Ambiguous Medical Image Segmentation using Diffusion Models

Aimon Rahman[1] Jeya Maria Jose Valanarasu[1] Ilker Hacihaliloglu[2] Vishal M. Patel[1]
[1]Johns Hopkins University [2]University of British Columbia
arahma30@jhu.edu

Abstract

Collective insights from a group of experts have always proven to outperform an individual's best diagnostic for clinical tasks. For the task of medical image segmentation, existing research on AI-based alternatives focuses more on developing models that can imitate the best individual rather than harnessing the power of expert groups. In this paper, we introduce a single diffusion model-based approach that produces multiple plausible outputs by learning a distribution over group insights. Our proposed model generates a distribution of segmentation masks by leveraging the inherent stochastic sampling process of diffusion using only minimal additional learning. We demonstrate on three different medical image modalities- CT, ultrasound, and MRI that our model is capable of producing several possible variants while capturing the frequencies of their occurrences. Comprehensive results show that our proposed approach outperforms existing state-of-the-art ambiguous segmentation networks in terms of accuracy while preserving naturally occurring variation. We also propose a new metric to evaluate the diversity as well as the accuracy of segmentation predictions that aligns with the interest of clinical practice of collective insights. Implementation code: https://github.com/aimansnigdha/Ambiguous-Medical-Image-Segmentation-using-Diffusion-Models.

1. Introduction

Diagnosis is the central part of medicine, which heavily relies on the individual practitioner assessment strategy. Recent studies suggest that misdiagnosis with potential mortality and morbidity is widespread for even the most common health conditions [32, 49]. Hence, reducing the frequency of misdiagnosis is a crucial step towards improving healthcare. Medical image segmentation, which is a central part of diagnosis, plays a crucial role in clinical outcomes. Deep learning-based networks for segmentation are now getting traction for assisting in clinical settings, however, most of the leading segmentation networks in the literature are deterministic [17, 23, 34, 36, 41, 42, 44], meaning they predict a single segmentation mask for each input image. Unlike natural images, ground truths are not deterministic in medical images as different diagnosticians can have different opinions on the type and extent of an anomaly [1,15,37,39]. Due to this, the diagnosis from medical images is quite challenging and often results in a low inter-rater agreement [22,24,56]. Depending on only pixel-wise probabilities and ignoring co-variances between the pixels might lead to misdiagnosis. In clinical practice, aggregating interpretations of multiple experts have shown to improve diagnosis and generate fewer false negatives [57].

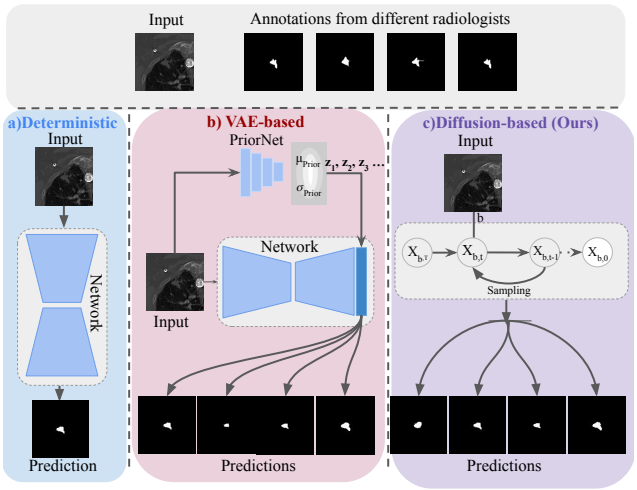

Figure 1. a) Deterministic networks produce a single output for an input image. b) c-VAE-based methods encode prior information about the input image in a separate network and sample latent variables from there and inject it into the deterministic segmentation network to produce stochastic segmentation masks. c) In our method the diffusion model learns the latent structure of the segmentation as well as the ambiguity of the dataset by modeling the way input images are diffused through the latent space. Hence our method does not need an additional prior encoder to provide latent variables for multiple plausible annotations.

In fact, utilizing the aptitude of multiple medical experts has been a part of long-standing clinical traditions such as case conferences, specialist consultations, and tumor boards. By harnessing the power of collective intelligence, team-based decision-making provides safer healthcare through improved diagnosis [32, 40]. Although collective insight is gaining traction in healthcare for its potential

in enhancing diagnostic accuracy, the method and its implication remain poorly characterized in automated medical vision literature. It has been suggested that the use of artificial intelligence can optimize these processes while considering physician workflows in clinical settings [40].

In recent times, there has been an outstanding improvement in specialized deterministic models for different medical image segmentation tasks [13, 44, 52–55, 61]. Deterministic models are notorious for choosing the most likely hypothesis even if there is uncertainty which might lead to sub-optimal segmentation. To overcome this, some models incorporate pixel-wise uncertainty for segmentation tasks, however, they produce inconsistent outputs [25,26]. Conditional variational autoencoders (c-VAE) [48], a conditional generative model, can be fused with deterministic segmentation networks to produce unlimited numbers of predictions by sampling from the latent space conditioned on the input image. Probabilistic U-net and its variants use this technique during the inference process. Here, the latent spaces are sampled from a prior network which has been trained to be similar to c-VAE [8, 29, 30]. This dependency on a prior network as well as injecting stochasticity only at the highest resolution of the segmentation network produces less diverse and blurry segmentation predictions [46]. To overcome this problem, we introduce a single inherently probabilistic model without any additional prior network that represents the collective intelligence of several experts to leverage multiple plausible hypotheses in the diagnosis pipeline (visualized in Figure 1).

