Addressing Challenges in Image Translation for Contrast-Enhanced Mammography using Generative Advesarial Networks

Mohammad Hosseinipour[1,†], Luca Bergamin[1,†], Harel Kotler[2,†], Gisella Gennaro[2] and Fabio Aiolli[1,*]

[1]Department of Mathematics, University of Padua, Via Trieste, 63, Padua, 35122, Italy
[2]Istituto Oncologico Veneto IOV – IRCCS, Via Gattamelata 64, Padua, 35128, Italy

Abstract

Medical imaging is a cornerstone of modern healthcare, facilitating early diagnosis and the development of efficient treatment plans. Breast imaging includes different imaging modalities, including mammography and MRI, each encompassing unique information. Unfortunately, improving diagnostic performance can be accompanied by an increase in patient-related risks. Specifically, Contrast-enhanced mammography (CEM) offers better performance while exposing women to the risk of adverse reactions from the contrast agents used for it. To reduce these risks, deep learning solutions have become one of the promising research frontiers in recent years. In image-to-image translation, a mapping function is learned to transform a given image from a source domain to a target domain. In medical imaging, the most common solutions are based on GANs, such as pix2pix. When applied to CEM, we found that pix2pix encounters specific challenges due to low data quality, insufficient model capacity, and domain-derived requirements. Thus, these models have low performance out-of-the-box. In this paper, we highlight these specific challenges, propose tailored evaluation strategies, and present preliminary results on a novel dataset, showcasing the need for specialized approaches in medical imaging translation.

Keywords

medical imaging, generative adversarial networks, AI in healthcare, image translation, breast cancer detection

1. Introduction

Medical imaging is a key pillar of modern medicine. It provides the foundation for diagnostics, the process of identifying and characterizing a disease, monitoring, and treatment planning. Medical breast imaging includes different imaging methods, such as mammography, ultrasound, and MRI [1].

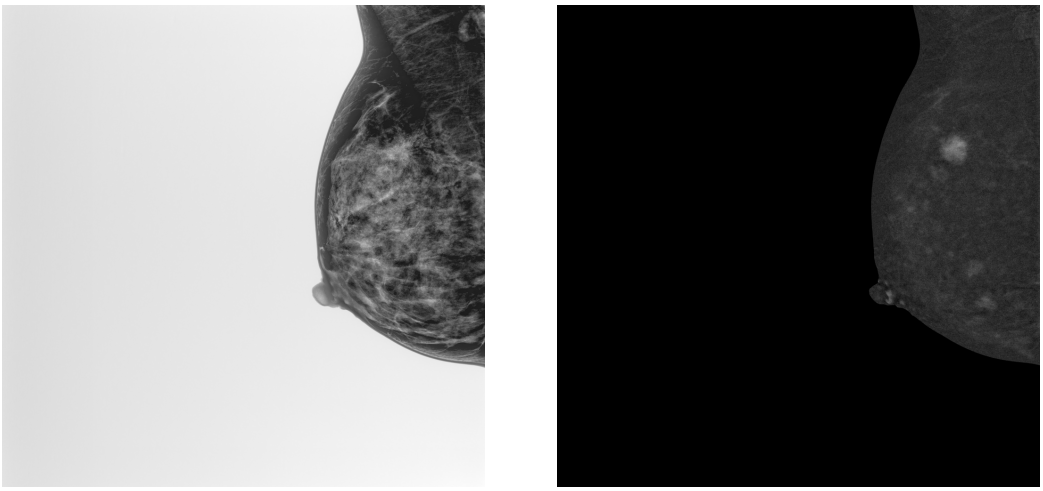
Mammography is an X-ray based method that projects the breast into a 2D image. Since the breast is a 3D object, overlapping tissues may mask underlying anomalies or generate false ones when it is projected to 2D. This is further emphasized in dense breasts, typical in younger women, where the elevated amount of breast tissue increases the overlap and reduces diagnostic performance [2].

Contrast-enhanced mammography (CEM) is a method developed to overcome this challenge. CEM uses intravenous iodinated contrast agents and energy subtraction to increase anatomical contrast and better represent potential malignancies, thus making it particularly strong in detecting masses in women who have dense breast tissue [3]. As demonstrated in figure 1, a CEM exam results in two main images used by radiologists to diagnose the breast [4]:

- 1. **Processed low-energy (pLE) image**: a mammography-equivalent image that does not show the enhancement of the contrast media [5].
- 2. **Dual-energy subtraction (DES)** image: the enhancement showing image.

3rd AIxIA Workshop on Artificial Intelligence For Healthcare (HC@AIxIA 2024) + 5th Data4SmartHealth (D4SH 2024), 25-28 November 2024, Bolzano, Italy

*Corresponding author.
†These authors contributed equally.
mohammad.hosseinipour@studenti.unipd.it (M. Hosseinipour); bergamin@math.unipd.it (L. Bergamin); harel.kotler@ioveneto.it (H. Kotler); gisella.gennaro@ioveneto.it (G. Gennaro); aiolli@math.unipd.it (F. Aiolli)
0009-0003-6573-8081 (M. Hosseinipour); 0000-0002-0662-7862 (L. Bergamin); 0000-0001-9299-082X (H. Kotler); 0000-0003-2444-1778 (G. Gennaro); 0000-0002-5823-7540 (F. Aiolli)



(a) pLE Image (Mammography Equivalent) (b) DES Image (Contrast Enhanced)

Figure 1: The two main CEM images: (a) processed low-energy (pLE) and dual-energy subtracted (DES).

Despite its advantages, CEM still requires an increased radiation dose compared to standard mammography and exposes the patient to the risk of adverse reactions related to the contrast agents [6, 7]. Additionally, some women who could benefit from CEM cannot undergo the procedure due to limitations for contrast agent use, such as renal disease [8]. We argue that reducing or eliminating these risks without compromising diagnostic performance could lead to broader adoption. Furthermore, this could allow women who currently cannot benefit from CEM to access its diagnostic capabilities, and potentially enable its use in general breast cancer screening programs in the future.

An important technique that holds promise in reducing these risks is image-to-image translation. In this technique, models learn the relationships between features in an image from the source domain (e.g., style, structure, or content) and how to associate them with corresponding features in the target domain [9]. Applying it to CEM holds the potential to create virtual contrast-enhanced images without the use of iodinated contrast agents.

