MULTIMIX: SPARINGLY SUPERVISED, EXTREME MULTITASK LEARNING FROM MEDICAL IMAGES
(SUPPLEMENTARY MATERIAL)

Ayaan Haque[1⋆], *Abdullah-Al-Zubaer Imran*[2,3⋆], *Adam Wang*[2], *Demetri Terzopoulos*[3,4]

[1]Saratoga High School, Saratoga, CA, USA
[2]Stanford University, Stanford, CA, USA
[3]University of California, Los Angeles, CA, USA
[4]VoxelCloud, Inc., Los Angeles, CA, USA

ABSTRACT

This supplemental document presents additional algorithmic details, architectural details, performance results, and visualizations of the outputs of the MultiMix model.

Index Terms— Classification, Segmentation, Multitasking, Semi-Supervised Learning, Data Augmentation, Saliency Bridge, Chest X-Ray, Lungs, Pneumonia

1. INTRODUCTION

Algorithm 1 presents the details of the MultiMix training procedure.

Fig. 1 visualizes the ground truth lung masks and the MultiMix model (MultiMix-50-1000) predicted masks for a number of images from the JSRT dataset (in-domain).

Fig. 2 visualizes the ground truth lung masks and the MultiMix model (MultiMix-50-1000) predicted masks for a number of images from the MCU dataset (cross-domain).

Architectural details of the MultiMix model are presented in Table 1 for the Encoder and Table 2 for the Decoder networks.

Evaluations of the performance of the MultiMix model against the baselines were performed with varying quantities of class (CheX) and segmentation (JSRT) labeled data. Accuracy and class-wise F1 scores were used to compare the classification performances, while the segmentation performances were compared using Dice Similarity (DS), Jaccard Similarity (JS), Structural Similarity (SSIM), average Hausdorff Distance (HD), Precision (P), and Recall (R). Tables 3 and 4 report the results on the in-domain (CheX, JSRT) and cross-domain (NIHX, MCU) evaluations, respectively.

Good agreement is observed between the ground truth lung masks and the MultiMix predicted segmentation masks, as is confirmed by the Bland-Altman plots—Figs. 3 and 4 for the MultiMix model with varying quantities of labeled data in in-domain and cross-domain evaluations, respectively.

Fig. 7 demonstrates the superiority and better consistency of the MultiMix models over the baselines in classifying normal and abnormal (pneumonia) X-rays in either domain.

Fig. 8 further showcases the superior classification performance of our MultiMix model over the baseline single-task or multitask models.

[⋆] Authors contributed equally

Algorithm 1 MultiMix Mini-Batch Training

```
Require:
  Training set of labeled data $x_l^c, y_l^c \in \mathcal{D}_l^c$
  Training set of labeled data $x_l^s, y_l^s \in \mathcal{D}_l^s$
  Training set of unlabeled data $x_{u_{weak}}^c \in \mathcal{D}_u^c$
  Training set of unlabeled data $x_{u_{strong}}^c \in \mathcal{D}_u^c$, where $x_{u_{weak}}^c$ and $x_{u_{strong}}^c$ augmented at different strengths
  Training set of unlabeled inputs $x_u^s \in \mathcal{D}_u^s$
  Network architecture $\mathcal{F}_\theta$ with learnable parameters $\theta$
  for each step do
    Sample minibatch $x_l^c(i); x_l^c(1), \ldots, x_l^c(m) \sim p_{\mathcal{D}^c(x)}$
    Sample minibatch $x_l^s(i); x_l^s(1), \ldots, x_l^s(m) \sim p_{\mathcal{D}^s(x)}$
    Sample minibatch $x_{u_{weak}}^c(i); x_{u_{weak}}^c(1), \ldots, x_{u_{weak}}^c(m) \sim p_{\mathcal{D}_u^c(x)}$ $x_{u_{strong}}^c(i); x_{u_{strong}}^c(1), \ldots, x_{u_{strong}}^c(m) \sim p_{\mathcal{D}_u^c(x)}$
    Sample minibatch $x_u^s(i); x_u^s(1), \ldots, x_u^s(m) \sim p_{\mathcal{D}^s(x)}$
    Compute model outputs for the labeled data: $\hat{c}_l, \hat{y} \leftarrow \mathcal{F}_\theta$
    Compute model outputs for the unlabeled data: $\hat{c}_{u_{weak}}, \hat{c}_{u_{strong}}, \hat{y}_u \leftarrow \mathcal{F}_\theta$
    Compute psuedo-label for weakly augmented classification predictions: $1\max(\hat{c}_{u_{weak}}) > t \leftarrow y_a$
    Update $\mathcal{F}_\theta$ along its gradient:
    $\nabla_{\theta_\mathcal{F}} \frac{1}{|\mathcal{M}_l|} \sum_{i \in \mathcal{M}_\mathcal{L}} \left[ L_{(\hat{c}_l, \hat{y}_l, y_c l, y_s l)} \right] + \alpha \frac{1}{|\mathcal{M}_u|} \sum_{i \in \mathcal{M}_u} \left[ L_{\left(\hat{c}_{u_{strong}}, y_l, \hat{y}_u, \hat{y}_l\right)} \right]$
  end for
```

