Supplementary Material for 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

A. Appendix Ablation Study

We additionally report the ablation study for both the Bone-US dataset and MS-MRI [4] dataset. As we can observe from Table 1 and Table 2 that CIMD improves the diffusion model performance in terms of both GED and CI scores. We visualize the ablation study results for the LIDC-IDRI dataset [1] in Figure 1. DDPM-det-Seg is the diffusion model [6, 9] trained using the average of all four segmentation masks. Although the sampling process is stochastic, we see minimal changes in generated segmentation masks. DDPM-Prob-Seg is trained using all the segmentation masks. In other words, different segmentation masks are used in each forward pass for an input image. It can be seen that although there are some variations in segmentation masks, most of them are empty. In contrast to that, CIMD is able to segment the lesion as well as produce different segmentation masks that match the ground truth distributions. This proves DDPM itself is not able to model the stochasticity of the dataset alone.

Table 1. Ablation study: we perform an ablation study on the Bone-US dataset to better understand the contributions incorporated in the CIMD method.

Method	GED (↓)	CI (↑)	$Dice_{max}$ (↑)
DDPM-det-Seg [9]	0.887	0.673	0.626
DDPM-Prob-Seg	0.798	0.675	0.627
CIMD (Ours)	**0.295**	**0.757**	**0.889**

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

Method	GED (↓)	CI (↑)	$Dice_{max}$ (↑)
DDPM-det-Seg [9]	0.799	0.507	0.497
DDPM-Prob-Seg	0.804	0.509	0.499
CIMD (Ours)	**0.733**	**0.560**	**0.562**

B. Appendix Network Architecture

AMN and ACN architecture. AMN (Ambiguity Modeling Network) and ACN (Ambiguity Controlling Network) have the same architecture, which is an encoder consisting of repeated application of four 3x3 convolution layers with 32, 64, 128, and 192 filters each followed by a rectified linear unit (ReLU) and a 2x2 average pooling with stride 2 for down-sampling. Then we add a 1x1 convolution layer that takes the global average pooled feature maps from the previous layer as input and predicts the Gaussian distribution which is parameterized by mean and variance. AMN takes the concatenation of the input image with the ground truths as input and predicts the Gaussian distribution of the segmentation masks conditioned on an input image. ACN takes the concatenation of the input image with the predictions as input and predicts the Gaussian distribution of predicted masks conditioned on the input image.

C. Appendix Training details

β **Parameter.** The regularization parameter β is empirically chosen to be 0.001, as higher β overwhelms the other loss terms and produces noisy outputs. Lower β that 0.001 ignores the KL divergence between ACN and AMN, hence network acts like a regular diffusion model with minimal variations in outputs.

D. Appendix Qualitative Result Analysis

Average Segmentation Quality. We visualize 16 samples for each input image from the test set distribution to assess their quality. For both Prob-Unet [5] and PHi-Seg [2] we can observe from Figure 4 and Figure 5 that although there are some segmentation masks that are close to ground truth (therefore, not affecting the quantitative metric much), not all segmentation masks are complete or consistent. This happens because they are sampled using different latent variables which might not always produce high-fidelity samples. However, CIMD is observed to consistently produce high-fidelity samples as the model doesn't

