A Deep Learning Approach for Efficient Registration of Dual View Mammography

Sina Famouri(✉), Lia Morra, and Fabrizio Lamberti

Dipartimento di Automatica e Informatica, Politecnico di Torino, Turin, Italy
{sina.famouri,lia.morra,fabrizio.lamberti}@polito.it

Abstract. In a standard mammography study, two views are acquired per breast, the Cranio-Caudal (CC) and Mediolateral-Oblique (MLO). Due to the projective nature of 2D mammography, tissue superposition may both mask or mimic the presence of lesions. Therefore, integrating information from both views is paramount to increase diagnostic confidence for both radiologists and computer-aided detection systems. This emphasizes the importance of automatically matching regions from the two views. We here propose a deep convolutional neural network for the registration of mammography images. The network is trained to predict the affine transformation that minimizes the mean squared error between the MLO and the registered CC view. However, due to the complex nature of the breast glandular pattern, deformations due to compression and the paucity of natural anatomic landmarks, optimizing the mean squared error alone yields suboptimal results. Hence, we propose a weakly supervised approach in which existing annotated lesions are used as landmarks to further optimize the registration. To this aim, the recently proposed Generalized Intersection over Union (GIoU) is exploited as loss. Experiments on the public CBIS-DDSM dataset show that the network was able to correctly realign the images in most cases; corresponding bounding boxes were spatially matched in 68% of the cases. Further improvements can be expected by incorporating an elastic deformation field in the registration network. Results are promising and support the feasibility of our approach.

Keywords: Mammography · Image registration · Spatial transformer · Convolutional neural networks

1 Introduction

Population screening by means of digital mammography was shown to reduce mortality associated to breast cancer. However, the 2D projective nature of mammography results in tissue superposition that may both mask and simulate the presence of lesions [13,20]. This is especially true when breast tissue is very dense, [18], as the fibrous and glandular components have higher attenuation than fatty tissue, and more similar to that of potential lesions, especially masses.

F.-P. Schilling and T. Stadelmann (Eds.): ANNPR 2020, LNAI 12294, pp. 162–172, 2020.
https://doi.org/10.1007/978-3-030-58309-5_13

In a standard screening examination, two projection views are acquired for each breast, named craniocaudal (CC) and mediolateral oblique (MLO) [3]. The breast is positioned between two compression plates; in the MLO view, the compression plates are rotated by 45°–50°, towards the axilla. The radiologist is thus able to locate suspicious areas on both views by triangulating from these projections. This increases the diagnostic confidence as false positives due to tissue superposition are likely to disappear in the contralateral view. Computer Aided Detection (CAD) algorithms have also shown reduced false positive rates when the two views are taken into account [3,14,19].

The objective of our research is to design and evaluate a registration network for CC-MLO registration based on emerging deep learning technologies. Applications range from enhancing image presentation to the radiologist, to improving the performance of lesion detection algorithms that operate on single-view images [12,14,19]. Unfortunately, registration of the breast is considerably more challenging than other imaging modalities as the soft tissues in the breast are compressed and distorted during the acquisition [4]. To the best of our knowledge, few authors have explored the registration of CC and MLO views, and no established deep learning approach exists for this task [4,5].

Given the difficulty of estimating the deformation field between the CC and MLO views, many works in literature have resorted to matching Regions of Interest instead. The goal is not necessarily to establish the exact correspondence between lesions, but to minimize the chance that true positives are matched with false positive detections. This technique has largely been explored in combination with CAD algorithms that detect candidate lesions, which are then matched based on a combination of position and visual similarity. Visual similarity can be estimated based on hand-crafted features such as texture, size, intensity, etc. [19] or, with the advent of deep learning, by training a Siamese Convolutional Neural Network (CNN) [14]. Compared to this standard candidate-matching approach, our proposed registration technique works directly on the input image, and can be applied before, after or independently of other lesion detection or classification networks. At the same time, it is a flexible and versatile module that can be incorporated and jointly trained in more complex pipelines.

Successfully training a registration CNN requires defining a robust loss while reducing the cost of annotation [5]. To this aim, we augment the standard Mean Squared Error (MSE) loss exploiting available lesion annotations in the form of bounding boxes. The Generalized Intersection over Union (GIoU) forces the registration to match true lesions across both views. Preliminary experiments on the CBIS-DDSM dataset (presented in Sect. 5) with an affine transformation support the feasibility of our approach.

