USC University of Southern California

[CSCI 461] AI for Sustainable Development

Robust Domain-Generalizable AI for Chest X-Ray Disease Classification

Group **AI4SD**

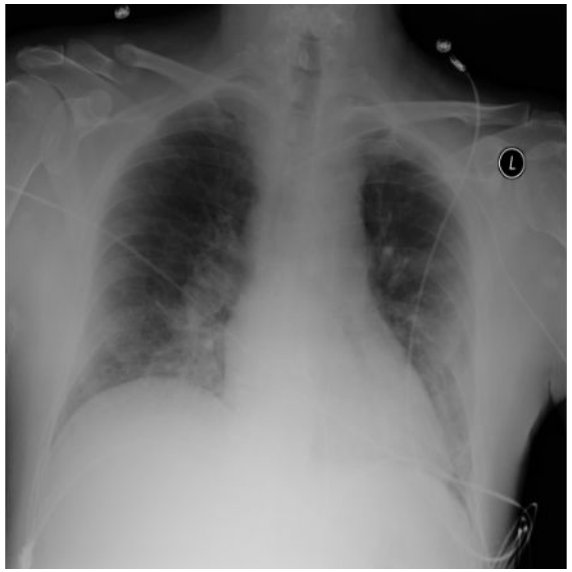
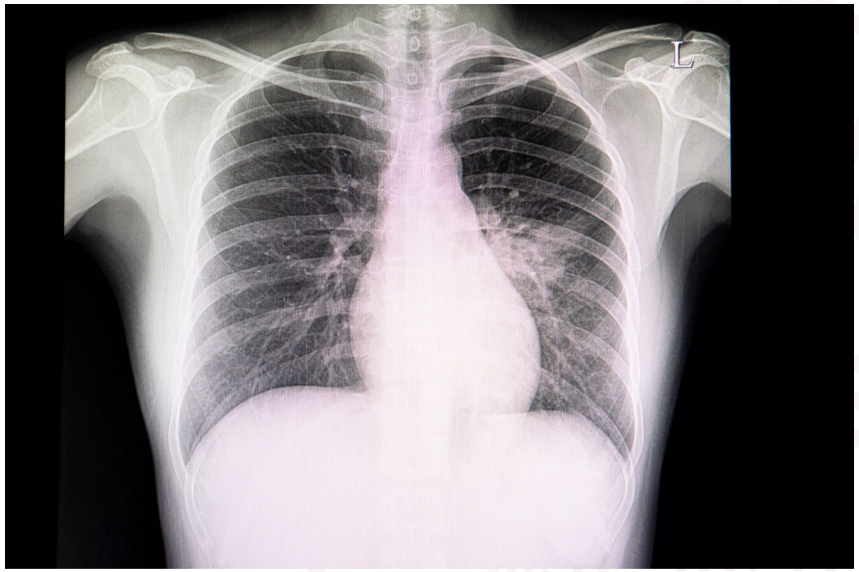
Yash Bansal (ybansal@usc.edu)
Kyoungeui Hong (hongkyou@usc.edu)
Hanie Kang (hanie.kang@usc.edu)
Juann Kim (juannkim@usc.edu)

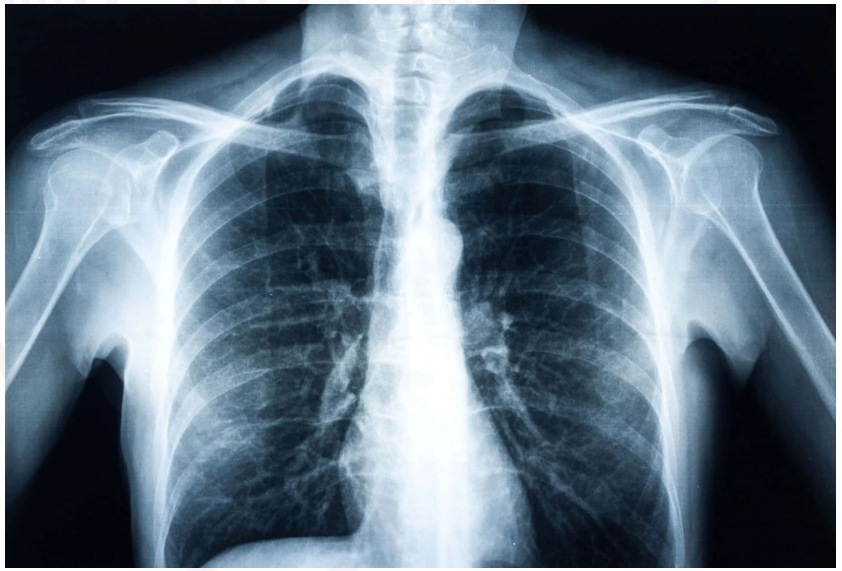
USC

The Challenge: Domain Shift in Medical Imaging

- Chest X-ray = globally essential + low-cost diagnostic tool
- AI models perform well only in controlled, single-hospital datasets
- Real-world hospitals → different scanners, noise, demographics
- Domain shift → performance collapse → patient risk

