28th International Conference on Medical Image Computing and Computer Assisted Intervention

25th September 2025

Multi-scale Attention-based Multiple Instance Learning for Breast Cancer Diagnosis

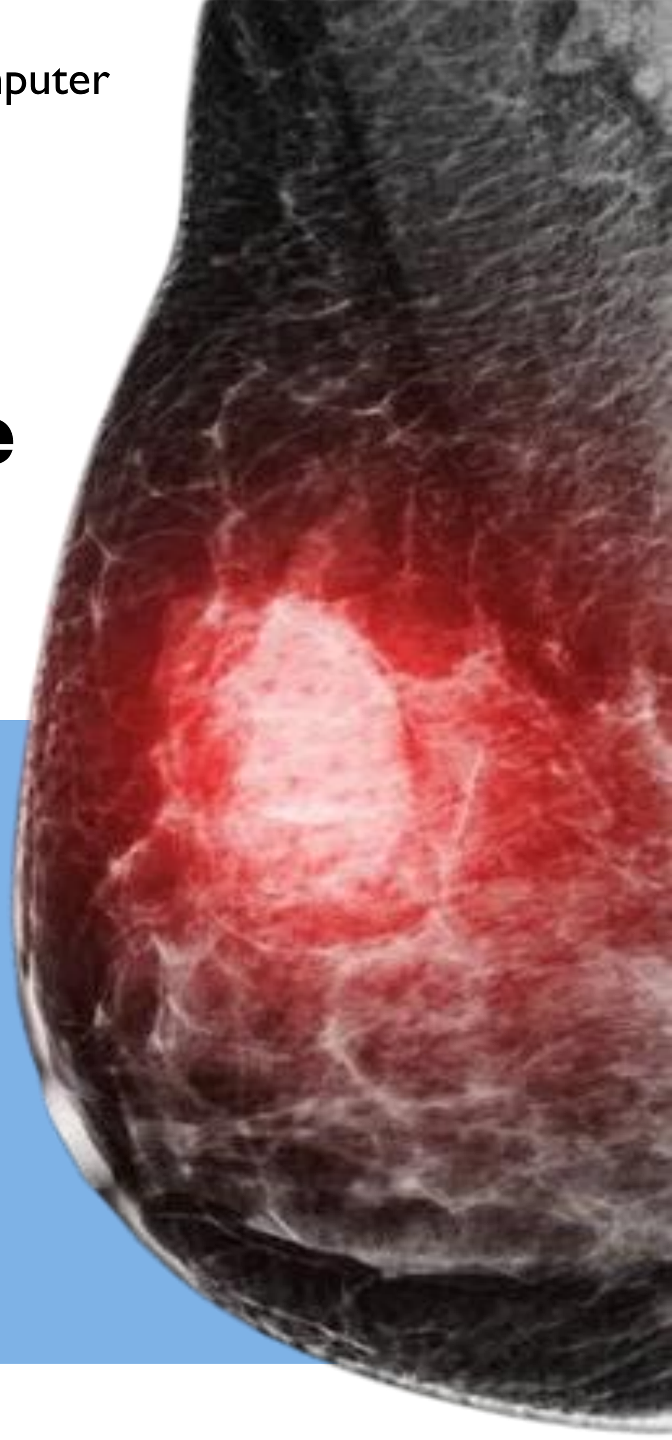
Mariana Mourão, Jacinto Nascimento, Carlos Santiago and Margarida Silveira

Institute for Systems and Robotics, Instituto Superior Técnico, Lisbon, Portugal

Institute for Systems and Robotics | LISBOA

TÉCNICO LISBOA

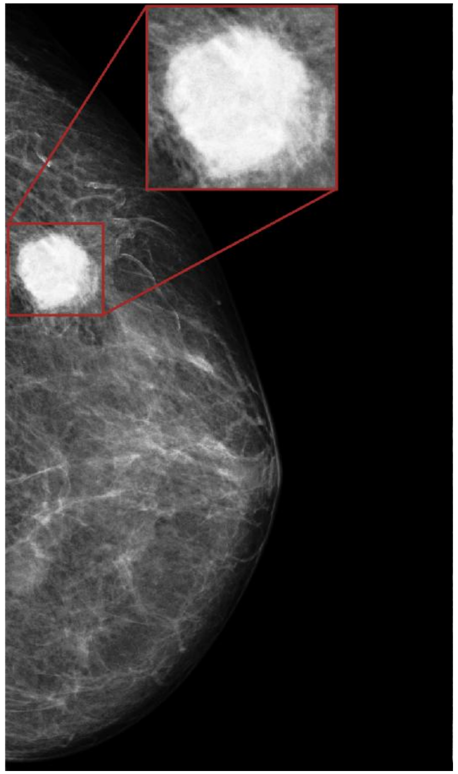
LARSyS Laboratory of Robotics and Engineering Systems



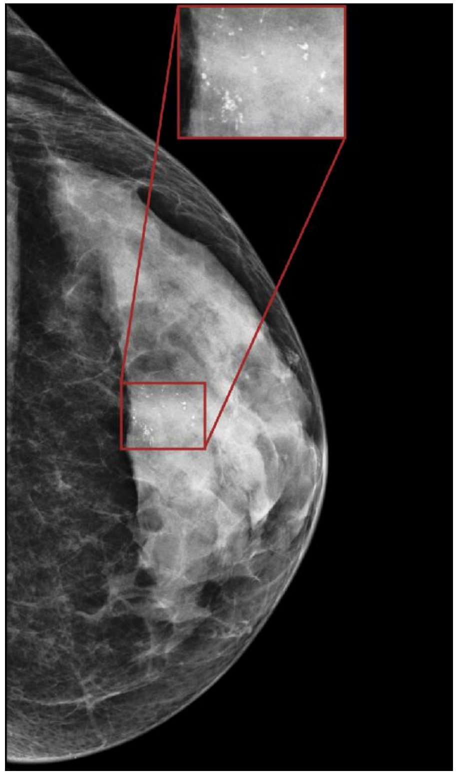
Mammographic Breast Cancer Diagnosis

Institute for Systems and Robotics | LISBOA

Mass



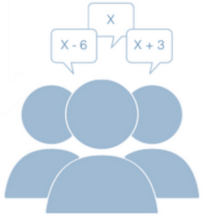
Calcifications



Challenges

High workload

Complex & diverse lesion features

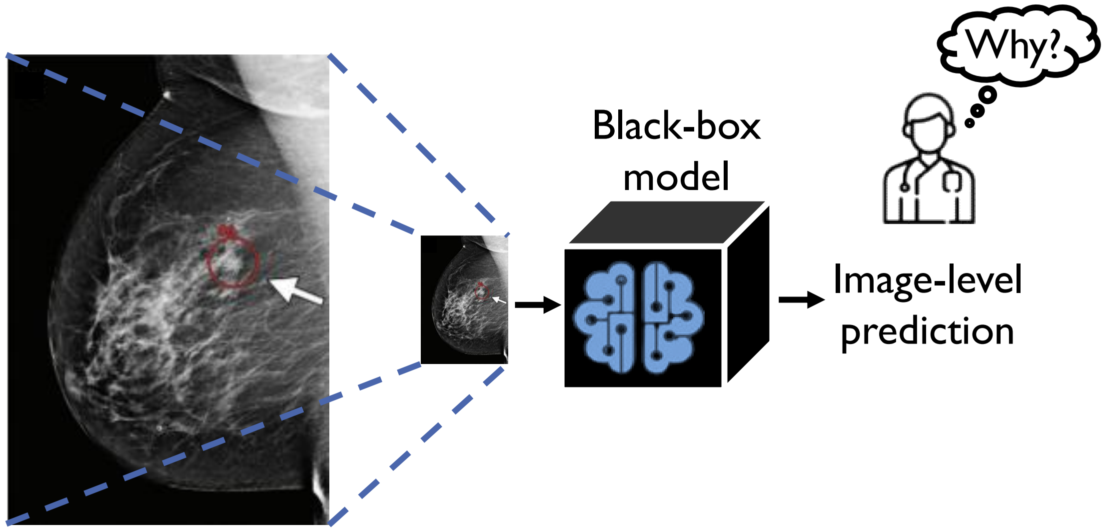


Intra- & inter-reader variability

Deep Learning (DL)-based Models

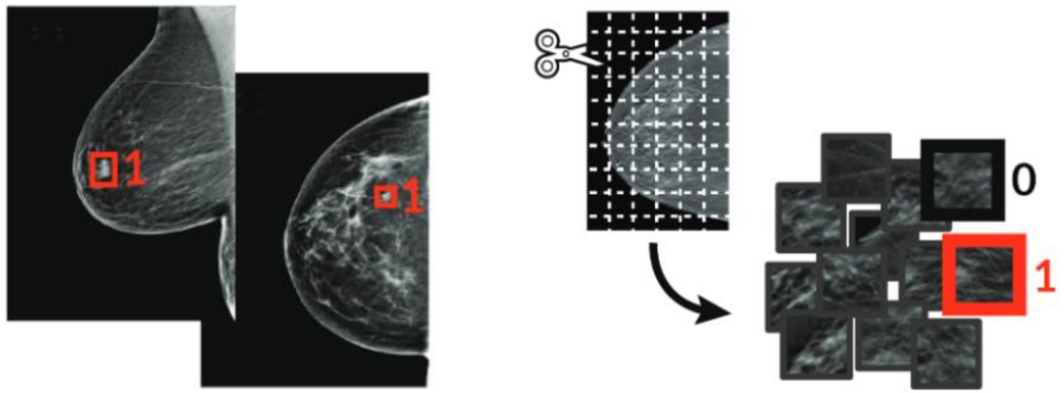
Full image-based DL models

- × Loss of detail from harsh downsampling
- × Lack interpretability



Region of Interest (ROI)-based DL models

- ✓ Improved performances & interpretability
- × Costly annotations under fully supervised learning

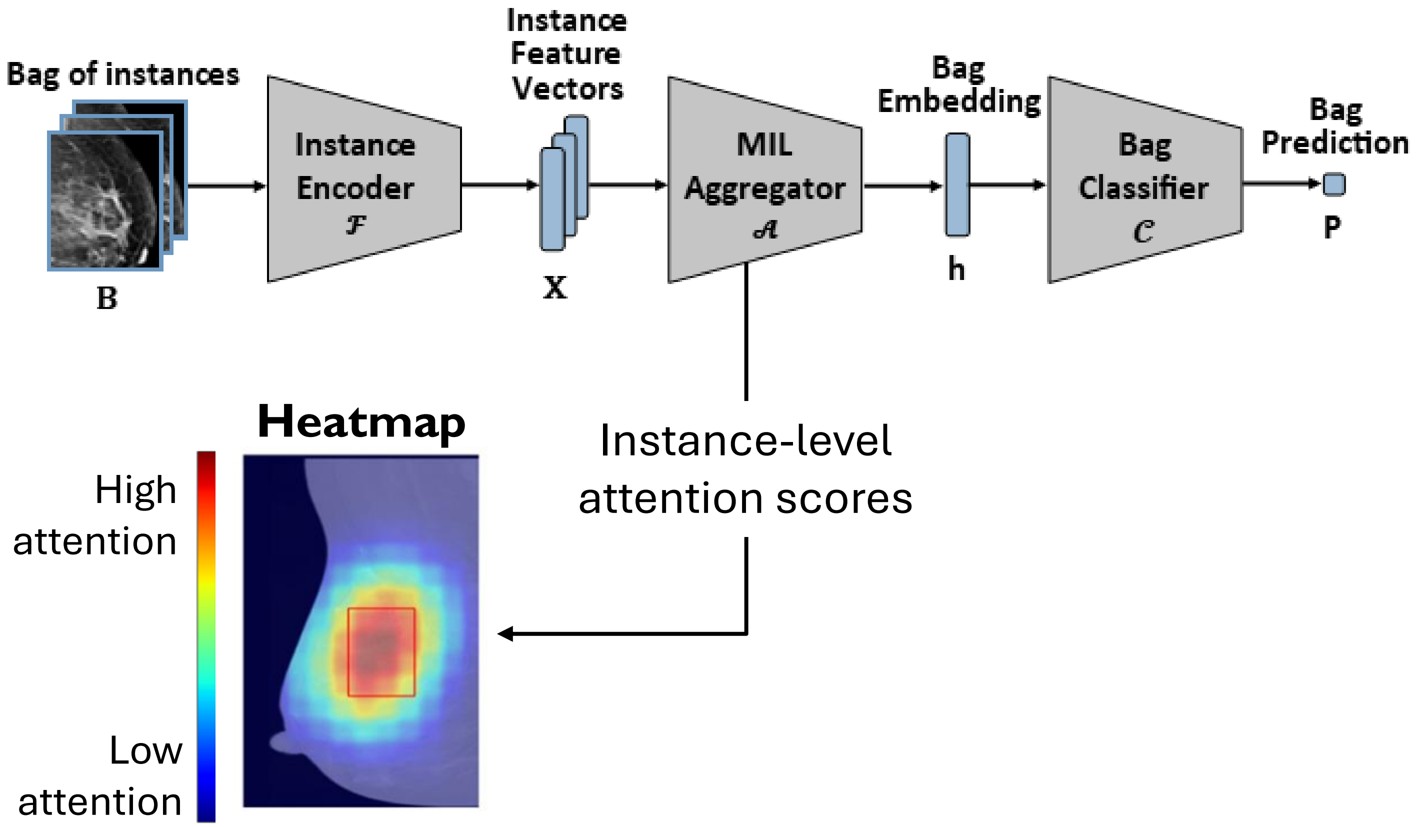


Bounding-box annotations

Patch-wise annotations

Multiple Instance Learning (MIL)

