From 2D to 3D Without Extra Baggage: Data-Efficient Cancer Detection in Digital Breast Tomosynthesis

Yen Nhi Truong Vu NHI@WHITERABBIT.AI
Dan Guo DANG@WHITERABBIT.AI
Sripad Joshi SRIPAD@WHITERABBIT.AI
Harshit Kumar HARSHIT@WHITERABBIT.AI
Jason Su JASON@WHITERABBIT.AI
Thomas Paul Matthews TOM@WHITERABBIT.AI
Whiterabbit.ai, Redwood City, CA, USA

Abstract

Digital Breast Tomosynthesis (DBT) enhances finding visibility for breast cancer detection by providing volumetric information that reduces the impact of overlapping tissues; however, limited annotated data has constrained the development of deep learning models for DBT. To address data scarcity, existing methods attempt to reuse 2D full-field digital mammography (FFDM) models by either flattening DBT volumes or processing slices individually, thus discarding volumetric information. Alternatively, 3D reasoning approaches introduce complex architectures that require more DBT training data. Tackling these drawbacks, we propose M&M-3D, an architecture that enables learnable 3D reasoning while remaining parameter-free relative to its FFDM counterpart, M&M. M&M-3D constructs malignancy-guided 3D features, and 3D reasoning is learned through repeatedly mixing these 3D features with slice-level information. This is achieved by modifying operations in M&M without adding parameters, thus enabling direct weight transfer from FFDM. Extensive experiments show that M&M-3D surpasses 2D projection and 3D slice-based methods by 11–54% for localization and 3–10% for classification. Additionally, M&M-3D outperforms complex 3D reasoning variants by 20–47% for localization and 2–10% for classification in the low-data regime, while matching their performance in high-data regime. On the popular BCS-DBT benchmark, M&M-3D outperforms previous top baseline by 4% for classification and 10% for localization.

Keywords: Mammography; Cancer Detection; Data Efficiency.

Data and Code Availability We utilize a multi-site in-house DBT dataset. This dataset was collected from 14 U.S. sites. We also evaluate on the publicly available BCS-DBT test set (Buda et al., 2021). Code is provided in Appendix D.

Institutional Review Board (IRB) The research does not require IRB approval.

1. Introduction

Breast cancer is a leading cause of cancer-related mortality among women (Ferlay et al., 2020), with early detection critical for patient outcomes (Duffy et al., 2020). Digital Breast Tomosynthesis (DBT) is increasingly adopted as an alternative to Full-Field Digital Mammography (FFDM) due to its offering of volumetric data that reduces the impact of overlapping tissues and enhances finding visibility (Friedewald et al., 2014; Richman et al., 2021). Yet, while deep learning has shown strong performance in FFDM, progress in DBT has been more limited due to two main reasons: (1) since DBT is a relatively recent technology, data remain scarce, and (2) the high cost of annotating DBT volumes further hinders the training of robust deep learning models.

A key strategy to address data limitations is to maximize the use of FFDM-pretrained models. For example, prior work reduces the 3D input to 2D (Fig. 1, left), such as via maximum intensity projection (MIP) (Samala et al., 2014), histogram matching (HM) (Singh et al., 2020), StyleGAN (Jiang et al., 2019, 2021) and maximum suspicious projection (MSP) (Lotter et al., 2021). This strategy makes z-axis localization impossible and risks 3D information loss. The most widely used alternative is to train

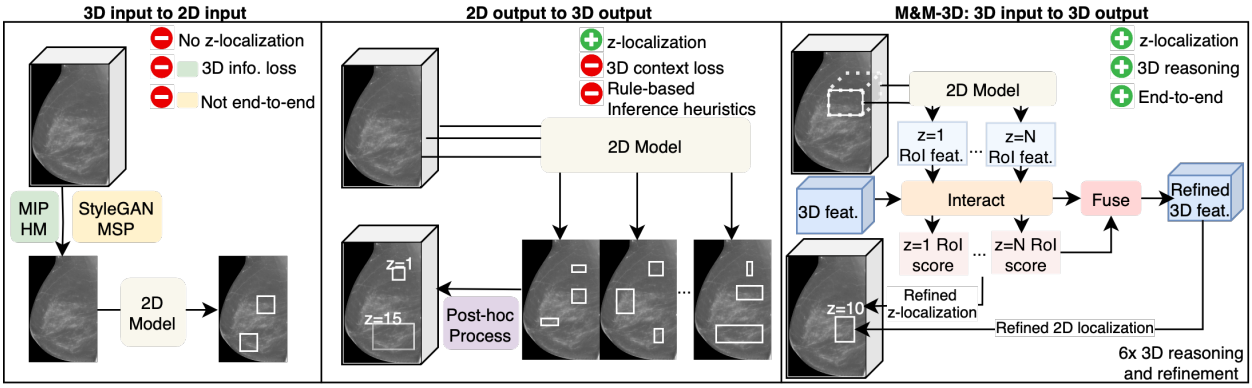

Figure 1: Common DBT approaches either (left) project the input into 2D, making z-localization impossible, or (middle) process the volume slice by slice, relying on heuristics for output aggregation. M&M-3D (right) enables seamless 3D reasoning by dynamically fusing slice-level features into 3D representations, which repeatedly interact with the slices to facilitate 3D information mixing.

with one slice at a time and then convert slice-by-slice 2D outputs into 3D outputs during inference (Fig. 1, middle) (Alberb et al., 2024; Buda et al., 2021; Fan et al., 2019; Konz et al., 2023; Lai et al., 2020; Li et al., 2021). This approach struggles to leverage the full 3D context and increases the number of candidate boxes that persist as false positives despite using post-hoc aggregation, such as Non-Maximum Suppression (NMS) (Ge and Dou, 2024; Luo et al., 2021; Yu et al., 2023). Conversely, recent methods introduce complex reasoning modules to incorporate 3D information (Lee et al., 2023; Park et al., 2023; Tardy and Mateus, 2021; Zhang et al., 2018). However, they come with more parameters not pretrained on FFDM and thus may struggle in the low-data regime.

Our work aims to tackle these limitations by designing an architecture that can perform *learnable* 3D reasoning while remaining *parameter-free* relative to its 2D counterpart, i.e., without introducing new modules that prevent direct transfer from FFDM and increase data requirements. To this end, we propose M&M-3D (Fig. 1, right), a simple yet effective extension of M&M (Truong Vu et al., 2023), a high-performing 2D mammography detector. M&M-3D fuses slice-level features into 3D features using a malignancy-driven weighting mechanism. These 3D features repeatedly interact with individual slices in the model's cascading heads using existing dynamic convolution modules. This design maintains volumetric awareness, provides implicit z-axis localization, and crucially enables 3D reasoning without adding parameters beyond the original 2D architecture. Through extensive experiments, we demonstrate two key findings:

1. M&M-3D provides learnable 3D reasoning without additional parameters. By preserving 3D information, M&M-3D outperforms 2D projection methods by 0.27 in recall at 0.25 false positives per volume (R@0.25) and 0.08 in area under the receiver operating characteristic curve (AUC). By learning 3D reasoning instead of relying on aggregation heuristics, M&M-3D outperforms slice-by-slice baselines by 0.09 in R@0.25 and 0.03 in AUC.

