PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

Improving the Generalization of Mammography Segmentation Models for Multiple Equipments

João Pedro Monteiro Maia

PROJETO FINAL DE GRADUAÇÃO
CENTRO TÉCNICO CIENTÍFICO - CTC
DEPARTAMENTO DE INFORMÁTICA
Curso de Graduação em Ciências da Computação

Rio de Janeiro Julho 2024

Pontifícia Universidade Católica
do Rio de Janeiro

João Pedro Monteiro Maia

Improving the Generalization of Mammography Segmentation Models for Multiple Equipments

Final Project

Final Project presented to the Computer Science Course of PUC-Rio in partial fulfillment of the requirements for the degree of Bachelor in Computer Science.

Advisor : Prof. Alberto Barbosa Raposo
Co-advisor: Dr. Jan Jose Hurtado Jauregui

Rio de Janeiro
June 2024

Abstract

Monteiro Maia, João Pedro; Barbosa Raposo, Alberto (Advisor); Hurtado Jauregui, Jan Jose (Co-Advisor). **Improving the Generalization of Mammography Segmentation Models for Multiple Equipments**. Rio de Janeiro, 2024. 57p. Final Project Proposal – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Mammography, a low-dose x-ray technology for breast examination, is the primary screening method for early detection of breast cancer, significantly improving treatment success rates. Segmenting key structures in mammography images can enhance medical assessment by evaluating cancer risk and the quality of image acquisition. We introduce a series of data-centric strategies to enrich the training data for deep learning-based segmentation of landmark structures, such as the nipple, pectoral muscle, fibroglandular tissue, and fatty tissue. Our approach involves augmenting training samples through annotation-guided image intensity manipulation and style transfer, aiming for better generalization than standard training methods. These augmentations are applied in a balanced manner to ensure the model processes a diverse range of images from different vendor equipment while maintaining efficacy on the original data. We present extensive numerical and visual results demonstrating the superior generalization capabilities of our methods compared to standard training. This evaluation uses a large dataset of mammography images from various vendors. Additionally, we present complementary results showing both the strengths and limitations of our methods in different scenarios. The accuracy and robustness demonstrated in the experiments suggest that our method is well-suited for integration into clinical practice.

Keywords

Mammography; Semantic segmentation; Deep learning; Generalization.

Resumo

Monteiro Maia, João Pedro; Barbosa Raposo, Alberto; Hurtado Jauregui, Jan Jose. **Improving the Generalization of Mammography Segmentation Models for Multiple Equipments**. Rio de Janeiro, 2024. 57p. Proposta de Projeto Final – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A mamografia, uma tecnologia de raio-X de baixa dose para exame das mamas, é o principal método de triagem para a detecção precoce do câncer de mama, melhorando significativamente as taxas de sucesso do tratamento. A segmentação de estruturas-chave nas imagens de mamografia pode aprimorar a avaliação médica ao avaliar o risco de câncer e a qualidade da aquisição de imagens. Introduzimos uma série de estratégias centradas em dados para enriquecer os dados de treinamento para a segmentação baseada em aprendizado profundo de estruturas de referência, como o mamilo, músculo peitoral, tecido fibroglandular e tecido adiposo. Nossa abordagem envolve o aumento das amostras de treinamento por meio da manipulação da intensidade da imagem guiada por anotação e transferência de estilo, visando uma melhor generalização do que os métodos de treinamento padrão. Essas ampliações são aplicadas de maneira balanceada para garantir que o modelo processe uma ampla gama de imagens de equipamentos de diferentes fornecedores, mantendo a eficácia nos dados originais. Apresentamos resultados numéricos e visuais que demonstram as capacidades superiores de generalização de nossos métodos em comparação com o treinamento padrão. Esta avaliação utiliza um grande conjunto de dados de imagens de mamografia de vários equipamentos. Além disso, apresentamos resultados complementares que mostram as vantagens e as limitações de nossos métodos em diferentes cenários. A precisão e robustez demonstradas nos experimentos sugerem que nosso método é adequado para integração na prática clínica.

Palavras-chave

Mamografia; Segmentação semântica; Aprendizado profundo; Generalização.

Table of contents

List of figures

List of tables

1
Introduction

Mammography is a low-dose x-ray exam of the breast that is one of the most effective screening tools available today [2]. Regular mammograms can help find breast cancer at an early stage when treatment is most likely to be successful. There are several types of mammography, and all of them produce images to be analyzed by professionals [3].

Since mammography is a screening exam, it always produces a result that needs to be interpreted visually. In this process, intelligent machines proved to be able to assist professionals to achieve an astonishing correct identification rate [4]. Several techniques are used to assist these professionals. One technique that showed significant results is image segmentation. Although it can happen in a 3D context, 2D is more common in mammography due to the fact that most of the equipment generates 2D images [3].

Using image segmentation, we can identify and separate landmark structures of interest in mammography images, such as the nipple, the pectoral muscle, the fibroglandular tissue, and the fatty tissue, which can be useful to assist healthcare specialists in better interpreting these images. More precisely, identifying these structures is useful in categorizing the risk of an abnormality and evaluating image acquisition adequacy.

However, segmentation of mammography images can be challenging due to various factors, such as the occlusion caused by the fibroglandular tissue, the inclusion of the minor pectoral muscle, and the inclusion of skin folds, among others. Although several methods have been proposed in the literature to address medical image segmentation [5], a few ones were proposed to address the segmentation of landmark structures in mammography images.

One of those methods was proposed in [6], where a deep neural network is used to segment some landmark structures of interest. This method considers a large private dataset for training and different model architectures, including the U-Net model. The results are promising in the processing of images generated by specific equipment, i.e., General Electric equipment. However, when processing non-similar images acquired using equipment of other vendors, this method presents some limitations.

We propose a set of data-centric strategies to achieve better generaliza-

tion on the processing of mammography images acquired using different vendor equipment. More precisely, we introduce augmentation procedures based on image intensity manipulation and style-transfer methods, incorporating samples during training that enable the model to learn from diverse hypothetical domains. We present extensive numerical and visual results on analyzing the reference method, i.e. [6], and highlighting the benefits of the proposed strategies. These results demonstrate the promising potential of our strategies, making them strong candidates for integration into clinical practice.

The remainder of this document is structured as follows. Chapter 2 introduces some related work relevant to our proposal. Chapter 3 explains the datasets that we will use for our experiments. Chapter 4 presents the reference method [6] and its results on the selected datasets. Chapter 5 presents the proposed methods. Chapter 6 presents numerical and visual results of the proposed methods. Finally, Chapter 7 concludes this work.

This document is based on our manuscript titled "Improving the generalization of deep learning models in the segmentation of mammography images", which was submitted to the journal "Biomedical Signal Processing and Control".

2
Related work

2.1
Mammography images segmentation

While the segmentation of abnormalities like masses or nodules in mammography images is a common focus [7], our attention is directed towards techniques that aim to identify key landmarks enabling the spatial description of breast tissues.

The identification of the pectoral muscle serves as a crucial reference point in the MLO view evaluation, aiding in the assessment of potential abnormalities and the quality of the image acquisition. This anatomical structures is typically represented by a triangular shape in the mammography image corner. Its accurate segmentation is difficult due to varying shapes resulting from diverse anatomical conditions, occlusion from fibroglandular tissue, interference from the minor pectoral muscle, and the presence of skin folds, among other factors. Various methodologies have been introduced in existing literature, leveraging conventional signal processing and statistical analysis [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. Rampun et al. propose a multi-step approach involving a deep learning model for delineating the pectoral muscle boundary, followed by post-processing steps to ensure precise demarcation [19]. Similarly, Soleimani et al. introduce a deep learning model for pectoral muscle boundary segmentation, complemented by graph-based analysis for enhancement [20]. In contrast, Ali et al. achieve complete pectoral muscle shape segmentation using a U-Net-based deep learning model instead of focusing solely on boundary segmentation [21]. Guo et al. suggest a two-step methodology utilizing a U-Net for identifying confident pectoral muscle regions and a GAN for final shape estimation [22], while Rubio and Montiel present a comparative study employing various deep learning models and metrics for pectoral muscle segmentation, also considering breast shape [23]. Furthermore, Yu et al. propose an innovative deep learning model incorporating an attention mechanism, yielding superior outcomes compared to standard encoder-decoder models.