2 Background and Related Work

2.1 Deep Learning for Medical Image Registration

Registration requires estimating the spatial coordinate transformation that maximizes some measure of similarity between two images, usually denoted as the *fixed* and *moving* images [4,12]. Conventional registration methods are based on numerical optimization techniques, and may differ based on the domain of the transformation (global, local), its nature (rigid, affine, or elastic) and the optimization procedure [4,22].

Recently, CNN-based techniques have been proposed to regress the registration transformation from pairs of unregistered images [5]. Available solutions include fully convolutional networks or encoder-decoder architectures for elastic transformations [2,8,11,15] and Spatial Transformer Networks for affine transformations [23]. For a comprehensive review on the topic, the reader is referred to a recent survey by Haskins and colleagues [5].

Compared to traditional optimization approaches, CNN-based approaches are poised to have a substantial advantage: even if the training process is slower and requires hundreds or thousands of image pairs, at inference time it is usually much faster than optimizing the transformation on each image pair.

One of the main obstacles to efficient CNN-based registration is defining a suitable loss. In principle, the registration can be trained from image pairs, without additional annotations, by defining a similarity metric, such as the MSE, and a regularization term (registration is a generally ill-posed inverse problem). This approach forms the basis of unsupervised approaches, such as Voxelmorph [2], which has been applied to the registration of several imaging modalities, such as brain, breast and cardiac magnetic resonance imaging [1]. However, defining a robust image similarity measurement is notoriously challenging, especially in the presence of different source modalities, anatomical deformations or temporal changes [5,8]. Unlike common registration tasks in brain, cardiac or abdominal images, mammography images are characterized by stronger changes in viewpoint and high tissue deformation induced by organ compression; this fact makes the task more complex and, to the best of our knowledge, the feasibility of registering mammographic images has yet to be established.

An alternative strategy is supervised training, which however requires marking an appropriate number of manually matched points. Such ground truth is usually difficult and expensive to obtain in the medical domain. In our case, the breast is highly compressible and lacks rigid structures, and hence very few anatomical landmarks can be accurately matched. Large calcifications have been used as landmarks for validating registration algorithms as their location and correspondence can be determined very precisely [21]. However, collecting a large number of such annotations would be time consuming, and such benign structures are usually disregarded in radiological reports.

Our methodology falls into the semi-supervised domain, exploiting existing partial annotations. A similar strategy was successfully applied to train prostate MR registration from organ segmentation maps [8]. Our setting is more challenging as we assume that only coarse bounding boxes are available for training.

Finally, our work shares some similarities with multi-task learning settings in which the registration task is jointly learned with another task. For instance, Qin *et al.* combined estimation of cardiac motion and segmentation for cardiac MRI in a single network with shared weights [15]. Our approach is complementary since the bounding boxes, which are in any case an approximate ground truth, are used to supervise directly the registration task.

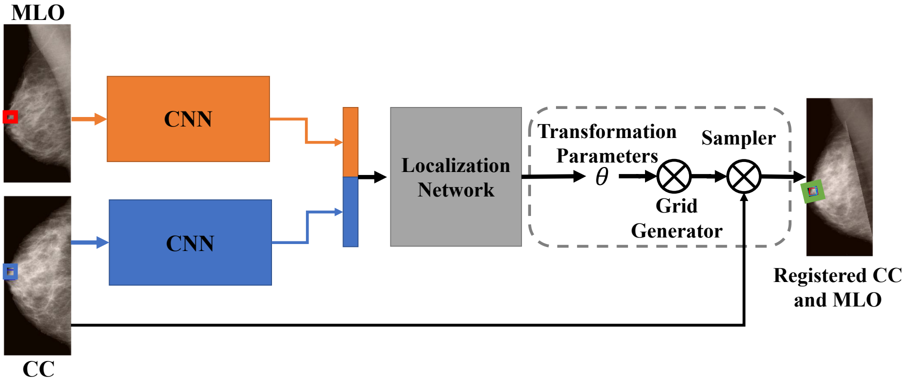


Fig. 1. Overall architecture of the registration network. From left to right: the CC and MLO views are passed through the shared convolutional layers; the feature map is concatenated and passed as input to the localization network; the CC image is registered by applying the estimated affine transformation parameters.

