Cross-view Correspondence Reasoning based on Bipartite Graph Convolutional Network for Mammogram Mass Detection

Yuhang Liu[1] Fandong Zhang[2] Qianyi Zhang[1] Siwen Wang[1] Yizhou Wang[3] Yizhou Yu[1,*]
[1]Deepwise AI Lab [2] Center for Data Science, Peking University
[3] Center on Frontiers of Computing Studies, Dept. of Computer Science & Technology,
Advanced Institute of Information Technology, Peking University
{liuyuhang, zhangqianyi, wangsiwen, yuyizhou}@deepwise.com
{fd.zhang, yizhou.wang}@pku.edu.cn

Abstract

Mammogram mass detection is of great clinical significance due to the high proportion of breast cancers. The information from cross views (i.e., mediolateral-oblique and cranio-caudal) is highly related and complementary, and is helpful to make comprehensive decisions. However, unlike radiologists who can recognize masses with reasoning ability in cross-view images, most existing methods lack the ability to reason under the guidance of domain knowledge, thus it limits the performance. In this paper, we introduce the bipartite graph convolutional network to endow existing methods with cross-view reasoning ability of radiologists in mammogram mass detection. The bipartite node sets are constructed by cross-view images respectively to represent relatively consistent regions in breasts, while the bipartite edge learns to model both inherent cross-view geometric constraints and appearance similarities between correspondences. Based on the bipartite graph, the information propagates methodically through correspondences and enables spatial visual features equipped with customized cross-view reasoning ability. Experimental results on DDSM dataset demonstrate the proposed algorithm achieves state-of-the-art performance. Besides, visual analysis shows the model has a clear physical meaning, which is helpful to radiologists in clinical interpretation.

1. Introduction

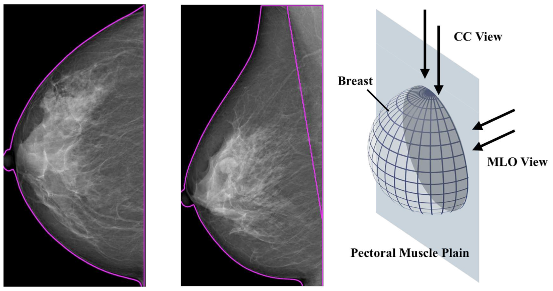
Breast cancer continues to have the highest incidence and mortality rates among women worldwide [49]. Screening mammography has been proved to effectively reduce breast cancer mortality [48]. Mass is one of the most important signs of breast cancer. However, mammogram mass detection is challenging for both radiologists and computer-aided detection (CAD) system, since masses can be partially obscured by high-intensity compacted glands especially in dense breasts. In clinical practice, cross-view images (i.e, as shown in Figure 1, cranio-caudal (CC) view which is a top-down view of the breast, and mediolateral oblique (MLO) view which is a side view of the breast taken at a certain angle) provide related and complementary information [46], and help to make comprehensive decisions.

To exploit relations of cross-view mammogram images, an intuitive idea is to adopt [21, 54] to model inter-image non-local relations. For example, CVR-RCNN [32] adds a relation module to the second stage of Faster RCNN [42] to learn inter-proposal relations. However, unlike radiologists who are able to reason with domain knowledge, the constraints of the relation learning is implicit and uncontrolled, while cross-view geometric constraints and semantic relations are not explicitly considered. Thus, the learned relations may be incorrect. Besides, the relation module relies on the quality of stage-one proposals. It may fail under severe gland occlusions, which also lead to poor performance.

We argue that the key issue is how to endow the current detection methods with the power of reasoning. When identifying masses, radiologists take the reasoning procedure explicitly. First extract suspicious regions in the examined image. And then search the regions in the auxiliary view with compatible locations and appearances. If reasonable correspondences are found, regions in both views are more likely to be mass, and vice versa. Therefore, the cross-view region-based reasoning procedure is helpful for mass detection.

Motivated by the above observations, in this paper, we introduce a novel Bipartite Graph convolutional Network (BGN) to provide the reasoning ability in mammogram mass detection. BGN can be embedded into any object detection frameworks [42, 18, 59]. It takes backbone im-

[*]Corresponding author.



(a) CC View (b) MLO View (c) Projection Model

Figure 1. Relations between CC and MLO views. Figure (a)-(b) indicate CC and MLO views of the breast. Line to the right side of figure (b) corresponds to the projected pectoral muscle plain. Figure (c) indicates an ideal projection model. The CC view is a top-down view taken along the pectoral muscle plain, while the MLO view is a side view taken at a certain angle.

age features as input and outputs cross-view enhanced features. To model cross-view region-based reasoning procedure, bipartite graph nodes are constructed by cross-view images respectively, each of which represents a relatively consistent region in breasts. The graph edges are designed to model both inherent geometric constraints and appearance similarities of cross-view nodes jointly. Therefore, only edges between nodes from different views exist, leading to a bipartite graph structure. After several layers of bipartite graph convolution, the node features are enhanced through correspondences and enables spatial visual features equipped with cross-view reasoning ability. Unlike existing methods that use none or weak cross-view constraints, the proposed model learns stronger customized cross-view reasoning with both geometric and semantic correspondences. Moreover, instead of applying the module after the proposal stage, the proposed graph module enhances backbone features before the proposal. Therefore, it suffers less from the proposal missing problem.

Experimental results on both a public dataset DDSM [20] and an in-house dataset demonstrate that the proposed algorithm achieves state-of-the-art performance. Besides, visual analysis shows the model has a clear physical meaning, which is helpful for clinical interpretation.

Our contributions are mainly two-fold: **Firstly**, we propose a novel mammogram mass detection framework that effectively exploits cross-view information and visual correspondences. **Next**, we build a bipartite graph convolutional network capable of performing reasoning about cross-view correspondences and modeling both geometric constraints and visual similarities across views.

