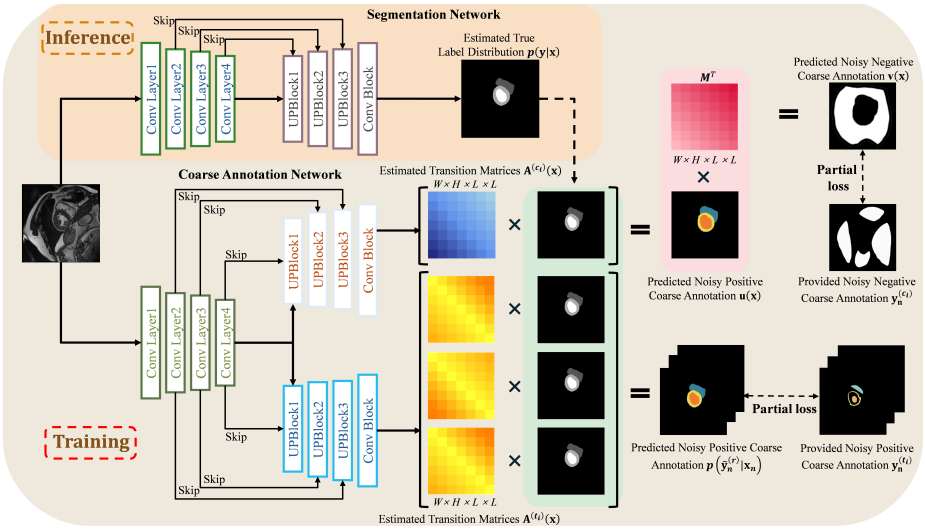

Fig. 1: Overview of our proposed coarse-to-fine segmentation framework that jointly learns from *positive* and *negative* coarse annotations.

to train large medical segmentation models, e.g., MedSAM [11], with minimal manual labeling effort while maintaining high performance.

2 Method

2.1 Problem Set-up

In this work, we address the scenario where a set of images $\{\mathbf{x}_n \in \mathbb{R}^{W\times H\times C}\}_{n=1}^{N}$ (with W, H, C denoting the width, height and channels of the image) are assigned **Positive** (target) and **Negative** (complementary) coarse labels $\{\mathbf{y}_n^{(t_i)}, \mathbf{y}_n^{(c_i)} \in Y^{W\times H}\}_{i=1}^{P}, \quad n = 1, \ldots, N$. Here, P denotes the total number of annotation strategies, N represents the total number of images in the dataset, and $Y = \{1, 2, \ldots, L\}$ denotes the set of possible classes. Figure 1 illustrates our proposed end-to-end joint training framework. Our problem can be formulated as estimating the unobserved true segmentation distribution $p(\mathbf{y}_n|\mathbf{x}_n)$ from the dataset $\mathcal{D} = \{\mathbf{x}_n, (\mathbf{y}_n^{(t_i)}, \mathbf{y}_n^{(c_i)})\}_{i=1}^{P}$ with multiple coarse drawing labels.

2.2 Joint Training with Multiple Coarse Annotations

In this section, we describe how to jointly learn the true segmentation distribution $p(\mathbf{y}_n|\mathbf{x}_n)$ alongside the transition matrix $A^{(t_i)}$ and $A^{(c_i)}$ from multiple coarse annotation networks. In short, we minimize the joint training loss functions of the probability model using the observed positive and negative coarse labels. A detailed description is provided below.

Pixel-wise Transition Matrix. Different from traditional methods that assume all images share a same transition matrix [7], our approach leverages the independent pixel-wise transition matrix [22] to refine segmentation predictions for each input image. Our transition matrix is built on the *Markov chain transition assumption* [4], which ensures the current segmentation state depends only on its immediate previous state. In particular, we refer to the $L \times L$ matrix, where each (m,k)-th element is defined by : $A^{(r)}(\mathbf{x},u,v)_{mk} := p(\tilde{\mathbf{y}}_{uv}^{(r)} = m \mid \mathbf{y}_{uv} = k, \mathbf{x}), \quad \forall m,k \in \{1,\ldots,L\}$, as the transition matrix at pixel (u,v) in image $\mathbf{x}$, $r \in \{t_i, c_i\}$ represents the coarse annotation strategies.

