Confidence-guided Semi-supervised Learning for Generalized Lesion Localization in X-ray Images

Abhijit Das[1], Vandan Gorade[1], Komal Kumar[1], Snehashis Chakraborty[1], Dwarikanath Mahapatra[2], and Sudipta Roy[1]*

[1] Artificial Intelligence Data Science, Jio Institute, Navi Mumbai-410206, India
[2] Inception Institute of AI, Abu Dhabi, United Arab Emirates

Abstract. In recent years pseudo label (PL) based semi-supervised (SS) methods have been proposed for disease localization in medical images for tasks with limited labeled data. However these models are not curated for chest x-rays containing anomalies of different shapes and sizes. As a result, existing methods suffer from biased attentiveness towards minor class and PL inconsistency. Soft labeling based methods filters out PLs with higher uncertainty but leads to loss of fine-grained features of minor articulates, resulting in sparse prediction. To address these challenges we propose AnoMed, an uncertainty aware SS framework with novel scale-invariant bottleneck (SIB) and confidence guided pseudo-label optimizer (PLO). SIB leverages base feature ($\mathcal{F}_b$) obtained from any encoder to capture multi-granular anatomical structures and underlying representations. On top of that, PLO refines hesitant PLs and guides them separately for unsupervised loss, reducing inconsistency. Our extensive experiments on cardiac datasets and out-of-distribution (OOD) fine-tuning demonstrate that AnoMed outperforms other state-of-the-art (SOTA) methods like Efficient Teacher and Mean Teacher with improvement of 4.9 and 5.9 in $AP_{50:95}$ on VinDr-CXR data. Code for our architecture is available at https://github.com/aj-das-research/AnoMed.

Keywords: Uncertainty Aware · Confidence Guided · Detection.

1 Introduction

Computer aided diagnosis (CAD), a key aspect to gain medical assistance in disease identification and treatment planning. Detection of abnormalities in CXRs often involves drawing a bounding-box (b-box) around the target lesion [5]. Supervised methods like Faster-RCNN [15], Mask-RCNN [4], SSD [7], YOLO [14] have shown remarkable performance in supervised setting. But the lack of well annotated datasets in medical imaging has raised the importance of semi-supervised (SS) b-box detection methods [8, 21, 24, 22, 2, 16, 1], especially in thorax x-rays.

In recent years PL based SS frameworks like Efficient Teacher [20], Mean Teacher [17] achieved significant performance on natural vision. Two key factors in these methods are- **(1) Designing algorithms that can exploit the**

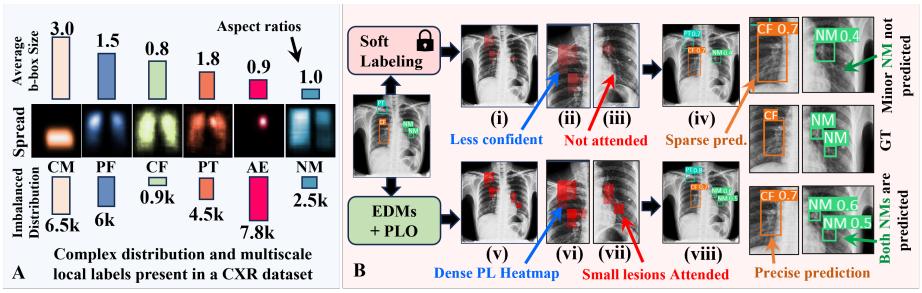

Fig. 1: **(A) Overview of a chest x-ray dataset validating the need of a scale-invariant feature extractor. (B) Effectiveness of proposed EDMs and PLO during PL generation.** Example input image from VinDr-CXR test set has 3 annotations of varying sizes: Plural Thickening (**PT**), Calcification (**CF**) and Nodule/ Mass (**NM**). Soft Labeling fails to attend the minor lesions **NM** (7x smaller than **CF**). Conversely, PLO generates dense PL heatmap, attends the minor lesions precisely.

inherent semantics of the unlabeled data, and (2) Maintaining the consistency of generated PLs across all classes. FPNs [6] has dominated over *Featurizing Image Pyramids* [13] in extracting semantics from encoder layers to form a feature pyramid. But, in FPNs information loss from the top level of the pyramid hinders the objective of learning coarse to fine-grained features. To tackle this, AugFPN [3] proposed residual augmentation based FPN. Nevertheless, this may not explicitly capture the scale-invariant information flow in medical images. On the other hand PL assignment is the key to mitigate PL inconsistency. While consistency regularization methods like ATSS [23] and AutoAssign [25] cannot be applied in SS detection frameworks, *soft labeling* [18] proposed a PL filtering technique. This discards the uncertain PLs based on only the objectness score while optimizing SS loss.