2. Related Work

Mammogram Mass Detection Mammogram mass detection has been studied for several years. Traditional methods use handcrafted features to represent masses and design complex classifiers for identification [37, 52, 11]. However, these methods are limited due to the lack of representation ability and end-to-end training ability. In recent years, deep learning has made great progress in medical image computing [58, 17, 10, 61]. Modern object detection networks are applied to enhance mass detection performance [43, 24, 30, 55, 31, 5]. However, cross-view mammogram images containing related and complementary information are not considered. Ma *et al.* [32] attempt to model cross-view property and integrate relation module [21] into Faster RCNN [42] to learn cross-view inter-proposal relations. However, the relation learning is implicit and uncontrolled, and lacks reasoning ability under the guidance of domain knowledge. Thus, it limits the performance. Besides, severe gland occlusion may lead to proposal-missing problem, which also causes poor performance. The BGN provides reasoning ability with domain knowledge , which helps to learn stronger customized feature enhancement.

Visual Reasoning based on Graph Convolutional Network Visual reasoning aims to combine different information or interactions between objects or things, and has been applied to many computer vision tasks, such as classification [1, 34], object detection [8, 22], segmentation [28] and so on [27, 35, 15, 33, 41]. Recently, researchers attempt to introduce graph convolutional network [62] for visual reasoning [56]. Li *et al.* [28] propose graph convolutional units to learn graph visual representation from 2D data. However, information propagates from all semantic correspondences which introduce noises for learning representation. Meanwhile, the reasoning procedure is implicit and uncontrolled, and it limits the performance as well as interpretation. Gao *et al.* [16] enhance target feature by introducing a spatial-temporal graph in visual tracking task. However, the nodes are represented by uniform grids, which are sensitive to object scales, image shapes, geometric distortions, etc. Noticing crucial semantic dependencies among objects, Xu *et al.* [57] attempt to reason with a class-to-class prior graph. However, the graph remains fixed and is hard to adapt to all cases. Besides, it cannot be applied to the single-class detection tasks (e.g., mass detection).

Multi-view Visual Recognition Understanding and representing 3D object is a fundamental problem in vision recognition [39, 14] and stereo vision [25, 47, 6, 7, 40]. Multi-view based approaches render the 3D object from multi views, and deploy image-based classifiers on individual view images [50, 23]. Feng *et al.* [14] propose a group-view convolutional network to model hierarchical correlation from multiple views. Yang *et al.* [60] learn to reinforce the information by exploiting region-level and view-level relations. Different from multi-view based approaches,

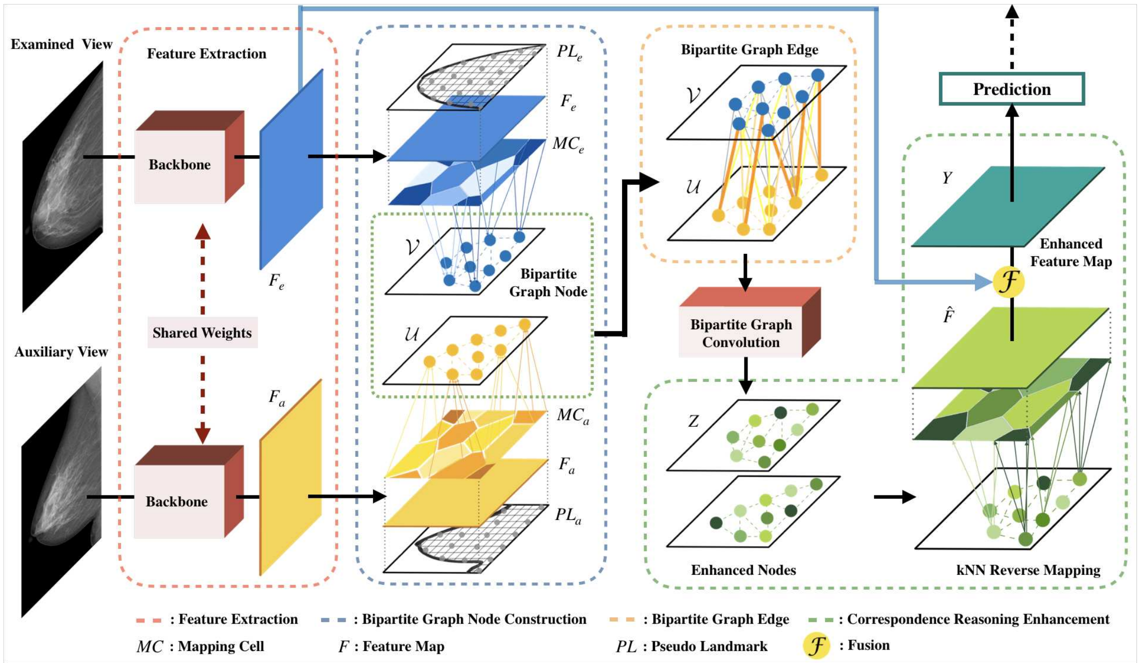


Figure 2. The pipeline of the proposed BGN. BGN takes cross-view backbone features as inputs, and outputs enhanced features for further prediction. First, bipartite graph nodes are constructed by mapping spatial visual features with pseudo landmarks. Each mapping cell is a representative region for each graph node. Then, the bipartite graph edge learns to model both geometric constraints and semantic similarities. Next, correspondence reasoning enhancement is conducted for feature enhancement by propagating information on the bipartite graph. Finally, the enhanced features are aggregated with original features for further prediction.

mammography cross-view images have more explicit correspondences, which helps to design stronger customized reasoning mechanisms. Explicit correspondences also exist in stereo vision [47, 6, 7, 40], which matches key points in general scenario as correspondences with calibrated cameras. However, different from stereo vision, we cannot get exact matched correspondences due to standard mammography screening mechanisms [46] . The key challenge is to utilize fuzzy correspondences for feature enhancement.