Given an image $\mathbf{x}_n$, under the assumption that annotations at different pixels are conditionally independent, the probability of the observed coarse labels on each pixel (u,v) can be formulated as:

$$p(\tilde{\mathbf{y}}_n^{(r)}(u,v)|\mathbf{x}_n) = \sum_{\mathbf{y}_n \in Y} A^{(r)}(u,v) \ \cdot p(\mathbf{y}_n(u,v)|\mathbf{x}_n), \tag{1}$$

where $p(\mathbf{y}_n(u,v)|\mathbf{x}_n)$ represents the predicted fine-grained label distribution, and $p(\tilde{\mathbf{y}}_n^{(r)}(u,v)|\mathbf{x}_n)$ represents the predicted coarse label distribution. Annotation network estimates the pixel-wise transition matrices $\left\{A^{(r)}(\mathbf{x}) \in [0,1]^{W\times H\times L\times L}\right\}_{r=1}^{P}$ for input image $\mathbf{x}$. Equation (1) describes the probabilistic transition process in which annotation network r adjusts $p(\mathbf{y}_n(u,v)|\mathbf{x}_n)$ to align with the coarse labels.

Learning with positive coarse label. Given a training input $\mathbf{x}_n$ and a positive coarse label $\mathbf{y}_n^{(t_i)}$, we optimize the transition matrix $A^{(t_i)}$ of the coarse annotation network by minimizing the following hybrid loss function:

$$\mathcal{L}_{pos}^{(t_i)} = \sum_{i=1}^{P_{pos}} \left[\alpha_i \mathcal{L}_{ce}^{(t_i)} + \beta_i \mathcal{L}_{dice}^{(t_i)}\right], \tag{2}$$

where $\mathcal{L}_{ce}^{(t_i)}$ and $\mathcal{L}_{dice}^{(t_i)}$ denote the *cross-entropy loss* and *Dice loss*, which together form the hybrid loss. P_{pos} represents the set of positive annotation strategies within P. α_i and β_i are weight parameters that balance the contribution of each loss term. Minimizing Equation (2) encourages the transition matrix $A^{(t_i)}$ adjusted segmentation probability map $p(\mathbf{y}_n|\mathbf{x}_n)$ to align closely with the provided positive coarse label $\mathbf{y}_n^{(t_i)}$. However, directly applying $\mathcal{L}_{ce}^{(t_i)}$ and $\mathcal{L}_{dice}^{(t_i)}$ to the entire image is ineffective due to severe class imbalance. The CE loss can be minimized by predicting all pixels as the most frequent background class. Although Dice loss mitigates class imbalance, annotation noise in coarse labels makes unannotated background regions unreliable, as they may still contain target information.

To address this issue, [18] proposed to restrict loss computation to only the annotated pixels while ignoring unverified regions. Building on this idea, we design the partial loss function, formulated as:

$$\mathcal{L}_{ce}^{(t_i)} = -\frac{1}{|\mathcal{R}_{pos}^{(t_i)}|} \sum_{(u,v)\in\mathcal{R}_{pos}^{(t_i)}} \mathbf{y}_n^{(t_i)}(u,v) \log\left[A^{(t_i)} p(\mathbf{y}_n|\mathbf{x}_n)(u,v)\right], \tag{3}$$

$$\mathcal{L}_{dice}^{(t_i)} = 1 - \frac{2\sum_{(u,v)\in\mathcal{R}_{pos}^{(t_i)}} A^{(t_i)} p(\mathbf{y}_n|\mathbf{x}_n)(u,v)\mathbf{y}_n^{(t_i)}(u,v)}{\sum_{(u,v)\in\mathcal{R}_{pos}^{(t_i)}} \left(A^{(t_i)} p(\mathbf{y}_n|\mathbf{x}_n)(u,v) + \mathbf{y}_n^{(t_i)}(u,v)\right)}, \tag{4}$$

where $\mathcal{R}_{pos}^{(t_i)}$ represents the set of pixels labeled as positive in $\mathbf{y}_n^{(t_i)}$, excluding negative pixels.