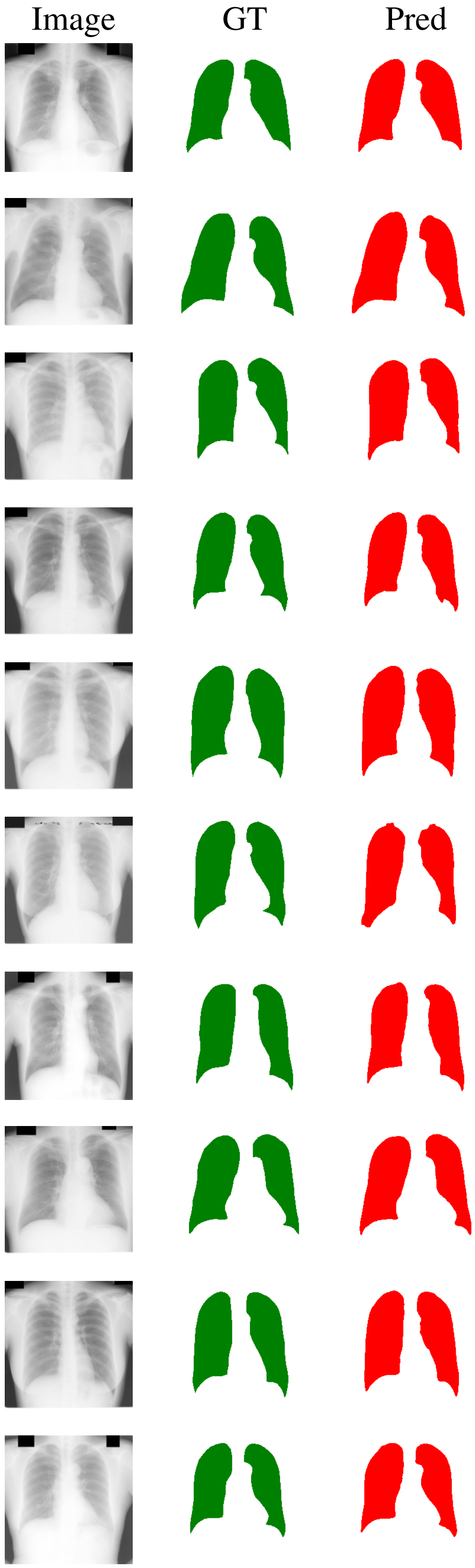

Fig. 1. Visualizations of the segmented lung masks by MultiMix-50-1000 on the in-domain JSRT dataset. The results show good agreement between the groundtruth and predicted masks.

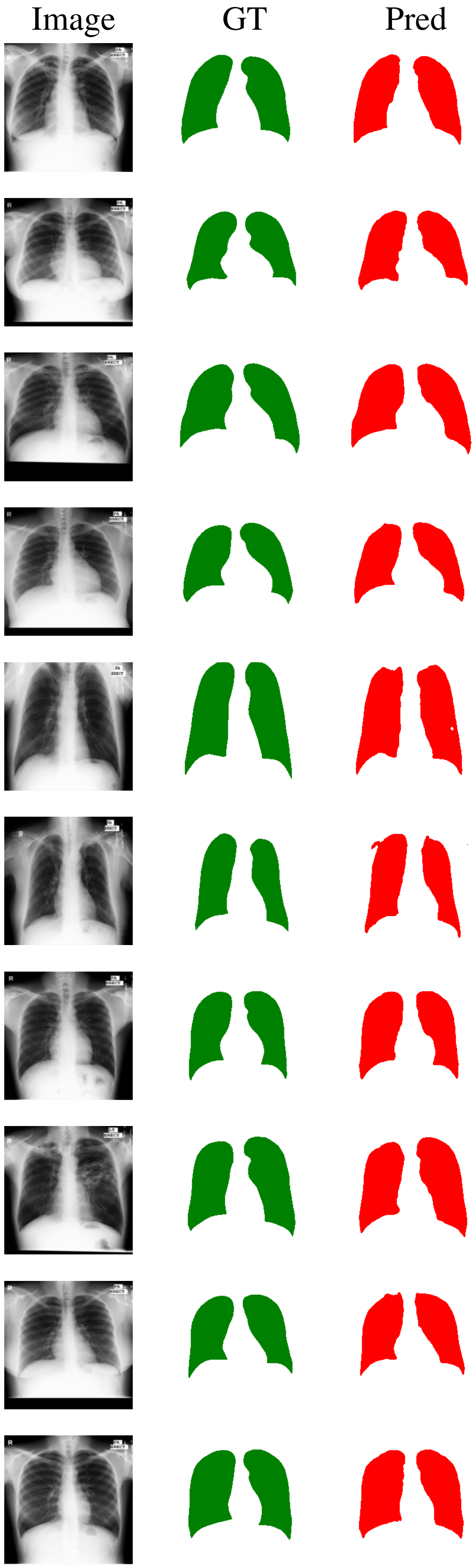

Fig. 2. Visualizations of the segmented lung masks by the MultiMix-50-1000 on the cross-domain test set. The results show good agreement between the groudtruth and segmented mask, especially on a difficult cross-domain evaluation.

Table 1. Architectural details of the Encoder in the MultiMix model: M denotes the minibatch size.