The nipple constitutes another important landmark structure, facilitating

the registration of multiple views or modalities by enabling efficient region matching and anatomical measurements. Various methodologies employ shape and texture analysis across different regions of the breast boundary [24, 25, 26, 27]. Techniques presented in [28] and [29] operate under the assumption that the fibroglandular tissue converges at the nipple, leading to the development of geometric descriptors for optimal convergence point detection. Casti et al. introduce a Hessian-based approach incorporating geometric descriptors and constraints to accurately define the nipple's position [30]. Jiang et al. propose a random forest classifier that leverages quantitative radiomic features to identify subtle nipples and determine relevant regions of interest [31]. Lin et al. propose a deep learning classification model applied to a series of candidate patches extracted from mammography images, where the region with the highest density of classified potential nipple patches is selected as the nipple position [32]. These methods are focused on pinpointing the precise location of the nipple rather than segmenting it.

The fibroglandular tissue represents a critical area of concern warranting specific attention during medical assessments. Depending on the patient's unique anatomy, this tissue can exhibit varying characteristics, ranging from dense formations to more dispersed patterns, with higher density correlating to increased risk. Numerous methodologies have been proposed in the existing literature for segmenting dense fibroglandular tissue regions [33], encompassing both handcrafted approaches [34, 35, 36, 37] and data-driven models [38, 39, 40, 41, 42]. While dense regions are of utmost concern, it is also crucial for clinical practitioners to be attentive to scattered areas, as abnormalities can manifest there as well. Therefore, the segmentation of both dense and scattered regions plays a vital role in accurately describing the spatial composition of the breast.

Several methodologies aim to integrate the segmentation of various landmark structures within a unified framework. Tiryaki et al. conduct experiments employing multiple U-Net-based models to segment the pectoral muscle, dense fibroglandular tissue regions, and adipose tissues [43]. In a similar vein, considering these structures and incorporating the nipple, Dubrovina et al. introduce a novel deep learning-based framework for comprehensive segmentation tasks [44]. By leveraging multiple deep learning models, Bou demonstrates segmentation results encompassing more intricate structures, including vessels, calcifications, and skin, among others [45]. However, due to the utilization of relatively small datasets, the robustness and reliability of these methods in real-world applications might be limited. Additionally, the first two approaches primarily concentrate on the MLO view, a trend shared by most segmentation methods

detailed in this section.

In a recent development, Sierra-Franco et al. [6] unveiled a sizable dataset alongside deep learning experiments for the segmentation of mammography images, encompassing both MLO and CC views. The study highlights four primary structures of interest in both views: nipple, pectoral muscle, fibroglandular tissue, and fatty tissue. We propose a data-centric approach for the improvement of the generalization of the solution introduced in this work on the processing of mammography images generated by different vendor equipments. Although our experiments are performed on the MLO view only, our method is extendable to the CC view also.

2.2 Data augmentation

Data augmentation plays a pivotal role in working with diverse datasets, especially in the medical field, where datasets often suffer from imbalances and limited samples for certain structures and classes. Basically, when dealing with images, we can divide the data augmentation in two types [46], transforming the original data and generating artificial samples.

Commonly used methods for image transformation include a combination of simple affine transformation (such as mirroring, zooming, and resizing), which are applied directly to the images. We can also use erasing transformations to "crop-out" certain portions of the image to avoid simplified detection patterns.

More advanced image transformation done in images are the Greedy Policy Search (GPS) and Mixup. GPS is a method that searches for the optimal augmentation policy by maximizing the validation accuracy of a model. Mixup is a method that generates new training samples as convex combinations of pairs of original samples [47]

Generating artificial/synthetic samples open room for more possibilities, since it overcomes the constrains of transformations. Generative networks are today the most common approach to medical image synthesis [46, 48].

Overall, the use of data augmentation techniques demonstrates a consistent pattern of advantages, as evidenced by the findings in Garcea's study [46]. These benefits are observed across various modalities and tasks within the domain of medical image segmentation and remain consistent across a broad spectrum of augmentation techniques, ranging from fundamental affine transformations to advanced generative methods. Consequently, the integration of data augmentation proves to be a valuable resource when confronted with diverse datasets.

2.3 Style transfer

Style transfer is a powerful technique in the realms of computer vision and graphics, enabling the creation of new images by blending the content of one image with the stylistic elements of another [49].

One common way to implement style transfer is by utilizing neural networks. These networks take as input a content image and one or more style images, alongside an additional vector that specifies the degree to which each style should be applied to the content image. The output is a visually image that combines the content of the source image with the artistic style of the other(s).

The application of style transfer extends beyond the realm of artistry; it has proven to be particularly beneficial in data augmentation, which is essential for machine learning tasks involving small datasets. Style transfer, when used as a form of data augmentation, adds diversity and variability to the training dataset, thus enhancing model performance. By preserving the high-level semantic content of the content image while adopting the style of the reference image, neural style transfer effectively enriches the training data, making it a valuable tool for improving the performance of models in image classification tasks.

A notable study in this context is the STaDA (Style Transfer as Data Augmentation) paper by Zheng et al. [50]. In their research, they explore the use of state-of-the-art neural style transfer algorithms as a data augmentation technique for image classification tasks. Their experiments, conducted on datasets like Caltech 101 and Caltech 256, yielded significant improvements in image classification accuracy when compared to traditional data augmentation methods. For instance, they reported an approximate 2% increase in accuracy with the VGG-16 model. To maximize the benefits, they also combined neural style transfer with conventional data augmentation strategies, achieving even better performance in image classification.

An alternative approach to data augmentation through style transfer is presented in [51]. This method offers an efficient and annotation-free way to enhance image datasets. The process involves targets mask generation, style transfer, and the addition of details to images. One distinct feature is that it doesn't require additional manual annotation work. This technique was successfully applied to create a dataset of military vehicle images, and the results were impressive. In high-contrast situations, they achieved improvements in precision by 0.101, while in low-contrast situations, they improved precision by 0.134. These improvements were observed when using both single-style and

multi-style stylized image datasets.

These works demonstrate the potential of style transfer in computer vision fields, such as helping to reduce the difficulty of collecting sufficient labeled data and improving the performance on small datasets.

3
Datasets

Typically, mammography examinations consist of two primary views: Medio-Lateral Oblique (MLO) and Cranio-Caudal (CC), which are taken for both breasts. These imaging modalities provide both a top-to-bottom and side-on perspective of the breast, enabling a comprehensive analysis from multiple angles. These views capture critical anatomical structures that are essential for detecting abnormalities and assessing the quality of image acquisition. For this work, we mainly consider the MLO view.

We consider MLO view digital mammography images from the private dataset introduced in [6] and the VinDr-Mammo dataset introduced in [52]. We compose four different datasets, each one representing a different vendor of mammography equipments. These datasets are named as GE, IMS, PLANMED, and HOLOGIC, containing MLO view mammography images generated by equipments of the General Electric, IMS Giotto, Planmed Oy, and Hologic vendors, respectively.

The primary purpose of these datasets is image segmentation, and it includes annotations for four major landmarks: the nipple, pectoral muscle, fibroglandular tissue, and fatty tissue. A team of eight annotators received training from two clinical experts to identify and delineate these structures accurately using a contour drawing tool. Then, these contours, represented as polygons, are rasterized to generate a multi-class label map for each structure. All the left breast images are horizontally flipped to simplify the input domain. For further details about this annotation process and how the label maps are generated, please refer to [6].

Trying to uniformize the input images, all the images follow the preprocessing stablished in [6]. The mammography images are normalized using the percentiles 2 and 98 as minimum and maximum values, then equalized using Contrast Limited Adaptive Histogram Equalization (CLAHE) [53] with kernel size being 1/8 of the height and width of the image, and finally re-scaled to the range $[0, 255]$. For the IMS and PLANMED datasets, we include additional processing due to the different image format that is adopted for these cases. Figure 3.1 shows an example of the pre-processed image and its corresponding label map.

