Supplementary Material for the Paper "Check and Link: Pairwise Lesion Correspondence Guides Mammogram Mass Detection"

Ziwei Zhao*[1,4,5], Dong Wang*[2], Yihong Chen[1], Ziteng Wang[3], and Liwei Wang[1,2]

[1] Center for Data Science, Peking University
[2] Key Laboratory of Machine Perception, MOE, School of Artificial Intelligence, Peking University
[3] Yizhun Medical AI Co., Ltd
[4] Pazhou Laboratory (Huangpu)
[5] Peng Cheng Laboratory

1 Performance on the In-house Dataset

Table 1: Performance on the in-house dataset (%).

Method	R@0.125	R@0.25	R@0.5	R@1.0	R@2.0
Faster RCNN [4]	78.8	84.0	88.6	91.2	93.6
Mask RCNN [2]	80.5	85.2	89.2	92.1	94.8
Deformable DETR [5]	79.2	84.7	88.6	**93.4**	95.3
CL-Net	**82.2**	**87.8**	**90.6**	**93.2**	**96.0**

2 Additional Ablation Study

Table 2: Ablation study on number of attention layers in lesion linker.

Number	R@0.25	R@0.5	R@1.0	R@2.0	R@4.0
1	77.1	81.4	**88.4**	**92.4**	94.4
2	77.1	82.4	88.0	91.4	93.4
3	**78.1**	**83.1**	88.0	**92.4**	**95.0**
4	75.1	80.4	86.4	90.7	93.0
5	75.1	80.4	85.0	90.0	92.0

Number of Attention Layers in Lesion Linker. Since several attention layers can be stacked in Lesion Linker for more powerful reasoning ability, we explore the effect of the number of attention layers in Table 2. From the table, with the number of attention layers increasing from 1 to 3, performance of CL-Net continues to improve and gets the best overall recalls at 3. Thus the number of layers is set to 3 as our default setting.

* Equal contribution.

3 Visualization and Analysis

Figure 1 illustrates an example of the detection and linking results of our proposed CL-Net. As shown in column (a - b), the two predicted lesion pairs are consistent with the ground truth. Column (c) and (d) present the attention weights between link queries and the detection embeddings outputted by object queries. For each link query, the highest responses are obtained from one lesion pair from the two views. This phenomenon tallies with our expectation that the lesion linker can learn the pairwise correspondence across views, which performs as the guidance for lesion detection. The error analysis can be found in the supplementary materials.

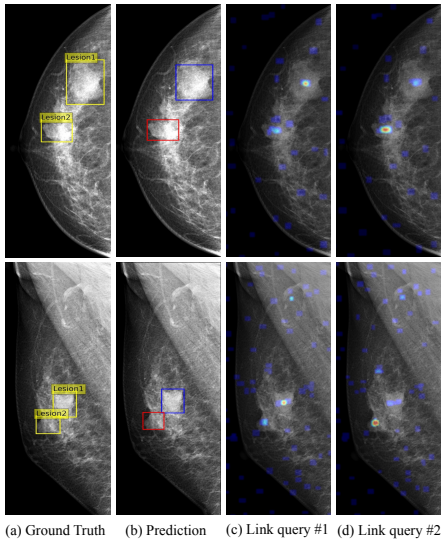

Fig. 1: Case visualization. We mark a lesion pair using the same id for ground truth and the same color for predictions. Column (c - d) visualize the attention weights between the two link queries that generate pair outputs in column (b) and the detection embeddings outputted by object queries of the two views. Red corresponds to larger weights.

4 Details of Match Learning Strategies

In the main text of our paper, in order to achieve precise pairwise lesion correspondence, lesion linker utilizes correspondence supervision to guide the interaction process across lesion candidates of MLO and CC views. Besides lesion linker, we also introduce two alternative approaches for pairwise correspondence learning in Section 3.5 of the main paper to highlight the advantages of our proposed method. We will elaborate on the details of the alternative approaches in the following.

