

Diversified and Personalized Multi-rater Medical Image Segmentation

Yicheng Wu^{1(✉),*}, Xiangde Luo^{2,*}, Zhe Xu³, Xiaoqing Guo⁴, Lie Ju¹, Zongyuan Ge¹,
 Wenjun Liao², Jianfei Cai¹

¹Monash University, ²University of Electronic Science and Technology of China,

³The Chinese University of Hong Kong, ⁴University of Oxford

Abstract

Annotation ambiguity due to inherent data uncertainties such as blurred boundaries in medical scans and different observer expertise and preferences has become a major obstacle for training deep-learning based medical image segmentation models. To address it, the common practice is to gather multiple annotations from different experts, leading to the setting of **multi-rater medical image segmentation**. Existing works aim to either merge different annotations into the “groundtruth” that is often unattainable in numerous medical contexts, or generate diverse results, or produce personalized results corresponding to individual expert raters. Here, we bring up a more ambitious goal for multi-rater medical image segmentation, i.e., obtaining both diversified and personalized results. Specifically, we propose a two-stage framework named **D-Persona** (first **Diversification** and then **Personalization**). In Stage I, we exploit multiple given annotations to train a Probabilistic U-Net model, with a bound-constrained loss to improve the prediction diversity. In this way, a common latent space is constructed in Stage I, where different latent codes denote diversified expert opinions. Then, in Stage II, we design multiple attention-based projection heads to adaptively query the corresponding expert prompts from the shared latent space, and then perform the personalized medical image segmentation. We evaluated the proposed model on our in-house Nasopharyngeal Carcinoma dataset and the public lung nodule dataset (i.e., LIDC-IDRI). Extensive experiments demonstrated our D-Persona can provide diversified and personalized results at the same time, achieving new SOTA performance for multi-rater medical image segmentation. Our code will be released at <https://github.com/ycwu1997/D-Persona>.

* Equal Contribution. Correspondence to yicheng.wu@monash.edu. This research is supported by the Monash FIT Start-up Grant.

1. Introduction

To build up a powerful computer-aided diagnosis (CAD) system, automatic medical image segmentation is an indispensable step in providing quantitative measurements for interested medical targets [13]. Despite deep-learning based medical image segmentation methods have achieved great progress [9, 12, 46], their deployment in clinical settings [4] remains skeptical since it is still hard for them to accurately segment challenging targets such as the primary gross tumor volume (GTVp) [26, 28], blurred lesions [6, 47] and complex anatomical structures [10, 22].

One key reason lies in the “**annotation ambiguity**”, which mainly comes from two aspects: *data-level uncertainties* and *observer-level preferences*. The first aspect refers to the inherent uncertainties in the data such as irregular shapes and varied locations of a medical target (e.g., MS lesions [6]) due to its great variability, and blurred boundaries around the target regions caused by limited imaging resolutions in medical scans. These inherent data uncertainties unavoidably affect the process of expert annotations [30], resulting in inconsistent training labels that are likely to confound data-driven segmentation models. The second aspect refers to the diversity in domain expertise and personal preferences among annotators, which leads to a range of interpretations for a single target. In other words, there is no “absolutely correct label”, which we call **meta segment** in this work. It has been shown in numerous medical contexts that the meta segment is unattainable [24].

To mitigate the “**annotation ambiguity**”, a standard practice in medical settings is to gather multiple annotations from different observers [28, 34], leading to multi-rater medical image segmentation [15] (see Fig. 1 Top-left). The existing works dealing with medical multi-rater annotations can be generally categorized into three groups: crowdsourcing methods, generation methods, and personalization methods, each aligning with a different objective: becoming meta expert, implicit experts, or explicit experts (Fig. 1 Top-right). In particular, the crowdsourcing methods [44] (Fig. 1 Middle-left) assume there exists one and only one meta segment and combining multiple annotations can ap-

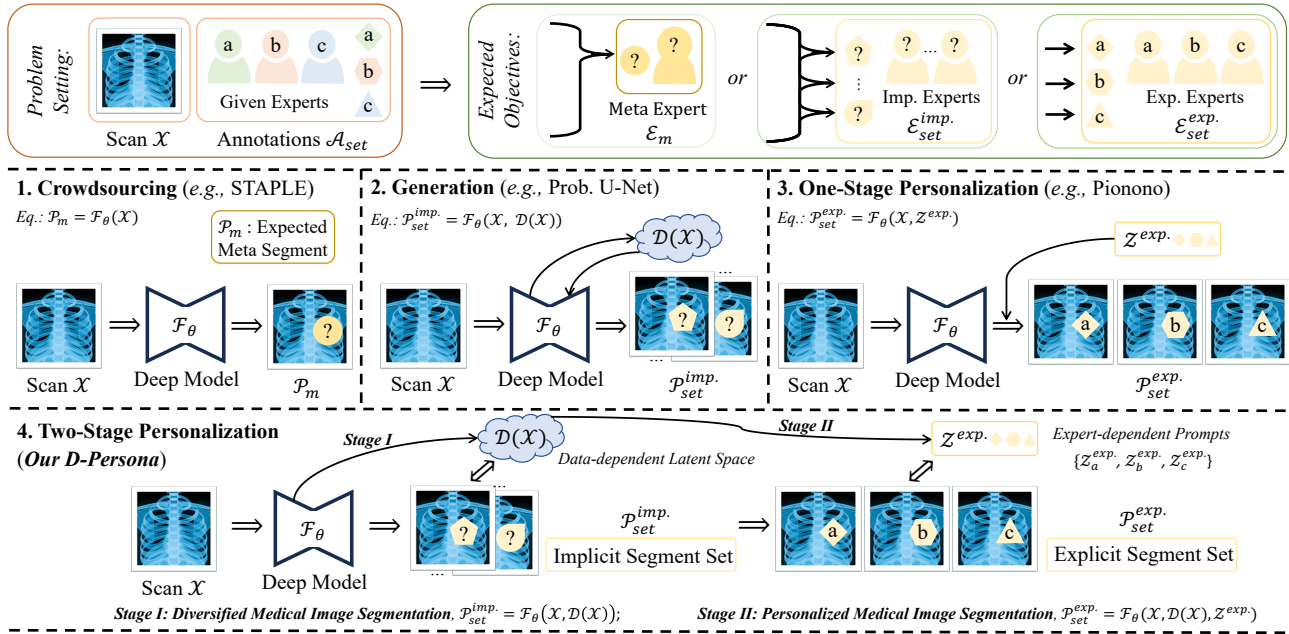


Figure 1. Overview of scheme designs in multi-rater medical image segmentation. Top: problem setting and expected objectives (i.e., meta, implicit, or explicit experts); Middle: existing methods including crowdsourcing, generation, and one-stage personalization; Bottom: our proposed two-stage framework, providing both diversified and specifically personalized segmentation simultaneously.