Diffusion probabilistic models are a class of generative models consisting of Markov chains trained using variational inference [21]. The model learns the latent structure of the dataset by modeling the diffusion process through latent space. A neural network is trained to denoise noisy image blurred using Gaussian noise by learning the reverse diffusion process [50]. Recently, diffusion models have been found to be widely successful for various tasks such as image generation [14], and inpainting [35]. Certain approaches have also been proposed to perform semantic segmentation using diffusion models [6,59]. Here, the stochastic element in each sampling step of the diffusion model using the same pre-trained model paves the way for generating multiple segmentation masks from a single input image. However, there is still no exploration of using diffusion models for ambiguous medical image segmentation despite its high potential. In this paper, we propose the CIMD (**C**ollectivly **I**ntelligent **M**edical **D**iffusion), which addresses ambiguous segmentation tasks of medical imaging. First, we introduce a novel diffusion-based probabilistic framework that can generate multiple realistic segmentation masks from a single input image. This is motivated by our argument that the stochastic sampling process of the diffusion model can be harnessed to sample multiple plausible annotations. The stochastic sampling process also eliminates the need for a separate 'prior' distribution during the inference stage, which is critical for c-VAE-based segmentation models to sample the latent distribution for ambiguous segmentation. The hierarchical structure of our model also makes it possible to control the diversity at each time step hence making the segmentation masks more realistic as well as heterogeneous. Lastly, in order to assess ambiguous medical image segmentation models, one of the most commonly used metrics is known as GED (Generalized Energy Distance), which matches the ground truth distribution with prediction distribution. In real-world scenarios for ambiguous medical image segmentation, ground truth distributions are characterized by only a set of samples. In practice, the GED metric has been shown to reward sample diversity regardless of the generated samples' fidelity or their match with ground truths, which can be potentially harmful in clinical applications [30]. In medical practice, individual assessments are manually combined into a single diagnosis and evaluated in terms of sensitivity. When real-time group assessment occurs, the participant generates a consensus among themselves. Lastly, the minimum agreement and maximum agreement among radiologists are also considered in clinical settings. Inspired by the current practice in collective insight medicine, we coin a new metric, namely the CI score (**C**ollective **I**nsight) that considers total sensitivity, general consensus, and variation among radiologists. In summary, the following are the major contributions of this work:

- We propose a novel diffusion-based framework: Collectively Intelligent Medical Diffusion (CIMD), that realistically models heterogeneity of the segmentation masks without requiring any additional network to provide prior information during inference unlike previous ambiguous segmentation works.
- We revisit and analyze the inherent problem of the current evaluation metric, GED for ambiguous models and explain why this metric is insufficient to capture the performance of the ambiguous models. We introduce a new metric inspired by collective intelligence medicine, coined as the CI Score (**C**ollective **I**nsight).
- We demonstrate across three medical imaging modalities that CIMD performs on par or better than the existing ambiguous image segmentation networks in terms of quantitative standards while producing superior qualitative results.

2. Related Work

Ambiguous Image Segmentation. Previous work [25] models the ambiguity using approximate Bayesian inference over the network weights. However, the method is shown to produce samples that only vary pixel by pixel and can not capture the complex correlation structure of the

ground truth distribution [29]. Probabilistic U-net is capable of capturing distribution over multiple annotations that can produce a wide variety of segmentation maps from a single image [29]. The model is a combination of a U-net with a conditional variational auto-encoder that uses its stochasticity to produce an unlimited number of plausible hypotheses. This method has been shown to produce samples with limited diversity as the stochasticity is only injected in the highest resolution of the backbone segmentation network, hence the network chooses to ignore the random draws from the latent space [8]. To increase sample diversity, PHi-SegNet [8] and Hierarchical Probabilistic U-Net [30] incorporate a series of hierarchical latent spaces to sample the feature maps. The key element of their works is that the backbone networks rely on variational inference to produce multiple annotations for an image by sampling from a distribution, which, if not sufficiently complex, may not produce realistic samples [10]. The diversity of segmentation masks of these c-VAE-like models relies on an axis-aligned Gaussian latent posterior distribution, which can be too restrictive and not expressive enough to model the rich variations [46].

Diffusion Model for Image Segmentation. Diffusion models have recently shown remarkable potential in various segmentation tasks [2, 5, 6, 11] including medical images [18,27,59,60]. In fact, the stochastic sampling process of the diffusion model has been utilized to generate an implicit ensemble of segmentations that ultimately boosts the segmentation performance [59]. However, the model is only trained using a single segmentation mask per input image, hence the model has no control over the variation as well as the produced masks are not necessarily diverse. To the best of our knowledge, CIMD is the first network specifically designed to model the ambiguity of medical images by harnessing its random sampling process. Moreover, the diffusion model's hierarchical structure makes it possible to govern the ambiguity at each time step, thereby eliminating the problem of low diversity of previous methods.

3. Proposed Method

3.1. Diffusion Model

Diffusion probabilistic models have gained a lot of attention in recent years due to their ability to generate extraordinarily high-quality images compared to Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), autoregressive models, and flows. Due to their superior ability in extracting key semantics from training images, diffusion models are also being utilized for image segmentation [2, 59]. Additionally, using random Gaussian noise, it is possible to generate a distribution of segmentation rather than a deterministic output. In this section, we present a brief overview of the diffusion model framework.

Gaussian Diffusion Process. Diffusion models perform variational inference on a Markovian process using T timesteps to learn the training data distribution $p(x_0)$. The framework consists of a forward and a reverse process. During the forward process for each timestep in T, Gaussian noise is added to the image $x_0 \sim p(x_0)$ until the image becomes an isotropic Gaussian. This forward noising process is denoted as,

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I), \quad (1)$$

where $(x_0, x_1, .., x_T)$ denotes T steps in the Markov chain and α is the noise scheduler that controls the variance of the noise. In the reverse process, a neural network (f_θ) is used to create a sequence of incremental denoising operations to obtain back the clean image. f_θ learns the parameters of the reverse distribution $p(x_{t-1}|x_t) := \mathcal{N}(x_{t-1} : \mu_\theta(x_t, t), \sum_\theta(x_t, t))$. The parameters for f_θ are obtained by minimizing the KL-divergence between the forward and the reverse distribution for all timesteps. The optimization requires sampling from the distribution $q(x_t|x_{t-1})$ that subsequently requires the knowledge of x_{t-1}. Given x_0, the marginal distribution of x_t can be obtained by marginalizing out the intermediate latent variable as,

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\gamma_t}x_0, (1-\gamma_t)I). \quad (2)$$

Here, $\gamma_t = \Pi_{i=1}^{t}\alpha_i$. However, minimizing the KL-divergence between the forward and the reverse distribution can be further simplified by using a posterior distribution $q(x_t|x_{t-1}, x_0)$ instead [47]. The posterior distribution can be derived using Eq. 1 and 2 under the Markovian assumptions,