Significant work has been done in the field of medical image translation using Generative Adversarial Networks (GANs). GANs have been applied in recent years in image-to-image translation in medical imaging, particularly in breast imaging. An example of that ability is MammoGANesis, a GAN-based framework that can synthesize mammograms [10]. Thanks to its demonstrated strong performance across similar tasks across the various imaging techniques, we believe GANs could be used to translate pLE images to DES images with contrast-enhancement [11].

While GANs excel at generating general images, we observed they can fail to reproduce fine-grained, location-specific details. Given the importance of these features in medical settings, we incorporated attention modules to enhance the representation of local features [12]. This approach could offer a better chance at generating contrast where it is needed and suppressing it where it is not required, which is critical for clinical use.

In this paper, we experimentally demonstrate that state-of-the-art attention modules are outperformed by the older U-Net based GAN architecture when performing CEM image-to-image translation. Therefore, we propose two novel solutions to overcome this issue, namely an attention-based improvement on the generator architecture and a tailor-made loss function for CEM images to promote the reproduction of bright details in the image.

2. Related Work

GANs have been widely used in various medical image translation tasks. For instance, the MedGAN framework has demonstrated the capability of GANs to generate realistic medical images across multiple

modalities [9]. In virtual contrast generation for breast imaging, Müller-Franzes et al. used GANs to enhance the effect of contrast media in contrast-enhanced MRI images [13]. This was done in the hope of reducing the dose of contrast agent used in this imaging method. Since mammograms and CEMs possess higher resolutions compared to MRIs, we aim to assess the feasibility of applying similar GAN-based techniques to these higher-resolution imaging modalities. Other works in the literature explicitly consider the creation of high-resolution images, and they can be considered in future extensions of this work [14].

While other approaches, such as CycleGAN and diffusion models, have emerged in the field, they are not well-suited to our application. CycleGAN [15] is designed for scenarios with unpaired datasets, i.e., where there is no clear match between input and output, whereas our study utilizes paired images, making traditional GANs such as pix2pix a more appropriate choice. Furthermore, although denoising diffusion probabilistic models (e.g., DDPMs) show promise [16], GANs are currently more mature and provide a more trustworthy technology for our purposes. In particular, an existing work considered a low-dose setting for breast MRI gathered some evidence that GAN-generated images are preferred over DDPM-generated images by radiologists for the lowest levels of contrast agent [13]. The study indicates that both models are promising, and they conclude that further development is needed. We argue that lower performance for DDPMs can be due to a small training set size and higher computing requirements. In fact, it was observed that GAN-based architectures still work well [17] even with hundreds of samples. Another important observation concerns the higher inference time required for diffusion models compared to GANs [18]; this issue can hinder their usage in real-time applications.

The Attention U-Net model, which introduces attention gates to focus on relevant regions, has already demonstrated improved performance in medical image segmentation tasks [12]. Our approach builds on these advancements by integrating channel, multi-scale channel, and spatial attention mechanisms. This enhancement enables our model to better capture the complex structures inherent in medical images, resulting in a more accurate generation of DES images.

Finally, the idea of reweighting the loss function has already been explored in other contexts, such as in object detection [19]. Our proposal specifically applies to CEM medical images, and is studied for reproducing bright details. It can be argued that this technique could also mitigate mode collapse issues, since healthy tissue is predominant in the pixels of the images, and the discriminator of the GAN could be fooled most of the times by reconstructing the mode of the data. Many works in the literature, such as BicycleGAN [20] specifically address this issue, while the present work does not investigate the mode collapse issue explicitly.

3. Background

A general overview of the techniques we considered follows.

3.1. Overview of GANs

For our model, we considered Generative Adversarial Networks (GANs). GANs are a class of machine learning frameworks designed by Goodfellow et al. in 2014 [21]. GANs consist of two neural networks, a generator and a discriminator, which are trained alternately through adversarial processes. The generator's goal is to create data that is indistinguishable from real data, while the discriminator's goal is to correctly identify whether the data is real or generated. The interplay between these two networks allows GANs to generate high-quality synthetic data.

3.2. Image-to-image translation using pix2pix

The pix2pix model was first introduced by Isola et al. in 2017 [17]. Their work extends the concept of GANs to the task of paired image-to-image translation. In this task, the objective is to learn a mapping function $f : S \to T$, where S is a source domain and T is a target domain. This model has been

successfully applied to a wide range of domains and used in image colorization, background removal, and semantic segmentation [17].

The pix2pix framework employs a conditional GAN [21, 22, 23], where the generator learns to map an observed image $x \in \mathbb{R}^{C_S \times H \times W}$ and a random noise vector $z \in \mathbb{R}^d$, to $y \in \mathbb{R}^{C_T \times H \times W}$, $G: \{x, z\} \rightarrow y$ to fool the discriminator; while the discriminator evaluates whether the imageis a real or fake image.

3.2.1. PatchGAN

In pix2pix, the discriminator is a PatchGAN [21], which classifies whether each N×N patch in an image is real or fake. This approach ensures that high-frequency structures are captured in the output images.

3.2.2. U-Net

The U-Net architecture, proposed by Ronneberger et al. [24], is a popular choice for image segmentation tasks, especially in the healthcare domain. It consists of an encoder-decoder structure with skip connections between corresponding layers in the encoder and decoder. These skip connections help in retaining spatial information that is often lost during downsampling in the encoder [24].

In pix2pix, the most effective generator proposed is the U-Net generator, with the encoder contracting the input image to a bottleneck layer and the decoder expanding it back to the original size while merging features from the encoder layers through skip connections. We show the architecture of U-Net in Fig. 2(a).

3.2.3. Loss Function

The pix2pix architecture employs a composite loss function that consists of two main components: an adversarial loss and a reconstruction loss.

- **Adversarial Loss (L_{cGAN})**: The adversarial loss is the core component of GANs, where the generator tries to fool the discriminator, and the discriminator tries to distinguish between real and fake images [17]. In the context of pix2pix, the conditional adversarial loss (L_{cGAN}) is used to condition the generation process on the input image, ensuring that the generated output is a plausible transformation of the input. The adversarial loss is defined as:

$$L_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))] \tag{1}$$

  where G is the generator, D is the discriminator, x is the input image, y is the real output image, and z is the noise vector.