2.2 Spatial Transformer Networks

A Spatial Transformer network is a lightweight block which predicts and applies a spatial transformation to an input feature map during a single forward pass. It was proposed as a way to enhance an image classification network by allowing the network to transform feature maps to a canonical, expected pose to simplify inference in the subsequent layers [9]. The spatial transformer is composed of a localization network, which predicts the parameters of an affine transformation, which only requires six output parameters. Then, a sampling grid is created, that is a set of points where the input map should be sampled to produce the transformed output. Finally, the input feature map is resampled and interpolated to produce the output image (see Fig. 1). Spatial Transformers include a differentiable implementation of the sampling grid and resampling layer, allowing for end-to-end training, with standard back-propagation, of the models they are injected in. The network learns how to actively transform the feature maps to help minimise the overall cost function of the network during training.

3 Methodology

The proposed registration network is an end-to-end architecture which accepts as input a pair of unregistered CC and MLO images, and outputs the resampled CC image. We chose the MLO as fixed image and the CC as moving image since the former includes also the pectoral muscle, which is outside of the CC field of view. Registering the MLO to the CC would push the pectoral muscle out of the image pixels grid, and it would be impossible to estimate the correct deformation for the pixel belonging to the pectoral muscle.

The overall architecture, depicted in Fig. 1, is divided in two parts: the feature extraction block, and the Spatial Transformer block. The feature maps are extracted for each view separately, before being concatenated and passed to the Spatial Transformer network (introduced in Sect. 2.2). The proposed architecture implements an affine transformation, but can be easily extended to support other types of deformations by substituting the localization network. The architecture is trained in an end-to-end fashion exploiting the ground truth lesion bounding boxes as additional supervision. This provides cues for higher quality registration compared to the plain MSE.

The feature extraction backbone, marked as CNN in Fig. 1, is based on a ResNet50 network [6]. Specifically, we include up to the Conv4_x blocks. Weights are shared between views to reduce the number of parameters.

The Spatial Transformer is formed by a localization network and a resampling module. The localization network is made of a residual block (corresponding to the Conv5_x block of the ResNet50) followed by a dense layer to predict the parameters of the affine transformation:

$$\theta = \begin{bmatrix} a_{1,1} & a_{1,2} & t_1 \\ a_{2,1} & a_{2,2} & t_2 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

In the case of image registration, the sampling grid is simply the pixel grid of the fixed image, which greatly simplifies the implementation of the *grid generator* [9]. The output warped CC image is obtained by applying the affine transformation to this sampling grid using a bi-linear interpolation scheme.

The above resampling scheme can be applied indifferently to the original images (as done here), as well as to the feature maps (which could be useful if the feature maps were used for other tasks). Bounding boxes are converted by applying the inverse affine transformation and then rectifying the results. All layers including the bounding box registration are differentiable and, hence, can be trained end-to-end.

3.1 Loss

We argue that the MSE cannot by itself achieve successful registration. One of the underlying reasons is that the pectoral muscle is visible only in the MLO view. Experimentally, we observe that the CC may be overstretched over the pectoral muscle to achieve lower loss. If the registration is correct, the border of the CC should align to that of the pectoral muscle (see Fig. 2(a)).

To counterbalance this fact, we include in the loss only the region in which the moving CC image and the fixed MLO overlap (see Fig. 2(c)). The effect of the pectoral muscle, as well as of external air, is thus minimized. The resulting loss is defined as:

$$L_{MSE}(X^{mlo}, X^{cc}) = \left\|(X^{mlo} - X^{cc_{reg}})M\right\|^2 \qquad (2)$$

where X^{mlo} is the MLO image, $X^{cc_{reg}}$ is the CC view after registration and M is the binary overlap mask.