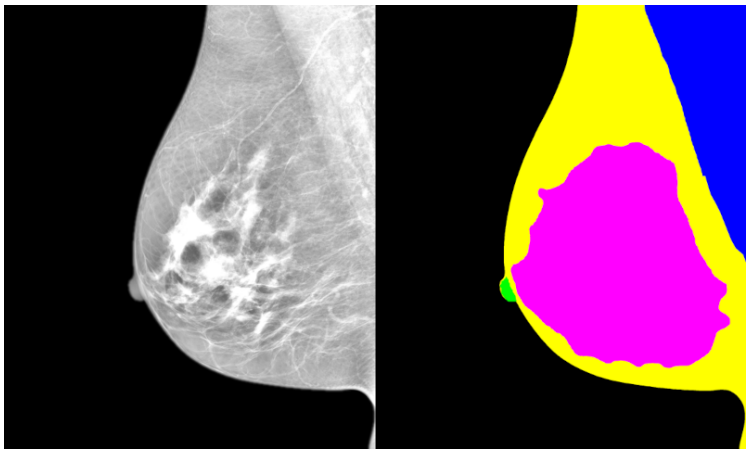

Figure 3.1: Pre-processed image and its corresponding label map. The nipple is colored in green, the pectoral muscle is colored in blue, the fibroglandular tissue is colored in magenta, the fatty tissue is colored in yellow, and the background is colored in black.

Although, we present different datasets, we just consider the GE dataset for the training task. The other datasets are used for testing purposes only. A fully detailed specification of each dataset is presented in the following sections.

3.1 GE dataset

A collection of 5214 MLO view mammography images was selected to construct this dataset, belonging to the acquisition of three types of GE equipments: Senographe Essential, Senograph DS, Senographe Pristina, and Senographe Crystal. All of these equipments present similar images that were fully annotated and pre-processed using the standard method explained above. Figure 3.2 shows some samples of this dataset.

The annotated samples are split into the three standard subsets considered in a conventional supervised learning pipeline: training, validation, and test. The splitting process follows a random behavior with certain balancing regarding the fibro-glandular tissue density and avoiding data leakage, i.e. we avoid including the same accession number in different sets. This distribution results in 3450 samples for training ($\sim 70\%$), 1206 samples for validation ($\sim 20\%$), and 557 samples for test ($\sim 10\%$). We use this dataset for both training and testing purposes.

3.2 IMS dataset

This dataset considers a collection of 52 MLO view mammography images acquired using the GIOTTO CLASS and GIOTTO IMAGE 3DL equipments.

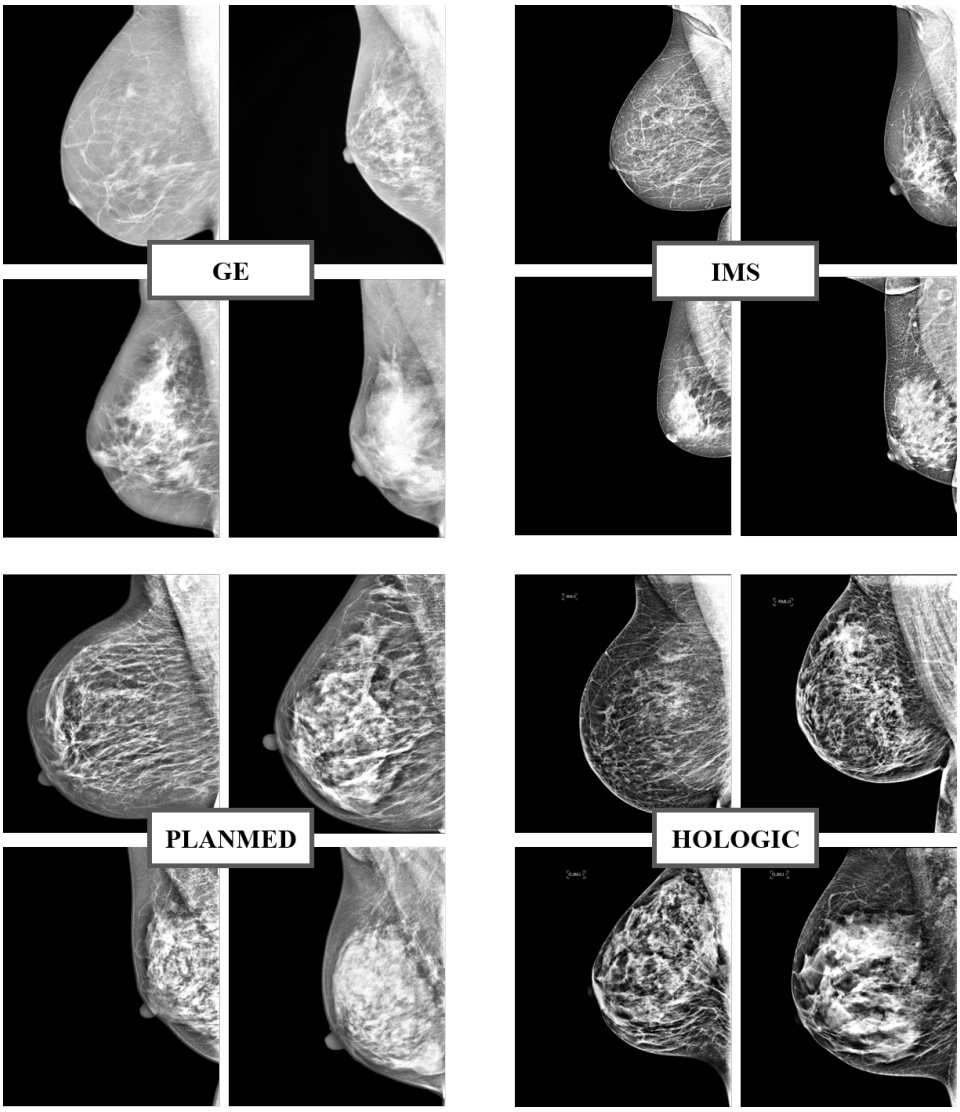

Figure 3.2: Image samples.

Differently from the GE images, for this dataset, we included an additional step for pre-processing because of the different formats. As the first pre-processing step, we use the window center and window width metadata to rescale the intensity values. Let us denote as c the window center and w the window width. The rescaling minimum $x_{\min}$ and maximum values $x_{\max}$ are computed as follows: $x_{\min} = c - \lfloor w/2 \rfloor - \lfloor 0.25w \rfloor$ and $x_{\max} = c + \lfloor w/2 \rfloor$. The rest of the steps are the same as explained above. Figure 3.2 shows some samples of this dataset. In this work, this dataset is used for testing purposes only.

3.3 PLANMED dataset

In this dataset, we include 48 MLO view mammography images acquired using the Planmed Nuance equipment. As in the IMS case, we also include an

additional first step for pre-processing due to the image format that presents inverted values. Thus, we adopt the following minimum and maximum values to rescale the negative version of the input image: $x_{\min} = -(c + \lfloor w/2 \rfloor + \lfloor 0.25w \rfloor)$ and $x_{\max} = -(c - \lfloor w/2 \rfloor)$. The rest of the steps are the same as explained above. Figure 3.2 shows some samples of this dataset, which is also used for testing purposes only.

3.4 HOLOGIC dataset

This dataset includes a collection of 34 MLO view mammography images acquired using Selenia Dimensions equipment. In this case, these images follow the standard pre-processing pipeline, as in the GE case. Figure 3.2 shows some samples of this dataset. We use this dataset for testing purposes only.

4
Mammography image segmentation

In this chapter, we present the reference approach proposed in [6], which modeled the problem as a semantic segmentation task that can be tackled using deep learning models. More precisely, we describe a baseline model and its corresponding training settings, numerical results on the different datasets, and visual results useful to discuss the benefits and drawbacks of this method in the processing of mammography images of different vendors' equipment.