Goal: build AI that works consistently across environments

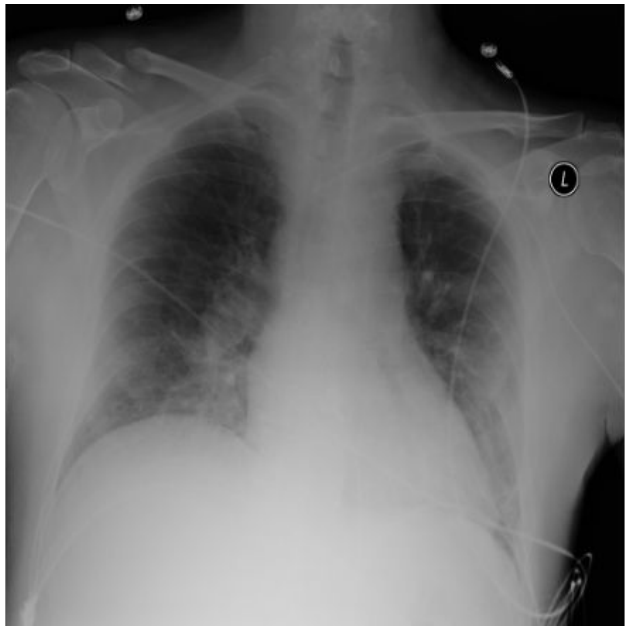



USC

Our Data Foundation

MIMIC-CXR

- high-quality, single institution
- train and validation
- 3,710 images with human-verified annotations



ChestX-ray14

- multi-hospital, noisy, realistic
- test only
- 112,120 images with NLP-mined disease labels (>90% precision)

USC

Our Data Foundation

- Source (Train): MIMIC-CXR
- Target (Test Only): ChestX-ray14
- 14-label multilabel task (Pneumonia, Effusion, etc.)
- No target-domain data used in training
  → Strict domain generalization setup

#	Label	Train	Test
0	Atelectasis	9247	2312
1	Cardiomegaly	2221	555
2	Consolidation	3734	933
3	Edema	1842	461
4	Effusion	10654	2663
5	Emphysema	2013	503
6	Fibrosis	1349	337
7	Hernia	182	45
8	Infiltration	15915	3979
9	Mass	4626	1156
10	Nodule	5065	1266
11	Pleural_Thickening	2708	677
12	Pneumonia	1145	286
13	Pneumothorax	4242	1060