Learning with negative coarse label. For some situations, it is easier to provide negative coarse labels to help the model predict the true label distribution. However, if we directly apply loss, as in Equation (2), when learning with these negative coarse labels, the model can only learn a mapping $\mathbb{R} \to Y$ that attempts to predict the conditional probability $p(\tilde{\mathbf{y}}_n^{(c_i)}|\mathbf{x}_n)$ and the corresponding negative pixel that predicts $\tilde{\mathbf{y}}_n^{(c_i)}(u,v)$ for a input image $\mathbf{x}_n$.

To address this issue, inspired by [19] [22], which summarizes all the probabilities into a transition matrix $M \in \mathbb{R}^{L\times L}$, where $M_{mk}(u,v) = p(\tilde{\mathbf{y}}_n^{(c)}(u,v) = m|\mathbf{y}_n(u,v) = k, \mathbf{x}_n)$ and $M_{mm}(u,v) = 0$, we introduce a transition-based negative learning approach. Here, M_{mk} denotes the entry in the m-th row and k-th column of M, representing the probability of flipping the true label k into the complementary label m. We achieve this by introducing a linear transformation layer in the negative coarse label learning channel. This layer outputs $\mathbf{v}(\mathbf{x}_n)$ by multiplying the output of the coarse annotation network $A^{(c_i)}p(\mathbf{y}_n|\mathbf{x}_n)$, denoted as $\mathbf{u}(\mathbf{x}_n)$, with the transposed transition matrix $M^\top$.

We also observe that $p(\tilde{\mathbf{y}}^{(c)}|\mathbf{x})$ can be transformed into $p(\tilde{\mathbf{y}}^{(t)}|\mathbf{x})$ using the transition matrix M,

$$p(\tilde{\mathbf{y}}_{uv}^{(t)} = k \mid \mathbf{x}_n) = \sum_{m\neq k} p(\tilde{\mathbf{y}}_{uv}^{(t)} = k \mid \tilde{\mathbf{y}}_{uv}^{(c)} = m, \mathbf{x}_n) p(\tilde{\mathbf{y}}_{uv}^{(c)} = m \mid \mathbf{x}_n). \tag{5}$$

To enable end-to-end learning rather than transferring after training, we define:

$$\mathbf{v}(\mathbf{x}_n) = M^\top \mathbf{u}(\mathbf{x}_n). \tag{6}$$

Here, we apply the transposed transition matrix $M^\top$ to ensure that the learned distribution $\mathbf{v}(\mathbf{x}_n)$ aligns with the negative coarse label. Then, $\mathcal{L}_{neg} = \sum_{i=1}^{P_{neg}} (M^\top (A^{(c_i)} p(\mathbf{y}_n|\mathbf{x}_n)), \mathbf{y}_n^{(c_i)})$, where P_{neg} represents the set of positive annotation strategies within P.

Regularizing for the Transition Matrices. In the joint training process, the model may rely excessively on the transition matrix to adjust $p(\mathbf{y}_n|\mathbf{x}_n)$ to fit the coarse labels rather than making subtle refinements to enhance the quality of the segmentation distribution. Existing studies [22] employ trace constraints to mitigate overfitting to coarse labels. However, negative coarse labels differ significantly from positive labels, making trace constraints insufficient to prevent unstable optimization caused by annotation heterogeneity. To address this, we propose a identity matrix regularization term to stabilize the transition matrices:

$$\mathcal{L}_{reg} = \sum_{i=1}^{P_{pos}} ||A^{(t_i)} - I||_F^2, \tag{7}$$

where I is the identity matrix, $||\cdot||_F^2$ denotes the Frobenius norm. This regularization preserves the structural integrity of segmentation predictions while allowing necessary refinements, preventing transition matrices from learning trivial mappings that overfit coarse labels instead of capturing meaningful features.

Finally, we combine the positive annotation loss $\mathcal{L}_{pos}$ and the negative annotation loss $\mathcal{L}_{neg}$ as our objective and optimize the following:

$$\mathcal{L}_{total} = \sum_{i=1}^{P_{pos}} \mathcal{L}_{pos}(A^{(t_i)} p(\mathbf{y}_n|\mathbf{x}_n), \mathbf{y}_n^{(t_i)}) + \sum_{i=1}^{P_{neg}} \mathcal{L}_{neg}(M^\top (A^{(c_i)} p(\mathbf{y}_n|\mathbf{x}_n)), \mathbf{y}_n^{(c_i)}) + \lambda \mathcal{L}_{reg}, \tag{8}$$

where λ is weight parameter for the regularization term.