4.1
Model training

While [6] presents diverse experiments involving various deep learning model architectures and training configurations, this study adopts as baseline a U-Net architecture in conjunction with an EfficientNetB3 model serving as a feature extractor (backbone). The network input consists of a single-channel image with dimensions 384×384, with intensity values in the range $[0, 1]$. The network's output takes the form of a $384 \times 384 \times C$ per-pixel probability map, where C is the number of classes, encompassing an implicit background class for unannotated pixels. Given that the segmentation task is treated as a multi-class per-pixel classification problem, the final layer incorporates a softmax activation function. For the training phase, we employe a hybrid loss function combining Categorical Focal Loss and Jaccard Loss functions, with a batch size of 4, learning rate of 10^{-3}, and a maximum of 200 epochs, integrating early stopping with a patience of 30.

The model is trained on the GE training set without considering augmentation operations and using the GE validation set to select the best weights regarding the loss function. Figure 4.1 shows the training evolution through the epoch, where we can see that the model rapidly converges due to the high amount of images and the best weights are obtained in the 15th epoch.

4.2
Numerical results

To evaluate the model, we consider the different datasets presented in the previous chapter that represent mammography images of different

Figure 4.1: Baseline model training evolution through the epochs. The x-axis represent the epochs and the y-axis represents the loss values. The training set loss is shown in blue while the validation set loss is shown in orange.

Table 4.1: Baseline approach IoU results

Dataset	Nipple	Pectoral	Fib. Tissue	Fat. Tissue	Mean
GE (validation)	0.7641	0.9695	0.9116	0.8401	0.8713
GE (test)	0.7488	0.9608	0.9069	0.8078	0.8561
IMS	0.7401	0.9165	0.7120	0.6070	0.7439
PLANMED	0.7015	0.9432	0.7736	0.5962	0.7536
HOLOGIC	0.1463	0.7677	0.6487	0.4192	0.4955

vendors' equipment. We use the metric Intersection Over Union (IoU) which is a widely used metric for the semantic segmentation evaluation (For more details about this metric, please refer to Chapter 5). This metric measures the degree of overlap between the segmentation prediction and the ground-truth segmentation (annotation). Thus, we can apply this metric to each class, obtaining IoU scores for each structure.

Table 4.1 shows the IoU results on the different datasets. As expected, the model presents good results on the validation and test sets of the GE dataset, similar to the results found in [6]. We can see pectoral muscle IoU scores close to 0.96 and fibroglandular tissue results close to 0.91. The nipple seems to be the most challenging structure; however, it presents lower values because it is a small structure that tends to be more sensitive to the metrics. Thus, this is a good model for the segmentation of mammography images generated by GE equipments. Recall that during training, we just use GE images.

In contrast, we can see that the results over the IMS, PLANMED and HOLOGIC datasets are considerably worse in average, especially in the HOLOGIC case. This means that the model do not present a good generalization when working with images generated by different vendors equipment. This is mainly caused by the image differences that are not considered during training.

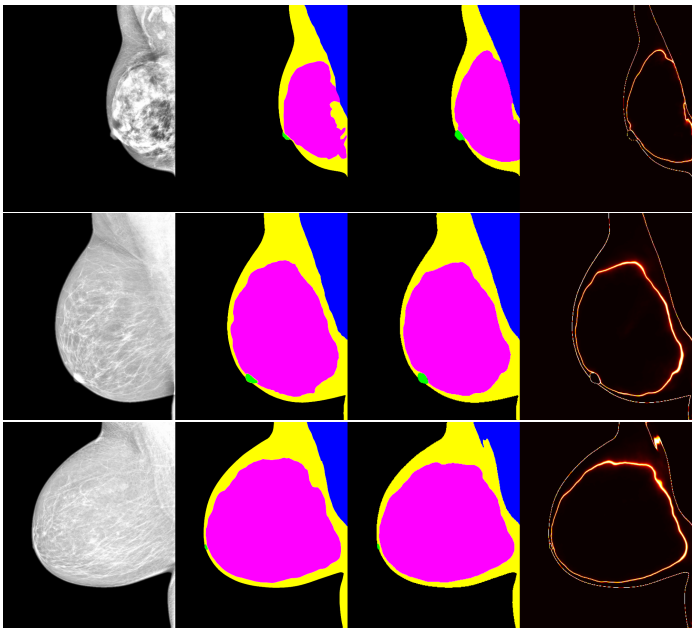

Figure 4.2: GE results. Each row represents a different case. First column: input image. Second column: ground-truth annotation. Third column: prediction. Fourth column: uncertainty map (hot color map with values in the range $[0, 1]$).

4.3 Visual results

Figures 4.2, 4.3, 4.4, and 4.5, show some visual results on GE, IMS, PLANMED, and HOLOGIC images, respectively. In addition to the predicted structures, we are showing an uncertainty map computed using Test Time Augmentation (TTA) that allows us to highlight the regions where the model presents high uncertainty (For more details about the uncertainty map computation, please refer to Chapter 5). In other words, the highlighted regions are the regions where the model presented more doubts.

In the GE case, we can see how the predictions are close to the ground-truth annotations and the uncertainty maps are well behaved, i.e. the highlighted regions are close to the prediction boundaries, which is an expected behavior of a good segmentation. Differently, for the other vendors, we can see noisy predictions and chaotic uncertainty maps that highlight thick regions in most cases. These prediction noise and uncertainty map chaoticity are indicators that the model is not performing well on these cases.

Further, Figure 4.6 shows some predictions on mammography images obtained from the DDSM dataset [54] that were generated using screen-film technology. All our datasets consider digital mammography technology. Notice how the model generates noisy and inaccurate predictions.

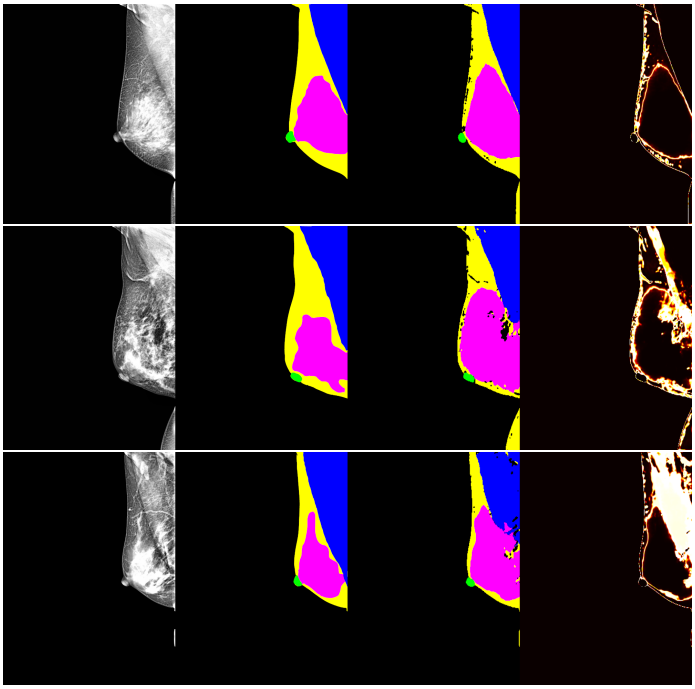

Figure 4.3: IMS results. Each row represents a different case. First column: input image. Second column: ground-truth annotation. Third column: prediction. Fourth column: uncertainty map (hot color map with values in the range $[0, 1]$).