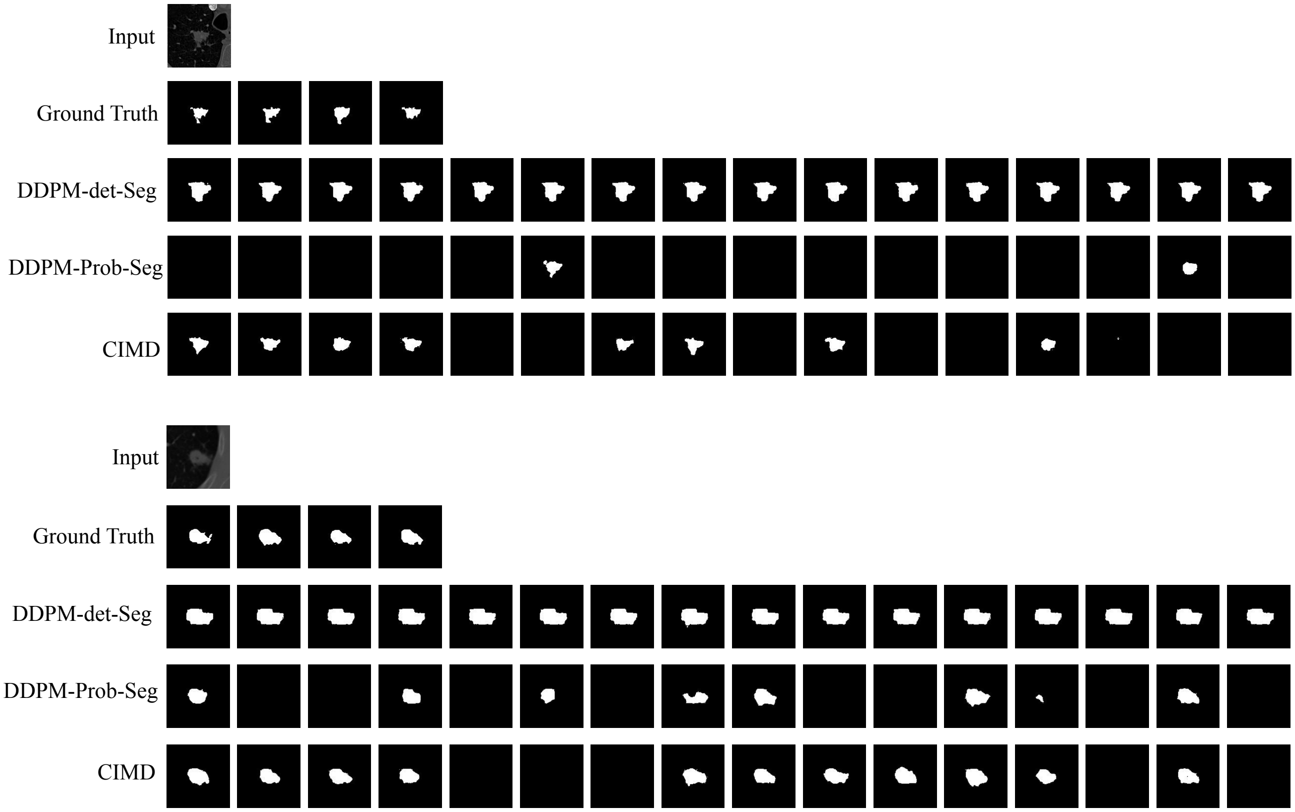

Figure 1. Visualization of ablation study for LIDC-IDRI [1] dataset. DDPM-det-Sg is trained using the average of all segmentation masks of one input image. DDPM-Prob-Seg is trained using all segmentation masks of one input image.

depend on latent variables from a prior model for segmentation.

Empty Segmentation in Bone-US dataset. In ultrasound, the high acoustic impedance mismatch between soft tissue and bone surface produces a high contrast curve-linear region. This high-contrast region indicates the presence of the bone surface. However, this response can be extremely noisy due to the nature of ultrasound imaging. In our dataset, some ultrasound scan doesn't have any bone surface response, hence all four raters annotated them as empty masks. From Figure 6 we can observe that some latent variables from both Prob-Unet and PHi-Seg struggle to ignore random contrast in ultrasound imaging, and segment those regions as bone surfaces. On the other hand, CIMD produces much more consistent results when the bone surface is not present with minimal error.

Fine Lesion segmentation. As MS-MRI [4] dataset contains images with very fine lesions, it is difficult for other networks to segment it. However, from Figure 2 it can be observed that CIMD is able to segment even the finest lesion from MRI scans.

E. Choice of Distribution

Table 3. Quantitative results using LIDC-IDRI [1] dataset using CIMD with axis-aligned Gaussian (CIMD-AA) and full-covariance matrix (CIMD-FC).