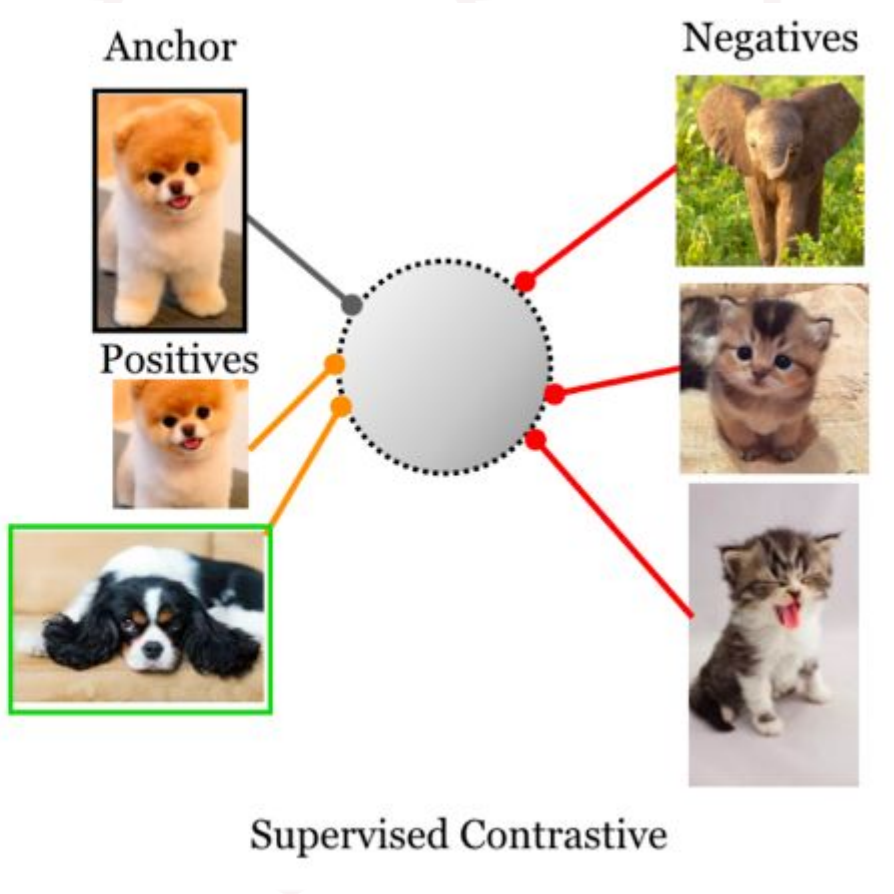
Supervised Contrastive Learning

- Positive pairs: same image + augmented views
- Negative pairs: images with different labels