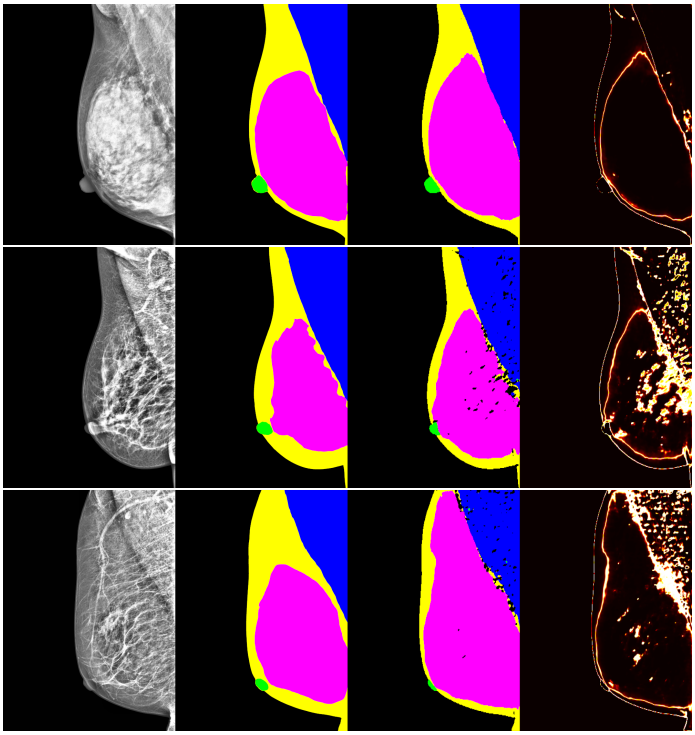

Figure 4.4: PLANMED results. Each row represents a different case. First column: input image. Second column: ground-truth annotation. Third column: prediction. Fourth column: uncertainty map (hot color map with values in the range $[0, 1]$).

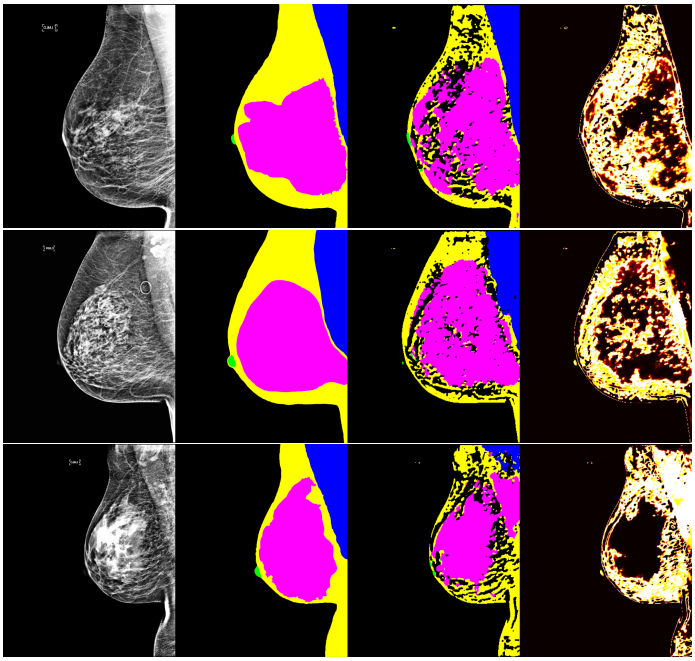

Figure 4.5: HOLOGIC results. Each row represents a different case. First column: input image. Second column: ground-truth annotation. Third column: prediction. Fourth column: uncertainty map (hot color map with values in the range $[0, 1]$).

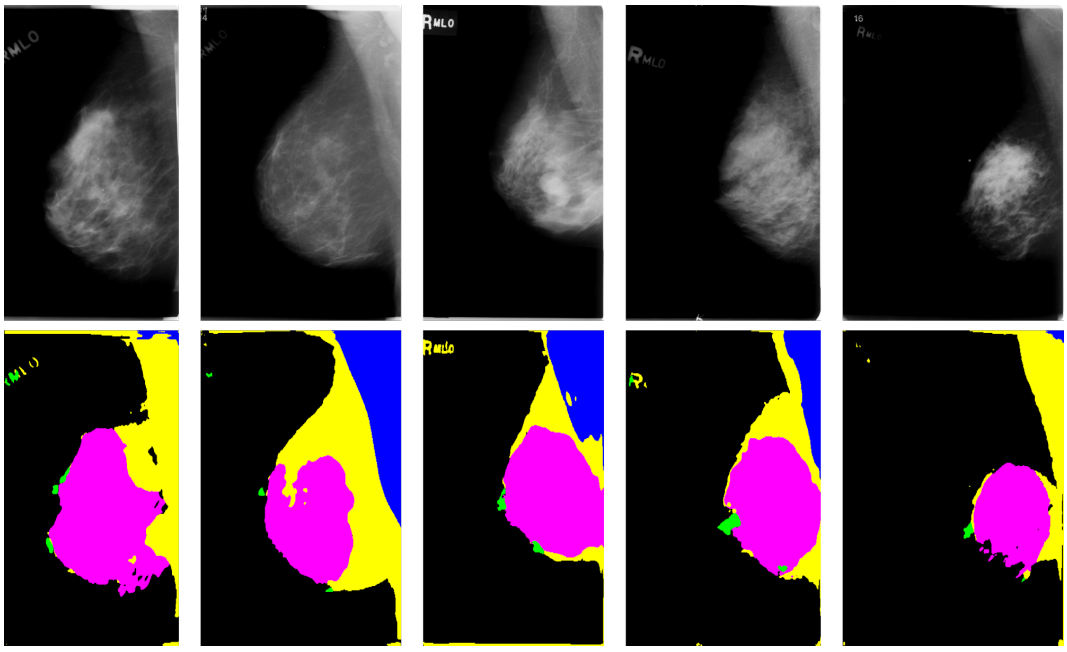

Figure 4.6: Screen-film mammography results. Each column represents a different case. First row: input images. Second row: predictions.

4.4 Discussion

Based on these results, we can see that training on the GE dataset presents several limitations when dealing with images generated by other vendor equipments, where the corresponding model presents noisy predictions and large high uncertainty areas. Enhancing this model to optimize its performance in processing images from diverse vendor equipment could significantly contribute to its successful integration into clinical practice.

5
Improving the generalization of segmentation models

The model proposed in [6] is effective, yet, as discussed in the previous chapters, there is room for improvement, particularly when handling datasets with images generated by equipment from non-GE vendors. Currently, the model performs well on one dataset (GE) but falls short on others, especially those with more diverse exam images.

As shown in the data analysis, the GE dataset contains nearly 5000 images, whereas the other datasets (from different equipment) only have 50-100 images each. This disparity makes it challenging to validate our models confidently, as there are insufficient samples for reliable training, testing, and validation. Additionally, training separate models for each type of equipment is not practical for real-world applications, as it introduces new issues, such as requiring users to know which model to use.

To address these challenges, we employ advanced data augmentation methods on the large GE dataset. Our goal is to develop a single model that performs well across various type of images. Instead of relying solely on basic affine transformations like cropping and resizing, we explore the use of image intensity manipulation and style transfer to incorporate samples that approximate a variate domain of images. By training the model with these synthetic samples, we aim to enhance its performance and generalization. We introduce two distinct methodologies for this improvement, along with a combined approach.

5.1
Annotation-guided image manipulation for data augmentation

Data augmentation is usually related to applying random image transformations to the existing samples to achieve better generalization and robustness. The characteristics of the target domain guide the selection of these transformations we aim to represent. Thus, we propose a set of operations for manipulating image intensity values, enabling better representation of non-GE images.

