Supplementary Material

S1. NoTeacher Method: Additional Details/Theory

S1.1. Derivation of NoT weights

In this appendix we derive the marginal density distribution of the network outputs when the consensus function is integrated out. Subsequently, the loss multipliers $\{\lambda^{\mathrm{L}}_{y,1}, \lambda^{\mathrm{L}}_{y,2}, \lambda^{\mathrm{L}}_{1,2}, \lambda^{\mathrm{U}}_{1,2}\}$ will be derived as functions of the hyperparameters $\{\sigma_1^2, \sigma_2^2, \sigma_y^2\}$. Given the NoT graphical model, it is necessary to integrate f_c out of the joint density distribution of the graph, because f_c is a latent variable. Consider a more general model, where there are M networks, each outputs a posterior, i.e., $\{f_m\}_{m=1}^M$. Graphically, each posterior is represented by a random variable connected only to the consensus function f_c via a zero-mean Gaussian potential

$$f_m - f_c \sim \mathcal{N}\left(0, \sigma_m^2\right). \tag{8}$$

Technically, even the target y can be considered as a posterior with potential $y - f_c \sim \mathcal{N}\left(0, \sigma_y^2\right)$. The joint density distribution function of the general graph is as follows

$$p\left(f_c, f_1, \ldots, f_M\right) = \frac{1}{\mathcal{Z}_1} \prod_{m=1}^{M} \exp\left[-\frac{\left(f_c - f_m\right)^2}{2\sigma_m^2}\right] \tag{9}$$

$$= \frac{1}{\mathcal{Z}_1} \exp\left(-\frac{\psi}{2} f_c^2 + \phi f_c + \chi\right), \tag{10}$$

where the normalizing factor $\mathcal{Z}_1$ is a constant w.r.t. $f_c, f_1, \ldots, f_M$ and

$$\psi = \sum_{m=1}^{M} \frac{1}{\sigma_m^2} \qquad \phi = \sum_{m=1}^{M} \frac{f_m}{\sigma_m^2} \qquad \chi = \sum_{m=1}^{M} -\frac{f_m^2}{2\sigma_m^2}. \tag{11}$$

Notice that ψ, ϕ, χ are constants w.r.t. f_c. In addition, we have the following integration rule

$$\int \exp\left(-\frac{1}{2} a x^2 + b x\right) dx = \sqrt{\frac{2\pi}{a}} \exp\left(\frac{b^2}{2a}\right), \tag{12}$$

where a is positive. With this rule, knowing that $\psi > 0$, we can integrate f_c out of the joint distribution in (10) to obtain the following marginal likelihood as follows

$$p\left(f_1, \ldots, f_M\right) = \int p\left(f_c, f_1, \ldots, f_M\right) df_c \tag{13}$$

$$= \frac{1}{\mathcal{Z}_2} \exp\left(\frac{\phi^2}{2\psi} + \chi\right) = \frac{1}{\mathcal{Z}_2} \exp\left[\frac{1}{2\psi}\left(\phi^2 + 2\psi\chi\right)\right] \tag{14}$$

$$= \frac{1}{\mathcal{Z}_2} \exp\left[\frac{1}{2\psi}\left(\sum_m \frac{f_m^2}{\sigma_m^4} + 2\sum_m \sum_{k>m} \frac{f_m f_k}{\sigma_m^2 \sigma_k^2} - \psi \sum_m \frac{f_m^2}{\sigma_m^2}\right)\right] \tag{15}$$

$$= \frac{1}{\mathcal{Z}_2} \exp\left[\frac{1}{2\psi}\left(\sum_m \sum_{k>m} -\frac{f_m^2 - 2 f_m f_k + f_k^2}{\sigma_m^2 \sigma_k^2}\right)\right] \tag{16}$$

$$= \frac{1}{\mathcal{Z}_2} \exp\left[\sum_m \sum_{k>m} -\lambda_{m,k}\left(f_m - f_k\right)^2\right], \tag{17}$$

where $\mathcal{Z}_2$ is another constant w.r.t. $f_c, f_1, \ldots, f_M$, and

$$\lambda_{m,k} = \left[2\sigma_m^2 \sigma_k^2 \left(\sum_{i=1}^{M} \frac{1}{\sigma_i^2}\right)\right]^{-1}. \tag{18}$$

This result implies that the marginal likelihood can be factorized as a product of $\binom{M}{2}$ components, each component is a Gaussian distribution on the difference between a pair of posteriors (f_m, f_k) with zero mean and variance $\lambda_{m,k}^{-1}$.