2. M&M-3D provides efficient 3D reasoning in low data regime. By avoiding additional 3D-specific parameters, M&M-3D maximizes the use of FFDM-pretrained weights. M&M-3D matches performance of complex 3D variants at 100% data and outperforms them by 0.10–0.20 in R@0.25 and 0.02–0.09 in AUC at 10% data.

Finally, we also evaluate M&M-3D on the widely used BCS-DBT benchmark (Buda et al., 2021). M&M-3D demonstrates excellent generalizability, achieving 0.97 AUC and 0.83 R@0.25, the highest results reported to date (Appendix A).

2. Prerequisites

Sparse R-CNN. Sun et al. (2021) proposed Sparse-RCNN, which utilizes a sparse set of N proposals parametrized by boxes $\mathbf{b}_0 \in \mathbb{R}^{N\times 4}$ and corresponding features $\mathbf{h}_0 \in \mathbb{R}^{N\times D}$, with D the hidden dimension. These proposals are refined through 6

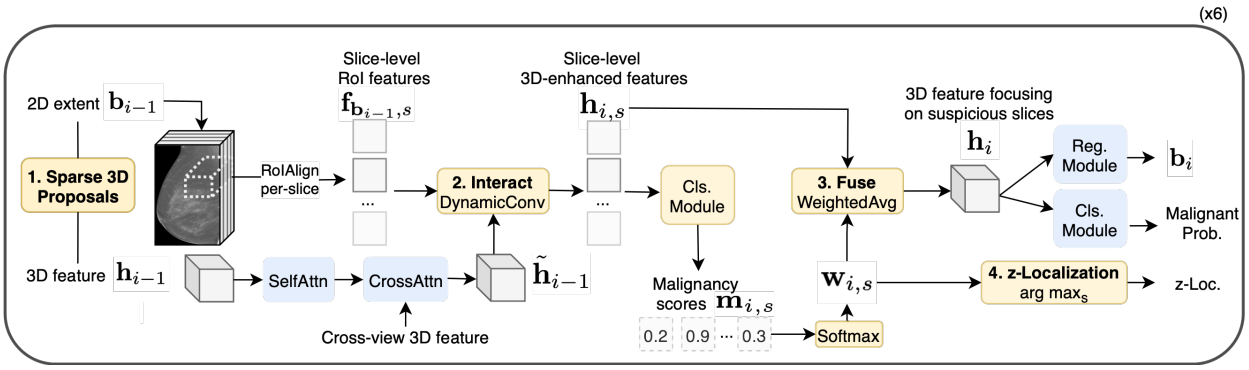

Figure 2: M&M-3D extends M&M (blue, section 2) with parameter-free 3D reasoning (yellow, section 3). 3D proposals, parameterized by 3D features $\mathbf{h}_{i-1}$ and 2D extent $\mathbf{b}_{i-1}$ spanning all slices s, are refined using 6 cascade heads ($1 \leq i \leq 6$). These 3D features interact with 2D RoI features $\mathbf{f}_{\mathbf{b}_{i-1},s}$ across slices, producing slice-level object features $\mathbf{h}_{i,s}$ enhanced with 3D context. The classification (Cls.) module is reused to produce finding-slice scores $\mathbf{w}_{i,s}$, which are used to fuse $\mathbf{h}_{i,s}$ into refined 3D features $\mathbf{h}_i$ focusing on the most suspicious slices. z-axis localization is obtained as the slice with maximum score, i.e., $\arg\max_s \mathbf{w}_{i,s}$.

cascading heads, where the i^{th} head refines $\mathbf{h}_{i-1}$ into new features $\mathbf{h}_i$ using two key operations

$$\mathbf{h}'_{i-1} = \text{SelfAttn}(\mathbf{h}_{i-1}), \quad (1)$$
$$\mathbf{h}_i = \text{DynamicConv}(\mathbf{h}'_{i-1}, \mathbf{f}_{\mathbf{b}_{i-1}}), \quad (2)$$

where $\mathbf{f}_{\mathbf{b}_{i-1}} \in \mathbb{R}^{N \times k^2 \times D}$ are region-of-interest (RoI) features pooled from boxes $\mathbf{b}_{i-1}$ and k is the pool size. Self-attention (SelfAttn) enables global feature mixing across proposals. Dynamic convolutions (DynamicConv) update $\mathbf{f}_{\mathbf{b}_{i-1}}$ using kernels dynamically generated by $\mathbf{h}'_{i-1}$, thus conditioning each proposal on its corresponding RoI features. A regression module is applied to $\mathbf{h}_i$ to obtain new boxes $\mathbf{b}_i$, completing the iterative refinement.

M&M. Truong Vu et al. (2023) proposed M&M, which adapts Sparse R-CNN to 2D mammography, where each breast is imaged under two complementary views: cranio-caudal (CC) and medio-lateral oblique (MLO). Let CrossAttn($\mathbf{q}, \mathbf{k}$) denote cross-attention from query $\mathbf{q}$ to key/value $\mathbf{k}$. Let $\mathbf{h}'^{a}_{i-1}$ denote the alternative view's post-SelfAttn feature. M&M introduces multi-view attention to mix information from the alternative view as follows:

$$\tilde{\mathbf{h}}_{i-1} = \text{CrossAttn}(\mathbf{h}'_{i-1}, \mathbf{h}'^{a}_{i-1}), \quad (3)$$
$$\mathbf{h}_i = \text{DynamicConv}(\tilde{\mathbf{h}}_{i-1}, \mathbf{f}_{\mathbf{b}_{i-1}}). \quad (4)$$

M&M also employs a multi-instance learning (MIL) formulation. Proposal malignancy scores $\mathbf{m}_i$ are aggregated by noisy-or pooling to yield image scores, and mean pooling to yield breast scores. Finding-level losses are disabled if annotations are absent.

3. M&M-3D Architecture

Our goal is to achieve learnable 3D reasoning, but without adding any additional learnable parameters. This design decision will allow us to maximize the pre-trained weights from the FFDM model. To implement such a design, we propose the following parameter free extensions of M&M (Fig. 2):

1. Sparse 3D proposals. Consider a DBT volume with S slices. A naive slice-by-slice application of M&M would yield an unstructured set of $N \times S$ 2D proposals. These 2D proposals are independent across slices, which complicates the learning process and necessitates post-hoc aggregation to consolidate 2D detections into 3D objects (Fig. 1, middle). To enforce spatial coherence, we reinterpret the initial proposals as *3D primitives*: in M&M-3D, $\mathbf{h}_0 \in \mathbb{R}^{N \times D}$ represents features associated with 3D proposals with 2D extents $\mathbf{b}_0 \in \mathbb{R}^{N \times 4}$ and depths that span the entirety of the DBT volume. We retain 6 cascading heads, where the i^{th} head incorporates slice features to refine the previous 3D feature $\mathbf{h}_{i-1}$ and box $\mathbf{b}_{i-1}$ into better 3D feature $\mathbf{h}_i$ and box $\mathbf{b}_i$. With this new interpretation, M&M-3D preserves a sparse set of N 3D proposals throughout refinement.

2. Slice-level feature interaction. During refinement, we reuse SelfAttn (Eq. 1) and CrossAttn

(Eq. 3) to obtain $\tilde{\mathbf{h}}_{i-1}$, a 3D proposal feature enhanced with information from other proposals across both views. DynamicConv (Eq. 2) now convolves slice-level features using kernels dynamically generated by 3D feature $\tilde{\mathbf{h}}_{i-1}$, effectively allowing for 2D and 3D feature interaction.

