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
zhaozw@stu.pku.edu.cn, chenyihong@pku.edu.cn
[2] Key Laboratory of Machine Perception, MOE, School of Artificial Intelligence, Peking University
wangdongcis@pku.edu.cn, wanglw@cis.pku.edu.cn
[3] Yizhun Medical AI Co., Ltd
ziteng.wang@yizhun-ai.com
[4] Pazhou Laboratory (Huangpu)
[5] Peng Cheng Laboratory

Abstract. Detecting mass in mammogram is significant due to the high occurrence and mortality of breast cancer. In mammogram mass detection, modeling pairwise lesion correspondence explicitly is particularly important. However, most of the existing methods build relatively coarse correspondence and have not utilized correspondence supervision. In this paper, we propose a new transformer-based framework CL-Net to learn lesion detection and pairwise correspondence in an end-to-end manner. In CL-Net, View-Interactive Lesion Detector is proposed to achieve dynamic interaction across candidates of cross views, while Lesion Linker employs the correspondence supervision to guide the interaction process more accurately. The combination of these two designs accomplishes precise understanding of pairwise lesion correspondence for mammograms. Experiments show that CL-Net yields state-of-the-art performance on the public DDSM dataset and our in-house dataset. Moreover, it outperforms previous methods by a large margin in low FPI regime.

Keywords: Pairwise Lesion Correspondence, Mammogram Mass, Object Detection

1 Introduction

With the highest incidence of cancers in women, breast cancer has become a serious threat to human health worldwide. In recent years, mammography screening has been used by most hospitals as a common examination for its effectiveness and non-invasiveness. Detecting mass is one of the core objectives for mammography screening since mass behaved spiculated and irregular is a typical sign

* Equal contribution.

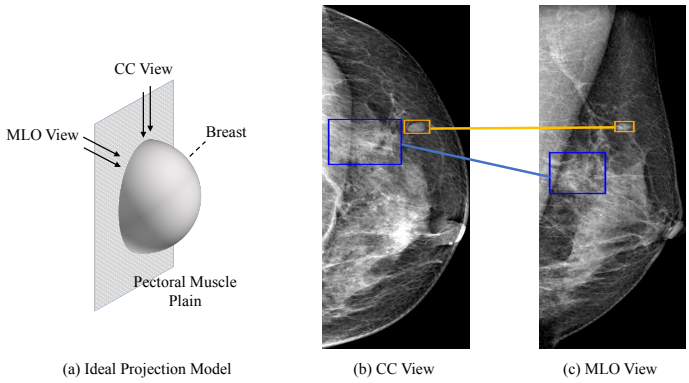

Fig. 1: (a) In mammograms, the CC view is a top-down view, while the MLO view is taken at a certain angle from the side. (b - c) show an example. Lesions connected by a line from different views are the projections of the same mass instance.

of breast cancer. However, gland overlap and occlusion are great obstacles for distinguishing mass from the gland, accordingly identifying suspicious lesions on mammogram is difficult for both radiologists and deep learning models.

In clinical practice, as shown in Figure 1, each breast is taken from two different angles, which are cranio-caudal (CC) view and mediolateral oblique (MLO) view, respectively. The complementary information of the ipsilateral view (CC view and MLO view of the same breast) will help radiologists to make better decisions for lesion detection. They usually cross-check the possible lesion locations in CC view and MLO view repeatedly. Once the relevant evidences are found in both views, the existence of the lesion can be confirmed. We call the co-existence of the same mass manifestations in both of the two views **pairwise lesion correspondence**. An example breast mammogram with two lesion pairs is shown in Figure 1 (b-c).

As for deep models, it is also particularly important to model the pairwise lesion correspondence explicitly for lesion detectors. Firstly, once the model is empowered with the ability to model pairwise lesion correspondences, the complementary information from the auxiliary view will help to distinguish the suspicious regions of the examined view, which is in line with the analysis logic of radiologists. Besides, the correspondences are also important supervision signals to train the network. The supervision of pairwise correspondences can guide the network to establish more accurate relations across the two views, which can further improve the detection performance.

Previous works have attempted to model lesion correspondences [20,17,18,35], however, the correspondence captured by these works is not accurate enough to represent pairwise lesion correspondence. For example, previous SOTA method BG-RCNN [17] divides the image into multiple parts and builds part-wise correspondence using a graph neural network (Figure 2(a)), which leads to relatively

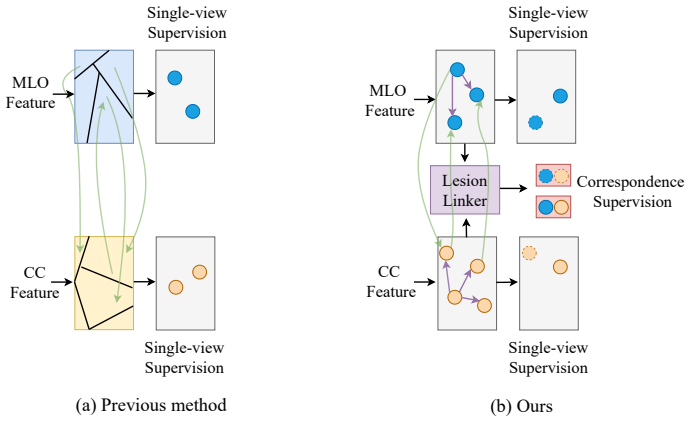

Fig. 2: (a) Previous methods [17] split mammograms into several parts and model part-level relationships without correspondence supervision. (b) We establish lesion-level interaction across views, and leverage correspondence supervision to guide the network training procedure explicitly.

coarse correspondence. Meanwhile, the utilization of correspondence supervision is not considered by previous works.