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}, \mu, \sigma^2 I), \quad (3)$$

where, $\mu(x_t, x_0) = \frac{\sqrt{\gamma_{t-1}}(1-\alpha_t)}{1-\gamma_t}x_0 + \frac{\sqrt{\alpha_t}(1-\gamma_{t-1})}{1-\gamma_t}x_t$ and $\sigma^2 = \frac{(1-\gamma_{t-1})(1-\alpha_t)}{1-\gamma_t}$. This posterior distribution is then utilized during the parameterization of the reverse Markov chain for formulating a variational lower bound on the log-likelihood of the reverse chain. During optimization, the covariance matrix for both distributions $q(x_{t-1}|x_t, x_0)$ and $p(x_{t-1}|x_t)$ are considered the same, and the mean of the distributions is predicted by f_θ. The denoising model f_θ takes noisy image x_t as input which is denoted by,

$$x_t = \sqrt{\gamma}x_0 + \sqrt{1-\gamma}\epsilon, \quad (4)$$

where, $\epsilon = \mathcal{N}(0, I)$. Now, the combination of p and q is a variational auto-encoder [28] and the variational lower bound (V_{lb}) can be expressed as,

$$\mathcal{L}_{vlb} := \mathcal{L}_0 + \mathcal{L}_1 + ... + \mathcal{L}_{T-1} + \mathcal{L}_T \quad (5)$$

$$\mathcal{L}_0 := -\log p_\theta(x_0|x_1) \quad (6)$$

$$\mathcal{L}_{t-1} := D_{KL}(q(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t)) \quad (7)$$

$$\mathcal{L}_T := D_{KL}(q(x_T|x_0)\|p(x_T)). \quad (8)$$

However, the training objective can be further simplified as [21],

$$\mathcal{L}_{simple} = \mathbb{E}_{x_0,\epsilon}|f_\theta(\tilde{x}, t) - \epsilon|_2^2, \quad (9)$$

where $\epsilon = \mathcal{N}(0, I)$. Now, log-likelihood is considered a good metric to evaluate generative models, and optimizing log-likelihood has been proven to force the model to cap-

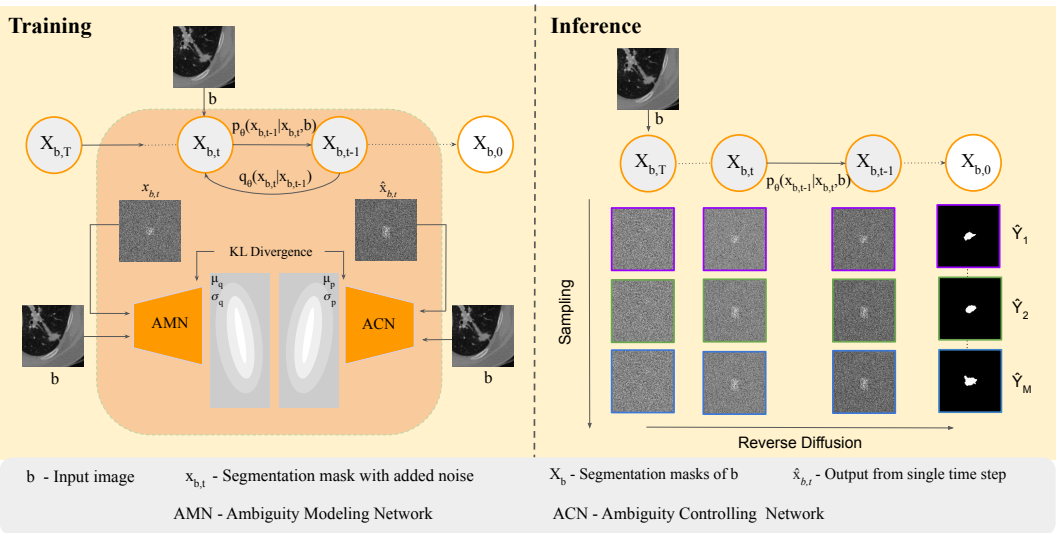

Figure 2. The graphical model of a) sampling and b) training procedure of our method. In the training phase, for every step t, the anatomical structure is induced by adding the input image b to the noisy segmentation mask $x_{b,t}$. Sampling n times with different Gaussian noise, n different plausible masks are generated.

ture all the data distribution [43] as well as improve sample quality [20]. Hence, we get the hybrid loss [38] by combining Eq. 5 and 9,

$$\mathcal{L}_{hybrid} = \mathcal{L}_{simple} + \lambda \mathcal{L}_{vlb}, \tag{10}$$

where, λ is a regularization parameter, which is used to prevent $\mathcal{L}_{vlb}$ from overwhelming $\mathcal{L}_{simple}$. Inference starts from a Gaussian noise x_t which at each timestep is iteratively denoised to get back x_{t-1} as follows,

$$x_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(x_t, t) \right) + \gamma_t z, \tag{11}$$

where, $z = \mathcal{N}(0, I)$ and $t = T, .., 1$.

3.2. Collectivly Intelligent Medical Diffusion

Let b be the given image with dimension of $C \times H \times W$ and x_b the corresponding segmentation mask. In a classical diffusion model, the input image x_b is required for training, which would result in an arbitrary segmentation mask x_0 when sampled from noise during inference. In contrast to that, to produce a segmentation mask $x_{b,0}$ for a given image b, an additional channel is concatenated to the input. This induces the anatomical information by concatenating it as an image prior to x_b and thus defining $X := b \oplus x_b$. During the noising process q, noise is added to the ground truth segmentation x_b only. As the sampling process is stochastic, the diffusion model produces different segmentation masks $x_{b,0}$ for an image b. When the diffusion model is trained using only one segmentation mask per input image, the model can implicitly generate an ensemble of segmentation masks that can be used to boost the performance of the model [59].