Concretely, in the 2D setting, DynamicConv (Eq. 2) operates between proposal features $\tilde{\mathbf{h}}_{i-1}$ and RoI features $\mathbf{f}_{\mathbf{b}_{i-1}} \in \mathbb{R}^{N\times k^2\times D}$. For DBT, we apply RoIAlign across all slices $1 \le s \le S$, obtaining slice-specific features $\mathbf{f}_{\mathbf{b}_{i-1},s}$. We then reuse the same dynamic convolution module as follows

$$\mathbf{h}_{i,s} = \text{DynamicConv}(\tilde{\mathbf{h}}_{i-1}, \mathbf{f}_{\mathbf{b}_{i-1},s}). \tag{5}$$

Here, $\mathbf{h}_{i,s} \in \mathbb{R}^{N\times D}$ represents 3D-enhanced slice-level features: while remaining specific to slice s, $\mathbf{h}_{i,s}$ now contains 3D context carried over from $\tilde{\mathbf{h}}_{i-1}$. This design is both simple and effective—by reusing dynamic convolutions, M&M-3D seamlessly adapts to 3D reasoning without adding parameters. More importantly, since this interaction occurs in every head, proposals repeatedly mix volumetric information, ensuring progressive refinement of 3D spatial cues.

3. Slice-to-volume feature fusion. The vector $[\mathbf{h}_{i,1}, \cdots, \mathbf{h}_{i,S}]$ contains enhanced features of the region $\mathbf{b}_{i-1}$ across all slices; thus, it is a natural candidate from which to derive $\mathbf{h}_i \in \mathbb{R}^{N\times D}$, the refined 3D representation. To this end, we propose a simple weighted average (Wgt. Avg.) fusion module. First, we pass $\mathbf{h}_{i,s}$ through the classification module of M&M to obtain malignancy scores of each finding across slices $\mathbf{m}_{i,s}$. We then compute

$$\mathbf{h}_i = \sum_{s=1}^{S} \mathbf{w}_{i,s} \odot \mathbf{h}_{i,s}, \qquad \mathbf{w}_{i,s} = \frac{e^{\mathbf{m}_{i,s}}}{\sum_{s'=1}^{S} e^{\mathbf{m}_{i,s'}}} \in \mathbb{R}^N, \tag{6}$$

where $\odot$ denotes element-wise multiplication. The refined vector $\mathbf{h}_i$ is a genuine 3D proposal feature: it contains information from all slices, where slices with higher suspicion contribute more strongly to the representation through their larger weights. Furthermore, for different proposal $n \in \{1, \ldots, N\}$, the weights across slices $\mathbf{w}_{i,\cdot}[n]$ are different, allowing each proposal to specialize in the subset of slices most relevant to its RoI.

4. z-axis localization. In clinical practice, radiologists primarily make callback and diagnostic decisions based on the slice where the finding is most clearly visible (see Appendix C). Following this workflow, our design focuses supervision on the most suspicious slice rather than the visibility range of findings. Concretely, the scores $\{\mathbf{w}_{i,s}\}_{s=1}^{S}$ encode finding malignancy across slices, implicitly providing z-axis localization. In particular, $\mathbf{z}_i = \arg\max_s \mathbf{w}_{i,s} \in \mathbb{R}^N$ is taken as the most suspicious slice for the findings. Given a proposal m matching a ground truth finding, if annotation of the most suspicious slice z is available, we can apply the cross entropy loss $-\log(\mathbf{w}_{i,z}[m])$. We do not apply losses on the remaining slices $\mathbf{w}_{i,s}$ with $s \neq z$ since it is not known whether the finding is visible in such slice. Thus, the total training loss for M&M-3D is

$$\mathcal{L} = \mathcal{L}_{\text{M\&M}} - \mathbf{1}_{\text{annotated finding}} \sum_{i=1}^{6} \log(\mathbf{w}_{i,z}[m]), \tag{7}$$

where $\mathcal{L}_{\text{M\&M}}$ contains finding-level 2D localization loss and image and breast-level classification loss.

Table 1: The number of malignant, benign and negative breasts in each dataset split. Last column reports the number of breasts with finding annotations.

Split	Malignant	Benign	Negative	Annotated
Train	2311	8524	10266	955
Val	302	1100	1489	246
Test	664	3462	2677	204

4. Experiments

M&M-3D avoids new parameters to fully leverage FFDM-pretraining. We compare against parameter-free baselines (§ 4.2), evaluate data efficiency, including against parameter-heavy variants (§ 4.3), and present ablation studies (§ 4.4).

4.1. Experiment details

Implementation details. To reduce memory, we crop the background region and resize the image to a width of 1100 and max length of 2200. We also downsample the z-resolution to 16 using MaxPool, corresponding to a resizing factor of 1.25–8.50. During inference, the z-coordinate of each prediction is mapped back by selecting the middle slice of the corresponding downsampled region. Horizontal and vertical flip augmentations are applied during training. We use AdamW optimizer with 2.5×10^{-5} learning

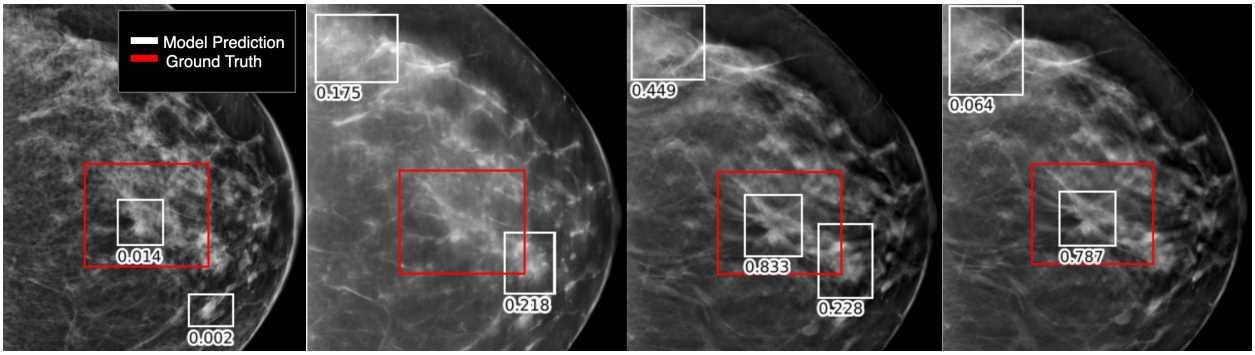

Figure 3: Qualitative results. From left to right: (1) **FFDM**: The finding is obscured by surrounding tissue, leading to low detection scores. (2) **MIP** still suffers from occlusion and introduces false positives. (3) **Buda*** assigns a high score to the correct finding on the most suspicious slice but also increases the number of false positives. (4) **M&M-3D** successfully assigns a high score to the malignant finding on the most suspicious slice while maintaining low scores in other regions.

rate and 0.0001 weight decay. Unless stated otherwise, all models are initialized with FFDM pretrained weights shared by authors of M&M.

Dataset. We utilize a multi-site in-house DBT dataset. This dataset was collected from 14 sites and was split into training, validation, and test sets at the site level. Tab. 1 summarizes dataset statistics for each split. Findings are labeled with bounding boxes on the most visible slice, and for the test split, their visibility range across slices is also annotated.