Name	Feature maps (input)	Feature maps (output)
Conv layer - 1	$M \times 256 \times 256 \times 1$	$M \times 256 \times 256 \times 16$
InstanceNorm - 1	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
LReLU - 1	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
Conv Layer - 2	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
InstanceNorm - 2	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
LReLU - 2	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
Dropout - 1	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
Maxpool - 1	$M \times 256 \times 256 \times 16$	$M \times 128 \times 128 \times 16$
Conv Layer - 3	$M \times 128 \times 128 \times 16$	$M \times 128 \times 128 \times 32$
InstanceNorm - 3	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
LReLU - 3	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
Conv Layer - 4	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
InstanceNorm - 4	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
LReLU - 4	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
Dropout - 2	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
Maxpool - 2	$M \times 128 \times 128 \times 32$	$M \times 64 \times 64 \times 32$
Conv Layer - 5	$M \times 64 \times 64 \times 32$	$M \times 64 \times 64 \times 64$
InstanceNorm - 5	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
LReLU - 5	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
Conv Layer - 6	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
InstanceNorm - 6	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
LReLU - 6	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
Dropout - 3	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
Maxpool - 3	$M \times 64 \times 64 \times 64$	$M \times 32 \times 32 \times 64$
Conv Layer - 7	$M \times 32 \times 32 \times 64$	$M \times 32 \times 32 \times 128$
InstanceNorm - 7	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
LReLU - 7	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
Conv Layer - 8	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
InstanceNorm - 8	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
LReLU - 8	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
Dropout - 4	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
Maxpool - 4	$M \times 32 \times 32 \times 128$	$M \times 16 \times 16 \times 128$
Conv Layer - 9	$M \times 16 \times 16 \times 128$	$M \times 16 \times 16 \times 256$
InstanceNorm - 9	$M \times 16 \times 16 \times 256$	$M \times 16 \times 16 \times 256$
LReLU - 9	$M \times 16 \times 16 \times 256$	$M \times 16 \times 16 \times 256$
Conv Layer - 10	$M \times 16 \times 16 \times 256$	$M \times 16 \times 16 \times 256$
InstanceNorm - 10	$M \times 16 \times 16 \times 256$	$M \times 16 \times 16 \times 256$
LReLU - 10	$M \times 16 \times 16 \times 256$	$M \times 16 \times 16 \times 256$
Dropout - 5	$M \times 16 \times 16 \times 256$	$M \times 16 \times 16 \times 256$
Maxpool - 5	$M \times 16 \times 16 \times 256$	$M \times 8 \times 8 \times 256$
Avgpool	$M \times 8 \times 8 \times 256$	$M \times 1 \times 1 \times 256$
GAP	$M \times 1 \times 1 \times 256$	$M \times 256$
Fully Connected Layer	$M \times 256$	$M \times 2$

Table 2. Architectural details of the Decoder in the MultiMix model: M denotes the minibatch size.

Name	Feature maps (input)	Feature maps (output)
Upsample - 1	$M \times 16 \times 16 \times 256$	$M \times 32 \times 32 \times 256$
Conv Layer - 1	$M \times 32 \times 32 \times 386$	$M \times 32 \times 32 \times 128$
InstanceNorm - 1	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
LReLU - 1	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
Conv Layer - 2	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
InstanceNorm - 2	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
LReLU - 2	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
Dropout - 1	$M \times 32 \times 32 \times 128$	$M \times 32 \times 32 \times 128$
Upsample - 2	$M \times 32 \times 32 \times 128$	$M \times 64 \times 64 \times 128$
Conv Layer - 3	$M \times 64 \times 64 \times 192$	$M \times 64 \times 64 \times 64$
InstanceNorm - 3	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
LReLU - 3	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
Conv Layer - 4	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
InstanceNorm - 4	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
LReLU - 4	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
Dropout - 2	$M \times 64 \times 64 \times 64$	$M \times 64 \times 64 \times 64$
Upsample - 3	$M \times 64 \times 64 \times 64$	$M \times 128 \times 128 \times 64$
Conv Layer - 5	$M \times 128 \times 128 \times 96$	$M \times 128 \times 128 \times 32$
InstanceNorm - 5	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
LReLU - 5	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
Conv Layer - 6	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
InstanceNorm - 6	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
LReLU - 6	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
Dropout - 3	$M \times 128 \times 128 \times 32$	$M \times 128 \times 128 \times 32$
Upsample - 4	$M \times 128 \times 128 \times 32$	$M \times 256 \times 256 \times 32$
Conv Layer - 7	$M \times 256 \times 256 \times 48$	$M \times 256 \times 256 \times 16$
InstanceNorm - 7	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
LReLU - 7	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
Conv Layer - 8	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
InstanceNorm - 8	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
LReLU - 8	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
Dropout - 4	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 16$
Final Conv Layer	$M \times 256 \times 256 \times 16$	$M \times 256 \times 256 \times 1$