3. Methodology

3.1. Overview

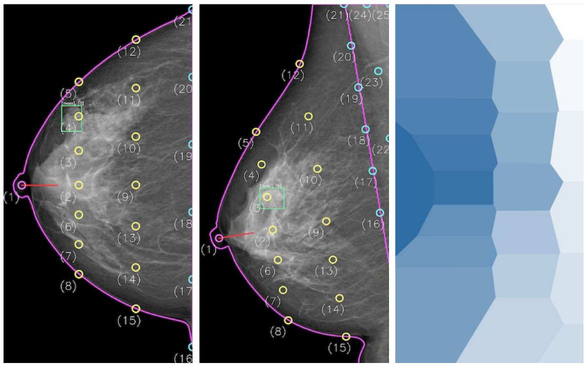
The proposed BGN aims to endow cross-view correspondence reasoning ability in the mammogram detection framework. BGN is stacked on the backbone to enhance feature representations, and can be integrated into any modern detection frameworks. As illustrated in Figure 2, there are three major steps. Firstly, to model the region-based reasoning procedure, bipartite graph nodes are mapped from cross-view backbone visual features, where each node denotes the representations of relative consistent region in breasts. Then, bipartite graph edges are designed to model both cross-view geometric constraints and appearance similarities of bipartite graph nodes. Finally, correspondence reasoning enhancement based on the pre-defined bipartite graph is designed to enhance feature representations. After information propagation through nodes, node representations are mapped to spatial visual domain reversely, which enables enhanced spatial features aware of cross-view correspondences. Both enhanced features and original backbone features are fused for further proposals.

Formally, we are given a paired 2D feature maps $F_e, F_a \in \mathbb{R}^{HW \times C}$ extracted from the examined view (where detection is performed) and its auxiliary view (the other view), where $e, a \in \{CC, MLO\}$ indicate the view types, H, W and C represent the height, width and channel of the feature map. Note either CC or MLO view can be treated as the examined view. As formulated in Equation 1, BGN learns a function f, parameterized by the bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{U}, \mathcal{E})$ with node sets as $\mathcal{V}, \mathcal{U}$ and edges as $\mathcal{E}$. $\mathcal{V}, \mathcal{U}$ indicate nodes from CC and MLO views respectively, each edge in $\mathcal{E}$ connects a node in $\mathcal{V}$ to one in $\mathcal{U}$.

$$Y = f(F_e, F_a; \mathcal{G}) \qquad (1)$$

3.2. Bipartite Graph Node

Bipartite graph node is designed to represent region-level correspondences in breasts. There are two issues: (1) Where to locate? (2) What to represent?



(a) CC View (b) MLO View (c) Mapping Cell

Figure 3. Illustration of pseudo landmarks and bipartite graph node mapping. (a)-(b) draw pseudo landmarks and the matched bounding boxes on CC and MLO views respectively. (c) illustrates how bipartite node mapping works when $k = 1$. Each mapping cell denotes the representative region of the node in the CC view.

Pseudo landmarks that preserve relative consistent location in breasts are defined to solve the first issue, while bipartite graph node mapping produces node representations from spatial visual features. We describe the details in the following parts.

3.2.1 Pseudo Landmarks

Landmarks are points in a shape object in which correspondences between and within the populations of the object are preserved [12]. However, there are no specialized landmarks for breasts. We have to define pseudo landmarks according to prior knowledge.

The pseudo landmarks should satisfy the following three properties: I. Each pseudo landmark represents a relatively consistent region in breasts; II. Different pseudo landmarks represent different regions; III. The union of all pseudo landmarks covers the breast region completely.

An intuitive idea is to treat uniform grids of the image as landmarks. However, property I. is not satisfied, leading to sensitiveness to image scale, geometric distortions, etc.

As illustrated in Figure 1, the design of the pseudo landmark embedding method is based on a basic observation: CC and MLO views of standard mammography screening have natural geometric correspondences. Ideally, a point in CC view approximately corresponds to a line parallel to projected pectoral muscle plane in MLO view.

To embed pseudo landmarks as shown in Figure 3, equidistant parallel lines are first inserted between the nipple and pectoral muscle line (projected by pectoral muscle plane). The parallel lines and the breast contour intersect, and we insert points uniformly between two intersection points. Finally, all the points are re-ordered based on intersections and defined as pseudo landmarks. Specially, as for MLO view which contains pectoral muscle areas additionally, a similar method is applied to define pseudo landmarks in pectoral muscle areas. With these processes, we obtain a set of pseudo landmarks for each view.

3.2.2 Bipartite Graph Node Mapping

Bipartite graph node mapping aims to project spatial visual features (F_{CC}, F_{MLO}) to node domain parameterized by matrices $X^{CC} \in \mathbb{R}^{|\mathcal{V}| \times C}$, $X^{MLO} \in \mathbb{R}^{|\mathcal{U}| \times C}$. The features at a node are region-level features of the region corresponding to the node.

The node mapping reveals the relation between graph nodes and all the pixels. As illustrated in the following equations, we design kNN (k Nearest Neighbor) forward mapping ϕ_k with its auxiliary matrix A for node visual representations. Each node corresponds to an irregular region satisfying the property that for any pixel in the region, the node is one of its k nearest nodes. ϕ_k performs region-level feature pooling within the regions corresponding to the graph nodes:

$$\phi_k(F, \mathcal{N}) = (Q^f)^T F, \tag{2}$$

$$Q^f = A(\Lambda^f)^{-1}, \tag{3}$$

$$A_{ij} = \begin{cases} 1 & \text{if } j \text{ th node is kNN of } i \text{ th pixel} \\ 0 & \text{Otherwise} \end{cases}, \tag{4}$$

where $\mathcal{N} \in \{\mathcal{V}, \mathcal{U}\}$ represents node set corresponding to spatial visual feature $F \in \mathbb{R}^{HW \times C}$, $A \in \mathbb{R}^{HW \times |\mathcal{N}|}$ is an auxiliary matrix to assign spatial features to top-k nearest graph nodes, $\Lambda^f \in \mathbb{R}^{|\mathcal{N}| \times |\mathcal{N}|}$ is a diagonal matrix, $\Lambda^f_{jj} = \sum_{i=1}^{HW} A_{ij}$, and $Q^f \in \mathbb{R}^{HW \times |\mathcal{N}|}$ which is a normalized form of A serves as the forward mapping matrix.

