This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Appendix

A.1 We evaluate all methods on their ability to discriminate ambiguous from non-ambiguous cases using AUROC, AUPRC, and F1 score, with all test cases held out from segmentation training and ambiguity post-hoc training of the head. For LIDC, the test set comprises 88 (1vs3), 47 (2vs2), 47(3vs1) and 80 non-ambiguous cases; the optimal F1 threshold is selected on the validation set. For LNDb, we evaluate on a class-balanced subset obtained by down-sampling each class to the minority-class size. We report the mean over 5 differently initialized runs. All models were trained for 100 epochs with early stopping of 15 on best val dice score, using gt masks with a minimum consensus of 2 and 0.5 threshold for foreground prediction. Ambiguity head post-hoc training is done for 10 epochs. The per-voxel aleatoric uncertainty $\mathrm{AU}(v_i)$ is derived via stochastic sampling for MC-dropout models and Probabilistic 3D-UNet (S=100 samples, dropout=0.1), across N=3 members for the ensemble, and across K per-annotator predictions $p_k(v_i)$ obtained from learned annotator-specific confusion matrices for the Annotator-Confusion 3D-UNet, where $\mathrm{AU}(v_i) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{H}(p_k(v_i))$ and $\mathcal{H}(\cdot)$ denotes the entropy.

A.2 In Tables 1 and 2 below we provide detailed metrics underlying the reported metrics in Figures 2 and 3. Tables 3 and 4 provide results as in Figures 3 and 4 where $U_{\mathrm{AU}}(x)$ was aggregated over the prediction mask instead of the ground truth mask.

Table 1. Presence ambiguity detection performance on disagreement level.

		AU (Ensemble)			AU (MC-Dropout)			Ambiguity Head		
Split	**Model**	**AUROC**	**AUPRC**	**F1**	**AUROC**	**AUPRC**	**F1**	**AUROC**	**AUPRC**	**F1**
LIDC 1vs3	Attn-3DUnet	0.45±0.02	0.53±0.02	0.40±0.03	0.48±0.04	0.54±0.02	0.38±0.09	**0.86**±0.04	**0.87**±0.03	**0.78**±0.05
	3D-ResUnet	0.48±0.02	0.57±0.02	0.44±0.03	0.49±0.03	0.57±0.03	0.47±0.04	**0.90**±0.01	**0.89**±0.03	**0.82**±0.04
	SegFormer3D	0.43±0.04	0.58±0.04	0.32±0.09	0.42±0.12	0.53±0.08	0.23±0.17	**0.61**±0.04	**0.62**±0.03	**0.57**±0.04
	3D-Unet	0.39±0.02	0.45±0.01	0.32±0.04	0.42±0.07	0.47±0.03	0.31±0.17	**0.91**±0.02	**0.92**±0.01	**0.83**±0.02
LIDC 2vs2	Attn-3DUnet	0.52±0.01	0.42±0.01	0.38±0.09	0.52±0.03	0.40±0.01	0.41±0.09	**0.76**±0.04	**0.66**±0.04	**0.60**±0.06
	3D-ResUnet	0.50±0.01	0.45±0.02	0.39±0.04	0.50±0.02	0.43±0.02	0.38±0.03	**0.82**±0.03	**0.72**±0.04	**0.67**±0.06
	SegFormer3D	0.50±0.03	0.46±0.06	0.27±0.05	0.48±0.07	0.42±0.06	0.28±0.14	**0.63**±0.03	**0.51**±0.02	**0.53**±0.04
	3D-Unet	0.43±0.02	0.34±0.01	0.36±0.09	0.46±0.05	0.36±0.04	0.38±0.06	**0.79**±0.03	**0.70**±0.03	**0.66**±0.05
LIDC 3vs1	Attn-3DUnet	0.51±0.02	0.43±0.02	0.45±0.05	0.51±0.02	0.41±0.02	0.43±0.08	**0.66**±0.06	**0.57**±0.05	**0.52**±0.08
	3D-ResUnet	0.52±0.03	0.41±0.02	0.43±0.01	0.51±0.03	0.40±0.03	0.43±0.05	**0.65**±0.02	**0.54**±0.03	**0.53**±0.02
	SegFormer3D	0.46±0.04	0.42±0.05	0.20±0.05	0.47±0.07	0.40±0.05	0.26±0.22	**0.60**±0.02	**0.48**±0.03	**0.51**±0.04
	3D-Unet	0.50±0.01	0.39±0.01	0.45±0.08	0.51±0.04	0.39±0.02	0.48±0.08	**0.69**±0.02	**0.57**±0.01	**0.56**±0.03
LNDb 1vs2	Attn-3DUnet	0.53±0.03	0.59±0.06	0.46±0.07	0.53±0.04	0.58±0.06	0.49±0.07	**0.69**±0.02	**0.68**±0.04	**0.64**±0.10
	3D-ResUnet	0.48±0.06	0.53±0.05	0.37±0.14	0.52±0.05	0.55±0.06	0.51±0.10	**0.69**±0.04	**0.67**±0.04	**0.68**±0.04
	SegFormer3D	0.34±0.04	0.42±0.04	0.40±0.37	0.38±0.06	0.45±0.04	0.39±0.32	**0.67**±0.05	**0.67**±0.06	**0.64**±0.11
	3D-Unet	0.47±0.03	0.50±0.02	0.39±0.05	0.51±0.05	0.55±0.04	0.47±0.07	**0.71**±0.01	**0.70**±0.04	**0.68**±0.04
LNDb 2vs1	Attn-3DUnet	0.34±0.02	0.45±0.03	0.09±0.10	0.36±0.03	0.45±0.03	0.48±0.27	**0.67**±0.02	**0.65**±0.04	**0.69**±0.05
	3D-ResUnet	0.40±0.06	0.45±0.04	0.45±0.29	0.40±0.04	0.47±0.03	0.58±0.16	**0.62**±0.04	**0.60**±0.04	**0.62**±0.06
	SegFormer3D	0.37±0.04	0.44±0.02	0.48±0.27	0.42±0.05	0.49±0.03	0.29±0.23	**0.67**±0.01	**0.67**±0.03	**0.64**±0.05
	3D-Unet	0.38±0.01	0.47±0.01	0.31±0.20	0.41±0.03	0.48±0.03	0.25±0.14	**0.66**±0.01	**0.64**±0.03	**0.70**±0.04