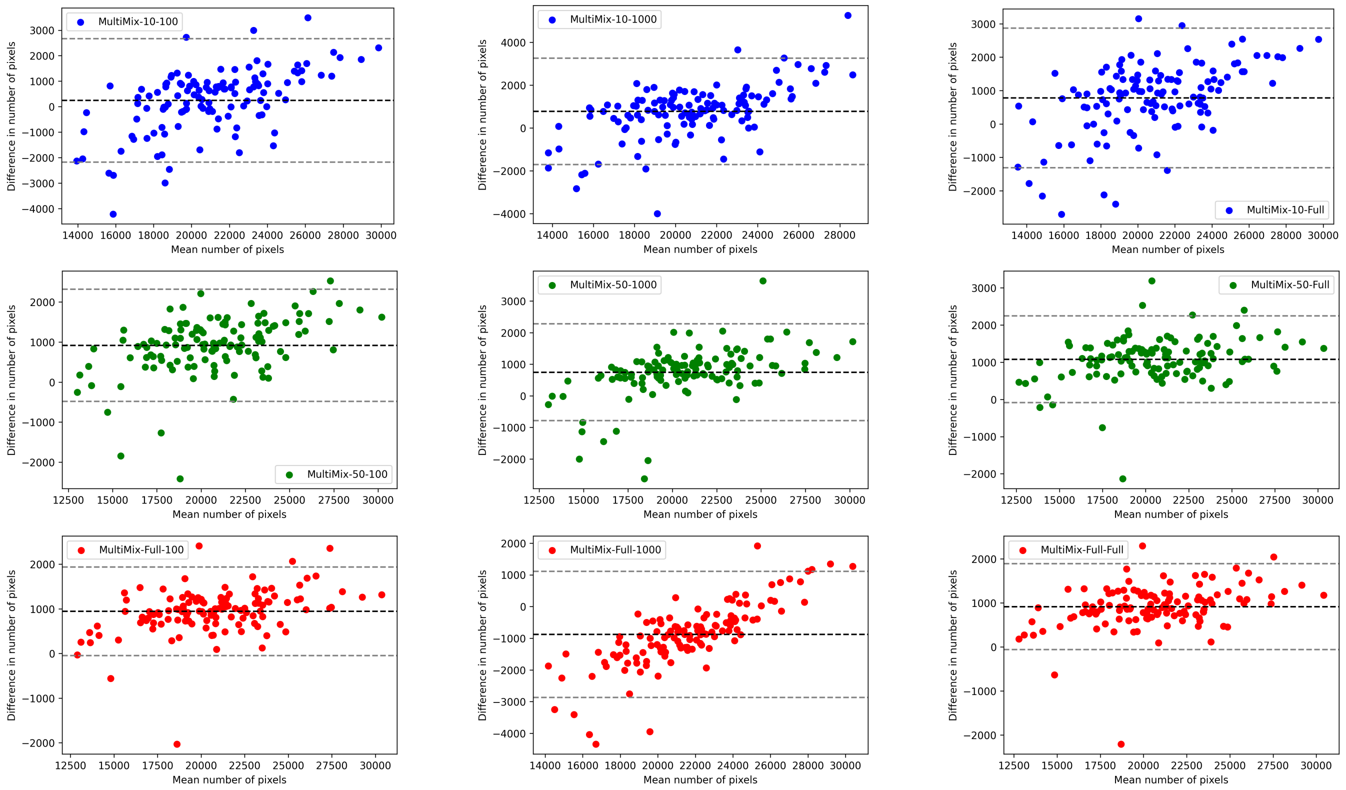

Fig. 3. Bland-Altman plots at varying training labels show good agreement between the number of ground truth pixels and MultiMix-predicted pixels for the in-domain evaluation, as well as consistent improvement with more labeled data.

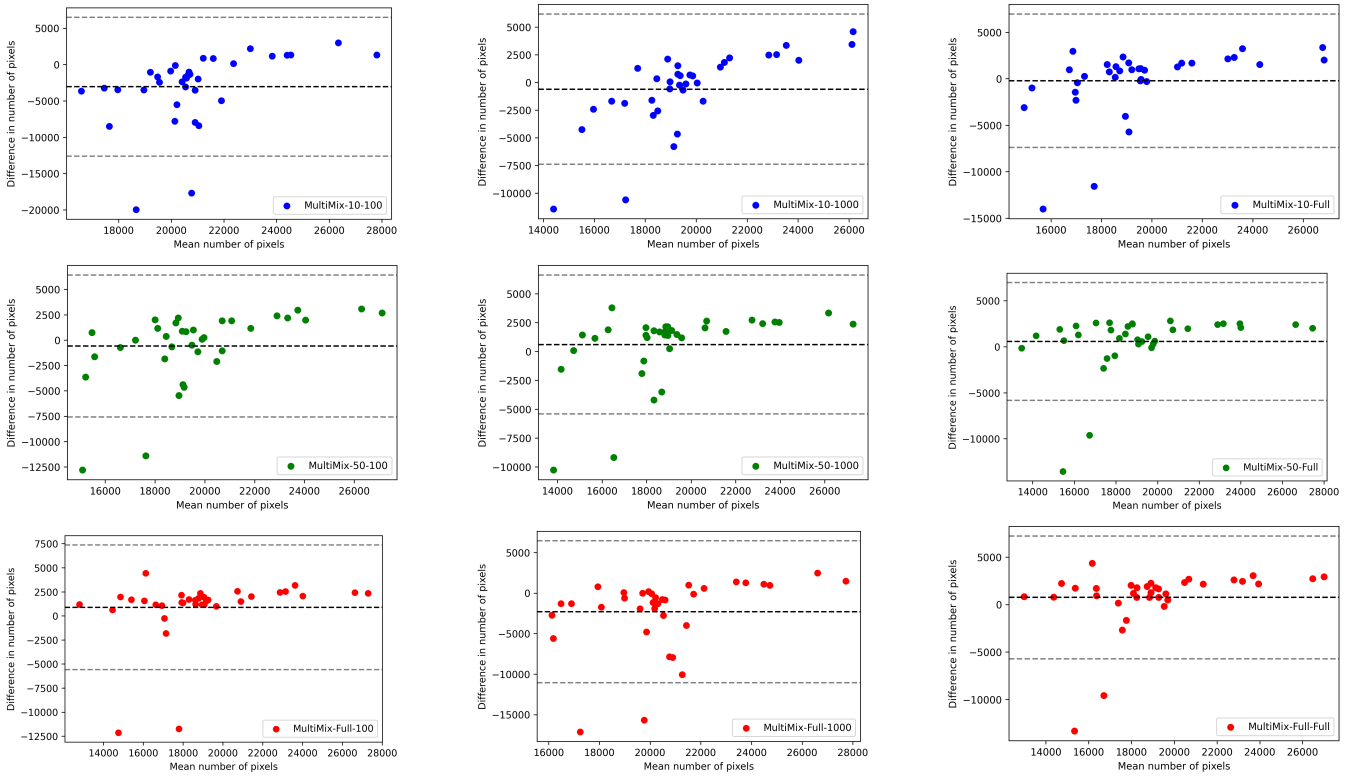

Fig. 4. Bland-Altman plots at varying training labels show good agreement between the number of ground truth pixels and MultiMix-predicted pixels for the cross-domain evaluation, as well as consistent improvement with more labeled data.