In CXR images, there can be multiple target lesions of different scales (as shown in Fig 1(A)) and datasets are often highly imbalanced. While adopting existing natural vision based SS methodologies on medical images there are two major concerns- **(1) Existing FPNs and AugFPN may not capture the scale-invariant features in CXRs**. In Fig 1(B) also, larger target regions has strong bias and enforces the detector to rarely attend the minor anomalies. (2) **To reduce PL inconsistency soft labeling methods simply drop uncertain PLs by a 2-way threshold.** And the unsupervised loss ($\mathcal{L}_{det}^{us}$) only accounts for the classification scores of each PL ignoring regression efficacy. This gradually increases the $\mathcal{L}_{det}^{us}$ at each iteration and model fails to converge, predicting inaccurate and sparse b-boxes (Fig 1(B)).

To overcome these challenges we propose **AnoMed**, an anomaly detection framework tailored for medical images. To the best of our knowledge, AnoMed is a pioneering effort to develop a consistency regularization based SS b-box regression method for cardiac diseases. AnoMed consists of a novel scale-invariant bottleneck (SIB). Inside SIB multi-layer encoding-decoding modules (EDMs)

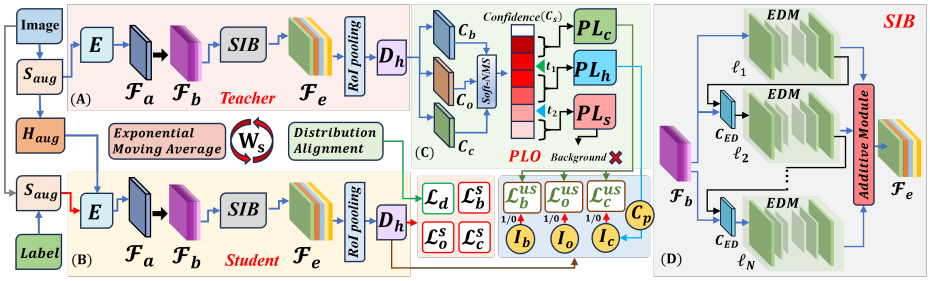

Fig. 2: **Holistic architecture of AnoMed**. A SS setup with Teacher-Student mutual learning, novel SIB module and curated confidence-guided loss function.

captures underlying representations in enhanced feature map ($\mathcal{F}_e$). Furthermore, AnoMed incorporates a new uncertainty aware optimizer PLO that categorizes the PLs into *Confident* (PL_c), *Hesitant* (PL_h) and *Scrap* (PL_s). Then, utilizing an uncertainty aware $\mathcal{L}_{det}^{us}$ PLO prevents hesitant PLs to participate in subsequent model updates and increases the confidence scores in PL_h. PLO exhibits denser PL heatmaps as shown in Fig 1(B) and detects precise b-box for minority classes too. To this end, the key contributions of our paper are summarized as:

1. **An unique layered encoding-decoding bottleneck, SIB is proposed.** SIB module captures local to global features of small to large diseases.
2. **Uncertainty guided PLO to reduce PL inconsistency.** Instead of filtration, PLO refines the PLs and boosts confidence into mutual learning. We also integrated distribution alignment ($\mathcal{DA}$) objective to stabilize training.
3. **Out-of-distribution fine-tuning of AnoMed for enhanced generalization.** AnoMed is evaluated on thin hairline fracture images besides CXRs.