The NoT graphical models are special cases of this general model. For a labeled sample, we have $M = 3$, i.e., there are three observed variables f_1, f_2 and y. Thus, we use $\{\lambda^{\mathrm{L}}_{y,1}, \lambda^{\mathrm{L}}_{y,2}, \lambda^{\mathrm{L}}_{1,2}\}$ as shorthand notations to denote $\{\lambda_{f_1,y}, \lambda_{f_2,y}, \lambda_{f_1,f_2}\}$ respectively. By applying (18), they can be expressed in terms of the hyperparameters $\{\sigma_1^2, \sigma_2^2, \sigma_y^2\}$ as

$$\lambda^{\mathrm{L}}_{y,1} = \lambda_{f_1,y} = \frac{\sigma_2^2}{2\left(\sigma_1^2\sigma_2^2 + \sigma_2^2\sigma_y^2 + \sigma_1^2\sigma_y^2\right)} \tag{19}$$

$$\lambda^{\mathrm{L}}_{y,2} = \lambda_{f_2,y} = \frac{\sigma_1^2}{2\left(\sigma_1^2\sigma_2^2 + \sigma_2^2\sigma_y^2 + \sigma_1^2\sigma_y^2\right)} \tag{20}$$

$$\lambda^{\mathrm{L}}_{1,2} = \lambda_{f_1,f_2} = \frac{\sigma_y^2}{2\left(\sigma_1^2\sigma_2^2 + \sigma_2^2\sigma_y^2 + \sigma_1^2\sigma_y^2\right)}. \tag{21}$$

Similarly, for an unlabeled sample, there are $M = 2$ observed variables f_1 and f_2, the formula of $\lambda^{\mathrm{U}}_{1,2}$ is therefore

$$\lambda^{\mathrm{U}}_{1,2} = \lambda_{f_1,f_2} = \frac{1}{2\left(\sigma_1^2 + \sigma_2^2\right)}. \tag{22}$$

S1.2. Derivation of NoT-GA loss function

The NoT-GA graphical model can be divided into two cases: (i) a labeled sample when $z = 1$ and (ii) an unlabeled sample when $z = 0$. Thus, we can compute the likelihood separately for labeled and unlabeled data. The joint distribution of a sample is

$$p(z, y, f_c, f_1, f_2) = \begin{cases} \gamma_y p\left(f_c, y, f_1, f_2\right), & \text{if } z = 1 \\ \left(1 - \gamma_y\right) p\left(f_c, y, f_1, f_2\right), & \text{if } z = 0 \end{cases}, \tag{23}$$

where we use γ_y to denote the γ_k value corresponding to the target label, i.e., $y = k$. By integrating out f_c – the latent consensus function, we obtain the data likelihood as

$$p(z, y, f_1, f_2) \propto \begin{cases} \gamma_y \int p\left(f_c, y, f_1, f_2\right) df_c, & \text{if } z = 1 \\ \iint \left(1 - \gamma_y\right) p\left(f_c, y, f_1, f_2\right) df_c dy, & \text{if } z = 0 \end{cases}. \tag{24}$$

Note that in (24), the ground-truth label is unobserved for the case $z = 0$, thus an additional integration over y is required. By taking the integration over f_c as performed in Subsection S1.1, the log likelihood can be computed. For a labeled sample, it is

$$\log\left[p\left(z = 1, y, f_1, f_2\right)\right] \propto -\lambda^{\mathrm{L}}_{y,1}\|f_1 - y\|^2 - \lambda^{\mathrm{L}}_{y,2}\|f_2 - y\|^2 - \lambda^{\mathrm{L}}_{1,2}\|f_1 - f_2\|^2 + \log(\gamma_y). \tag{25}$$

For an unlabeled sample, it is

$$\log\left[p\left(z = 0, y, f_1, f_2\right)\right] \propto -\lambda^{\mathrm{L}}_{1,2}\|f_1 - f_2\|^2 + \log\left[\sum_y \exp\left(-\lambda^{\mathrm{L}}_{y,1}\|f_1 - y\|^2 - \lambda^{\mathrm{L}}_{y,2}\|f_2 - y\|^2\right)\left(1 - \gamma_y\right)\right]. \tag{26}$$

Notice that in equation (25), the last term is a constant w.r.t. the observed variables and can be removed from the optimization. By combining the negative log likelihood functions over all labeled and unlabeled samples in the dataset, we obtain the NoT-GA loss function in (5).

