

FEATURES FUSION FOR DUAL-VIEW MAMMOGRAPHY MASS DETECTION

Arina Varlamova* Valery Belotsky* Grigory Novikov* Anton Konushin† Evgeny Sidorov*‡

* Third Opinion Platform; † Higher School of Economics; ‡ Lomonosov Moscow State University;

ABSTRACT

Detection of malignant lesions on mammography images is extremely important for early breast cancer diagnosis. In clinical practice, images are acquired from two different angles, and radiologists can fully utilize information from both views, simultaneously locating the same lesion. However, for automatic detection approaches such information fusion remains a challenge. In this paper, we propose a new model called MAMM-Net, which allows the processing of both mammography views simultaneously by sharing information not only on an object level, as seen in existing works, but also on a feature level. MAMM-Net's key component is the Fusion Layer, based on deformable attention and designed to increase detection precision while keeping high recall. Our experiments show superior performance on the public DDSM dataset compared to the previous state-of-the-art model, while introducing new helpful features such as lesion annotation on pixel-level and classification of lesions malignancy.

Index Terms— instance segmentation, mammography, lesion detection

1. INTRODUCTION

Breast cancer is one of the most common diseases, ranking first among the causes of cancer mortality among women[1]. Approximately 12% of all diagnosed cases of cancer in women are related to breast cancer, making it the dominant cancer-related disease in many countries.

Screening digital mammography is the most widely employed modality for the successful diagnosis of breast cancer[2]. Presently, Artificial Intelligence (AI), mostly in the form of computer vision, is extensively utilized for tasks involving the automatic detection of breast cancer features, demonstrating high diagnostic accuracy comparable to or surpassing that of a radiologist[3].

One of the primary tasks in forming radiological descriptions of mammography studies is the identification and comprehensive description of breast tissue abnormalities in both

projections [4]. The differential diagnosis of lesions involves a comparative analysis of two views of each breast and between the two breasts. However, most current neural network architectures do not take into account the context of both views, limiting model performance. To the best of our knowledge, existing two-view approaches fuse information between different angles only on object level, using features obtained independently [5, 6, 7]. In our study we propose new approach called **MAMM-Net**, where network is able to fuse information on features level in addition to object level, imitating radiologist's diagnosis process more naturally. We observe that additional information fusion helps model to effectively filter false positive detections while preserving the high recall, thus allowing to achieve state-of-the-art results.

2. RELATED WORK

2.1. Dual-view lesion segmentation

The idea of dual-view segmentation was considered in several approaches. Liu *et al.* [5] use a bipartite graph convolutional network to incorporate the intrinsic geometric and semantic relations of ipsilateral views. Ma *et al.* [7] propose a relation module to model correspondence between mass ROIs from different mammography images. In CL-Net [6] authors used cross-attention between object queries, generated by Deformable DETR [8] and a special module called Lesion Linker to verify object pairs.

2.2. Object recognition

Transformer-based architectures gained popularity in tasks related to computer vision, including object detection and instance segmentation. Detection models originate from DETR [9], which was the first to introduce the concept of query-based instance proposal. The idea was further adopted for instance segmentation [10], resulting in SOTA architecture at the time of writing.

2.3. Multi-view segmentation

3D segmentation in multi-camera space was actively explored in BEV Transformers [11]. Feature fusion here was utilized through reprojection into the same 2D space and usage of deformable attention in the local proximity to points of interest.