Metric. We report recall at X false postive per volume (R@X) with $X \in \{0.25, 0.5\}$. A proposal is a true positive if the intersection over union (IoU) of the 2D proposal and 2D ground truth is at least 0.25 and the predicted slice is within the visibility range. For classification, we report area under the receiver operating characteristic curve (AUC). Standard errors are bootstrapped for R@0.25 and R@0.5, and computed via DeLong et al. (1988) for AUC.

4.2. M&M-3D provides 3D reasoning without additional parameters

An important design motivation for M&M-3D is to maximize transfering ability from FFDM to DBT, thus we do not introduce any additional parameters. This allows the model to operate even without training on DBT data. In this section, we evaluate M&M-3D against other parameter-free methods and show that it provides meaningful 3D reasoning where prior approaches fall short.

Parameter-free baselines. We compare against three representative baselines:

1. 2D projection methods. Maximum Intensity Projection (MIP) (Samala et al., 2014) and Histogram Matching (HM) (Singh et al., 2020) collapse the DBT volume into a single FFDM-like image.

2. Slice-based aggregation methods. The most common DBT strategy trains slice-level models, then aggregates outputs during inference (Alberb et al., 2024; Buda et al., 2021; Fan et al., 2019; Konz et al., 2023; Lai et al., 2020; Li et al., 2021). This approach relies on sampling positive slices during training. Since 60% of our malignant volumes are unannotated (Tab. 1), we assess two versions of the representative baseline by Buda et al. (2021): (1) *Buda*, which samples a random slice when annotations are unavailable, risking noisy supervision, and (2) *Buda**, which discards all unannotated malignant volumes, drastically reducing training data.

3. FFDM baseline. Because our dataset provides paired DBT and FFDM images, we also report M&M performance when evaluated directly on FFDM. This acts as a non-transfer benchmark, reflecting performance without cross-modality adaptation, and serves as reference for evaluating transfer to DBT.

Inference-only performance (Tab. 2). Projection baselines perform poorly because collapsing volumes into 2D images discards depth information and amplifies occlusion (Fig. 3, second column). By preserving depth information, M&M-3D improves over MIP/HM by 0.20–0.39 R@0.25 and 0.08–0.20 AUC.

Table 2: Inference-only performance.

Method	Input	R@0.25	R@0.5	AUC
2D localization (2D IoU $\geq$ 0.25)				
M&M	FFDM	**0.61**$_{\pm 0.025}$	**0.67**$_{\pm 0.025}$	**0.89**$_{\pm 0.009}$
MIP	Proj.	0.14$_{\pm 0.018}$	0.16$_{\pm 0.019}$	0.65$_{\pm 0.012}$
HM	Proj.	0.33$_{\pm 0.024}$	0.41$_{\pm 0.026}$	0.77$_{\pm 0.010}$
M&M-3D	Vol.	0.53$_{\pm 0.025}$	0.57$_{\pm 0.025}$	0.85$_{\pm 0.008}$
3D localization (2D IoU $\geq$ 0.25 + slice in visible range)				
Buda	Vol.	0.45$_{\pm 0.026}$	**0.49**$_{\pm 0.026}$	0.84$_{\pm 0.009}$
M&M-3D	Vol.	**0.46**$_{\pm 0.026}$	0.48$_{\pm 0.026}$	**0.85**$_{\pm 0.008}$

Table 3: Fine-tuned performance. (*) models are trained without unannotated malignant volumes.

Method	Input	R@0.25	R@0.5	AUC
2D localization (2D IoU $\geq$ 0.25)				
M&M	FFDM	0.62$_{\pm 0.027}$	0.69$_{\pm 0.026}$	0.89$_{\pm 0.008}$
MIP	Proj.	0.52$_{\pm 0.026}$	0.60$_{\pm 0.025}$	0.86$_{\pm 0.008}$
HM	Proj.	0.53$_{\pm 0.027}$	0.60$_{\pm 0.026}$	0.87$_{\pm 0.008}$
M&M-3D	Vol.	**0.80**$_{\pm 0.020}$	**0.84**$_{\pm 0.019}$	**0.95**$_{\pm 0.004}$
3D localization (2D IoU $\geq$ 0.25 + slice in visible range)				
Buda*	Vol.	0.68$_{\pm 0.025}$	0.74$_{\pm 0.023}$	0.92$_{\pm 0.006}$
Buda	Vol.	0.68$_{\pm 0.023}$	0.73$_{\pm 0.023}$	0.92$_{\pm 0.006}$
M&M-3D*	Vol.	0.73$_{\pm 0.023}$	0.79$_{\pm 0.021}$	0.94$_{\pm 0.005}$
M&M-3D	Vol.	**0.77**$_{\pm 0.021}$	**0.82**$_{\pm 0.019}$	**0.95**$_{\pm 0.004}$

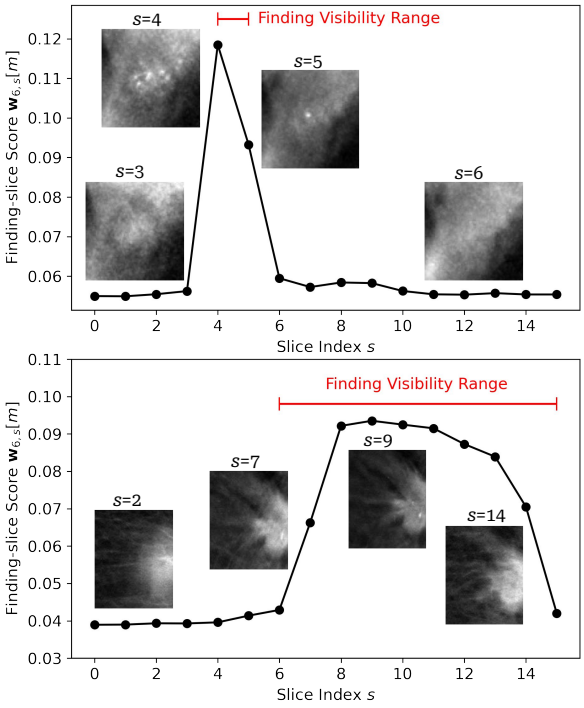

Figure 4: Finding-slice scores produced by M&M-3D. For a malignant finding, we identify the highest scoring proposal m that matches it, and plot the scores $\mathbf{w}_{6,s}[m]$. The red bar denotes the range of slices where the finding is visible based on ground truth annotations. Top: a small cluster of calcifications visible in only 2/16 slices, yielding a sharp localized peak in $\mathbf{w}_{6,s}[m]$. Bottom: a mass visible on 10/16 slices, yielding a broad span of elevated $\mathbf{w}_{6,s}[m]$ values. Insets show representative slices.

For 3D localization, M&M-3D performs on par with Buda (R@0.25: 0.46 vs. 0.45, AUC: 0.85 vs. 0.84), indicating that even without finetuning, M&M-3D's built-in 3D reasoning capability is already as effective as post-hoc aggregation methods (Tab. 2). While the FFDM baseline remains the strongest in this regime, M&M-3D substantially closes the gap, demonstrating its ability to transfer from FFDM to DBT.