In this paper, we propose **CL-Net** to model more precise pairwise lesion correspondence. We design a transformer-based network structure to learn the lesion detection and pairwise correspondence in an end-to-end manner. As shown in Figure 2(b), our CL-Net can not only build pairwise lesion correspondence for lesion candidates detected by single-view lesion detectors, but also leverage correspondence supervision to guide the network training procedure for discovering accurate and sensible pairwise lesion correspondence.

Specifically, we first propose **View-Interactive Lesion Detector** (VILD) to achieve dynamic interaction across lesion candidates of MLO and CC views. We build our model upon modern transformer-based object detectors (*e.g.*DETR [5]). These detectors often adopt the query mechanism, where each object query can be regarded as an abstract representation of a lesion candidate, and the information flow across queries is suitable for capturing the correspondences for lesion pairs. Therefore, we apply the inter-attention layer between the object queries' outputs of MLO and CC views to build relationships for the two views, which captures pairwise lesion correspondences in an elegant and efficient way.

Furthermore, we propose **Lesion Linker** to learn the precise pairwise lesion correspondence during network training. Lesion linker summarizes all lesion information from MLO and CC views by taking all lesion candidates generated by object queries as inputs, and then employs the **link query** and a decoder-like net structure to produce paired lesion outputs. Like DETR, we use a set prediction approach to output lesion pairs, where each pair is a point in the set. Hence, the lesion linker can also be trained with a set matching loss. Under the guidance of

pair supervision, the inter-attention layer in VILD can obtain more direct and precise correspondence, which will benefit the training of the detector.

Experimentally, the results show that the proposed approach outperforms the previous SOTA methods by a large margin on both the public dataset DDSM [13] and an in-house dataset. The ablation study validates the effectiveness of each part in our design.

In a nutshell, our contributions are three-folds:

- To the best of our knowledge, our work is the first to model and learn pairwise lesion correspondence explicitly for mammogram mass detection, which is essential for cross-view reasoning.
- VILD and lesion linker are proposed to achieve precise lesion correspondence.
- We propose a novel framework, which achieves a new SOTA performance for mammogram mass detection with ipsilateral views and surpasses all previous methods by a large margin.

2 Related Work

2.1 Mammogram Mass Detection

Traditional approaches [3,11,21,31] usually use complex preprocessing and design hand-crafted features for mammogram mass detection. However, due to the low representation ability, the performance of these methods is not satisfactory. In the past few years, deep learning has been introduced to this area. Most of works [4,1,34,26] only use a single view for detection, while recently several studies [20,17,18,35] attempt to establish cross-view reasoning mechanism for mammogram mass detection. Ma *et al*. [20] and Yang *et al*. [35] use relation module [14] to model the relationships of lesion proposals across views. Liu *et al*. [17] seeks to leverage bipartite graph convolutional network to achieve part-level correspondence. C2-Net [18] preprocesses the mammograms for column-wise alignment and performs column-wise correspondence between cross-views, since they assume that the perpendicular distance to the chest of the same lesion in CC view and MLO view is roughly the same. Although these methods model the correspondence of the two views to a certain extent, however, the correspondence is generated freely without any pairwise supervision. Perek *et al*. [24] proposes a Siamese approach to achieve cross-view mass matching, while the performance of mass detection is not considered. Different from above approaches, our CL-Net can model and learn the pairwise lesion correspondence explicitly, which significantly improves the detection performance.

2.2 Object Detection and HOI Detection with Transformer

Transformer [32] has drawn great attention in computer vision recently [5,38,9,33]. In the area of object detection, the first representative of the transformer-based detector is DETR [5]. DETR employs a transformer encoder-decoder architecture with object queries to hit the instances in the images. It regards object

detection as a set prediction task, and uses a set matching method [5] to train the network. Afterwards, Deformable DETR [38] is proposed as a variant of DETR. Deformable DETR uses the local receptive fields for attention layers, which reduces computational complexity significantly and speeds up convergence. Moreover, DETR has also been appied to the task of Human-Object Interaction detection [15,6,39,37]. Chen *et al*. [6] and Zou *et al*. [39] reformulate HOI detection as a set prediction task and predict humans, objects and their interactions directly. HOTR [15] utilizes HO Pointers to associate the outputs of two parallel decoders, which leverages the self-attention mechanisms to exploit the contextual relationships between humans and objects. It is worth mentioning that HOI detection focuses on predicting the associations of humans and objects, while in this paper mammogram mass detection is evaluated by the detection results of each single image view. Different from HOI detection, the correspondence of MLO and CC views is regarded as the auxiliary supervision to promote the detection model. Our proposed lesion linker takes the advantage of this supervision to guide the training of VILD.