4.1 Pair Verification

Verifying whether each two lesion candidates from ipsilateral views are truly paired lesions is a straightforward method to achieve lesion matching. We instantiate this model based on View-Interactive Lesion Detector (VILD). Given output embeddings $E^c \in \mathbb{R}^{N\times D}$ and $E^m \in \mathbb{R}^{N\times D}$ from VILD as inputs, a MLP layer firstly transforms them to new representation space which can be expressed as:

$$E^{c*} = \mathrm{MLP}(E^c),\ E^{m*} = \mathrm{MLP}(E^m). \tag{1}$$

Similar with lesion linker, we also set a learnable dustbin embedding $e^d \in \mathbb{R}^D$ to obtain the complete version $\hat{E}^c$, $\hat{E}^m \in \mathbb{R}^{(N+1)\times D}$:

$$\hat{E}^c = \mathrm{Concat}(E^{c*}, e^d),\ \hat{E}^m = \mathrm{Concat}(E^{m*}, e^d). \tag{2}$$

We construct the 2D matching matrix which represents the match probabilities of all possible pairs by calculating the similarity score for every two embeddings from ipsilateral views:

$$S_{i,j} = \langle \hat{E}_i^c, \hat{E}_j^m \rangle, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ is the inner product. $(i, j) \in [1, N+1] \times [1, N+1]$. $\hat{E}_i^t$ denotes the i-th row of $\hat{E}^t$, and $t \in \{c, m\}$.

To learn pairwise lesion correspondence in a supervised manner, we introduce the ground truth matching matrix $M \in \mathbb{R}^{(N+1)\times(N+1)}$. Through the label assignment rule for detection in DETR [1], pairwise ground truth boxes can be naturally converted to pairwise ground truth embeddings. We denote the ground truth match ids for the embeddings of ipsilateral views as $\mathcal{GT} = \{(\mu_i, v_i)\}_{i=1}^n$, which represent that the μ_i-th embedding in CC view and the v_i-th embedding in MLO view are a pair of ground truth embeddings. If a lesion corresponding to the μ_i-th embedding can only be viewed in CC view, we denote its match id as $(\mu_i, N+1)$. Lesions that can only be viewed in MLO view can also be processed in a similar way. Thus M can be obtained as follows:

$$M[i, j] = \begin{cases} 1, & \text{if } (i, j) \in \mathcal{GT} \\ 0, & \text{else} \end{cases} \tag{4}$$

The final loss function for this method can be written as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{D}} + \mathcal{L}_{\mathrm{V}}(S, M), \tag{5}$$

where $\mathcal{L}_D$ is the loss function in DETR, and $\mathcal{L}_{\mathrm{V}}$ is used to supervise the matching results. We adopt focal loss [3] as the loss function $\mathcal{L}_{\mathrm{V}}$ to achieve the supervision between predicted matching matrix S and ground truth matching matrix M.

4.2 Paired Lesion Query

We can also predict pairwise lesions with query mechanism directly. For this method, we use the same architecture of backbone and transformer encoder as VILD. In transformer decoder, as illustrated in Figure 2, we initialize a set of learnable paired lesion queries. At first paired lesion queries are passed through the multi-head self-attention layer. Then they will interact with image features for CC and MLO views outputted by transformer encoder sequentially to extract pairwise lesion information. Finally, a FFN layer is utilized to enhance the representative ability. Above layers can be stacked several times, and we denote the number of stacked layers as $\mathrm{N^d}$. At the top of decoder, the classes and bounding boxes of MLO and CC views are directly predicted through several FFN layers for each query.

The output of decoder can be reformulated as a set of N quads $\{\langle \hat{b}_i^c, \hat{b}_i^m, \hat{c}_i^c, \hat{c}_i^m \rangle\}_{i=1}^N$, where $\hat{b}_i^t \in \mathbb{R}^4$ and $\hat{c}_i^t \in \mathbb{R}^1$ are the predicted bounding box and classification score. Here $t \in \{\mathrm{CC}, \mathrm{MLO}\}$ denotes CC view or MLO view. The set of N quads

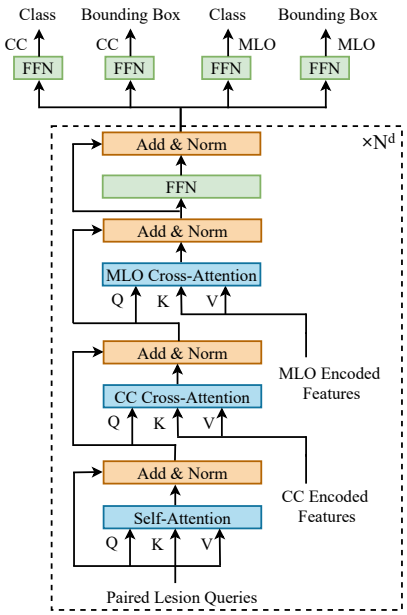

Fig. 2: Architecture of Decoder in Paired Lesion Query.