Compared with fixed-grid assign methods [16], the proposed node representations are more robust to image scales, geometric distortions, etc, since ϕ_k selects representative region adaptively according to relations among node locations. Besides, the mapping mechanism has a clear physical meaning, which is helpful for visual interpretation. Specifically, Figure 3(c) shows the mapping degenerates to Voronoi grids [2] when $k = 1$.

Based on kNN forward mapping, we can obtain visual representations of bipartite graph node sets.

$$X^{CC} = \phi_k(F_{CC}, \mathcal{V}) \tag{5}$$

$$X^{MLO} = \phi_k(F_{MLO}, \mathcal{U}) \tag{6}$$

3.3. Bipartite Graph Edge

If given a mass locating at one certain node in the examined view, it is obvious that different nodes in the auxiliary view can have different probabilities representing the same mass instance as the given mass. Thus, bipartite graph edge aims to reveal such underlying relations between nodes. We characterize the edge in two aspects: geometric constraints and appearance similarities. The two aspects describe the inherent constraints caused by mammogram screening mechanism and visual similarities between nodes, respectively.

Formally, bipartite graph edge is represented as an adjacency matrix $\mathcal{E} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{U}|}$ composed of a geometric graph $\mathcal{E}^g \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{U}|}$ and a semantic graph $\mathcal{E}^s \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{U}|}$. The geometric graph as a global prior graph reveals the geometric constraints across views. The semantic graph as an instance dependent graph represents the semantic similarities between nodes. The two graphs jointly affect cross-view information propagation. Equation 7 illustrates the relations of these matrices, where $\circ$ indicates element-wise dot.

$$\mathcal{E} = \mathcal{E}^g \circ \mathcal{E}^s \tag{7}$$

3.3.1 Geometric Relation Learning

How to represent geometric constraints? Though the CC and MLO views have standard camera pose, the exact geometric correspondence is not well-defined due to tissue deformation under pressure and lack of visual cues. We hereby model the geometric correspondence using masses as visual cues. Each edge in the geometric graph represents the correlation of the linked nodes that denote the same mass instance from different views. To approximate the correlation, for each mass, if the node is the closest to the center of the bounding box, it will be selected to represent this mass. Then we link the nodes that represent the same mass instance from different views (e.g. 4 th node in CC view and 3 th node in MLO view in Figure 3).

We take two steps to obtain the geometric graph. Firstly, we obtain a frequent statistics matrix $\epsilon \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{U}|}$ based on the annotated masses in the training set by calculating occurrences of cross-view node pairs representing the same mass instances. Then, we perform a column-row normalization method [57] to obtain $\mathcal{E}^g$.

3.3.2 Semantic Relation Learning

The geometric graph provides global geometric prior correlations. However, it is not precise enough to find exact correspondence pairs across views. Thus, noises can be involved in the reasoning procedure. The semantic graph is designed to learn the semantic relation between nodes, and can help filter the noisy relations.

How to define semantic similarities between nodes? An intuitive idea is to measure by inner product or cosine distance [3, 53]. However, relations between nodes that represent backgrounds are unknown, and may enhance background representations. Thus we release the weights, and allow the module to learn its own similarity:

$$\mathcal{E}^s_{ij} = \sigma([(X_i^{CC})^T, (X_j^{MLO})^T] w_s), \tag{8}$$

where $X_i^{CC}, X_j^{MLO} \in \mathbb{R}^C$ represent i th and j th node features of CC and MLO views respectively, $w_s \in \mathbb{R}^{2C}$ indicates the fusion parameter, and σ means the sigmoid activation function.

3.4. Correspondence Reasoning Enhancement

Correspondence reasoning enhancement, based on the defined bipartite graph $\mathcal{G}$, is designed to fully take the advantage of cross-view reasoning procedure for customized feature enhancement. There are three major steps. Firstly, we augment bipartite graph convolution to adapt to the modern graph convolutional manner. Then map node representations to spatial domain reversely, which enables spatial feature aware of the correspondences. Last, we concatenate with the original features to enhance the representations.

Bipartite Graph Convolution To adapt to the manner of modern graph convolutional network [16, 26], we first give the augmented form of the bipartite graph:

$$\hat{X} = [(X^{CC})^T, (X^{MLO})^T]^T, \tag{9}$$

$$\hat{\mathcal{E}} = \begin{pmatrix} \mathbf{0} & \mathcal{E} \\ \mathcal{E}^T & \mathbf{0} \end{pmatrix}, \tag{10}$$

where $\hat{X} \in \mathbb{R}^{|\mathcal{V} \cup \mathcal{U}| \times C}, \hat{\mathcal{E}} \in \mathbb{R}^{|\mathcal{V} \cup \mathcal{U}| \times |\mathcal{V} \cup \mathcal{U}|}$ indicate the augmented form of bipartite nodes and edges respectively.

We adopt similar fashion [16] to define graph convolution. An iteration of graph convolution layer with convolution parameters $W_g \in \mathbb{R}^{C \times D}$ is formulated in Equation 11. Intuitively, we can stack multiple graph convolutional layers in graph convolutional network.

$$Z = \sigma(\hat{\mathcal{E}} \hat{X} W_g) \tag{11}$$

kNN Reverse Mapping. To enhance spatial features, we build a kNN reverse mapping function ψ_k to project graph node features to the spatial domain. The mapping follows similar design principles as the kNN forward mapping and keeps the same number (k) of nearest neighbors.

Formally, ψ_k is formulated as :

$$\psi_k(Z, \mathcal{N}_e) = Q^r[Z]_e, \tag{12}$$

$$Q^r = (\Lambda^r)^{-1} A, \tag{13}$$

where $\mathcal{N}_e \in \{\mathcal{V}, \mathcal{U}\}$ represents the unipartite node set from the examined view, $A \in \mathbb{R}^{HW \times |\mathcal{N}|}$ follows the similar definition of the Equation 4, $[\cdot]_e$ indicates an indexing operator which selects nodes in the examined view from all bipartite nodes, $\Lambda^r \in \mathbb{R}^{HW \times HW}$ is a diagonal matrix, $\Lambda^r_{ii} = \sum_{j=1}^{|\mathcal{N}|} A_{ij}$, and $Q^r \in \mathbb{R}^{HW \times |\mathcal{N}|}$ is the reverse mapping matrix which is the normalized form of A.