Fine-tuning performance (Tab. 3). With supervision, the advantages of explicit 3D reasoning become more pronounced. M&M-3D surpasses the FFDM model by 0.18 R@0.25 and 0.06 AUC, reflecting its ability to exploit lesions that are visible in DBT slices but obscured in FFDM projections (Fig. 3, left). It also outperforms projection baselines by $\sim$0.25 in recall and $\sim$0.09 in AUC, demonstrating that simple 2D proxies cannot capture volumetric detail. Against slice-based methods, M&M-3D outperforms both Buda variants by about 0.10 in R@0.25, R@0.5 and by 0.03 in AUC. Even when evaluated under the same data constraints, M&M-3D* remains superior to Buda*, highlighting the advantage of learned 3D reasoning over heuristic aggregation. Notably, Buda gains no benefit over Buda* from including unannotated malignant volumes due to the noise introduced by random slice sampling. Meanwhile, M&M-3D significantly outperforms M&M-3D*, proving its effective use of unannotated data. The key advantage of M&M-3D is evident in Fig. 3: while slice-based models introduce more false positives during aggregation, M&M-3D leverages learned 3D reasoning to effectively filter out spurious detections.

Interpretable z-axis localization. Fig. 4 illustrates that the finding-slice scores $\mathbf{w}_{i,s}$ provide an interpretable signal of how the model localizes findings along the z-axis. In particular, the scores align closely with finding's visibility in DBT slices. In the top example, a small cluster of calcifications is visible in only 2 of 16 slices, and the score distribution exhibits a sharp, localized peak precisely within this range. In the bottom example, a mass extends across

10 of 16 slices, and the score profile remains elevated across the full span before tapering off as the lesion disappears. These patterns demonstrate that M&M-3D adapts its scoring to the lesion's extent rather than relying on fixed aggregation heuristics.

Generalizability. To assess generalizability, we evaluate M&M-3D on the BCS-DBT test set (Buda et al., 2021). Without training on DBT, M&M-3D achieves 0.91 AUC and 0.56 R@0.5, comparable to top performing methods (Konz et al., 2023; Terrassin et al., 2024). After finetuning, M&M-3D achieves 0.97 AUC, 0.83 R@0.25 and 0.85 R@0.5. To our knowledge, these are the highest performance results on this dataset (Appendix A).

4.3. M&M-3D provides efficient 3D reasoning in low data regime

Given that DBT model development is hindered by scarce data and costly annotation, we assess the data efficiency of M&M-3D in two ways: (1) by reducing the fraction of training data to simulate limited data availability, and (2) by reducing the proportion of annotated data while keeping the dataset size fixed at 100% to simulate limited annotation availability. When varying data size, we report both detection (R@0.25) and classification metrics (AUC). When varying annotation size, we discard only box annotations while keeping all images available. This keeps classification performance (AUC) stable across models, so we report only detection performance.

Complex 3D reasoning baselines. We benchmark M&M-3D against alternative designs that incorporate more sophisticated trainable components for 3D reasoning. These baselines have the potential to capture richer 3D representations, but at the cost of being more data-hungry.

1. TimeSform. Following Lee et al. (2023), *TimeSform* performs joint spatial and depth attention on slice-level RoI features $\mathbf{f}_{\mathbf{b}_{i-1},s}$ (Bertasius et al., 2021). Attention is applied to fuse RoI features across slices:

$$\mathbf{f}_{\mathbf{b}_{i-1}} = \mathrm{Attn}\big(\{\mathbf{f}_{\mathbf{b}_{i-1},s}\}_{s=1}^{S}\big)\,, \tag{8}$$

where $\mathbf{f}_{\mathbf{b}_{i-1}} \in \mathbb{R}^{N\times k^2\times D}$ encodes volumetric RoI context. From here, DynamicConv can be applied directly as in Eq. 4. z-axis localization is derived from the attention weights of Eq. 8.

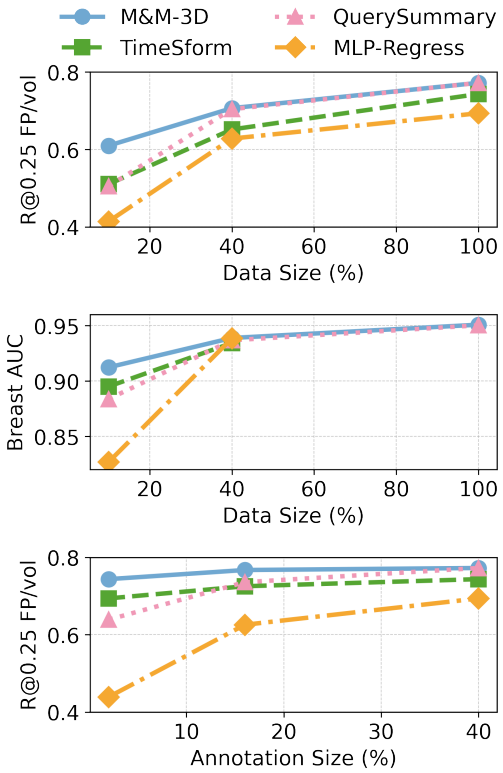

Figure 5: Comparison of M&M-3D with complex 3D reasoning variants. M&M-3D performs similarly to all alternatives in high-data regime but outperforms them significantly in low-data regime, illustrating its data efficiency. See Appendix B for figure data.

2. QuerySummary utilizes learnable queries that act as adaptive pooling operators (Devlin et al., 2019) to perform slice-to-volume feature fusion:

$$\mathbf{h}_i = \mathrm{Attn}\big(\mathbf{q}_i, \{\mathbf{h}_{i,s}\}_{s=1}^{S}\big)\,, \tag{9}$$

where $\mathbf{q}_i \in \mathbb{R}^{N\times D}$ are learnable query embeddings. Similar to M&M-3D, the attention weights from Eq. 9 are reused as finding–slice scores $\mathbf{w}_i$.

3. MLP-Regress replaces our malignancy-guided weighted average with an MLP-based summarization:

$$\mathbf{h}_i = \mathrm{MLP}\big(\{\mathbf{h}_{i,s}\}_{s=1}^{S}\big)\,, \tag{10}$$

and extends the regression module to explicitly predict central slices $\mathbf{z}_i = \mathrm{Regress}(\mathbf{h}_i)$.

Performance under limited data. As shown in Fig. 5, M&M-3D consistently outperforms all complex variants in low-data regimes. With only 10%

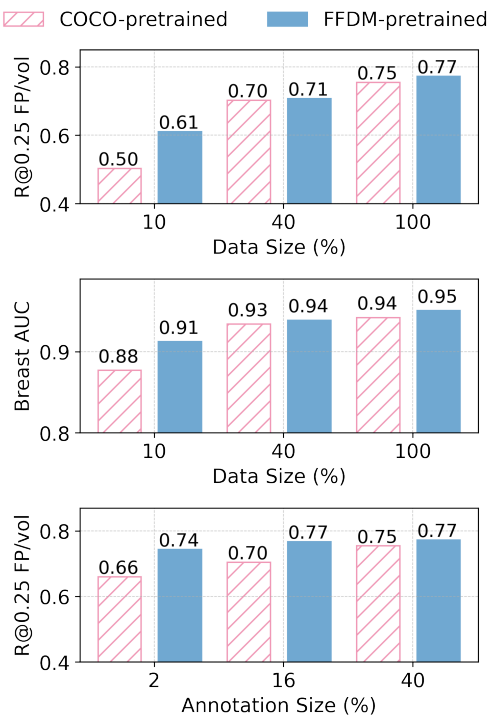

Figure 6: M&M-3D with FFDM vs. COCO pretrained weights. FFDM-pretrained models outperform COCO-pretrained models in detection and classification across all dataset and annotation sizes, with the largest gaps in low-data regimes.

of training data (∼230 malignant breasts), M&M-3D achieves at least 0.10 gain in R@0.25 and 0.02 in AUC compared to all alternatives. Notably, in this extreme setting, M&M-3D is the only method to surpass the FFDM-only M&M baseline (Tab. 3, row 1). This highlights M&M-3D's superior data efficiency, making it an ideal DBT solution for institutions with large FFDM datasets but limited DBT data.