2 Method

2.1 Preliminaries

Distribution Alignment: In SS learning $\mathcal{DA}$ learns a transformation $\mathcal{D}$ that minimizes the distribution discrepancy between the feature representations of labeled and unlabeled data using the KL divergence [10], Formulated as:

$$\min_{\mathcal{D}} \mathrm{KL}\left(P(X_l) \,\middle\|\, P(X_u)\right) = \sum_{x_l \in X_l} P(x_l) \log \frac{P(x_l)}{Q(x_l)} \tag{1}$$

where, $P(X_l)$ and $P(X_u)$ represent the probability distributions for labeled and unlabeled data, respectively. Here, $P(x_l)$ and $Q(x_l)$ are the probability density functions of feature representations. $\mathcal{DA}$ quantifies the amount of information loss when $P(X_l)$ is used to approximate $P(X_u)$ and aligns the representations from labeled and unlabeled data.

2.2 Proposed Architecture

Overview: Proposed SS framework **AnoMed** is based on a student-teacher mutual learning approach driven by exponential moving average. 4 major components are encoder (**E**), **SIB**, **PLO** and a curated **SS loss function** ($\mathcal{L}_{det}$). Labeled ($\{(\mathbf{x_i}, \mathbf{y_i})\}_{\mathbf{i=1}}^{\mathbf{N}}$) and unlabeled ($\{x_j\}_{j=1}^{M}$) images are fed to **E** (Default is ResNet50). Extracted features from the encoder layers are scaled and aggregated by the *Feature Aggregator* ($\mathcal{F}_a$) capturing multi-level information. $\mathcal{F}_b$ is fed to the SIB to obtain instance-level scale-invariant features and intrinsic relationships among the large and small anomalies. SIB employs spatial attention ($\mathcal{A}_s$) gated shallow **EDMs** and features from the decoder are fused by the *Additive Module* along the channel to obtain $\mathcal{F}_e$. Then anchor boxes are generated across the $\mathcal{F}_e$. Region of Interest (RoI) pooling extracts fixed-size feature maps from $\mathcal{F}_e$ corresponding to each proposed RoI defined by the anchor boxes. The detection head ($\mathbf{D_h}$) processes the RoI-pooled features to predict b-box offsets, classification scores, and objectness scores for each RoI. Then proposed **PLO** is employed to refine the generated PLs. Followed by a soft non-max-suppression b-boxes are predicted. Proposed loss function $\mathcal{L}_{det}^{ss}$ calculates the weighted loss for labeled and unlabeled data and back-propagates through the student.

Feature Aggregator ($\mathcal{F}_a$): Fusion of feature maps of the backbone enhances semantic information into $\mathcal{F}_b$. This can be formulated as:

$$\mathcal{F}_b = \mathcal{F}_a(\mathbf{X}_1, \mathbf{X}_2, .., \mathbf{X}_n) = \mathbf{F}\left(\bigoplus_{i=1}^{n} \mathbf{U}(\mathbf{X}_i)\right) \tag{2}$$

where, $\bigoplus$ represents the concatenation along the channel dimension, $\mathbf{U}(\mathbf{X}_i)$ denotes upsampling, X_i denotes the feature map of scale i, and $\mathbf{F}(\cdot)$ is the feature fusion operation.

Scale Invariant Bottleneck (SIB): As shown in Fig 2 SIB takes $\mathcal{F}_b$ as input and outputs enhanced spatial and channel-wise attended scale-invariant multilevel enhanced feature map $\mathcal{F}_e$. In SIB, **N** layers of shallow **EDMs** are stacked in parallel. These EDMs are designed to capture multiscale features from the input $\mathcal{F}_b$. In each EDM encoding layers gradually decrease the spatial dimensions and increase the number of channels. Then spatial attention gated decoder gives a attended feature representation capturing dense features. Encoder-decoder connector ($\mathbf{C_{EDs}}$) scales the features extracted from one EDM to match the shape of next EDM. alternating combination of EDMs and $\mathbf{C_{EDs}}$ generates multiscale feature maps. Features obtained from all layers are fused by the *Additive Module* through a scale-wise concatenation and channel attention to obtain the $\mathcal{F}_e$. This can be formulated as:

$$\mathcal{F}_e = \text{softmax}\left(\frac{\mathbf{w}_2\delta\left(\mathbf{w}_1\mathcal{Z}\right)}{\sigma\left(\mathbf{w}_2\delta\left(\mathbf{w}_1\mathcal{Z}\right)\right)}\right) \odot \mathcal{F}_j \tag{3}$$

where, $\mathcal{F}_j$ is the input feature map for a specific channel, $\mathcal{Z}$ denotes the global average-pooled feature map $\mathbf{w}_1$ and $\mathbf{w}_2$ are weight matrices. $\odot$ denotes element-wise multiplication. The differentiating factor of SIB from the FPN based feature extraction is, **at each pyramid level FPN only contains features from**

a single scale, but SIB contains features from each scale and each level combined. This helps to detect multi-granular lesions and learn inter-class relationships at an instance level.