Encourages disease-level clustering

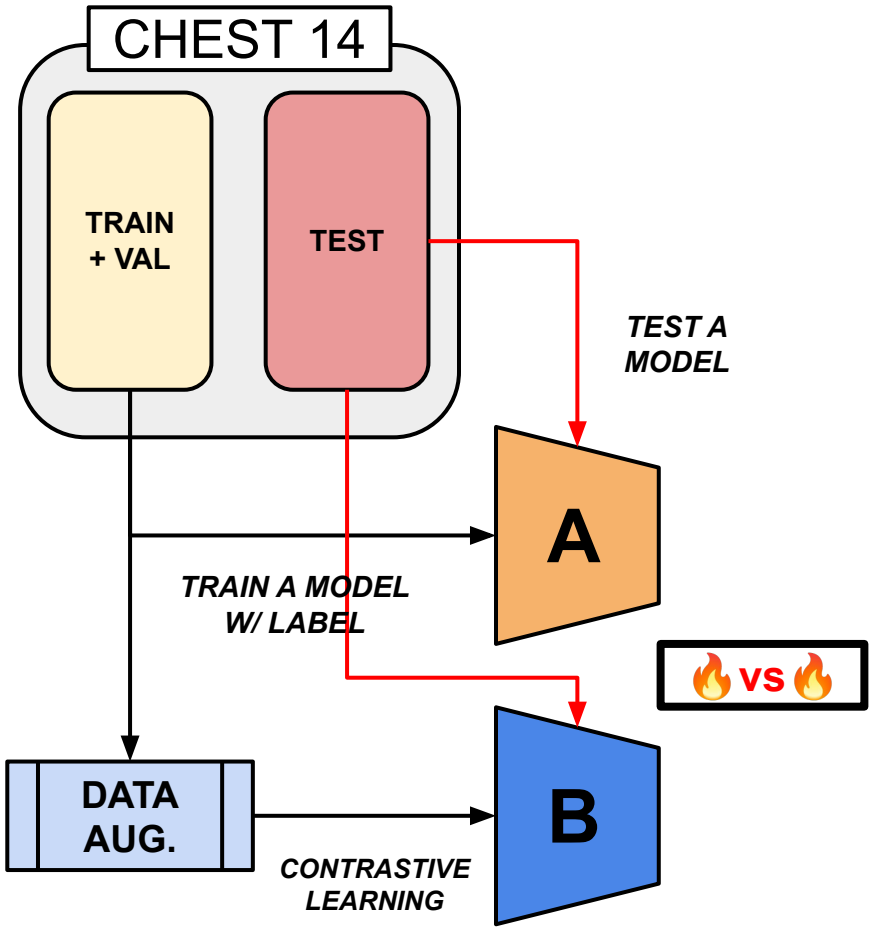
Reduces overfitting to domain artifacts

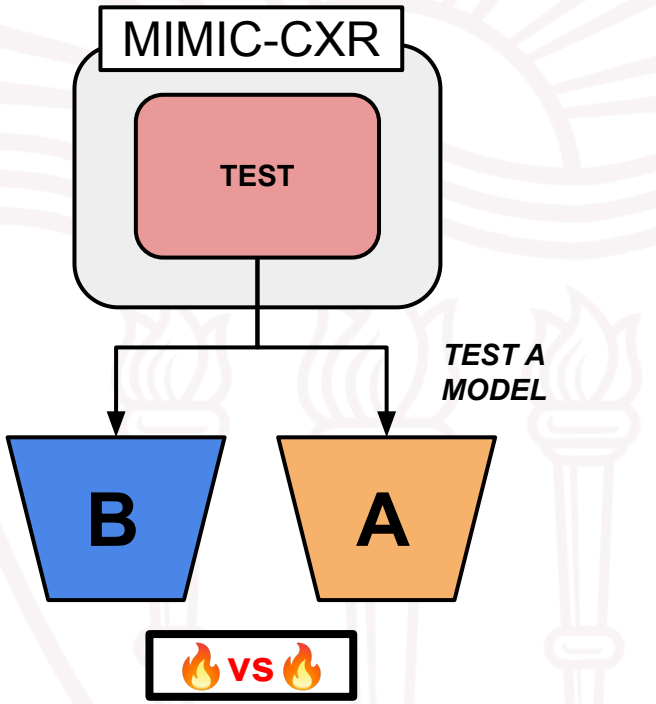
$$\mathcal{L}_{out}^{sup} = \sum_{i \in I} \mathcal{L}_{out,i}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau\right)}$$



USC

Overview Pipeline





Expected Results: *A model "B" recorded the highest accuracy. The "B" can be applied to another benchmark successfully even if it's trained by specific dataset.*

USC

Experimental Setup & Architecture Comparison

Training Configuration

- Compared three models: **ResNet-50**, **EfficientNet-B0**, and **ViT-B/16**
- All models initialized with ImageNet pretrained weights.
- NVIDIA A100 GPU / batch size: 128 / 10 epochs (Preliminary runs showed that models converge within a few epochs; 10 epochs offer stable convergence without excessive compute cost.)
- Linear Probing Setting: backbone weights frozen, only the final classification layer is trained.

Training

Supervised Learning: “Binary Cross-Entropy” loss (treats each label as an independent Bernoulli outcome)

Supervised Contrastive Learning: trains a representation model with a supervised contrastive loss. Two images are considered a positive pair if they share at least one disease label; otherwise they are treated as negatives. After contrastive pretraining, a linear classifier (again only the last layer) is trained on top of the frozen backbone using BCE loss.

$$\mathcal{L}_{out}^{sup} = \sum_{i \in I} \mathcal{L}_{out,i}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau\right)}$$

USC

Evaluation Metric

- **F1 micro:** Compute TP/FP/FN over all labels and samples jointly. Emphasizes performance on frequent labels; good overall indicator.
- **F1 macro:** Compute F1 per disease label, then average over 14 labels. Gives equal weight to common and rare diseases; highlights performance on under-represented conditions.
- **ROC-AUC macro:** Area under ROC curve for each label, then macro average. Threshold-free ranking quality; important when prevalence is low and operating thresholds may change.
- **mAP (mean Average Precision):** Average precision per label from the precision–recall curve, then mean over labels. Focuses on precision–recall trade-off, which is more informative than ROC in highly imbalanced medical data.

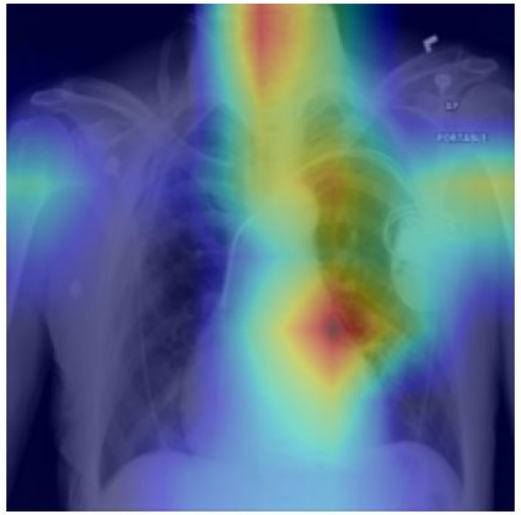
Supervised Learning Result				
model_name	f1_micro	f1_macro	auc_macro	map
ViT	27.84%	7.25%	71.70%	18.82%
EfficientNet	26.13%	5.21%	64.94%	15.50%
ResNet50	19.70%	4.74%	67.83%	16.32%

Supervised Contrastive Learning Result				
model_name	f1_micro	f1_macro	auc_macro	map
ViT	30.06%	9.41%	74.63%	21.79%
EfficientNet	29.14%	7.60%	67.82%	17.47%
ResNet50	27.95%	6.87%	69.68%	18.22%