Algorithm 1 Image manipulation

procedure MANIPULATE($\mathbf{I}_{\text{in}}$,$\mathbf{M}_{\text{nip}}$,$\mathbf{M}_{\text{fib}}$,$\mathbf{M}_{\text{fat}}$,$\mathbf{M}_{\text{b}}$)
  $\mathbf{I} = \text{rand}(0.8, 1.2) * \mathbf{I}_{\text{in}}$
  if $\text{rand}(0, 1) < 0.5$ **then**
    return $\mathbf{I}$
  $\mu_{\text{nip}} = \text{mean}(\mathbf{I}, \mathbf{M}_{\text{nip}})$
  $\mu_{\text{fat}} = \text{mean}(\mathbf{I}, \mathbf{M}_{\text{fat}})$
  $\mu_{\text{fib}} = \text{mean}(\mathbf{I}, \mathbf{M}_{\text{fib}})$
  $p_{\text{fat}} = \text{percentile}_5(\mathbf{I}, \mathbf{M}_{\text{fat}})$
  $a_{\text{min}} = \text{clip}(\text{rand}((p_{\text{fat}} - 20), (p_{\text{fat}} + 20)), 0, 255)$
  $b = 0.7\mu_{\text{fat}} + 0.3\mu_{\text{fib}}$
  if $a_{\text{min}} > (\mu_{\text{nip}} - 5)$ **then**
    $a_{\text{min}} = \max(0, (\mu_{\text{nip}} - 5))$
  else if $\text{rand}(0, 1) < 0.5 \wedge (\mu_{\text{nip}} - 5) < b$ **then**
    $a_{\text{min}} = \text{rand}(\max(0, (\mu_{\text{nip}} - 5)), \mu_{\text{nip}})$
  $a_{\text{max}} = \text{percentile}_{98}(\mathbf{I})$
  $\mathbf{I}_{\text{out}} = \text{rescale\_intensity}(\mathbf{I}, (a_{\text{min}}, a_{\text{max}}), (0, 255))$
  if $\text{rand}(0, 1) < 0.5$ **then**
    $\mathbf{I}_{\text{out}}[\mathbf{M}_{\text{b}}] = 0$
  if $\text{rand}(0, 1) < 0.5$ **then**
    $\mathbf{I}_{\text{out}} = \text{add\_label}(\mathbf{I}_{\text{out}})$
  return $\mathbf{I}_{\text{out}}$

5.1.1 Generation of augmented samples

The general idea of our custom image augmentation procedure is to rescale the intensity values using the information of the annotated structures. Algorithm 1 summarizes this procedure, which receives as input an image $\mathbf{I}_{\text{in}}$, the binary mask $\mathbf{M}_{\text{nip}}$ of the nipple, the binary mask $\mathbf{M}_{\text{fib}}$ of the fibroglandular tissue, the binary mask $\mathbf{M}_{\text{fat}}$ of the fatty tissue, and the binary mask $\mathbf{M}_{\text{b}}$ of the background, and returns a manipulated version of $\mathbf{I}$, i.e. $\mathbf{I}_{\text{out}}$.

The function $\text{mean}(\mathbf{I}, \mathbf{M})$ computes the mean intensity of $\mathbf{I}$ considering the values within the mask $\mathbf{M}$ only. The function $\text{percentile}_5(\mathbf{I}, \mathbf{M})$ computes the 5th percentile of the values of $\mathbf{I}$ within the mask $\mathbf{M}$. The function $\text{percentile}_{98}(\mathbf{I})$ computes the 98th percentile of the intensity values of $\mathbf{I}$. The function $\text{rand}(x_{\text{init}}, x_{\text{end}})$ generates a random float value between x_{init} and x_{end}. The function $\text{clip}(x_{\text{val}}, x_{\text{init}}, x_{\text{end}})$ is the classic clip function that limits the value x_{val} within the range $[x_{\text{init}}, x_{\text{end}}]$. The function

$$\text{rescale\_intensity}(\mathbf{I}, (x_{\text{min}}, x_{\text{max}}), (y_{\text{min}}, y_{\text{max}}))$$

rescales the intensity values of $\mathbf{I}$ to the range $[y_{\text{min}}, y_{\text{max}}]$, considering x_{min} and x_{max} the minimum and maximum values for $\mathbf{I}$, respectively. The operation

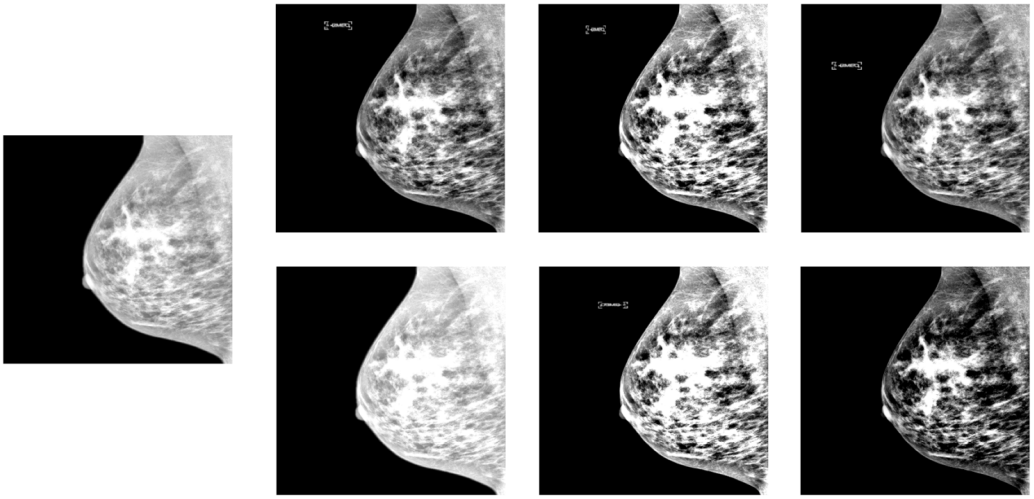

Figure 5.1: Image manipulation example. The most left image is the image $\mathbf{I}_{\text{in}}$. The other images are different results of applying the image manipulation algorithm.

$\mathbf{I}[\mathbf{M}] = x$ assigns the value x to all elements of the image $\mathbf{I}$ that fall within the mask $\mathbf{M}$. Finally, the function add_label($\mathbf{I}$) adds a synthetic view label to the image, considering a random location close to the top-left corner.

The intuition of this augmentation procedure is to achieve higher contrast within the breast region, simulating the behavior noticed in the non-GE equipment images. This manipulation uses local intensity statistics of the annotated structures to achieve robustness and avoid erasing regions of interest from the image, such as the nipple. Further, as shown in Figure 4.5, HOLOGIC images always include a label describing laterality and view position. For this reason, we randomly add a synthetic label to simulate this case. Figure 5.1 shows some examples of our custom augmentation procedure.

5.1.2 Segmentation model training settings

To train the segmentation model, we use the same settings described in Chapter 4 and include the custom image intensity manipulation procedure across all training and validation images. This annotation-guided augmentation method allows us to modify images in a context-aware manner, enhancing the model's ability to generalize across multiple vendor scenarios.

5.2 Style transfer-based data augmentation

Style transfer synthesizes novel images by merging the content of one image with the style of another. Various deep learning frameworks provide pre-trained models for style transfer, which can be fine-tuned to specific styles. Once trained, these models can effectively transfer the learned style to any input image, serving as an effective tool for data augmentation.

We aim to use style transfer to generate images resembling those from the non-GE equipment datasets, creating three different stylization models that adapt GE images to the IMS, PLANMED, and HOLOGIC styles. Then, using these models, we augment the training dataset to enhance generalization.

5.2.1 Style transfer model training

First, we select a reference image for each non-GE dataset, i.e. IMS, PLANMED, and HOLOGIC. These images are shown in Figure 5.2. Then, we fine-tune the model MLStyleTransfer from Apple's CreateML framework [55] to capture the style of each selected image. This fine-tuning process results in three distinct models, each capable of processing a 512×512 3-channel image and producing a similarly dimensioned stylized output. The models are fine-tuned over 550 iterations, a style strength of 6, and a style density of 256. We validated the training process by visually assessing the stylized results on GE images.

5.2.2 Generation of augmented samples

After training the models, we apply them to the entire GE training set to create synthetic images based on IMS, PLANMED, and HOLOGIC styles. To prepare these images for segmentation model training, we convert the stylized images into 384×384 single-channel images, which are the required input of our segmentation model. Additionally, to mitigate artifacts generated during the stylization process, we zero out all pixels within the annotated background region. Figure 5.3 illustrates the post-processing operation, while Figure 5.4 shows examples of the stylization process using the three models.

