Generalized Multi-Task Learning from Substantially Unlabeled Multi-Source Medical Image Data

Ayaan Haque*   ayaanzhaque@gmail.com
Saratoga High School, Saratoga, CA, USA

Abdullah-Al-Zubaer Imran*   aimran@stanford.edu
Stanford University, Stanford, CA, USA

Adam Wang   adamwang@stanford.edu
Stanford University, Stanford, CA, USA

Demetri Terzopoulos   dt@cs.ucla.edu
University of California, Los Angeles, CA, USA
VoxelCloud, Inc., Los Angeles, CA, USA

Abstract

Deep learning-based models, when trained in a fully-supervised manner, can be effective in performing complex image analysis tasks, although contingent upon the availability of large labeled datasets. Especially in the medical imaging domain, however, expert image annotation is expensive, time-consuming, and prone to variability. Semi-supervised learning from limited quantities of labeled data has shown promise as an alternative. Maximizing knowledge gains from copious unlabeled data benefits semi-supervised learning models. Moreover, learning multiple tasks within the same model further improves its generalizability. We propose MultiMix, a new multi-task learning model that jointly learns disease classification and anatomical segmentation in a semi-supervised manner, while preserving explainability through a novel saliency bridge between the two tasks. Our experiments with varying quantities of multi-source labeled data in the training sets confirm the effectiveness of MultiMix in the simultaneous classification of pneumonia and segmentation of the lungs in chest X-ray images. Moreover, both in-domain and cross-domain evaluations across these tasks further showcase the potential of our model to adapt to challenging generalization scenarios.

Keywords: Multi-Task Learning, Semi-Supervised Learning, Data Augmentation, Saliency Bridge, Classification, Segmentation, Chest X-Ray, Lungs, Pneumonia

1. Introduction

Learning-based medical image analysis has become widespread with the advent of deep learning. However, most deep learning models rely on large pools of labeled data. Especially in the medical imaging domain, obtaining copious labeled imagery is often infeasible, as annotation requires substantial domain expertise and manual labor. Therefore, developing large-scale deep learning methodologies for medical image analysis tasks is challenging. In confronting the limited labeled data problem, Semi-Supervised Learning (SSL) has been gaining attention. In semi-supervised learning, unlabeled training examples are leveraged in combination with labeled examples to maximize information gains (Chapelle et al.,

---

*. Co-first author

2009). Specifically within the medical domain, where collecting data is generally easier than annotating those data, the use of deep learning for medical image analysis tasks can be fostered by leveraging semi-supervised learning.

Recent research has yielded a variety of semi-supervised learning techniques (Imran, 2020). Pseudo-labeling (Lee, 2013) trains a model with labeled data and unlabeled data simultaneously, generating labels for the unlabeled data by assuming the model-predicted labels to be reliable. Similarly, entropy minimization (Grandvalet and Bengio, 2005) trains so as to match the predicted data distribution of unlabeled data with that of the labeled data, under the assumption that unlabeled examples should yield prediction distributions that are similar to those from labeled examples (Ouali et al., 2020). Domain adaptation (Beijbom, 2012) is a form of inductive transfer learning, where a model is trained on labeled data from the source domain as well as labeled plus unlabeled data from the target domain, which improves model generalization for the target domain, but lacks clinical value if the target domain data is inaccessible during training.

Thus, progress has been made in learning from limited labeled data, although mainly within the confines of single-task learning. In particular, individual medical imaging tasks, such as diagnostic classification and anatomical segmentation, have been addressed using state-of-the-art Convolutional Neural Network (CNN) models (Anwar et al., 2018); e.g., for medical image segmentation, encoder-decoder networks (Ronneberger et al., 2015), variational auto-encoder networks, (Myronenko, 2018), context encoder networks (Gu et al., 2019), multiscale adversarial learning (Imran and Terzopoulos, 2021a), etc.

By contrast, Multi-Task Learning (MTL) is defined as optimizing more than one loss in a single model such that multiple related tasks are performed by sharing the learned representation (Ruder, 2017). Jointly training multiple tasks within a model improves the generalizability of the model as each of the tasks regularizes the others (Caruana, 1993). Assuming that training data with limited annotations come from different distributions for different tasks, multi-task learning may be useful in such scenarios for learning in a scarcely-supervised manner (Imran et al., 2020; Imran and Terzopoulos, 2021b).

Combining the objectives of substantially unlabeled data training and multi-task learning, Semi-Supervised Multi-Task Learning (SSMTL) is a promising research area in the context of medical image analysis. While there have been prior efforts on multi-tasking (Mehta et al., 2018; Girard et al., 2019), rarely do they focus on incorporating semi-supervised learning particularly within the medical realm. Liu et al. (2008) proposed a general semi-supervised multi-tasking method that uses soft-parameter sharing to allow multiple classification tasks in a single model. Gao et al. (2019) performed multi-tasking on tasks within the same medical domain by exploiting feature transfer. Adversarial learning (Salimans et al., 2016) combines a classifier with a discriminator to perform semi-supervised, adversarial multi-tasking. Imran and Terzopoulos (2019) introduced semi-supervised multi-task learning using adversarial learning and attention masking. Zhou et al. (2019) proposed a semi-supervised multi-tasking model that uses an attention mechanism to grade segmented retinal images. None of the aforecited works, however, take into consideration the disparity in the training data distributions for multiple tasks.

To learn diagnostic classification and anatomical segmentation jointly from substantially unlabeled multi-source data, we propose *MultiMix*, a novel, better-generalized multi-tasking model that incorporates confidence-based augmentation and a module that bridges the

classification and segmentation tasks. This *saliency bridge module* produces a saliency map by computing the gradient of the class score with respect to the input image, thus not only enabling the analysis of the model's predictions, but also improving the model's performance of both tasks. While the explainability of any deep learning model can be based on visualizing saliency maps (Simonyan et al., 2014; Zhang et al., 2016; Hu et al., 2019), to our knowledge a saliency bridge between two shared tasks within a single model has not previously been explored. We demonstrate that the saliency bridge module in conjunction with a simple yet effective semi-supervised learning method in a multi-tasking setting can yield improved and consistent performance across multiple domains.

This article is a revised and extended version of our ISBI publication (Haque et al., 2021).[1] Our main contributions may be summarized as follows:

- A new semi-supervised learning model, MultiMix, that exploits confidence-based data augmentation and consistency regularization to jointly learn diagnostic classification and anatomical segmentation from multi-source, multi-domain medical image datasets.

- Incorporation of an innovative saliency bridge module connecting the segmentation and classification branches of the model, resulting in the improved performance of both tasks.

- Substantiation of the improved generalizability (both in-domain and cross-domain) of the proposed model via experimentation with varied quantities of labeled data and mixed data sources related to multiple tasks, specifically in the classification of pneumonia and the simultaneous segmentation of the lungs in chest X-ray images.

- MultiMix software made available at `https://github.com/ayaanzhaque/MultiMix`.

2. The MultiMix Model

To formulate our approach, we assume unknown data distributions $p(X^c, C)$ over images X^c and class labels C as well as $p(X^s, S)$ over images X^s and segmentation labels S. Hence, segmentation labels for the X^c images and class labels for the X^s images are unavailable. We also assume access to labeled training sets $\mathcal{D}_l^c$ sampled i.i.d. from $p(X^c, C)$ and $\mathcal{D}_l^s$ sampled i.i.d. from $p(X^s, S)$, along with unlabeled training sets $\mathcal{D}_u^c$ sampled i.i.d. from $p(X^c)$ and $\mathcal{D}_u^s$ sampled i.i.d. from $p(X^s)$, after marginalizing out C and S, respectively.

In our MultiMix model (Figure 1), we utilize a U-Net-like (Ronneberger et al., 2015) encoder-decoder architecture for image deconstruction and reconstruction. The encoder functions similarly to a standard CNN. To perform multi-tasking, we use pooling layers followed by fully-connected layers, allowing the encoder to output class predictions through the classification branch of the model. Furthermore, in the segmentaton branch of the model, the segmentation predictions are obtained as the output of the decoder.

MultiMix performs multi-tasking in a semi-supervised learning manner, assuming the training data for the two tasks come from disparate distributions. It is well established

---

1. With an augmented literature review, a more detailed explanation of the methods, model architecture, and training algorithm, further details about the datasets, saliency map visualizations from multiple datasets, and additional results and discussion supported by quantitative (performance metrics tables) and qualitative (mask predictions, Bland Altman plots, ROC curves, consistency plots) characteristics.
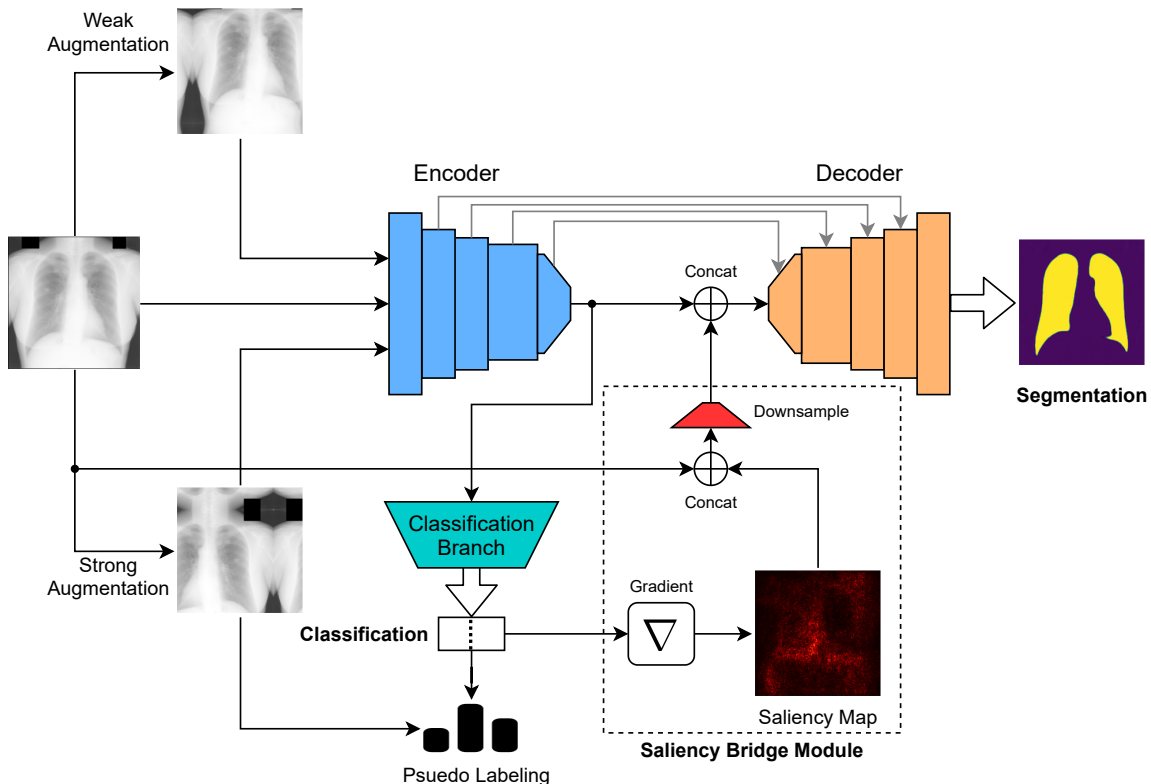
Figure 1: Schematic of the MultiMix model. *Classification:* Using predictions on unlabeled weakly augmented images, pseudo-labels are generated with confidence, and loss is computed with these labels and the strongly augmented versions of those images. *Segmentation:* Saliency maps generated from the class predictions are concatenated via the saliency bridge module to guide the decoder in generating the segmentation masks.

that a multi-tasking model usually outperforms its single-task counterparts (Imran and Terzopoulos, 2019; Imran et al., 2020; Imran, 2020). The shared encoder in the MultiMix model learns features useful for addressing both the classification and segmentation tasks. This joint representation learning enables the model to avoid overfitting and generalize better. Most importantly, it exploits the relatedness of the tasks, which is crucial for effective multi-tasking.