Table 1. Performance on DDSM dataset(%).

Method	R@t
Campanini *et at.* [4]	80@1.1
Eltonsy *et at.* [13]	92@5.4, 88@2.4, 81@0.6
Sampat *et at.* [45]	88@2.7, 85@1.5, 80@1.0
Faster RCNN [32]	85@2.1, 75@1.8, 73@1.2
CVR-RCNN [32]	92@4.4, 88@1.9, 85@1.2
BG-RCNN	**95@4.4**, **92@1.9**, **89@1.2**

Table 2. Performance on DDSM dataset(%).

Method	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
Faster RCNN, FPN	75.3	81.5	87.3	89.8	91.4
Faster RCNN, FPN, DCN	75.7	82.5	88.4	90.1	91.4
Mask RCNN, FPN	76.0	82.5	88.7	90.8	91.4
Mask RCNN, FPN, DCN	76.7	83.9	89.4	91.4	91.8
BG-RCNN	**79.5**	**86.6**	**91.8**	**92.5**	**94.5**

Feature fusion. We finally fuse and obtain the enhanced feature Y, parameterized by $W_f \in \mathbb{R}^{D\times(D+C)}$:

$$\hat{F} = \psi_k(Z, \mathcal{N}) \tag{14}$$

$$Y = [F, \hat{F}]W_f^T \tag{15}$$

4. Experiments

4.1. Implementation Details

The mammogram images are first segmented by OTSU [38], and the foreground region is treated as input. We apply hough transform to detect pectoral muscle line and the nipple for pseudo landmark embedding. To avoid over-fitting during training, we conduct several specific augmentation methods (e.g. random flip, random crop, multi-scaling).

We build the proposed BG-RCNN by integrating BGN into Mask RCNN object detection framework [18]. ResNet50 [19] which is pretrained on ImageNet [44] is taken as a backbone network. Our implementation is based on PyTorch deep learning framework [10]. We adopt SGD with a learning rate 0.02, weight decay 10^{-4}, momentum 0.9 and nesterov set True. The whole training procedure takes 30 epoches. As for stacked bipartite graph model, we keep the same number of nearest neighbors k for both ϕ_k and ψ_k for bipartite node mapping and reverse mapping.

4.2. Datasets

Our experiments are conducted on both a public dataset called DDSM [20] and an in-house dataset. We do not choose other dataset such as INBreast [36] , MIAS [51], because the amount of dataset is insufficient.

DDSM dataset. DDSM dataset contains 2620 mammography cases. For most cases, each contains two views of images for both breasts. As in other approaches [32, 13, 4, 45], we adopt the same method to split training, validation and testing set. There are 512 cases used in the evaluation.

Table 3. Performance on in-house dataset(%).

Method	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
Faster RCNN, FPN	82.9	84.7	88.0	89.1	89.6
Faster RCNN, FPN, DCN	83.1	86.9	88.7	89.8	90.3
Mask RCNN, FPN	83.1	85.9	89.6	90.3	90.7
Mask RCNN, FPN, DCN	84.2	87.8	90.2	91.6	92.1
BG-RCNN	**87.8**	**90.5**	**92.8**	**93.9**	**94.1**

In-house dataset. We collect an in-house dataset, which contains 3000 cases and 12000 images. Each case contains cross-view images of each breast. The annotations, namely the mask of each breast lesion, are labeled by 3 radiologists with experiences of more than 10 years. When disagreement meets, we take the majority opinion of radiologists. The dataset is randomly divided into training, validation and testing sets by 8:1:1.

4.3. Baselines

Faster RCNN, FPN. Faster RCNN [42] with Feature Pyramid Network (FPN) [29] is a solid baseline in object detection task. We use ResNet50[19] as the backbone.

Faster RCNN, FPN, DCN. Deformable Convolution Network (DCN) [9] is used to enhance the transformation modeling capability of convolutional networks. DCN is integrated into baselines to enhance the performance.

Mask RCNN, FPN. Mask RCNN [18] is a state-of-the-art model on both object detection and instance segmentation. To exploit the mask supervision for localization, we employ Mask RCNN as a baseline.

Mask RCNN, FPN, DCN. DCN is also integrated into Mask RCNN baselines.

4.4. Comparison with state-of-the-art methods

We evaluate the performance by recall (R) at t false postive per image (FPI), simplified as $R@t$, where $t \in \{0.5, 1.0, 2.0, 3.0, 4.0\}$.

Table 1 and Table 2 report the experimental performance on DDSM dataset. Results in Table 1 are reported from [4, 13, 45, 32], and baselines in Table 2 are implemented by us. We do not compare with [31], as they do not release the dataset split method. We keep the same FPI and compare the recall with a strong competitor CVR-RCNN[32]. We can conclude that the proposed model outperforms state-of-the-art methods. The same conclusion can be drawn on the in-house dataset from Table 3. To understand how the

Table 4. Effectiveness of pseudo landmarks on DDSM dataset(%).

Method	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
Uniform	76.4	84.6	90.4	**92.5**	93.2
BG-RCNN	**79.5**	**86.6**	**91.8**	**92.5**	**94.5**

Table 5. Effectiveness of node number on DDSM dataset(%).

Method	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
$\mathcal{V}_1, \mathcal{U}_1$	75.7	83.9	88.7	92.1	93.2
$\mathcal{V}_9, \mathcal{U}_{13}$	79.5	86.0	90.4	92.5	93.2
$\mathcal{V}_{21}, \mathcal{U}_{25}$	**79.5**	**86.6**	**91.8**	92.5	**94.5**
$\mathcal{V}_{42}, \mathcal{U}_{46}$	76.0	85.6	90.8	**93.5**	94.2
$\mathcal{V}_{66}, \mathcal{U}_{71}$	78.8	85.6	90.1	92.1	93.2

proposed model benefits from the correspondence reasoning mechanism, we analyze the cases in Figure 4. We compare the recall when keeping the same FPI. We can see that the proposed method can significantly improve the recall (the 2nd and 3rd row) and localization ability (the 1st row).