Pseudo Label Optimizer (PLO): $\mathcal{F}_e$ followed by a consecutive operation of region proposal and RoI pooling enters the detection head. $\mathbf{D}_h$ processes the RoI-pooled features to predict bounding b-box scores ($\mathbf{C_b}$), objectness scores ($\mathbf{C_o}$), and classification scores ($\mathbf{C_c}$) for each RoI. By soft NMS the confidence scores ($\mathcal{C}_s$) of PLs are obtained from RoIs. During SS training extraction of $\mathcal{C}_s$ remains identical for both student and teacher module. As illustrated in Fig 2, PLs with $\mathcal{C}_s$ above $\mathbf{t_1}$ are considered as *confident*, PLs with $\mathcal{C}_s$ below $\mathbf{t_2}$ are discarded and considered as background and the rest is the *hesitant* ($\mathbf{PL_h}$). Proposed PLO effectively leverages the $\mathbf{PL_h}$s by utilizing the uncertainly aware confidence guided SS loss function $\mathcal{L}_{det}^{ss}$. $\mathcal{L}_{det}^{ss}$ can be formulated as:

$$\mathcal{L}_{det}^{ss} = \frac{1}{M}\sum_{j=1}^{M} \underbrace{(\mathrm{CE}(S_c, \hat{S}_c) + (1 - \mathrm{IoU}(S_b, \hat{S}_b)) + (\mathrm{CE}(S_o, \hat{S}_o)) + \mathcal{L}_d}_{\mathcal{L}_{det}^{s} \text{ between PLO outputs and labels in Student}} + \underbrace{\lambda \cdot \mathcal{L}_{det}^{us}(\hat{S}, \hat{T})}_{\mathcal{L}_{det}^{us} \text{ loss}} \tag{4}$$

where, (S_c, $\hat{S}_c$), (S_b, $\hat{S}_b$), (S_o, $\hat{S}_o$) are the target-prediction pairs for classification, b-box and objectness scores respectively. $\mathcal{L}_d$ is the $\mathcal{DA}$ loss, calculated using equation 1. $\hat{S}$ and $\hat{T}$ are predictions by Student and PLs generated by the Teacher. $\mathcal{L}_{det}^{us}$ employs a checkpoint C_p to soften the t_1 and t_2 constraints and introduce flexibility to guide PL_h for unsupervised optimization. This effectively reduce false positives. When a PL is categorized as $\mathbf{PL_h}$, C_p outputs an indicator signal I_b or I_o or I_c (as shown in Fig 2). If any two of (C_b, C_o, C_c) are greater than t_2 then the one of the indicator activates and outputs 1. That means, if a hesitant PL has higher confidence in two of $\mathbf{C_c}$, $\mathbf{C_b}$ and $\mathbf{C_o}$ but suffers a deficit in $\mathcal{C}_s$, then the hesitant PL will be guided to the unsupervised loss for refinement with special attention towards the uncertain confidence attribute. The $\mathcal{L}_{det}^{us}$ can be formulated as:

$$\mathcal{L}_{det}^{us} = \frac{1}{M}\sum_{j=1}^{M} \underbrace{I_c \cdot \mathrm{CE}(\hat{S}_c, \hat{T}_c) + I_b \cdot (1 - \mathrm{IoU}(\hat{S}_b, \hat{T}_b)) + I_o \cdot \mathrm{CE}(\hat{S}_o, \hat{T}_o)}_{\mathcal{L}^{us} \text{ for hesitant pseudo labels } PL_h} + \underbrace{\mathrm{MSE}(\hat{S}, \hat{T}))}_{\mathcal{L}^{us} \text{ for } PL_c} \tag{5}$$

where, a single hesitant PL is guided to any one of the three uncertainty aware conditions. This potentially eliminates the inconsistency and sub-optimal model updates by boosting confidence. Impact of PLO and curated loss function is illustrated in Fig **??** in supplementary.

