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

Yuhang Liu\textsuperscript{1}  Fandong Zhang\textsuperscript{2}  Qianyi Zhang\textsuperscript{1}  Siwen Wang\textsuperscript{1}  Yizhou Wang\textsuperscript{3}  Yizhou Yu\textsuperscript{1,}\textsuperscript{*}

\textsuperscript{1}Deepwise AI Lab  \textsuperscript{2}Center for Data Science, Peking University  \textsuperscript{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-
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,
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 $\mathcal{U}$.

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

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?
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 \( X^{CC} \in \mathbb{R}^{|V| \times C} \), \( X^{MLO} \in \mathbb{R}^{|M| \times |C|} \). 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 \( \phi_k \) 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. \( \phi_k \) performs region-level feature pooling within the regions corresponding to the graph nodes:

\[
\phi_k(F, N) = (Q^f)^TF, \quad (Q^f = A(A^f)^{-1}, \quad (A_{ij}) = \begin{cases} 1 & \text{if } j \text{ th node is kNNe of } i \text{ th pixel} \\ 0 & \text{Otherwise} \end{cases},
\]

where \( N \in \{V, U\} \) represents node set corresponding to spatial visual feature \( F \in \mathbb{R}^{HW \times C} \), \( A \in \mathbb{R}^{HW \times |N|} \) is an auxiliary matrix to assign spatial features to top-k nearest graph nodes, \( A^f \in \mathbb{R}^{|N| \times |N|} \) is a diagonal matrix, \( A^f = \sum_{i=1}^{HW} A_{ij} \), and \( Q^f \in \mathbb{R}^{HW \times |N|} \) 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 \( \phi_k \) 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.

\[
X^{CC} = \phi_k(F_{CC}, V) \quad (5) \\
X^{MLO} = \phi_k(F_{MLO}, U) \quad (6)
\]

### 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 \( E \in \mathbb{R}^{|V| \times |U|} \) composed of a geometric graph \( E^g \in \mathbb{R}^{|V| \times |U|} \) and a semantic graph \( E^s \in \mathbb{R}^{|V| \times |U|} \). The geometric graph as a global prior graph reveals the geometric 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.

\[
E = E^g \circ E^s \quad (7)
\]

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., 4th node in CC view and 3rd node in MLO view in Figure 3).

We take two steps to obtain the geometric graph. Firstly, we obtain a frequent statistics matrix \( e \in \mathbb{R}^{|V| \times |U|} \) 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 \( E^g \).

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 process. 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:

\[
\mathcal{E}^s_{ij} = \sigma\left(\langle (X^CC_i)^T, (X^MLO_j)^T \rangle w_s \right), \quad (8)
\]

where \( X^CC, X^MLO \in \mathbb{R}^C \) represent the node features of\( CC \) and \( MLO \) views respectively, \( w_s \in \mathbb{R}^{2C} \) indicates the fusion parameter, and \( \sigma \) means the sigmoid activation function.

3.4. Correspondence Reasoning Enhancement

Correspondence reasoning enhancement, based on the defined bipartite graph \( G \), 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:

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

\[
\hat{\mathcal{E}} = \begin{bmatrix} 0 & \mathcal{E} \\ \mathcal{E}^T & 0 \end{bmatrix}, \quad (10)
\]

where \( \hat{X} \in \mathbb{R}^{(|V|+|U|) \times |V∪U|} \), \( \hat{\mathcal{E}} \in \mathbb{R}^{(|V|+|U|) \times (|V|+|U|)} \) 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.

\[
Z = \sigma(\hat{\mathcal{E}} \hat{X} W_g) \quad (11)
\]

kNN Reverse Mapping. To enhance spatial features, we build a kNN reverse mapping function \( \psi_k \) 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, \( \psi_k \) is formulated as:

\[
\psi_k(Z, \mathcal{N}_c) = Q^* [Z]_c, \quad (12)
\]

\[
Q^* = (\Lambda^r)^{-1} A, \quad (13)
\]