Performance under limited annotations. Fig. 5, last row illustrates detection performance when varying the fraction of annotated cases while fixing the dataset size. At maximum, 40% of malignant cases in our training set is annotated (Tab. 1), and we further simulate settings with only 2% and 16% annotated cases. Again, we observe larger performance gap when annotations are more limited: at 2% annotations (∼50 breasts), M&M-3D achieves 0.74 R@0.25, while the strongest baseline trails at 0.70. M&M-3D's ability to succeed with minimal annotations is especially beneficial for institutions that may already possess DBT data but lack the resources or expertise to annotate them extensively.

Table 4: Effect of z-resolution on M&M-3D. **Mem.**: training GPU memory required. **Time**: wall clock training time.

z-Res.	Mem. (GB)	Time (h)	R@0.25	R@0.5	AUC
8	42	11	$0.69_{\pm 0.022}$	$0.76_{\pm 0.020}$	$0.94_{\pm 0.005}$
16	84	12	$0.77_{\pm 0.021}$	$0.82_{\pm 0.019}$	$0.95_{\pm 0.004}$
32	168	12	$0.79_{\pm 0.021}$	$0.83_{\pm 0.019}$	$0.95_{\pm 0.004}$

Table 5: Effect of fusion methods on M&M-3D.

Method	R@0.25	R@0.5	AUC
MeanPool	$0.72_{\pm 0.021}$	$0.76_{\pm 0.020}$	$0.95_{\pm 0.004}$
MaxPool	$0.77_{\pm 0.021}$	$0.81_{\pm 0.021}$	$0.95_{\pm 0.005}$
Wgt. Avg.	$0.77_{\pm 0.021}$	$0.82_{\pm 0.019}$	$0.95_{\pm 0.004}$

Benefits of transfer learning. We examine the impact of initializing M&M-3D with FFDM-pretrained weights versus COCO-pretrained weights (Lin et al., 2014). Fig. 6 shows FFDM pretraining consistently outperforms COCO pretraining, achieving higher detection and classification performance across all data sizes. The gap is largest at 10% data (∼230 malignant breasts), where FFDM pretraining improves R@0.25 by 0.11 and AUC by 0.03. FFDM pretraining also boosts annotation efficiency (Fig. 6, bottom): at 2% annotations (∼50 breasts), M&M-3D initialized with FFDM weights loses only 0.03 in R@0.25 vs. 0.09 with COCO pretraining. These results underscore again the importance of leveraging FFDM-pretrained weights as much as possible, since they provide a stronger inductive bias for DBT and yield substantial gains in both low-data and low-annotation regimes.

4.4. Ablation Studies

z-Resolution. Tab. 4 examines the impact of downsampling z-resolution to 8, 16, or 32 slices. Higher resolution boosts recall and AUC, reinforcing the value of richer 3D information. However, increasing z-resolution also affects computational cost. If training resources pose a constraint, M&M-3D with 8 slices requires 42 GB GPU memory—only modestly higher than projection and slice-by-slice methods (30 GB)—while maintaining similar training time of 11 hours and significantly improving both classification

and detection accuracy (0.94 AUC versus 0.92 AUC, 0.76 R@0.5 versus 0.74 R@0.5). Overall, we choose a z-resolution of 16 slices as it provides the best balance between performance and computational efficiency.

Fusion method. Tab. 5 compares parameter-free strategies for fusing 3D features. MeanPool significantly harms detection, suggesting that equally weighting all slices causes non-suspicious slices to dilutes the representation with non-discriminative signal. MaxPool performs on-par with Wgt. Avg., suggesting that the representation of the most suspicious slice is the most important for malignancy determination. Notably, detection improves significantly, while AUC remains stable, suggesting that emphasizing suspicious slices in the 3D representation benefits detection more than classification.

5. Conclusion

We introduced M&M-3D, a data-efficient detector for DBT that enables learned 3D reasoning without additional learnable parameters. M&M-3D forms 3D representations through malignancy-weighted feature fusion and facilitates information mixing through the cascading heads. This design leverages 3D information while allowing direct transfer of FFDM weights. Our experiments highlight the benefits of learnable 3D reasoning (§ 4.2) and show that thoughtful model design can unlock its potential without compromising data efficiency (§ 4.3).

Looking ahead, M&M-3D paves the way for unified 2D-3D learning as the model design allows handling both DBT and FFDM without architectural changes. This flexibility enhances generalizability by enabling training on larger combined 2D-3D datasets. The unified representation also supports creation of models that can seamlessly reason across exams from different years despite heterogeneous modalities, mirroring the way radiologists leverage prior exams (Hayward et al., 2016).

References

M Alberb, M Elbatel, A Elgebaly, R Montoya del Angel, X Li, and R Martí. Comoto: Unpaired cross-modal lesion distillation improves breast lesion detection in tomosynthesis. In *ADSMI @ MICCAI 2024*, 2024.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

Mateusz Buda, Ashirbani Saha, Ruth Walsh, Sujata Ghate, Nianyi Li, Albert Swiecicki, Joseph Y Lo, and Maciej A Mazurowski. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. *JAMA network open*, 4(8): e2119100–e2119100, 2021.

Marco Cantone, Ciro Russo, Federico VL Dell'ascenza, Claudio Marrocco, and Alessandro Bria. Deep learning for dbt classification with saliency-guided 2d synthesis. *Pattern Recognition*, page 112316, 2025.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Yuexi Du, Regina J Hooley, John Lewin, and Nicha C Dvornek. Sift-dbt: Self-supervised initialization and fine-tuning for imbalanced digital breast tomosynthesis image classification. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

Stephen W Duffy, László Tabár, Amy Ming-Fang Yen, Peter B Dean, Robert A Smith, Håkan Jonsson, Sven Törnberg, Sam Li-Sheng Chen, Sherry Yueh-Hsia Chiu, Jean Ching-Yuan Fann, et al. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. *Cancer*, 126(13):2971–2979, 2020.

Ming Fan, Yuanzhe Li, Shuo Zheng, Weijun Peng, Wei Tang, and Lihua Li. Computer-aided detection of mass in digital breast tomosynthesis using a faster region-based convolutional neural network. *Methods*, 166:103–111, 2019.

Jacques Ferlay, M Ervik, F Lam, M Colombet, L Mery, M Piñeros, A Znaor, I Soerjomataram, and F Bray. Global cancer observatory: cancer today. *Lyon: International agency for research on cancer*, 20182020, 2020.

Sarah M. Friedewald, Elizabeth A. Rafferty, Stephen L. Rose, Melissa A. Durand, Donna M. Plecha, Julianne S. Greenberg, Mary K. Hayes, Debra S. Copit, Kara L. Carlson, Thomas M. Cink, Lora D. Barke, Linda N. Greer, Dave P. Miller, and Emily F. Conant. Breast cancer screening using tomosynthesis in combination with digital mammography. *JAMA*, 311(24):2499–2507, 2014.