Method	GED (↓)	CI (↑)	$Dice_{max}$ (↑)
CIMD-FC [9]	0.447	**0.774**	0.718
CIMD-AA	**0.321**	0.759	**0.915**

In this section, we discuss the choice of distribution for AMN and ACN. The previous approach modeled the ambiguity of the segmentation masks using multivariate Gaussian with diagonal covariance matrix [2, 5]. It has been assumed that the choice of a simple distribution restricts the sample diversity [7]. It has been hypothesized that the use of a full covariance matrix will produce a more diverse sample [3]. Generalized probabilistic U-net proposed the use of a full covariance matrix to model the distribution of segmentation masks [3]. Since the constraint of a valid covariance matrix is difficult to impose while training a network, the covariance matrix Σ is built using Cholesky decompo-

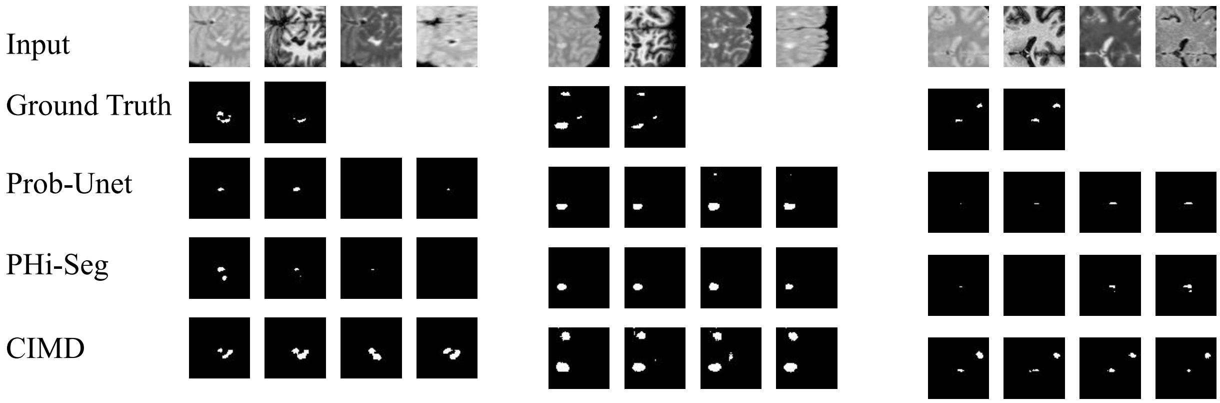

Figure 2. Qualitative comparison of MS-MRI [4] dataset between Probabilistic U-net [5], PHi-Seg [2] and CIMD. We can observe that MS lesions have a very fine structure, hence both Prob-Unet and PHi-Seg are both failing to capture them. On the other hand, CIMD is able to capture even the smallest lesion that is present in the scan.

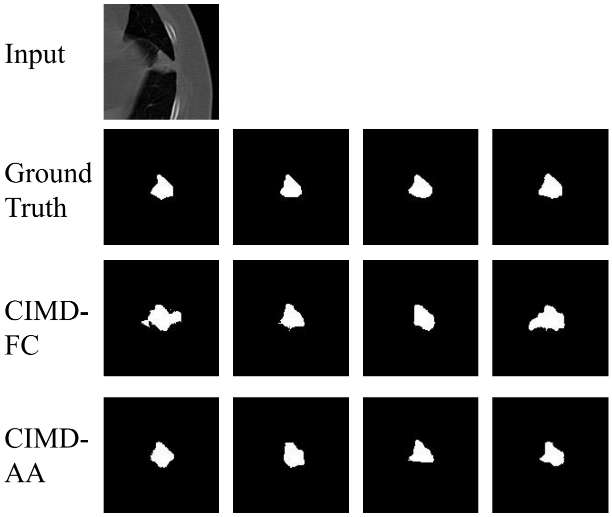

Figure 3. Comparative qualitative analysis between CIMD-FC and CIMD-FC. CIMD-FC denotes CIMD with a full covariance matrix and CIMD-AA denotes CIMD with axis-aligned Gaussian.

sition L [8].

$$\Sigma = LL^T \tag{1}$$

Here, L is a positive valued diagonal lower-triangular matrix, which is computed by a neural network. The samples are drawn using the reparametrizing trick,

$$z = \mu + L * \epsilon, \epsilon \sim \mathcal{N}(0, I) \tag{2}$$

Hence, we model the Gaussians of AMN and ACN with a full covariance matrix to observe its effect in the CIMD network. Although the CIMD with full covariance matrices were able to produce outputs with high diversity, they are always not close to ground truth distribution which can be observed from the D_{max} in Table 3. Moreover, CIMD with a full covariance matrix produces more coarse outputs hence the combined sensitivity is higher in this case. This skews the CI score, however, from D_{max} and qualitative results in Figure 3 we can observe that axis-aligned performed better.