4.5. Ablation study

Ablation of Pseudo Landmarks As shown in Table 4, We first investigate the effectiveness of pseudo landmarks versus uniform grids. We keep the same number of uniform grids and pseudo landmarks. The results have demonstrated that pseudo landmarks are rather more effective than uniform grids. We also investigate how node number affects the performance. "$\mathcal{V}_i, \mathcal{U}_j$" means there are i nodes in CC view and j nodes in MLO view. Specifically, the setting "$\mathcal{V}_1, \mathcal{U}_1$" is equivalent to two-branch Faster RCNN. As shown in Table 5, we choose "$\mathcal{V}_{21}, \mathcal{U}_{25}$" as our final results.

Ablation of Bipartite Graph Node Mapping. To investigate the effectiveness of bipartite node mapping, we first compare with a simple method, which directly crop a fixed region for graph node representation. We also evaluate how k influences the results. We keep the same k for both ϕ_k and ψ_k. As shown in Table 6, we can see that bipartite graph node mapping is effective and necessary. Meanwhile, when dense nodes embedded, the model works better with larger k, because larger k can abstract more context feature for the node which may lack sufficient context representation.

Ablation of Bipartite Graph Edge. We analyze the influence of $\mathcal{E}^s$ and $\mathcal{E}^g$ in the bipartite graph edge. It degenerates to naive Mask RCNN when neither $\mathcal{E}^s$ nor $\mathcal{E}^g$ are used, since no information propagates across views. When either $\mathcal{E}^s$ or $\mathcal{E}^g$ is used, we set $\mathcal{E}$ to $\mathcal{E}^s$ or $\mathcal{E}^g$ respectively. As shown in Table 7, either $\mathcal{E}^s$ or $\mathcal{E}^g$ can make an improvement, and combining both parts achieves the best performance.

Table 6. Effectiveness of bipartite graph node mapping on DDSM dataset(%).

Method	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
$\mathcal{V}_{21}, \mathcal{U}_{25}$, crop	76.4	84.2	90.1	91.8	93.2
$\mathcal{V}_{21}, \mathcal{U}_{25}$, k=1	**79.8**	86.3	**91.8**	**92.5**	94.2
$\mathcal{V}_{21}, \mathcal{U}_{25}$, k=2	79.5	**86.6**	**91.8**	**92.5**	**94.5**
$\mathcal{V}_{21}, \mathcal{U}_{25}$, k=3	79.5	86.3	87.7	91.8	93.5
$\mathcal{V}_{66}, \mathcal{U}_{71}$, crop	75.3	84.2	89.4	91.8	92.1
$\mathcal{V}_{66}, \mathcal{U}_{71}$, k=1	77.7	85.6	90.1	91.4	93.2
$\mathcal{V}_{66}, \mathcal{U}_{71}$, k=2	78.8	85.6	90.1	92.1	93.2
$\mathcal{V}_{66}, \mathcal{U}_{71}$, k=3	79.1	86.0	90.1	92.1	93.8

Table 7. Effectiveness of each component in bipartite graph edge on DDSM dataset(%).

$\mathcal{E}^s$	$\mathcal{E}^g$	R@0.5	R@1.0	R@2.0	R@3.0	R@4.0
×	×	76.7	83.9	89.4	91.4	91.8
×	√	78.4	83.9	91.1	92.1	93.8
√	×	77.7	86.3	89.4	91.8	93.5
√	√	**79.5**	**86.6**	**91.8**	**92.5**	**94.5**

4.6. Visualization

Our visualization experiments mainly answer two questions: (1) Where does the bipartite graph focus on auxiliary view? (2) How does the correspondence reasoning mechanism enhance the feature representations?

To answer the first question, we design a specialized method for correspondence visualization. The main purpose is to find representative regions of correlated nodes in the auxiliary view when given a query mass in the examined view. We first define a one-hot representative vector $x \in \mathbb{R}^{|\mathcal{V} \cup \mathcal{U}|}$ to represent locations of the mass in the examined area. The index of the node which is nearest to the center of the analyzed mass in the examined image is set to 1. Then visualize the feature by Equation 16, where $o \in \mathbb{R}^{HW}$ represents the response vector, and $[\cdot]_e$ indicates indexing operator which selects nodes in the examined view from bipartite graph nodes. As shown in Figure 4, we can see that the bipartite graph focuses on the mated mass area in the auxiliary view, which helps to learn complementary feature representations. Besides, our model has a clear physical meaning and provides visual cues of mated masses. Thus it can help radiologists in clinical interpretation.

$$o = Q^r [\hat{\mathcal{E}} x]_e \tag{16}$$

To answer the second question, we compare the response map before and after feature enhancement. Specifically, channel-wise max pooling is conducted on F_e and Y respectively. As shown in Figure 4, feature response map activates more prominently on mass located area after enhancement. As a result, the corresponding reasoning enhancement method can help to improve the detection performance and make comprehensive and sufficient judgment.