Now, we model the ambiguity of the ground truths using **Ambiguity Modelling Network (AMN)**. AMN models the distribution of ground truth masks given an input image. We embed this ambiguity of the segmentation masks in the latent space by parameterizing the weight of AMN by ν, given the image b and the ground truth x_b. This probability distribution denoted as Q is modeled as a Gaussian with mean $\mu(b, x_b; \nu) \in R^N$ and variance $\sigma(b, x_b; \nu) \in R^{N \times N}$ where N denotes the low dimensional latent space. The latent space is characterized by,

$$z_q \sim Q(.|b, x_b) = \mathcal{N}(\mu(b, x_b; \nu), \sigma(b, x_b; \nu)). \tag{12}$$

Similarly, we model the ambiguity of predicted masks using **Ambiguity Controlling Network (ACN)**. ACN models the noisy output from the diffusion model conditioning on an input image. For each time step t, assuming $\hat{x_b} = f_\theta(\tilde{x_b}, t)$, we estimate the ambiguity of our diffusion model by parameterizing the weight of ACN, ω as a probability distribution P with mean $\mu(b, \hat{x_b}; \omega) \in R^N$ and variance $\sigma(b, \hat{x_b}; \omega) \in R^{N \times N}$ as follows

$$z_p \sim P(.|b, \hat{x_b}) = \mathcal{N}(\mu(b, \hat{x_b}, t); \omega), \sigma(b, \hat{x_b}, t); \omega)). \tag{13}$$

Both networks **AMN** and **ACN** are modeled using an axis-aligned gaussian distribution with diagonal covariance matrices. The architectural details of both networks can be found in the supplementary. We penalize the difference between two distributions by imposing a Kullback-Leibler divergence,

$$\mathcal{L}_{amb} = D_{KL}(Q(z|x_b, b) \| P(z|\hat{x_b}, b)). \tag{14}$$

Finally, by modifying Eq. 10, all losses are combined as a weighted sum with a regularizing factor β as

$$\mathcal{L}_{total} = \mathcal{L}_{simple} + \lambda \mathcal{L}_{vlb} + \beta \mathcal{L}_{amb}. \tag{15}$$

During the sampling process, for $X_t := b \oplus x_{b,t}$, Eq. 11 is modified as,

$$x_{b,t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(x_{b,t} - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(X_t, t) \right) + \gamma_t z, \tag{16}$$

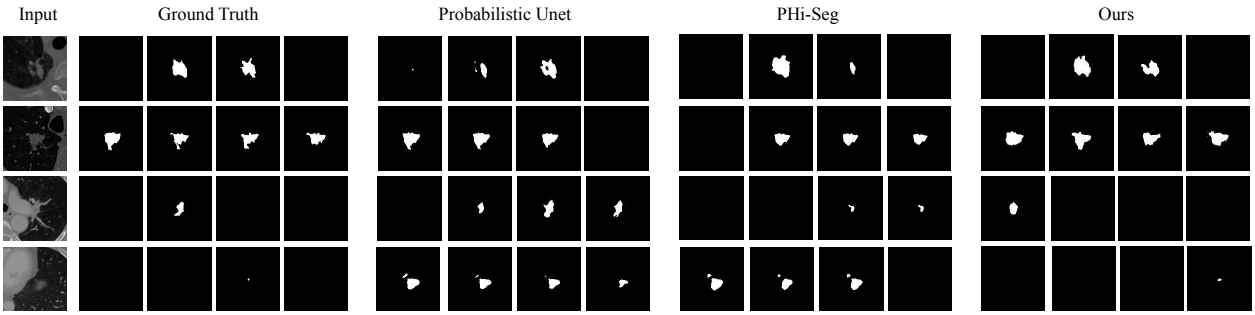

Figure 3. Comparative qualitative analysis with the two baseline methods – Probabilistic U-net [29] and PHi-Seg [8]. Sample images from the LIDC-IDRI dataset with 4 available expert gradings are shown on the left. Note that empty segmentation masks are also valid grading. For a fair comparison, we visualize only the first 4 sampled segmentation masks from the segmentation networks.

where, $z = \mathcal{N}(0, I)$ and $t = T, .., 1$. The graphical model of proposed approach is illustrated in Figure 2.

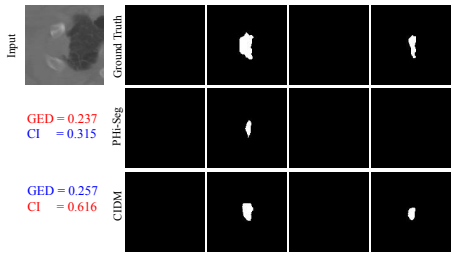

Figure 4. Visual analysis of the GED vs. the CI score for the LIDC-IDRI lung CT dataset. It can be observed that GED is lower for PHi-Seg even though it failed to segment most of the lesions. However, the combined sensitivity penalizes under segmentation hence the CI score is lower in that case. Red corresponds to better and blue corresponds to a lower score.

3.3. Collective Insight Score

To recap, ambiguous segmentation models generate a distribution of predictions rather than a deterministic one and are evaluated against a distribution of ground truths. Although Generalized Energy Distance (GED) has been used before for assessing ambiguous segmentation models, this metric has been found to be inadequate as it disproportionately rewards sample diversity regardless of its match with the ground truth samples [30]. This can be potentially dangerous, particularly in pathological cases. In Figure 4, we can observe how GED is unduly rewarding PHi-Seg even though CIDM outputs are qualitatively better. To this end, we propose an alternative evaluation metric called the CI score (**C**ollective **I**nsight), and explain the motivation behind each component of the metric in this section. The CI score is defined as,

$$CI = \frac{3 \times S_c \times D_{max} \times D_a}{S_c + D_{max} + D_a}, \tag{17}$$

where, S_c denotes the combined sensitivity, D_{max} denotes the maximum Dice matching score and D_a is the diversity agreement score. CI takes the harmonic mean of each component to equalize the weights of each part. In the following sections, true positive, false negative, and false positive are denoted as TP, FN, and FP, respectively. The collection of all ground truths is denoted as $\mathbf{Y} = \{Y_1, Y_2, .., Y_M\}$ where $Y_1, Y_2, .., Y_M$ corresponds to the individual ground truth of each sample. M here is the number of ground truths. Similarly, the collection of all predictions is denoted as $\hat{\mathbf{Y}} = \{\hat{Y}_1, \hat{Y}_2, .., \hat{Y}_N\}$ where $\hat{Y}_1, \hat{Y}_2, .., \hat{Y}_N$ corresponds to the individual predictions of each sample. N here is the number of predictions. Although the number of ground truths is limited, the model can generate an unlimited number of predictions, hence M and N are not necessarily equal. A visual diagram in Figure 6 illustrates the operation of each component.