proach it. The generation methods [18, 34] (Fig. 1 Middle) aim to generate diversified outputs, providing varying and plausible expert opinions as in the specialist consultation or case conference. Their generated segment set is implicit and unordered. The personalization methods [23, 36, 48] (Fig. 1 Middle-right) are recent ones, aiming to generate the explicit segment set, i.e., the corresponding expert results, which enables the tracking and analysis of individual raters’ performance and preferences, so as to enhance the efficacy of quality control and clinical training.

Given the interrelated nature and distinct values of both implicit and explicit segment sets in the medical domain [28, 40], we are prompted to ask: “Can we achieve the goals of both generation and personalization methods, i.e., producing both diversified and personalized results in multi-rater medical image segmentation?” To achieve this goal, in this paper, we propose a two-stage framework **D-Persona**, as shown in Fig. 1 Bottom. Our basic idea is to divide the task into two stages. Stage I focuses on learning a common latent space to capture the annotation variability, which can be drawn to generate diversified results as potential expert opinions. On the other hand, Stage II focuses on generating personalized segmentation results by reusing the learned latent space in Stage I and extracting the corresponding expert prompts from it.

Our major contributions can be summarized as follows.

- We set a new grand goal for multi-rater medical image

segmentation, i.e., aiming to generate diverse expert opinions and personalized results simultaneously. We propose a novel unified two-stage framework to address the diversity and personalization requirements, respectively.

- In Stage I, we propose a bound-constrained training loss, which predicts the intersections and unions of annotations and introduces flexibility in uncertain areas. This substantially enhances segmentation diversity and increases the variety in the latent space.
- In Stage II, we introduce an attention-based projection mechanism for personalized medical image segmentation, which can adaptively query expert prompts from the shared latent space.
- Extensive experiments on two datasets demonstrate our proposed framework can provide diversified and personalized segmentation results simultaneously, setting new state-of-the-art (SOTA) performance for multi-rater medical image segmentation.

2. Related work

In this section, we review existing multi-rater medical image segmentation methods as well as noisy learning works that are related to annotation ambiguity.

2.1. Crowdsourcing-based Segmentation

Crowdsourcing [31] is a common solution to reduce the expensive annotation costs by eliciting imperfect labels from

crowds. In medical scenarios, crowdsourced labels can be collected from junior medical students and are usually quite noisy. To facilitate the model training, crowdsourcing methods [20, 39] typically adopt several label-fusing strategies to achieve consensus and reduce noise, *e.g.*, the conventional majority-voting strategy. However, this intuitive way ignores the annotators’ differences either in domain expertise or personal preferences. Therefore, the Simultaneous Truth and Performance Level Estimation (STAPLE) [44] was further proposed to utilize the annotators’ performance as weights to fuse crowdsourced labels, which is a more convincing way. It has been widely used in many medical tasks [27] to produce the “golden standard”.

Nevertheless, in numerous medical contexts, there is no such golden standard and the meta segmentation remains elusive [20] as in the case of challenging GTVp delineation. Thus, recent multi-rater image segmentation efforts have shifted the focus to generate either diversified implicit or personalized explicit segments, instead of the meta segment. Note that, there are also other methods, not aiming for explicit meta segments. For example, MAX-MIG [5] proposed to maximize the mutual information between the input data and crowdsourced labels for the model training. To better represent the expert disagreements, randomly selecting a label for training rather than using consensus could achieve better model calibration performance [14, 17].

2.2. Generation-based Segmentation

Various generative techniques have emerged for medical tasks [7]. For example, [41] used generative adversarial networks (GANs) to synthesize missing brain modalities, significantly improving the diagnosis performance. In terms of segmentation, Probabilistic U-Net [18] proposed a conditional Variational Autoencoder (cVAE) for ambiguous medical image segmentation, which learns a latent space to represent the segmentation ambiguity. To increase the segmentation diversity, multi-scale improvements [2, 19] were proposed to enhance the original architecture. Moreover, [34] used a powerful diffusion-based model for ambiguous medical image segmentation, achieving superior performance.

Despite the progress, these methods mainly focus on generating diversified implicit segmentation results and they are unable to achieve personalized segmentation. In contrast, our work focuses on generating both diversified and personalized segments in a unified way.

2.3. One-stage Personalization Segmentation

Personalized segmentation directly incorporates expert preferences into the model training process [48]. For example, [15] used an explicit expertness code to denote the modified training labels. [23] employed learnable queries to represent corresponding expert annotations. Furthermore, [36] constructed several normal distributions with learnable

parameters to represent the inter- and intra-observer variability. All these one-stage works strive for a one-to-one correspondence in multi-rater medical image segmentation. Conversely, our proposed D-Persona model not only delivers improved personalized results but also captures a variety of expert segmentation opinions.

2.4. Noisy Learning

Label noise is highly related to annotation ambiguity [38]. The common practice for performing noisy learning can be divided into two categories: (1) detecting label noise and then rectifying the training samples; and (2) developing a noise-robust framework from noisy distribution. For example, in the first category, the noise transition matrix [8, 25, 29, 33] can be estimated and used to build a statistically consistent classifier with the inferred clean distribution. Also, sample selection works [16, 21, 32, 45] were proposed to design effective selection strategies to detect and reduce the impact of noisy samples during the training. For the second category, noise-robust frameworks are developed with a variety of robust loss functions [11, 43, 49] and modified architecture designs [37, 51], which focus on the adaptive training from the original distribution.

In contrast, similar to the recent efforts [23, 36] in multi-rater medical image segmentation, we do not consider label noise here. Instead, we consider label diversity caused by different expert preferences and inherent data uncertainty, aiming to model the diversity and personalize the results.