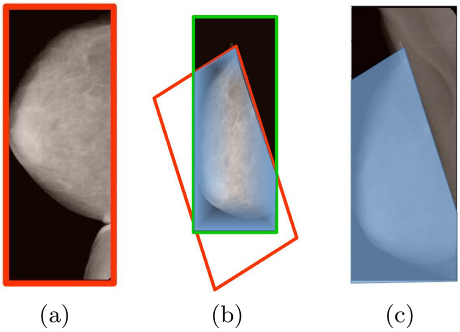


Fig. 2. Calculation of the overlap mask for the MSE loss. Unregistered (red box) and registered (green box) CC views are shown in (a) and (b). The shaded blue area is included in the calculation of the loss (b). In (c) the registered CC, fixed MLO and overlap mask are shown superimposed. It can be noticed how the margin of the CC view aligns with the pectoral muscle, outside of the overlap area. (Color figure online)

In order to exploit the lesion bounding boxes, we need a loss which reflects to which extent corresponding views are matched by the registration. The Intersection over Union (IoU) is a widely used measure to compare bounding boxes, but when the two bounding boxes do not overlap, the IoU is undefined. The recently proposed GIoU overcomes this limitation [16]. Given a pair of bounding boxes, it is defined as:

$$GIoU(B_i^{mlo}, B_i^{cc_{reg}}) = IoU(B_i^{mlo}, B_i^{cc_{reg}}) - \frac{A_c - U}{A_c} \qquad (3)$$

where B_i^{mlo} and $B_i^{cc_{reg}}$ are the two bounding boxes, A_c is the area of the smallest enclosing box that includes both and U is their union. In short, when the bounding boxes don't overlap significantly, the GIoU takes their relative distance into account.

The GIoU loss ($L_{GIoU} = 1 - GIoU$) was initially proposed as a regression loss to train object detection networks. To the best of our knowledge, this is the first time it is used for the purpose of registration. To conclude, for each pair of mammographic views the total loss is calculated as

$$L_{total} = L_{MSE} + \lambda L_{GIoU} \qquad (4)$$

where λ is a rescaling parameter.

4 Experimental Setup

Dataset. Our experiments were performed on the curated CBIS-DDSM collection [7,10]. Each study comprises up to 4 images including both CC and MLO orientation. We selected cases with benign and malignant lesions visible on both views. Based on the standard training/test split, we obtained 985 cases for the development set and 122 for the test set. The training set was further split into a training (75%) and validation (25%) set. Images were downsampled so that the largest dimension was equal to 600 pixels. We did not exploit metadata available in the DICOM images; although in digital mammography patient positioning and other useful information would be available in the image headers, the DDSM collection comprises only scannerized screen-film mammography. Images were converted to grayscale by replicating the intensity values across the RGB channels and normalized by subtracting the ImageNet mean. No other pixel normalization was applied.

Pretraining. The ResNet50 backbone is pretrained on the ImageNet dataset and finetuned for the task of object detection on the same CBIS-DDSM dataset. Specifically, it is pretrained using the Faster R-CNN for 80 epochs before transferring to the registration [17]. This allows faster convergence than transferring directly from ImageNet (results not reported due to space limitations). This observation opens interesting prospects for feature sharing across multiple tasks, which however are outside of the scope of these experiments.

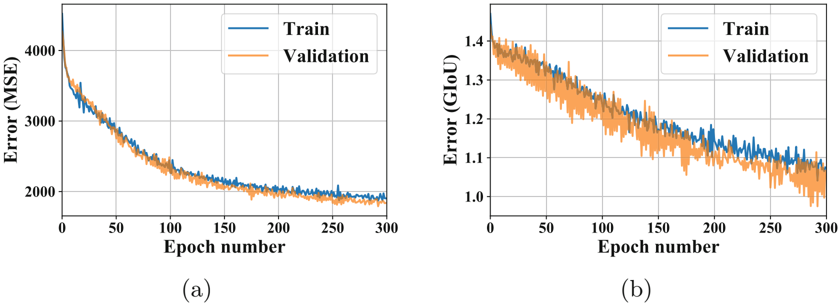
Hyperparameter Setup. Hyperparameters were experimentally finetuned on a smaller dataset. For the final training, we used the Adam optimizer (learning rate 10^{-4}, batch size 1). The network was trained for 300 epochs, each comprising 500 batches. The λ parameter (see Eq. 4) is set to 1000. The output dense layer of the Spatial Transformer is randomly initialized using Glorot initialization. The affine transformation parameters bias parameters are initialized to a 45 degree counterclockwise rotation, which is based on prior knowledge of the acquisition process. The network was implemented in Keras 2.2 with Tensorflow 1.13.1. All experiments were conducted on an AWS px2.large GPU instance.