2.3 Learnable Image Matching

The well-known image matching in computer vision aims to establish dense correspondences across images for camera pose recovery and scene structure estimation in geometric vision tasks, such as Structure-from-Motion (SfM) and Simultaneous Localization and Mapping (SLAM) [8,10,22,36,25,2,28,30]. These methods rely on dense interest points as local descriptors to build pixel-to-pixel dense correspondences for multiple views. However, in mammograms, the two views are two different projections of 3D breast, which means there is no precise pixel-to-pixel correspondence. Therefore, we can only model the sparse lesion-to-lesion correspondence for accurate lesion detection. Compared with pixel-level matching, extracting the pairwise lesion correspondence is a high-level vision task that requires the network to understand lesion instances in advance. Therefore, we design the lesion linker and use link queries after the detector to learn lesion matching.

3 Methodology

In this section, we will elaborate on the design of the proposed CL-Net. The name CL stands that our method can Cross-**C**heck the two views and **L**ink the corresponding lesions across views. An overview of the whole pipeline is illustrated in Figure 3. To be specific, we first explain how the proposed View-Interactive Lesion Detector (VILD) along with the Lesion Linker establishes pairwise lesion correspondence. Then we discuss how to effectively train the network by presenting the training details including label assignment rules for lesion correspondence and the final loss function.

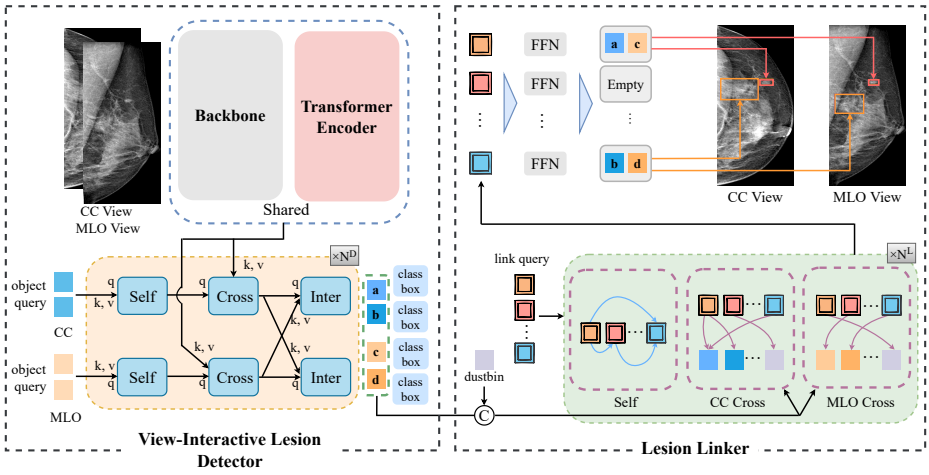

Fig. 3: Overview of our proposed CL-Net. A pair of mammograms are firstly processed by View-Interactive Lesion Detector (VILD) to achieve dynamic interaction across lesion candidates of MLO and CC views. Then, the embeddings outputted by the last decoder layer in VILD are provided for Lesion Linker to learn the precise pairwise lesion correspondence by learnable link queries.

3.1 Reviewing DETR

Recently, DETR [5] has drawn great attention since it proposes a novel paradigm for object detection through transformer encoder-decoder architecture. It reformulates object detection as a set prediction task and adopts one-to-one label assignment between ground truth and predicted objects, which achieves an end-to-end object detector.

Multi-head Attention. Attention is the core component in Transformer architecture. The standard version of attention can be written as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{1}$$

where Q, K, V stand for query vector, key vector and value vector, respectively. d_k is the vector dimension.

Multi-head attention is the extension of the standard version:

$$\text{MultiHeadAttn}(Q, K, V) = \text{Concat}(H_1, H_2, ..., H_m), \tag{2}$$

$$H_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{3}$$

where m is the number of heads, W_i^Q, W_i^K, W_i^V are projection matrices in the i-th head to map the original vector into a vector with lower dimension. For convenience, we use $\mathcal{M}$ to denote MultiHeadAttn in the following.

Object Query. In DETR, object queries can be regarded as abstract representations of objects. After image features are extracted by the transformer

encoder, queries will interact with these features through self-attention and cross-attention layers in the transformer decoder to aggregate instance information. Finally, several feed-forward network (FFN) layers are applied to decode box location and class information for each object query.

3.2 View-Interactive Lesion Detector

Dynamic interaction across lesion candidates of MLO and CC views is very helpful in establishing pairwise lesion correspondence. Therefore, we first propose the View-Interactive Lesion Detector (VILD) which aims to transfer lesion information across views effectively. The architecture of VILD is elaborated in the left part of Figure 3. VILD is a transformer-based detector that also employs object query as an abstraction of object. VILD takes mammogram of MLO and CC views as input and passes them through the shared backbone and feature encoder to encode the image content. Afterwards, two sets of object queries (one for MLO and one for CC) are fed into a specially designed decoder to predict lesions' position and class for each view while taking the lesion information from the ipsilateral view into consideration.