S1.3. Connections to Other Methods

The training process of MT and NoT are compared in Figure S1, with the networks in NoT no longer being connected by the EMA update. Mean Teacher (MT) sets up two neural networks with identical architecture: a student model F_S and a teacher model F_T. Given a batch $\mathbf{x}$ of training data, MT employs random augmentations η_S, η_T to generate augmented inputs $\mathbf{x}_S$ and $\mathbf{x}_T$ for the student and teacher models correspondingly. During the feed-forward pass, MT computes a weighted sum of a supervised classification loss and a consistency loss

$$\mathfrak{L}_{\mathrm{MT}} = \mathrm{CE}\left(\mathbf{y}, \mathbf{f}^{\mathrm{L}}_S\right) + \lambda_{\mathrm{cons}}\mathrm{MSE}\left(\mathbf{f}_S, \mathbf{f}_T\right), \tag{27}$$

where $\mathbf{f}_S, \mathbf{f}_T$ are posterior outputs from the student and teacher networks, $\mathbf{f}^{\mathrm{L}}_S$ is the student's posterior output on the labeled data, and λ_{cons} is a consistency weight hyperparameter. The classification loss is usually cross-entropy (CE), while the consistency loss is typically mean-squared error on the posteriors (MSE). The student model backpropagates directly using gradients from the loss $\mathfrak{L}_{\mathrm{MT}}$. In the meantime, the teacher model is updated via computing an exponential moving average (EMA) over the parameters of the student network. Recent papers have adapted MT for medical imaging tasks such as MR segmentation (Yu et al., 2019; Perone and Cohen-Adad, 2018) and nuclei classification (Su et al., 2019). We note that $\mathbf{f}_S, \mathbf{f}_T$ are similar to the views $\mathbf{f}_1, \mathbf{f}_2$ of NoT, except that MT uses the EMA update to compute $\mathbf{f}_T$, while NoT uses backpropagation to update both views.

Figure S2 illustrates the iterative process of co-training, a multi-view learning technique.

S2. Additional Experiment Setup Details

We summarize the data split statistics for the NIH-14 Chest X-Ray, RSNA Brain CT and Knee MRNet datasets in Table S1. We also provide the hyperparameters used for the different labeling budgets, methods and datasets in our experiments in Table S2. Finally, we provide the adapted pipeline for the 3D MRNet classification task in Figure S3.

S3. Additional Results: Comparisons to Baselines

For ease of assessment on quantitative results, we provide the detailed breakdown of AUROC scores for each dataset, labeling budget and method in Table S3. We

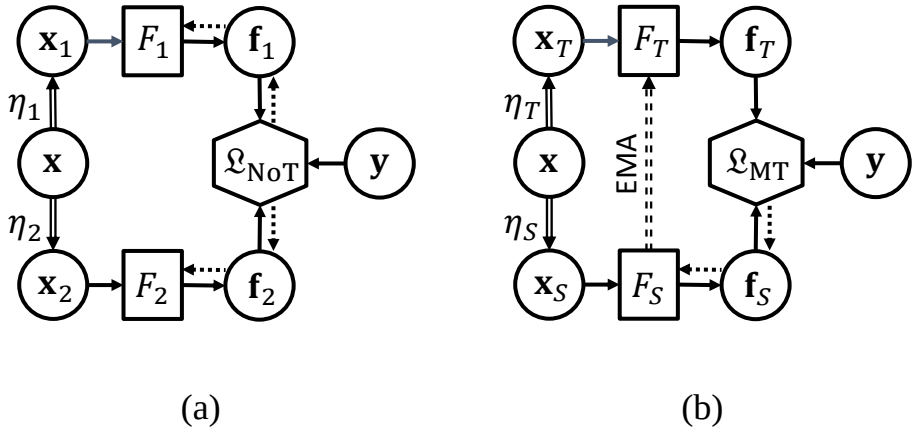


Figure S1: **Training process of (a) MT and (b) NoT** on a batch of semi-supervised data. Solid and dotted arrows denote forward and backward passes, respectively. Double-line arrows denote random data augmentations, while double dashed arrow represents EMA update.