3 Experiments

Dataset. Two public cardiac datasets, ACDC [1] and MSCMRseg [24] [6], along with UK Biobank (UKBB) [13] dataset are adopted to evaluate our method. ACDC contains cine MRI scans of 100 patients, MSCMRseg includes LGE MRI scans of 45 cardiomyopathy patients, and UKBB cardiac dataset comprises short-axis CMR images from 600 subjects. All three datasets are provided with ground-truth annotations meticulously performed by experienced cardiovascular imaging specialists. To obtain positive-negative coarse annotations, we erode the available segmentation masks for ACDC and MSCMRseg datasets, following the approach in [2]. For UKBB, we obtain the realistic coarse annotations by manually annotating the data following the principles in [15]. We also follow the approaches in [20], [17] and [14] to obtain the box and point annotations on ACDC and MSCMRseg datasets, and obtain scribbles on UKBB dataset for the comparison experiments. Across all datasets, we uniformly use 80% of each dataset for training and 20% for testing. Note that, during training, only coarse annotations are adopted in our framework.

Implementation Settings. Our framework was implemented in PyTorch and employed the 2D U-Net [12] as the network architecture. For all datasets, we resized or padded all images to a uniform size of 224 $\times$ 224 pixels. For data augmentation, zero-mean and unit-variance normalization, random flipping, and random rotation were applied. Before being input to the model, each image is normalized using min-max scaling to bring pixel values into the $[0, 1]$ range. The optimizer used was AdamW, with an initial learning rate of 1e-3 and a weight decay of 2e-5. In Equation (2), we empirically set $\alpha = 0.6$ and $\beta = 0.4$. For Equation (8), we set $\lambda = 0.2$. All models were trained using one single NVIDIA A100 40GB GPU for 200 epochs.

Baseline settings and Evaluation metrics. We conduct our experiments under the assumption that no ground truth labels are available. Specifically, we compare multiple weakly supervised approaches that leverage scribble annotations [21] [10], coarse annotations [16], box-level annotations [14] [17], and point labels [20]. Additionally, we include a semi-supervised method [9] that utilizes

Table 1: The results of Dice score on ACDC and MSCMRseg datasets. Bold denotes the best performance among all methods except nnUNet. Numbers in bold indicate the best method that statistically ($p < 0.01$) better than other methods by computing the p values of paired t-tests on Dice score.

Methods	**Annotations**	ACDC				MSCMRseg			
		LV	MYO	RV	Avg	LV	MYO	RV	Avg
Weakly-supervised									
PA-Seg [20]	points	.841	.723	.729	.764	.771	.609	.534	.638
WeakPolyp [17]	box	.836	.718	.632	.728	.767	.566	.516	.615
BoxInst [14]	box	.803	.674	.583	.686	.747	.503	.494	.581
ScribFormer [10]	scribbles	.922	.871	.871	.888	.896	.813	.807	.839
CycleMix [21]	scribbles	.883	.798	.863	.848	.870	.739	.791	.800
LC-MIL [16]	coarse	.873	.684	.561	.706	.723	.537	.520	.593
Ours	coarse	**.938**	**.884**	**.881**	**.901**	**.904**	**.839**	**.812**	**.852**
Semi-supervised									
PointWSSIS [9]	5%mask+point	.901	.777	.807	.828	.844	.748	.705	.765
Fully supervised									
nnUNet [8]	mask	.943	.901	.915	.920	.944	.882	.880	.902

point annotations, alongside the fully supervised nnU-Net [8], which represents the state-of-the-art in cardiac segmentation when trained with ground truth labels. To evaluate segmentation performance, we use the Dice score for RV, LV, and MYO. In addition, we report the average Dice across these three regions to provide a comprehensive measure of segmentation accuracy.