3 Experiments

Datasets: We utilized two b-box annotated open-source datasets. The VinDr-CXR dataset [11] includes 18000 CXR scans of 14 diseases. To mitigate potential disparity a *Weighted Box-Fusion (WBF)* preprocessing is applied. The TBX11K dataset [12] consists of 11200 X-rays, including 4 classes. We have used the

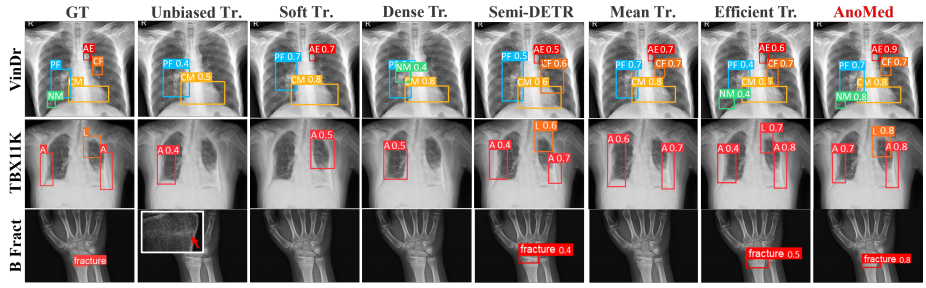

Fig. 3: Comparison of AnoMed against the SS SOTA methods on all three datasets.

official splits for experiments. For OOD fine-tuning the Hairline Bone Fracture (B-Fract) dataset [19] consisting 4346 images is used with 8:2 train-test split.

Implementation Details: Our models were implemented using PyTorch and trained on a 16GB NVIDIA RTX A4000 GPU. The input image size was 224×224, batch size of 8 with learning rate 0.01. We employed Adam optimizer with momentum 0.9 and weight decay of 0.0001 for optimization. Only mosaic is used for soft augmentation and for hard augmentation, flip, color-jittering, cut-mix are used. The max training epoch is 100 and EMA smoothing of 0.999.

Techniques of Comparison: We utilize $\mathbf{AP_{50}}$ and $\mathbf{AP_{50:95}}$ as standard metrics for b-box detection. AnoMed is evaluated against SOTA SS frameworks. For fair evaluation of proposed SIB we have evaluated AnoMed in supervised settings too. To further prove our hypothesis of scale-invariant learning we evaluated AnoMed for large and small target classes separately on VinDr-CXR. Performance with 5%, 10%, 15% and 20% labeled images is also reported. We have also visualized the attention heatmaps throughout the paper using gradient independent EigenCAM [9]. An interpretable analysis is also shown in Fig **??** in supplementary to understand AnoMed's decision making based on the learned salient features.

4 Results

VinDr-CXR: Table 1 states that AnoMed outperforms existing methods by a significant margin. With 20% labeled data, AnoMed achieves **4.9** and **5.9** $AP_{50:95}$ improvement over Efficient Teacher and Mean Teacher, respectively. It outperforms other methods in learning large anomalies (e.g., Cardiomegaly (CM), Pleural Fibrosis (PF)) and fine-grained anatomical structures (e.g., Aortic Enlargement (AE), Nodule or Mass (NM)). As shown in Fig 3, methods Efficient Teacher and Mean Teacher exhibits comparable results for large target lesions. But fails to detect the minor small sized anomalies due to the inherent feature loss from FPN and PL inconsistency. Transformer based Semi-DETR suffers from lack of local attentiveness. Additionally, AnoMed achieves more performance gain over existing methods with increasing data scarcity. Detailed analysis reported in Table **??** in supplementary.

Table 1: **Quantitative results of AnoMed against SOTA methods with 20% labeled data**. Shaded regions denotes supervised setting. **A**:VinDr-CXR, **B**:TBX11K-CXR, **C**:B-Fract dataset. Upper, middle, and lower halves of the table represent 2-stage anchor-based, 1-stage anchor-free, and 1-stage anchor-based methods, respectively.