Table S1: Data statistics for the NIH-14, RSNA-CT and MRNet datasets.

Dataset	$L_T : L_V$	**Split**	**Patients**	**Scans**	**Images**
NIH-14	70 : 10	Train	21528	78468	78468
		Val.	3090	11219	11219
		Test	6187	22433	22433
RSNA-CT	60 : 20	Train	10247	10247	352839
		Val.	3416	3416	117986
		Test	3416	3416	117907
MRNet	64 : 16	Train	904	904	31156
		Val.	226	226	7622
		Test	120	120	4118

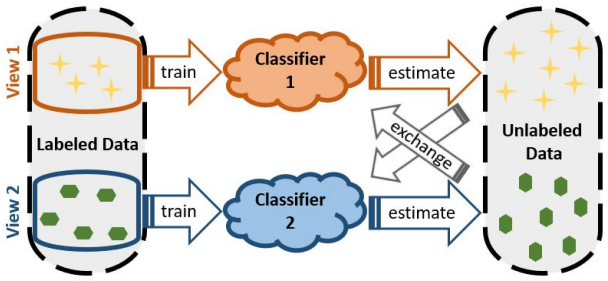


Figure S2: The iterative process of co-training in a two-view setup. This process continues until two views achieve a high level of agreement on unlabeled data.

also highlight comparisons to previously published state-of-the-art results where available. We note that, for the NIH-14 Chest X-Ray dataset, our fully-supervised baseline outperforms the numbers reported in the SRC-MT paper (Liu et al., 2020). We suspect that this is because we report the best results from either the trained model or its EMA copy. Also we highlight that the results of GraphX$^{\text{NET}}$ (Aviles-Rivero et al., 2019) and SRC-MT (Liu et al., 2020) are not directly comparable as they report metrics on a subset of the classes we considered and/or employ different backbones.

S4. Additional Results: NoT-GA Experiments

We have designed two other variants of the "DM-3311" setups, namely "DM-1133" and "DM-1313". First, the

Table S2: Hyperparameter tuning based on average validation AUROC. For VAT, multi-task KL divergence (KL_{mt}) offers > 10% AUROC boost. Other implementation details of NoT match with MT.

		NIH-14			RSNA-CT			MRNet		
		Train	**Val.**	**Test**	**Train**	**Val.**	**Test**	**Train**	**Val.**	**Test**
Labeling budget (%)		**1-5**	**10-20**	**25-100**	**0.25-0.5**	**1-2.5**	**5-100**	**1-10**	**15-25**	**30-100**
MT	$\alpha \in \{0.91, 0.93, \ldots, 0.99\}$	0.91	0.95	0.99	0.93			0.97	0.97	0.99
	$\lambda_{cons} \in \{1, 2, \ldots, 196\}$	196			100			10		
VAT	$\epsilon \in \{1, 2, \ldots, 6\}$	2						3		
	LDS $\in$ {KL, MSE, KL_{mt}}	KL_{mt}								
NoT	$\sigma_1^2 = \sigma_2^2 = 2^{-2}$ $\sigma_y^2 \in \{2^{-2}, 2^{-3} \ldots, 2^{-7}\}$	2^{-2}						2^{-7}	2^{-6}	
Early stopping		15	7	3	15	7	3	11	8	5
Reduce learning rate patience		5	3	1	5	3	1	5	4	3
Min no. of validation samples		113			268			25		

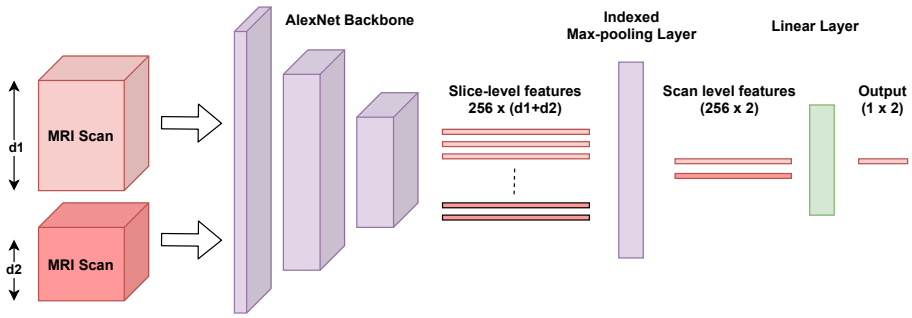


Figure S3: Adapted MR-Net Training Pipeline. As the MRI scans have variable slice depths, the original MR-Net can process only one scan at a time. We extend this architecture using indexed max-pooling, where the maximization is performed on the depth dimension of the scan. This allows us to process a batch of scans at a time.