Lei Ge and Lei Dou. Non-maximum suppression for rotated object detection during merging slices of high-resolution images. *IEEE Access*, 2024.

Jessica H Hayward, Kimberly M Ray, Dorota J Wisner, John Kornak, Weiwen Lin, Bonnie N Joe, and Edward A Sickles. Improving screening mammography outcomes through comparison with multiple prior mammograms. *American Journal of Roentgenology*, 207(4):918–924, 2016.

Gongfa Jiang, Yao Lu, Jun Wei, and Yuesheng Xu. Synthesize mammogram from digital breast tomosynthesis with gradient guided cgans. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Proceedings, Part VI 22*, pages 801–809. Springer, 2019.

Gongfa Jiang, Jun Wei, Yuesheng Xu, Zilong He, Hui Zeng, Jiefang Wu, Genggeng Qin, Weiguo Chen, and Yao Lu. Synthesis of mammogram from digital breast tomosynthesis using deep convolutional neural network with gradient guided cgans. *IEEE Trans. Med. Imaging*, 40(8), 2021.

Idan Kassis, Dror Lederman, Gal Ben-Arie, Maia Giladi Rosenthal, Ilan Shelef, and Yaniv Zigel. Detection of breast cancer in digital breast tomosynthesis with vision transformers. *Scientific Reports*, 14(1): 22149, 2024.

Nicholas Konz, Mateusz Buda, Hanxue Gu, Ashirbani Saha, Jichen Yang, Jakub Chłedowski, Jungkyu Park, Jan Witowski, Krzysztof J Geras, Yoel Shoshan, et al. A competition, benchmark, code, and data for using artificial intelligence to detect lesions in digital breast tomosynthesis. *JAMA network open*, 6(2):e230524–e230524, 2023.

Xiaobo Lai, Weiji Yang, and Ruipeng Li. Dbt masses automatic segmentation using u-net neural networks. *Computational and mathematical methods in medicine*, 2020(1):7156165, 2020.

Weonsuk Lee, Hyeonsoo Lee, Hyunjae Lee, Eun Kyung Park, Hyeonseob Nam, and Thijs Kooi. Transformer-based deep neural network for breast cancer classification on digital breast tomosynthesis images. *Radiology: Artificial Intelligence*, 5(3):e220159, 2023.

Yue Li, Zilong He, Yao Lu, Xiangyuan Ma, Yanhui Guo, Zheng Xie, Genggeng Qin, Weimin Xu, Zeyuan Xu, Weiguo Chen, et al. Deep learning of mammary gland distribution for architectural distortion detection in digital breast tomosynthesis. *Phys. Med. Biol.*, 66(3):035028, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

William Lotter, Abdul Rahman Diab, Bryan Haslam, Jiye G Kim, Giorgia Grisot, Eric Wu, Kevin Wu, Jorge Onieva Onieva, Yun Boyer, Jerrold L Boxerman, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nature medicine*, 27(2):244–249, 2021.

Zekun Luo, Zheng Fang, Sixiao Zheng, Yabiao Wang, and Yanwei Fu. Nms-loss: Learning with non-maximum suppression for crowded pedestrian detection. In *proceedings of the 2021 international conference on multimedia retrieval*, pages 481–485, 2021.

Jungkyu Park, Jakub Chłedowski, Stanisław Jastrzebski, Jan Witowski, Yanqi Xu, Linda Du, Sushma Gaddam, Eric Kim, Alana Lewin, Ujas Parikh, et al. An efficient deep neural network to classify large 3d images with small objects. *IEEE Transactions on Medical Imaging*, 2023.

Ilana B Richman, Jessica B Long, Jessica R Hoag, Akhil Upneja, Regina Hooley, Xiao Xu, Natalia Kunst, Jenerius A Aminawung, Kelly A Kyanko, Susan H Busch, et al. Comparative effectiveness of digital breast tomosynthesis for breast cancer

screening among women 40-64 years old. *JNCI: Journal of the National Cancer Institute*, 113(11): 1515–1522, 2021.

Ravi K Samala, Heang-Ping Chan, Yao Lu, Lubomir M Hadjiiski, Jun Wei, and Mark A Helvie. Digital breast tomosynthesis: computer-aided detection of clustered microcalcifications on planar projection images. *Physics in Medicine & Biology*, 59(23):7457, 2014.

Sadanand Singh, Thomas Paul Matthews, Meet Shah, Brent Mombourquette, Trevor Tsue, Aaron Long, Ranya Almohsen, Stefano Pedemonte, and Jason Su. Adaptation of a deep learning malignancy model from full-field digital mammography to digital breast tomosynthesis. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314. SPIE, 2020.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021.

Mickael Tardy and Diana Mateus. Trainable summarization to improve breast tomosynthesis classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 140–149. Springer, 2021.

Paul Terrassin, Mickael Tardy, Hassan Alhajj, Nathan Lauzeral, and Nicolas Normand. Thick slices for optimal digital breast tomosynthesis classification with deep-learning. In *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care*, pages 127–136. Springer, 2024.

Yen Nhi Truong Vu, Dan Guo, Ahmed Taha, Jason Su, and Thomas Paul Matthews. M&m: Tackling false positives in mammography with a multi-view and multi-instance learning sparse detector. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 778–788. Springer, 2023.

Xiang Yu, Shui-Hua Wang, and Yu-Dong Zhang. Multiple-level thresholding for breast mass detection. *Journal of King Saud University-Computer and Information Sciences*, 35(1):115–130, 2023.

Xiaofei Zhang, Yi Zhang, Erik Y Han, Nathan Jacobs, Qiong Han, Xiaoqin Wang, and Jinze Liu. Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks. *IEEE transactions on nanobioscience*, 17 (3):237–242, 2018.

Appendix A. Comparison to Literature on BCS-DBT

Tab. 6 demonstrates that M&M-3D achieves state-of-the-art (SOTA) classification performance on BCS-DBT, with an AUC of 0.97 after finetuning on our internal dataset. Remarkably, even without any DBT-specific training data, the model attains an AUC of 0.91, which surpasses several methods from recent years, including Cantone et al. (2025) and Kassis et al. (2024). This finding highlights the strong transferability of representations learned from FFDM and their robustness to domain shift. With just 10% of DBT training data, M&M-3D reaches an AUC of 0.93, placing it on par with the top 3 methods on BCS-DBT from the DBTex challenge (Konz et al., 2023): NYU BTeam, Zedus and Vicorob. Note that all 3 DBTex methods utilize internal data and then fine-tune on BCS-DBT train set. Meanwhile, M&M-3D is only trained using our internal data. Our results suggest that M&M-3D offers a scalable path to high-performance DBT classification, even in low-data regimes.

Table 6: Classification performance on BCS-DBT. M&M-3D performs comparably to top methods even with 0% DBT data used. After finetuning with 100% data, M&M-3D achieves the highest AUC reported to date.(*) Methods not evaluated on the official test set. (†) As reported in Terrassin et al. (2024).