To be specific, we append an additional inter-attention layer at the end of each transformer decoder block to achieve dynamic interaction across views. Object queries can be regarded as abstract representations of objects, thus directly applying cross-view inter-attention can be realized as an elegant and efficient way to capture pairwise lesion correspondence. The cross-view inter-attention is also instantiated as a multi-head attention block which takes intermediate embedding of one view as queries and intermediate embedding of the other view as keys and values. Formally, suppose the number of object queries of each view is N and denote the embeddings output by cross-attention layer in the i-th decoder layer as $E_i^c,\ E_i^m \in \mathbb{R}^{N\times D}$ for CC view and MLO view respectively, then the enhanced embeddings are obtained through attention mechanism which could be expressed as (take CC view for example),

$$E_i^{c*} = E_i^c + \mathcal{M}(E_i^c + P^c, E_i^m + P^m, E_i^m), \tag{4}$$

where P^m and P^c denote the positional encodings for MLO and CC view's embedding, respectively. The positional encodings are learnable vectors, which are the same as Deformable DETR. $\mathcal{M}$ is MultiHeadAttn as defined in 3.1. The enhanced embedding for MLO view is obtained vice versa:

$$E_i^{m*} = E_i^m + \mathcal{M}(E_i^m + P^m, E_i^c + P^c, E_i^c). \tag{5}$$

By passing through the decoder layer for several times, cross-view lesion correspondence is gradually transferred and formed bidirectionally with the help of inter-attention block. This aligns with how radiologists identify lesions. They usually search for potential lesions in both views back and forth. Once a suspicious region is discovered in one view, they will check all possible positions in the other view in order to find the corresponding lesion with similar spatial and visual information.

Denote the embeddings for CC view and MLO view outputted by the last decoder layer as E^c, $E^m \in \mathbb{R}^{N\times D}$, the detection results are then predicted by FFN layers with E^c and E^m as input.

3.3 Lesion Linker

In VILD, the establishment of cross-view dynamic interaction endows lesions from one view with the ability to form correspondence with lesions from the other view. We argue that by explicitly utilizing the guidance of the pair supervision, a more accurate pairwise lesion correspondence could be achieved and the detection ability of the network could be further boosted. We propose Lesion Linker, a transformer decoder-like structure to take full advantage of the pair supervision. The architecture of lesion linker is illustrated in the right part of Figure 3. Lesion linker adopts link query, which is initialized as a set of learnable vectors, as abstract representations of possible pairwise relationships. Given output embeddings E^c and E^m from VILD as input, link queries will interact with them to extract lesion information and gradually focus on specific lesion pairs. Each link query will finally predict a triplet including link embeddings for CC and MLO views and lesion pair score through FFN layers. Given these embeddings, corresponding detection results in MLO view and CC view could be linked together to form pairwise lesion detection results. In the following, we will elaborate on the key designs of our lesion linker.

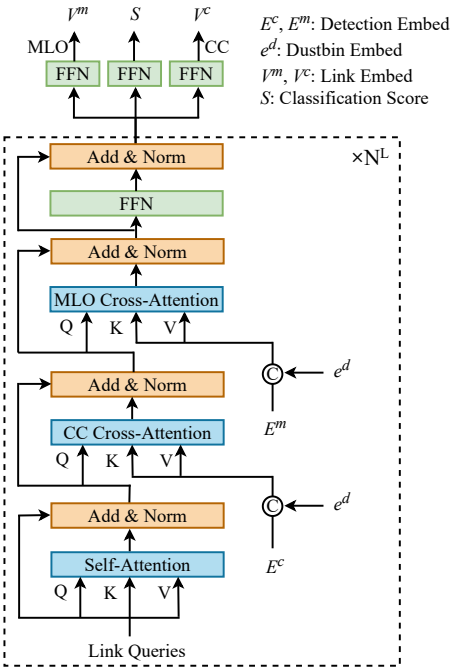

Fig. 4: Architecture of Lesion Linker.

Dustbin Embedding. In clinical practice, mammogram is a projection along the X-ray direction in which lots of information is lost, thus some mass instances can only be seen in one view. To cope with this special situation for lesion linker, we set a learnable vector $e^d \in \mathbb{R}^D$, named dustbin embedding. Detection embeddings that have no correspondence should be linked to it.

We concatenate detection embeddings from CC view and MLO view with dustbin embedding to obtain the complete version $\tilde{E}^c, \tilde{E}^m \in \mathbb{R}^{(N+1)\times D}$:

$$\tilde{E}^c = \text{Concat}(E^c, e^d), \tag{6}$$

$$\tilde{E}^m = \text{Concat}(E^m, e^d). \tag{7}$$

More explanations about dustbin embedding can be found in the supplementary materials.

Architecture. As illustrated in Figure 4, at first link queries Q are passed through the multi-head self-attention layer. Then they will interact with detection embeddings from CC and MLO views sequentially to extract view-dependent information to form pairwise relationships. The process could be written as

$$\dot{Q} = Q + \mathcal{M}(Q, Q, Q), \tag{8}$$

$$\ddot{Q} = \dot{Q} + \mathcal{M}(\dot{Q}, \tilde{E}^c + P^c, \tilde{E}^c), \tag{9}$$

$$\hat{Q} = \ddot{Q} + \mathcal{M}(\ddot{Q}, \tilde{E}^m + P^m, \tilde{E}^m), \tag{10}$$

where P^c and P^m denote the positional encoding for $\tilde{E}^c$ and $\tilde{E}^m$. Finally, $\hat{Q}$ is processed by a FFN layer to further enhance the representative ability. Above layers can be stacked for several times. Link queries transformed by stacked attention have fully explored lesion-level relationships from MLO view and CC view and the pairwise lesion correspondence is gradually formed.