3. Methods

Fig. 2 shows the overall pipeline of our proposed D-Persona framework for multi-rater medical image segmentation, which consists of two major stages. Stage I aims to learn a common latent space from n experts’ annotations $A_{set} = \{A_1, \dots, A_n\}$, which can then be sampled to generate diverse segmentation results. We use the Probabilistic U-Net [18] as the baseline, including a U-Net backbone F_{θ}^b for feature extraction, a prediction head F_{θ}^{head} for the mapping from the latent space to segmentation results, and two separate encoders F_{θ}^{prior} and $F_{\theta}^{post.}$ for the prior and posterior distribution generations, respectively. Stage II aims to find specific expert prompts in the latent space corresponding to the n experts for personalized segmentation, where we learn n individual projection heads $F_{\theta_i}^{proj.}$, $i = \{1, \dots, n\}$, while reusing F_{θ}^b , F_{θ}^{prior} and F_{θ}^{head} trained in Stage I.

3.1. Stage I: Diversified Segmentation

In the conventional Probabilistic U-Net model [18], there are two separate parts to achieve diversified segmentation. In the first part, the input X alone is first used to generate the prior distribution $\mathbb{D}^{prior}(X)$, which is set as a multivariate normal distribution with a diagonal covariance ma-

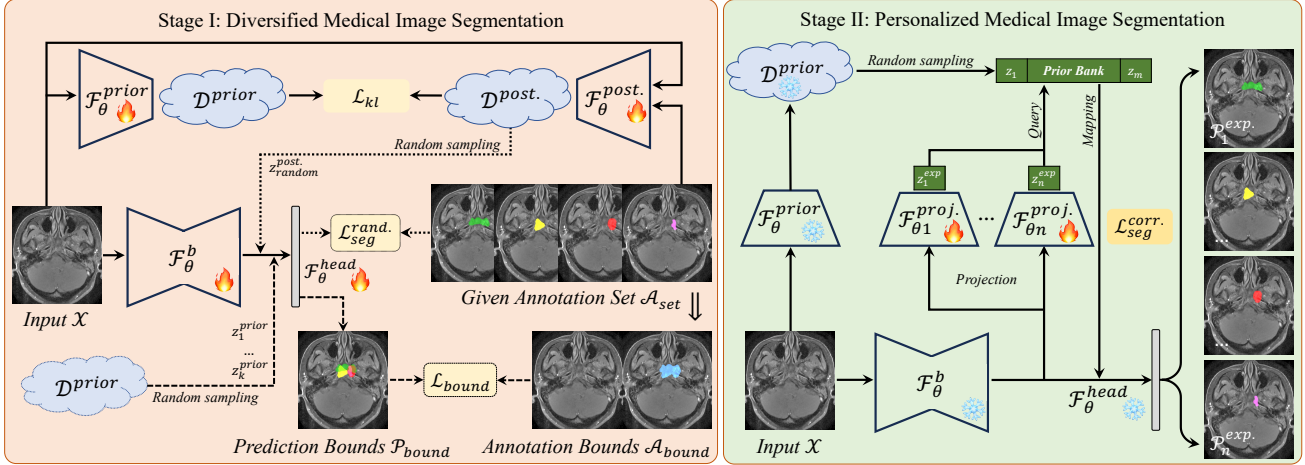


Figure 2. Pipeline of our proposed D-Persona framework for multi-rater medical image segmentation. Left: *Stage I* is designed to construct a common latent space where different latent codes lead to diversified segmentation results; Right: *Stage II* performs the personalized segmentation by individual projection heads to mimic the corresponding expert raters.

trix, denoted as $\mathcal{N}(\mu^{prior}, \sigma^{prior})$. The posterior distribution $\mathbb{D}^{post.}(X, A_{set}) = \mathcal{N}(\mu^{post.}, \sigma^{post.})$ is defined as the joint distribution of X and all annotations A_{set} . The corresponding mean values and variances are obtained as

$$\begin{aligned} \mu^{prior}, \sigma^{prior} &= F_{\theta}^{prior}(X), \\ \mu^{post.}, \sigma^{post.} &= F_{\theta}^{post.}(X, A_{set}), \end{aligned} \quad (1)$$

where F_{θ}^{prior} and $F_{\theta}^{post.}$ are two learnable neural encoders as shown in Fig. 2 (Left-top). Here, X and A_{set} are concatenated together along the channel axis as the inputs of $F_{\theta}^{post.}$. A common Kullback–Leibler divergence (KL) loss is used to align the two distributions:

$$\mathcal{L}_{kl} = KL(\mathbb{D}^{prior}(X), \mathbb{D}^{post.}(X, A_{set})). \quad (2)$$

Then, the second part is for the target generation. Given a posterior distribution $\mathbb{D}^{post.}$, we randomly sample a latent code $z_{random}^{post.} \in \mathbb{R}^{D \times 1 \times 1}$, which is further scaled to $\hat{z}_{random}^{post.} \in \mathbb{R}^{D \times H \times W}$ as the image size. Here, D is the dimension of all latent codes. We then concatenate the scaled latent code and U-Net features together and feed them into a prediction head F_{θ}^{head} to generate the segmentation result:

$$P^{post.} = F_{\theta}^{head}(\hat{z}_{random}^{post.}, F_{\theta}^b(X)). \quad (3)$$

Since we do not have the correspondence between $\hat{z}_{random}^{post.}$ and the label, we randomly select an annotation A_{random} from A_{set} as [14, 17], and supervise the model training with a segmentation loss (*i.e.*, Dice loss as [36]):

$$\mathcal{L}_{seg}^{rand.} = DSC(P^{post.}, A_{random}). \quad (4)$$

Following [36], we exploit a re-parameterization trick to sample $\hat{z}_{random}^{post.}$ from the posterior distribution $\mathbb{D}^{post.}$ so as to back-propagate the gradients for the model training.