Typical MIL Framework in Mammography



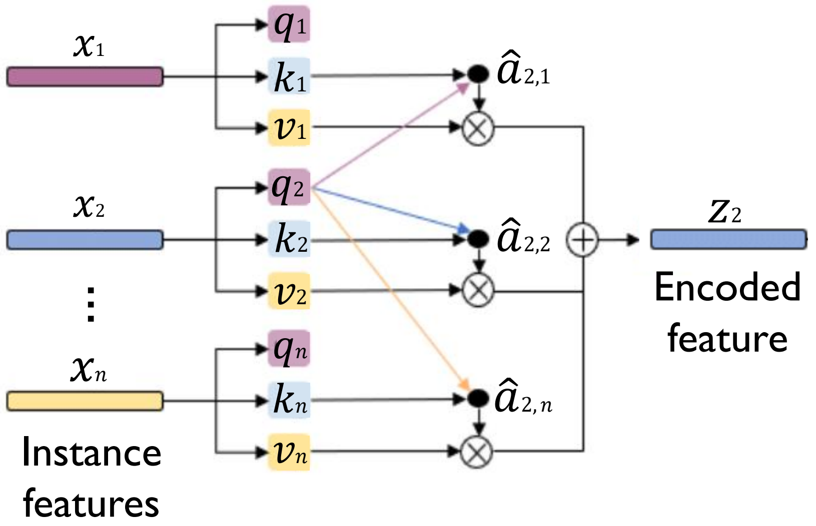
- ✓ Handles high-resolution images
- ✓ Attention-based aggregators enable image classification and instance detection
- ✓ Supervision with weak image-level labels

Limitations

- ✗ Neglects contextual information between instances
- ✗ Non-adaptive to multi-scale lesions

Related Works

Transformer Architectures



✓ Accounts for instance interactions

✗ $\mathcal{O}(n^2)$ computational complexity

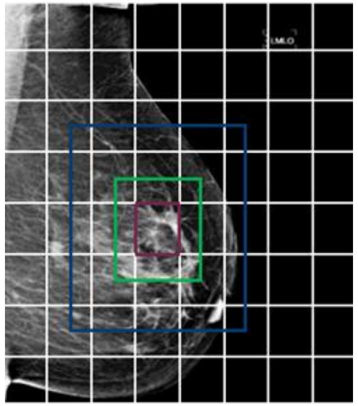
⇓

Efficient Transformers !

Multi-scale MIL models

Based on Multi-scale Patches (MSP)

MuSTMIL[1]; MSAA-Net[2]



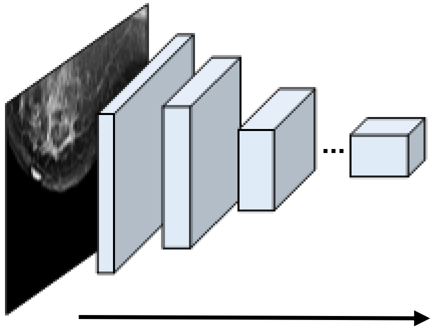
✓ High representational power across scales

✗ Increases computational burden

✗ Coarse patch-level detection granularity

Based on Feature Pyramids

Swin-MIL[3]



✓ Enhanced pixel-level detection granularity

✗ Operates on downsampled images

✗ Large semantic gap across scales

Contributions

Proposed a novel **multi-scale attention-based MIL framework** for weakly supervised classification and detection of breast lesions in high-resolution mammograms.

Multi-scale Instance Encoder

Builds a refined feature pyramid from single-scale patches

Flexible Instance Aggregators

Investigated localized and context-aware attention mechanisms

Multi-scale Aggregator

Adaptive scale fusion for robustness to lesion size variability

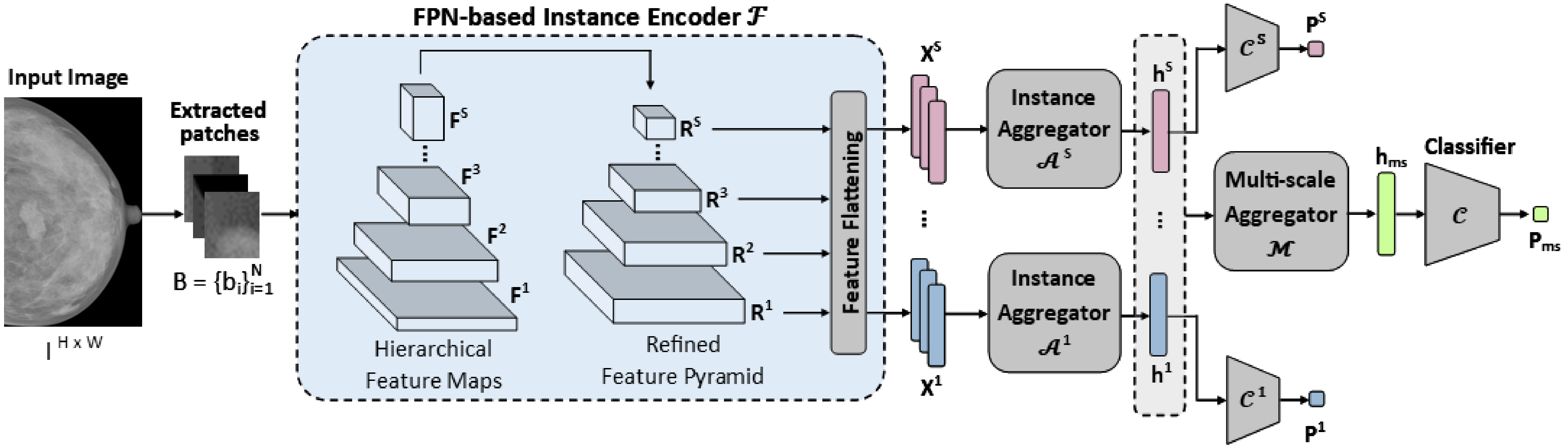
Comparison with Baselines & SoTA

Benchmark against baselines and SoTA models

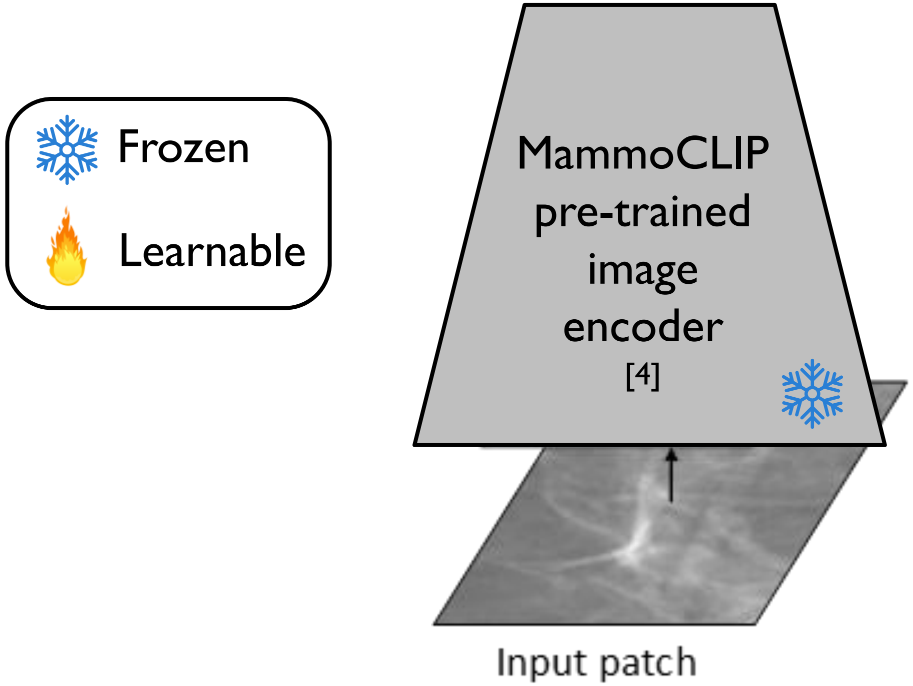
Ablation Studies

Evaluate the effectiveness of the main modules

Multi-scale Attention-based MIL Framework



Multi-scale Instance Encoder



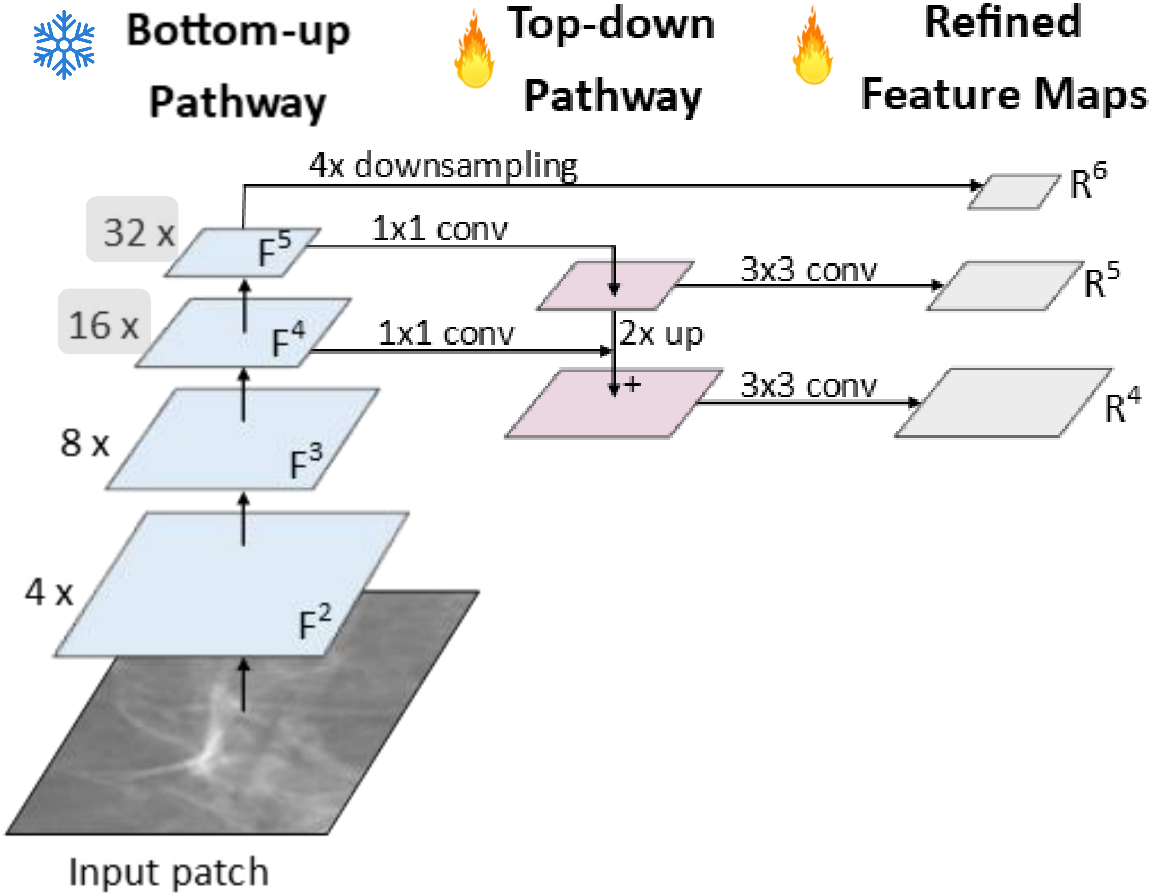
Multi-scale Instance Encoder

Institute for Systems and Robotics | LISBOA

Feature Pyramid Network (FPN) [5]