5.2.3 Segmentation model training settings

To train the segmentation model using the stylized images, we keep the same settings described in Chapter 4. Throughout the training process, we aim

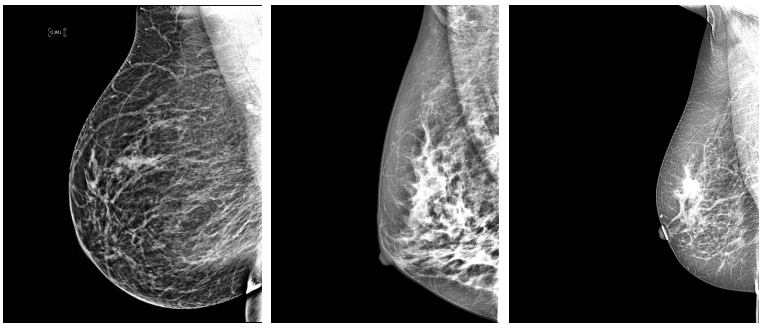

Figure 5.2: HOLOGIC, PLANMED and IMS base images used to train style transfer models

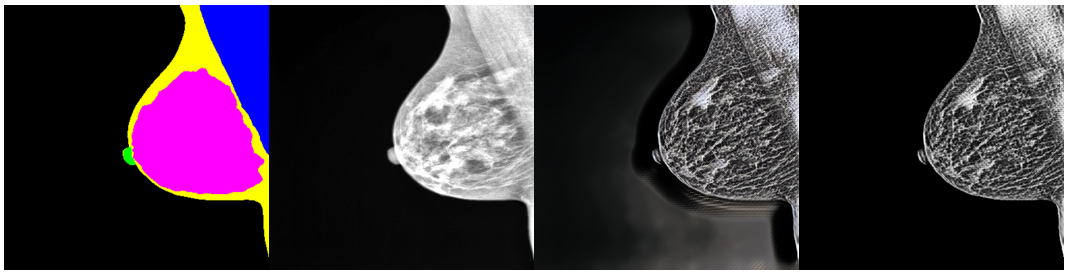

Figure 5.3: Postprocessing of stylized images. First column: annotated regions, where the background is colored in black. Second column: original image. Third column: stylized image. Fourth column: postprocessed image.

to achieve a balanced distribution between the original and various stylized image versions. As a result, each sampled input is equally likely to be either an original, IMS stylized image, PLANMED stylized image, or HOLOGIC stylized image, with a 25% probability for each category. The same processing is applied to the validation set.

5.3 Combining image manipulation with style transfer

Both image manipulation and style transfer strategies offer unique advantages and drawbacks. Combining these methods can yield a more robust approach that enhances the segmentation model's generalization capabilities. We implement a straightforward combination by allocating a 20% probability to each category of images: original, image manipulation results, and stylized images from IMS, PLANMED, and HOLOGIC datasets. As with previous cases, we preserve the same settings presented in Chapter 4 to train the segmentation model, applying this augmented approach to both the training and validation sets.

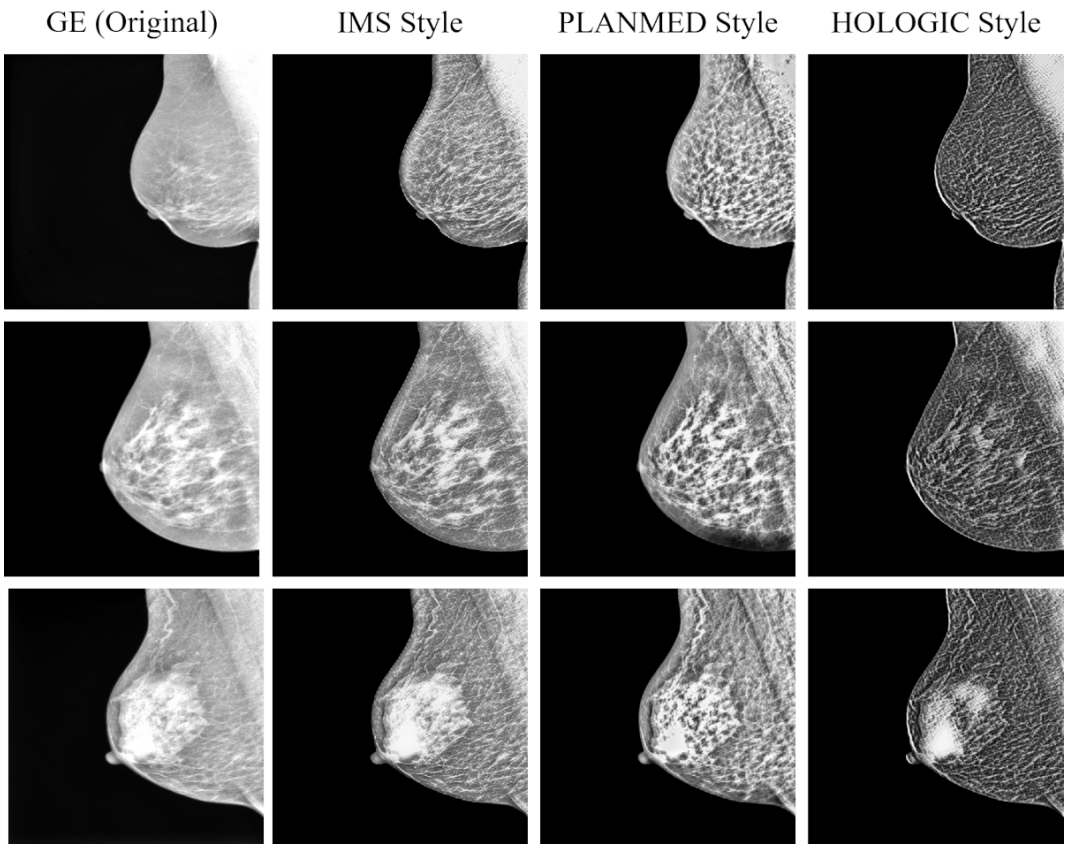

Figure 5.4: Style transfer examples. Each row is a different case. First column: original GE image. Second column: IMS stylization results. Third column: Second column: PLANMED stylization results. Second column: HOLOGIC stylization results.

6
Results

6.1
Evaluation methodology

To evaluate the performance of the trained models, we consider numerical metrics and visual representations that are explained in the following.

6.1.1
Numerical metrics

In our proposal, evaluation techniques play a crucial role. Our aim is to create a model that outperforms the baseline model introduced in [6] on the same dataset, across various structures and datasets. To gauge the effectiveness of our solution, we will evaluate it on different metrics, compare the results with the baseline model, and assess the performance of our solution.

We will consider a range of metrics that highlight different aspects of image segmentation. The following subsections will outline these metrics, their basic aspects, and their workings.

6.1.1.1
Precision and Recall

Precision and recall are important metrics for evaluating many machine learning models. These metrics rely on a confusion matrix that summarizes a model's performance by comparing the true labels to the predicted ones.

According to the Google Machine Learning Developer Site [56], precision addresses the question: "What proportion of positive identifications was actually correct" This can be expressed using the following formula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Here, TP represents true positives (correctly identified positive instances), and FP stands for false positives (incorrectly identified positive instances).

On the other hand, recall attempts to answer the question: "What proportion of actual positives was identified correctly?" Mathematically, it is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

In this equation, FN represents false negatives (actual positive instances that were incorrectly classified as negative).

These two metrics are particularly helpful in the classification context, however, we can adapt it to work in our image segmentation scenario. One way of doing so is to compare the number of correct pixels for each label in both the ground-truth and predicted masks.

Our proposal includes these two metrics to draw a precision and recall curve over different datasets and compare their performance.

6.1.1.2 Accuracy score

Accuracy is a metric calculated by dividing the number of correct predictions by the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

It is important to note that accuracy may not always be the best metric to use, especially in cases where the classes are imbalanced which is our case. However, this metric can still be used for comparison purposes. Our proposal includes using it to evaluate the proposed performance over the baseline as well as to compare it with different training techniques.