Quantitative Comparison. Table 1 compares the Dice performance of models trained with different supervision strategies and backbones on the ACDC and MSCMRseg datasets, including results reported in [10] and [21]. Our proposed framework outperforms multiple weakly supervised approaches. Specifically, in the first section, our method surpasses the ScribFormer [10] by 1.4% in average Dice on ACDC dataset (0.902 vs. 0.888) and by 1.3% on MSCMRseg dataset (0.852 vs. 0.839). The second and third sections of Table 1 further present the comparison of our framework with semi-supervised [9] and fully-supervised [8] methods. Results indicate that our approach outperforms semi-supervised training approach with partial masks and achieves performance close to full supervision approach. This demonstrates the effectiveness of jointly training with both *positive* and *negative* coarse annotations. The inclusion of negative coarse annotations further enhances the model's ability to extract positive features while imposing stronger regularization, leading to more robust feature representation. Figure 4 highlights our model not only achieves a higher average Dice score but also exhibits greater performance stability compared to other weakly supervised approaches. Notably, our study is the first weakly supervised benchmark compared to fully supervised approach on the UKBB dataset, providing a valuable reference for future research in this area.

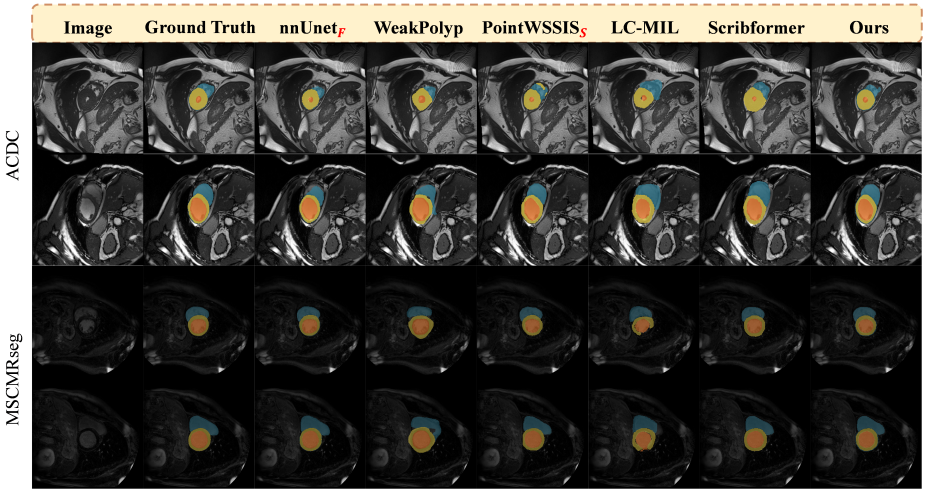

Fig. 2: Qualitative comparisons of our framework with state-of-the-art weakly supervised methods on the ACDC and MSCMRseg datasets. Subscripts F and S indicate segmentation models trained with full supervision and semi-supervision.

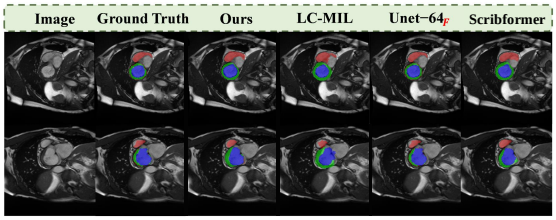

Fig. 3: Visualization of the segmentation performance of different supervision strategies on UKBB datasets. Subscripts F denote segmentation models trained with full supervision.

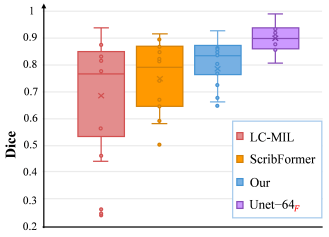

Fig. 4: The Dice distribution of different supervision strategies on the UKBB dataset.

Qualitative Comparison. Figure 2 presents the segmentation visualization of different methods on the ACDC and MSCMRseg datasets, and Figure 3 shows the performance of the approaches with different supervision strategies on UKBB dataset. As shown in Figure 2, both WeakPolyp and LC-MIL struggle to accurately preserve structural shape. In contrast, our approach effectively integrates and optimizes features from different coarse annotations, resulting in a more comprehensive representation. This capability mitigates the inherent limitations of Unet, which tends to focus primarily on localized features. Moreover, compared to the weakly supervised results shown in Figure 3, our approach not only produces results that are closer to the ground truth but also maintains shape integrity comparable to fully supervised method.