In the following sections, we describe the classification and segmentation branches of the MultiMix model, explain the saliency bridge module that bridges the two branches, and specify the MultiMix training procedure.

2.1 Classification Branch

For semi-supervised classification, we leverage data augmentation and pseudo-labeling. Inspired by the work of Sohn et al. (2020), for each unlabeled image we perform two degrees of augmentation: weak and strong. The former consists of standard augmentations—both random horizontal flipping and random cropping—and is applied to the labeled data as

well, whereas the latter is performed by randomly applying any number of augmentations from a pool of "heavy" augmentations.[2] An unlabeled image x_u^c is first weakly augmented, $x_w^c = \text{WAug}(x_u^c)$, and a pseudo-label $c_p = \arg\max(\hat{c}_w) \geq t$ is synthesized from x_w^c using the model prediction $\hat{c}_w$. The image-label pair is retained only if the confidence with which the model generates the pseudo-label exceeds the experimentally tuned threshold t, thus deterring learning from poor and incorrect labels. Second, $x_g^c = \text{GAug}(x_u^c)$ are strongly augmented versions of x_u^c.

Our training strategy promotes effective learning from large amounts of unlabeled data, which is challenging. At first, the predictions are less reliable as the model begins to learn mainly from the labeled data, but the model gains confidence with the generation of labels for the unlabeled images and, as a result, it becomes more proficient. Since the unlabeled examples are incrementally added to the training set, subject to the threshold, the model learns to predict more accurately in a progressive manner and, with increasing confidence, the performance of the model improves at an increasingly higher rate. Furthermore, employing two degrees of augmentation enables the model to maximize its knowledge gain from the unlabeled data due to the enhanced image diversity through what is known as consistency learning as, in theory, two augmented versions of the same image should yield the same prediction, which is encouraged using an unsupervised loss. In other words, imposing on the model, through an unsupervised loss, to produce the same predictions on images subjected to two different degrees of augmentation results in better classification performance.

The classification objective

$$L^c(c_l, \hat{c}_l, c_p, \hat{c}_g) = L_l(c_l, \hat{c}_l) + \lambda L_u(c_p, \hat{c}_g) \tag{1}$$

includes a supervised loss component L_l for the labeled data, which uses cross-entropy between the reference class label c_l and the model prediction $\hat{c}_l$, as well as an unsupervised loss component L_u for the unlabeled data, which uses cross-entropy between the pseudo-label c_p and the model prediction $\hat{c}_g$.

Note that the model is trained to ignore GAug as it is provided the pseudo-label c_p. Since the underlying data distributions are the same for both augmentations, it is compelled to learn that for the sake of consistency. Weak augmentations are used to produce reliable and usable pseudo-labels whereas strong augmentations are used to provide a difficult challenge for the model. This difficulty forces the model to learn more effective representations in order to be accurate, and it also prevents overfitting from minimizing the loss too early. With the assumption that the weakly augmented image has the correct label to be associated with the strongly augmented image, the model is empowered to discern the augmentations in the image, and its performance improves as a result, by learning the underlying features crucial to the diagnosis. This helps achieve better generalization despite the differences in data distributions across different domains. By teaching the model to learn only the more salient representations that will exist to some extent in all domains, it can generalize and be effective across domains.

---

2. This pool includes augmentations such as horizontal flip, crop, autocontrast, brightness, contrast, equalize, identity, posterize, rotate, sharpness, shearX, shearY, solarize, translateX, and TranslateY. Autocontrast, brightness, contrast, and equalize are all severe image intensity modifications.

2.2 Segmentation Branch

For segmentation, the predictions are made through the encoder-decoder architecture with skip connections. For the labeled samples x_l^s, we calculate the direct segmentation loss in the form of Dice loss $L_l(s_l, \hat{s}_l)$ between the reference lung mask s_l and predicted segmentation $\hat{s}_l$. Since we do not have the segmentation masks for the unlabeled examples x_u^s, we cannot directly calculate the segmentation loss for them. To ensure consistency, we compute the KL divergence $L_u(\hat{s}_l, \hat{s}_u)$ between segmentation predictions for the labeled examples and unlabeled examples $\hat{s}_u$. This penalizes the model for making predictions that increasingly differ from those of the labeled data, which helps the model fit the unlabeled data. The total segmentation objective is therefore

$$L^s(s_l, \hat{s}_l, \hat{s}_u) = \alpha L_l(s_l, \hat{s}_l) + \beta L_u(\hat{s}_l, \hat{s}_u), \tag{2}$$

where α and β are weights.

2.3 Saliency Bridge Module

We incorporate a saliency bridge module to bridge between the classification and segmentation branches of the MultiMix model, as indicated in Figure 1. To learn which image regions are most relevant to classification, saliency maps

$$y_l = \text{Saliency}(\hat{c}_l^s) \quad \text{and} \quad y_u = \text{Saliency}(\hat{c}_u^s), \tag{3}$$

where $\hat{c}_l^s$ and $\hat{c}_u^s$ denote the class predictions for the input images x_l^s and x_u^s, respectively, are generated from the classification branch by computing the gradient of the predicted class with respect to the input image.[3] It cannot be directly known if the image samples in $\mathcal{D}^s$ represent normal or diseased cases, thus x_l^s and x_u^s are considered to be unlabeled for the classification task. Therefore, the saliency maps generated via the class prediction are not true segmentation maps, but they will nonetheless highlight the lungs or lung regions relevant to the particular disease class (see Appendix C).

The outputs of the saliency bridge module,

$$b_l = y_l \oplus x_l^s \quad \text{and} \quad b_u = y_u \oplus x_u^s, \tag{4}$$

obtained by concatenating the saliency maps with the associated input images, are further downsampled before they are concatenated with the encoder-decoder bottleneck in the segmentation branch. This results in a tighter connection between the classification and segmentation tasks and improves the effectiveness of the bridge module, which retains important information from the encoder that may otherwise be lost because of the repeated convolutions. The saliency maps serve to guide the segmentation during the decoding phase, yielding improved segmentation while learning from limited labeled data. With improving classification performance, the saliency maps become more accurate, thus yielding improved

---

3. These saliency maps should not be confused with simultaneous segmentation and saliency detection or prediction, where a semantic segmentation model is trained to produce saliency maps to accompany the output segmentation masks; e.g., (Zeng et al., 2019). Our saliency bridge module is novel in that it performs a saliency analysis of MultiMix's classification branch and leverages it to improve the performance of its semantic segmentation branch.

Algorithm 1 MultiMix Mini-Batch Training

Require:

  Training set of labeled classification data $\mathcal{D}_l^c$

  Training set of labeled segmentation data $\mathcal{D}_l^s$

  Training set of unlabeled classification data $\mathcal{D}_u^c$

  Training set of unlabeled segmentation data $\mathcal{D}_u^s$

  Network architecture $\mathcal{F}_\theta$ with learnable parameters θ

  Minibatch size m

  repeat

    Create labeled classification minibatch: $\{^1x_l^c, \ldots, {}^mx_l^c\} \sim \mathcal{D}_l^c$

    Create labeled segmentation minibatch: $\{^1x_l^s, \ldots, {}^mx_l^s\} \sim \mathcal{D}_l^s$

    Create unlabeled classification minibatch: $\{^1x_u^c, \ldots, {}^mx_u^c\} \sim \mathcal{D}_u^c$

    Create unlabeled segmentation minibatch: $\{^1x_u^s, \ldots, {}^mx_u^s\} \sim \mathcal{D}_u^s$

    Compute predictions for the labeled data: ${}^i\hat{c}_l \leftarrow \mathcal{F}_\theta({}^ix_l^c)$; ${}^i\hat{s}_l \leftarrow \mathcal{F}_\theta({}^ix_l^s)$

    Generate weakly-augmented samples: ${}^ix_w^c \leftarrow \mathrm{WAug}({}^ix_u^c)$

    Generate strongly-augmented samples: ${}^ix_g^c \leftarrow \mathrm{GAug}({}^ix_u^c)$

    Compute predictions for the unlabeled data: ${}^i\hat{c}_w \leftarrow \mathcal{F}_\theta({}^ix_w^c)$; ${}^i\hat{c}_g \leftarrow \mathcal{F}_\theta({}^ix_g^c)$; ${}^i\hat{s}_u \leftarrow \mathcal{F}_\theta({}^ix_u^s)$

    Compute pseudo label: ${}^ic_p \leftarrow \arg\max({}^i\hat{c}_w) \geq t$

    Update $\mathcal{F}_\theta$ by backpropagating the loss gradient $\nabla_\theta L$

  until convergence

---

segmentations, since the shared parameters responsible for improved classification produce a feedback loop that allows both tasks to improve jointly.

Conventionally, saliency maps are used to analyze which features and areas of the image are most relevant for classification, thereby enhancing understanding of the model's learning process. Similarly, our saliency module is explainable, as it is a relevant connection between the classification and segmentation tasks (although model explainability in and of itself is not the main focus of our work). Since the saliency maps are comparable to segmentation masks, it is sensible to employ them to guide the decoder in the task of segmentation. Multi-tasking requires the tasks to be somewhat related, so our task-relevant bridge fosters a tighter bond between classification and segmentation.

2.4 MultiMix Training Procedure

Algorithm 1 presents the main steps of the MultiMix training procedure applied to labeled and unlabeled classification and segmentation training data. The model is trained simultaneously on the classification objective (1) and segmentation objective (2) using the following total loss for a minibatch size of m:

$$L = \frac{1}{m} \sum_{i=1}^m \left(L^c \left({}^ic_l, {}^i\hat{c}_l, {}^ic_p, {}^i\hat{c}_g \right) + L^s \left({}^is_l, {}^i\hat{s}_l, {}^i\hat{s}_u \right) \right) \tag{5}$$
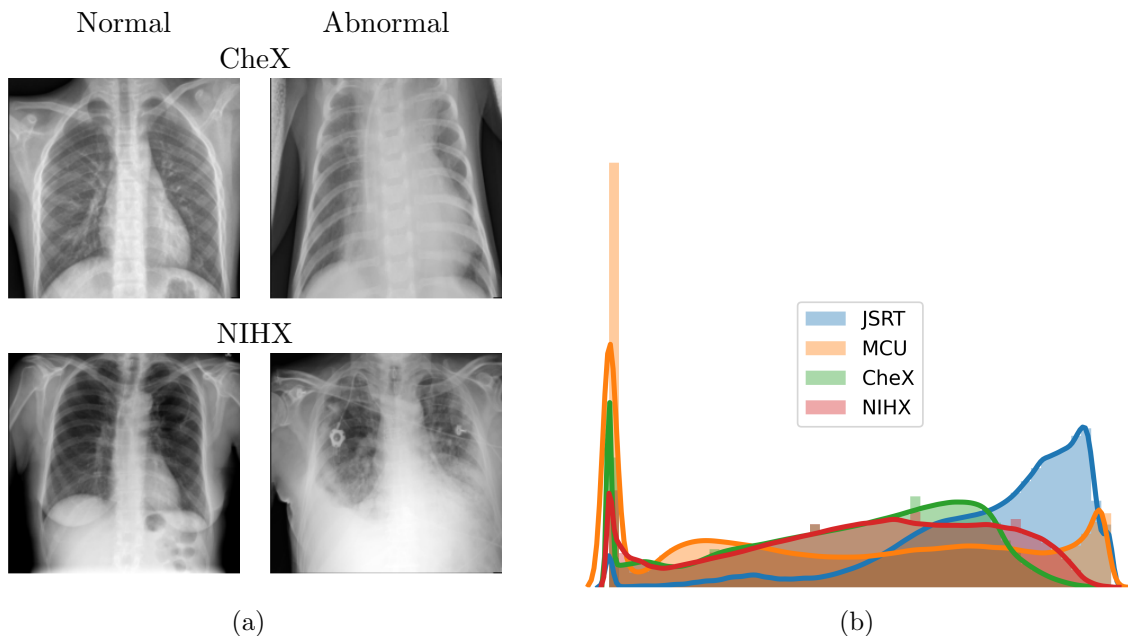
Figure 2: (a) Sample (normal, abnormal) images from the CheX and NIHX datasets. (b) Intensity distributions of the four chest X-ray image datasets.