Method	A						B		A → C*	
	All Classes		Large		Small		All Classes		All Classes	
			CM	PF	AE	NM				
	AP_{50}	$AP_{50:95}$	$AP_{50:95}$	$AP_{50:95}$	$AP_{50:95}$	$AP_{50:95}$	AP_{50}	$AP_{50:95}$	AP_{50}	$AP_{50:95}$
Fast-RCNN	40.6	26.4	41.6	29.8	17.4	25.2	52.3	39.0	57.8	42.3
Faster-RCNN	42.3	27.7	40.4	31.8	18.9	24.2	54.4	40.4	60.6	43.7
Unbiased Teacher	50.1	24.4	46.3	28.6	22.2	28.9	52.5	38.9	59.2	42.4
Soft Teacher	47.0	31.7	43.8	35.7	21.2	22.8	56.8	46.2	65.9	51.3
YOLOv8	45.2	30.2	41.3	33.7	24.5	21.0	55.2	39.8	61.5	43.2
DETR	46.8	31.2	43.0	35.2	27.4	25.4	57.6	41.4	63.4	45.4
Dense Teacher	49.5	33.6	45.0	37.3	28.6	26.1	68.2	52.0	75.5	57.4
Semi-DETR	52.4	35.9	47.3	39.0	29.8	28.6	70.8	54.5	77.8	59.2
SSD	53.5	34.2	43.5	35.1	27.4	31.0	66.4	52.3	77.4	55.2
RetinaNet	56.1	40.5	45.2	38.2	31.2	33.3	69.3	54.3	**79.4**	58.0
AnoMed^{S}	61.2	**43.2**	51.6	46.2	32.4	35.2	73.8	**58.1**	78.6	**61.5**
Efficient Teacher	**75.5**	46.5	56.1	50.1	**42.5**	43.9	78.0	62.8	81.5	68.7
Mean Teacher	73.6	45.9	55.8	49.4	41.0	39.3	79.6	61.8	**85.2**	64.4
AnoMed^{R18}	74.4	**49.2**	58.3	**51.6**	42.1	**44.4**	**81.9**	**65.8**	84.6	**70.0**
AnoMed	**76.8**	**51.4**	**59.4**	**54.8**	**44.3**	**46.1**	**82.2**	**67.0**	**86.6**	**71.6**

Note: * indicates out-of-distribution transfer learning, respectively. S indicates the supervised counterpart, and $R18$ represents AnoMed with ResNet18 encoder.

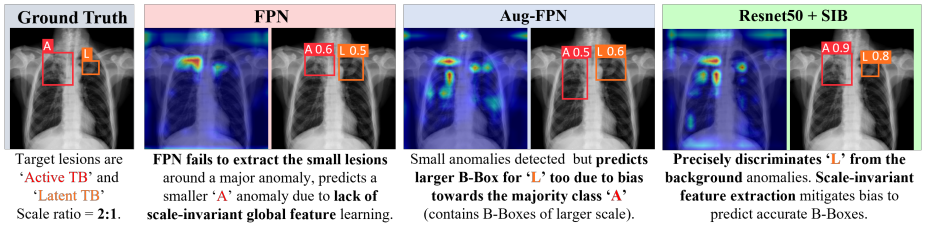

Fig. 4: Ablation of backbones showing effectiveness of SIB with ResNet50.

TBX11K: From Table 1, we note that AnoMed outperforms all the baselines with an improvement of **2.6** AP_{50} and **4.2** $AP_{50:95}$ over Mean Tr. and Efficient Tr.. Improvement is more significant in $AP_{50:95}$. That means, for AnoMed, the confidence score of predicted bounding boxes are higher than the baselines. Fig 3 shows that Efficient Tr. struggles with the bias induced by majority class A (Active TB), Unbiased Tr. outputs sparse predictions.

Out of Distribution (OOD) Fine-tuning on B-Fract: OOD fine-tuning on B-Fract dataset allows AnoMed to generalize on unseen data. As reported in the Table 1, AnoMed gains $AP_{50:95}$ of 3.4 and 5.2 over Efficient Tr and Mean Tr.. Efficient Tr. produces a comparable AP_{50} but $AP_{50:95}$ is poor. Due to PL uncertainty these models often fail to detect minute fractures. In Fig 3) while, Semi-DETR predicts the bone joint that resembles a fracture, AnoMed precisely detects with 0.8 confidence.