Frozen

Learnable

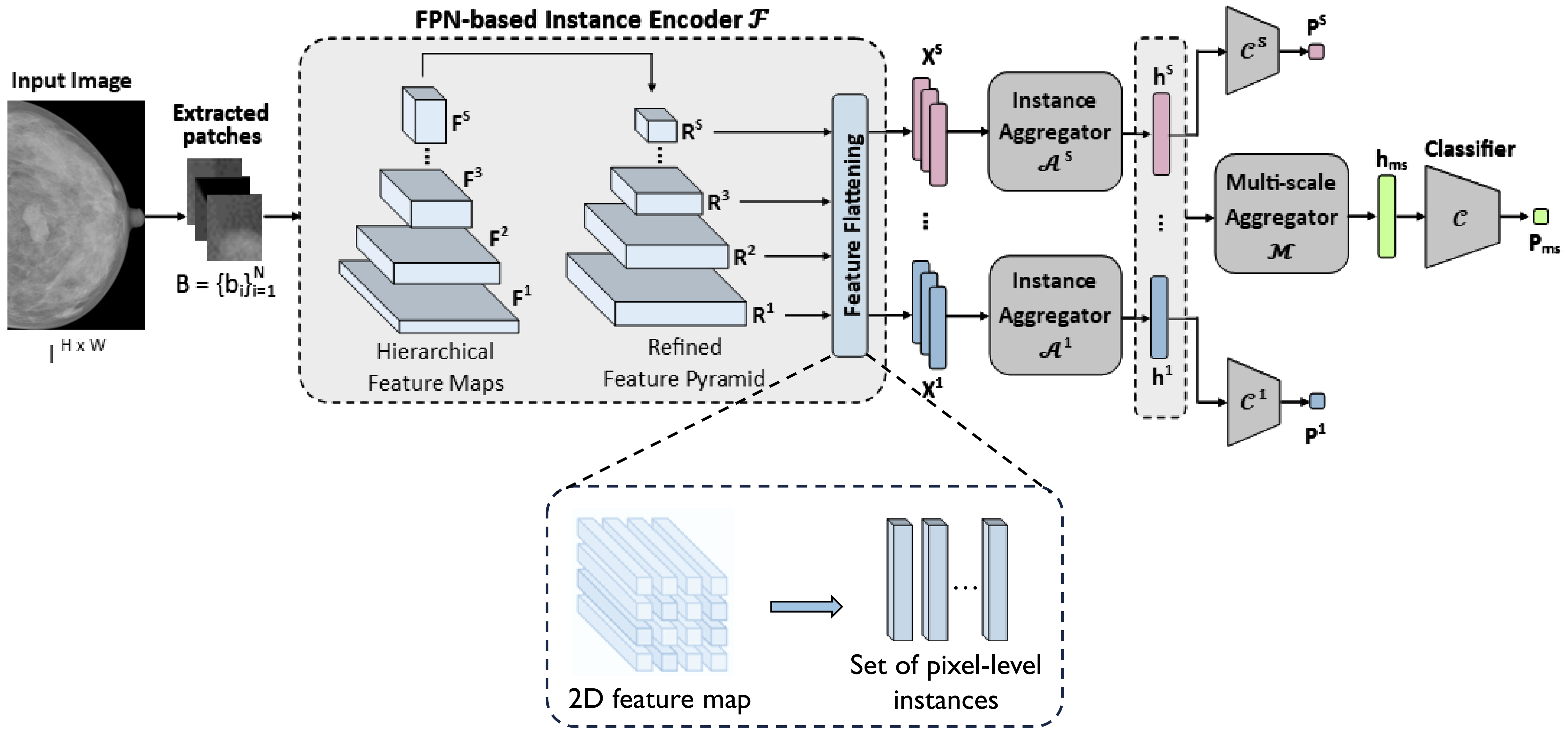


Large receptive-field

Medium receptive-field

Small receptive-field

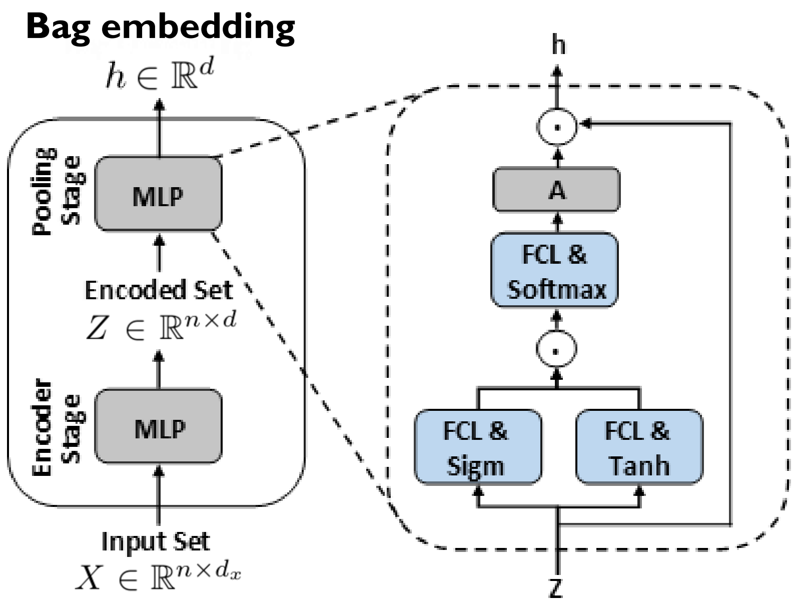
Multi-scale Attention-based MIL Framework



Institute for Systems and Robotics | LISBOA

Attention-based MIL (AbMIL) [6]

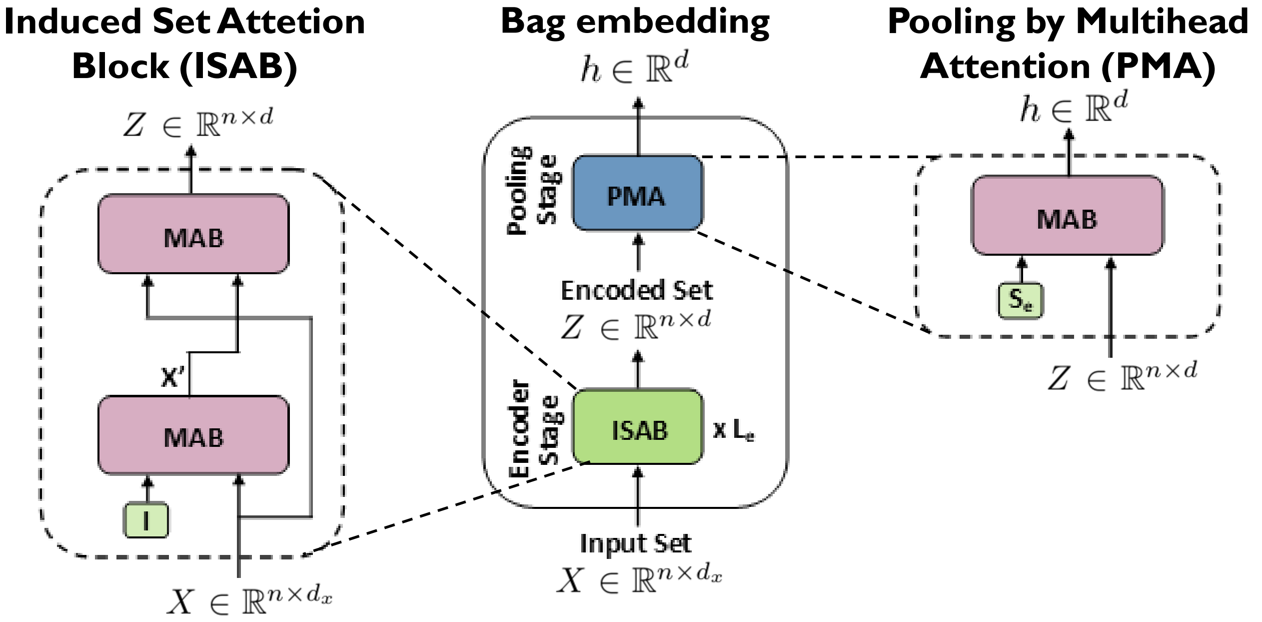
Localized attention instance aggregation, computing instance-level attention weights independently.



$$h = \sum_{i=1}^{n} a_i z_i$$

Set Transformer (SetTrans) [7]

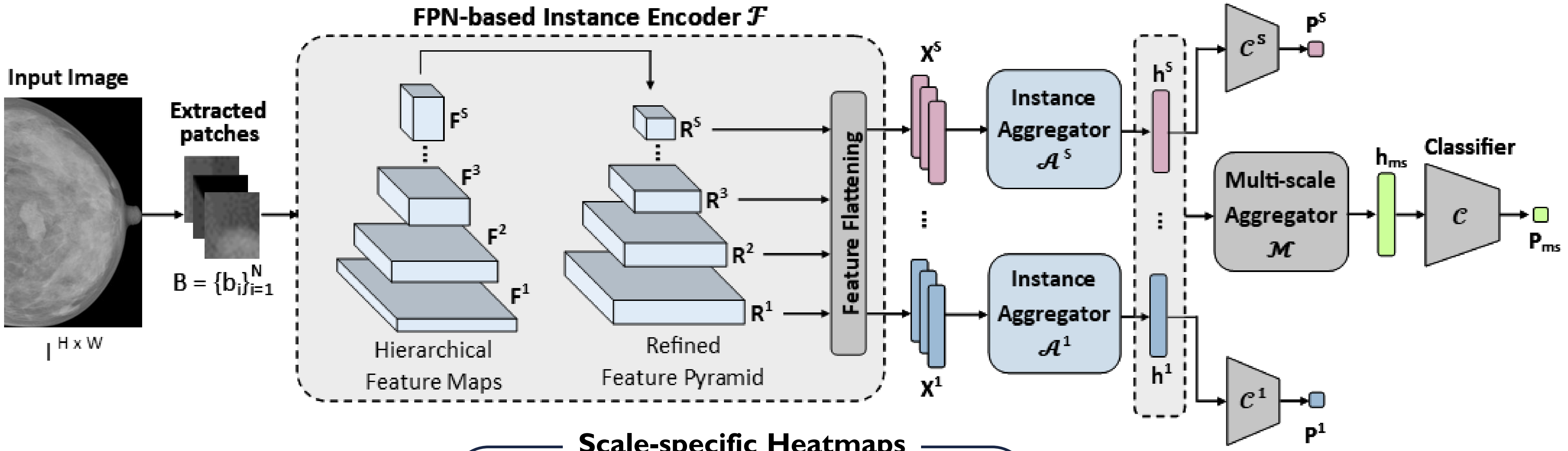
Efficient context-aware aggregation, with its basic operation – Multihead Attention Block (MAB) – being the vanilla transformer encoder.