USC

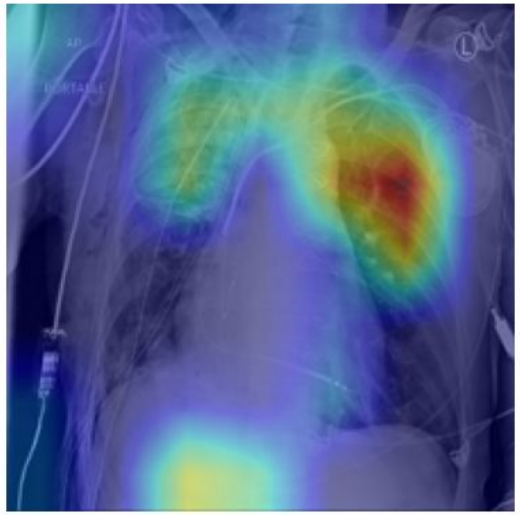
Attention Map (label results on the same model)

efficientnet - Atelectasis



efficientnet - Edema

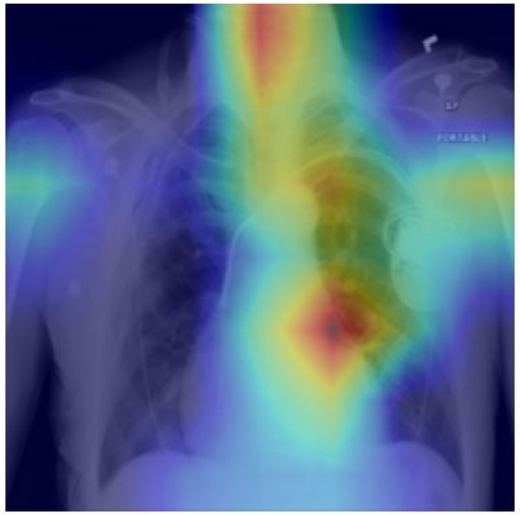
efficientnet - Pneumothorax



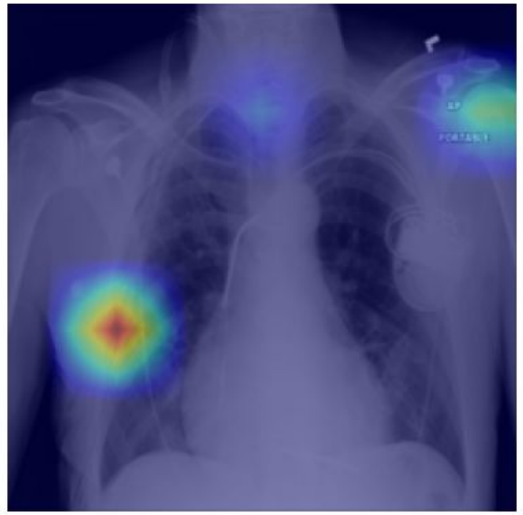
USC

Attention Map (model results on the same label)

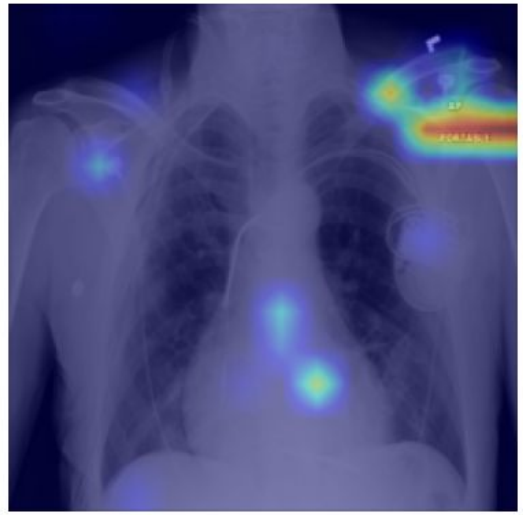
efficientnet - Atelectasis



resnet50 - Atelectasis



vit - Atelectasis



USC

Impact & Clinical Applications

Equitable Healthcare Access

- By developing models that generalize across institutions, we reduce diagnostic inequities and enable developing countries to effectively provide adequate health treatments, directly supporting SDG 10.

Robust Clinical Deployment

- Our domain generalization approach ensures AI models remain reliable across different imaging devices, patient populations, and clinical settings, addressing both technical robustness and healthcare fairness.

Improved Patient Outcomes

- Reliable AI-assisted diagnosis can streamline triage and reporting, improving accuracy and workflow efficiency while maintaining performance across diverse clinical environments.

Conclusion & Future Works

- Build a chest X-ray classification model that generalizes across hospitals and imaging conditions.
- Use supervised contrastive learning to encourage domain-invariant and pathology-relevant representations.
- Our method improves performance on cross-domain datasets and shows more clinically meaningful attention patterns.

Attention maps are still inconsistent across images and models.

Even for the same disease, models sometimes attend to different anatomical regions.

Clinical AI requires more stable, consistent attention patterns.

USC

Thank You!

USC