Nevertheless, the prediction diversity of such a conventional design is limited [2, 19]. To address this, inspired by [50] that leverages the prior distribution with a complementary task to improve the diversity of image inpainting, we propose a bound prediction task as the complement and also involve the prior distribution into the segmentation training. Specifically, the intersection and union (*i.e.*, $A_{bound} = \{A_{inter.}, A_{union}\}$) of the annotation set are calculated as supervision labels. Then, by sampling K latent codes from the prior distribution \mathbb{D}^{prior} , we can obtain K segmentation results and their intersection and union as $P_{bound}^{prior} = \{P_{inter.}^{prior}, P_{union}^{prior}\}$, based on which we can compute the complementary segmentation loss:

$$\mathcal{L}_{bound} = DSC(P_{inter.}^{prior}, A_{inter.}) + DSC(P_{union}^{prior}, A_{union}). \quad (5)$$

Overall, the model in Stage I is trained by a weighted sum of \mathcal{L}_{kl} , $\mathcal{L}_{seg}^{rand.}$ and \mathcal{L}_{bound} as:

$$\mathcal{L}_{stage1} = \mathcal{L}_{kl} + \alpha \times \mathcal{L}_{seg}^{rand.} + \beta \times \mathcal{L}_{bound}, \quad (6)$$

where α and β are two hyperparameters to balance the three losses. Such a design has several advantages: (1) we relax the predictions in uncertain regions to facilitate more diversified results; (2) the prior distribution is being trained anyway and we do not introduce any new model parameters; (3) the design is flexible and can be easily combined with other advanced model architectures [2, 19].

3.2. Stage II: Personalized Segmentation

Once Stage I learns a common latent space that can produce diversified and plausible segmentation predictions, the purpose of Stage II is to find specific expert prompts in the

latent space that correspond to the individual expert annotators. As shown in Fig. 2 (Right), our idea is to employ n individual projection heads to generate n expert prompts from the U-Net features $F_\theta^b(X)$ as

$$z_i^{exp.} = Pooling[F_{\theta_i}^{proj.}(F_\theta^b(X)), i = 1, \dots, n] \quad (7)$$

where $z_i^{exp.} \in \mathbb{R}^{D \times 1 \times 1}$ denotes the specific latent code relevant for the i -th expert. It is obtained via a global average pooling operation for the 2D outputs of the i -th projection head $F_{\theta_i}^{proj.}$, to capture global annotation characteristics.

Then, to ensure all generated expert prompts are in the learned common latent space, we further introduce a cross-attention operation to map $z_i^{exp.}$ into $\tilde{z}_i^{exp.}$. In particular, we first randomly sample M latent codes from the fixed prior distribution D^{prior} as a prior bank $z_{bank}^{prior} \in \mathbb{R}^{D \times 1 \times M}$. We then use $z_i^{exp.}$ as the query and z_{bank}^{prior} as the key and value for cross-attention:

$$\tilde{z}_i^{exp.} = Softmax(z_i^{exp. \top} \cdot z_{bank}^{prior}) \cdot z_{bank}^{prior \top} \quad (8)$$

In this way, $\tilde{z}_i^{exp.}$ is constrained into the frozen prior distribution $D^{prior}(X)$, with respect to the input X . After that, we use the fixed F_θ^{head} to transform the expert prompt into the final segmentation result $P_i^{exp.}$, which is used to compare with the corresponding expert annotation A_i for computing the segmentation loss as

$$\begin{aligned} \mathcal{L}_{seg}^{corr.} &= \sum_{i=1}^n DSC(P_i^{exp.}, A_i) \\ s.t., P_i^{exp.} &= F_\theta^{head}(\tilde{z}_i^{exp.}, F_\theta^b(X)), \end{aligned} \quad (9)$$

where $\tilde{z}_i^{exp.}$ is the scaled version of $z_i^{exp.}$. Note that, $\mathcal{L}_{seg}^{corr.}$ is only used to train the n projection heads, while other modules are fixed in Stage II. This design principle of our Stage II personalized segmentation part is general and can be extended into other powerful generative methods [34].

4. Experiments and Results

4.1. Dataset

We evaluate our method on the two segmentation datasets: our in-house dataset (NPC-170) and the public lung nodule segmentation dataset (LIDC-IDRI) [1].

The NPC-170 dataset contains 170 subjects with Nasopharyngeal Carcinoma in three Magnetic Resonance Imaging (MRI) contrasts (T1, T1-Contrast, and T2). For each sample, four senior radiologists from different locations (with experiences of around 5-10 years) individually annotated the GTVp of Nasopharyngeal Carcinoma. We divide the 170 subjects into separate training, validation, and testing sets (*i.e.*, 100, 20, and 50 subjects, respectively). 2D medical image segmentation is employed in the experiments and, in total, there are 6134, 1126, and 3058 slices for the model training, validation, and testing, respectively.

The LIDC-IDRI dataset contains 1609 2D thoracic CT scans belonging to 214 subjects. Each scan is equipped with four binary masks to indicate the lung nodule. Note that, 12 radiologists participated in the annotation process [1]. To simulate the expert preferences, we manually rank the areas of four provided annotations, as the setting in [48]. In this way, the first expert is supposed to be conservative in the sense of only annotating the minimum regions while the fourth expert prefers to annotate any suspicious lung nodule area, *i.e.*, producing the largest annotation regions. Following [42], a four-fold cross-validation setting at the patient level is used in the experiments.

4.2. Implementation Details

The inputs from both the NPC-170 and LIDC-IDRI datasets are center-cropped into a fixed size of 128×128 . Then, for the samples on NPC-170, we normalize them into zero mean and unit variance. Random flips, rotation, or adding random noise are applied to augment the training data. On the LIDC-IDRI dataset, we follow [42] to perform the data pre-processing and use the common flip and rotation operations to augment the samples on the LIDC-IDRI dataset.

For the model training, we adopt the Adam optimizer with an initial learning rate of $1e-4$. The loss weight α and the distribution dimension D are set as 1 and 6, as [36]. The overall training epochs are 300, including 100 epochs for Stage I and 200 epochs for Stage II. The size of z_{bank}^{prior} , *i.e.*, M is set as 100. K and β are set as 10 and 0.5, which will be discussed in Section 5.2. An L2-regularization term is also used to avoid over-fitting as [18, 36]. All experiments are conducted in an identical environment with a single NVIDIA GeForce RTX 3090 GPU. The overall number of parameters is 28.46 M and the projection head is relatively lightweight (*i.e.*, 22.03 K for each head).