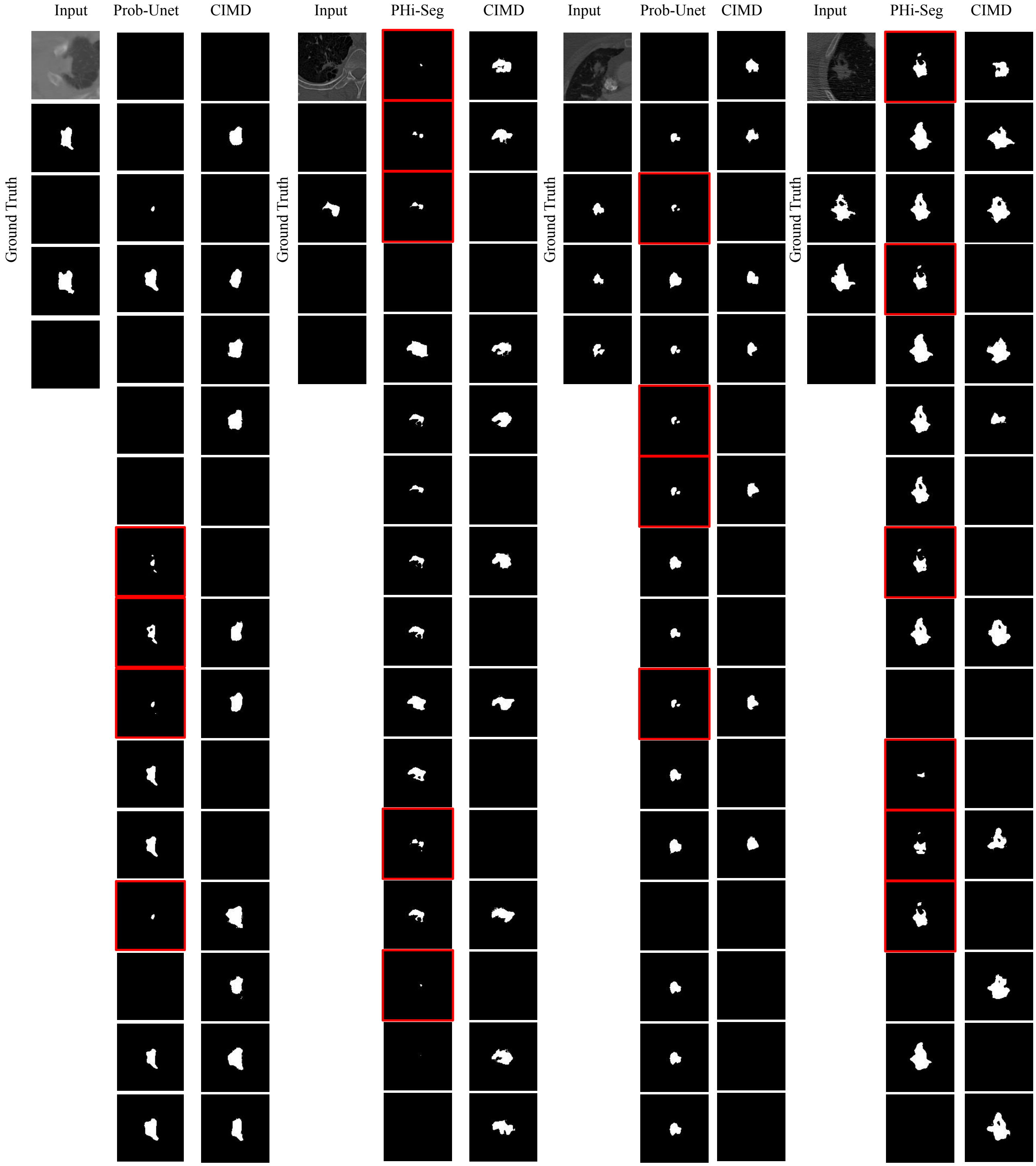

Figure 4. Comparative qualitative analysis with the two baseline methods Probabilistic U-net [5] and PHi-Seg [2] for LIDC-IDRI [1] dataset. Here we show 16 samples from each model. The red boxes indicate incomplete or noisy segmentation masks. Here we can observe some incomplete or noisy output from baseline methods while all 16 samples from CIMD have high fidelity.