"DM-1133" has a class imbalance ratio of $\boldsymbol{\alpha}^{\mathrm{L}} \propto [3, 3, 1, 1]$ on labeled data and a class imbalance ratio of $\boldsymbol{\alpha}^{\mathrm{U}} \propto [1, 1, 3, 3]$ on unlabeled data, which is the reverse setup of "DM-3311". In contrast to the earlier setups, No Finding and Infiltration are now the classes with higher γ values. Second, the "DM-1313" has a class imbalance ratio of $\boldsymbol{\alpha}^{\mathrm{L}} \propto [3, 1, 3, 1]$ on labeled data and a class imbalance ratio of $\boldsymbol{\alpha}^{\mathrm{U}} \propto [1, 3, 1, 3]$ on unlabeled data, which intentionally mixes up the (naturally) rare and common classes together. By comparing NoT against NoT-GA on these new setups, we further strengthen our understanding of the NoT-GA behaviors. The results are reported in Table S7, Table S8, Figure S4 and Figure S5.

S5. Additional Results: MT vs. NoT

We provide additional plots from various seeds for the disagreement count comparison between MT and NoT. Figure S6 shows results on NIH-14 and Figure S7 shows results on RSNA Brain CT.

Table S3: Average AUROC scores (multiplied by 100) vs. labeling budget (%) for NIH-14 (top) and RSNA Brain CT (center) and Knee MRNet (bottom). At the best case budgets (bolded) of 5% in NIH-14, 1% in RSNA-CT and 25% in MRNet, NoT has 2.5%, 2.2% and 1.2% higher AUPRC than other SSL methods respectively. We include the average AUROC scores reported for SRC-MT (Liu et al., 2020) and GraphX^{NET} (Aviles-Rivero et al., 2019).

NIH-14							
Budget (Images)	**SUP**	**PSU**	**VAT**	**MT**	**NoT**	**GraphX^{NET}**	**SRC-MT**
1.3 (1177)	66.53 ± 0.38	66.48 ± 0.43	69.56 ± 0.17	68.97 ± 0.63	**70.69 ± 0.15**	—	—
2 (1569)	66.85 ± 1.45	67.64 ± 0.70	69.63 ± 0.23	70.42 ± 0.58	**72.60 ± 0.18**	53.00	66.95
5 (3923)	70.68 ± 1.31	70.93 ± 0.99	73.94 ± 0.14	73.60 ± 0.70	**77.04 ± 0.22**	58.00	72.29
10 (7846)	75.69 ± 1.17	76.00 ± 0.67	77.15 ± 1.06	76.98 ± 0.02	**77.61 ± 0.54**	63.00	75.28
20 (15693)	77.19 ± 0.75	78.06 ± 0.47	79.38 ± 0.10	78.66 ± 0.64	**79.49 ± 0.89**	78.00	79.23
50 (39234)	80.57 ± 0.68	81.46 ± 0.34	82.01 ± 0.14	81.78 ± 0.11	**82.10 ± 0.05**	—	—
100 (78468)	**83.33 ± 0.38**	—	—	—	—	—	81.75
RSNA-CT							
Budget (Slices)	**SUP**	**PSU**	**VAT**	**MT**	**NoT**	**GraphX^{NET}**	**SRC-MT**
0.25 (749)	70.77 ± 3.52	72.19 ± 1.81	71.07 ± 2.72	74.86 ± 2.23	**76.53 ± 1.84**	—	—
0.5 (1777)	80.55 ± 0.97	80.89 ± 1.43	82.50 ± 0.60	82.15 ± 1.43	**83.57 ± 0.98**	—	—
1 (3495)	80.01 ± 1.19	81.53 ± 0.32	81.41 ± 2.17	80.90 ± 1.91	**84.50 ± 1.86**	—	—
2.5 (6744)	86.13 ± 1.10	87.01 ± 0.15	87.22 ± 0.40	86.26 ± 0.58	**89.81 ± 0.23**	—	—
5 (17242)	91.31 ± 0.29	**91.84 ± 0.37**	90.53 ± 0.38	91.24 ± 0.18	91.79 ± 0.73	—	—
10 (33560)	92.74 ± 0.38	93.27 ± 0.03	92.84 ± 0.44	92.62 ± 0.64	**93.31 ± 0.80**	—	—
100 (352839)	**96.69 ± 0.11**	—	—	—	—	—	—
MRNet							
Budget (Scans)	**SUP**	**PSU**	**VAT**	**MT**	**NoT**	**GraphX^{NET}**	**SRC-MT**
5 (32)	66.26 ± 3.13	67.31 ± 3.72	71.37 ± 2.42	68.36 ± 2.70	**73.25 ± 2.78**	—	—
7.5 (55)	68.51 ± 6.09	73.81 ± 3.57	75.38 ± 5.23	73.33 ± 2.27	**77.54 ± 2.11**	—	—
15 (137)	83.72 ± 2.08	85.63 ± 1.27	85.75 ± 2.85	87.12 ± 0.44	**88.35 ± 1.38**	—	—
25 (227)	85.07 ± 5.50	85.94 ± 5.39	85.12 ± 3.12	85.15 ± 3.86	**89.19 ± 0.89**	—	—
50 (452)	89.81 ± 1.08	90.47 ± 1.31	90.23 ± 0.47	90.89 ± 1.72	**91.49 ± 0.90**	—	—
100 (904)	**91.03 ± 0.33**	—	—	—	—	—	—