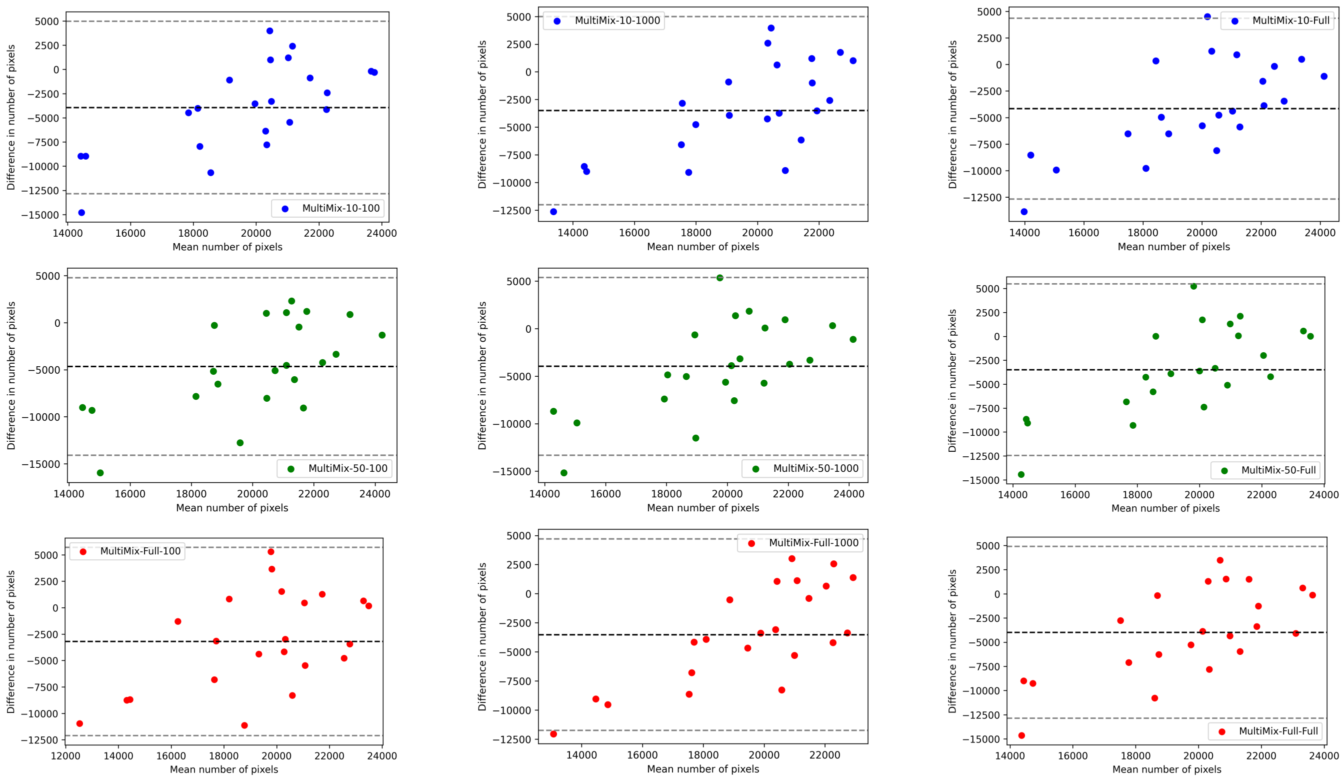

Fig. 5. Bland-Altman plots at varying training labels show good agreement between the number of ground truth pixels and MultiMix-predicted pixels for in-domain classification datasets, as well as consistent improvement with more labeled data.