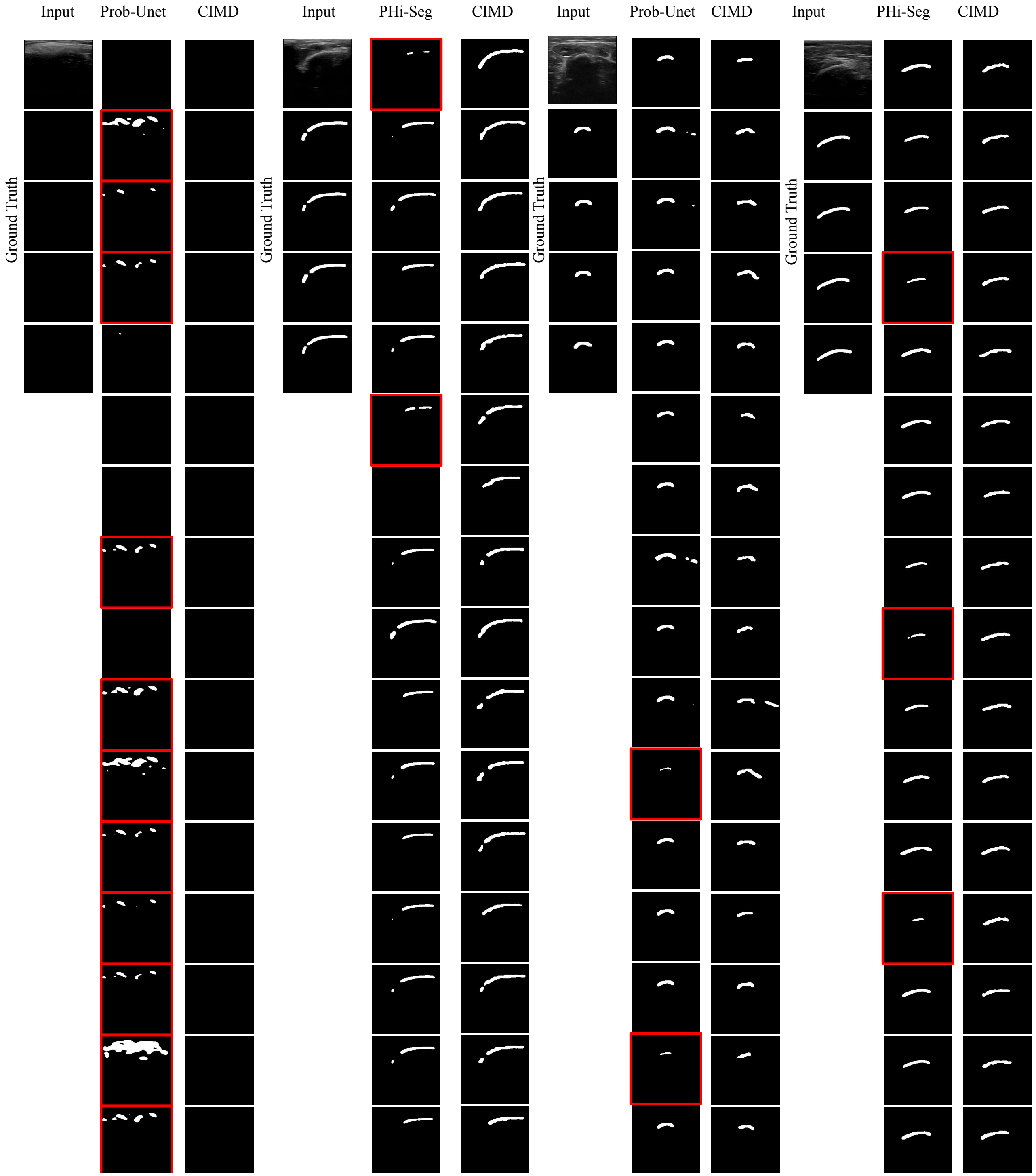

Figure 5. Comparative qualitative analysis with the two baseline methods Probabilistic U-net [5] and PHi-Seg [2] for Bone-US dataset. Here we show 16 samples from each model. The red boxes indicate incomplete or noisy segmentation masks. Here we can observe some incomplete or noisy output from baseline methods while all 16 samples from CIMD have high fidelity.