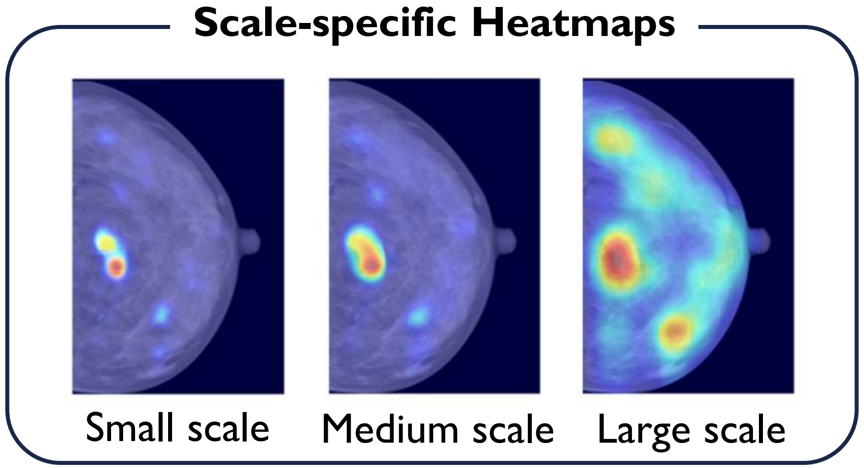


Number of inducing points : $m = 10 \times \log(n)$

Computational Complexity : $\mathcal{O}(m.n)$

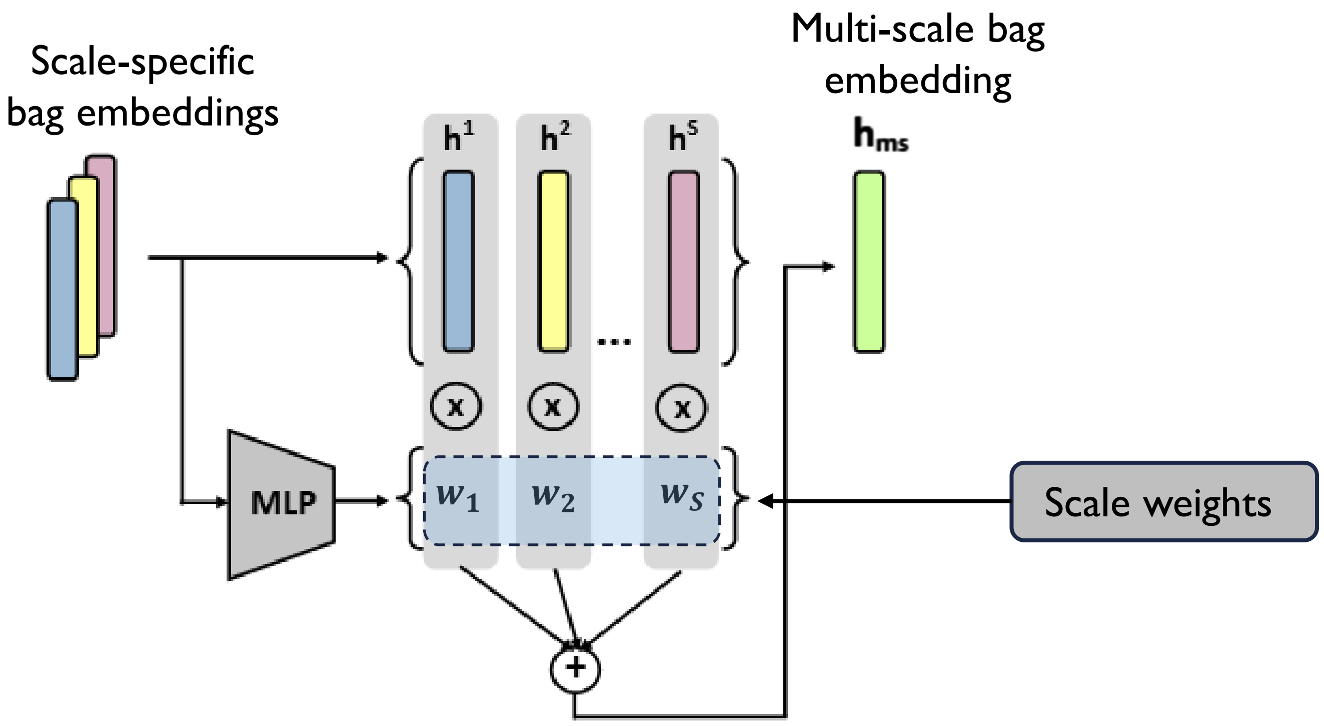
Multi-scale Attention-based MIL Framework

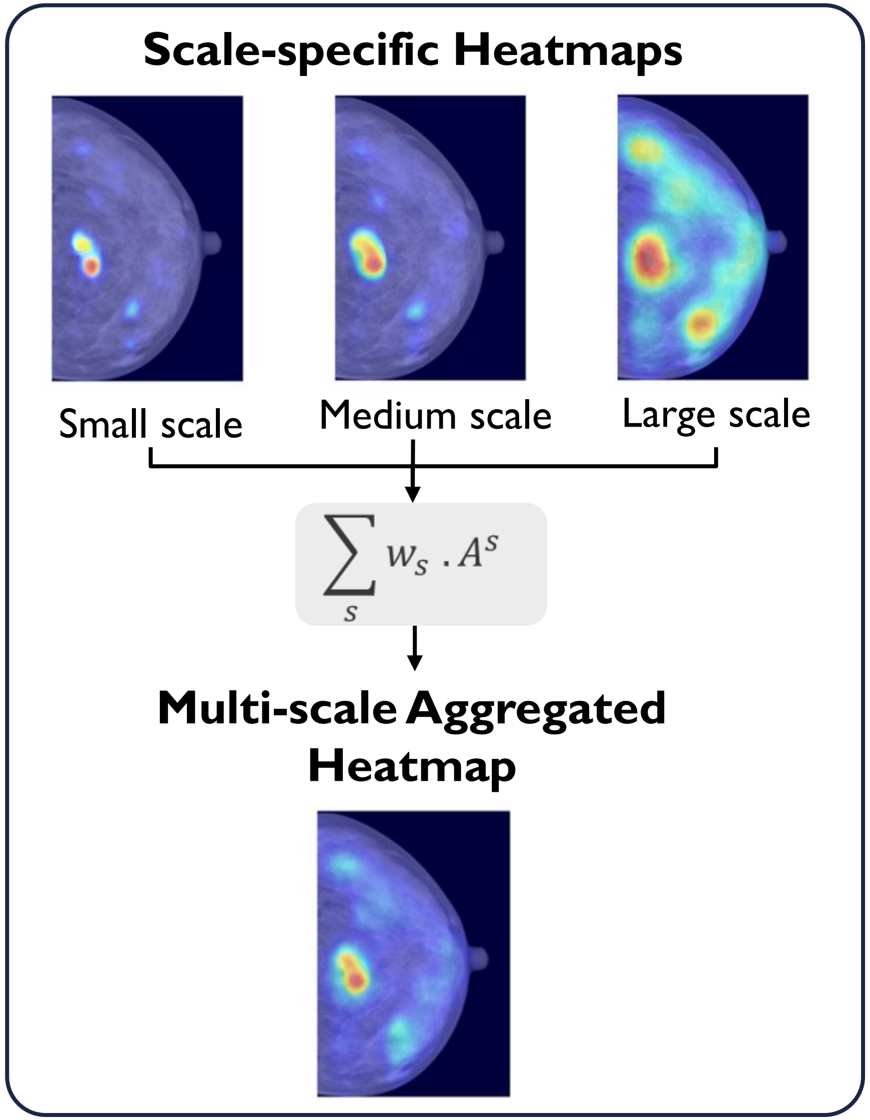




Multi-scale Aggregator

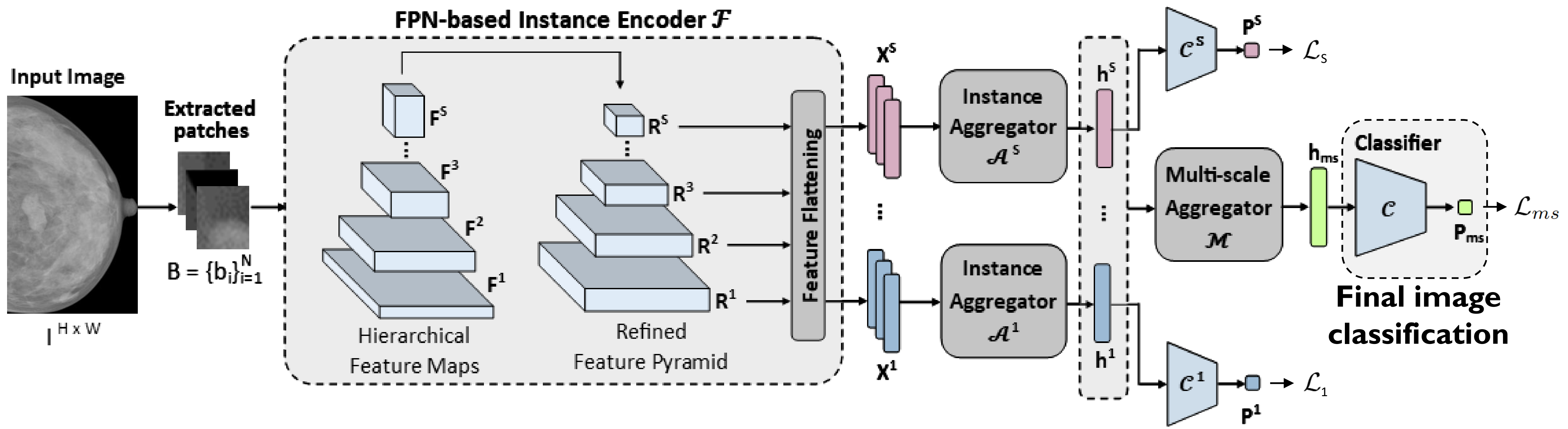
Attention-based MIL (AbMIL) [6]





Multi-scale Attention-based MIL Framework

Institute for Systems and Robotics | LISBOA



Loss function : $\mathcal{L}_{mil} = \mathcal{L}_{ms} + \sum_{s=1}^{S} \mathcal{L}_s$

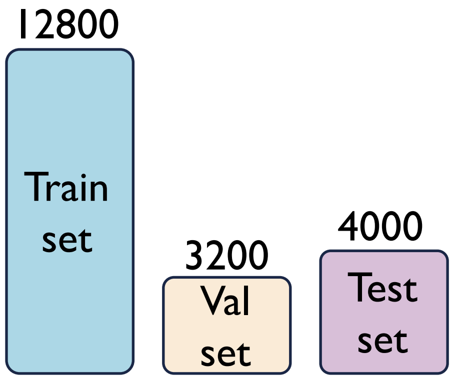
Multi-scale loss term

Deep-supervised scale-specific loss terms

Experimental Setup

VinDr-Mammo Dataset

- Used original train-test slipt
- 80%–20% class-stratified patient-wise train-validation split



Available annotations

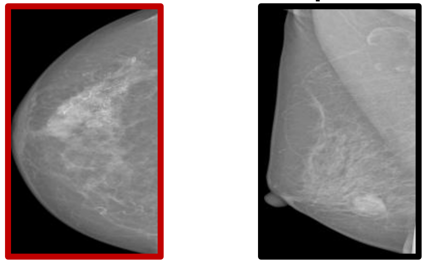
Image-level labels for training & classification evaluation

Bounding-boxes for detection evaluation

Image Classification

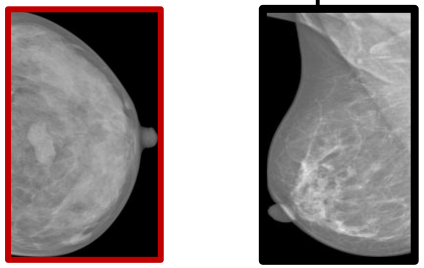
Calcifications

Present · Not present



Mass

Present · Not present

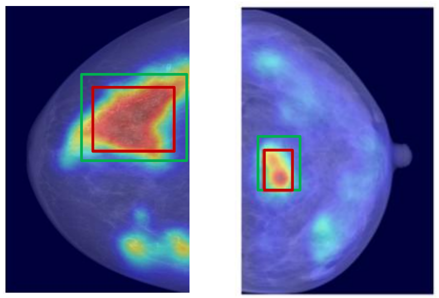


Evaluation metric

AUC-ROC

Lesion Detection

Multi-scale Aggregated Heatmaps



Calcifications · Mass

Predicted Bounding-box

Ground-truth Bounding-box

Evaluation metric

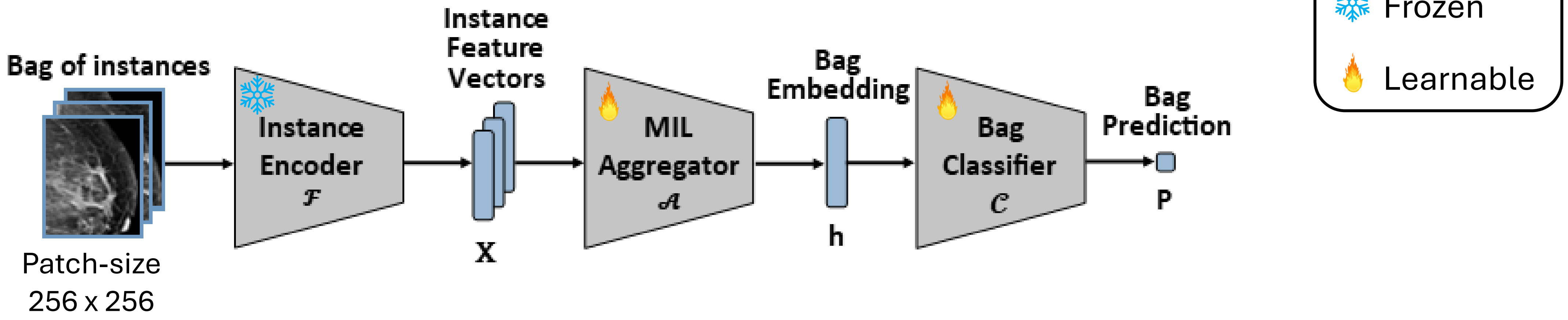
Mean Average Precision (mAP)

Lesion size categories

- Small Lesions: area $\leq 128^2$
- Medium Lesions: $128^2 <$ area $\leq 256^2$
- Large Lesions: area $> 256^2$