- **Reconstruction Loss (L_{L1})**: To reduce the distance between two images, the L1 and L2 distances were investigated. The L2 distance makes the generator create results that are more blurry compared to the L1 distance, producing sharper images [17]. This loss encourages the generator to produce images close to the real images in pixel space, promoting accurate reconstruction of the target output. The L1 loss is defined as:

$$L_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1] \tag{2}$$

The final loss function for the generator is a combination of the adversarial loss (L_{cGAN}) and the L1 distance loss (L_{L1}). The total loss is given by:

$$L_G = L_{cGAN}(G, D) + \lambda L_{L1}(G) \tag{3}$$

where λ is the weighting factor that balances the contribution of the L1.

3.3. Attention U-Net

Attention U-Net, proposed by Oktay et al. [12], introduces attention gates to the standard U-Net architecture. These attention gates allow the model to focus on relevant regions of the image, thereby improving quality performance. We show the architecture of Attention U-Net in Fig. 2(b).

Attention gates are inserted into the skip connections of the U-Net, enabling the network to suppress irrelevant regions and highlight salient features useful for a specific task. This should result in enhanced model's sensitivity and prediction accuracy without significantly increasing computational overhead.

4. Problem definition

We present the main challenges we found while attempting to solve this problem, give appropriate context, and discuss a number of mitigation strategies.

1. **Data quality**
    a) **Noise**: CEM images, and X-ray based medical images in general, are usually affected by noise [25]. This is because ensuring safety requires limiting the ionizing radiation exposure. Denoising techniques, such as bilateral filtering [26], offer a trade-off between sharpness and noise removal. While the introduction of noise in input data has been argued to not affect much the generalization capabilities of deep neural networks due to emergent self-denoising capabilities [27], it is yet to be understood whether applying denoising to medical images improves the capabilities of generative models.
    b) **Scarsity**: as outlined by [17], it is argued that as little as 400 images are needed to train a good model. This can be crucial for medical applications since the cost of acquisition and legal requirements can severely affect the size of the dataset. It is still unclear if this can apply to high-resolution images.
    c) **Imbalance**: due to the different settings in which medical data is acquired (e.g., screening data vs. at-risk patient monitoring), data can be affected by different selection biases. The most important factor is the ratio between negative and positive cases, which can affect the sensibility and sensitivity of the application.
    d) **Diversity**: different conditions can affect each image, ranging from different breast densities to the presence of foreign objects such as breast implants or surgical clips.
2. **Model capacity**
    a) **Architecture**: Different deep learning architectures have been proposed to process medical data effectively, with the U-Net being the most popular one [24]. For generative applications, it is yet to be understood which kind of deep architecture works best.
    b) **Objective function**: the identification of a differentiable objective that aligns with the requirements of radiologists is not defined unanimously. Literature shows that some objectives, different from classic L1/L2 distances, align more with human perception of quality [28].
3. **Domain-related challenges**
    a) **Reproducing bright areas from low-energy images**: bright spots in the DES image often highlight the presence of lesions. Making sure those areas are preserved increases the sensitivity of the instrument.
    b) **Suppressing dark areas from low-energy images**: dark areas in the DES image help the reader not get confused by irrelevant information. Making sure those areas are suppressed increases the specificity of the instrument.
    c) **Reproduction of small bright details**: specifically in CEM applications, bright spots in the subtracted image are associated with the presence of lesions. Their size can range from many centimeters down to a few millimeters in the case of micro-calcifications [29]. It has been shown that small details in images can be hard to reproduce in GANs using a standard MSE loss [30]. Thus, these small details are likely to be ignored by generative models if not taken into account.

d) **High resolution**: CEM images have high resolution (>2048x2048) and high bit depth (12-13 bits). This can severely impact training times. Current works in medical image-to-image translation do not address these issues [9].
e) **Evaluation**: the evaluation of medical images can vary differently from reader to reader. Thus, finding a proper quality metric for generated images is challenging.

In summary, our present work investigates the following aspects:

1. we show how a state-of-the-art architecture is not effective in this specific task, being outperformed by its U-Net baseline;
2. we provide a new proposal that mitigates this issue using an attention module, and we explore different U-Net model capacities;
3. we propose a novel loss function that focuses more on reproducing small, bright details in the image.

Note that many other solutions could be investigated, such as resampling minority classes, data augmentation, and data denoising. In this paper, we chose to address the challenges related to the model architecture and training, and we left other related problems to future works.

5. Method

In this section, we cover the main workings of the proposed techniques.

5.1. Inner Attention Module Network (IAMNet)

We aim to enhance the capability of a model to focus on relevant features by incorporating attention mechanisms. Our work starts with the U-Net architecture and uses some readapted ideas from the Attention U-Net. To this end, we combine the Convolutional Block Attention Module (CBAM) [31], which contains Channel Attention and Spatial Attention, with a third mechanism named Multi-Scale Channel Attention.

Channel Attention Block The Channel Attention Block is designed to highlight important feature channels, which correspond to particular types of information within an image such as edges, textures, or colors [31]. It works by applying two pooling operations, global average pooling and max pooling, across the entire spatial domain of each channel. These pooling operations produce a channel-wise descriptor that summarizes the significance of each channel. These descriptors are then passed through a small neural network, which outputs a set of weights that are applied to the channels via a sigmoid activation function. The resulting attention map selectively emphasizes the most relevant channels, allowing the model to focus on key details. The Channel Attention Block is defined as:

$$\text{ChannelAttention}(x) = x \times \sigma(\text{MLP}\,(\text{AvgPool}(x)) + \text{MLP}\,(\text{MaxPool}(x))) \tag{4}$$

Where $x \in \mathbb{R}^{B\times C\times H\times W}$ is the input feature map, B is the batch size, C is the number of channels, and H and W are the spatial dimensions of the input.