arXiv:2404.16718v1 [eess.IV] 25 Apr 2024

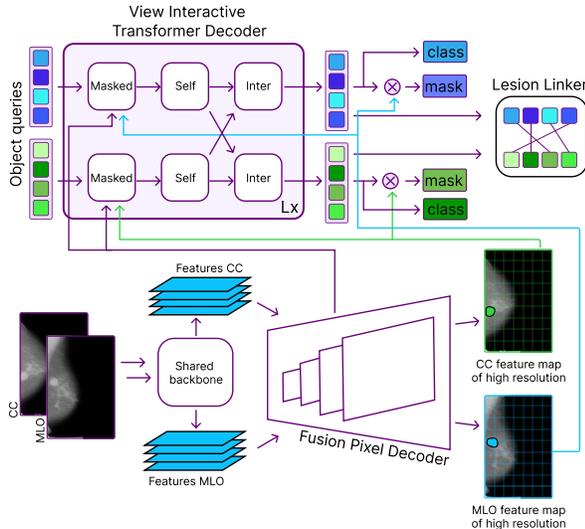


Fig. 1. General overview of **MAMM-Net**: 1) Two different views are processed by a shared backbone independently; 2) Generated feature maps are processed by Fusion Pixel Decoder, which provides fused feature maps for View-Interactive Transformer Decoder’s masked attention and feature maps of high resolution of both views for masks generation; 3) View-Interactive Transformer Decoder (VITD), consisting of blocks of masked-, self- and inter-attention, which outputs object queries, masks for both CC and MLO view, classification of found objects along with their malignancy scores; 4) Lesion Linker uses object queries from VITD to set correspondence between objects in CC and MLO views and outputs triplets of embeddings and pair classification.

However, such approach is tricky in mammography application, since angles between images are in general unknown. In this settings there is a need for more flexible approach for generation of reference points, which can be achieved through deformable attention introduced in [12].

3. METHODS

Our proposed architecture will be explored further in this section, which we will reference as **MAMM-Net** (Multi-view Attention for Mass Matching). A brief overview of its structure is shown in Fig. 1.

We organize methods as follows: firstly, we briefly introduce key points of Mask2Former [10] and CL-Net [6] architectures, which our model is based on, and then we explain our proposed Fusion Layer and View-Interactive Transformer Decoder (VITD) in more details.

3.1. Mask2Former

Our model mainly inherits the Mask2Former structure in its main components: backbone, pixel decoder, and transformer decoder with masked attention. The transformer decoder’s key component is the masked attention operator, which allows the usage of spatial features restricted to the foreground area of the predicted masks. We left this part mostly unchanged except for using two object query branches for crano-caudal (CC) and mediolateral oblique (MLO) mammography views and adding additional cross-attention logic between them.

Our Fusion Pixel Decoder, however, has more significant differences. Instead of generating feature maps independently, we combine them each time resolution increases. To provide intermediate feature maps of different resolutions, at each step we fuse features of CC and MLO projections into each other using our proposed Fusion Layers.

3.2. CL-Net

Key components of CL-Net we were interested in are the VILD (View-Interactive Lesion Detector) and LL (Lesion Linker) modules. The main idea behind VILD is to add a cross-view inter-attention step in addition to self- and cross-attention in DETR’s transformer decoder in order to help the model capture the relationship between objects in CC and MLO views. LL module at the same time aims to model correspondence between detected objects, outputting classification of the presence of object pair and link embeddings.

3.3. Fusion Pixel Decoder

The Fusion Pixel Decoder is designed to provide both high-resolution mask feature maps for CC and MLO views as well as multi-scale fused feature maps for the VITD’s masked attention. A brief overview of module structure is shown in Fig. 2.

The module consists of several consequent blocks for combining information between projections. Each block takes two feature maps of different resolutions from the shared backbone, which are passed through the Fusion layer in two branches. Generated feature maps are further concatenated and convolved using 1×1 convolution to maintain the same channel dimension. Such feature maps are generated from low to high resolution and merged by element-wise addition in an FPN [13] manner.

A key component of our pixel decoder is the Fusion layer, which is based on deformable attention [12]. The Fusion layer uses two feature maps: main to provide queries and reference from another view to provide key and values components. Since classic cross-attention between two feature maps is computationally demanding, it’s essential to focus on a limited subset of spatial positions for key-values. In order to achieve that, we use the predefined amount of uniformly distributed points to initialize the start position of candidates.