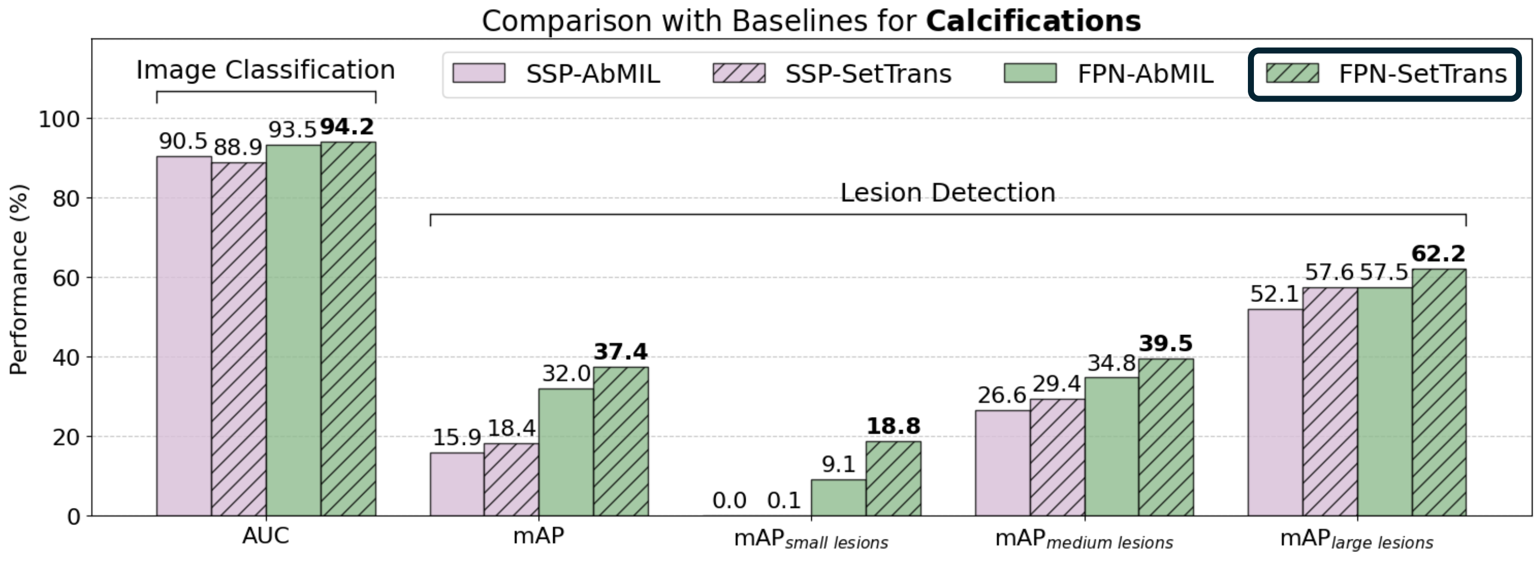
Comparison with Baselines

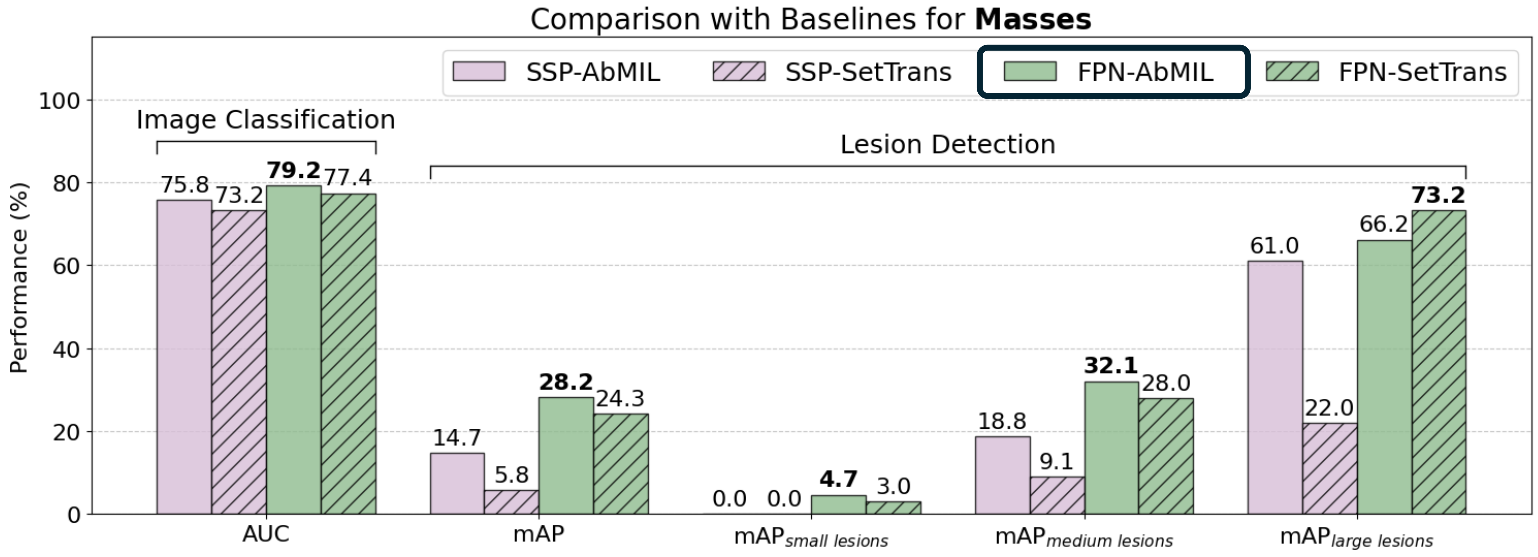
Single-Scale Patch-based (SSP)-MIL Baselines



- **Instance Encoder**: Frozen MammoCLIP [4] backbone
- **MIL Aggregator**: AbMIL[6] or SetTrans [7]

Comparison with Baselines

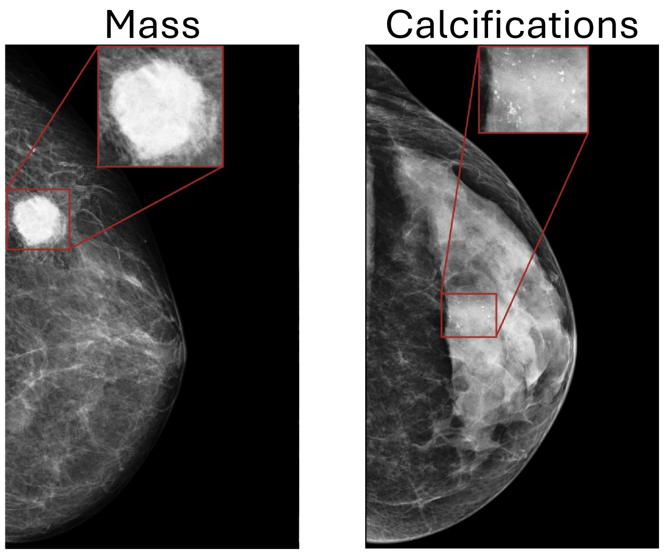




Best instance aggregator is lesion-dependent

→ SetTrans for calcifications

→ AbMIL for masses



Our FPN-MIL models significantly outperform the **SSP-MIL baselines**

Comparison with State-of-the-Art Models

Learning Paradigms	Model	Calcifications		Mass	
		AUC	mAP	AUC	mAP
Fully Supervised Classification (FSC)	EfficientNet-B2 **[4]**	92.0	---	73.0	---
Fully Supervised Object Detection (FSOD)	RetinaNet **[4]**	---	17.0	---	37.0
Weakly Supervised Object Detection (WSOD)	Mammo-FActOR **[4]**	---	20.0	---	**38.0**
Multipe Instance Learning (MIL)	**FPN-MIL (Ours)**	**94.2**	**37.4**	**79.2**	28.2

Our best-performing models...

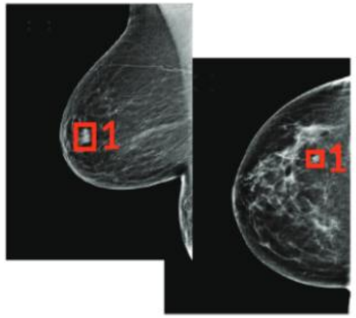
- ✓ Outperformed FSC baseline in image-level classification
- ✓ Outperformed FSOD & WSOD baselines in calcification detection
- ! Underperformed FSOD & WSOD baselines in mass detection

Weakly Supervised Object Detection

Radiology Reports
Suspicious mass in upper left...

Sentence-level annotations

Fully Supervised Object Detection



Bounding-box annotations

Multiple Instance Learning

Image-level annotations

Heatmap Visualization: Our Best-Performing Models

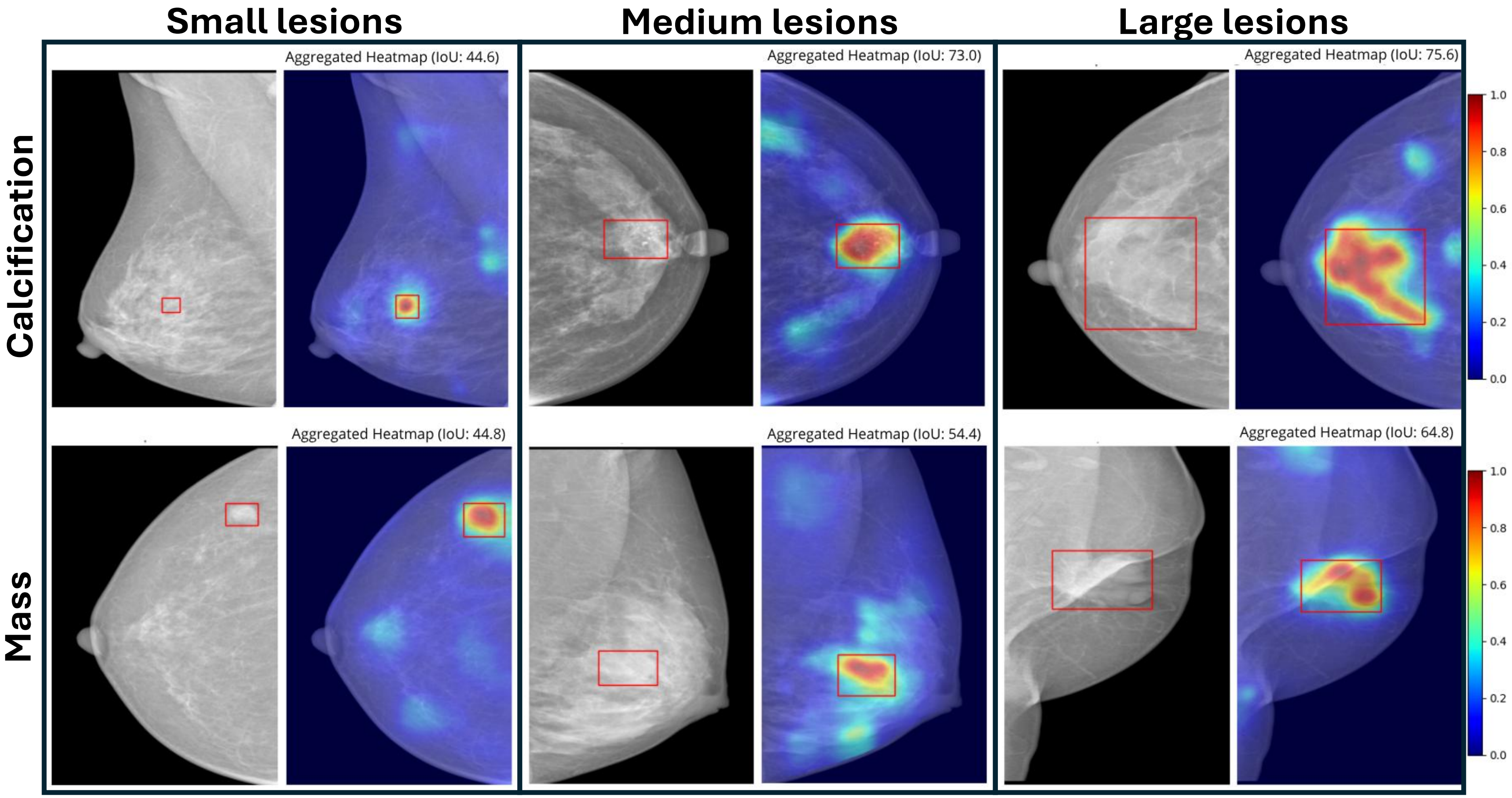
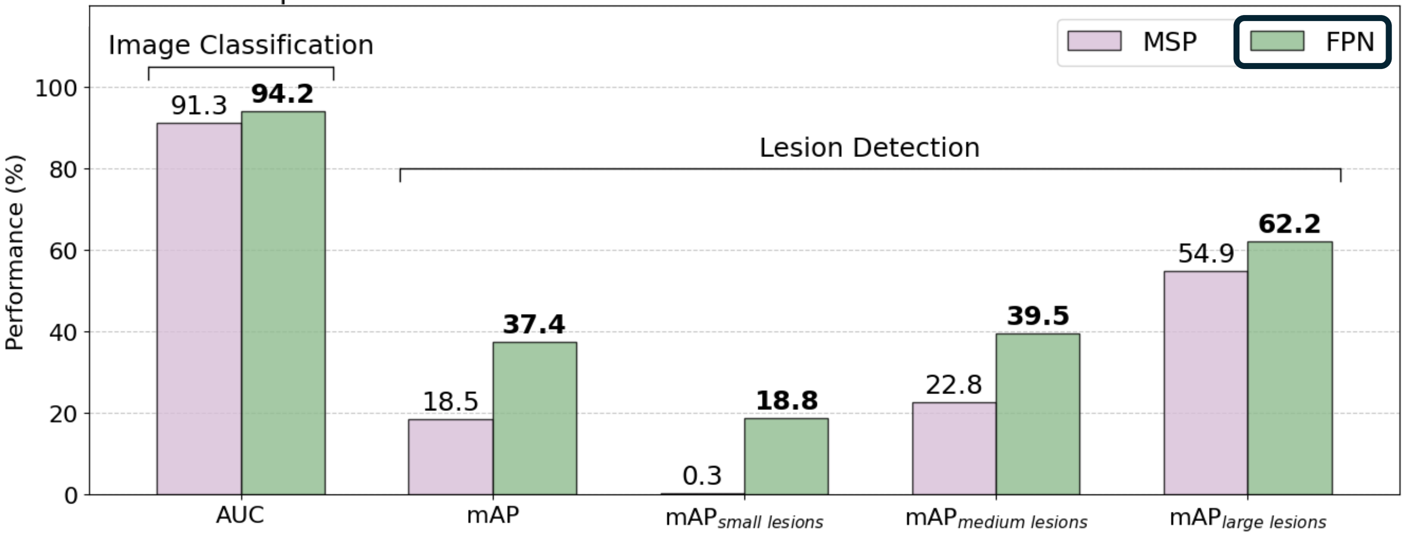