4.3. Evaluation Metrics

Considering our D-Persona model consists of two stages: diversified and personalized segmentation, we thus evaluate them separately. First, we use two common set-to-set metrics for the diversity estimation, *i.e.*, the Generalized Energy Distance GED [3, 18] and the soft Dice score $Dice_{soft}$ [15, 42]. The former is used to measure the prediction diversity and the latter indicates the reliability of different generated results. Specifically, GED is defined as

$$GED = 2\mathbb{E}[d(P, A)] - \mathbb{E}[d(P, P')] - \mathbb{E}[d(A, A')], \quad (10)$$

where P , P' and A , A' are independent samples from the prediction set P_{set} and annotation set A_{set} , respectively, and d denotes the distance function $d(a, b) = 1 - IoU(a, b)$ as in [18]. In general, a lower GED indicates greater dispersion and variability in the segmentation results. Moreover, $Dice_{soft}$ receives the soft predictions P_{soft} and soft annotations A_{soft} (*i.e.*, the average of multiple predictions

Table 1. Performance of our proposed D-Persona framework and several public methods for multi-rater medical image segmentation on our in-house NPC-170 dataset. We show the results of individually trained U-Net models (*i.e.*, using annotations from a single expert, Top), generation-based methods (with different sampling numbers, Middle), and personalized segmentation works (Bottom), respectively.

Method	Diversity Performance		Personalized Bounds (%)		Personalized Segmentation Performance (%)				
	$GED \downarrow$	$Dice_{soft} \uparrow$ (%)	$Dice_{max} \uparrow$	$Dice_{match} \uparrow$	$Dice_{A1} \uparrow$	$Dice_{A2} \uparrow$	$Dice_{A3} \uparrow$	$Dice_{A4} \uparrow$	$Dice_{mean} \uparrow$
U-Net (A_1)	0.4888	73.69	N/A		80.90	70.51	67.14	72.57	72.78
U-Net (A_2)	0.5086	73.64			72.05	75.62	69.93	71.85	72.36
U-Net (A_3)	0.5077	72.37			69.46	73.09	76.25	69.86	72.17
U-Net (A_4)	0.5433	71.63			73.71	68.77	66.08	73.54	70.52
Prob. U-Net [18] (#10)	0.4478	75.33	74.18	74.17	N/A				
Prob. U-Net [18] (#30)	0.4466	75.34	74.23	74.22					
Prob. U-Net [18] (#50)	0.4465	75.34	74.24	74.24					
D-Persona (Stage I, #10)	0.2553	79.99	76.54	76.40					
D-Persona (Stage I, #30)	0.2385	80.40	77.50	77.43					
D-Persona (Stage I, #50)	0.2359	80.43	77.70	77.67					
CM-Global [38]	0.4375	76.28	75.00	75.00	77.92	74.65	71.97	75.47	75.00
CM-Pixel [48]	0.4400	76.39	74.92	74.92	78.97	73.93	71.65	75.12	74.92
TAB [23]	0.2760	78.90	76.68	75.15	77.33	73.45	74.83	71.67	74.32
Pionono [36]	0.3924	76.96	75.58	75.44	78.87	74.11	71.97	75.41	75.09
D-Persona (Stage II)	0.3199	79.01	77.26	76.61	79.78	74.60	75.22	75.17	76.19

and annotations) and uses varied thresholds to conduct T times of binary evaluations as:

$$Dice_{soft} = \frac{1}{T} \sum_{i=1}^T Dice([P_{soft} > \tau_i], [A_{soft} > \tau_i]) \quad (11)$$

where τ is a threshold selected from the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $T = 5$.

Dice	Four Given Annotations			
Six Sampled Predictions	0.5532	0.8994	0.8624	0.6338
	0.5552	0.6694	0.5210	0.7575
	0.7135	0.8476	<u>0.8500</u>	0.5928
	0.5556	0.5949	0.6366	0.7836
	0.5099	0.6071	0.5953	0.8973
	0.8094	0.8452	0.5741	0.7156
Metrics	$Dice_{max} = \{0.8094, 0.8994, 0.8624, 0.8973\}_{mean}$			
	$Dice_{match} = \{0.8094, 0.8994, 0.8500, 0.8973\}_{mean}$			

Figure 3. Exemplar explanation of $Dice_{max}$ and $Dice_{match}$ in a given 4×6 Dice matrix. $Dice_{max}$ averages the maximum scores of individual columns and $Dice_{match}$ further constrains a one-to-one matching between the prediction and annotation sets.

Second, we further use the set-to-set metrics $Dice_{max}$ and $Dice_{match}$ to evaluate the personalized segmentation in our framework. As Fig. 3 shows, $Dice_{max}$ quantifies the optimal overlap between P_{set} and A_{set} . $Dice_{match}$ further imposes a stringent one-to-one matching criterion between the sets. In other words, $Dice_{max}$ is an upper bound of $Dice_{match}$ and their differences serve as an indicator to de-

termine whether each expert annotation corresponds with a unique and closely aligned prediction.

Additionally, we also provide each annotator a Dice score (denoted as $Dice_{A(i)}$, $i = 1, \dots, n$), to reflect the personalized segmentation performance for the respective human expert. Then, $Dice_{mean}$ further gives the average personalized performance of the n experts.

4.4. Performance on NPC-170

Table 1 gives the quantitative performance of our proposed D-Persona in comparison with other existing works on the in-house NPC-170 dataset, indicating several findings: (1) the individually trained U-Net models, where only annotations from a single expert are used, achieve superior results for the targeted expert but significantly poor performance for the others, highlighting the great inter-observer variability; (2) for generation-based methods, our Stage I exhibits remarkable segmentation diversity than the baseline model, as evidenced by leading scores in GED and $Dice_{soft}$; (3) we also observe an increase in personalized bounds in Stage I (*e.g.*, $Dice_{max}$ improves from 74.18% of the baseline to 76.54% of our model), and further improvements can be achieved by increasing the number of sampled latent codes (*e.g.*, a 1.16% gain in $Dice_{max}$ with 50 random samples); (4) in Stage II, our two-stage personalization strategy surpasses the performance of other one-stage methods (*e.g.*, a 1.10% gain in $Dice_{mean}$ than the second-best model [36]), demonstrating the superiority of our approach. Note that, when compared with the individually trained models, our D-Persona Stage II does not perform better for the personalized results specific to the particular annotator except A_4 . This also happens to all the compared one-stage methods.

Table 2. Performance of our proposed D-Persona framework and several public methods for multi-rater medical image segmentation on the public LIDC-IDRI dataset. We show the results of individually trained U-Net models (*i.e.*, using annotations from a single expert, Top), generation-based methods (with different sampling numbers, Middle), and personalized segmentation works (Bottom), respectively.