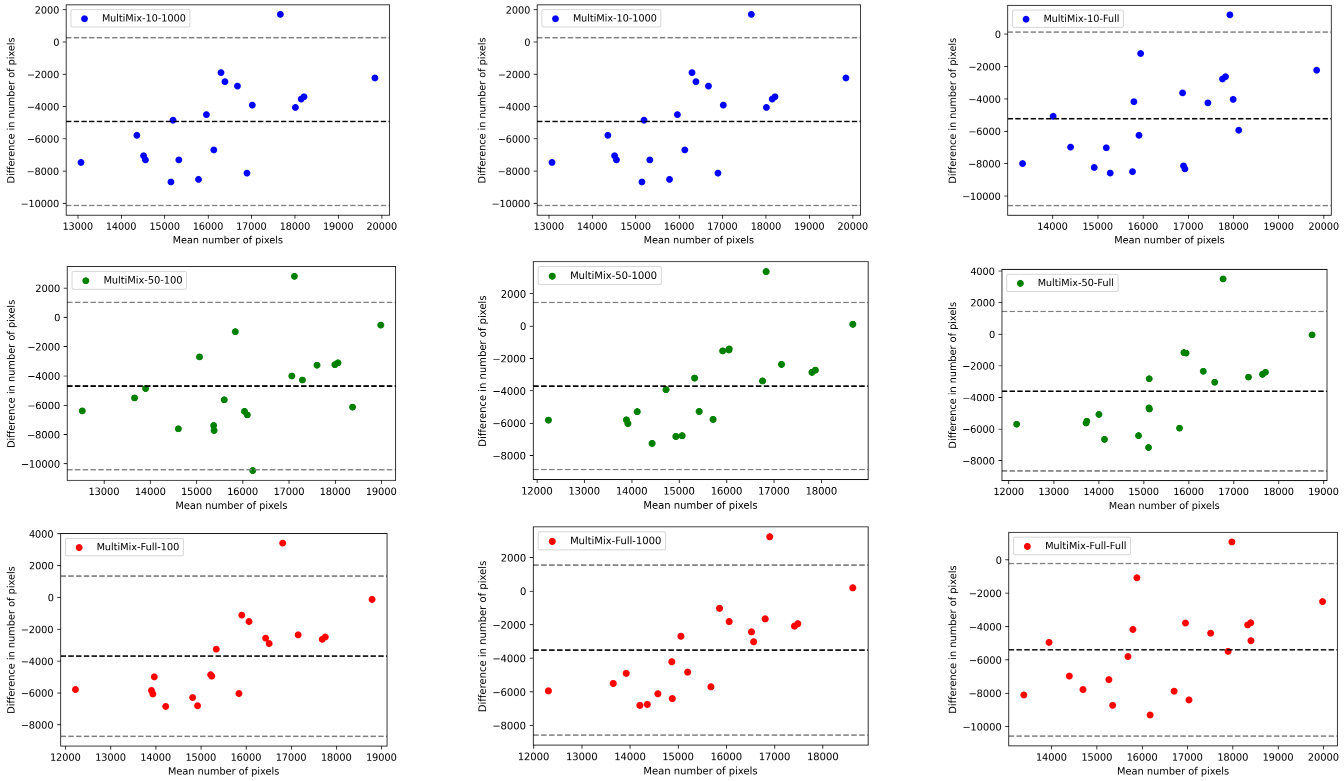

Fig. 6. Bland-Altman plots at varying training labels show good agreement between the number of ground truth pixels and MultiMix-predicted pixels for cross-domain classification datasets, as well as consistent improvement with more labeled data.

Table 3. Classification and segmentation performance with varying label proportions in in-domain evaluations: CheX (classification) and JSRT (segmentation) datasets. The best scores from fully-supervised models are underlined and the best scores from semi-supervised models are bolded.

Model	$\lvert\mathcal{D}_l^c\rvert$	Classification			$\lvert\mathcal{D}_l^s\rvert$	Segmentation					
		Acc	F1-Nor	F1-Abn		DS	JS	SSIM	HD	P	R
U-Net	—	—	—	—	10	0.634	0.695	0.810	2.899	0.779	0.865
	—	—	—	—	50	0.855	0.854	0.904	0.341	0.918	0.925
	—	—	—	—	Full	0.915	0.906	0.929	0.104	0.949	0.953
Enc	100	0.732	0.424	0.806	—	—	—	—	—	—	—
	1000	0.773	0.546	0.842	—	—	—	—	—	—	—
	Full	0.737	0.534	0.838	—	—	—	—	—	—	—
Enc-SSL	100	0.780	0.570	0.844	—	—	—	—	—	—	—
	1000	0.822	0.692	0.876	—	—	—	—	—	—	—
	Full	0.817	0.680	0.872	—	—	—	—	—	—	—
UMTL	100	0.707	0.443	0.797	10	0.626	0.871	0.908	4.323	0.900	0.964
	100	0.655	0.683	0.853	50	0.647	0.854	0.881	4.733	0.864	0.989
	100	0.706	0.416	0.804	Full	0.696	0.872	0.911	3.908	0.892	0.986
	1000	0.750	0.490	0.825	10	0.761	0.904	0.926	3.050	0.924	0.977
	1000	0.749	0.510	0.833	50	0.768	0.927	0.938	2.606	0.940	0.985
	1000	0.747	0.530	0.840	Full	0.759	0.928	0.930	2.955	0.924	0.981
	Full	0.744	0.515	0.828	10	0.909	0.919	0.521	0.903	0.912	0.962
	Full	0.738	0.438	0.820	50	0.930	0.948	0.954	0.444	0.969	0.977
	Full	0.731	0.447	0.822	Full	0.932	0.951	0.957	0.372	0.965	0.977
UMTL-S	100	0.704	0.358	0.806	10	0.922	0.848	0.891	4.005	0.871	0.966
	100	0.701	0.336	0.796	50	0.926	0.867	0.894	4.393	0.873	0.891
	100	0.713	0.442	0.794	Full	0.931	0.890	0.920	3.983	0.906	0.980
	1000	0.740	0.482	0.828	10	0.948	0.908	0.924	2.546	0.931	0.972
	1000	0.771	0.566	0.844	50	0.965	0.931	0.941	2.083	0.949	0.981
	1000	0.742	0.497	0.830	Full	0.962	0.925	0.935	1.758	0.958	0.985
	Full	0.747	0.500	0.830	10	0.955	0.914	0.936	0.568	0.954	0.956
	Full	0.737	0.433	0.820	50	0.972	0.944	0.953	0.560	0.966	0.977
	Full	0.723	0.413	0.817	Full	0.974	0.953	0.957	0.539	0.967	0.981
UMTL-SSL	100	0.790	0.618	0.856	10	0.906	0.925	0.940	0.626	0.954	0.953
	100	0.818	0.688	0.872	50	0.919	0.946	0.952	0.561	0.962	0.963
	100	0.852	0.670	0.868	Full	0.937	0.954	0.958	0.613	0.969	0.981
	1000	0.794	0.630	0.860	10	0.893	0.926	0.941	0.524	0.961	0.962
	1000	0.822	0.693	0.877	50	0.903	0.945	0.952	0.712	0.963	0.980
	1000	0.818	0.707	0.867	Full	0.899	0.953	0.958	0.724	0.968	0.982
	Full	0.812	0.688	0.870	10	0.905	0.921	0.935	0.627	0.946	0.973
	Full	0.813	0.683	0.873	50	0.927	0.947	0.954	**0.397**	0.968	0.977
	Full	0.816	0.678	0.873	Full	0.935	0.954	0.958	0.625	0.970	0.981
UMTL-SSL-S	100	0.798	0.628	0.860	10	0.951	0.911	0.935	0.792	0.940	0.963
	100	0.834	0.696	0.874	50	0.972	0.946	0.952	0.727	0.965	0.977
	100	0.817	0.688	0.860	Full	0.975	0.951	0.954	0.812	0.968	0.981
	1000	0.806	0.652	0.872	10	0.956	0.916	0.937	0.852	0.943	0.966
	1000	0.808	0.662	0.862	50	0.971	0.944	0.952	0.917	0.965	0.978
	1000	0.801	0.646	0.862	Full	0.975	0.952	0.954	0.753	0.969	0.981
	Full	0.796	0.632	0.864	10	0.960	0.923	0.940	0.782	0.954	0.967
	Full	0.808	0.662	0.868	50	0.972	0.945	0.953	0.645	0.966	0.978
	Full	0.800	0.632	0.628	Full	0.961	0.924	0.940	0.392	0.948	0.969
MultiMix	100	0.800	0.594	0.856	10	0.954	0.920	0.938	0.695	0.949	0.969
	100	0.824	0.613	0.854	50	0.971	0.943	0.951	0.681	0.964	0.976
	100	0.792	0.593	0.854	Full	0.973	0.948	0.954	0.636	0.966	0.981
	1000	0.817	0.647	0.865	10	0.954	0.910	0.932	0.902	0.942	0.968
	1000	0.825	0.650	0.860	50	0.970	0.941	0.950	0.811	0.964	0.977
	1000	0.830	0.586	0.856	Full	**0.974**	0.919	0.953	0.643	0.933	0.984
	Full	0.840	0.730	0.880	10	0.954	0.913	0.935	0.621	0.949	0.968
	Full	**0.854**	**0.760**	**0.890**	50	0.972	**0.950**	**0.956**	0.692	**0.970**	**0.980**
	Full	0.843	0.740	0.890	Full	0.975	0.952	0.960	0.528	0.970	0.982