Table 1: Details of the datasets used for training and testing.

Mode	Dataset	Total	Normal	Abnormal	Train	Val	Test
in-domain	JSRT	247	–	–	111	13	123
	CheX	5,856	1583	4273	5216	16	624
cross-domain	MCU	138	–	–	93	10	35
	NIHX	4185	2754	1431	–	–	4185

3. Experimental Evaluation

3.1 Data

Models were trained and tested in the combined classification and segmentation tasks using chest X-ray images from two different sources: pneumonia detection (CheX) (Kermany et al., 2018) and the Japanese Society of Radiological Technology (JSRT) (Shiraishi et al., 2000). We further validated the models using the Montgomery County chest X-rays (MCU) (Jaeger et al., 2014) and a subset of the NIH chest X-ray dataset (NIHX) (Wang et al., 2017) (Figure 2a). Table 1 presents some details about the datasets used in our experiments. In addition to the diversity in the source, image quality, size, and proportion of normal and abnormal images, the disparity in the intensity distributions of the four datasets is also evident (Figure 2b). All the images were normalized and resized to $256 \times 256 \times 1$ before passing them to the models.

8

3.2 Implementation Details

Baselines: We used the U-Net and encoder-only (Enc) networks separately for the single-task baseline models in both the fully supervised and semi-supervised schemes. Using the same backbone network, we also trained a multi-tasking U-Net with the classification branch (UMTL). All these models incorporate an INorm, LReLU, and dropout at every convolutional block (see Appendix A). Moreover, we performed ablation experiments to assess the impact of each key piece of our MultiMix model: single-task Enc-SSL (encoder with confidence-based augmentation SSL), single-task Enc-MM (an implementation of MixMatch (Berthelot et al., 2019)), UMTL-S (UMTL with saliency bridge), UMTL-SSL (UMTL with SSL classification), and UMTL-SSL-S (UMTL with saliency bridge and confidence-based augmentation).

Augmentations: We performed random horizontal flip and 32×32 crop in WAug for the examples in $\mathcal{D}_u^c$. On the other hand, GAug was applied through a random combination from the pool of augmentations: random horizontal flip, crop (32×32), autocontrast, brightness, contrast, equalize, identity, posterize, rotate ($30°$), sharpness, shearX, shearY, solarize, translateX (30%), and translateY (30%).

Training: All the models (single-task or multi-task) were trained on varying $|\mathcal{D}_l^s|$ (10, 50, full), and $|\mathcal{D}_l^c|$ (100, 1000, full). Each experiment was repeated 5 times and the average performance is reported. We implemented the models using Python and the PyTorch framework and trained using an Nvidia K80 GPU.

Hyper-parameters: We used the Adam optimizer with adaptive learning rates of 0.1 every 8 epochs and an initial learning rate of 0.0001. A negative slope of 0.2 was applied to Leaky ReLU, and the dropout was set to 0.25. We set $t = 0.7$, $\lambda = 0.25$, $\alpha = 5.0$ (for smaller $|\mathcal{D}_l^s|$) and $\beta = 0.01$. Each model was trained with a mini-batch size of $m = 10$. All model-specific hyperparameters were experimentally tuned. We found that the performance of the model varied only minimally subject to the different choices.

Evaluation: For classification, along with the overall accuracy (Acc), we recorded the class-wise F1 scores (F1-N for normal and F1-P for pneumonia). To evaluate segmentation performance, we used the Dice similarity (DS), Jaccard similarity (JS), structural similarity measure (SSIM), average Hausdorff distance (HD), precision (P), and Recall (R) scores.

3.3 Results and Discussion

As is revealed by the results in Table 2, the performance of our model improves with the inclusion of each of the novel components in the backbone network. For the classification task, our confidence-based augmentation approach for semi-supervised learning yields significantly improved performance compared to the baseline models. Even with the min $|\mathcal{D}_l^c|$ and min $|\mathcal{D}_l^s|$, our MultiMix-100-10 model outperforms the fully-supervised baseline (Enc) in classifying the normal and abnormal chest X-rays. As is confirmed by the Student's t-test, MultiMix exhibits significant improvements over the classification baselines Enc, Enc-SSL, and UMTL ($p < 0.05$).

For the segmentation task, the inclusion of the saliency bridge module yields large improvements over the baseline U-Net and UMTL models. Again, with min $|\mathcal{D}_l^s|$, we observed a 30% performance gain over its counterparts, which proves the effectiveness of our

Table 2: Classification and segmentation performance figures 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. Scores are given as mean ± std.