5. Conclusions and Future Work

In this paper, we introduce the bipartite graph convolutional network to provide customized reasoning ability in

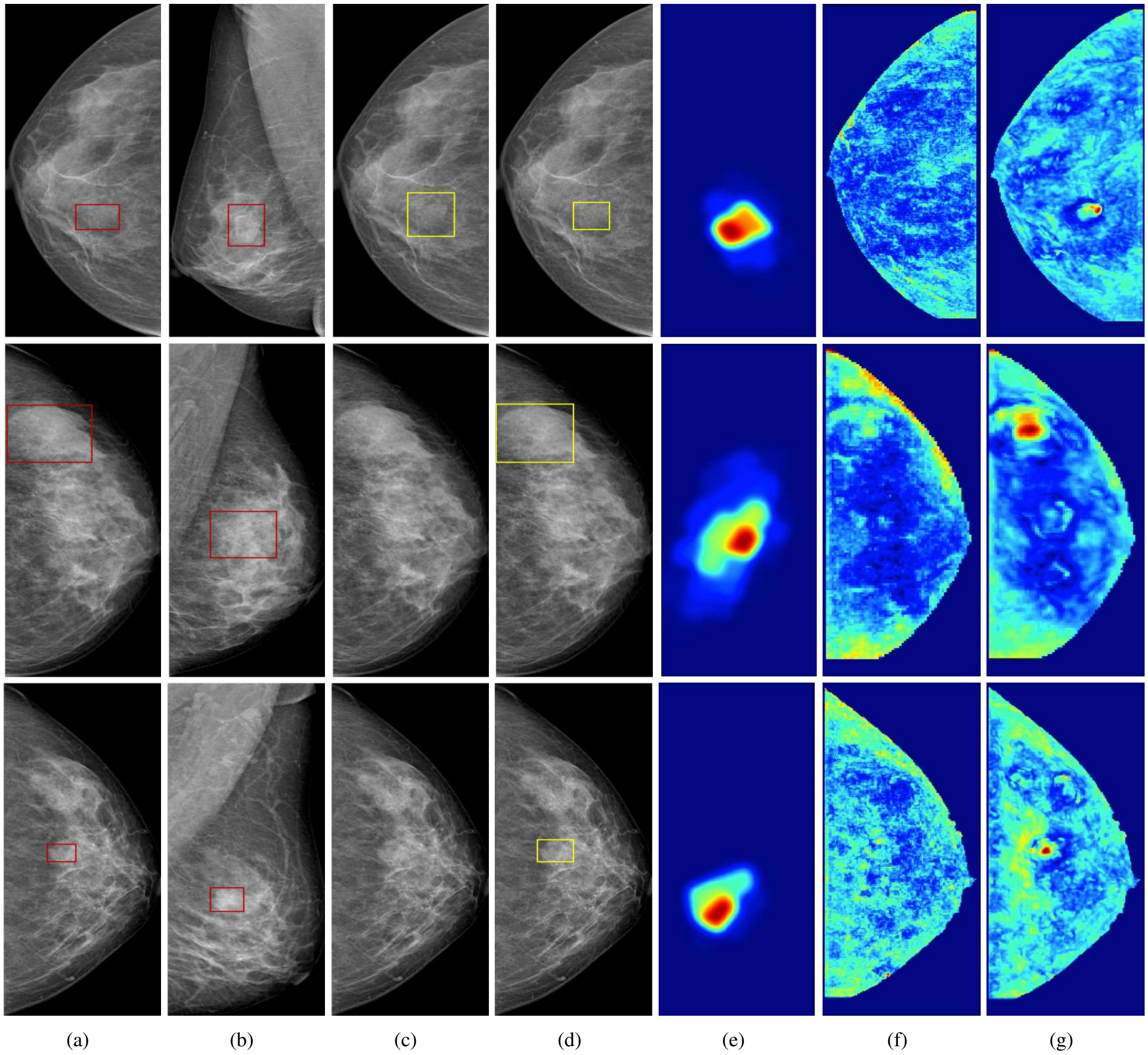


Figure 4. Detection results of BG-RCNN. Each row shows a representative case. Column (a)-(b) refer to the examined image and its auxiliary view image with annotations. Column (c)-(d) indicate detection results by Mask-RCNN and BG-RCNN. Column (e) visualizes the attention area on the auxiliary view. Column (f)-(g) visualize the response maps before and after correspondence visual reasoning.

mammogram mass detection. To model the cross-view reasoning procedure, bipartite graph nodes induced by pseudo landmarks are constructed from cross-view images respectively, which are able to represent the relatively consistent regions. Then bipartite graph edge learns both cross-view inherent geometric constraints and semantic relations. Finally, in correspondence reasoning enhancement, information propagates on the bipartite graph and it enables spatial visual features aware of cross-view correspondences. Thus feature representations are enhanced. Experiments on both public and in-house datasets demonstrate that the proposed method achieves state-of-the-art performance. Besides, visual analysis shows that the proposed model has a clear physical meaning, which is helpful to radiologists in clinical interpretation.

Future work will include: (1) exploring learnable forms of pseudo landmarks; (2) integrating bilateral (same view of left and right breasts) domain knowledge to the model; (3) exploiting more powerful graph mechanisms to facilitate information propagation.

Acknowledgment

This work was supported in part by Zhejiang Province Key Research Development Program (No. 2020C03073), MOST-2018AAA0102004, NSFC-61625201, 61527804, DFG TRR169 / NSFC Major International Collaboration Project ”Crossmodal Learning”. We would like to thank Shu Zhang and Yuchun Chen for valuable discussions.

References

[1] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2552–2566, 2014. 2

[2] Franz Aurenhammer and Rolf Klein. Voronoi diagrams. *Handbook of computational geometry*, 5(10):201–290, 2000. 4

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005. 5

[4] Renato Campanini, Danilo Dongiovanni, Emiro Iampieri, Nico Lanconelli, Matteo Masotti, Giuseppe Palermo, Alessandro Riccardi, and Matteo Roffilli. A novel featureless approach to mass detection in digital mammograms based on support vector machines. *Physics in Medicine & Biology*, 49(6):961, 2004. 6

[5] Zhenjie Cao, Zhicheng Yang, Xiaoyan Zhuo, Ruei-Sung Lin, Shibin Wu, Lingyun Huang, Mei Han, Yanbo Zhang, and Jie Ma. Deeplima: Deep learning based lesion identification in mammograms. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[6] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[7] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals using stereo imagery for accurate object class detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1259–1272, 2017. 2, 3

[8] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7239–7248, 2018. 2

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 6

[10] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. Automated mass detection in mammograms using cascaded deep learning and random forests. In *2015 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–8. IEEE, 2015. 2, 6