Table 4. Lung segmentation performance comparison with varying label proportions in cross-domain evaluations: NIH (classification) and NLM (segmentation) datasets. The best scores from fully-supervised models are underlined and the best scores from semi-supervised models are bolded.

Model	$\lvert\mathcal{D}_l^c\rvert$	Classification			$\lvert\mathcal{D}_l^s\rvert$	Segmentation					
		Acc	F1-Nor	F1-Abn		DS	JS	SSIM	HD	P	R
U-Net	—	—	—	—	10	0.555	0.480	0.680	8.691	0.553	0.866
	—	—	—	—	50	0.763	0.736	0.870	2.895	0.752	0.887
	—	—	—	—	Full	0.838	0.906	0.929	1.414	0.793	0.910
Enc	100	0.352	0.070	0.506	—	—	—	—	—	—	—
	1000	0.390	0.192	0.508	—	—	—	—	—	—	—
	Full	0.434	0.296	0.524	—	—	—	—	—	—	—
Enc-SSL	100	0.402	0.222	0.510	—	—	—	—	—	—	—
	1000	0.486	0.380	0.530	—	—	—	—	—	—	—
	Full	0.510	0.472	0.538	—	—	—	—	—	—	—
UMTL	100	0.350	0.045	0.510	10	0.586	0.708	0.836	7.156	0.731	0.950
	100	0.363	0.085	0.515	50	0.580	0.684	0.825	7.013	0.697	0.975
	100	0.342	0.015	0.508	Full	0.607	0.742	0.863	6.398	0.759	0.968
	1000	0.413	0.263	0.507	10	0.676	0.674	0.833	3.268	0.712	0.927
	1000	0.400	0.203	0.513	50	0.704	0.811	0.896	3.232	0.828	0.964
	1000	0.430	0.293	0.517	Full	0.638	0.795	0.890	3.893	0.810	0.966
	Full	0.455	0.365	0.525	10	0.737	0.765	0.879	0.917	0.801	0.930
	Full	0.444	0.332	0.522	50	0.868	0.793	0.894	0.742	0.898	0.946
	Full	0.443	0.328	0.520	Full	0.854	0.828	0.913	0.792	0.866	0.942
UMTL-S	100	0.344	0.006	0.510	10	0.797	0.670	0.807	5.754	0.698	0.938
	100	0.364	0.098	0.506	50	0.828	0.715	0.826	6.412	0.731	0.971
	100	0.342	0.008	0.510	Full	0.838	0.715	0.834	6.321	0.740	0.966
	1000	0.378	0.138	0.512	10	0.844	0.718	0.854	3.921	0.754	0.939
	1000	0.392	0.186	0.514	50	0.883	0.793	0.888	3.017	0.821	0.959
	1000	0.370	0.130	0.510	Full	0.898	0.831	0.905	4.150	0.845	0.970
	Full	0.470	0.398	0.524	10	0.881	0.785	0.888	0.862	0.830	0.939
	Full	0.413	0.270	0.510	50	0.917	0.848	0.919	0.658	0.966	0.888
	Full	0.433	0.315	0.513	Full	0.916	0.850	0.921	0.882	0.886	0.952
UMTL-SSL	100	0.442	0.316	0.524	10	0.833	0.778	0.884	0.895	0.810	0.948
	100	0.398	0.166	0.520	50	0.853	0.839	0.907	0.851	0.864	0.952
	100	0.385	0.165	0.515	Full	0.841	0.818	0.911	0.853	0.854	0.949
	1000	0.445	0.333	0.525	10	0.818	0.781	0.892	1.085	0.825	0.938
	1000	0.526	0.486	0.544	50	0.826	0.804	0.904	0.811	0.792	0.949
	1000	0.485	0.413	0.538	Full	0.843	0.837	0.924	0.983	0.882	0.953
	Full	0.526	0.504	0.546	10	0.824	0.765	0.873	0.994	0.790	0.943
	Full	0.530	0.514	0.542	50	0.867	0.839	0.917	0.566	0.881	0.945
	Full	$\underline{0.520}$	$\underline{0.490}$	0.542	Full	0.884	0.884	0.934	0.599	0.918	0.955
UMTL-S-SSL	100	0.370	0.114	0.510	10	0.853	0.747	0.866	1.048	0.782	0.944
	100	0.400	0.192	0.518	50	0.889	0.799	0.899	0.854	0.834	0.950
	100	0.370	0.114	0.514	Full	0.915	0.848	0.920	0.987	0.880	0.956
	1000	0.432	0.286	0.524	10	0.871	0.785	0.884	1.327	0.818	0.944
	1000	0.458	0.342	0.530	50	0.893	0.803	0.895	1.123	0.835	0.946
	1000	0.462	0.350	0.536	Full	0.930	0.860	0.925	1.042	0.912	0.955
	Full	0.482	0.412	0.536	10	0.880	0.765	0.885	0.745	0.818	0.941
	Full	0.490	0.426	0.540	50	0.912	0.845	0.909	0.956	0.881	0.952
	Full	0.510	0.474	0.540	Full	0.875	0.809	0.875	0.722	0.851	0.944
MultiMix	100	0.440	0.164	0.510	10	0.857	0.732	0.863	1.227	0.767	0.943
	100	0.370	0.036	0.510	50	0.889	0.790	0.890	1.061	0.866	0.947
	100	0.500	0.300	0.510	Full	0.899	0.825	0.906	0.647	0.852	0.952
	1000	0.520	0.386	0.530	10	0.862	0.775	0.878	1.307	0.816	0.939
	1000	0.540	0.500	0.536	50	0.912	0.831	0.907	1.293	0.865	0.955
	1000	**0.570**	**0.620**	0.510	Full	**0.936**	**0.880**	**0.932**	0.803	0.917	**0.979**
	Full	0.550	0.430	0.534	10	0.886	0.802	0.894	0.746	0.839	0.948
	Full	0.560	0.570	**0.550**	50	0.935	0.878	0.930	**0.515**	**0.928**	0.957
	Full	$\underline{0.520}$	$\underline{0.490}$	$\underline{0.550}$	Full	$\underline{0.943}$	$\underline{0.892}$	$\underline{0.937}$	$\underline{0.417}$	$\underline{0.928}$	$\underline{0.958}$