Model	$|\mathcal{D}_l^c|$	Classification			$|\mathcal{D}_l^s|$	Segmentation					
		Acc	F1-Nor	F1-Abn		DS	JS	SSIM	HD	P	R
U-Net	—	—	—	—	10	0.634 ± 0.017	0.695 ± 0.024	0.810 ± 0.019	2.899 ± 0.272	0.779 ± 0.021	0.865 ± 0.023
	—	—	—	—	50	0.855 ± 0.004	0.854 ± 0.008	0.904 ± 0.003	0.341 ± 0.071	0.918 ± 0.009	0.925 ± 0.009
	—	—	—	—	Full	0.915 ± 0.001	0.906 ± 0.002	0.929 ± 0.002	0.104 ± 0.025	0.949 ± 0.007	0.953 ± 0.005
Enc	100	0.732 ± 0.044	0.424 ± 0.122	0.806 ± 0.026	—	—	—	—	—	—	—
	1000	0.773 ± 0.037	0.546 ± 0.020	0.842 ± 0.018	—	—	—	—	—	—	—
	Full	0.737 ± 0.021	0.534 ± 0.058	0.838 ± 0.012	—	—	—	—	—	—	—
Enc-MM	100	0.738 ± 0.043	0.452 ± 0.103	0.800 ± 0.024	—	—	—	—	—	—	—
	1000	0.745 ± 0.036	0.560 ± 0.078	0.584 ± 0.101	—	—	—	—	—	—	—
	Full	0.751 ± 0.036	0.605 ± 0.065	0.846 ± 0.025	—	—	—	—	—	—	—
Enc-SSL	100	0.780 ± 0.035	0.570 ± 0.083	0.844 ± 0.021	—	—	—	—	—	—	—
	1000	0.822 ± 0.027	0.692 ± 0.058	0.876 ± 0.015	—	—	—	—	—	—	—
	Full	0.817 ± 0.016	0.680 ± 0.042	0.872 ± 0.013	—	—	—	—	—	—	—
UMTL	100	0.707 ± 0.024	0.443 ± 0.071	0.797 ± 0.016	10	0.626 ± 0.008	0.871 ± 0.014	0.908 ± 0.010	4.323 ± 1.268	0.900 ± 0.014	0.964 ± 0.007
	100	0.655 ± 0.052	0.683 ± 0.129	0.853 ± 0.030	50	0.647 ± 0.022	0.854 ± 0.040	0.881 ± 0.025	4.733 ± 1.238	0.864 ± 0.043	0.989 ± 0.004
	100	0.706 ± 0.046	0.416 ± 0.138	0.804 ± 0.028	Full	0.696 ± 0.014	0.872 ± 0.0252	0.911 ± 0.016	3.908 ± 0.795	0.892 ± 0.025	0.986 ± 0.004
	1000	0.750 ± 0.010	0.490 ± 0.020	0.825 ± 0.005	10	0.761 ± 0.004	0.904 ± 0.009	0.926 ± 0.002	3.050 ± 0.531	0.924 ± 0.001	0.977 ± 0.009
	1000	0.749 ± 0.024	0.510 ± 0.064	0.833 ± 0.009	50	0.768 ± 0.001	0.927 ± 0.003	0.938 ± 0.002	2.606 ± 0.205	0.940 ± 0.004	0.985 ± 0.001
	1000	0.747 ± 0.045	0.530 ± 0.140	0.840 ± 0.024	Full	0.759 ± 0.005	0.928 ± 0.010	0.930 ± 0.008	2.955 ± 0.483	0.924 ± 0.015	0.981 ± 0.006
	Full	0.744 ± 0.011	0.515 ± 0.022	0.828 ± 0.016	10	0.909 ± 0.021	0.919 ± 0.037	0.521 ± 0.028	0.903 ± 0.296	0.912 ± 0.050	0.962 ± 0.013
	Full	0.738 ± 0.004	0.438 ± 0.013	0.820 ± 0.000	50	0.930 ± 0.000	0.948 ± 0.000	0.954 ± 0.000	0.444 ± 0.142	0.969 ± 0.001	0.977 ± 0.001
	Full	0.731 ± 0.018	0.447 ± 0.067	0.822 ± 0.009	Full	0.932 ± 0.000	0.951 ± 0.000	0.957 ± 0.000	$\underline{0.372} \pm 0.052$	0.965 ± 0.001	0.977 ± 0.001
UMTL-S	100	0.704 ± 0.052	0.358 ± 0.223	0.806 ± 0.024	10	0.922 ± 0.007	0.848 ± 0.013	0.891 ± 0.009	4.005 ± 0.413	0.871 ± 0.017	0.966 ± 0.005
	100	0.701 ± 0.033	0.336 ± 0.130	0.796 ± 0.019	50	0.926 ± 0.002	0.867 ± 0.003	0.894 ± 0.003	4.393 ± 0.217	0.873 ± 0.005	0.891 ± 0.002
	100	0.713 ± 0.041	0.442 ± 0.164	0.794 ± 0.025	Full	0.931 ± 0.003	0.890 ± 0.006	0.920 ± 0.003	3.983 ± 0.375	0.906 ± 0.008	0.980 ± 0.004
	1000	0.740 ± 0.020	0.482 ± 0.052	0.828 ± 0.012	10	0.948 ± 0.001	0.908 ± 0.003	0.924 ± 0.003	2.546 ± 0.217	0.931 ± 0.005	0.972 ± 0.002
	1000	0.771 ± 0.041	0.566 ± 0.010	0.844 ± 0.024	50	0.965 ± 0.001	0.931 ± 0.003	0.941 ± 0.001	2.083 ± 0.217	0.949 ± 0.005	0.981 ± 0.002
	1000	0.742 ± 0.019	0.497 ± 0.059	0.830 ± 0.014	Full	0.962 ± 0.005	0.925 ± 0.010	0.935 ± 0.008	1.758 ± 0.132	0.958 ± 0.015	0.985 ± 0.005
	Full	0.747 ± 0.006	0.500 ± 0.021	0.830 ± 0.006	10	0.955 ± 0.020	0.914 ± 0.035	0.936 ± 0.027	0.568 ± 0.136	0.954 ± 0.037	0.956 ± 0.005
	Full	0.737 ± 0.016	0.433 ± 0.054	0.820 ± 0.008	50	0.972 ± 0.006	0.944 ± 0.011	0.953 ± 0.009	0.560 ± 0.427	0.966 ± 0.014	0.977 ± 0.004
	Full	0.723 ± 0.005	0.413 ± 0.019	0.817 ± 0.005	Full	0.974 ± 0.000	0.953 ± 0.000	0.957 ± 0.000	0.539 ± 0.437	0.967 ± 0.002	0.981 ± 0.001
UMTL-SSL	100	0.790 ± 0.043	0.618 ± 0.105	0.856 ± 0.024	10	0.906 ± 0.002	0.925 ± 0.004	0.940 ± 0.002	0.626 ± 0.280	0.954 ± 0.006	0.953 ± 0.003
	100	0.818 ± 0.039	0.688 ± 0.087	0.872 ± 0.024	50	0.919 ± 0.001	0.946 ± 0.001	0.952 ± 0.001	0.561 ± 0.115	0.962 ± 0.003	0.963 ± 0.002
	100	0.852 ± 0.039	0.670 ± 0.095	0.868 ± 0.022	Full	0.937 ± 0.001	0.954 ± 0.001	0.958 ± 0.001	0.613 ± 0.386	0.969 ± 0.004	0.981 ± 0.002
	1000	0.794 ± 0.020	0.630 ± 0.046	0.860 ± 0.012	10	0.893 ± 0.000	0.926 ± 0.001	0.941 ± 0.001	0.524 ± 0.107	0.961 ± 0.002	0.962 ± 0.001
	1000	0.822 ± 0.038	0.693 ± 0.096	0.877 ± 0.026	50	0.903 ± 0.000	0.945 ± 0.000	0.952 ± 0.001	0.712 ± 0.167	0.963 ± 0.002	0.980 ± 0.002
	1000	0.818 ± 0.005	0.707 ± 0.019	0.867 ± 0.005	Full	0.899 ± 0.001	0.953 ± 0.001	0.958 ± 0.001	0.724 ± 0.400	0.968 ± 0.005	0.982 ± 0.003
	Full	0.812 ± 0.022	0.688 ± 0.050	0.870 ± 0.012	10	0.905 ± 0.005	0.921 ± 0.008	0.935 ± 0.004	0.627 ± 0.150	0.946 ± 0.014	0.973 ± 0.009
	Full	0.813 ± 0.012	0.683 ± 0.020	0.873 ± 0.008	50	0.927 ± 0.001	0.947 ± 0.001	0.954 ± 0.001	$\mathbf{0.397} \pm 0.172$	0.968 ± 0.001	0.977 ± 0.001
	Full	0.816 ± 0.008	0.678 ± 0.019	0.873 ± 0.004	Full	0.935 ± 0.001	$\underline{0.954} \pm 0.001$	0.958 ± 0.001	0.625 ± 0.208	$\underline{0.970} \pm 0.002$	0.981 ± 0.001
UMTL-SSL-S	100	0.798 ± 0.030	0.628 ± 0.081	0.860 ± 0.018	10	0.951 ± 0.004	0.911 ± 0.008	0.935 ± 0.004	0.792 ± 0.313	0.940 ± 0.006	0.963 ± 0.006
	100	0.834 ± 0.033	0.696 ± 0.074	0.874 ± 0.019	50	0.972 ± 0.001	0.946 ± 0.001	0.952 ± 0.001	0.727 ± 0.340	0.965 ± 0.002	0.977 ± 0.002
	100	0.817 ± 0.036	0.688 ± 0.100	0.860 ± 0.021	Full	0.975 ± 0.001	0.951 ± 0.001	0.954 ± 0.001	0.812 ± 0.315	0.968 ± 0.002	0.981 ± 0.002
	1000	0.806 ± 0.020	0.652 ± 0.055	0.872 ± 0.014	10	0.956 ± 0.002	0.916 ± 0.004	0.937 ± 0.003	0.852 ± 0.275	0.943 ± 0.005	0.966 ± 0.003
	1000	$0.808 \pm .013$	0.662 ± 0.038	0.862 ± 0.010	50	0.971 ± 0.000	0.944 ± 0.001	0.952 ± 0.000	0.917 ± 0.239	0.965 ± 0.002	0.978 ± 0.003
	1000	0.801 ± 0.020	0.646 ± 0.049	0.862 ± 0.010	Full	0.975 ± 0.001	0.952 ± 0.001	0.954 ± 0.001	0.753 ± 0.228	0.969 ± 0.001	0.981 ± 0.001
	Full	0.796 ± 0.033	0.632 ± 0.086	0.864 ± 0.018	10	0.960 ± 0.002	0.923 ± 0.004	0.940 ± 0.002	0.782 ± 0.229	0.954 ± 0.005	0.967 ± 0.003
	Full	0.808 ± 0.014	0.662 ± 0.030	0.868 ± 0.007	50	0.972 ± 0.001	0.945 ± 0.001	0.953 ± 0.001	0.645 ± 0.196	0.966 ± 0.003	0.978 ± 0.003
	Full	0.800 ± 0.016	0.632 ± 0.038	0.628 ± 0.009	Full	0.961 ± 0.008	0.924 ± 0.016	0.940 ± 0.009	0.392 ± 0.337	0.948 ± 0.014	0.969 ± 0.007
MultiMix	100	0.800 ± 0.025	0.594 ± 0.064	0.856 ± 0.015	10	0.954 ± 0.004	0.920 ± 0.008	0.938 ± 0.004	0.695 ± 0.198	0.949 ± 0.010	0.969 ± 0.007
	100	0.824 ± 0.022	0.613 ± 0.056	0.854 ± 0.014	50	0.971 ± 0.001	0.943 ± 0.002	0.951 ± 0.001	0.681 ± 0.086	0.964 ± 0.003	0.976 ± 0.002
	100	0.792 ± 0.035	0.593 ± 0.101	0.854 ± 0.021	Full	0.973 ± 0.012	0.948 ± 0.022	0.954 ± 0.015	0.636 ± 0.070	0.966 ± 0.025	0.981 ± 0.004
	1000	0.817 ± 0.016	0.647 ± 0.038	0.865 ± 0.006	10	0.954 ± 0.004	0.910 ± 0.008	0.932 ± 0.004	0.902 ± 0.186	0.942 ± 0.005	0.968 ± 0.007
	1000	0.825 ± 0.016	0.650 ± 0.033	0.860 ± 0.011	50	0.970 ± 0.001	0.941 ± 0.002	0.950 ± 0.001	0.811 ± 0.112	0.964 ± 0.004	0.977 ± 0.002
	1000	0.830 ± 0.048	0.586 ± 0.138	0.856 ± 0.029	Full	$\mathbf{0.974} \pm 0.011$	0.919 ± 0.020	0.953 ± 0.014	0.643 ± 0.126	0.933 ± 0.024	0.984 ± 0.004
	Full	0.840 ± 0.025	0.730 ± 0.060	0.880 ± 0.016	10	0.954 ± 0.002	0.913 ± 0.004	0.935 ± 0.001	0.621 ± 0.123	0.949 ± 0.006	0.968 ± 0.006
	Full	$\mathbf{0.854} \pm 0.022$	$\mathbf{0.760} \pm 0.055$	$\mathbf{0.890} \pm 0.014$	50	0.972 ± 0.001	$\mathbf{0.950} \pm 0.001$	$\mathbf{0.956} \pm 0.001$	0.692 ± 0.036	$\mathbf{0.970} \pm 0.003$	$\mathbf{0.980} \pm 0.003$
	Full	$\underline{0.843} \pm 0.024$	$\underline{0.740} \pm 0.065$	$\underline{0.890} \pm 0.017$	Full	$\underline{0.975} \pm 0.000$	0.952 ± 0.001	$\underline{0.960} \pm 0.001$	0.528 ± 0.037	$\underline{0.970} \pm 0.001$	$\underline{0.982} \pm 0.001$

Table 3: Classification and segmentation performance figures with varying label proportions in cross-domain evaluations: NIHX (classification) and MCU (segmentation) datasets. The best scores from fully-supervised models are underlined and the best scores from semi-supervised models are bolded. Scores are given as mean ± std.