Motivated by [15], at the top of lesion linker, we decode the correspondence by applying three FFN layers to predict the link embedding for CC and MLO views $V^c \in \mathbb{R}^{M\times D}$, $V^m \in \mathbb{R}^{M\times D}$ and lesion pair classification score $S \in \mathbb{R}^{M\times 1}$, respectively. M is the number of the link queries and D is the feature dimension. The predicted link embeddings V^c and V^m are used for indexing the detection results, which will be introduced later. The classification score S denotes the confidence that whether the pair of detection results captured by the link query is true positive.

Lesion Correspondence Extracting. The output of lesion linker can be reformulated as a set of M triplets, $\{\langle v_i^c, v_i^m, s_i\rangle\}_{i=1}^M$, where $v_i^c, v_i^m \in \mathbb{R}^D$ and $s_i \in \mathbb{R}^1$ are the i-th row of V^c, V^m and S. The pairwise lesion correspondence could be explicitly established by first calculating the feature similarity between detection embeddings $\tilde{e}_j^t \in \mathbb{R}^D$ and link embeddings v_i^t for each view and then taking the index of the detection embedding with the highest similarity as result. Here $t \in \{c, m\}$ denotes CC view or MLO view, and $\tilde{e}_j^t$ is the j-th row of $\tilde{E}^t$. Formally, this process could be expressed as

$$c_i = \arg\max_j(\text{sim}(v_i^c, \tilde{e}_j^c)),\ \ m_i = \arg\max_j(\text{sim}(v_i^m, \tilde{e}_j^m)), \tag{11}$$

where we use cosine similarity to measure the feature similarity:

$$\text{sim}(x, y) = \frac{x^T y}{||x||_2||y||_2}. \tag{12}$$

Finally, for each link query q_i, we could obtain its extracted lesion correspondence pair $\langle c_i,\ m_i\rangle$ as result. Next we will discuss how to effectively train our network.

3.4 Training Details

We will elaborate on training details of our proposed CL-Net in this section. To be specific, we first explain the label assignment rule for pairwise lesion correspondence. Then we introduce the loss function of our CL-Net.

Label Assignment for Lesion Correspondence. In original DETR, a one-to-one label assignment based on bipartite matching is used to assign training targets for the predicted bounding boxes. In our CL-Net, we also aim to establish a similar rule to assign the pairwise ground truth lesion boxes to the set of link triplets predicted by lesion linker.

Our VILD shares a similar structure and training strategy with DETR, therefore through the label assignment rule for detection, we can obtain the assignment relationships between ground truth boxes and detection embeddings. Thus the pairwise ground truth boxes can be naturally converted to pairwise detection embeddings. We denote the conversion results as $y = \langle e^c, e^m, a = 1\rangle$. e^c and e^m denote the detection embedding converted from ground truth boxes for CC view and MLO view. For a lesion that can only be viewed in CC view, the converted result is $y = \langle e^c, e^d, a = 1\rangle$, in which e^d denotes the dustbin embedding defined in Section 3.3. The same is for the lesion that can only be viewed in MLO view.

Suppose the number of unique ground truth lesions is K. Then the set of converted lesion triplets from the ground truth lesions could be denoted as $Y = \{y_i\}_{i=1}^{K}$. The set of M predictions from lesion linker could be similarly denoted as $\hat{Y} = \{\hat{y}_j = \langle v_j^c, v_j^m, s_j\rangle\}_{j=1}^{M}$. Since K is less than M in mammogram, we pad the ground truth set Y with $\langle \varnothing, \varnothing, a = 0\rangle$ (no lesion pair) to the size of M, similar to DETR. We aim to find an optimal bipartite matching between these two sets by searching for a permutation of M elements $\pi \in \Pi_M$ with the lowest cost:

$$\hat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^{M} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\pi(i)}), \tag{13}$$

where $\mathcal{L}_{\text{match}}$ is a matching cost between ground truth y_i and prediction $\hat{y}_{\pi(i)}$. We consider two aspects when calculating the matching cost, which are the prediction scores and the similarity of ground truth embeddings and predicted link embeddings:

$$\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\pi(i)}) = -\mathbf{1}_{\{a_i \neq 0\}} \cdot [\mathcal{L}_{\text{emd}}(i, \pi(i))]^{\alpha} \cdot [\mathcal{L}_{\text{score}}(i, \pi(i))]^{1-\alpha}, \tag{14}$$

where $\mathcal{L}_{\text{emd}}$ and $\mathcal{L}_{\text{score}}$ denote cost of feature similarity and classification score. The operation of $+\,1$ in Eq. 15 aims to guarantee that $\mathcal{L}_{\text{emd}}$ is positive.

$$\mathcal{L}_{\text{emd}}(i, j) = \beta \text{sim}(e_i^c, v_j^c) + (1 - \beta)\text{sim}(e_i^m, v_j^m) + 1, \tag{15}$$

$$\mathcal{L}_{\text{score}}(i, j) = s_j. \tag{16}$$

We adopt the weighted geometric mean of the feature similarity $\mathcal{L}_{\text{emd}}$ and classification score $\mathcal{L}_{\text{score}}$, in which $\alpha \in [0, 1]$ is the balance hyper-parameter. The ablation study of the cost function and analysis can be found in Table 5. β is set to 0.5 by default to adjust the ratio of feature similarity in CC view and MLO view. The optimal bipartite assignment can be obtained through the Hungarian algorithm efficiently as in [29].