Spatial Attention Block While the Channel Attention Block prioritizes feature channels, the Spatial Attention Block focuses on identifying critical spatial locations within the feature map [31]. This block operates by compressing the channel information into a single map, which highlights regions that contain significant information. By combining the maximum and average values across the channels, the module generates a spatial attention map. A convolutional layer followed by a sigmoid activation function is applied to this map, which is then multiplied element-wise with the original feature map. The result is an enhanced representation that emphasizes the most important spatial regions, enabling

the network to concentrate on key areas in the image, such as regions of interest. The Spatial Attention Block is defined as:

$$\text{SpatialAttention}(x) = x \times \sigma(\text{Conv}([\text{AvgPool}(x), \text{MaxPool}(x)])) \tag{5}$$

Again, where $x \in \mathbb{R}^{B\times C\times H\times W}$ is the input feature map, B is the batch size, C is the number of channels, and H and W are the spatial dimensions of the input.

Multi-Scale Channel Attention Block In addition to the Channel and Spatial Attention blocks, we incorporated the Multi-Scale Channel Attention Block to enhance feature representation by processing the input feature map at multiple scales. This approach is akin to viewing an object through magnifying glasses of different strengths, allowing the network to capture both coarse and fine details simultaneously. In this block, average pooling is applied to the feature maps, reducing spatial dimensions and enabling the model to focus on important features at each scale. Subsequently, Channel Attention is employed to highlight significant feature channels, ensuring that the most relevant information is prioritized. After applying attention mechanisms, the feature maps are upsampled back to the original dimensions. This process generates multiple attention maps, which are then averaged to create a comprehensive attention map that enhances relevant features across the entire image [32, 33, 34].

$$\text{MultiScaleChannelAttention}(x) = x \times \frac{1}{N}\sum_{i=1}^{N} \sigma\big(\text{ChannelAttention}\left(\text{AvgPool}_i(x)\right)\big)_{\uparrow} \tag{6}$$

Where $x \in \mathbb{R}^{B\times C\times H\times W}$ is the input feature map, with B as the batch size, C the number of channels, and H and W representing height and width, respectively. The term N denotes the number of different scales applied. $\text{AvgPool}_i(x)$ represents the average pooling operation at the i-th scale, which reduces the spatial dimensions of the feature map. The output is then passed through the Channel Attention mechanism, and σ is the sigmoid activation function that produces the attention map. The symbol $(\cdot)_{\uparrow}$ denotes upsampling the attention-modulated feature map back to the original spatial dimensions using bilinear interpolation. Finally, the attention maps from all scales are averaged and applied element-wise to the original input x, allowing the model to emphasize relevant features across multiple scales. By integrating different combinations of attention blocks, we explored a higher-performance generator that employs an encoder-decoder structure with a focus on integrating attention mechanisms at its bottleneck. We represent its architecture in Fig. 2(c). This design allows the model to effectively highlight key feature channels and/or important spatial regions in the feature maps via different combinations of attention blocks. The full inner Attention Module is comprised of three components, as shown in Fig. 3(b). With IAMNet enhances the model's ability to pick out relevant information, ultimately boosting its performance in tasks like segmentation and detection.

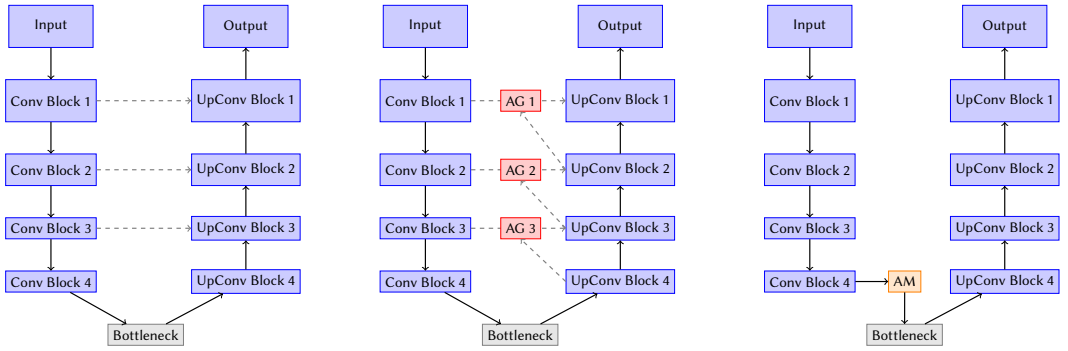
5.2. Exponential dampening loss function

As noted in Section 4, the quality of generated images is dependent on the presence of small, bright details. Thus, giving equal weight to the fidelity of dark and bright areas is undesirable. In particular, using the L1 distance between the source and the target image is not a sensible choice due to its symmetry represented in Fig. 4 (left). Therefore, we provide a novel loss function, defined in Eq. 7.

$$L_{L1}(G;\tau) = \mathbb{E}_{x,y,z}[w \cdot \|y - G(x,z)\|_1], \quad w = e^{-\tau(1-y)} \tag{7}$$

The loss considers the normalized intensity of the target pixel (i.e., between 0 and 1). If the target pixel is not bright (closer to 0), then the value of the L1 loss term is rescaled using an exponential function. We show in Fig. 4 (middle) a simplifying example, which shows that dark target pixels predicted as high are penalized less, while dark pixels are not penalized as much.

The τ hyperparameter can be fixed, selected through tuning, and can be annealed during training. Note that, when $\tau = 0$, the function matches the original L_{L1} definition. A higher τ value makes the objective further to the original L1 loss (Fig. 4, right).



(a) U-Net (b) Attention U-Net (c) Inner Attention Module Network (our proposal)

Figure 2: (a) A simplified U-Net with 4 convolutional blocks. The U-Net encodes data until it reaches a bottleneck. Then, data is decoded with the help of skip-connections. (b) A simplified Attention U-Net with 3 attention gates (AGs). The gating mechanism receives queries from the decoding blocks on the right and selects data from the encoding path on the left. (c) A simplified Inner Attention Module Network. The overall architecture has fewer connections and contains a single Attention Module at the bottleneck.

(a) Attention Gate (AG) (b) Attention Module (AM, our proposal)

Figure 3: Schematic explanation of the (a) Attention Gate (AG) and (b) Attention Module (AM).