Model	$|\mathcal{D}_l^c|$	Classification			$|\mathcal{D}_l^s|$	Segmentation					
		Acc	F1-Nor	F1-Abn		DS	JS	SSIM	HD	P	R
U-Net	—	—	—	—	10	0.555 ± 0.047	0.480 ± 0.053	0.680 ± 0.059	8.691 ± 1.100	0.553 ± 0.070	0.866 ± 0.032
	—	—	—	—	50	0.763 ± 0.026	0.736 ± 0.037	0.870 ± 0.019	2.895 ± 0.832	0.752 ± 0.035	0.887 ± 0.019
	—	—	—	—	Full	0.838 ± 0.023	0.906 ± 0.035	0.929 ± 0.017	1.414 ± 0.529	0.793 ± 0.041	0.910 ± 0.013
Enc	100	0.352 ± 0.035	0.070 ± 0.131	0.506 ± 0.008	—	—	—	—	—	—	—
	1000	0.390 ± 0.037	0.192 ± 0.124	0.508 ± 0.007	—	—	—	—	—	—	—
	Full	0.434 ± 0.026	0.296 ± 0.068	0.524 ± 0.005	—	—	—	—	—	—	—
Enc-MM	100	0.360 ± 0.030	0.110 ± 0.065	0.500 ± 0.024	—	—	—	—	—	—	—
	1000	0.406 ± 0.078	0.242 ± 0.114	0.460 ± 0.010	—	—	—	—	—	—	—
	Full	0.452 ± 0.040	0.316 ± 0.068	0.502 ± 0.030	—	—	—	—	—	—	—
Enc-SSL	100	0.402 ± 0.052	0.222 ± 0.136	0.510 ± 0.012	—	—	—	—	—	—	—
	1000	0.486 ± 0.050	0.380 ± 0.109	0.530 ± 0.010	—	—	—	—	—	—	—
	Full	0.510 ± 0.024	0.472 ± 0.056	0.538 ± 0.004	—	—	—	—	—	—	—
UMTL	100	0.350 ± 0.034	0.045 ± 0.123	0.510 ± 0.003	10	0.586 ± 0.023	0.708 ± 0.035	0.836 ± 0.020	7.156 ± 2.316	0.731 ± 0.032	0.950 ± 0.014
	100	0.363 ± 0.029	0.085 ± 0.096	0.515 ± 0.005	50	0.580 ± 0.040	0.684 ± 0.061	0.825 ± 0.038	7.013 ± 1.576	0.697 ± 0.065	0.975 ± 0.012
	100	0.342 ± 0.049	0.015 ± 0.153	0.508 ± 0.006	Full	0.607 ± 0.023	0.742 ± 0.037	0.863 ± 0.021	6.398 ± 1.331	0.759 ± 0.041	0.968 ± 0.007
	1000	0.413 ± 0.004	0.263 ± 0.033	0.507 ± 0.005	10	0.676 ± 0.020	0.674 ± 0.030	0.833 ± 0.016	3.268 ± 0.768	0.712 ± 0.025	0.927 ± 0.013
	1000	0.400 ± 0.022	0.203 ± 0.069	0.513 ± 0.005	50	0.704 ± 0.025	0.811 ± 0.040	0.896 ± 0.015	3.232 ± 0.229	0.828 ± 0.034	0.964 ± 0.014
	1000	0.430 ± 0.041	0.293 ± 0.110	0.517 ± 0.005	Full	0.638 ± 0.015	0.795 ± 0.024	0.890 ± 0.013	3.893 ± 0.465	0.810 ± 0.031	0.966 ± 0.010
	Full	0.455 ± 0.015	0.365 ± 0.038	0.525 ± 0.004	10	0.737 ± 0.037	0.765 ± 0.056	0.879 ± 0.028	0.917 ± 0.478	0.801 ± 0.059	0.930 ± 0.013
	Full	0.444 ± 0.026	0.332 ± 0.075	0.522 ± 0.007	50	0.868 ± 0.022	0.793 ± 0.038	0.894 ± 0.014	0.742 ± 0.212	0.898 ± 0.0378	0.946 ± 0.004
	Full	0.443 ± 0.038	0.328 ± 0.099	0.520 ± 0.010	Full	0.854 ± 0.018	0.828 ± 0.030	0.913 ± 0.012	0.792 ± 0.490	0.866 ± 0.026	0.942 ± 0.008
UMTL-S	100	0.344 ± 0.021	0.006 ± 0.074	0.510 ± 0.005	10	0.797 ± 0.029	0.670 ± 0.042	0.807 ± 0.027	5.754 ± 1.047	0.698 ± 0.043	0.938 ± 0.009
	100	0.364 ± 0.035	0.098 ± 0.019	0.506 ± 0.006	50	0.828 ± 0.032	0.715 ± 0.049	0.826 ± 0.033	6.412 ± 1.753	0.731 ± 0.053	0.971 ± 0.009
	100	0.342 ± 0.016	0.008 ± 0.070	0.510 ± 0.000	Full	0.838 ± 0.041	0.715 ± 0.063	0.834 ± 0.043	6.321 ± 1.573	0.740 ± 0.068	0.966 ± 0.009
	1000	0.378 ± 0.017	0.138 ± 0.057	0.512 ± 0.007	10	0.844 ± 0.018	0.718 ± 0.027	0.854 ± 0.014	3.921 ± 0.480	0.754 ± 0.023	0.939 ± 0.011
	1000	0.392 ± 0.024	0.186 ± 0.078	0.514 ± 0.005	50	0.883 ± 0.016	0.793 ± 0.025	0.888 ± 0.011	3.017 ± 0.191	0.821 ± 0.027	0.959 ± 0.002
	1000	0.370 ± 0.014	0.130 ± 0.057	0.510 ± 0.000	Full	0.898 ± 0.016	0.831 ± 0.027	0.905 ± 0.014	4.150 ± 1.269	0.845 ± 0.031	0.970 ± 0.010
	Full	0.470 ± 0.026	0.398 ± 0.015	0.524 ± 0.010	10	0.881 ± 0.019	0.785 ± 0.031	0.888 ± 0.014	0.862 ± 0.199	0.830 ± 0.034	0.939 ± 0.009
	Full	0.413 ± 0.009	0.270 ± 0.008	0.510 ± 0.014	50	0.917 ± 0.009	0.848 ± 0.014	0.919 ± 0.006	0.658 ± 0.227	0.966 ± 0.012	0.888 ± 0.007
	Full	0.433 ± 0.026	0.315 ± 0.007	0.513 ± 0.015	Full	0.916 ± 0.008	0.850 ± 0.013	0.921 ± 0.005	0.882 ± 0.151	0.886 ± 0.014	0.952 ± 0.002
UMTL-SSL	100	0.442 ± 0.045	0.316 ± 0.004	0.524 ± 0.007	10	0.833 ± 0.025	0.778 ± 0.042	0.884 ± 0.019	0.895 ± 0.415	0.810 ± 0.042	0.948 ± 0.007
	100	0.398 ± 0.010	0.166 ± 0.004	0.520 ± 0.019	50	0.853 ± 0.023	0.839 ± 0.038	0.907 ± 0.017	0.851 ± 0.157	0.864 ± 0.039	0.952 ± 0.007
	100	0.385 ± 0.062	0.165 ± 0.006	0.515 ± 0.013	Full	0.841 ± 0.023	0.818 ± 0.039	0.911 ± 0.015	0.853 ± 0.410	0.854 ± 0.039	0.949 ± 0.007
	1000	0.445 ± 0.038	0.333 ± 0.091	0.525 ± 0.005	10	0.818 ± 0.032	0.781 ± 0.051	0.892 ± 0.023	1.085 ± 0.464	0.825 ± 0.051	0.938 ± 0.008
	1000	0.526 ± 0.080	0.486 ± 0.017	0.544 ± 0.016	50	0.826 ± 0.018	0.804 ± 0.029	0.904 ± 0.014	0.811 ± 0.137	0.792 ± 0.031	0.949 ± 0.006
	1000	0.485 ± 0.043	0.413 ± 0.097	0.538 ± 0.008	Full	0.843 ± 0.010	0.837 ± 0.018	0.924 ± 0.007	0.983 ± 0.429	0.882 ± 0.016	0.953 ± 0.006
	Full	0.526 ± 0.023	0.504 ± 0.036	0.546 ± 0.012	10	0.824 ± 0.020	0.765 ± 0.0312	0.873 ± 0.017	0.994 ± 0.379	0.790 ± 0.031	0.943 ± 0.010
	Full	0.530 ± 0.030	0.514 ± 0.058	0.542 ± 0.012	50	0.867 ± 0.020	0.839 ± 0.034	0.917 ± 0.013	0.566 ± 0.282	0.881 ± 0.033	0.945 ± 0.005
	Full	$\underline{0.520} \pm 0.031$	$\underline{0.490} \pm 0.061$	0.542 ± 0.007	Full	0.884 ± 0.011	0.884 ± 0.021	0.934 ± 0.008	0.599 ± 0.201	0.918 ± 0.021	0.955 ± 0.003
UMTL-SSL-S	100	0.370 ± 0.046	0.114 ± 0.129	0.510 ± 0.011	10	0.853 ± 0.026	0.747 ± 0.041	0.866 ± 0.019	1.048 ± 0.186	0.782 ± 0.041	0.944 ± 0.010
	100	0.400 ± 0.067	0.192 ± 0.171	0.518 ± 0.012	50	0.889 ± 0.019	0.799 ± 0.031	0.899 ± 0.014	0.854 ± 0.298	0.834 ± 0.031	0.950 ± 0.006
	100	0.370 ± 0.061	0.114 ± 0.159	0.514 ± 0.016	Full	0.915 ± 0.019	0.848 ± 0.032	0.920 ± 0.013	0.987 ± 0.328	0.880 ± 0.033	0.956 ± 0.003
	1000	0.432 ± 0.043	0.286 ± 0.019	0.524 ± 0.010	10	0.871 ± 0.034	0.785 ± 0.054	0.884 ± 0.025	1.327 ± 0.135	0.818 ± 0.053	0.944 ± 0.008
	1000	0.458 ± 0.077	0.342 ± 0.186	0.530 ± 0.013	50	0.893 ± 0.008	0.803 ± 0.014	0.895 ± 0.006	1.123 ± 0.215	0.835 ± 0.015	0.946 ± 0.005
	1000	0.462 ± 0.050	0.350 ± 0.115	0.536 ± 0.014	Full	0.930 ± 0.008	0.860 ± 0.013	0.925 ± 0.006	1.042 ± 0.206	0.912 ± 0.013	0.955 ± 0.003
	Full	0.482 ± 0.042	0.412 ± 0.087	0.536 ± 0.008	10	0.880 ± 0.031	0.765 ± 0.050	0.885 ± 0.022	0.745 ± 0.219	0.818 ± 0.045	0.941 ± 0.014
	Full	0.490 ± 0.021	0.426 ± 0.041	0.540 ± 0.006	50	0.912 ± 0.014	0.845 ± 0.024	0.909 ± 0.009	0.956 ± 0.215	0.881 ± 0.025	0.952 ± 0.002
	Full	0.510 ± 0.038	0.474 ± 0.071	0.540 ± 0.011	Full	0.875 ± 0.032	0.809 ± 0.053	0.875 ± 0.026	0.722 ± 0.335	0.851 ± 0.057	0.944 ± 0.008
MultiMix	100	0.440 ± 0.058	0.164 ± 0.019	0.510 ± 0.014	10	0.857 ± 0.028	0.732 ± 0.044	0.863 ± 0.018	1.227 ± 0.534	0.767 ± 0.044	0.943 ± 0.016
	100	0.370 ± 0.086	0.036 ± 0.003	0.510 ± 0.013	50	0.889 ± 0.021	0.790 ± 0.036	0.890 ± 0.015	1.061 ± 0.434	0.866 ± 0.035	0.947 ± 0.008
	100	0.500 ± 0.080	0.300 ± 0.002	0.510 ± 0.006	Full	0.899 ± 0.022	0.825 ± 0.036	0.906 ± 0.017	0.647 ± 0.074	0.852 ± 0.040	0.952 ± 0.012
	1000	0.520 ± 0.041	0.386 ± 0.009	0.530 ± 0.014	10	0.862 ± 0.017	0.775 ± 0.026	0.878 ± 0.011	1.307 ± 0.325	0.816 ± 0.029	0.939 ± 0.006
	1000	0.540 ± 0.018	0.500 ± 0.036	0.536 ± 0.005	50	0.912 ± 0.018	0.831 ± 0.031	0.907 ± 0.012	1.293 ± 0.375	0.865 ± 0.030	0.955 ± 0.007
	1000	$\mathbf{0.570} \pm 0.088$	$\mathbf{0.620} \pm 0.003$	0.510 ± 0.008	Full	$\mathbf{0.936} \pm 0.026$	$\mathbf{0.880} \pm 0.043$	$\mathbf{0.932} \pm 0.022$	0.803 ± 0.178	0.917 ± 0.050	$\mathbf{0.979} \pm 0.008$
	Full	0.550 ± 0.038	0.430 ± 0.006	0.534 ± 0.010	10	0.886 ± 0.013	0.802 ± 0.022	0.894 ± 0.011	0.746 ± 0.284	0.839 ± 0.028	0.948 ± 0.007
	Full	0.560 ± 0.040	0.570 ± 0.008	$\mathbf{0.550} \pm 0.007$	50	0.935 ± 0.017	0.878 ± 0.030	0.930 ± 0.012	$\mathbf{0.515} \pm 0.232$	$\mathbf{0.928} \pm 0.033$	0.957 ± 0.005
	Full	$\underline{0.520} \pm 0.022$	$\underline{0.490} \pm 0.064$	$\underline{0.550} \pm 0.008$	Full	$\underline{0.943} \pm 0.009$	$\underline{0.892} \pm 0.015$	$\underline{0.937} \pm 0.006$	$\underline{0.417} \pm 0.181$	$\underline{0.928} \pm 0.016$	$\underline{0.958} \pm 0.002$

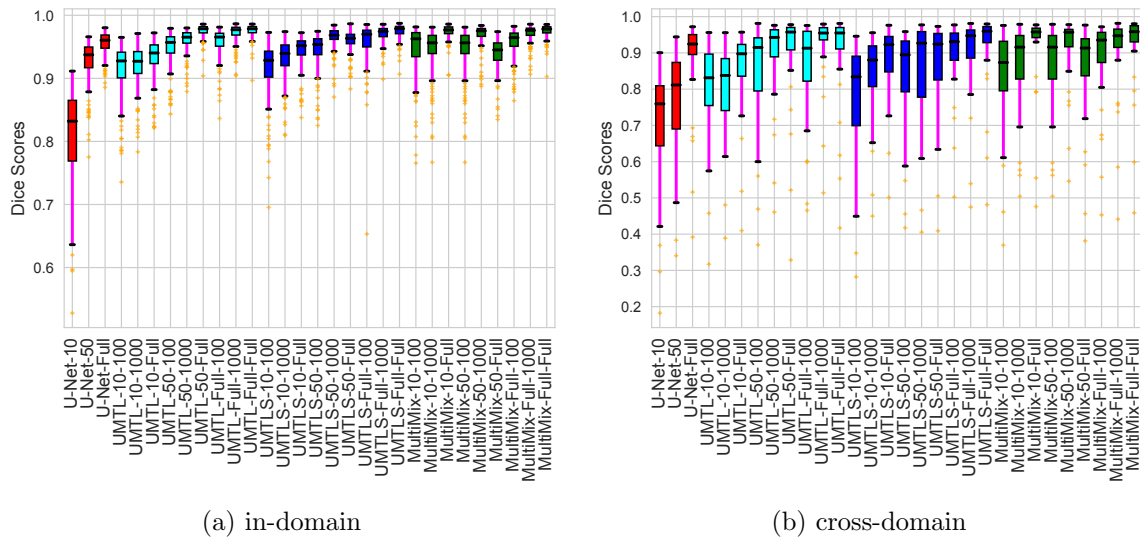(a) in-domain   (b) cross-domain

Figure 3: Distributions of the Dice scores demonstrate the superiority of the MultiMix model over the baseline models in segmenting lungs from the chest X-ray images in both domains.