* For GraphX^{NET} (Aviles-Rivero et al., 2019), the average AUROC scores were reported from 8 labels, namely Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia and Pneumothorax, whereas the rest of the methods were averaged over all of the 14 labels.

* While the rest of methods use a DenseNet-121 backbone, the SRC-MT reports its results using a DenseNet-169 backbone and the GraphX^{NET} utilizes graph based representation.

Table S4: Validation AUC while training on 5% annotation budget on the Chest-XRay14 dataset for the selection of best loss function

	LDS divergence function		
eps	MSE	KL Multiclass	KL
2	0.7425	**0.7503**	0.6532
4	0.7436	0.7499	0.5297
6	0.7269	0.7450	0.6510
8	0.7484	0.7472	0.5346
10	0.7393	0.7457	0.5459

Table S5: **Class Distribution Mismatch**: Average Per-Class AUPRC scores (multiplied by 100)

Methods	**No Finding**	**Infiltration**	**Pneumothorax**	**Mass**	**Average**
SUP	61.18 ± 0.66	44.58 ± 1.48	33.44 ± 2.37	19.99 ± 2.77	39.80 ± 1.82
MT	60.93 ± 1.53	45.31 ± 1.18	35.58 ± 1.41	22.91 ± 3.92	41.18 ± 2.01
VAT	61.58 ± 0.88	45.10 ± 1.02	34.75 ± 0.87	20.74 ± 2.54	40.54 ± 1.33
NoT	62.30 ± 0.83	48.27 ± 0.61	41.71 ± 1.25	34.81 ± 3.17	46.77 ± 1.47
NoT-GA	**62.79 ± 0.71**	**48.83 ± 0.52**	**42.33 ± 1.13**	**38.66 ± 2.75**	**48.15 ± 1.28**
Fully Supervised	65.51 ± 1.22	55.05 ± 0.88	46.30 ± 0.70	41.44 ± 1.05	52.08 ± 0.96