Method	AUC
Tardy and Mateus (2021)*	0.73
Cantone et al. (2025) 3D-Vol.	0.81
Cantone et al. (2025) 2D-Proj.	0.84
Kassis et al. (2024)	0.66
Park et al. (2023)*	0.85
Terrassin et al. (2024)	0.90
Zedus (Konz et al., 2023)†	0.92
NYU BTeam (Konz et al., 2023)†	0.93
Vicorob (Konz et al., 2023)†	0.93
Du et al. (2024)	0.93
M&M-3D-0% data	0.91
M&M-3D-10% data	0.93
M&M-3D-100% data	**0.97**

Tab. 7 reports malignant lesion recall, with false positive computed over all non-benign volumes. We remove benign cases from metric computation as DBTex challenge considers benign findings as true positive, whereas our setup considers these findings as hard false positives. For this comparison, we obtained released predictions from the top-5 DBTex methods and computed R@0.25 and R@0.5. The results show that M&M-3D consistently outperforms top methods from the DBTex competition (Konz et al., 2023). Direct transfer from FFDM (0% DBT data) already achieves recalls superior to the 4th and 5th ranking teams (Prarit and UCLA-MII). With 10% finetuning data, M&M-3D surpasses all methods in R@0.5. When trained with the full DBT dataset, M&M-3D achieves the best performance across both R@0.25 and R@0.50 metrics.

Table 7: Detection performance on BCS-DBT. We compare with top-5 methods from the DBTex Competition Konz et al. (2023). Direct transfer from FFDM using M&M-3D outperforms the 4th and 5th ranking methods. With 10% finetuning data, M&M-3D outperforms all methods for R@0.5. With 100% finetuning data, M&M-3D achieves best performance across all metrics.

Method	R@0.25	R@0.5
1. NYU B-Team	0.70	0.74
2. ZeDuS	0.70	0.76
3. Vicorob	0.74	0.77
4. Prarit	0.41	0.56
5. UCLA-MII	0.41	0.50
M&M-3D-0% data	0.52	0.56
M&M-3D-10% data	0.67	0.79
M&M-3D-100% data	**0.83**	**0.85**

Our results highlight that M&M-3D provides consistent improvements across both classification and detection tasks, setting a new benchmark for cancer detection in DBT.

Appendix B. Detailed Performance Results Across Data and Annotation Regimes

We report the performance results used to generate Fig. 5. In particular, Tab. 8 and Tab. 9 report detection and classification performance, respectively, when varying data sizes from 10% to 100%. Tab. 10 reports detection performance when varying annotation sizes from 2% to 40% while keeping dataset size fixed at 100%.

Table 8: Detection performance in terms of R@0.25 across different data regimes.

Method	10% data	40% data	100% data
TimeSform	0.51	0.65	0.74
QuerySummary	0.51	0.70	0.77
MLP-Regress	0.41	0.63	0.69
M&M-3D	**0.61**	**0.71**	**0.77**

Table 9: Classification performance in terms of breast AUC across different data regimes.

Method	10% data	40% data	100% data
TimeSform	0.90	0.93	0.95
QuerySummary	0.88	0.94	0.95
MLP-Regress	0.83	0.94	0.95
M&M-3D	**0.91**	**0.94**	**0.95**

Appendix C. Usability Study on z-axis Localization

To assess the clinical sufficiency of single-slice z-axis localization, we conducted semi-structured interviews with five board-certified radiologists. Participants were asked about their decision-making process when reviewing DBT exams and their preferences for how AI-assisted findings should be presented. Through this usability study, we found that:

- 5/5 radiologists emphasized that callback and diagnostic decisions are made primarily on the slice where the finding is most clearly visible.
- 5/5 radiologists prefered the idea of indicating each finding with the most representative slice on the 3D volume, which allows them to quickly find the finding in the volume and make their diagnostic determination.
- 2/5 radiologists reported using z-extent mainly to verify whether a finding is "real" or tissue overlap.
- 5/5 radiologists agreed that highlighting the most suspicious slice would substantially reduce navigation time without compromising diagnostic confidence.

These results support our design choice of supervising z-axis localization on a single representative slice for each finding, showing that our formulation is consistent with established radiologist workflow.

Table 10: Detection performance in terms of R@0.25 across different annotation (ann.) regimes.

Method	2% ann.	16% ann.	40% ann.
TimeSform	0.69	0.73	0.74
QuerySummary	0.64	0.74	0.77
MLP-Regress	0.44	0.63	0.69
M&M-3D	**0.74**	**0.77**	**0.77**

Appendix D. M&M-3D Algorithm

Algorithm 1 highlights the key modifications introduced in M&M-3D relative to M&M for handling DBT data. For simplicity, we illustrate the pseudocode for a single DBT view. The same procedure is applied symmetrically across both CC and MLO views during training and inference.

Algorithm 1: Pseudo code showing key modifications of M&M-3D on top of M&M to maximize transfer learning from FFDM to DBT.

```
class MM3DModel:
    def __init__(self, mm_2d):
        """init with mm_2d, which bundles:
          - proposal initializer -> (b0: Nx4, h0: NxD)
          - SelfAttn, CrossAttn # Eqs. (1), (3)
          - DynamicConv # Eq. (5)
          - ClsModule, RegModule # box cls. & reg.
              heads
          - MIL pooling (noisy-or, mean across CC/MLO)
          - get_alt_view_feats # retrieved feature from
              the other ipsilateral view
        No new learnable params are introduced.
        """
        self.mm_2d = mm_2d

    def forward_one_view(self, dbt_image,
         cross_view_image):
        """simplify forward pass on one image
        """
        b = self.mm_2d.init_boxes
        h = self.mm_2d.init_features
        backbone_feat = self.mm_2d.backbone(dbt_image)

        w_last, z_last = None, None

        for i in range(self.mm_2d.num_heads):
            # Global & multi-view mixing
            h_prime = self.mm_2d.SelfAttn(h) # Eq. (1)
            alt_view_feats = self.mm_2d.
                 get_alt_view_feats(cross_view_image)
```

```
            h_tilde = self.mm_2d.CrossAttn(h_prime,
                 alt_view_feats) # Eq. (2)

            # Slice-level Feature Interaction
            slice_h, slice_scores = [], []
            for s in range(backbone_feat.shape[1]):
                f_b_s = self.mm_2d.RoIAlign(
                     backbone_feat, b)
                # mixing of 3D feature h_tilde and
                     slice features f_b_s
                h_i_s = self.mm_2d.DynamicConv(h_tilde,
                      f_b_s) # Eq. (5)
                slice_h.append(h_i_s)
                m_i_s = self.mm_2d.ClsModule(m_i_s)
                slice_scores.append(m_i_s)

            # Slice-to-Volume Feature Fusion
            w = torch.softmax(torch.tensor(scores_slice
                 ), dim=0)
            # fusing slice_h to get a refined 3D
                 representation
            h = (w * torch.tensor(slice_h)).sum(dim=0)
                 # Eq. (6)

            # Box and score refinement
            b = self.mm_2d.RegModule(h, b)
            m = self.mm_2d.ClsModule(h)

            # z-axis localization
            z = torch.argmax(w, dim=0)
            w_last, z_last = w, z

        # MIL aggregation (noisy-or for volume)
        image_score  = self.mm_2d.image_noisy_or(m)

        return {
            "boxes_2d": b,
            "boxes_score": m,
            "boxes_z": z_last,
            "slice_weights": w_last,
            "image_score": image_score,
        }

    def loss(self, outputs, targets):
        """modified loss with z-localization
             supervision
        """
        L_mm = self.mm_2d.loss(outputs, targets)
        if targets.has_slice_annotation():
            L_z = F.cross_entropy(outputs["
                 slice_weights"], targets.gt_slice) #
                 Eq. (7)
            return L_mm + L_z
        return L_mm
```