Training Loss. The final loss function can be written as follows:

$$\mathcal{L} = \mathcal{L}_{\text{D}} + \mathcal{L}_{\text{Link}}, \tag{17}$$

where $\mathcal{L}_{\text{D}}$ is the loss function in DETR, $\mathcal{L}_{\text{Link}}$ is defined as

$$\mathcal{L}_{\text{Link}} = \sum_{i=1}^{M} [\mathbf{1}_{\{a_i \neq 0\}} \lambda_{\text{sim}} \mathcal{L}_{\text{sim}}(i, \pi(i)) + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}(a_i, s_{\pi(i)})], \tag{18}$$

where λ_{sim} and λ_{cls} are weight hyper-parameters. We adopt focal loss [16] as the loss function $\mathcal{L}_{\text{cls}}$ for lesion pair classification.

Following [15], we first calculate the similarity scores $S^t \in \mathbb{R}^{N+1}$, where $t \in \{c, m\}$ denotes the CC view and MLO view, and the j-th item of S^t is $\text{sim}(v^t_{\pi(i)}, \tilde{e}^t_j)$. Then, we use Cross-Entropy Loss to localize the ground truth embeddings:

$$\mathcal{L}_{\text{sim}}(i, \pi(i)) = \text{CrossEntropyLoss}(S^c, i) + \text{CrossEntropyLoss}(S^m, i). \tag{19}$$

3.5 Discussion of Match Learning Strategy

Our lesion linker learns the paired relationships of lesions by learning to predict MLO and CC embeddings which are close to the corresponding lesion pairs. Our match learning strategy is a soft way, which gradually pushes the link embeddings to get closer to the ground-truth embeddings during training. Although there are also other alternative approaches for this task, learning lesion matching is not trivial. In this subsection, we compare our method with two other seemingly reasonable solutions, to strengthen the advantages of our method.

Pair Verification. A straightforward solution to predict match pairs is to verify whether every two lesions from ipsilateral views are truly paired lesions. Following this design, we need to output a 2D matrix that represents the match probabilities of all possible pairs. The shape of the matrix should be $N \times N$, where N is number of lesion candidates per view. However, since the number of possible lesion pairs is much larger than the number of truly paired lesions, it is difficult to extract useful training signals from such a small amount of pairwise annotation information.

Compared to the verification approach, the introduction of link queries decouples the pairwise training from the number of object queries N. The number of link queries M could be in the same order of magnitude as N, thus the pair supervision signal could be fully utilized, leading to an easier optimization process.

Paired Lesion Query. Another seemingly straightforward way is to predict pairwise lesions with query mechanism directly. With this paired lesion query, the network can output a pair of detected boxes in the two views for each query. Then, the form of outputted lesion pairs is similar to the extracted lesion pairs of lesion linker. Therefore, we can also adopt a similar set matching loss to train the network. With the paired lesion query, the object query for each view is not required anymore.

However, the optimization of paired lesion query is much harder than our lesion linker. Our CL-Net first detects lesion candidates from each view (in VILD), therefore the lesion linker only focus on extracting the pairwise correspondence.

Table 1: Comparison with baselines and previous SOTA on DDSM dataset (%).

Method	R@0.25	R@0.5	R@1.0	R@2.0	R@4.0
Mask RCNN [17]	-	76.0	82.5	88.7	91.4
Mask RCNN, DCN [17]	-	76.7	83.9	89.4	91.8
Deformable DETR [38]	73.8	78.4	83.7	88.7	93.7
BG-RCNN [17]	-	79.5	86.6	91.8	94.5
CL-Net	**78.1**	**83.1**	**88.0**	**92.4**	**95.0**

Table 2: Comparison with previous works on DDSM dataset (%).

Method	R@t
Campanini *et al.* [3]	80@1.1
Eltonsy *et al.* [11]	92@5.4, 88@2.4, 81@0.6
Sampat *et al.* [27]	88@2.7, 85@1.5, 80@1.0
CVR-RCNN [20]	92@4.4, 88@1.9, 85@1.2
CL-Net	**96@4.4, 92@1.9, 89@1.2**

While using the paired lesion query, the detection of objects and pairing are performed in the same step, which increases the difficulty of network training and results in inferior performance.

We elaborate on the implementation details of above two methods in the supplementary materials. The experimental results are presented in section 4.4.

4 Experiments

4.1 Implementation Details

Our model is based on Deformable DETR [38] for its flexibility and fast convergence. We adopt ResNet-50 [12] pre-trained from ImageNet [7] as backbone. The number of object queries N and link queries M are set to 125 and 16, respectively. The loss weights λ_{sim} and λ_{cls} are 0.125 and 1.0 by default. We set $\alpha = 0.5$ and $\gamma = 2.0$ for the focal loss $\mathcal{L}_{\text{cls}}$. It is worth mentioning that since we mainly focus on the task of lesion detection, the final predictions come from VILD in the inference process.

We implement our network with PyTorch [23]. We train the network in an end-to-end manner on 8 GPUs for 25k iterations. For each GPU, we use 4 images containing two mammogram pairs. Following Deformable DETR [38], we train our model using AdamW Optimizer [19] with base learning rate of 2×10^{-4}. We use the same multiplied factors for learning rates as [38], while the learning rates of lesion linker parameters are multiplied by 0.25. In addition, we adopt cosine learning rate schedule with warm-up. To avoid overfitting, we use several data augmentation methods (random flip, random crop, random normalization) in training.