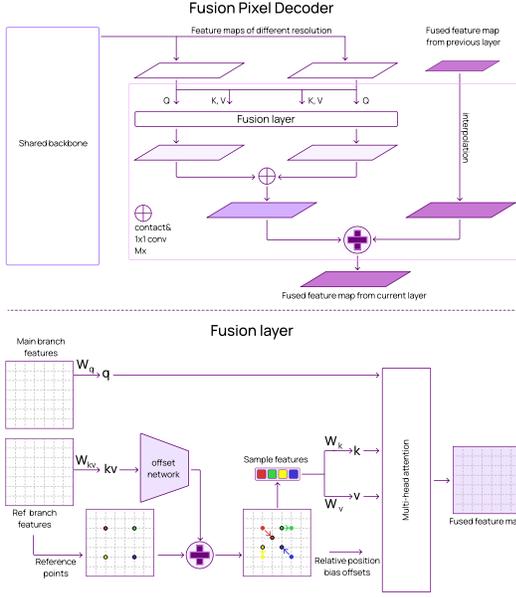


Fig. 2. Architecture of Fusion Pixel Decoder (upper) and Fusion Layer (lower). 1) Fusion Pixel Decoder: The module uses feature maps of different resolutions for both CC and MLO views. Starting from the lowest resolution, feature maps are fused into each other using a special Fusion Layer and then are combined in a FPN manner. Fused feature maps of low resolution are transferred to the VITD to use in masked attention. The last fused feature map is used to generate masks of high resolution; 2) Fusion Layer: the main feature map (Q in Fusion Pixel Decoder) is used as queries in the multi-head attention module. Key and values are sampled from the reference feature map (K, V in Fusion Pixel Decoder). Generated queries, keys, and values are processed by a multi-head attention block.

The special network generates their relative offsets, which are used to sample deformed key-values. Produced queries, keys, and values are then processed by a standard multi-head attention block.

It is worth noting that we use the Fusion Layer only in two blocks with feature maps with the lowest resolution. As resolution increases, we replace it with two independent 1×1 convolutions. Mask feature maps are generated by separately convolving the fused feature map with the highest resolution.

3.4. VITD

Following [10], blocks of our VITD consist of masked attention, self-attention, and FFN layers. We added an additional inter-attention layer to share object information between different views, similar to [6]. More formally, at i^{th} decoder's iteration, module uses $Q_i^v = f_Q(X^{v_{i-1}}) \in \mathbb{R}^{N \times C}$ and K_i ,

$V_i \in \mathbb{R}^{H_i W_i \times C}$ which are fused feature maps with applied transformations $f_K(\cdot)$ and $f_V(\cdot)$ respectively. X_i^v refers to query features at layer i^{th} and specific view (CC or MLO). In these terms, output of the masked attention layer may be defined as follows:

$$X_{m_i}^v = softmax(\mathcal{M}^{v_{i-1}} + Q^v K_i^T) V_i + X^{v_{i-1}}, \quad (1)$$

where attention mask at spatial location (x, y) $\mathcal{M}^{v_{i-1}}(x, y) = 0$ if the binarized output of the resized mask of corresponding view from a previous layer of decoder equals 1, and ∞ otherwise. To utilize both high- and low-level features, fused feature maps are fed to the decoder in a round-robin fashion.

After applying the self-attention layer to $X_{m_i}^v$ for both CC and MLO, query vector is passed as Q to inter-attention layer with the same view, and as K and V to inter-attention layer for another projection. Outputs are processed by the feed-forward network, providing X_i^v for the next iteration of the decoder.

Finally, we apply three feed-forward networks to each X_i^v to provide a classification of objects, their malignancy class, and masks embeddings E_i^v , which are further multiplied to feature maps of high resolution from Fusion Pixel Decoder in order to produce objects masks.

3.5. Loss

Our training loss is a combination of detection loss [10], linker loss [6] and malignancy loss. Detection and linker losses were implemented as in original publications, and a malignancy loss is a binary cross-entropy loss, where the target class is objects being malignant.

4. EXPERIMENTS

4.1. Implementation Details

We adopt EfficientNet-b3 [14] as a backbone. We set the number of object queries and link queries to 100 and 50 respectively. In our experiments, we used 10 blocks in VITD. A downsample factor of 4 was used to generate reference points in Fusion layers. We set the number of heads of multi-head attention to 8 in both VITD and Fusion Pixel Decoder.