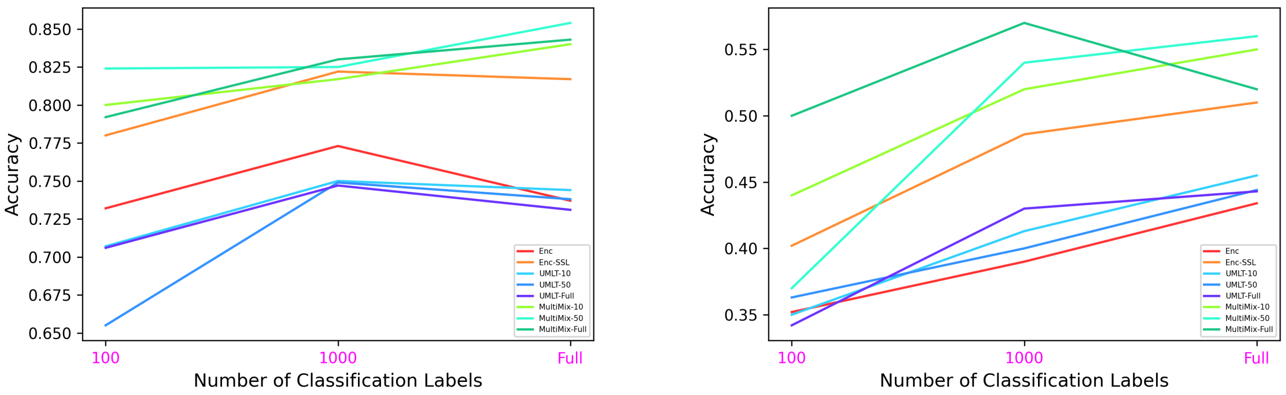

Fig. 7. Classification accuracies of different supervised and semi-supervised baselines at different training datasizes. The in-domain (left) and cross-domain (right) plots show that MultiMix has higher accuracy and consistency.

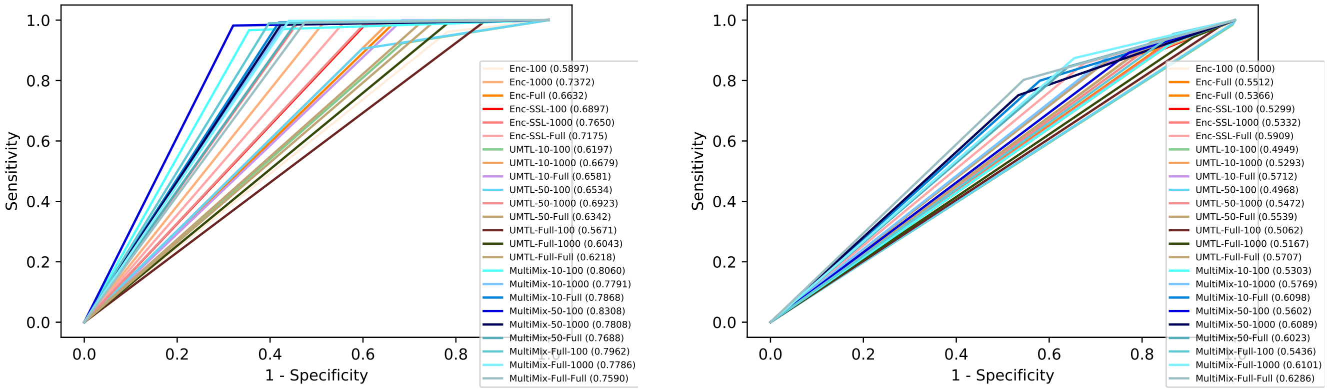

Fig. 8. ROC Curves for supervised and semi-supervised baselines and MultiMix labels shows higher AUC values from our MultiMix for in-domain (left) and cross-domain (right) evaluations.