[11] João Otávio Bandeira Diniz, Pedro Henrique Bandeira Diniz, Thales Levi Azevedo Valente, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, and Marcelo Gattass. Detection of mass regions in mammograms by bilateral analysis adapted to breast density using similarity indexes and convolutional neural networks. *Computer methods and programs in biomedicine*, 156:191–207, 2018. 2

[12] I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis, with Applications in R. Second Edition.* John Wiley and Sons, Chichester, 2016. 4

[13] Nevine H Eltonsy, Georgia D Tourassi, and Adel S Elmaghraby. A concentric morphology model for the detection of masses in mammography. *IEEE transactions on medical imaging*, 26(6):880–889, 2007. 6

[14] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2018. 2

[15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 2

[16] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4649–4659, 2019. 2, 4, 5

[17] Zhihui Guo, Ling Zhang, Le Lu, Mohammadhadi Bagheri, Ronald M Summers, Milan Sonka, and Jianhua Yao. Deep logismos: Deep learning graph-based 3d segmentation of pancreatic tumors on ct scans. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1230–1233. IEEE, 2018. 2

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 6

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[20] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and W Philip Kegelmeyer. The digital database for screening mammography. In *Proceedings of the 5th international workshop on digital mammography*, pages 212–218. Medical Physics Publishing, 2000. 2, 6

[21] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018. 1, 2

[22] Chenhan Jiang, Hang Xu, Xiaodan Liang, and Liang Lin. Hybrid knowledge routed modules for large-scale object detection. In *Advances in Neural Information Processing Systems*, pages 1552–1563, 2018. 2

[23] Edward Johns, Stefan Leutenegger, and Andrew J Davison. Pairwise decomposition of image sequences for active multi-view recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3813–3822, 2016. 2

[24] Hwejin Jung, Bumsoo Kim, Inyeop Lee, Minhwan Yoo, Junhyun Lee, Sooyoun Ham, Okhee Woo, and Jaewoo Kang. Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. *PloS one*, 13(9):e0203355, 2018. 2

[25] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in neural information processing systems*, pages 365–376, 2017. 2

[26] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 5

[27] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009. 2

[28] Yin Li and Abhinav Gupta. Beyond grids: Learning graph representations for visual recognition. In *Advances in Neural Information Processing Systems*, pages 9225–9235, 2018. 2

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 6

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[31] Yuhang Liu, Zhen Zhou, Shu Zhang, Ling Luo, Qianyi Zhang, Fandong Zhang, Xiuli Li, Yizhou Wang, and Yizhou Yu. From unilateral to bilateral learning: Detecting mammogram masses with contrasted bilateral network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 477–485. Springer, 2019. 2, 6

[32] Jiechao Ma, Sen Liang, Xiang Li, Hongwei Li, Bjoern H Menze, Rongguo Zhang, and Wei-Shi Zheng. Cross-view relation networks for mammogram mass detection. *arXiv preprint arXiv:1907.00528*, 2019. 1, 2, 6

[33] Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *Proceedings of the IEEE international conference on computer vision*, pages 2533–2541, 2015. 2

[34] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844*, 2016. 2

[35] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017. 2

[36] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012. 6

[37] Naga R Mudigonda, Rangaraj M Rangayyan, and JE Leo Desautels. Detection of breast masses in mammograms by density slicing and texture flow-field analysis. *IEEE Transactions on Medical Imaging*, 20(12):1215–1227, 2001. 2

[38] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, Jan 1979. 6

[39] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016. 2

[40] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. *International Journal of Computer Vision*, 126(12):1394–1414, 2018. 2, 3

[41] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2, 6

[43] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):4165, 2018. 2

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252. 6

[45] Mehul P Sampat, Alan C Bovik, Gary J Whitman, and Mia K Markey. A model-based framework for the detection of spiculated masses on mammography a. *Medical physics*, 35(5):2110–2123, 2008. 6

[46] Mehul P Sampat, Mia K Markey, Alan C Bovik, et al. Computer-aided detection and diagnosis in mammography. *Handbook of image and video processing*, 2(1):1195–1217, 2005. 1, 3

[47] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017. 2, 3

[48] Edward A Sickles. Breast cancer screening outcomes in women ages 40-49: clinical experience with service screening using modern mammography. *JNCI Monographs*, 1997(22):99–104, 1997. 1

[49] Rebecca Siegel, Jiemin Ma, Zhaohui Zou, and Ahmedin Jemal. Cancer statistics, 2014. *CA: a cancer journal for clinicians*, 64(1):9–29, 2014. 1

[50] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 2

[51] P SUCKLING J. The mammographic image analysis society digital mammogram database. *Digital Mammo*, pages 375–386, 1994. 6

[52] Shen-Chuan Tai, Zih-Siou Chen, and Wei-Ting Tsai. An automatic mass detection system in mammograms based on complex texture features. *IEEE journal of biomedical and health informatics*, 18(2):618–627, 2014. 2

[53] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Iccv*, volume 98, page 2, 1998. 5

[54] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. 1

[55] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. Deep neural networks improve radiologists' performance in breast cancer screening. 2019. 2

[56] Hang Xu, Chenhan Jiang, Xiaodan Liang, and Zhenguo Li. Spatial-aware graph relation network for large-scale object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9298–9307, 2019. 2

[57] Hang Xu, ChenHan Jiang, Xiaodan Liang, Liang Lin, and Zhenguo Li. Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6419–6428, 2019. 2, 5

[58] Zhoubing Xu, Yuankai Huo, JinHyeong Park, Bennett Landman, Andy Milkowski, Sasa Grbic, and Shaohua Zhou. Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 711–719. Springer, 2018. 2

[59] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. *arXiv preprint arXiv:1904.11490*, 2019. 1

[60] Ze Yang and Liwei Wang. Learning relationships for multiview 3d object recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7505–7514, 2019. 2

[61] Fandong Zhang, Ling Luo, Xinwei Sun, Zhen Zhou, Xiuli Li, Yizhou Yu, and Yizhou Wang. Cascaded generative and discriminative learning for microcalcification detection in breast mammograms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12578–12586, 2019. 2

[62] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018. 2