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

<table>
<thead>
<tr>
<th>Method</th>
<th>R@0.5</th>
<th>R@1.0</th>
<th>R@2.0</th>
<th>R@3.0</th>
<th>R@4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster RCNN, FPN</td>
<td>92.5</td>
<td>89.8</td>
<td>93.9</td>
<td>94.1</td>
<td>94.3</td>
</tr>
<tr>
<td>Faster RCNN, FPN, DCN</td>
<td>92.8</td>
<td>93.6</td>
<td>94.1</td>
<td>94.4</td>
<td>94.5</td>
</tr>
<tr>
<td>Mask RCNN, FPN</td>
<td>90.5</td>
<td>90.7</td>
<td>91.4</td>
<td>91.8</td>
<td>92.0</td>
</tr>
<tr>
<td>Mask RCNN, FPN, DCN</td>
<td>91.0</td>
<td>91.8</td>
<td>92.1</td>
<td>92.4</td>
<td>92.6</td>
</tr>
<tr>
<td>BG-RCNN</td>
<td>96.0</td>
<td>96.5</td>
<td>96.9</td>
<td>97.3</td>
<td>97.5</td>
</tr>
</tbody>
</table>

Table 2. Performance on DDSM dataset(\%).

<table>
<thead>
<tr>
<th>Method</th>
<th>R@0.5</th>
<th>R@1.0</th>
<th>R@2.0</th>
<th>R@3.0</th>
<th>R@4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster RCNN, FPN</td>
<td>75.3</td>
<td>81.5</td>
<td>87.3</td>
<td>89.8</td>
<td>91.4</td>
</tr>
<tr>
<td>Faster RCNN, FPN, DCN</td>
<td>75.7</td>
<td>82.5</td>
<td>88.4</td>
<td>90.1</td>
<td>91.4</td>
</tr>
<tr>
<td>Mask RCNN, FPN</td>
<td>76.0</td>
<td>82.5</td>
<td>88.7</td>
<td>90.8</td>
<td>91.4</td>
</tr>
<tr>
<td>Mask RCNN, FPN, DCN</td>
<td>76.7</td>
<td>83.9</td>
<td>89.4</td>
<td>91.4</td>
<td>91.8</td>
</tr>
<tr>
<td>BG-RCNN</td>
<td>79.5</td>
<td>86.6</td>
<td>91.8</td>
<td>92.5</td>
<td>94.5</td>
</tr>
</tbody>
</table>

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 positive 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

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

$$ F = \psi_k(Z, N) $$

$$ Y = [F; \hat{F}] W_f^T $$

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.

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.

**4.1. 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 epochs. As for stacked bipartite graph model, we keep the same number of nearest neighbors k for both $\phi_k$ and $\psi_k$ 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.

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 positive 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
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 demonstrate that pseudo landmarks are rather more effective than uniform grids. We also investigate how node number affects the performance. “V$i$, U$t_i$” means there are i nodes in CC view and j nodes in MLO view. Specifically, the setting “V$1$, U$1$” is equivalent to two-branch Faster RCNN. As shown in Table 5, we choose “V$25$, 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 $\phi_k$ and $\psi_k$. 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 $\mathcal{E}^a$ and $\mathcal{E}^o$ in the bipartite graph edge. It degenerates to naive Mask RCNN when neither $\mathcal{E}^a$ nor $\mathcal{E}^o$ are used, since no information propagates across views. When either $\mathcal{E}^a$ or $\mathcal{E}^o$ is used, we set $\mathcal{E}$ to $\mathcal{E}^a$ or $\mathcal{E}^o$ respectively. As shown in Table 7, either $\mathcal{E}^a$ or $\mathcal{E}^o$ can make an improvement, and combining both parts achieves the best performance.

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 $x \in \mathbb{R}^{1 \times |\mathcal{U}|}$ 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.

$$o = Q^e[\hat{\mathcal{E}}x]_e$$  \hspace{1cm} (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
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