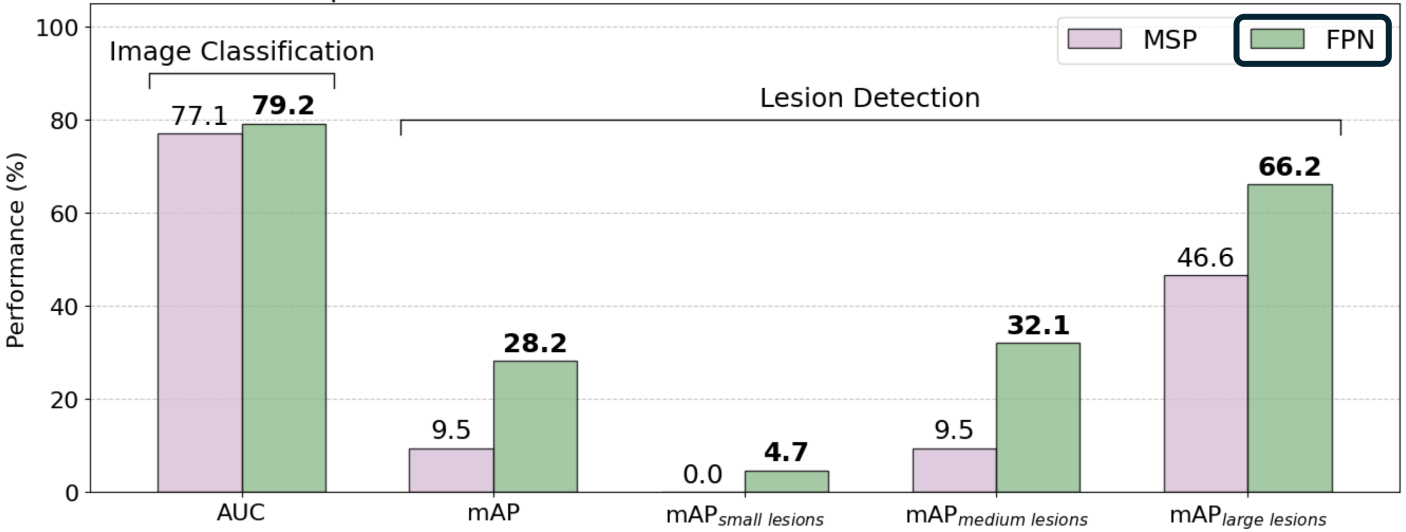
Fig. 1. Multi-scale aggregated heatmaps produced by the proposed framework, namely the FPN-SetTrans for calcifications and FPN-AbMIL for masses.

Ablation Studies

Impact of Different Multi-scale Instance Encoders for **Calcifications**



Impact of Different Multi-scale Instance Encoders for **Masses**



The proposed **FPN-based instance encoder achieves ...**

✓ Improved classification performance

⇩

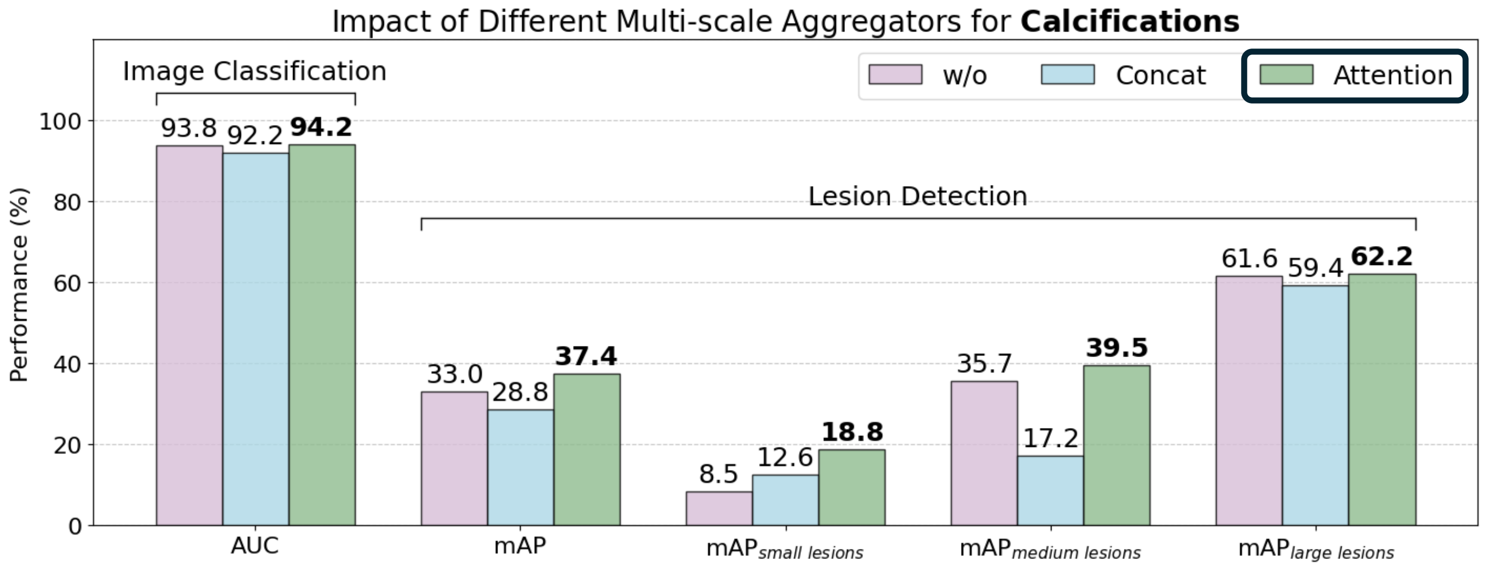
More discriminative instance features

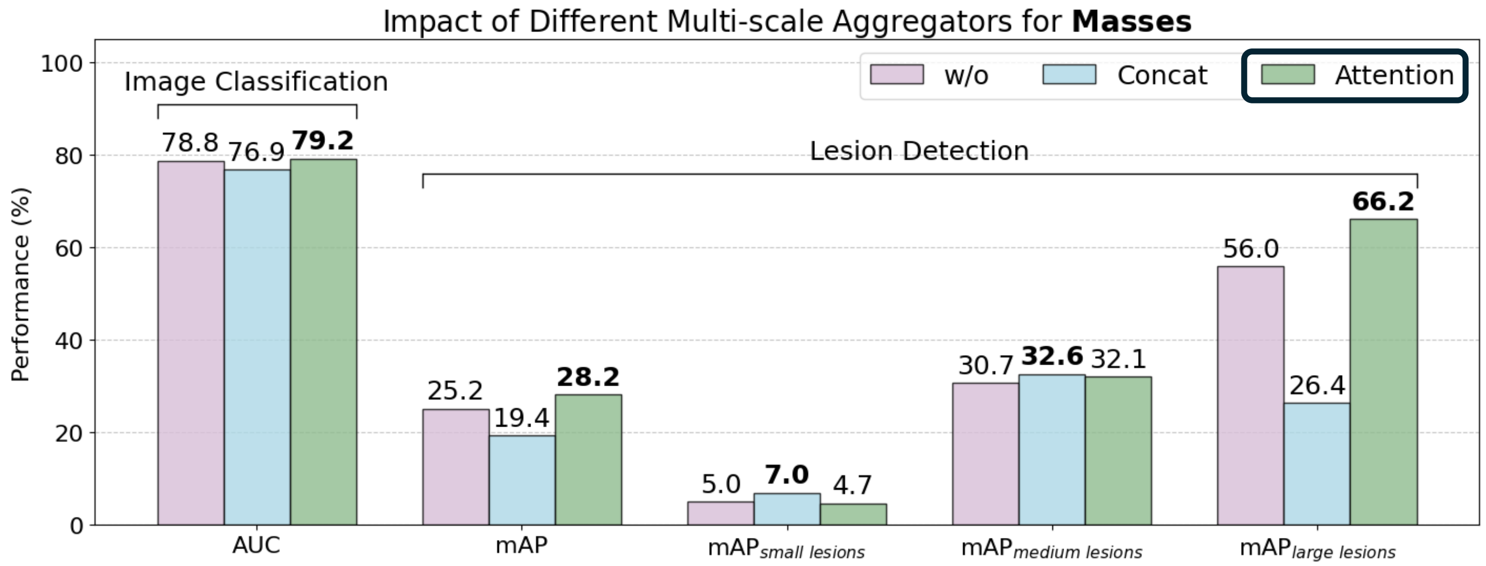
✓ Improved detection performance

⇩

Finer-grained instance features across different receptive-fields

Ablation Studies





Attention gives the best classification and detection trade-off.

- ✓ Better preserves relevant features across scales.
- ✓ Improves robustness to lesion size variability.

Conclusions & Future Work

- This work proposed a novel **multi-scale attention-based MIL framework** for weakly supervised classification and detection of breast lesions in high-resolution mammograms.
- It has a modular and adaptable design, robust across different lesion types and sizes.
- Outperformed or achieved competitive performance against baselines and SoTA models.
- Provides an extensible and strong framework for computationally and label-efficient mammographic lesion detection.

In the future:

- Investigate more instance aggregators (e.g., with positional encodings).
- Jointly analyze multi-view mammograms.

Multi-scale Attention-based Multiple Instance Learning For Breast Cancer Diagnosis

Institute for Systems and Robotics | LISBOA

Mariana Mourão

MSc in Biomedical Engineering

marianamourao@tecnico.ulisboa.pt

Paper & Code

Thank you !

Join me on Poster Session 3: **Poster C183**

Acknowledgements

LARSyS
Laboratory of Robotics and Engineering Systems

fct Fundação para a Ciência e a Tecnologia

TÉCNICO LISBOA
SIPG SIGNAL AND IMAGE PROCESSING GROUP

References

[1] Marini, N., et al.: Multi-scale task multiple instance learning for the classification of digital pathology images with global annotations. In: Proceedings of the MIC CAI Workshop on Computational Pathology. Proceedings of Machine Learning Research, vol. 156, pp. 170–181. PMLR (2021)

[2] Takeshi Yoshida, Kazuki Uehara, Hidenori Sakanashi, Hirokazu Nosato, and Masahiro Murakawa, "Multi-scale feature aggregation based mul tiple instance learning for pathological image classification," in International Conference on Pattern Recognition Applications and Methods, 2023, pp. 619–628.