4.2 Datasets

We conduct experiments on the public DDSM dataset and our in-house dataset.

Table 3: Ablation study on different components of CL-Net on DDSM dataset (%). VILD: View-Interactive Lesion Detector. LL: Lesion Linker. "not using VILD" means we use Deformable DETR directly.

VILD	LL	R@0.25	R@0.5	R@1.0	R@2.0	R@4.0
		73.8	78.4	83.7	88.7	93.7
✓		76.1	81.7	86.4	91.7	94.4
	✓	74.1	80.1	86.0	88.7	93.4
✓	✓	**78.1**	**83.1**	**88.0**	**92.4**	**95.0**

DDSM dataset. DDSM [13] is a widely used public dataset. It contains 2620 patient cases, and each case has four images, including MLO view and CC view for both breasts. We use the same data split method with previous studies [17,20,27,3]. The original dataset does not provide lesion correspondence annotations, hence we fulfill the annotations with experienced radiologists.

In-house dataset. We collect an in-house mammography dataset with 3,160 cases. Each case is annotated by at least two experts. We randomly split the dataset into train, validation, and test set with the ratio of 8:1:1.

Evaluation Metric. We report recall (R) at t false positives per image (FPI) to evaluate the performance following [17,20]. The metric can be simplified as $R@t$.

4.3 Compare with State-of-the-art Methods

We compare our methods with previous works on DDSM dataset in Table 1 and Table 2. In Table 1, the results of Mask RCNN, Mask RCNN DCN, and BG-RCNN are from [17], and Deformable DETR is implemented by ourselves. In Table 2, we use the same FPIs as in CVR-RCNN [20] and compare our method with previous works. From these two tables, we can draw a conclusion that our CL-Net outperforms all baselines by a large margin and surpasses previous SOTA [17,20]. The performance of our method is more significant in low FPIs, outperforming BG-RCNN [17] by 3.6 at $R@0.5$, which could benefit clinical practice a lot. The results on the in-house dataset are reported in the supplementary materials. Similar improvement over baselines on in-house dataset also demonstrates the superiority of our approach.

4.4 Ablation Study

In this section, we elaborate on ablation studies for CL-Net. Other ablation experiments are presented in the supplementary materials.

Different Components of CL-Net. We ablate the impact of different components of CL-Net on detection performance in Table 3. There are mainly two important modules in CL-Net, VILD and lesion linker. As shown in the table, using VILD can significantly improve the detection performance, while the improvement of employing lesion linker alone is marginal. Considering that learning accurate correspondences relies on the expression ablitily of VILD, it

Table 4: Different strategies for match learning on DDSM dataset (%). PV: Pair Verification. PL Query: Paired Lesion Query.

Method	R@0.25	R@0.5	R@1.0	R@2.0
VILD	76.1	81.7	86.4	91.7
PV	75.7	82.1	86.4	90.7
PL Query	68.8	75.1	81.7	87.0
CL-Net	**78.1**	**83.1**	**88.0**	**92.4**

Table 5: Ablation study on cost function of label assignment. Default parameter is marked by *.

Method	α	R@0.25	R@0.5	R@1.0	R@2.0
Add	0.25	77.7	82.7	87.7	90.4
	0.5	73.8	81.1	86.4	90.0
	0.75	75.1	80.1	86.7	91.7
Mul	0.25	76.1	**83.7**	**89.7**	91.7
	0.5*	**78.1**	83.1	88.0	**92.4**
	0.75	74.8	80.4	85.4	90.0

is explainable that the contribution of lesion linker is limited without VILD. The effect of lesion linker in CL-Net is guiding the interaction process more precisely in the inter-attention layer of VILD. The experimental results also verifies our conjecture. The joint contributions of VILD and lesion linker improve the detection performance of VILD significantly (+2.0 at R@0.25).

Different Strategies for Match Learning. We present the results of different strategies for match learning in Table 4. The methods described in section 3.5 are adopted. Pair verification method yields similar performance as VILD solely, which indicates that it is hard for the model to mine useful correspondences from plenty of feasibilities. In addition, paired lesion query performs much worse than VILD (-7.3 at R@0.25) due to the difficulty of optimization. Our CL-Net achieves the best performance attributing the success to the design of link query.

Cost Function of Label Assignment. We investigate the effect of different forms of cost function for label assignment on our model in Table 5. Method 'Mul' denotes the cost function in Eq. 14, while 'Add' refers to the weighted sum of $\mathcal{L}_{\text{emd}}$ and $\mathcal{L}_{\text{score}}$, where α is also the weighting factor. The experimental results show that method 'Mul' achieves better performance than 'Add', which could be mainly attributed to the sensitivity to both $\mathcal{L}_{\text{emd}}$ and $\mathcal{L}_{\text{score}}$ in the form of multiplication.

The visualization and error analysis can be found in the supplementary materials.

5 Conclusion

In this work, we present CL-Net, a novel mammogram mass detector based on transformer architecture. Our CL-Net can not only model precise pairwise lesion correspondence, but also leverage correspondence supervision to guide the network training. The experimental results conducted on the public DDSM dataset and an in-house dataset show that CL-Net surpasses the state-of-the-art methods by a large margin.