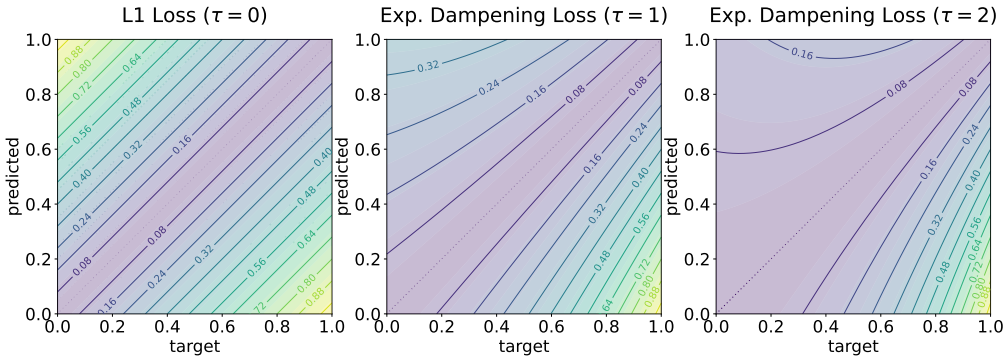


Figure 4: Simplifying example of the proposed loss. The bottom right area shows a high penalty for bright pixels wrongly predicted as dark.

6. Evaluation

In this section, we evaluate our proposal on a benchmark dataset. First, we describe the characteristics of the dataset. Then, we review the chosen quality metrics. Finally, we report our experimental setting and our results.

6.1. Dataset

We obtained the dataset from the Istituto Oncologico Veneto (IOV – IRCCS), which includes images from 550 patients, resulting in approximately 2000 image pairs of low-energy images and DES. The low-energy images were initially acquired as raw images, which underwent processing step for standard contrast adjustment and noise reduction for better object visibility [35]. This results in a processed low-energy image (pLE), that is equivalent to standard mammography [5]. The dataset, provided in DICOM file format, cannot be made publicly available due to legal requirements. The resolution of the images is 2850 × 2394 pixels. The dataset went through comprehensive reviews to remove outliers, artifacts, and abnormalities. Subsequently, the dataset was divided into training and test subgroups in a 95:5 ratio.

Benchmark: We created a benchmark of 11 patients with a mass in at least one of their DES images. To better understand the performance, we compared them based on L1 distance, L2 distance, ΔCNR, and Peak Signal-to-Noise Ratio (PSNR) metrics in three different settings. The first setting is the segmented breasts without visible masses, the second setting is the segmented breasts with visible masses, and the third setting is the segmented masses only.

6.2. Contrast vs. Pixel Value Distance

We argue that a good contrast-enhanced image is one that effectively shows the contrast between the mass and the surrounding tissues rather than just having a lower L1 or L2 distance. Although a generated image may have an average pixel value closer to that of the target image, it may fail to capture the sensible contrast between the mass and the surrounding tissue, reducing its clinical relevance. Therefore, we used the ΔCNR metric, which can help us quantitatively compare the models based on their ability to provide clear contrast for mass visualization.

Contrast-to-Noise Ratio (CNR): CNR is a quantitative measure commonly used in medical imaging to evaluate the contrast of a region of interest (ROI), such as a mass, against its surrounding background. It is defined as the difference in the average pixel intensity between the ROI and its surrounding background, normalized by the standard deviation of the background. The general equation for CNR can be expressed as:

$$\mathrm{CNR} = \frac{\mu_{\mathrm{ROI}} - \mu_{\mathrm{background}}}{\sigma_{\mathrm{background}}} \tag{8}$$

where:

- μ_{ROI} is the mean pixel intensity of the region of interest,
- $\mu_{\mathrm{background}}$ is the mean pixel intensity of the surrounding background,
- $\sigma_{\mathrm{background}}$ is the standard deviation of the pixel intensities in the background.

In our study, we adapt this general definition of CNR to assess the contrast specifically in our three different benchmark settings. To do this, we employ two different approaches to compute CNR:

- **Mass CNR:** For the ”mass only” setting, two readers have segmented all the masses and their surroundings in the real DES images separately, and applied the segmentations to both real and generated DES images. For each DES image, we compute the CNR as the difference between the mean pixel intensity of all segmented mass regions and the mean pixel intensity of all the surrounding regions of the masses, normalized by the standard deviation of the surroundings. This can be expressed mathematically as:

$$\mathrm{CNR} = \frac{\mu_{\mathrm{mass\_regions}} - \mu_{\mathrm{surrounding\_regions}}}{\sigma_{\mathrm{surrounding\_regions}}} \tag{9}$$

  where:

- $\mu_{\text{mass\_regions}}$ is the mean pixel intensity of all segmented mass regions,
- $\mu_{\text{surrounding\_regions}}$ is the mean pixel intensity of the surrounding regions of all the segmented masses,
- $\sigma_{\text{surrounding\_regions}}$ is the standard deviation of the pixel intensities in the surrounding regions of all the segmented masses.

- **Breast CNR:** For both real and generated DES images, we segment the breast and apply square patches with patch size $= 64 \times 64$ and stride $= \frac{\text{patch size}}{4}$. The CNR is computed for each patch and then averaged across all patches. This method is used for the "segmented breast with mass" and "segmented breast without mass" settings. Mathematically, for each patch on the segmented breast, the CNR is computed as follows:

$$\text{CNR}_i = \frac{\mu_{a,i} - \mu_{b,i}}{\sigma_{b,i}} \tag{10}$$

where:

- $\mu_{a,i}$ is the mean pixel intensity of the i-th target patch (the center patch),
- $\mu_{b,i}$ is the mean pixel intensity of the surrounding 8 patches for the i-th target patch,
- $\sigma_{b,i}$ is the standard deviation of the pixel intensities of the surrounding 8 patches for the i-th target patch.

The overall CNR for the entire image is then averaged across all N patches in the segmented breast region, which can be written as:

$$\text{CNR} = \frac{1}{N} \sum_{i=1}^{N} \text{CNR}_i \tag{11}$$

where N is the total number of patches in the segmented breast region.