Method	Diversity Performance		Personalized Bounds (%)		Personalized Segmentation Performance (%)				
	$GED \downarrow$	$Dice_{soft} \uparrow$ (%)	$Dice_{max} \uparrow$	$Dice_{match} \uparrow$	$Dice_{A1} \uparrow$	$Dice_{A2} \uparrow$	$Dice_{A3} \uparrow$	$Dice_{A4} \uparrow$	$Dice_{mean} \uparrow$
U-Net (A_1)	0.3062	86.59	N/A		87.80	87.47	85.49	80.67	85.36
U-Net (A_2)	0.2459	88.43			87.16	89.08	88.59	85.15	87.50
U-Net (A_3)	0.2436	88.20			85.29	88.48	89.40	87.20	87.59
U-Net (A_4)	0.2962	85.83			80.80	85.48	88.22	88.90	85.85
Prob. U-Net [18] (#10)	0.2181	88.79	88.60	88.43	N/A				
Prob. U-Net [18] (#30)	0.2169	88.79	88.80	88.73					
Prob. U-Net [18] (#50)	0.2168	88.80	88.87	88.81					
D-Persona (Stage I, #10)	0.1461	90.24	90.75	90.51					
D-Persona (Stage I, #30)	0.1375	90.42	91.23	91.16					
D-Persona (Stage I, #50)	0.1358	90.45	91.37	91.33					
CM-Global [38]	0.2432	88.53	87.51	87.51	86.13	88.76	88.99	86.18	87.51
CM-Pixel [48]	0.2407	88.64	87.72	87.72	85.99	88.81	89.31	86.77	87.72
TAB [23]	0.2322	86.35	87.11	86.08	85.00	86.35	86.77	85.77	85.97
Pionono [36]	0.1502	90.00	90.10	88.97	87.94	89.11	89.55	88.76	88.84
D-Persona (Stage II)	0.1444	90.31	90.38	89.17	88.54	89.50	90.03	88.60	89.17

4.5. Performance on LIDC-IDRI

Table 2 extends the evaluation of our model to the LIDC-IDRI dataset, where the average performance of a four-fold cross-validation is given. The individually trained results indicate that the inter-observer variance is less pronounced compared to our in-house NPC-170 dataset. Then, echoing the findings from NPC-170, our D-Persona model consistently outperforms the baseline Probabilistic U-Net [18] in achieving superior diversified segmentation performance. Furthermore, the second stage of our proposed D-Persona model exhibits better performance in both the set-to-set evaluations and the one-to-one comparisons. Meanwhile, in Stage II, we can see that the average personalized performance has reached the upper bound (*i.e.*, both $Dice_{match}$ and $Dice_{mean}$: 89.17%) on LIDC-IDRI, highlighting its effectiveness in personalized medical image segmentation.

5. Discussions

5.1. Visual Results

Fig. 4 shows the diversified segmentation results of Stage I in our D-Persona model. We can see that, on the two datasets, our model can generate varying and plausible predictions for the target regions, which brings potential benefits for providing different opinions in clinical scenarios.

Fig. 5 further illustrates the personalized segmentation results of Stage II in the proposed model. First, on the NPC-170 dataset, our D-Persona can detect most of the targets, producing predictions in complex anatomical structures as the given expert annotations, indicated by the green arrows in Fig. 5. Second, on LIDC-IDRI, as we simulated the

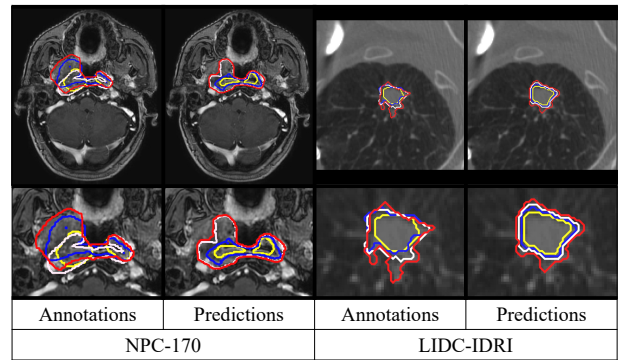


Figure 4. Diversified segmentation results of Stage I in our proposed D-Persona framework on the NPC-170 (Left) and LIDC-IDRI (Right) datasets. Different colors denote different delineations, which shows that our model can generate diverse and plausible predictions.

Table 3. Ablation studies of different K and β in Stage I of our proposed D-Persona framework on the NPC-170 dataset. The sampling number is set as 50 for all comparisons.

K	β	$GED \downarrow$	$Dice_{soft} \uparrow$ (%)	K	β	$GED \downarrow$	$Dice_{soft} \uparrow$ (%)
6	0.5	0.2545	78.81	10	0.01	0.4378	75.80
8		0.2512	79.48		0.1	0.3358	77.95
10		0.2359	80.43		0.5	0.2359	80.43
12		0.2389	80.17		1	0.2170	80.20
14		0.2368	80.08		2	0.2581	77.73

expert preferences from conservative to aggressive styles (see Section 4.1), the right part of Fig. 5 reveals that our model can successfully capture the underlying preferences, demonstrating the superiority of our proposed D-Persona model for personalized segmentation.

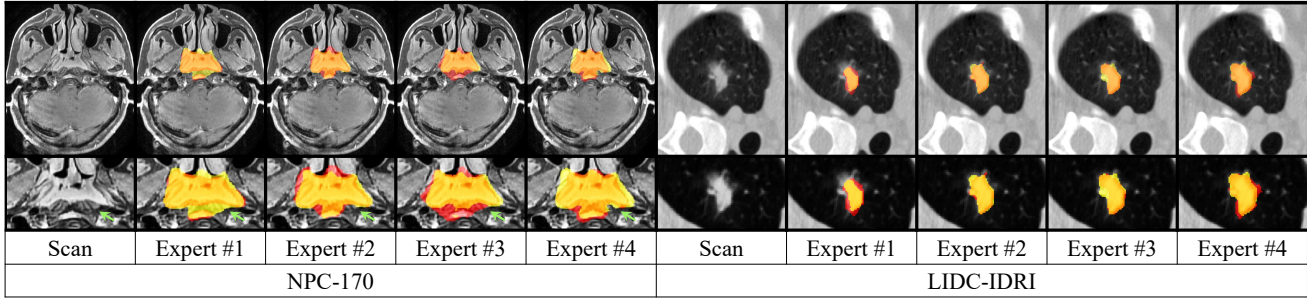


Figure 5. Personalized segmentation results of Stage II in our proposed D-Persona framework on the NPC-170 (Left) and LIDC-IDRI (Right) datasets. Compared to expert annotations (Red), our model can generate corresponding personalized segmentation results (Yellow). Particularly, our model successfully captures the underlying annotation preferences, *i.e.*, from conservative to aggressive styles, as shown in the four LIDC-IDRI results on the right.