Acknowledgement This work is supported by Exploratory Research Project of Zhejiang Lab (No. 2022RC0AN02), Project 2020BD006 supported by PKUBaidu Fund.

References

1. Agarwal, R., Diaz, O., Lladó, X., Yap, M.H., Martí, R.: Automatic mass detection in mammograms using deep convolutional neural networks. Journal of Medical Imaging **6**(3), 031409 (2019) 4
2. Brachmann, E., Rother, C.: Neural-guided ransac: Learning where to sample model hypotheses. In: ICCV (2019) 5
3. Campanini, R., Dongiovanni, D., Iampieri, E., Lanconelli, N., Masotti, M., Palermo, G., Riccardi, A., Roffilli, M.: A novel featureless approach to mass detection in digital mammograms based on support vector machines. Physics in Medicine & Biology **49**(6), 961 (2004) 4, 12, 13
4. Cao, Z., Yang, Z., Zhuo, X., Lin, R.S., Wu, S., Huang, L., Han, M., Zhang, Y., Ma, J.: Deeplima: Deep learning based lesion identification in mammograms. In: ICCV Workshops (2019) 4
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020) 3, 4, 5, 6
6. Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., Qian, C.: Reformulating hoi detection as adaptive set prediction. In: CVPR (2021) 5
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 12
8. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: CVPR workshops (2018) 5
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4
10. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint detection and description of local features. arXiv preprint arXiv:1905.03561 (2019) 5
11. Eltonsy, N.H., Tourassi, G.D., Elmaghraby, A.S.: A concentric morphology model for the detection of masses in mammography. IEEE transactions on medical imaging **26**(6), 880–889 (2007) 4, 12
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 12
13. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, W.P.: The digital database for screening mammography. In: Proceedings of the 5th international workshop on digital mammography. p. 212–218. Medical Physics Publishing (2000) 4, 13
14. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: CVPR (2018) 4
15. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: Hotr: End-to-end human-object interaction detection with transformers. In: CVPR (2021) 5, 9, 11
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017) 11
17. Liu, Y., Zhang, F., Zhang, Q., Wang, S., Wang, Y., Yu, Y.: Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection. In: CVPR (2020) 2, 3, 4, 12, 13
18. Liu, Y., Zhou, C., Zhang, F., Zhang, Q., Wang, S., Zhou, J., Sheng, F., Wang, X., Liu, W., Wang, Y., et al.: Compare and contrast: Detecting mammographic soft-tissue lesions with c2-net. Medical image analysis **71**, 101999 (2021) 2, 4

19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 12
20. Ma, J., Liang, S., Li, X., Li, H., Menze, B.H., Zhang, R., Zheng, W.S.: Cross-view relation networks for mammogram mass detection. In: ICPR (2020) 2, 4, 12, 13
21. Mudigonda, N.R., Rangayyan, R.M., Desautels, J.L.: Detection of breast masses in mammograms by density slicing and texture flow-field analysis. IEEE Transactions on Medical Imaging **20**(12), 1215–1227 (2001) 4
22. Ono, Y., Trulls, E., Fua, P., Yi, K.M.: Lf-net: Learning local features from images. arXiv preprint arXiv:1805.09662 (2018) 5
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Zeming Lin, e.a.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) NeurIPS (2019) 12
24. Perek, S., Hazan, A., Barkan, E., Akselrod-Ballin, A.: Siamese network for dual-view mammography mass matching. In: Image Analysis for Moving Organ, Breast, and Thoracic Images, pp. 55–63. Springer (2018) 4
25. Ranftl, R., Koltun, V.: Deep fundamental matrix estimation. In: ECCV (2018) 5
26. Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with deep learning. Scientific reports **8**(1), 1–7 (2018) 4
27. Sampat, M.P., Bovik, A.C., Whitman, G.J., Markey, M.K.: A model-based framework for the detection of spiculated masses on mammography a. Medical physics **35**(5), 2110–2123 (2008) 12, 13
28. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: CVPR (2020) 5
29. Stewart, R., Andriluka, M., Ng, A.Y.: End-to-end people detection in crowded scenes. In: CVPR (2016) 10
30. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: CVPR (2021) 5
31. Tai, S.C., Chen, Z.S., Tsai, W.T.: An automatic mass detection system in mammograms based on complex texture features. IEEE journal of biomedical and health informatics **18**(2), 618–627 (2013) 4
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017) 4
33. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: CVPR (2021) 4
34. Xi, P., Shu, C., Goubran, R.: Abnormality detection in mammography using deep convolutional neural networks. In: 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA). pp. 1–6. IEEE (2018) 4
35. Yang, Z., Cao, Z., Zhang, Y., Tang, Y., Lin, X., Ouyang, R., Wu, M., Han, M., Xiao, J., Huang, L., et al.: Momminet-v2: mammographic multi-view mass identification networks. Medical Image Analysis **73**, 102204 (2021) 2, 4
36. Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: CVPR (2018) 5
37. Zhang, A., Liao, Y., Liu, S., Lu, M., Wang, Y., Gao, C., Li, X.: Mining the benefits of two-stage and one-stage hoi detection. arXiv preprint arXiv:2108.05077 (2021) 5
38. 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) 4, 5, 12

39. Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., et al.: End-to-end human object interaction detection with hoi transformer. In: CVPR (2021) 5