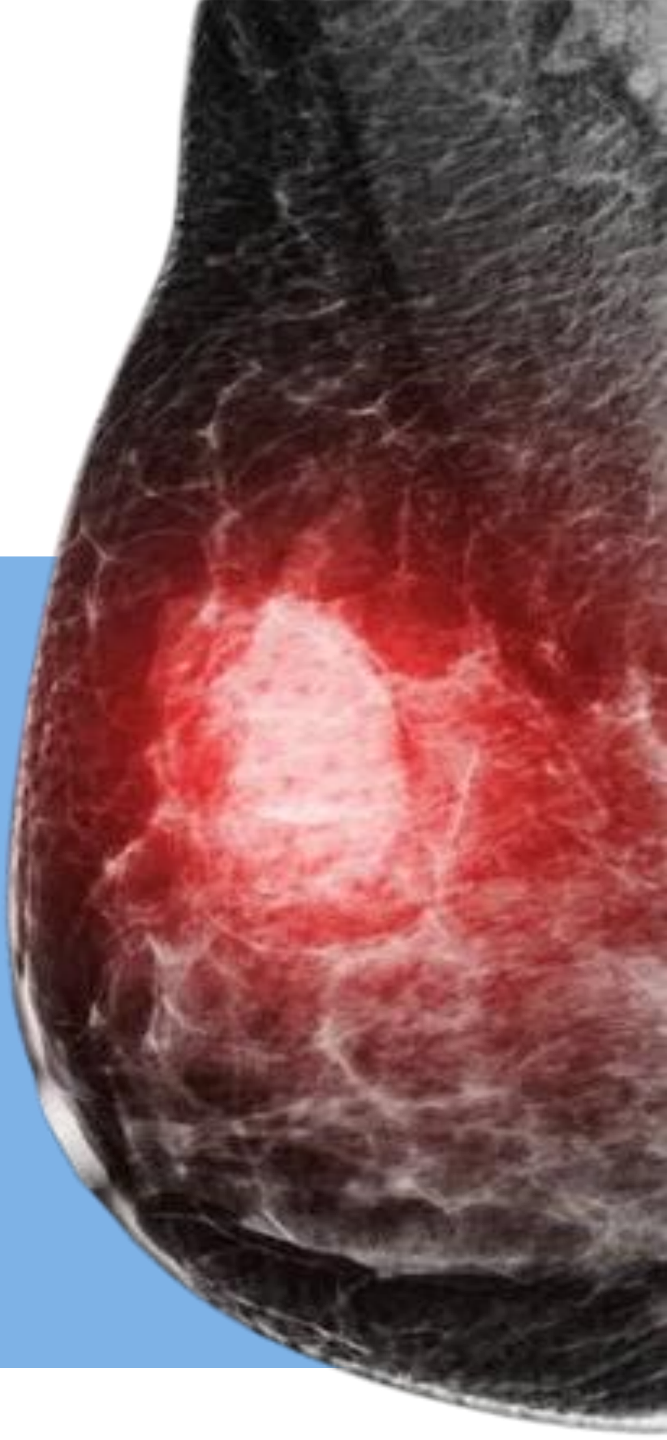
[4] Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-CLIP: a vision language foundation model to enhance data efficiency and robustness in mammography. In: Medical Image Computing and Computer Assisted Intervention– MICCAI 2024, pp. 632–642. Springer (2024)

[5] 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)

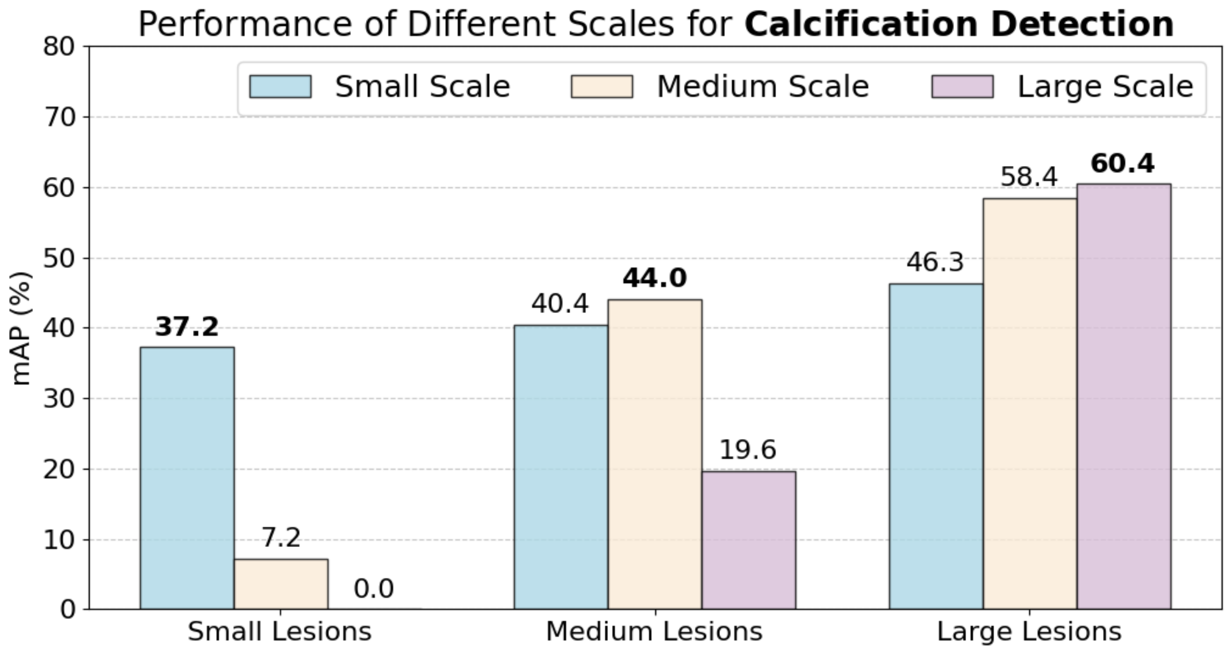
[6] Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learn ing. In: International Conference on Machine Learning (2018)

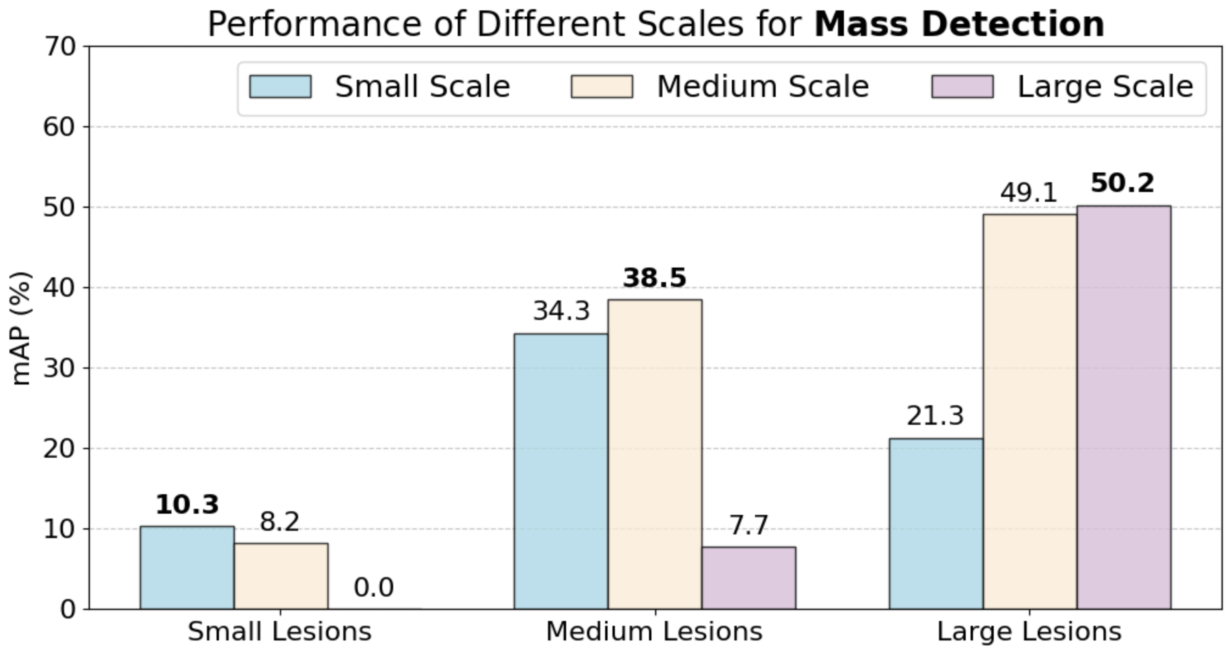
[7] Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: a framework for attention-based permutation-invariant neural networks. In: Pro ceedings of the 36th International Conference on Machine Learning, vol. 97, pp. 3744–3753. PMLR (2019)