Table S6: Results on DM-3311 Setup

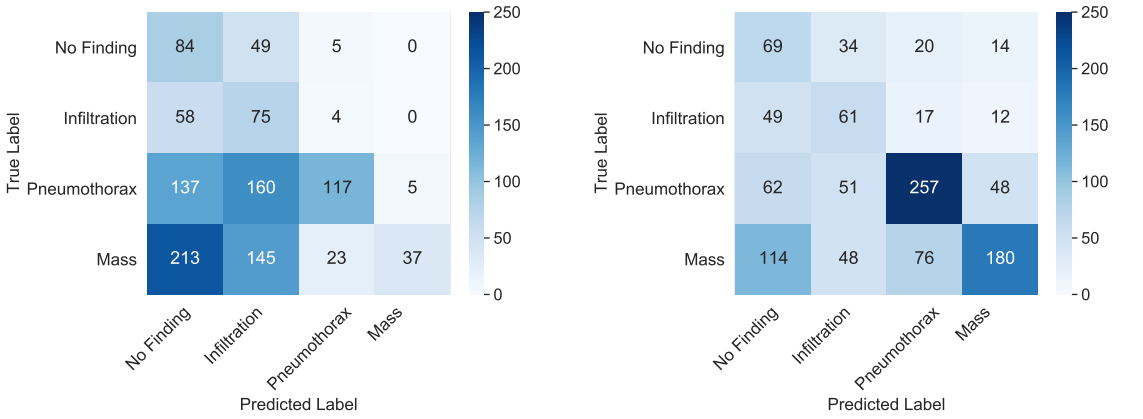
Metric	**Method**	**No Finding**	**Infiltration**	**Pneumothorax**	**Mass**	**Average**
AUPRC	NoT	51.25 ± 0.68	53.39 ± 0.68	50.04 ± 1.26	46.7 ± 1.89	50.34 ± 1.13
	NoT-GA	51.07 ± 0.45	53.36 ± 0.77	53.28 ± 0.74	60.58 ± 2.75	54.57 ± 1.18
Precision (threshold 0.5)	NoT	0.585 ± 0.033	0.638 ± 0.081	0.283 ± 0.021	0.216 ± 0.028	–
	NoT-GA	0.517 ± 0.01	0.548 ± 0.002	0.431 ± 0.028	0.527 ± 0.053	–
Recall (threshold 0.5)	NoT	0.16 ± 0.076	0.115 ± 0.082	0.803 ± 0.021	0.792 ± 0.034	–
	NoT-GA	0.497 ± 0.03	0.419 ± 0.05	0.705 ± 0.018	0.619 ± 0.025	–

Table S7: Results on DM-1133 Setup

Metric	Method	No Finding	Infiltration	Pneumothorax	Mass	Average
AUPRC	NoT	28.31 ± 0.83	30.72 ± 1.63	70.75 ± 1.15	64.48 ± 2.18	48.57 ± 1.45
	NoT-GA	26.67 ± 0.45	30.68 ± 1.12	72.66 ± 0.81	67.02 ± 1.45	49.26 ± 0.96
Precision (threshold 0.5)	NoT	0.174 ± 0.018	0.177 ± 0.027	0.788 ± 0.034	0.886 ± 0.069	-
	NoT-GA	0.239 ± 0.024	0.315 ± 0.025	0.693 ± 0.022	0.714 ± 0.051	-
Recall (threshold 0.5)	NoT	0.606 ± 0.054	0.541 ± 0.054	0.279 ± 0.123	0.090 ± 0.058	-
	NoT-GA	0.497 ± 0.048	0.4358 ± 0.047	0.614 ± 0.019	0.430 ± 0.087	-

Table S8: Results on DM-1313 Setup

Metric	Method	No Finding	Infiltration	Pneumothorax	Mass	Average
AUPRC	NoT	23.03 ± 1.03	55.7 ± 1.68	47.51 ± 3.02	57.99 ± 6.84	46.06 ± 3.15
	NoT-GA	25.97 ± 0.62	58.57 ± 1.42	55.59 ± 0.68	63.76 ± 1.93	50.97 ± 1.16
Precision (threshold 0.5)	NoT	0.161 ± 0.008	0.552 ± 0.311	0.259 ± 0.015	0.641 ± 0.360	-
	NoT-GA	0.203 ± 0.022	0.618 ± 0.033	0.396 ± 0.028	0.640 ± 0.018	-
Recall (threshold 0.5)	NoT	0.700 ± 0.024	0.035 ± 0.021	0.823 ± 0.052	0.082 ± 0.053	-
	NoT-GA	0.549 ± 0.062	0.303 ± 0.047	0.781 ± 0.031	0.385 ± 0.051	-



(a) NoTeacher (NoT) (b) NoTeacher with Generative Assumption (NoT-GA)

Figure S4: **Results on DM-1133 Setup**: Average confusion matrix (threshold 0.5) over 5 seeds with vanilla NoT (a) and with NoT-GA (b).

(a) NoTeacher (NoT) (b) NoTeacher with Generative Assumption (NoT-GA)

Figure S5: **Results on DM-1313 Setup**: Average confusion matrix (threshold 0.5) over 5 seeds with vanilla NoT (a) and with NoT-GA (b).