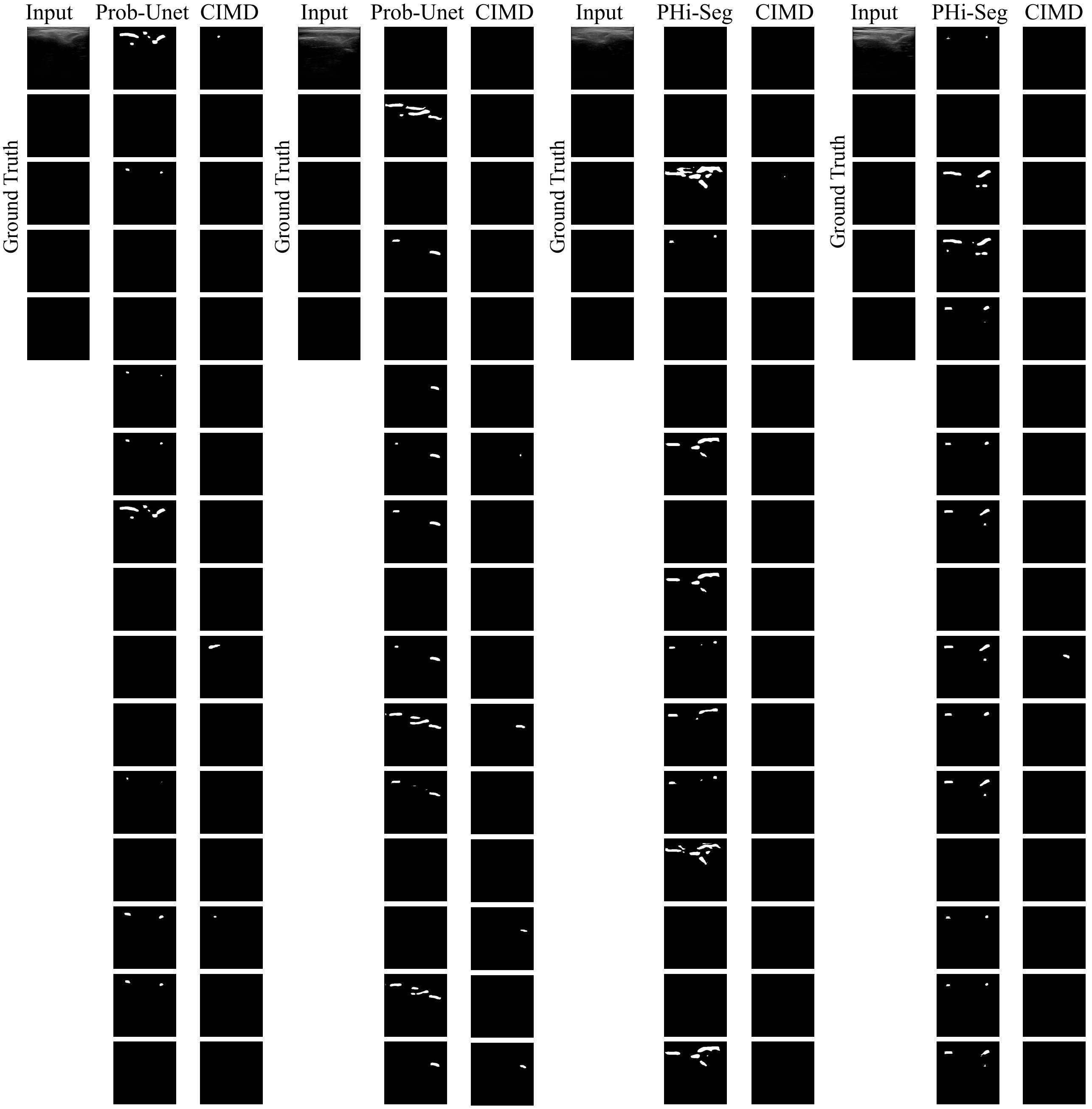

Figure 6. Comparative qualitative analysis with the two baseline methods Probabilistic U-net [5] and PHi-Seg [2] for blank segmentations from all experts in Bone-US dataset. We sample 16 masks from each model. We can observe that for blank annotation Prob-Unet and PHi-Seg both struggles as the noisy contrast resemble bone surface response in ultrasound images.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] Peter M Williams. Using neural networks to model conditional multivariate densities. *Neural computation*, 8(4):843–854, 1996.

[9] 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.