Appendix

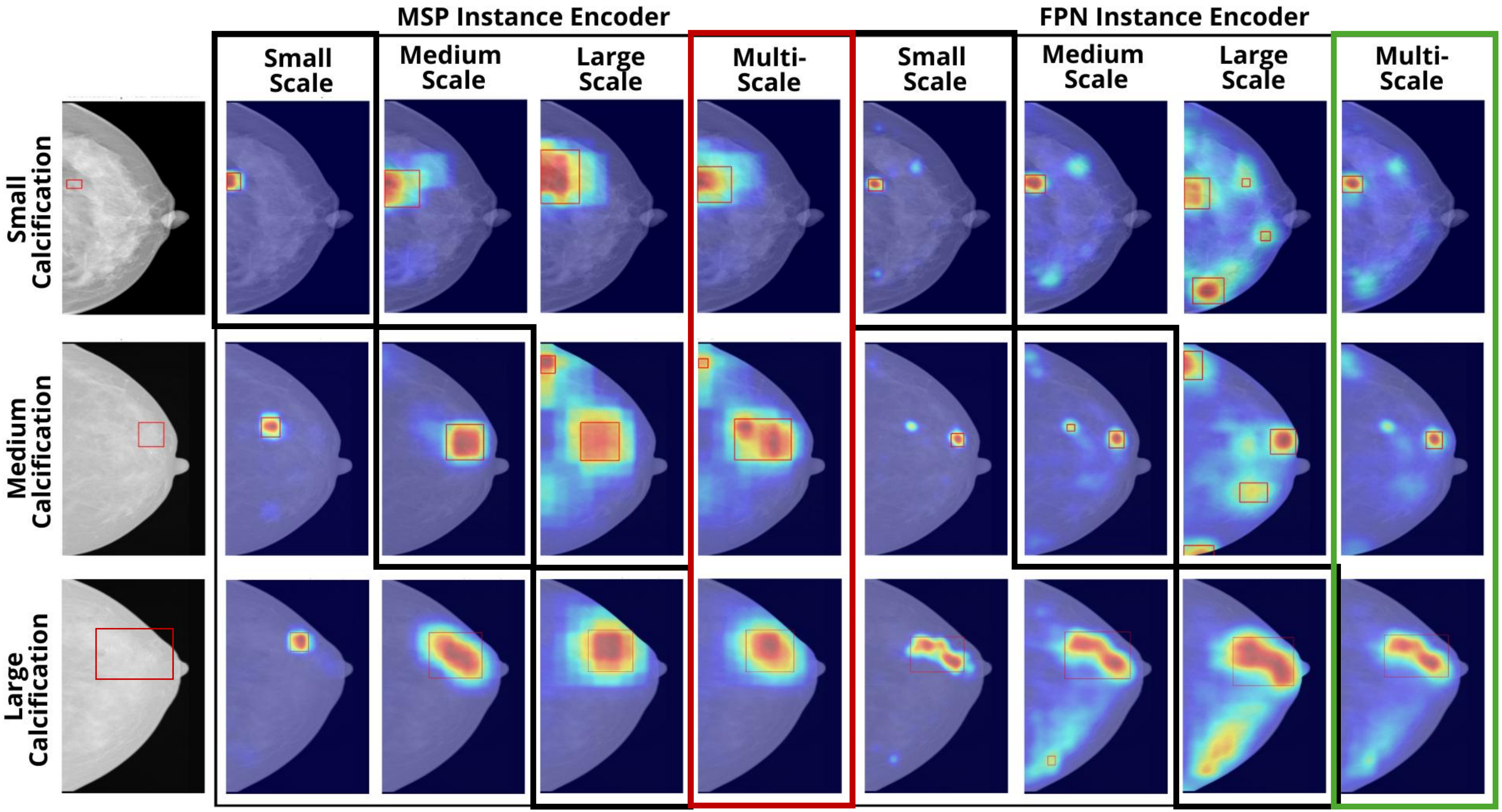


Detection Performance across Scales

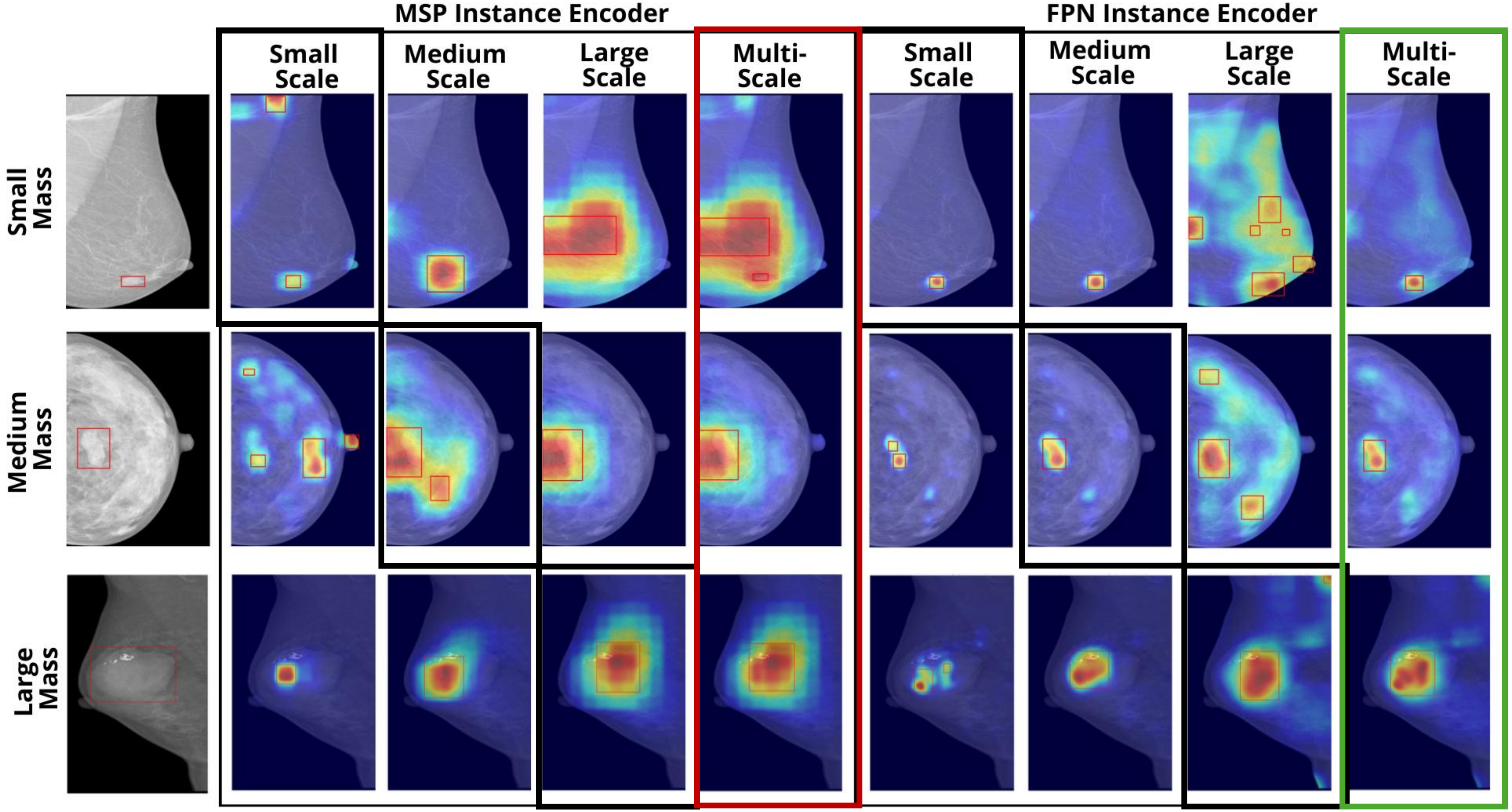




Impact of Different Multi-scale Instance Encoders



Impact of Different Multi-scale Instance Encoders



Set Transformer: Permutation Invariance Property

Institute for Systems and Robotics | LISBOA

Encoder Stage: Permutation-equivariant function

$$f(\pi(X)) = \pi(f(X))$$

Pooling Stage: Permutation-invariant function

$$g(\pi(X)) = g(X)$$

Set Transformer → Composition of functions

$$\text{Model} = g(f(X))$$

$$\text{Model}(\pi(X)) = g(f(\pi(X))) = g(\pi(f(X))) = g(f(X)) \Rightarrow \boxed{\text{Final model is pemutation-invariant}}$$

TÉCNICO LISBOA | SIPG SIGNAL AND IMAGE PROCESSING GROUP

Set Transformer: Efficient Context-aware Aggregation

Number of instances n_s and corresponding number of inducing points m_s for all analyzed scales when using scale-specific instance aggregators modeled by SetTrans in the proposed framework. The number of patches $N = 6$ extracted from the input mammograms are analyzed across three different scales $s = \{small,\ medium,\ large\}$, each associated with a specific reduction factor r_s relative to the original patch size dimensions $H_p = W_p = 512$.

Scales s	**Reduction Factor** r_s	**Number of instances** $n_s = N \times \frac{H_p}{r_s} \times \frac{W_p}{r_s}$	**Number of Inducing Points** $m_s = 10 \times \log(n_s)$
Small	16	6144	38
Medium	32	1536	32
Large	128	96	20

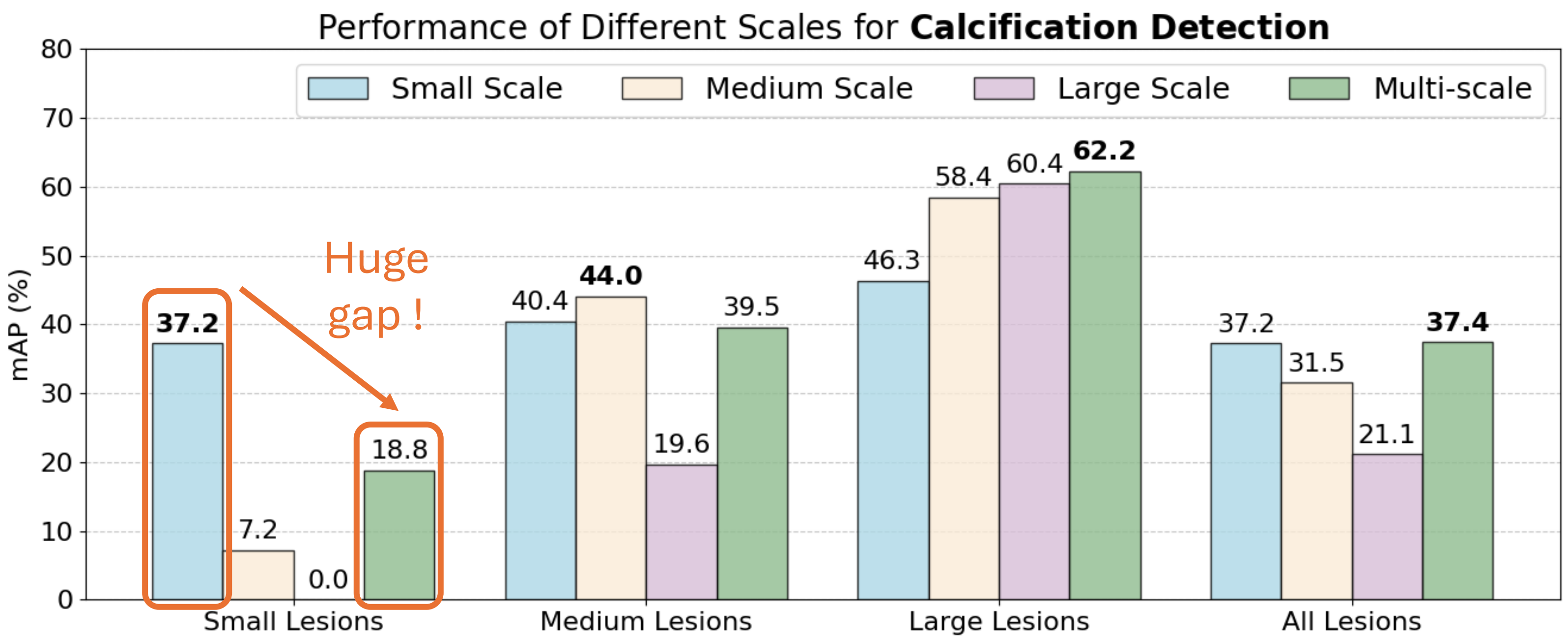
Future Work: Multi-level Instance Aggregators

Gif adapted from: https://research.google/blog/nested-hierarchical-transformer-towards-accurate-data-efficient-and-interpretable-visual-understanding/

Limitations: Multi-scale Aggregator

The **multi-scale aggregator** is optimized for MIL classification

Can learn non-optimal scale weights for the post-hoc detection analysis



Fine-grained details captured by the small-scale branch are **not fully preserved**

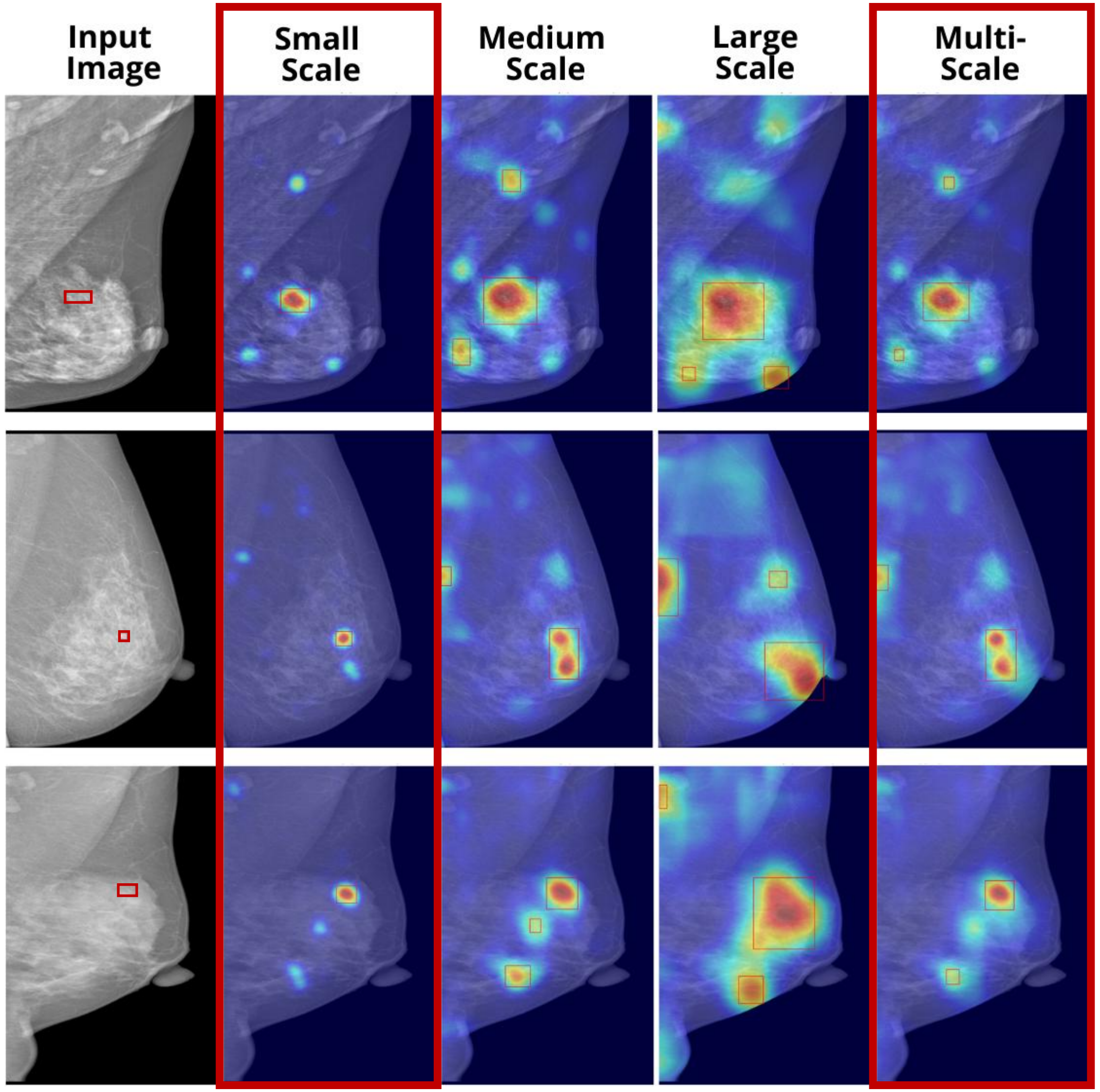
Limitations: Multi-scale Aggregator

Institute for Systems and Robotics | LISBOA

! The **multi-scale aggregator** is optimized for MIL classification

Can learn non-optimal scale weights for the post-hoc detection analysis

Fine-grained details captured by the small-scale branch are **not fully preserved**



TÉCNICO LISBOA SIPG SIGNAL AND IMAGE PROCESSING GROUP