5.2. Selection of Hyper-parameters

To control the diversity of generated segmentation results, we here discuss the selection of hyper-parameters K and β in Stage I of our model. K is the size of the sampled latent codes in the prior distribution \mathbb{D}^{prior} . Since we applied the bound-constrained loss to supervise the learning of \mathbb{D}^{prior} and a smaller K could make the segmentation results approach the fixed annotation bounds. The top of Table 3 shows that, on the NPC-170 dataset, setting K as 10 can achieve better diversified performance for the GTVP delineation of Nasopharyngeal Carcinoma.

At the same time, β is used to balance the proposed bound loss with other losses in Equation (6). The bottom results of Table 3 indicate that selecting β as 0.5 for the model training achieves the highest $Dice_{soft}$ and second-best GED values on the NPC-170 dataset. We thus set β as 0.5 for all experiments in this paper.

5.3. Comparisons with Crowdsourcing Methods

Table 4. Comparisons between several crowdsourcing methods including majority voting (MV), random selection (RS) [17], STAPLE [44] and our proposed D-Persona model on NPC-170.

Method	$Dice_{A1} \uparrow$	$Dice_{A2} \uparrow$	$Dice_{A3} \uparrow$	$Dice_{A4} \uparrow$	$Dice_{mean} \uparrow$
MV	76.80	71.28	70.24	73.64	72.99
RS [17]	78.45	74.92	72.04	75.54	75.24
STAPLE [44]	75.75	75.76	72.29	74.07	74.47
D-Persona	79.78	74.60	75.22	75.17	76.19

As aforementioned, crowdsourcing methods aim to combine multiple annotations for the model training. Table 4 shows the performance of several conventional crowdsourcing methods, including majority voting (MV), Random Selection (RS) [17] and STAPLE [44] on the NPC-170 dataset. We can see that, our proposed D-Persona model achieves the highest performance in $Dice_{mean}$ than all heuristic crowdsourcing works (*e.g.*, a 0.95% gain than the second-

best model [17]). Furthermore, our model can provide diversified segmentation results at the same time, facilitating more clinical usage through a unified model.

5.4. Limitation and Future Work

As mentioned in Section 4.4, on the NPC-170 dataset, it remains difficult for our model to surpass the individually trained U-Net models for personalized segmentation. This suggests that preserving the distinctive preference of a particular expert is still a challenge in this task, especially when the training labels vary significantly. In other words, improving the predictions having an individual expert’s style by leveraging different annotations from other experts is not fully resolved, which is worth future investigation.

6. Conclusion

In this paper, we have presented a two-stage framework (D-Persona) for the multi-rater medical image segmentation problem, pointing out a new perspective that diversified and personalized segmentation can be addressed at the same time. Our main idea is to conduct diversified segmentation first and then query personalized results from the diversified ones. Specifically, an effective bound-constrained loss was proposed to improve the segmentation diversity, and a common latent space was constructed in Stage I to capture the annotation variability. Then, in the second stage, an attention-based projection mechanism was used to find specific expert prompts from the common latent space, so as to perform personalized segmentation. Extensive experiments on the two datasets have demonstrated that our proposed D-Persona model significantly outperforms other existing works in multi-rater medical image segmentation.

Societal Impacts. Our proposed model was trained and evaluated on the two available but limited datasets and may be impacted by the dataset bias [35], which might result in unconvincing predictions in practical applications.

References

- [1] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. 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, 2011. **5**
- [2] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötter, Urs J Muehlethaler, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *MICCAI*, pages 119–127, 2019. **3, 4**
- [3] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017. **5**
- [4] Olivier Bernard, Alain Lalonde, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE TMI*, 37(11):2514–2525, 2018. **1**
- [5] Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. Maxmig: an information theoretic approach for joint learning from crowds. In *ICLR*, 2018. **3**
- [6] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017. **1**
- [7] Pedro Celard, EL Iglesias, JM Sorribes-Fdez, Rubén Romero, A Seara Vieira, and L Borrajo. A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. *Neural Computing and Applications*, 35(3):2291–2323, 2023. **3**
- [8] De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, and Masashi Sugiyama. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In *CVPR*, pages 16630–16639, 2022. **3**
- [9] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature Methods*, 16(1):67–70, 2019. **1**
- [10] Fan Fu, Jianyong Wei, Miao Zhang, Fan Yu, Yueting Xiao, Dongdong Rong, Yi Shan, Yan Li, Cheng Zhao, Fangzhou Liao, et al. Rapid vessel segmentation and reconstruction of head and neck angiograms using 3d convolutional neural network. *Nature Communications*, 11(1):4829, 2020. **1**
- [11] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, pages 1919–1925, 2017. **3**
- [12] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *WACV*, pages 574–584, 2022. **1**
- [13] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. **1**
- [14] Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In *MICCAI*, pages 540–548, 2019. **3, 4**
- [15] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *CVPR*, pages 12341–12351, 2021. **1, 3, 5**
- [16] Lie Ju, Xin Wang, Lin Wang, Dwarikanath Mahapatra, Xin Zhao, Quan Zhou, Tongliang Liu, and Zongyuan Ge. Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE TMI*, 41(6):1533–1546, 2022. **3**
- [17] Alain Jungo, Raphael Meier, Ekin Ermiş, Marcela Blatti-Moreno, Evelyn Herrmann, Roland Wiest, and Mauricio Reyes. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *MICCAI*, pages 682–690, 2018. **3, 4, 8**
- [18] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *NeurIPS*, 2018. **2, 3, 5, 6, 7**
- [19] Simon AA Kohl, Bernardino Romera-Paredes, Klaus H Maier-Hein, Danilo Jimenez Rezende, SM Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*, 2019. **3, 4**
- [20] Anne W Lee, Wai Tong Ng, Jian Ji Pan, Sharon S Poh, Yong Chan Ahn, Hussain AlHussain, June Corry, Cai Grau, Vincent Grégoire, Kevin J Harrington, et al. International guideline for the delineation of the clinical target volumes (ctv) for nasopharyngeal carcinoma. *Radiotherapy and Oncology*, 126(1):25–36, 2018. **3**
- [21] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2019. **3**
- [22] Rongjian Li, Tao Zeng, Hanchuan Peng, and Shuiwang Ji. Deep learning segmentation of optical microscopy images improves 3-d neuron reconstruction. *IEEE TMI*, 36(7):1533–1541, 2017. **1**
- [23] Zehui Liao, Shishuai Hu, Yutong Xie, and Yong Xia. Transformer-based annotation bias-aware medical image segmentation. In *MICCAI*, pages 24–34, 2023. **2, 3, 6, 7**
- [24] Li Lin, Qi Dou, Yue-Ming Jin, Guan-Qun Zhou, Yi-Qiang Tang, Wei-Lin Chen, Bao-An Su, Feng Liu, Chang-Juan Tao, Ning Jiang, et al. Deep learning for automated contouring of primary tumor volumes by mri for nasopharyngeal carcinoma. *Radiology*, 291(3):677–686, 2019. **1**