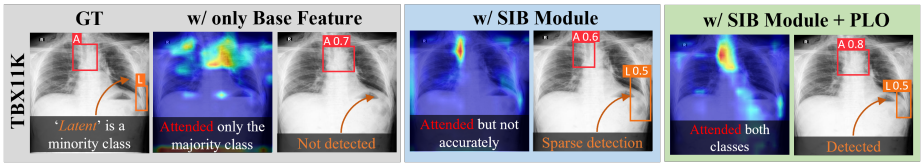

Fig. 5: Interpretation of effect of proposed modules on TBX11K.

Table 2: Ablation study of contribution of different modules (results in $AP_{50:95}$). ✓= *in use*.

Resnet18	✓							
Resnet50		✓	✓	✓	✓	✓	✓	✓
2EDMs			✓					
3EDMs				✓		✓	✓	✓
4EDMs					✓			
$\mathcal{A}_{cw}$						✓	✓	✓
PLO							✓	✓
$\mathcal{DA}$								✓
VinDr	38.5	39.7	41.27	44.7	45.6	46.5	48.1	**51.4**
TBX11K	56.4	58.3	58.6	59.2	60.1	62.6	64.3	**67.0**

Table 3: Ablation of different threshold values in PLO.

t_2	t_1	VinDr-CXR	TBX11K
0.2	0.5	34.6	52.3
0.2	0.6	39.5	63.4
0.2	0.7	48.7	48.9
0.3	0.5	49.3	62.8
0.3	0.6	**50.3**	**65.0**
0.3	0.7	**51.4**	**67.0**
0.4	0.5	45.4	50.5
0.4	0.6	46.8	51.5
0.4	0.7	45.5	53.4.5

5 Ablation Studies

Backbone Analysis: Fig 4 presents the impact of SIB with ResNet50 encoder over FPN and Aug-FPN. FPN fails to attend the minority class, while Aug-FPN seems to overfit due to deep supervision based approach. Comparatively SIB detects the minority class well and predicts the b-boxes with confidence 0.9 and 0.8 for major and minority class.

Depth Analysis of SIB: As presented in Table 2, with 3 EDMs AnoMed achieves a significant improvement of 5% and 1% on VinDr and TBX11K, respectively. While 4 EDMs increases the performance marginally, we choose 3 for computational efficiency. However, no of EDMs can be adjusted as per need.

Threshold Analysis in PLO: As reported in Table 3, for VinDr and TBX11K the optimal combination is $t_1 = 0.7$, $t_2 = 0.3$. Contribution of PLO with best setting is shown in 5. As we lower the value of t_1, $AP_{50:95}$ decreases. With low t_1 and t_2 values, PLs with less confidence score are treated as the confident PLs and guided to the unsupervised MSE loss as mentioned in equation 5. This results in severe PL inconsistency.

Contribution of Distribution Alignment: As reported in Table 2 Distribution Alignment ($\mathcal{DA}$) increases the performance by 3.3% and 2.7% of $AP_{50:95}$ on VinDr and TBX11K, respectively. Qualitative results are shown in Fig 6.

6 Conclusion

In summary, AnoMed stands out as a pioneering effort in PL based SS cardiac disease detection. Proposed SIB learns scale-invariant features through multi-layer EDMs and PLO with uncertainty aware un-supervision, guides hesitant PLs

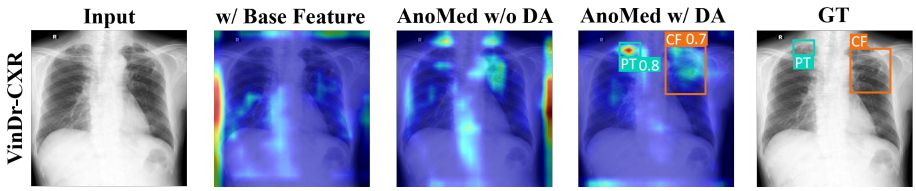

Fig. 6: **Effect of distribution alignment ($\mathcal{DA}$).** In semi supervised settings, $\mathcal{DA}$ allows the model exploit the maximum information from the unlabeled data and aligns the distribution of labeled and unlabeled data during PL based SS training. It makes model robust to noise also.

towards confident and consistent learning. AnoMed outputs precise detection of large to small anomalies in CXRs and generalizes well in OOD data as well. We also recognize the rise of Fourier Transformer based methods for local to global learning. In future work, we will extend our strategy by encoding the concept of discrete and continuous space based learning in the same consistency regularization SS problem setting.