The network was trained with a batch size equal to 5, each item in a batch containing one mammography pair. We train our model with a learning rate set to 1×10^{-4} . To avoid overfitting we used L2-regularization with a weight of 1×10^{-5} and a variety of augmentations, including image flip, rotation, brightness and contrast modification, and random scaling.

4.2. Datasets

We conducted our experiments using the DDSM dataset [15, 16]. Originally, the dataset does not provide matching between objects in different views. In order to obtain it, cases

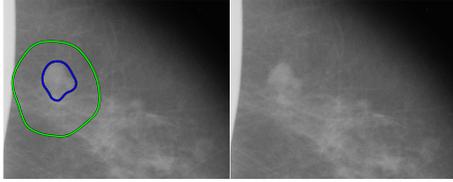


Fig. 3. Example of prediction (blue) and ground true (green). Intersection over Union for those two objects equals 0.15 although contours clearly indicate the same object.

Table 1. Comparison with previous SOTA on DDSM dataset %.

Method	R@0.25	R@0.5	R@1.0
Liu <i>et al.</i> [5] (AG-RCNN)		82.0	89.0
Zhao <i>et al.</i> [6] (CL-Net)	78.1	83.1	88.0
MAMM-Net (ours)	81.6	87.9	90.6

were annotated by a skilled radiologist. In such annotations, we didn’t change any contour properties or coordinates, but rather made a classification of objects being pair or not.

Similar to [7, 17, 6, 5], we use recall (R) at t false positives per image (FPI) to compare the performance of our model with other studies ($R@t$). Unlike in previous methods, we recall an object if it has IoU with ground true more than 0.1 since we have masks instead of bounding boxes and observe a relatively high proportion of true positives in diapason of [0.1, 0.2]. An example of such an object is shown in Fig. 3.

Data splits are another matter of discussion. Train splits proposed in [6, 17, 5] is ambiguous since Liu *et al.* [17, 5] mention 512 test cases and [6] doesn’t specify test set size. At the same time, they both refer to [18, 7], which uses 512 images. We decided to follow the test split proposed in [7, 18] with some modifications. Originally, images from cancer volumes without any masses were excluded. Since additional images without ground true objects can only worsen $R@t$, we selected all cases from these volumes for more representative comparison, resulting in 270 test cases and 1080 images.

4.3. Comparison with other studies

We show the comparison of our **MAMM-Net** with other methods in Table 1. Result from Table 1 are reported from [5, 6]. We keep the same FPIs as in [6] as the previous SOTA (CL-Net). It can be concluded that our model surpasses CL-Net by a large margin at all reported FPIs. It is worth noting, that we use a lighter backbone compared to ResNet-50 used in [5, 6], since we achieved similar performance as CL-Net in setting with fusion on object level only. We show more details on that in section 4.4.1. We believe that the result on $R@0.25$

Table 2. Comparison different components of our model on DDSM dataset %.

Method	R@0.25	R@0.5	R@1.0
VITD	78.2	83.3	87.3
Fusion Pixel Decoder	77.9	80.1	85.9
MAMM-Net (ours)	81.6	87.9	90.6

is of the most significance and additionally provide binary malignancy metrics (per mammary gland) for this setting, such as ROC-AUC (85.3), sensitivity (80.2) and specificity (76.2).

4.4. Ablation Study

We evaluate the performance of the networks that use only one fusion component, either the VITD or the Fusion Pixel Decoder, against the performance of the **MAMM-Net**, which incorporates both fusion blocks. Table 2 illustrates the differences in recall for selected FPI thresholds. We further discuss the implementation details of networks with only one fusion component.

4.4.1. VITD

In this setting, we used a pixel decoder from Mask2Former, which outputs two independent sets of multi-scale feature maps for both views. We got similar values to CL-Net (78.2 vs 78.1, 83.3 vs 83.1, 87.3 vs 88.0 at respectively), which is to be expected since both models have similar key components.