ΔCNR Metric: The ΔCNR metric is defined as the difference between the CNR values of the generated (fake) DES images and the ground truth (real) DES images in each of our three analytical benchmarking settings. This can be expressed mathematically as:

$$\Delta\text{CNR} = \text{CNR}_{\text{fake}} - \text{CNR}_{\text{real}} \tag{12}$$

where:

- CNR_{real} = Contrast-to-Noise Ratio of the real DES images,
- CNR_{fake} = Contrast-to-Noise Ratio of the generated DES images.

This metric allows for a quantitative comparison of the mass visibility between real and generated images at the whole breast and mass only levels.

6.3. Experimental setting

We trained our models on a single NVIDIA V100 GPU, with every model taking 12-24 hours to complete its training, depending on the architecture size. We report in Table 1 the hyperparameters used for the trained models. To train the models, we used a fixed learning rate for $E_f = 100$ epochs, then we linearly decayed it for $E_d = 200$ additional epochs. We employed Adam as optimizer. For models trained using the exponential dampening function, we initialized the value to $\tau(t = 0) = \tau$, then we linearly decayed it until $\tau(t = N) = 0$, using $\tau(t) = (1 - \frac{t}{N})\tau$, where t is the current epoch number and N is the total number of epochs.

Name	Value	Name	Value	Name	Value
GAN objective	Least-square	Resolution (train)	512 × 512	Resolution (test)	1024 × 1024
Batch size	10	Learning rate	0.0002	Dropout rate	0.5
λ	100	Generator	U-Net 512	Discriminator	PatchGAN (70×70)
Epochs (fixed LR)	100	Epochs (LR decay)	200	LR annealing type	linear
#Conv. (Attn U-Net)	9	#Conv. (IAMNet)	5	ngf/ndf	64/64

Table 1
Hyperparameters considered for training our models.

	With mass				Without mass				Mass only			
Type of attention in IAMNet	L1	L2	PSNR	ΔCNR	L1	L2	PSNR	ΔCNR	L1	L2	PSNR	ΔCNR
Channel	**9.34**	**303.97**	**26.07**	**0.274**	**8.28**	**214.97**	**26.49**	**0.248**	33.79	2712.04	36.10	-2.782
Spatial	9.89	370.33	25.44	0.310	8.95	278.42	25.38	0.282	32.85	2687.00	36.34	-2.839
Multi-Scale Channel	9.25	335.12	25.86	0.285	8.34	248.56	25.89	0.253	**32.68**	**2620.17**	**36.53**	**-2.643**

Table 2
Performance comparison of different attention blocks in the IAMNet model using L1, L2, PSNR, and ΔCNR to see the potential of each attention block alone.

6.4. Attention Module analysis

The attention blocks work as coefficients on the feature maps in the bottleneck of IAMNet, applied sequentially, without changing the data dimensionality. Therefore, we can test each of them individually to see how they perform and which one is the most effective one. As shown in Table 2, the Channel Attention Block performs best in the segmented breast; however, when it comes to the contrast of the mass, the Multi-Scale Channel Attention outperforms the Channel Attention. Moreover, Spatial Attention showed worse results in both mass and breast-segmented settings for all metrics. This is potentially due to the fact that at the bottleneck, hence the spatial dimension of feature maps is already shrunk and can not provide valuable information.

6.5. Attention Module comparison

We investigated the use of attention mechanisms to guide the model in focusing on more relevant features in Table 4. In contradiction to the potential performance improvement expected from the Attention U-Net, in this specific study of medical imaging, we see even lower performance results compared to the U-Net baseline. By comparing IAMNet with U-Net and U-Net with exponential dampening, we observed that IAMNet performed worse in both "With mass" and "Without mass" breast segmented scenarios. Our initial hypothesis was that IAMNet might perform better in reproducing the specific regions containing the masses. To test this, we segmented the images and computed the performance metrics exclusively for the mass regions. In this targeted analysis, IAMNet showed improved performance, suggesting that it is better suited for focusing on specific lesions. This result indicates a trade-off: while IAMNet's overall reconstruction performance across the entire image was inferior, it demonstrated enhanced performance in identifying and segmenting the lesion areas.

6.6. Exponential dampening loss

We report in Table 3 our results, computed using a U-Net baseline architecture as the generator. We mainly observe similar and consistent results across different selected τ values. We found a small but consistent improvement, with higher τ values, with best results obtained with $\tau = 2$ and $\tau = 3$. If we take into account only the segmented mass area, we find higher errors, meaning that it is generally harder for the model to reconstruct the correct intensity. Nonetheless, we observe the highest improvement, which shows that higher τ values improve the reconstruction of bright spots in the image. Interestingly, we found that annealing the τ value from a high to a low value had the best performance. This could be intended as a form of curriculum learning, where the model is first trained to solve a specific task and then moves to other tasks to improve its overall performance.

Model	With mass				Without mass				Mass only			
	L1	L2	PSNR	ΔCNR	L1	L2	PSNR	ΔCNR	L1	L2	PSNR	ΔCNR
$\tau = 0.0$	8.65	192.19	28.26	0.177	7.87	129.97	28.60	0.138	35.42	2955.36	35.74	-2.87
$\tau = 0.5$	8.76	190.78	28.26	0.179	7.87	129.36	28.62	0.139	34.35	2785.45	**36.06**	-2.85
$\tau = 1.0$	8.71	192.23	28.25	0.183	7.82	127.46	28.67	0.142	34.74	2849.54	35.88	-2.88
$\tau = 2.0$	**8.63**	189.59	**28.32**	0.173	**7.72**	**124.46**	28.79	**0.129**	35.16	2884.89	35.80	-2.85
$\tau = 3.0$	8.68	**188.74**	28.26	**0.170**	7.86	131.11	**28.57**	0.134	**33.81**	**2667.09**	**36.06**	**-2.73**

Table 3
Performance comparison for different U-Net-512 models using L1, L2, PSNR, and ΔCNR. We apply the exponential dampening loss function with different starting τ values, and linearly anneal τ to 0.

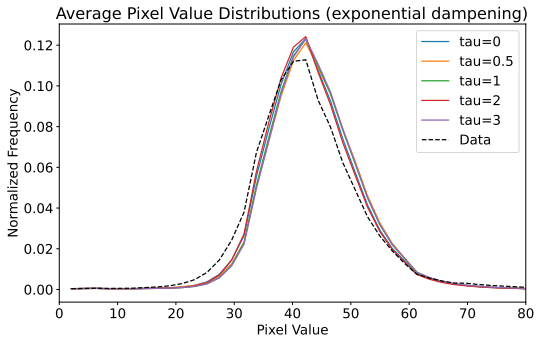


Figure 5: Color distribution matching property of the U-Net model, for models with different τ values.