Figure S6: Performance as a function of training as indicated by AUROC scores on the validation set (left vertical axis) and disagreement statistics (right vertical axis). Results are from NIH-14 Chest X-Ray with 5% labeling budget, $X = 40$, $\tau = 0.25$ with various seeds.

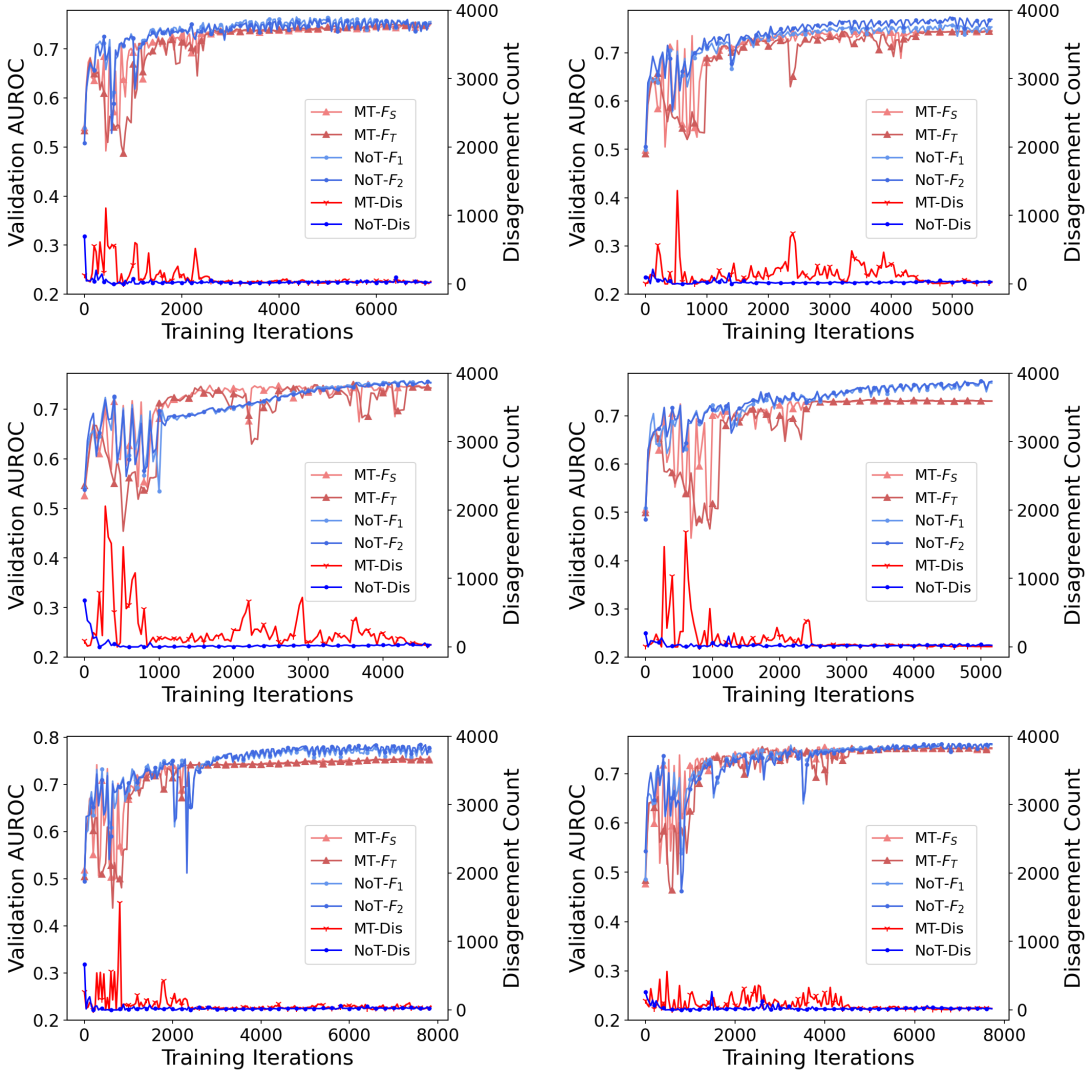


Figure S7: Performance as a function of training as indicated by AUROC scores on the validation set (left vertical axis) and disagreement statistics (right vertical axis). Results are from RSNA Brain CT with 2.5% labeling budget, $X = 40$, $\tau = 0.2$ with various seeds.