4.4.2. Fusion Pixel Decoder

In this setup the Fusion Pixel Decoder was left intact, while the VITD was significantly modified. Instead of two branches for different views, we used a single branch for object queries and excluded the inter-attention layer. For the mask in masked attention, we used a logical union of binarized mask outputs for both views. We skipped the LL block, using VITD’s output directly in the loss computation and forcing matched objects to be predicted in the same position. Similar to [6], we observe a significant drop in performance compared to the complete model. We support that forced prediction at the same position wasn’t the best choice for modeling relationships between different objects. However, it was observed that for lower FPIs Fusion Pixel Decoder performs better than our full model (*e.g.* 73.3 vs 69.7 $R@t=0.14$). We hypothesize that Fusion Layers help the model to effectively filter candidates that look worthy of attention only from one view.

5. CONCLUSION

The main novelty introduced in our paper is the Fusion Layer, which enables feature-level fusion of two projections, leading to a decrease in false positive predictions without an increase in false negatives. This component is integrated into our newly proposed MAMM-Net architecture, designed for effective object recognition across two projections. Experiments on the DDSM dataset have demonstrated that our architecture outperforms previous state-of-the-art models.

6. ACKNOWLEDGMENTS

No funding was received for conducting this study.

7. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data sourced from DDSM. Ethical approval was not required as confirmed by the license attached with the open access data.

8. REFERENCES

- [1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray, “Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.
- [2] Amanda Dibden, Judith Offman, Stephen W Duffy, and Rhian Gabe, “Worldwide review and meta-analysis of cohort studies measuring the effect of mammography screening programmes on incidence-based breast cancer mortality,” *Cancers*, vol. 12, no. 4, pp. 976, 2020.
- [3] Jung Hyun Yoon, Kyungwha Han, Hee Jung Suh, Ji Hyun Youk, Si Eun Lee, and Eun-Kyung Kim, “Artificial intelligence-based computer-assisted detection/diagnosis (ai-cad) for screening mammography: Outcomes of ai-cad in the mammographic interpretation workflow,” *European Journal of Radiology Open*, vol. 11, pp. 100509, 2023.
- [4] EA Sickles and Bassett LW D’Orsi CJ, “Acr bi-rads@ mammography,” *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*, 2013.
- [5] Yuhang Liu, Fandong Zhang, Chaoqi Chen, Siwen Wang, Yizhou Wang, and Yizhou Yu, “Act like a radiologist: towards reliable multi-view correspondence reasoning for mammogram mass detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5947–5961, 2021.
- [6] Ziwei Zhao, Dong Wang, Yihong Chen, Ziteng Wang, and Liwei Wang, “Check and link: Pairwise lesion correspondence guides mammogram mass detection,” in *European Conference on Computer Vision*. Springer, 2022, pp. 384–400.
- [7] Jiechao Ma, Xiang Li, Hongwei Li, Ruixuan Wang, Bjorn Menze, and Wei-Shi Zheng, “Cross-view relation networks for mammogram mass detection,” in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 8632–8638.
- [8] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” 2021.
- [9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, “End-to-end object detection with transformers,” 2020.
- [10] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar, “Masked-attention mask transformer for universal image segmentation,” *CoRR*, vol. abs/2112.01527, 2021.
- [11] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai, “Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers,” 2022.
- [12] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang, “Vision transformer with deformable attention,” 2022.
- [13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, “Feature pyramid networks for object detection,” 2017.
- [14] Mingxing Tan and Quoc V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” 2020.
- [15] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and W. Philip Kegelmeyer, “The digital database for screening mammography,” *Proceedings of the Fifth International Workshop on Digital Mammography*, 2001.
- [16] Michael Heath, Kevin Bowyer, Daniel Kopans, W. Philip Kegelmeyer, Richard Moore, Kyong Chang, and S. MunishKumaran, “Current status of the digital database for screening mammography,” *Digital Mammography*, 1998.

- [17] Yuhang Liu, Fandong Zhang, Qianyi Zhang, Siwen Wang, Yizhou Wang, and Yizhou Yu, “Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3812–3822.
- [18] 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*, vol. 49, no. 6, pp. 961, 2004.