Evaluation. Evaluation is not straightforward given the absence of a ground truth. Since the GIoU takes into account both the intersection and the distance of each pair of bounding boxes, we consider it as a viable evaluation metric. In addition, we visually inspected the registration results for the test set.

5 Results

The network was trained for 300 epochs without showing signs of overfitting (see Fig. 3). Both MSE and GIoU decreased indicating a synergistic behaviour of the two losses. When two bounding boxes do not overlap ($IoU = 0$), the

GIoU loss simplifies to $L_{GIoU} = 2 - \frac{U}{A_c} \geq 1$ [16]. In order to minimize $\frac{U}{A_c}$, the distance between the two bounding boxes must be reduced to the point where they eventually overlap.



(a) (b)

Fig. 3. Evolution of the loss during training: MSE (a) and GIoU (b)

The distribution of the L_{GIoU} on the test set is shown in Fig. 4. The bounding boxes for the registered CC and MLO overlap in 66.7% of the cases, which is an encouraging result. When L_{GIoU} approaches 0, the bounding boxes tend to perfectly overlap. In practice, due to the rectification process, the bounding boxes are unlikely to achieve perfect overlap, and lower IoU values are to be expected. Visually, in the large majority of cases the registration was successful in aligning the two views in terms of shape and global features, although an evaluation by a trained radiologist would be needed for confirmation.

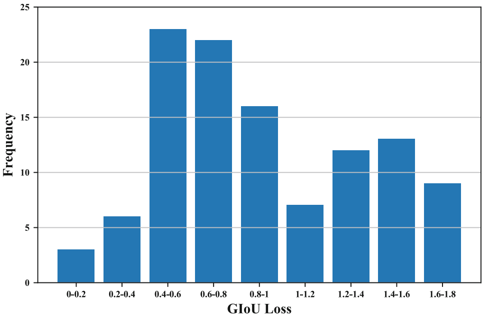
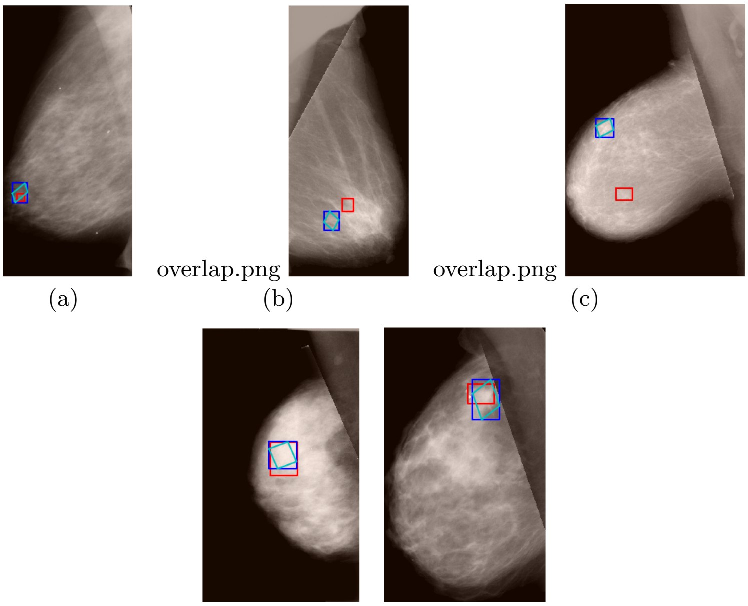


Fig. 4. Histogram of the GIoU loss for the test set

Examples of successful and unsuccessful registration results are shown in Fig. 5. In roughly 10% of the cases, the CC is still slightly overstretched to cover the pectoral muscle (Fig. 5a). It can be shown that in two cases, even if global alignment is successful, the bounding boxes do not overlap, sometimes by a large amount (Fig. 5c): this indicates that certain deformations cannot be recovered with the proposed affine transformation.