Combined Sensitivity. In clinical practice, all the diagnoses from different raters are combined into a single collective decision. The final decision is then usually assessed in terms of true positive rate (sensitivity) [16]. In many branches of medical diagnosis, the primary goal is to maximize the true positive rate while maintaining a tolerable degree of false positive rate [3, 33, 58]. Hence, we argue that assessing the combined sensitivity directly aligns with the interest of common clinical practice. Since empty ground truth is also a valid prediction, we consider sensitivity to be 1 in those instances. First, we define the combined ground truth Y_c which is the union of all ground truths maps. Similarly, we define combined predictions $\hat{Y}_c$ which is the union of all prediction maps. Y_c and $\hat{Y}_c$ are mathematically formulated as follows:

$$Y_c = \bigcup_{i=1}^{M} Y_i, \qquad \hat{Y}_c = \bigcup_{j=1}^{N} \hat{Y}_j. \tag{18}$$

We calculate the combined sensitivity S_c between the combined predictions and combined ground truths as follows:

$$S_c(\hat{Y}_c, Y_c) = \begin{cases} \frac{TP}{TP+FN}, & \text{if } \hat{Y}_c \cup Y_c \neq \emptyset \\ 1, & \text{if otherwise.} \end{cases} \tag{19}$$

Maximum Dice Matching. In medical diagnosis cases, empty sets, which indicate no abnormalities are also valid diagnoses. However, in this case, the Dice metric will be undefined, hence we set Dice = 1 in those cases. Thus, the

Dice score is defined as:

$$Dice(\hat{Y}, Y) = \begin{cases} \frac{2|Y \cap \hat{Y}|}{|Y|+|\hat{Y}|}, & \text{if } Y \cup \hat{Y} \neq \emptyset \\ 1, & \text{otherwise.} \end{cases} \quad (20)$$

In collective intelligence practice, it is common to assess the diagnosis of students against the experts [7, 31]. We emulate the process by calculating the Dice scores of individual predictions with all the ground truths and then calculating the maximum Dice score among all these pairs. First, we define the set of all Dice scores $\mathbf{D}_i$ for each individual ground truth Y_i as follows:

$$\mathbf{D}_i = \{Dice(\hat{Y}_1, Y_i), Dice(\hat{Y}_2, Y_i), ...Dice(\hat{Y}_N, Y_i)\}, \quad (21)$$

where $\mathbf{D}_i$ is a collection of Dice scores calculated between each ground truth Y_i and all the provided predictions. Then, we take the maximum Dice score among this set and report the average as the maximum Dice match D_{max}. D_{max} is formulated as follows:

$$D_{max} = \frac{1}{M} \sum_{i=1}^{M} \max(\mathbf{D}_i). \quad (22)$$

Diversity Agreement. For ambiguous models, the evaluation of the diversity of the predicted outputs can be tricky. Disproportionally rewarding diversity regardless of their match with the ground truth samples can be potentially misleading. On the other hand, the lack of diversity in predicted samples can indicate that model is rather deterministic than stochastic. Hence we consider matching maximum and minimum variance between two raters as they indicate minimum agreement and maximum agreement between two raters respectively. Here, we first calculate the variance between all pairs in ground truth distribution for a single input image. Then, we take the minimum and maximum variance. We define the minimum variance as V^Y_{min} and maximum variance as V^Y_{max}. Similarly, we calculate the variance between all pairs in the prediction distribution for that input and take the minimum and maximum variance. We define them as $V^{\hat{Y}}_{min}$ and $V^{\hat{Y}}_{max}$ respectively. The difference between the minimum variance of ground truth and prediction distribution for a particular input can be expressed as $\Delta V_{min} = |V^Y_{min} - V^{\hat{Y}}_{min}|$. Similarly, the difference between the maximum variance of ground truth and prediction distribution for a particular input is expressed as $\Delta V_{max} = |V^Y_{max} - V^{\hat{Y}}_{max}|$. Finally, we define the diversity agreement D_a as,

$$D_a = 1 - \left(\frac{\Delta V_{max} + \Delta V_{min}}{2} \right). \quad (23)$$

4. Experiments

4.1. Datasets

Lung lesion segmentation (LIDC-IDRI). This publicly available dataset contains 1018 lung CT scans from 1010 subjects with manual annotations from four domain experts making it a good representation of typical CT image ambiguity [4]. A total of 12 radiologists provided annotation masks for this dataset. We use the dataset after the second reading where the experts were shown the annotations of other radiologists that allowed them to make new adjustments. The training set contains 13511 and the test set contains 1585 lesion images with 4 expert gradings.

Bone surface segmentation (B-US). Bone segmentation from ultrasound (US) imaging often results in a low inter-rater agreement [19]. After obtaining institutional review board (IRB) approval 2000 US scans were collected from 30 healthy subjects. The scans were collected using a 2D C5-2/60 curvilinear probe and an L14-5 linear probe using the Sonix-Touch US machine (Analogic Corporation, Peabody, MA, USA). Depth settings and image resolutions varied between 3–8 cm, and 0.12–0.19 mm, respectively. All the collected scans were manually segmented by an expert ultrasonographer and three novice users who were trained to perform bone segmentation. The training set contains 1769 and the test set contains 211 bone ultrasound scans.

Multiple sclerosis lesion segmentation (MS-MRI). This publicly available dataset contains 84 longitudinal MRI scans from 5 subjects with a mean of 4.4 time-points [12]. The white matter lesions associated with MS are delineated by two domain expert raters one with four and the other with ten years of experience. Both experts were blinded to the temporal ordering of the MRI scans. From the volumetric MRI, we convert each slice into a 2D image with corresponding segmentation masks. Each data point contains proton density (PD), Flair, MP RAGE, and T2 MRI scans. The training set contains 6012 from 4 patients and the test set contains 1411 scans from 1 patient.

4.2. Implementation Details

The proposed method is implemented using the PyTorch framework. We set the time step as $T = 1000$ with a linear noise schedule for all the experiments. The U-Net-like diffusion model's weights and biases are optimized using an Adam optimizer with a learning rate of 10^{-4}. For all the experiments $\lambda = 0.001$ is set to regularize $\mathcal{L}_{vlb}$. We also set $\beta = 0.001$ to prevent $\mathcal{L}_{amb}$ from overwhelming both $\mathcal{L}_{simple}$ and $\mathcal{L}_{vlb}$. We chose 128×128 as the resolution for LIDC-IDRI, 256×256 as the resolution for Bone-US, and 64×64 as the resolution for the MS-MRI dataset. For MS-MRI we concatenate all four MRI scans and use them as an image prior before feeding it to the diffusion model. The rest of the parameters for diffusion in the model are the same as [38].