MultiMix model. The improvement in Dice scores of MultiMix with minimal supervision over the segmentation baselines of U-Net, UMTL, and UMTL-S is statistically significant ($p < 0.05$), confirming the quantitative efficacy of MultiMix. Figure 3 shows improved and consistent segmentation performance by the MultiMix model over the baselines. For a fair comparison, we used the same backbone U-Net and the same classification branch for all the models.

In Figure 4, the segmented lung boundary visualizations also show good agreement with the reference masks by MultiMix over the other models (also see Appendix B). For both the in-domain and cross-domain segmentations, we observe that the predicted boundaries are almost identical with the reference boundaries, as they substantially overlap. Moreover, the noise in the predictions is mitigated with the introduction of each additional component into the intermediate models, which justifies the value of those components in the MultiMix model. The good agreement between the ground truth lung masks and the MultiMix predicted segmentation masks is confirmed by the Bland-Altman plots for varying quantities of labeled data, shown in Figure 5a.

The generalization test through the cross-domain datasets (MCU and NIHX) demonstrates the effectiveness of the MultiMix model. It consistently performs well against both domains with improved generalizability in either task. As reported in Table 3, the performance of MultiMix is as promising as in the in-domain evaluations. MultiMix achieved better scores in the classification task over all the baseline models. Due to the significant differences in the NIHX and CheX datasets, the scores are not as good as the in-domain results, yet our model performs significantly better than the other classification models Enc, Enc-SSL, and UMTL ($p < 0.05$). For the segmentation task, our MultiMix model again achieved better scores in all the various metrics, with improved consistency over the baselines (Figure 3). Just like for the in-domain results, MultiMix shows significant improvements in Dice scores over the segmentation baselines U-Net, UMTL, and UMTL-S ($p < 0.05$), thus proving the

Ground Truth   Ground Truth

$|\mathcal{D}^s| = 10$ $|\mathcal{D}^s| = 50$ $|\mathcal{D}^s| = $ Full $|\mathcal{D}^s| = 10$ $|\mathcal{D}^s| = 50$ $|\mathcal{D}^s| = $ Full

U-Net

UMTL

UMTLS

UMTL-SSL

UMTL-SSL-S

MultiMix

(a) JSRT (in-domain)   (b) MCU (cross-domain)
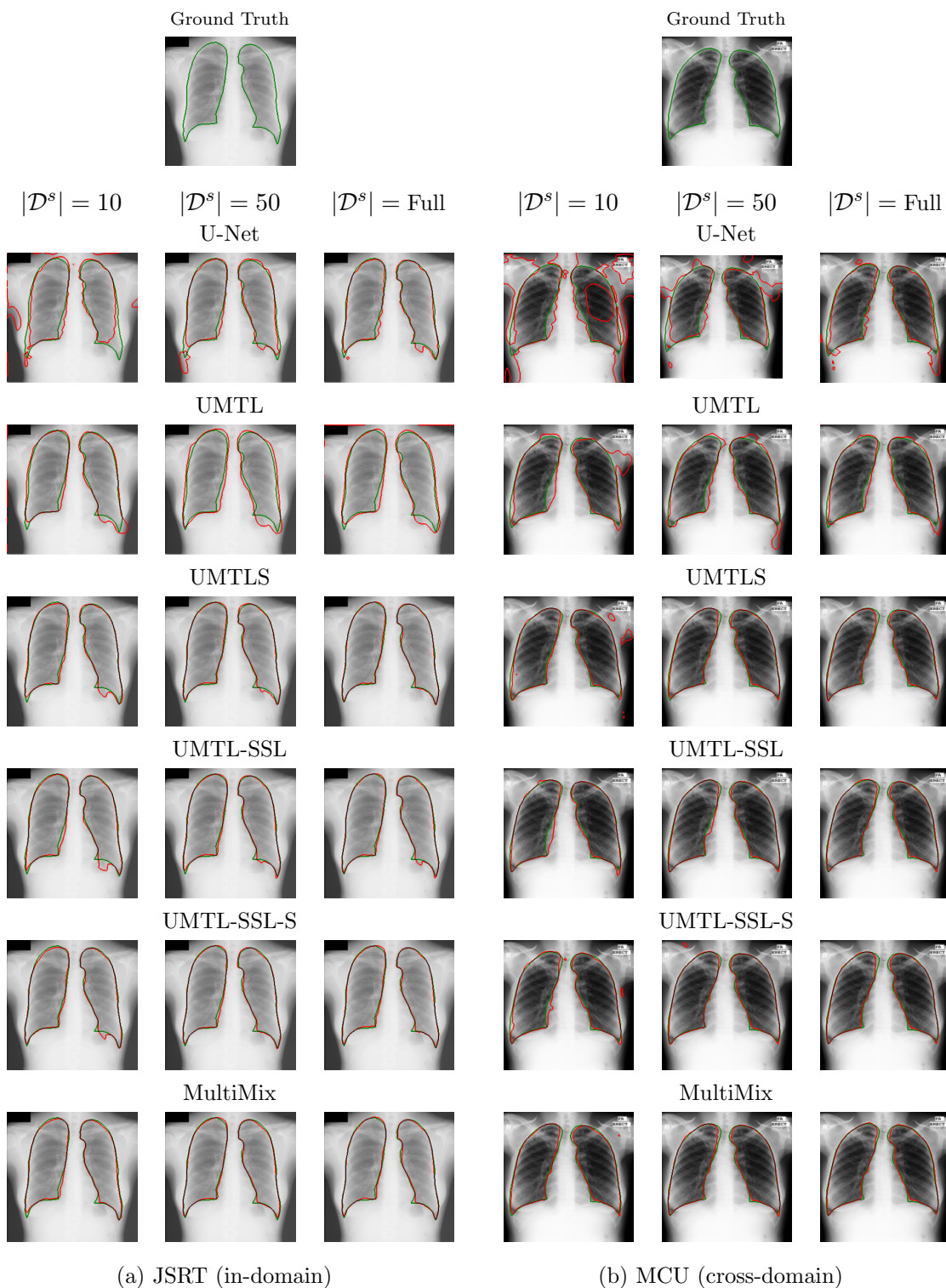
Figure 4: Visualization of the ground truth reference (green) and predicted (red) segmentation boundaries in a chest X-ray reveals the superiority of MultiMix.
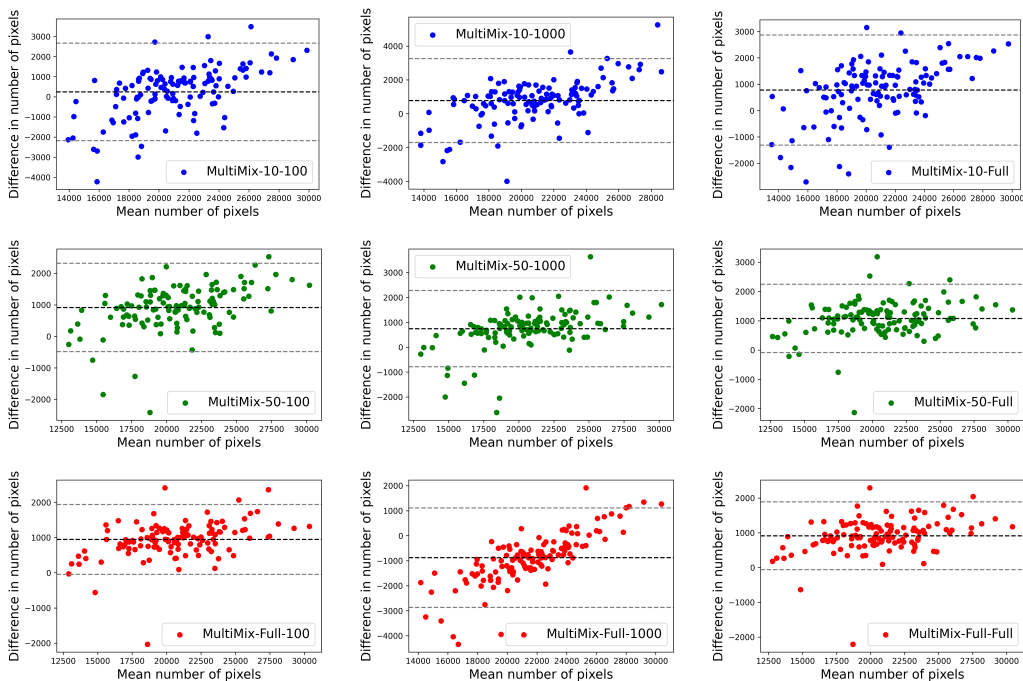
(a) in-domain
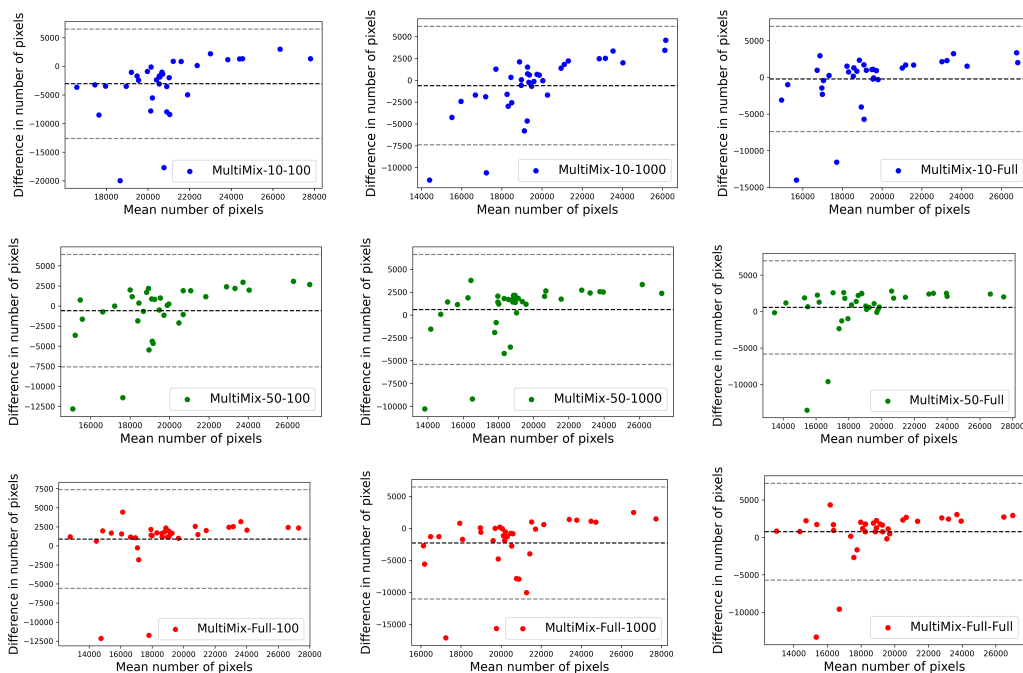
(b) cross-domain

Figure 5: Bland-Altman plots at varying training labels show good agreement between the number of ground truth pixels and MultiMix-predicted pixels for the (a) in-domain and (b) cross-domain evaluations, as well as consistent improvement with increasing quantities of labeled data.

(a) in-domain
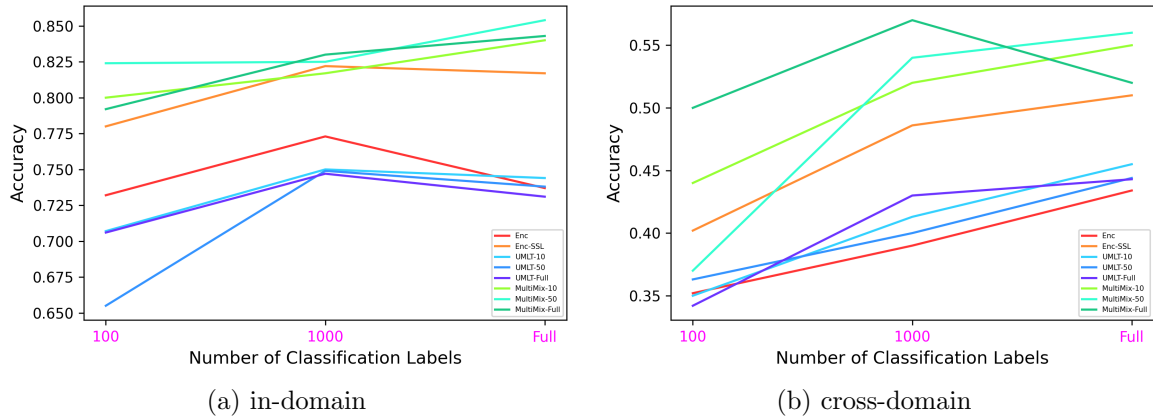
(b) cross-domain

Figure 6: Classification accuracies of different supervised and semi-supervised baselines at different training datasizes. The in-domain and cross-domain plots show that MultiMix has higher accuracy and consistency over the baselines.