(a) (b) (c)

(d) (e)

Fig. 5. Registration examples: the MLO and registered CC views are shown overlapped. The MLO bounding box is shown in red, the CC in blue, before and after rectification. (Color figure online)

6 Conclusion and Future Works

The presented work tackles the challenge of registering CC and MLO views by designing a fully trainable registration network. Weakly supervision that exploits available lesion annotations achieves promising results both in terms of visual alignment and lesion registration. The proposed technique has been demonstrated using an affine transformation. As a consequence, the network cannot fully capture the complex deformations occurring due to breast compression. Further improvements can be expected by substituting the Spatial Transformer with a different module to estimate a pixel-wise deformation field. This work lays the basis for several future developments. We will investigate how to combine the proposed network with other architectures, e.g., for object detection, to achieve multi-view analysis of mammographic images. The proposed technique could also be adapted to related tasks, such as the temporal registration of images from subsequent screening rounds.

Acknowledgements. The authors wish to thank Antonio Gaetano Barletta for his contributions in the implementation of the registration module.

References

1. Alfano, F., et al.: Prone to supine surface based registration workflow for breast tumor localization in surgical planning. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1150–1153. IEEE (2019)
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
3. van Engeland, S., Snoeren, P., Hendriks, J., Karssemeijer, N.: A comparison of methods for mammogram registration. IEEE Trans. Med. Imaging **22**(11), 1436–1444 (2003)
4. Guo, Y., Sivaramakrishna, R., Lu, C.C., Suri, J.S., Laxminarayan, S.: Breast image registration techniques: a survey. Med. Biol. Eng. Comput. **44**(1–2), 15–26 (2006). https://doi.org/10.1007/s11517-005-0016-y
5. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Mach. Vis. Appl. **31**(1), 1–18 (2020). https://doi.org/10.1007/s00138-020-01060-x
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. 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, pp. 212–218. Medical Physics Publishing (2000)
8. Hu, Y., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. Med. Image Anal. **49**, 1–13 (2018)
9. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28. pp. 2017–2025. Curran Associates, Inc. (2015). http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf
10. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A Curated Mammography Data Set for Use in Computer-Aided Detection and Diagnosis Research, vol. 4, p. 170177. Nature Publishing Group, Berlin (2017)
11. Li, H., Fan, Y.: Non-rigid image registration using self-supervised fully convolutional networks without training data. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1075–1078. IEEE (2018)
12. Morra, L., Delsanto, S., Correale, L.: Artificial Intelligence in Medical Imaging: From Theory to Clinical Practice. CRC Press, Boca Raton (2019)
13. Morra, L., et al.: Breast cancer: computer-aided detection with digital breast tomosynthesis. Radiology **277**(1), 56–63 (2015)
14. Perek, S., Hazan, A., Barkan, E., Akselrod-Ballin, A.: Siamese network for dual-view mammography mass matching. In: Stoyanov, D., et al. (eds.) RAMBO/BIA/TIA -2018. LNCS, vol. 11040, pp. 55–63. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00946-5_6

15. Qin, C.: Joint learning of motion estimation and segmentation for cardiac MR image sequences. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 472–480. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_53
16. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 658–666 (2019)
17. Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with deep learning. Sci. Rep. **8**(1), 4165 (2018)
18. Sacchetto, D., et al.: Mammographic density: comparison of visual assessment with fully automatic calculation on a multivendor dataset. Eur. Radiol. **26**(1), 175–183 (2016). https://doi.org/10.1007/s00330-015-3784-2
19. Samulski, M., Karssemeijer, N.: Optimizing case-based detection performance in a multiview CAD system for mammography. IEEE Trans. Med. Imaging **30**(4), 1001–1009 (2011)
20. Sechopoulos, I.: A review of breast tomosynthesis. Part I. The image acquisition process. Med. Phys. **40**(1), 014301 (2013)
21. Van Schie, G.: Correlating locations in ipsilateral breast tomosynthesis views using an analytical hemispherical compression model. Phys. Med. Biol. **56**(15), 4715 (2011)
22. Viergever, M.A., Maintz, J.A., Klein, S., Murphy, K., Staring, M., Pluim, J.P.: A survey of medical image registration-under review. Med. Image Anal. **33**, 140–144 (2016)
23. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I.: End-to-end unsupervised deformable image registration with a convolutional neural network. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 204–212. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_24