4.3. Evaluation Metrics

Generalized Energy Distance. A commonly used metric in ambiguous image segmentation tasks that leverages distance between observations by comparing the distribution of segmentations [29]. It is given by [9, 45, 51],

$$D^2_{GED}(P_{gt}, P_{out}) = 2\mathbb{E}[d(S, Y)] - \mathbb{E}[d(S, S')] - \mathbb{E}[d(Y, Y')], \quad (24)$$

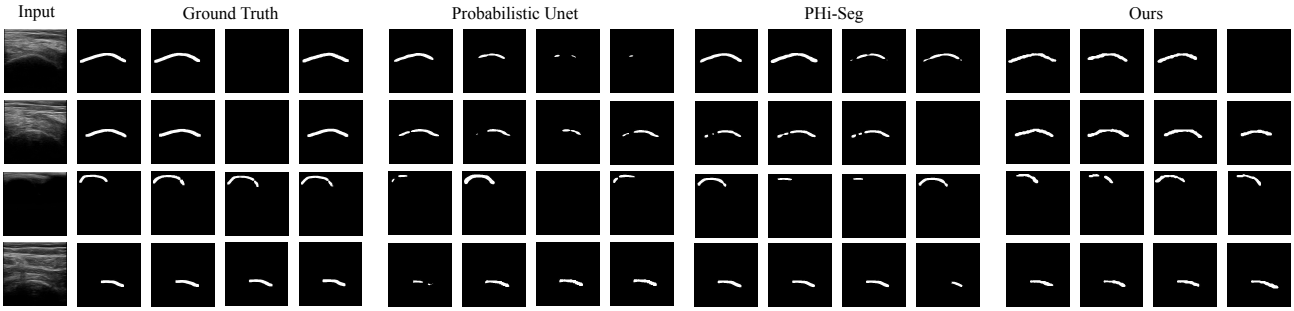

Figure 5. Comparative qualitative analysis with the two baseline methods Probabilistic U-net [29] and PHi-Seg [8]. Examples of the Bone-US dataset with 1 expert and 3 novice gradings are shown on the left. The bone-US dataset has a comparatively small inter-rater disagreement. We sample the first 4 segmentation masks from prediction distribution.

Table 1. Comparison of quantitative results in terms of GED, CI, and D_{max} for all the datasets with state-of-the-art ambiguous segmentation networks. The best results are in **Bold** and we achieve state-of-the-art results in terms of D_{max} and CI score across all datasets.

Method	LIDC-IDRI [4]			Bone Segmentation			MS-Lesion [12]		
	GED (↓)	CI(↑)	D_{max}(↑)	GED (↓)	CI(↑)	D_{max}(↑)	GED (↓)	CI(↑)	D_{max}(↑)
Probabilistic Unet [29]	0.353	0.731	0.892	0.390	0.738	0.844	0.749	0.514	0.502
PHi-Seg [8]	**0.270**	0.736	0.904	0.312	0.7544	0.848	0.681	0.518	0.506
Generalized Probabilistic U-net [10]	0.299	0.707	0.905	**0.289**	0.7501	0.863	**0.678**	0.522	0.513
$CIMD$ (Ours)	0.321	**0.759**	**0.915**	0.295	**0.7578**	**0.889**	0.733	**0.560**	**0.562**

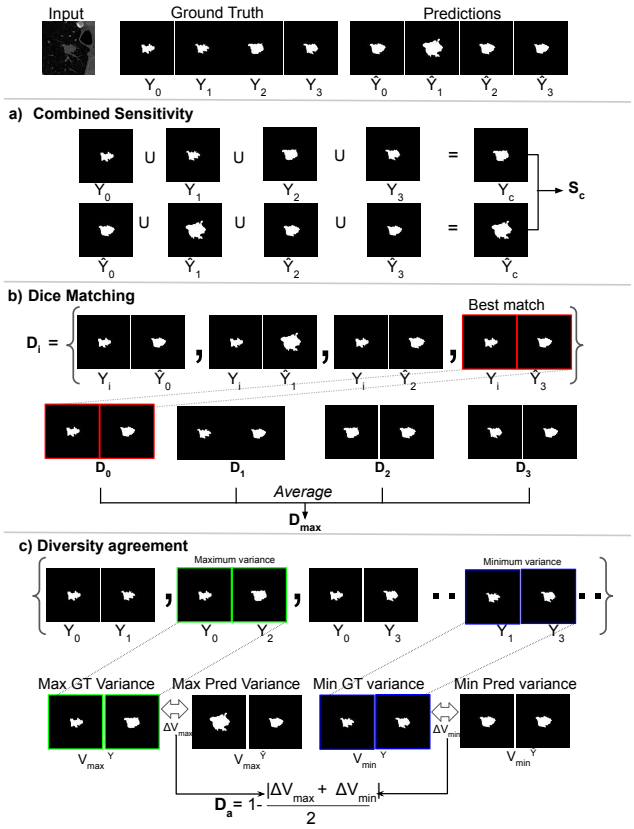

Figure 6. Visual representation of each component a) Combined Sensitivity, b) Dice Matching, and c) Diversity agreement of the proposed CI score metric. GT denotes ground truth and Pred denotes predictions. Here, the number of ground truth $M = 4$. In c) we only demonstrate the variance in ground truth distribution, however, in practice we calculate it for both ground truth and prediction distribution.

where, d corresponds to the distance measure $d(x, y) = 1 - IoU(x, y)$, Y and Y' are independent samples of P_{gt} and S and S' are sampled from P_{out}. Lower energy indicates better agreement between prediction and the ground truth distribution of segmentations.

4.4. Comparison with the Baseline methods.

To the best of our knowledge, there exists no other work that has considered explicitly modeling the ambiguity of medical images to produce multiple segmentation maps using diffusion models. We compare our approach with current state-of-the-art methods that are specifically designed to capture a distribution over multi-modal segmentation.