(a) in-domain
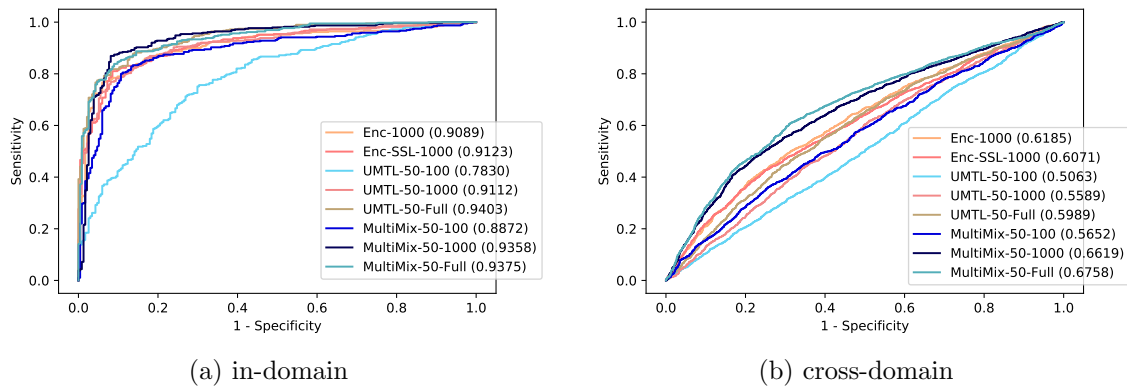
(b) cross-domain

Figure 7: ROC curves for supervised and semi-supervised baselines with 50 segmentation labels show higher AUC values from our MultiMix for in-domain and cross-domain evaluations.

generalizability of our method. The Bland-Altman plots in Figure 5b for the MultiMix model in the cross-domain segmentation evaluations with varying quantities of labeled data confirms the observed good agreement between the ground truth lung segmentation masks and the MultiMix-predicted segmentation masks.

Figure 6 demonstrates the superiority and better consistency of our MultiMix models over the baselines in classifying normal and abnormal (pneumonia) X-rays. Figure 7 further showcases the superior classification performance of MultiMix over the baseline single-task and multi-task models. Figure 6 and Figure 7 together show that MultiMix outperforms all baselines in many different metrics, as the Accuracy and Area Under Curve (AUC) values confirm the superiority of the MultiMix model. With regard to the cross-domain ROC

curves, however, although MultiMix has the best relative performance when compared to the baselines, the absolute performance of the algorithm indicates room for improvement.

4. Conclusions

We have presented MultiMix, a novel semi-supervised, multi-task learning model that jointly learns classification and segmentation tasks. Through the incorporation of confidence-guided data augmentation and a novel saliency bridge module, MultiMix performs improved and consistent pneumonia detection and lung segmentation when trained on multi-source chest X-ray datasets with varying quantities of ground truth labels. Our thorough experimentation using four different chest X-ray datasets demonstrated the effectiveness of MultiMix both in in-domain and cross-domain evaluations, for both tasks; in fact, outperforming a number of baseline models.

Beyond chest X-rays—which is the most frequently performed radiologic procedure worldwide, comprising 40% of all imaging tests, or 1.4 billion annually (World Health Organization, 2016)—our future work will focus on generalizing the MultiMix concept, particularly the saliency bridge module, to other applications and imaging modalities, including volumetric images.

Ethical Standards

Appropriate ethical standards were maintained in writing this manuscript and conducting the reported research, following all applicable laws and regulations regarding the treatment of animals or human subjects.

Conflicts of Interest

DT is a founder of VoxelCloud, Inc., Los Angeles, CA, USA.

Appendix A. Model Architecture

Architectural details of the MultiMix model are presented in Table 4 for the encoder network and in Table 5 for the decoder network. The encoder and decoder incorporate double-convolution blocks; the encoder has 5 blocks and the decoder has 4. Each block includes a 2D convolutional layer, an instance normalization layer, and a Leaky ReLU activation, and this sequence repeats in each block.

In the encoder, each double-convolution block is followed by a dropout layer and a maxpooling layer. The encoder finally branches to a classification branch, which includes a maxpooling layer (5), an average pooling layer, followed by a fully-connected layer for classification prediction.

The decoder begins with an upsampling layer. Next, in the first double-convolution layer, the downsampled saliency maps and original inputs are concatenated. The increase in dimensions at the beginning of each decoder block are due to the skip connections. These convolutional layers are also followed by a dropout layer. This sequence is repeated for 3 more layers. To output the final segmentation prediction, the decoder finishes with a single convolutional layer that downsamples to a single channel.

Appendix B. Segmentation Visualization

Figure 8 shows the ground truth lung masks and masks predicted by the MultiMix model (MultiMix-50-1000) for a number of images from the JSRT dataset (in-domain) and MCU dataset (cross-domain). Both parts of the figure display the accuracy in the predicted segmentation masks, both in-domain and cross-domain, as there is almost no noise in these predictions, proving the effectiveness of our algorithm even when being trained with limited labeled data.

Appendix C. Saliency Visualization

Figure 9 shows the class-specific saliency maps generated by our MultiMix-50-1000 model for both in-domain and cross-domain classification data (X^c). The maps consistently highlight particular regions in the input X-rays for the Normal and Pneumonia classes. Similarly, Figure 10 shows the saliency maps for the in-domain and cross-domain segmentation data (X^s). While the class labels are not available, two distinct types of saliency maps are generated like for the classification data.

Class-specific saliency maps generated for images in X^c consistently highlight regions responsible for predicting the particular classes of the images (Figure 9), enabling the use of these maps to improve the segmentation of images in X^s (Figure 10).

Table 4: Architectural details of the MultiMix Encoder for minibatch size m.

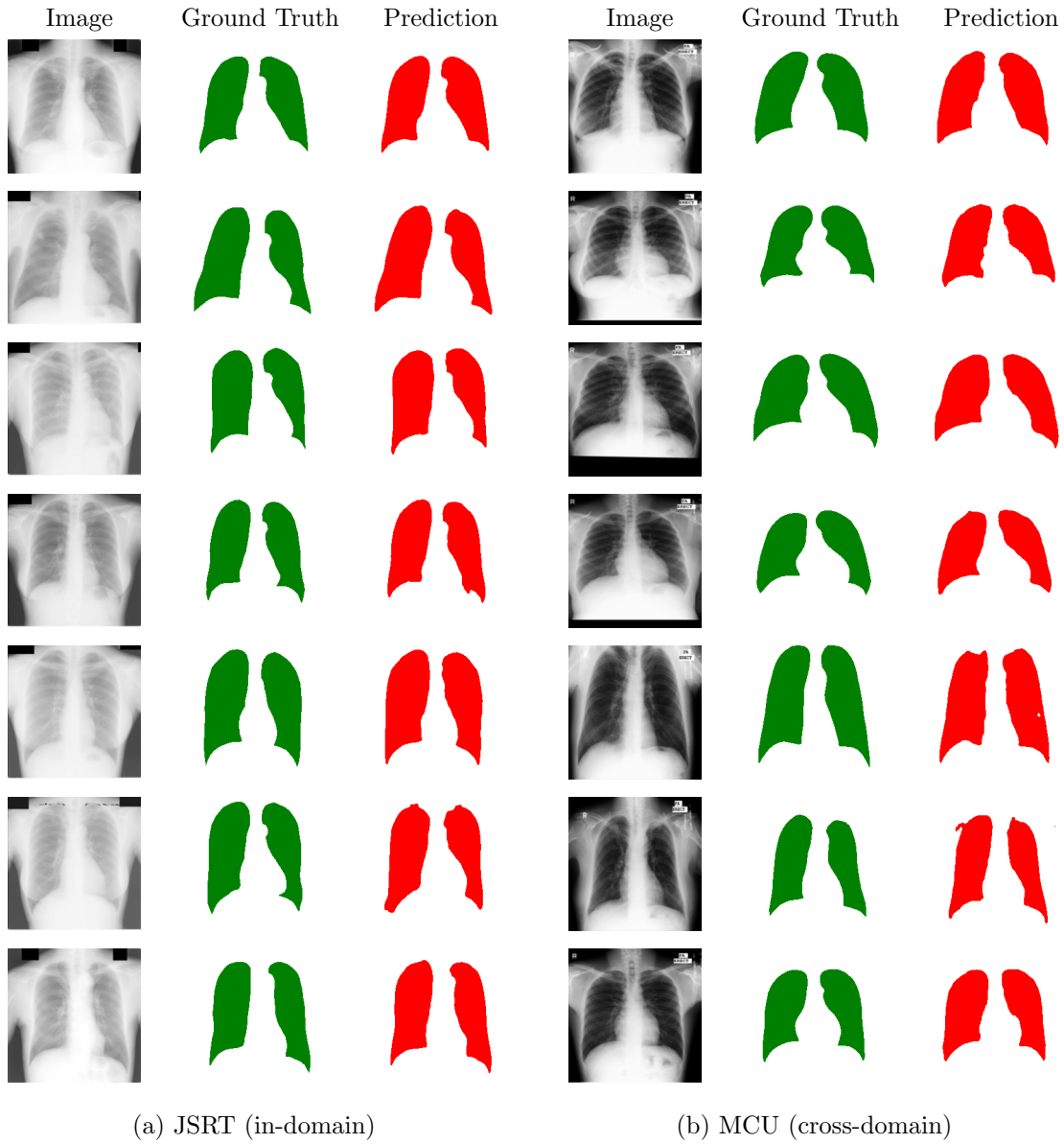| Name	Input Feature Maps	Output Feature Maps
Conv layer - 1	$m \times 256 \times 256 \times 1$	$m \times 256 \times 256 \times 16$
InstanceNorm - 1	$m \times 256 \times 256 \times 16$	$m \times 256 \times 256 \times 16$
LReLU - 1	$m \times 256 \times 256 \times 16$	$m \times 256 \times 256 \times 16$
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 5: Architectural details of the MultiMix Decoder for minibatch size m.