7 Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chakraborty, S., Kumar, K., Tadepalli, K., Pailla, B.R., Roy, S.: Unleashing the power of explainable ai: sepsis sentinel's clinical assistant for early sepsis identification. Multimedia Tools and Applications pp. 1–29 (2023)
2. Gorade, V., Mittal, S., Singhal, R.: Pacl: Patient-aware contrastive learning through metadata refinement for generalized early disease diagnosis. Computers in Biology and Medicine **167**, 107569 (2023)
3. Guo, C., Fan, B., Zhang, Q., Xiang, S., Pan, C.: Augfpn: Improving multi-scale feature learning for object detection (2019)
4. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
5. Kumar, K., Chakraborty, S., Tadepalli, K., Roy, S.: Weakly supervised learning based bone abnormality detection from musculoskeletal x-rays. Multimedia Tools and Applications pp. 1–26 (2024)
6. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944 (2017). https://doi.org/10.1109/CVPR.2017.106
7. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 21–37. Springer (2016)

8. Liu, Y., Ma, C., He, Z., Kuo, C., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. CoRR **abs/2102.09480** (2021), https://arxiv.org/abs/2102.09480
9. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE (Jul 2020). https://doi.org/10.1109/ijcnn48605.2020.9206626, http://dx.doi.org/10.1109/IJCNN48605.2020.9206626
10. Nguyen, A.T., Tran, T., Gal, Y., Torr, P.H.S., Baydin, A.G.: Kl guided domain adaptation (2022)
11. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T.T., Dinh, D.H., Do, C.D., Doan, L.T., Nguyen, C.N., Nguyen, B.T., Nguyen, Q.V., Hoang, A.D., Phan, H.N., Nguyen, A.T., Ho, P.H., Ngo, D.T., Nguyen, N.T., Nguyen, N.T., Dao, M., Vu, V.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations (2022)
12. Pan, C., Zhao, G., Fang, J., Qi, B., Liu, J., Fang, C., Zhang, D., Li, J., Yu, Y.: Computer-aided tuberculosis diagnosis with attribute reasoning assistance (2022)
13. Pang, Y., Wang, T., Anwer, R.M., Khan, F.S., Shao, L.: Efficient featurized image pyramid network for single shot detector. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7328–7336 (2019). https://doi.org/10.1109/CVPR.2019.00751
14. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection (2016)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
16. Singh, A., Gorade, V., Mishra, D.: Optiml: Dense semantic invariance using optimal transport for self-supervised medical image representation. arXiv preprint arXiv:2404.11868 (2024)
17. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results (2018)
18. Vyas, N., Saxena, S., Voice, T.: Learning soft labels via meta learning (2020)
19. Wang, W., Huang, W., Lu, Q., Chen, J., Zhang, M., Qiao, J., Zhang, Y.: Attention mechanism-based deep learning method for hairline fracture detection in hand x-rays. Neural Computing and Applications **34**(21), 18773–18785 (2022)
20. Xu, B., Chen, M., Guan, W., Hu, L.: Efficient teacher: Semi-supervised object detection for yolov5 (2023)
21. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
22. Zhang, J., Lin, X., Zhang, W., Wang, K., Tan, X., Han, J., Ding, E., Wang, J., Li, G.: Semi-detr: Semi-supervised object detection with detection transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23809–23818 (2023)
23. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection (2020)
24. Zhou, H., Ge, Z., Liu, S., Mao, W., Li, Z., Yu, H., Sun, J.: Dense teacher: Dense pseudo-labels for semi-supervised object detection (2022)
25. Zhu, B., Wang, J., Jiang, Z., Zong, F., Liu, S., Li, Z., Sun, J.: Autoassign: Differentiable label assignment for dense object detection (2020)