is similar with the extracted lesion pairs in lesion linker, thus we can also adopt a similar one-to-one matching assignment between the ground truth lesion pairs and the predicted lesion pairs. The cost function can be expressed as:

$$\begin{aligned}\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\pi(i)}) =& \mathbf{1}_{\{a_i \neq 0\}} \cdot \max\{\mathcal{L}_{\text{CC}}(i, \pi(i)), \\ & \mathcal{L}_{\text{MLO}}(i, \pi(i))\},\end{aligned} \tag{6}$$

where $\mathcal{L}_t$ is the same cost function as in [1] which contains classification cost and regression cost. $t \in \{\text{CC}, \text{MLO}\}$ denotes CC view or MLO view. The notation in Equation 6 is the same as Equation 17 in the main text. Instead of calculating the average of $\mathcal{L}_{\text{CC}}$ and $\mathcal{L}_{\text{MLO}}$, we adopt the larger one of the two costs. Since if one cost is significantly lower than the other, averaging them will let matching process be biased to the lower one. The final loss function can be written as follows:

$$\mathcal{L} = \mathcal{L}_{\text{D}}^{\text{CC}} + \mathcal{L}_{\text{D}}^{\text{MLO}}, \tag{7}$$

where $\mathcal{L}_{\text{D}}^{t}$ is the loss function in DETR, and $t \in \{\text{CC}, \text{MLO}\}$ denotes CC view or MLO view.

4.3 Implementation Details

Experimental details of Pair Verification and Paired Lesion Query are almost the same as CL-Net in the main paper. We will elaborate on the different parts in the following.

Pair Verification. We set $\alpha = 0.75$ and $\gamma = 2.0$ for the focal loss $\mathcal{L}_{\text{V}}$.

Paired Lesion Query. The number of layers in decoder N^{d} is set to 3 by default.

5 Error Analysis

To analyze the limitations of the proposed approach, we have visualized some typical cases with obvious errors made by CL-Net. As shown in the figure 3, there are mainly two types of mistakes: 1). the ground truth lesion could only be visible in MLO view (or CC view), while a FP box was detected in CC view (or MLO view). The FP box was associated with the TP box wrongly by Lesion Linker; 2). the model successfully discovered the ground truth lesion in one view while failed in the other view.

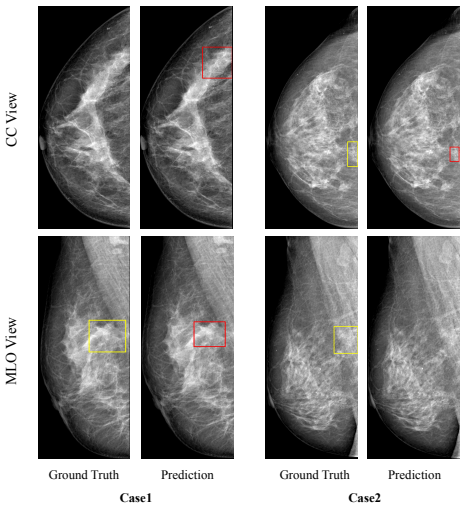

Fig. 3: Visualization of the cases with errors.

The main reason for these errors is that the ipsilateral information from mammogram is not sufficient for both deep learning models and radiologists to make an accurate diagnosis. One of the possible solutions is to leverage the bilateral (the same view of left and right breasts) information to further improve the detection performance, which could be our future work.

6 More Explanations about Dustbin Embedding

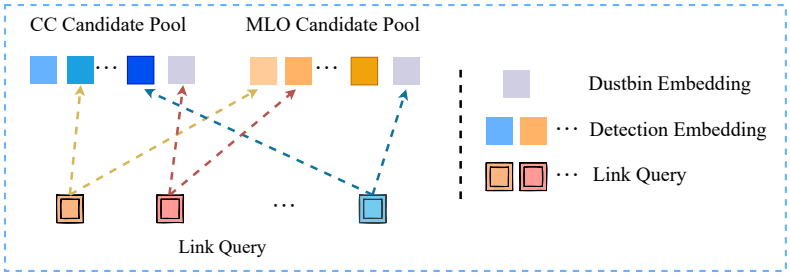

Fig. 4: The mechanism of dustbin embedding.

In Lesion Linker, we have firstly constructed the candidate pools of detection embeddings for MLO and CC views. Then, the link queries are responsible for cross-checking the suspicious detections and linking the same lesions in the two candidate pools. To handle the lesions which are visible only in one view, the mechanism of dustbin embedding is introduced. The dustbin embedding is concatenated with the detection embeddings to endow Lesion Linker the ability to make predictions only in one view, which means the detection embedding is associated with the dustbin embedding by link query. The mechanism is illustrated in Figure 4. Benefiting from the design of dustbin embedding, Lesion Linker can deal with different situations flexibly.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020) 3, 4
2. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017) 1
3. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017) 3
4. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS (2015) 1
5. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020) 1