Below we provide results as in Figures 3 and 4 aggregated on prediction mask instead of ground truth mask.

Table 2. Presence ambiguity detection performance compared with supervised competitors.

Split	Model	AUROC	AUPRC	F1
LIDC 1vs3	**Ambiguity Head (mean)**	**0.82**±0.14	**0.82**±0.14	**0.75**±0.12
	Ann-Conf-3DUnet	0.54±0.09	0.68±0.09	0.55±0.08
	Prob-3DUnet	0.49±0.03	0.56±0.01	0.48±0.10
LIDC 2vs2	**Ambiguity Head (mean)**	**0.75**±0.08	**0.65**±0.10	**0.61**±0.06
	Ann-Conf-3DUnet	0.63±0.12	0.63±0.14	0.51±0.13
	Prob-3DUnet	0.48±0.01	0.40±0.02	0.40±0.03
LIDC 3vs1	**Ambiguity Head (mean)**	**0.65**±0.04	**0.54**±0.04	**0.53**±0.02
	Ann-Conf-3DUnet	0.61±0.09	**0.54**±0.10	0.47±0.16
	Prob-3DUnet	0.55±0.04	0.43±0.02	0.47±0.06
LNDb 1vs2	**Ambiguity Head (mean)**	**0.69**±0.01	**0.68**±0.01	**0.66**±0.02
	Ann-Conf-3DUnet	0.50±0.08	0.57±0.05	0.35±0.18
	Prob-3DUnet	0.55±0.06	0.60±0.07	0.46±0.04
LNDb 2vs1	**Ambiguity Head (mean)**	**0.66**±0.02	**0.64**±0.03	**0.66**±0.04
	Ann-Conf-3DUnet	0.52±0.04	0.57±0.02	0.38±0.14
	Prob-3DUnet	0.44±0.07	0.53±0.07	0.39±0.26

Table 3. Presence ambiguity detection performance across architectures and disagreement thresholds, on LIDC and the external LNDb cohort for aggregation of $U_{\mathrm{AU}}(x)$ on the prediction mask instead of ground truth mask.

	LIDC			LNDb		
Method	AUROC	AUPRC	F1	AUROC	AUPRC	F1
Ambiguity Head	**0.74**	**0.67**	**0.63**	**0.67**	**0.66**	**0.66**
Ann-Conf-3DUnet	0.73	0.60	0.62	0.65	0.60	0.66
Prob-3DUnet	0.64	0.56	0.60	0.62	0.61	0.55
MC-Dropout	0.58	0.53	0.49	0.54	0.55	0.56
Ensemble	0.54	0.51	0.42	0.49	0.52	0.47

Table 4. Ordinal correlation (Spearman ρ) between uncertainty scores and disagreement severity for aggregation of $U_{\mathrm{AU}}(x)$ on the prediction mask instead of ground truth mask.

Method	Backbone	LIDC ρ	LNDb ρ
AU Aggregation	Ann. Confusion	0.43 ± 0.08	0.08 ± 0.03
	3D-ResUnet	0.32 ± 0.02	0.16 ± 0.01
	Prob-UNet	0.29 ± 0.03	0.18 ± 0.01
	Attn-3DUnet	0.21 ± 0.02	0.18 ± 0.01
	3D-Unet	0.15 ± 0.04	0.16 ± 0.01
	SegFormer3D	-0.01 ± 0.03	0.02 ± 0.02
Ambiguity Head	3D-Unet	$\mathbf{0.63 \pm 0.01}$	$\mathbf{0.22 \pm 0.01}$
	3D-ResUnet	0.60 ± 0.00	0.20 ± 0.01
	Attn-3DUnet	0.51 ± 0.02	0.13 ± 0.02
	SegFormer3D	0.13 ± 0.02	0.02 ± 0.01