Probabilistic U-net and its variants. We report the results for the current state-of-the-art method for ambiguous medical image segmentation network probabilistic U-net [29]. We train a probabilistic U-net for the LIDC-IDRI dataset using the same parameter reported in the paper. For Bone-US and MS-MRI, we train the probabilistic U-net with $\beta = 10$ until the loss doesn't improve. Additionally, we compare with a variant of probabilistic U-net, namely generalized probabilistic U-net [10], where instead of using axis-aligned Gaussian distribution to model prior and posterior network, they use a mixture of full covariance Gaussian distributions.

PHi-Seg. One of the major problems of probabilistic U-net is the lack of diversity in the predicted sample, as the stochasticity is only injected into the highest resolution. To solve this, PHi-Seg [8] adopts a hierarchical structure inspired by Laplacian Pyramids, where the model generated conditional segmentation by refining the distribution at increasingly higher resolution, hence producing better quality samples. We train PHi-Seg using the parameters reported in the paper for all the datasets.

Quantitative Comparison: We consider both GED and CI

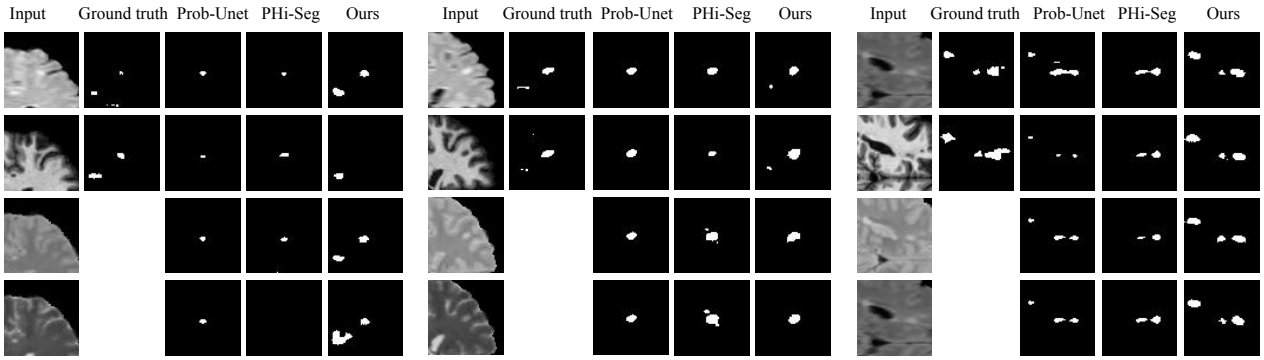

Figure 7. Comparative qualitative analysis with the two baseline methods Probabilistic U-net [29] and PHi-Seg [8]. Examples of the MS-MRI dataset with 2 expert gradings are shown here. We sample the first 4 segmentation masks from the prediction distribution.

Table 2. Ablation study: We perform an ablation study on LIDC-IDRI dataset [4] to better understand the contributions incorporated in the CIMD method.

Method	GED (↓)	CI (↑)	D_{max} (↑)
DDPM-det-Seg [59]	1.081	0.616	0.548
DDPM-Prob-Seg	0.417	0.683	0.689
CIMD (Ours)	**0.321**	**0.759**	**0.915**

scores to evaluate the performance of different ambiguous medical image segmentation models. We tabulate the results in Table 1 evaluating 4 samples from the prediction distribution. We also separately report maximum Dice-matching scores. Our method outperforms other state-of-the-art networks in terms of both D_{max} and CI scores. In terms of GED, we get on par performance with the Probabilistic U-net. A high D_{max} score indicates the generated samples are in good match with the ground truth distribution and the CI score ensures the sample diversity matches with ground truth diversity.

Qualitative Comparison: As the evaluation of ambiguous networks is difficult to characterize, we argue that qualitative results can be a good indicator of network performance, especially for difficult cases. We show the predictions from the test dataset for all the models in Figure 3, 5, and 7. It can be seen that CIMD achieves visually superior and diverse results compared to the previous state-of-the-art methods. From Figure 3 it can be observed that the model was able to capture the frequencies of blanks as well as maintained diversity in difficult cases. CIMD works especially well on ultrasound modalities with minimal error as can be seen in Figure 5. From Figure 7 it can be seen that CIMD is able to capture all the lesions even if they have small structures while maintaining diversity in segmentation masks. As CIMD injects stochasticity at each hierarchical feature representation, it demonstrates diverse and accurate segmentation in all datasets.

5. Discussion

Ablation Study For the ablation study, we evaluate the proposed contribution. In particular, we compare our contribution with the original DDPM Model, which is inherently stochastic, hence capable of generating several segmentation masks from a single input image. We demonstrate the contribution in detail in our ablation study.

DDPM for segmentation. We report two sets of results for DDPM-based segmentation. *DDPM-det-Seg* refers to the model where DDPM is trained using the average of all segmentation masks for an input image. *DDPM-Prob-Seg* refers to training the DDPM with different segmentation masks for an input image. The results for the LIDC dataset can be observed in Table 2. It can be observed that even though the DDPM sampling process is stochastic, the distribution of generated segmentation masks is not diverse enough as well as are not similar to the ground truth distribution. Additional ablation results are in supplementary.

Limitations: Although our proposed method produces diverse-yet-meaningful predictions, it is quite slow at training and inference due to the inherent nature of the diffusion process. Also, training a model for ambiguous segmentation requires annotations from multiple radiologists which is costly and time-consuming. In the future, we plan on extending our method trying to tackle these limitations while also extending the approach to more modalities and 3D volumetric datasets.

6. Future work and Conclusion

In this work, we introduce a diffusion-based ambiguous segmentation network that can generate multiple plausible annotations from a single input image. Unlike traditional cVAE-based networks, CIMD uses its hierarchical structure to incorporate stochasticity at each level and doesn't require a separate network to encode prior information about the image during the inference stage. Our method can be incorporated into any diffusion-based framework with minimal additional training. For future work, it is possible that CIMD can be extended to more general computer vision problems as well as tested for other medical imaging modalities. Lastly, our approach can be incorporated in other specialized diffusion-based segmentation networks, e.g. MedSegDiff [60], SegDiff [60], etc for higher fidelity segmentation outputs.

References

[1] Hillel R Alpert and Bruce J Hillman. Quality and variability in diagnostic radiology. *Journal of the American College of Radiology*, 1(2):127–132, 2004.