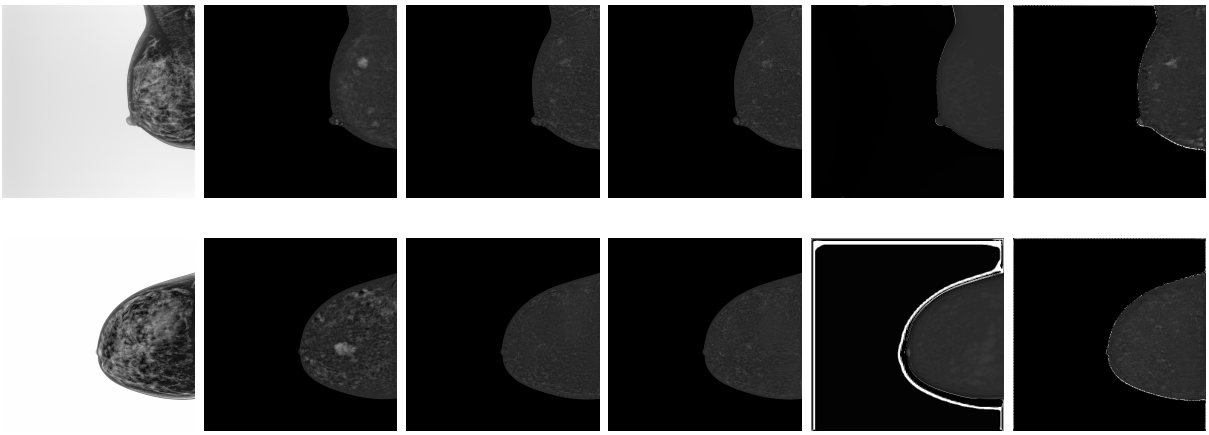
Model	With mass				Without mass				Mass only			
	L1	L2	PSNR	ΔCNR	L1	L2	PSNR	ΔCNR	L1	L2	PSNR	ΔCNR
U-Net	8.65	192.19	28.26	0.177	7.87	129.97	28.60	0.138	35.42	2955.36	35.74	-2.867
Attn. U-Net	9.55	432.77	25.10	0.365	8.13	357.98	25.16	0.339	35.85	2951.78	35.81	**-2.541**
U-Net ($\tau = 2.0$)	**8.63**	**89.59**	**28.32**	**0.173**	**7.72**	**124.46**	**28.79**	**0.129**	35.16	2884.89	35.80	-2.845
IAMNet	9.57	359.06	25.51	0.287	8.69	289.33	25.34	0.255	**33.71**	**2387.87**	**35.83**	-2.725

Table 4
Final performance comparison between main 4 models, U-Net, Attn. U-Net, U-Net ($\tau = 2.0$), and IAMNet.

Finally, in Figure 5, we analyze the frequency of the pixel intensities of the generated images and compare them to the original DES images. We find results consistent with the existing literature [17], as our models are consistently able to reproduce most of the target data distribution.

6.7. Qualitative comparison

In Fig. 6, we show a comparison of input images (pLE, mammogram-equivalent images without contrast enhancement), ground-truth images (DES, contrast-enhanced images), and generated images. The ground-truth DES images (b) demonstrate multiple bright spots, which in CEM images can represent the breast border, normal tissue, or masses. The presence, position, and intensity of these bright areas are crucial for radiologists, as they guide the interpretation of the images and aid in the detection of potential abnormalities and the identification of potentially cancerous masses. We can observe that the U-Net usually reduces many bright details. Using $\tau = 2$, bright spots are more preserved, but still not closely resembling when compared to the ground truth. Attention U-Net presents some white artifacts and lacks fine detail definition. Finally, our IAMNet shows better performance by effectively highlighting the bright details in the DES image, closely resembling both brightness and location of the bright area found in the ground truth, potentially enough to raise suspicion of potential abnormalities.



(a) pLE (b) DES (c) U-Net (d) U-Net ($\tau = 2$) (e) Attn. U-Net (f) IAMNet

Figure 6: Comparison of low energy (pLE), dual-energy subtracted (DES), U-Net generated, U-Net ($\tau = 2$) generated, Attention U-Net generated, IAMNet generated.

7. Conclusions

The results in this paper suggest that investigating novel architecture and losses is an effective way to address many of the challenges in image-to-image translation for CEM medical images. We provide some preliminary experimental data that shows our approach has potential, but requires further improvement to reach clinical utility. Despite achieving a high quantitative metrics, the generated images miss or inaccurately represent the anatomical details and occasionally fail to clearly delineate critical features such as tumor boundaries or vascular structures, limiting their current diagnostic value.

The first step to address these shortcomings is to explore whether incorporating skip connections into the IAMNet architecture could enhance the preservation of clinically relevant details. Interestingly, despite lacking skip connections, IAMNet has shown the ability to outperform other models that rely on them. Furthermore, integrating the proposed loss function into new architectures must be explored further to refine image quality and maintain diagnostic integrity.

In future work, we want to address more of the challenges we stated, considering better ways to evaluate our results, both in a qualitative and quantitative way. Moreover, we want to consider in a systematic way whether mode collapse is a measurable issue for generative models applied to medical images.

References

[1] T. Shah, S. Guraya, Breast cancer screening programs: Review of merits, demerits, and recent recommendations practiced across the world, Journal of Microscopy and Ultrastructure 5 (2017) 59. doi:10.1016/j.jmau.2016.10.002.

[2] D. S. Al Mousa, E. A. Ryan, C. Mello-Thoms, P. C. Brennan, What effect does mammographic breast density have on lesion detection in digital mammography?, Clinical Radiology 69 (2014) 333–341. doi:10.1016/j.crad.2013.11.014.

[3] N. Fico, G. D. Grezia, V. Cuccurullo, et al., Breast imaging physics in mammography (part ii), Diagnostics 13 (2023) 3582.