Name	Input Feature Maps	Output Feature Maps
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	$m \times 256 \times 256 \times 16$	$m \times 256 \times 256 \times 1$

| Image | Ground Truth | Prediction | | Image | Ground Truth | Prediction |

(a) JSRT (in-domain) (b) MCU (cross-domain)

Figure 8: Visualizations of the segmented lung masks by MultiMix-50-1000 on the in-domain JSRT dataset and cross-domain MCU dataset. The results show good agreement between the groundtruth and predicted masks.
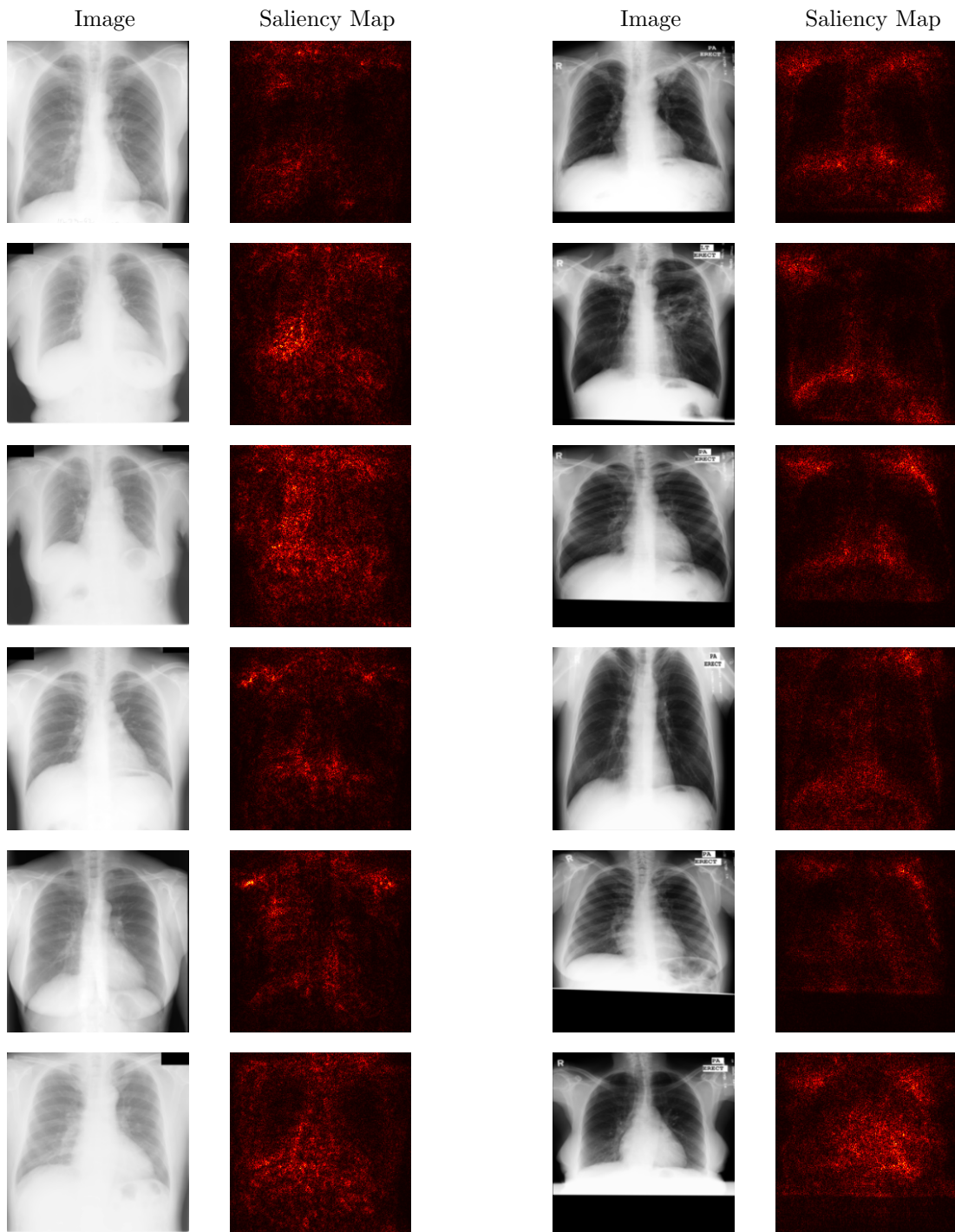
(a) CheX (in-domain) (b) NIHX (cross-domain)
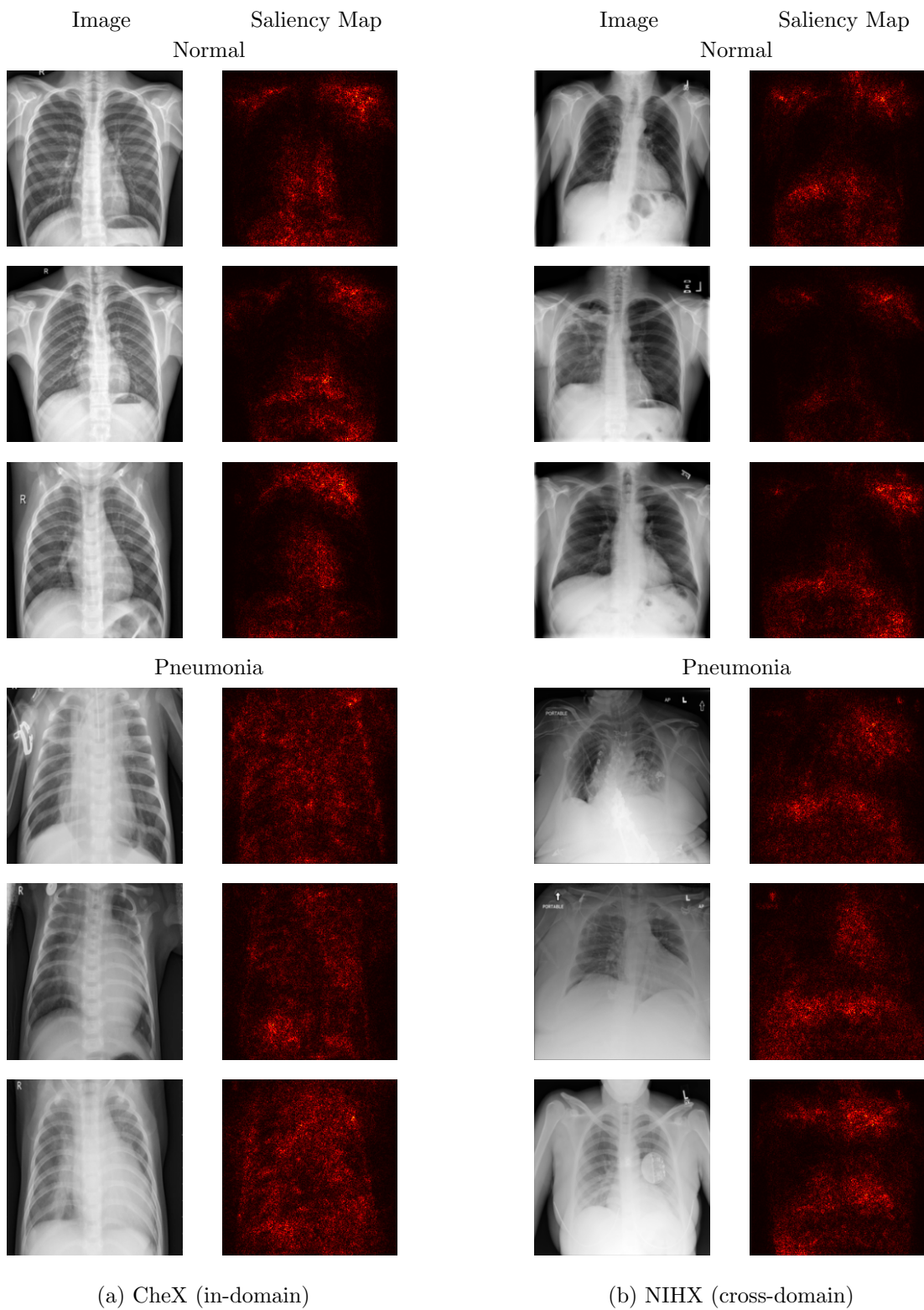
Figure 9: Examples from X^c. Class-specific MultiMix saliency maps highlight crucial regions in the input X-ray images in detecting pneumonia, demonstrating the effective predictions by the classifier and providing useful information for improved segmentation.

(a) JSRT (in-domain)   (b) MCU (cross-domain)

Figure 10: Examples from X^s. MultiMix saliency maps consistently highlight the crucial regions in the input X-ray images, thus providing useful information for improved segmentation.

References

S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan. Medical image analysis using convolutional neural networks: A review. *Journal of Medical Systems*, 42(11):226, 2018.

O. Beijbom. Domain adaptations for computer vision applications. *arXiv Preprint arXiv:1211.4860*, 2012.

D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv Preprint arXiv:1905.02249*, 2019.

R. A. Caruana. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the International Conference on Machine Learning*, pages 41–48, 1993.

O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

F. Gao, H. Yoon, T. Wu, and X. Chu. A feature transfer enabled multi-task deep learning model on medical imaging. *arXiv Preprint arXiv:1906.01828*, 2019.

F. Girard, C. Kavalec, and F. Cheriet. Joint segmentation and classification of retinal arteries/veins from fundus images. *Artificial Intelligence in Medicine*, 94:96–109, Mar 2019.

Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, pages 529–536, 2005.

Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu. CE-Net: Context encoder network for 2D medical image segmentation. *IEEE Transactions on Medical Imaging*, 38(10):2281–2292, 2019.

A. Haque, A.-A.-Z. Imran, A. Wang, and D. Terzopoulos. Multimix: Sparingly supervised, extreme multitask learning from medical images. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 693–696, Nice, France, April 2021.

J. Hu, H. Wang, S. Gao, M. Bao, T. Liu, Y. Wang, and J. Zhang. S-unet: A bridge-style U-Net framework with a saliency mechanism for retinal vessel segmentation. *IEEE Access*, 7:174167–174177, 2019. doi: 10.1109/ACCESS.2019.2940476.

A.-A.-Z. Imran. *From Fully-Supervised, Single-Task to Scarcely-Supervised, Multi-Task Deep Learning for Medical Image Analysis*. PhD thesis, Computer Science Department, University of California, Los Angeles, 2020.

A.-A.-Z. Imran and D. Terzopoulos. Semi-supervised multi-task learning with chest X-ray images. In *Proceedings of the 10th International Workshop on Machine Learning in Medical Imaging (MLMI 2019)*, volume 11861 of *Lecture Notes in Computer Science*, pages 98–105. Springer Nature, 2019.

A.-A.-Z. Imran and D. Terzopoulos. Progressive adversarial semantic segmentation. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, pages 4910–4917, Milan, Italy, January 2021a.

A.-A.-Z. Imran and D. Terzopoulos. Multi-adversarial variational autoencoder nets for simultaneous image generation and classification. In *Deep Learning Applications, Volume 2*, pages 249–271. Springer, 2021b.

A.-A.-Z. Imran, C. Huang, H. Tang, W. Fan, Y. Xiao, D. Hao, Z. Qian, and D. Terzopoulos. Partly supervised multi-task learning. In *Proceedings of the 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 769–774, Miami, FL, December 2020.

S. Jaeger, S. Candemir, et al. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 2014.

D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, volume 3, pages 1–6, 2013.

Q. Liu, X. Liao, and L. Carin. Semi-supervised multitask learning. In *Advances in Neural Information Processing Systems*, pages 937–944, 2008.

S. Mehta, E. Mercan, J. Bartlett, D. Weave, J. G. Elmore, and Shapiro L. Y-Net: Joint segmentation and classification for diagnosis of breast biopsy images. *arXiv Preprint arXiv:1806.01313*, 2018.

A. Myronenko. 3D MRI brain tumor segmentation using autoencoder regularization. *arXiv Preprint arXiv:1810.11654*, 2018.

Y. Ouali, C. Hudelot, and M. Tami. An overview of deep semi-supervised learning. *arXiv Preprint arXiv:2006.05278*, 2020.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. *arXiv Preprint arXiv:1505.04597*, 2015.

S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv Preprint arXiv:1706.05098*, 2017.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. *arXiv Preprint arXiv:1606.03498*, 2016.

J. Shiraishi, S. Katsuragawa, et al. Development of a digital image database for chest radiographs with and without a lung nodule. *American Journal of Roentgenology*, 174: 71–74, 2000.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv Preprint arXiv:1312.6034*, 2014.

K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv Preprint arXiv:2001.07685*, 2020.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.

World Health Organization. Communicating radiation risks in paediatric imaging, 2016. URL `https://www.who.int/publications/i/item/978924151034`.

Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7223–7233, 2019.

D. Zhang, D. Meng, L. Zhao, and J. Han. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

Y. Zhou, X. He, L. Huang, L. Liu, F. Zhu, S. Cui, and L. Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.