- [25] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE TPAMI*, 38(3):447–461, 2015. [3](#)
- [26] Xiangde Luo, Wenjun Liao, Yuan He, Fan Tang, Mengwan Wu, Yuanyuan Shen, Hui Huang, Tao Song, Kang Li, Shichuan Zhang, et al. Deep learning-based accurate delineation of primary gross tumor volume of nasopharyngeal carcinoma on heterogeneous magnetic resonance imaging: A large-scale and multi-center study. *Radiotherapy and Oncology*, 180:109480, 2023. [1](#)
- [27] Amirreza Mahbod, Gerald Schaefer, Benjamin Bancher, Christine Löw, Georg Dorffner, Rupert Ecker, and Isabella Ellinger. Cryonuseg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images. *Computers in Biology and Medicine*, 132:104349, 2021. [3](#)
- [28] Thibault Marin, Yue Zhuo, Rita Maria Lahoud, Fei Tian, Xiaoyue Ma, Fangxu Xing, Maryam Moteabbed, Xiaofeng Liu, Kira Grogg, Nadya Shusharina, et al. Deep learning-based gtv contouring modeling inter-and intra-observer variability in sarcomas. *Radiotherapy and Oncology*, 167:269–276, 2022. [1](#), [2](#)
- [29] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *ICML*, pages 125–134, 2015. [3](#)
- [30] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE TMI*, 34(10):1993–2024, 2014. [1](#)
- [31] Silas Ørting, Andrew Doyle, Arno van Hilten, Matthias Hirth, Oana Inel, Christopher R Madan, Panagiotis Mavridis, Helen Spiers, and Veronika Cheplygina. A survey of crowdsourcing in medical image analysis. *arXiv preprint arXiv:1902.09159*, 2019. [2](#)
- [32] Deep Patel and PS Sastry. Adaptive sample selection for robust learning under label noise. In *WACV*, pages 3932–3942, 2023. [3](#)
- [33] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 1944–1952, 2017. [3](#)
- [34] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *CVPR*, pages 11536–11546, 2023. [1](#), [2](#), [3](#), [5](#)
- [35] María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, 13(1):4581, 2022. [8](#)
- [36] Arne Schmidt, Pablo Morales-Álvarez, and Rafael Molina. Probabilistic modeling of inter-and intra-observer variability in medical image segmentation. In *ICCV*, pages 21097–21106, 2023. [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
- [37] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014. [3](#)
- [38] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *CVPR*, pages 11244–11253, 2019. [3](#), [6](#), [7](#)
- [39] Vincenzo Valentini, Maria Antonietta Gambacorta, Brunella Barbaro, Giuditta Chiloire, Claudio Coco, Prajnan Das, Francesco Fanfani, Ines Joye, Lisa Kachnic, Philippe Maingon, et al. International consensus guidelines on clinical target volume delineation in rectal cancer. *Radiotherapy and Oncology*, 120(2):195–201, 2016. [3](#)
- [40] Julie van der Veen, Akos Gulyban, Siri Willems, Frederik Maes, and Sandra Nuyts. Interobserver variability in organ at risk delineation in head and neck cancer. *Radiation Oncology*, 16:1–11, 2021. [2](#)
- [41] Bao Wang, Yongsheng Pan, Shangchen Xu, Yi Zhang, Yang Ming, Ligang Chen, Xuejun Liu, Chengwei Wang, Yingchao Liu, and Yong Xia. Quantitative cerebral blood volume image synthesis from standard mri using image-to-image translation for brain tumors. *Radiology*, 308(2):e222471, 2023. [3](#)
- [42] Lin Wang, Xiufen Ye, Lie Ju, Wanji He, Donghao Zhang, Xin Wang, Yelin Huang, Wei Feng, Kaimin Song, and Zongyuan Ge. Medical matting: Medical image segmentation with uncertainty from the matting perspective. *Computers in Biology and Medicine*, 158:106714, 2023. [5](#)
- [43] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, pages 322–330, 2019. [3](#)
- [44] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE TMI*, 23(7):903–921, 2004. [1](#), [3](#), [8](#)
- [45] Qi Wei, Haoliang Sun, Xiankai Lu, and Yilong Yin. Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In *ECCV*, pages 516–532, 2022. [3](#)
- [46] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022. [1](#)
- [47] Yicheng Wu, Zhonghua Wu, Hengcan Shi, Bjoern Picker, Winston Chong, and Jianfei Cai. Coactseg: Learning from heterogeneous data for new multiple sclerosis lesion segmentation. In *MICCAI*, pages 3–13, 2023. [1](#)
- [48] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Joseph Jacob, Olga Ciccarelli, Frederik Barkhof, and Daniel C. Alexander. Disentangling human error from the ground truth in segmentation of medical images. In *NeurIPS*, pages 15750–15762, 2020. [2](#), [3](#), [5](#), [6](#), [7](#)
- [49] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, pages 8778–8788, 2018. [3](#)
- [50] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, pages 1438–1447, 2019. [4](#)
- [51] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *CVPR*, pages 1237–1246, 2019. [3](#)