[4] I. P. L. Houben, P. Van De Voorde, C. R. L. P. N. Jeukens, et al., Contrast-enhanced spectral mammography as work-up tool in patients recalled from breast cancer screening has low risks and might hold clinical benefits, European Journal of Radiology 94 (2017) 31–37. doi:10.1016/j.ejrad.2017.07.00.

[5] M. B. I. Lobbes, M. L. Smidt, J. Houwers, V. C. Tjan-Heijnen, J. E. Wildberger, Contrast enhanced

mammography: techniques, current results, and potential indications, Clinical Radiology 68 (2013) 935–944. doi:10.1016/j.crad.2013.04.009.

[6] G. Gennaro, A. Cozzi, S. Schiaffino, F. Sardanelli, F. Caumo, Radiation dose of contrast-enhanced mammography: A two-center prospective comparison, Cancers 14 (2022) 1774. doi:10.3390/cancers14071774.

[7] W. Bottinor, P. Polkampally, I. Jovin, Adverse reactions to iodinated contrast media, International Journal of Angiology 22 (2013) 149–154. doi:10.1055/s-0033-1348885.

[8] American College of Radiology, Acr manual on contrast media, version 10.3, 2017. URL: https://www.acr.org/-/media/ACR/Files/Clinical-Resources/Contrast_Media.pdf.

[9] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, B. Yang, Medgan: Medical image translation using gans, Computerized Medical Imaging and Graphics 79 (2020) 101684. URL: http://dx.doi.org/10.1016/j.compmedimag.2019.101684. doi:10.1016/j.compmedimag.2019.101684.

[10] C. Zakka, G. Saheb, E. Najem, G. Berjawi, Mammoganesis: Controlled generation of high-resolution mammograms for radiology education, arXiv (2020). doi:10.48550/ARXIV.2010.05177, published online.

[11] M. Gong, S. Chen, Q. Chen, Y. Zeng, Y. Zhang, Generative adversarial networks in medical image processing, Current Pharmaceutical Design 27 (2021) 1856–1868. doi:10.2174/1381612826666201125110710.

[12] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention u-net: Learning where to look for the pancreas, 2018. URL: https://arxiv.org/abs/1804.03999. arXiv:1804.03999.

[13] G. Müller-Franzes, L. Huck, M. Bode, et al., Diffusion probabilistic versus generative adversarial models to reduce contrast agent dose in breast mri, European Radiology Experimental 8 (2024). doi:10.1186/s41747-024-00451-3.

[14] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017) 8798–8807. URL: https://api.semanticscholar.org/CorpusID:41805341.

[15] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020. URL: https://arxiv.org/abs/1703.10593. arXiv:1703.10593.

[16] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, ArXiv abs/2006.11239 (2020). URL: https://api.semanticscholar.org/CorpusID:219955663.

[17] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, 2018. URL: https://arxiv.org/abs/1611.07004. arXiv:1611.07004.

[18] X. Liu, C. Gong, Q. Liu, Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL: https://arxiv.org/abs/2209.03003. arXiv:2209.03003.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2018. URL: https://arxiv.org/abs/1708.02002. arXiv:1708.02002.

[20] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, in: Neural Information Processing Systems, 2017. URL: https://api.semanticscholar.org/CorpusID:19046372.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014. URL: https://arxiv.org/abs/1406.2661. arXiv:1406.2661.

[22] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014. URL: https://arxiv.org/abs/1411.1784. arXiv:1411.1784.

[23] J. Gauthier, Conditional generative adversarial nets for convolutional face generation, 2015. URL: https://api.semanticscholar.org/CorpusID:3559987.

[24] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[25] S. V M, S. George, A review on medical image denoising algorithms, Biomedical Signal Processing

and Control 61 (2020) 102036. doi:10.1016/j.bspc.2020.102036.
[26] C. Tomasi, R. Manduchi, Bilateral filtering for gray and color images, Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271) (1998) 839–846. URL: https://api.semanticscholar.org/CorpusID:14308539.
[27] G. Charpiat, N. Girard, L. Felardos, Y. Tarabalka, Input similarity from the neural network perspective, in: Neural Information Processing Systems, 2019. URL: https://api.semanticscholar.org/CorpusID:202779680.
[28] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, 2016. URL: https://arxiv.org/abs/1603.08155. arXiv:1603.08155.
[29] C. Depretto, A. Borelli, A. Liguori, G. Presti, A. Vingiani, F. Cartia, C. Ferranti, G. P. Scaperrotta, Contrast-enhanced mammography in the evaluation of breast calcifications: preliminary experience, Tumori Journal 106 (2020) 491 – 496. URL: https://api.semanticscholar.org/CorpusID:219553317.
[30] W. Lotter, G. Kreiman, D. Cox, Unsupervised learning of visual structure using predictive generative networks, 2016. URL: https://arxiv.org/abs/1511.06380. arXiv:1511.06380.
[31] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, 2018. URL: https://arxiv.org/abs/1807.06521. arXiv:1807.06521.
[32] H. Li, P. Xiong, J. An, L. Wang, Pyramid attention network for semantic segmentation, 2018. URL: https://arxiv.org/abs/1805.10180. arXiv:1805.10180.
[33] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, 2018. URL: https://arxiv.org/abs/1711.07971. arXiv:1711.07971.
[34] P. Cao, F. Xie, S. Zhang, Z. Zhang, J. Zhang, Msanet: Multi-scale attention networks for image classification, Multimedia Tools and Applications 81 (2022) 34325 – 34344. URL: https://api.semanticscholar.org/CorpusID:248782198.
[35] J. Phillips, J. Steinkeler, K. Talati, A. Brook, V. Dialani, M. Fishman, P. J. Slanetz, T. S. Mehta, Workflow considerations for incorporation of contrast-enhanced spectral mammography into a breast imaging practice, Journal of the American College of Radiology 15 (2018) 881–885. URL: https://www.sciencedirect.com/science/article/pii/S1546144018302059. doi:https://doi.org/10.1016/j.jacr.2018.02.012.

Acknowledgments

We would like to thank the Istituto Oncologico Veneto (IOV – IRCCS) for providing the dataset used in this research. https://www.ioveneto.it/en/