4 Conclusion

Pixel-level annotation remains a major challenge in medical image segmentation, constraining further progress in this field. To address this challenge, we propose a novel weakly supervised segmentation framework that enables end-to-end training using *positive-negative* coarse annotations by introducing transition matrices. Experimental results show our method outperforms state-of-the-art weakly supervised approaches and closely matches fully supervised model. By reducing annotation costs without compromising performance, our method enhances efficiency and underscores the potential of weakly supervised learning for cost-effective, high-precision segmentation. Furthermore, it presents a promising approach for facilitating the training of large medical segmentation models, including MedSAM [11].

Disclosure of Interests

The authors declare that there are no competing interests related to this work.

References

1. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
2. Castro, D.C., Tan, J., Kainz, B., Konukoglu, E., Glocker, B.: Morpho-mnist: Quantitative assessment and diagnostics for representation learning. Journal of Machine Learning Research **20**(178), 1–29 (2019)
3. Deng, R., Yao, T., Tang, Y., Guo, J., Lu, S., Xiong, J., Yu, L., Cap, Q.H., Cai, P., Lan, L., et al.: Kpis 2024 challenge: Advancing glomerular segmentation from patch-to slide-level. Medical Image Analysis (2025)
4. Gagniuc, P.A.: Markov chains: from theory to implementation and experimentation. John Wiley & Sons (2017)
5. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. Medical Image Analysis **89**, 102889 (2023)
6. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. Medical Image Analysis **89**, 102889 (2023)
7. Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., Welling, M.: Argmax flows and multinomial diffusion: Learning categorical distributions. Advances in Neural Information Processing Systems **34**, 12454–12465 (2021)
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
9. Kim, B., Jeong, J., Han, D., Hwang, S.J.: The devil is in the points: Weakly semi-supervised instance segmentation via point-guided mask representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11360–11370 (2023)

10. Li, Z., Zheng, Y., Shan, D., Yang, S., Li, Q., Wang, B., Zhang, Y., Hong, Q., Shen, D.: Scribformer: Transformer makes cnn work better for scribble-based medical image segmentation. IEEE Transactions on Medical Imaging (2024)
11. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
13. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al.: Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine **12**(3), e1001779 (2015)
14. Tian, Z., Shen, C., Wang, X., Chen, H.: Boxinst: High-performance instance segmentation with box annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5443–5452 (2021)
15. Valvano, G., Leo, A., Tsaftaris, S.A.: Learning to segment from scribbles using multi-scale adversarial attention gates. IEEE Transactions on Medical Imaging **40**(8), 1990–2001 (2021)
16. Wang, Z., Saoud, C., Wangsiricharoen, S., James, A.W., Popel, A.S., Sulam, J.: Label cleaning multiple instance learning: Refining coarse annotations on single whole-slide images. IEEE transactions on medical imaging **41**(12), 3952–3968 (2022)
17. Wei, J., Hu, Y., Cui, S., Zhou, S.K., Li, Z.: Weakpolyp: You only look bounding box for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 757–766. Springer (2023)
18. Wen, H., Cui, J., Hang, H., Liu, J., Wang, Y., Lin, Z.: Leveraged weighted loss for partial label learning. In: International conference on machine learning. pp. 11091–11100. PMLR (2021)
19. Yu, X., Liu, T., Gong, M., Tao, D.: Learning with biased complementary labels. In: Proceedings of the European conference on computer vision (ECCV). pp. 68–83 (2018)
20. Zhai, S., Wang, G., Luo, X., Yue, Q., Li, K., Zhang, S.: Pa-seg: Learning from point annotations for 3d medical image segmentation using contextual regularization and cross knowledge distillation. IEEE transactions on medical imaging **42**(8), 2235–2246 (2023)
21. Zhang, K., Zhuang, X.: Cyclemix: A holistic strategy for medical image segmentation from scribble supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11656–11665 (2022)
22. Zhang, L., Tanno, R., Xu, M.C., Jin, C., Jacob, J., Cicarrelli, O., Barkhof, F., Alexander, D.: Disentangling human error from ground truth in segmentation of medical images. Advances in Neural Information Processing Systems **33**, 15750–15762 (2020)
23. Zhang, L., Wu, F., Bronik, K., Papiez, B.W.: Diffuseg: Domain-driven diffusion for medical image segmentation. IEEE Journal of Biomedical and Health Informatics (2025)
24. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE transactions on pattern analysis and machine intelligence **41**(12), 2933–2946 (2018)