[2] Tomer Amit, Eliya Nachmani, Tal Shaharbany, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

[3] Giuseppe Argenziano, Gabriella Fabbrocini, Paolo Carli, Vincenzo De Giorgi, Elena Sammarco, and Mario Delfino. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Archives of dermatology*, 134(12):1563–1570, 1998.

[4] Samuel G Armato III, Geoffrey McLennan, Michael F McNitt-Gray, Charles R Meyer, David Yankelevitz, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, Ella A Kazerooni, Heber MacMahon, et al. Lung image database consortium: developing a resource for the medical imaging research community. *Radiology*, 232(3):739–748, 2004.

[5] Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal M Patel. Ddpm-cd: Remote sensing change detection using denoising diffusion probabilistic models. *arXiv preprint arXiv:2206.11892*, 2022.

[6] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[7] Michael L Barnett, Dhruv Boddupalli, Shantanu Nundy, and David W Bates. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. *JAMA network open*, 2(3):e190096–e190096, 2019.

[8] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer, 2019.

[9] 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.

[10] Ishaan Bhat, Josien PW Pluim, and Hugo J Kuijf. Generalized probabilistic u-net for medical image segementation. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 113–124. Springer, 2022.

[11] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4175–4186, 2022.

[12] 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.

[13] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[14] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[15] Kunio Doi, Heber MacMahon, Shigehiko Katsuragawa, Robert M Nishikawa, and Yulei Jiang. Computer-aided diagnosis in radiology: potential and pitfalls. *European journal of Radiology*, 31(2):97–109, 1999.

[16] Stephan D Fihn. Collective intelligence for clinical diagnosis—are 2 (or 3) heads better than 1? *JAMA network open*, 2(3):e191071–e191071, 2019.

[17] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019.

[18] Xutao Guo, Yanwu Yang, Chenfei Ye, Shang Lu, Yang Xiang, and Ting Ma. Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation. *arXiv preprint arXiv:2210.17408*, 2022.

[19] Ilker Hacihaliloglu. Ultrasound imaging and segmentation of bone surfaces: A review. *Technology*, 5(02):74–80, 2017.

[20] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[22] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–548. Springer, 2019.

[23] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255. IEEE, 2019.

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021.

[25] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[26] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[27] Boah Kim, Yujin Oh, and Jong Chul Ye. Diffusion adversarial representation learning for self-supervised vessel segmentation. *arXiv preprint arXiv:2209.14566*, 2022.

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[29] 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. *Advances in neural information processing systems*, 31, 2018.

[30] 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.

[31] Olga Kunina-Habenicht, Wolf E Hautz, Michel Knigge, Claudia Spies, and Olaf Ahlers. Assessing clinical reasoning (asclire): Instrument development and validation. *Advances in Health Sciences Education*, 20(5):1205–1224, 2015.

[32] Ralf HJM Kurvers, Stefan M Herzog, Ralph Hertwig, Jens Krause, Patricia A Carney, Andy Bogart, Giuseppe Argenziano, Iris Zalaudek, and Max Wolf. Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, 113(31):8777–8782, 2016.

[33] Ralf HJM Kurvers, Jens Krause, Giuseppe Argenziano, Iris Zalaudek, and Max Wolf. Detection accuracy of collective intelligence assessments for skin cancer diagnosis. *JAMA dermatology*, 151(12):1346–1353, 2015.

[34] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020.

[35] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[36] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[37] Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in Neural Information Processing Systems*, 33:12756–12767, 2020.

[38] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[39] Di Qiu and Lok Ming Lui. Modal uncertainty estimation via discrete latent representation. *arXiv preprint arXiv:2007.12858*, 2020.

[40] Kate Radcliffe, Helena C Lyson, Jill Barr-Walker, and Urmimala Sarkar. Collective intelligence in medical decision-making: a systematic scoping review. *BMC medical informatics and decision making*, 19(1):1–11, 2019.

[41] Aimon Rahman, Wele Gedara Chaminda Bandara, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Orientation-guided graph convolutional network for bone surface segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 412–421. Springer, 2022.

[42] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Simultaneous bone and shadow segmentation network using task correspondence consistency. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*, pages 330–339. Springer, 2022.

[43] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[45] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

[46] Raghavendra Selvan, Frederik Faye, Jon Middleton, and Akshay Pai. Uncertainty quantification in medical image segmentation with normalizing flows. In *International Workshop on Machine Learning in Medical Imaging*, pages 80–90. Springer, 2020.

[47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[48] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[49] Katharina Sonderegger-lseli, Stefanie Burger, Jörg Muntwyler, and Franco Salomon. Diagnostic errors in three medical eras: a necropsy study. *The Lancet*, 355(9220):2027–2031, 2000.

[50] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[51] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

[52] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International*

Conference on Medical Image Computing and Computer-Assisted Intervention, pages 36–46. Springer, 2021.

[53] Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: Mlp-based rapid medical image segmentation network. *arXiv preprint arXiv:2203.04967*, 2022.

[54] Jeya Maria Jose Valanarasu, Vishwanath A Sindagi, Ilker Hacihaliloglu, and Vishal M Patel. Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 363–373. Springer, 2020.

[55] Jeya Maria Jose Valanarasu, Vishwanath A Sindagi, Ilker Hacihaliloglu, and Vishal M Patel. Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation. *IEEE Transactions on Medical Imaging*, 41(4):965–976, 2021.

[56] M Visser, DMJ Müller, RJM van Duijn, Marion Smits, N Verburg, EJ Hendriks, RJA Nabuurs, JCJ Bot, RS Eijgelaar, M Witte, et al. Inter-rater agreement in glioma segmentations on longitudinal mri. *NeuroImage: Clinical*, 22:101727, 2019.

[57] Max Wolf, Jens Krause, Patricia A Carney, Andy Bogart, and Ralf HJM Kurvers. Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. *PloS one*, 10(8):e0134269, 2015.

[58] Max Wolf, Ralf HJM Kurvers, Ashley JW Ward, Stefan Krause, and Jens Krause. Accurate decisions in an uncertain world: collective cognition increases true positives while decreasing false positives. *Proceedings of the Royal Society B: Biological Sciences*, 280(1756):20122777, 2013.

[59] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. *arXiv preprint arXiv:2112.03145*, 2021.

[60] Junde Wu, Huihui Fang, Yu Zhang, Yehui Yang, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. *arXiv preprint arXiv:2211.00611*, 2022.

[61] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.