6.1.1.3 Dice coefficient (F1-score)

The F1-score is widely used for measuring the performance of image segmentation algorithms. It builds on top of both precision and recall metrics to calculate the Jaccard Index (IoU), which we'll discuss later, and the Dice Similarity Coefficient (DSC). Both of these metrics measure the overlap between the predicted and ground truth masks.

The Dice Coefficient, also called the "Sørensen–Dice coefficient", measures the similarity between two sets, and in our context, measures the similarity between two masks. It is calculated by the following formula:

$$\text{Dice Coefficient} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$$

The Dice Coefficient is very appropriate for imbalanced datasets since it is more sensitive. Therefore, it is going to be an important metric to compare our model over the baseline.

6.1.1.4 Intersection over Union

The Intersection over Union (IoU), also known as Jaccard Index, is calculated using the area of the intersection over the union of the predicted segmentation and the ground truth. It was the main metric used in [6] for evaluation. The score can be computed using the following formula:

$$\text{IoU (Jaccard Index)} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

Nevertheless, both the F1 and the IoU are positively correlated for any fixed ground-truth. That is to say, if our model is better than the other one under the F1 metric, it is also better under the IoU metric. However, this does not mean that we do need both of them or we can arbitrarily which one we will use.

The difference between these metrics appears when dealing with a set of inferences that are not very similar. When quantifying how much worse one model is compared to another. In general, the IoU metric tends to penalize isolated instances of poor classification more severely than the F1-Score, even when both metrics agree that a specific instance is bad. Therefore, the F1-Score tends to measure something closer to the average performance, while the IoU score leans towards measuring worst-case performance.

Since we have a highly imbalanced dataset, both metrics will be used to evaluate the performance of our model. We want to have a better overall performance, but achieving a higher worst-case performance is also promising.

6.1.1.5 Hausdorff distance

The Hausdorff distance measures how far two subsets of a metric space are from each other. Two sets of points are close in the Hausdorff distance if every point of their set is close to some point of the other set. We can express this distance as the longest distance one needs to travel anywhere in set A to reach the set B.

The Hausdorff distance is a widely used performance measure to calculate the distance between two point sets in medical image segmentation. It is used

to compare ground truth images with segmentation results and allows ranking different segmentation results.

Another important property is that Hausdorff distance should be preferred for in segmentation tasks with complex boundaries and small, thin segments, as demonstrated in the literature [57]. This characteristic makes it suitable for our scenario where fine details matter, such as the boundaries of anatomical structures like nipples. Therefore, we will use it to evaluate our model performance.

6.1.2 Visual evaluation

There is no way we can fully interpret the results of a vision model without a visual inspection. To perform a visual inspection we show the predicted labels for a given input image as well as the corresponding ground-truth labels. This method proved to be very effective in [6] since it can show where we can improve our model.

6.1.2.1 Uncertainty map

Uncertainty maps reveal areas where our model encounters difficulty in determining the appropriate mask. Typically, these uncertainties are more pronounced in the border regions of our labels. One method for generating an uncertainty map involves employing Test-time Augmentation (TTA). This process entails applying a series of random transformations to the input image during evaluation, resulting in multiple variations of the original image. The model then makes predictions on this augmented dataset, and the final prediction is derived by averaging the predictions from all the augmented images.

TTA can also serve as a tool to gauge the uncertainty of the model's predictions. By employing TTA, we can compute an uncertainty map that accentuates regions where the model exhibits significant uncertainty. This map is generated by calculating the variance or entropy of predictions across the augmented images. Regions displaying high variance or entropy correspond to areas where the model struggles to ascertain the correct segmentation [58].

This technique was used in Chapter 4, particularly in Figures 4.2, 4.3, 4.4, and 4.5. As mentioned earlier, some datasets exhibit noisy uncertainty maps, signifying that this model is unsure about most parts of those datasets. Since our goal is to enhance generalization to other models in our proposal, achieving less noisy results would represent a significant improvement.

6.2 Experiment settings

We use the datasets described in Chapter 3 to compare the methods we proposed in this work with the baseline method outlined in Chapter 4. We assess numerical performance using six standard metrics commonly used for evaluating semantic segmentation methods. These metrics include precision, recall, accuracy, Dice coefficient (F1-Score), IoU, and Hausdorff distance. Specifically, for the Hausdorff distance, we calculate the average of the two one-sided Hausdorff distances between the prediction and ground-truth structure contours in meters. We then present the metric values for each structure of interest by averaging these measurements across all tested images. We also present a mean value representing all the structures, excluding the background class. For visual analysis, we present the predictions and uncertainty maps as illustrated in Chapter 4.

6.3 Generalization and segmentation accuracy results

Tables 6.1, 6.2, 6.3, and 6.4 present the numerical results for the GE, IMS, PLANMED, and HOLOGIC datasets, while Figures 6.1, 6.2, 6.3, and 6.4 illustrates the corresponding visual results.

The numerical results from the GE dataset (Table 6.1) show that the proposed data augmentation methods perform comparably to the baseline method. This results suggests that the proposed approaches do not degrade performance on this type of images. Moreover, they yield improved segmentation for the pectoral muscle, which is one the most important structures for mammography positioning analysis. Figure Figure 6.1 illustrate the consistent quality of predictions across various anatomies, even in complex cases when the nipple overlap other tissues.

The image manipulation method presents better mean numerical results than the baseline on the PLANMED and IMS datasets, as shown in Tables 6.2 and 6.3. However, in the specific case of the nipple structure, we can see performance degradation. Thus, this is not the ideal method for these images when considering applications that require a good approximation of the nipple. In the case of the HOLOGIC dataset, which is the most distant from the GE dataset in terms of image similarity, we can see superior performance when compared to the baseline and to the style transfer method. This is more evident in the nipple structure, where the Dice, IoU, and Hausdorff metrics highlight this superiority. Further, as shown in Figure 6.4, the synthetic labels

Table 6.1: Numerical results on the GE dataset (test)

Metric	Method	Nipple	Pectoral	Fib. Tissue	Fat. Tissue	Mean
Precision	Baseline	0.8349	**0.9909**	0.9491	**0.8990**	**0.9185**
	Image manipulation	**0.8441**	0.9830	0.9557	0.8892	0.9180
	Style transfer	0.8212	0.9866	**0.9614**	0.8707	0.9100
	Combination	0.8409	0.9871	0.9443	0.8972	0.9174
Recall	Baseline	**0.8867**	0.9695	0.9543	0.8880	0.9246
	Image manipulation	0.8632	**0.9799**	0.9479	0.9068	0.9244
	Style transfer	0.8807	0.9773	0.9341	**0.9165**	**0.9272**
	Combination	0.8610	0.9741	**0.9546**	0.8890	0.9197
Accuracy	Baseline	**0.9995**	0.9971	0.9798	0.9757	0.9880
	Image manipulation	0.9994	0.9972	**0.9799**	**0.9760**	**0.9881**
	Style transfer	0.9994	0.9972	0.9783	0.9743	0.9873
	Combination	0.9994	**0.9972**	0.9786	0.9750	0.9876
Dice	Baseline	**0.8464**	0.9780	0.9496	0.8882	**0.9156**
	Image manipulation	0.8367	0.9799	**0.9497**	**0.8931**	0.9149
	Style transfer	0.8344	**0.9808**	0.9450	0.8877	0.9120
	Combination	0.8358	0.9789	0.9473	0.8878	0.9124
IoU	Baseline	**0.7488**	0.9608	0.9069	0.8078	**0.8561**
	Image manipulation	0.7344	0.9634	**0.9070**	**0.8150**	0.8550
	Style transfer	0.7316	**0.9644**	0.8988	0.8061	0.8502
	Combination	0.7333	0.9623	0.9024	0.8061	0.8510
Hausdorff	Baseline	**0.0019**	0.0038	0.0192	0.0141	0.0098
	Image manipulation	0.0020	0.0046	**0.0106**	**0.0134**	**0.0076**
	Style transfer	0.0020	**0.0036**	0.0110	0.0137	0.0076
	Combination	0.0020	0.0039	0.0110	0.0139	0.0077

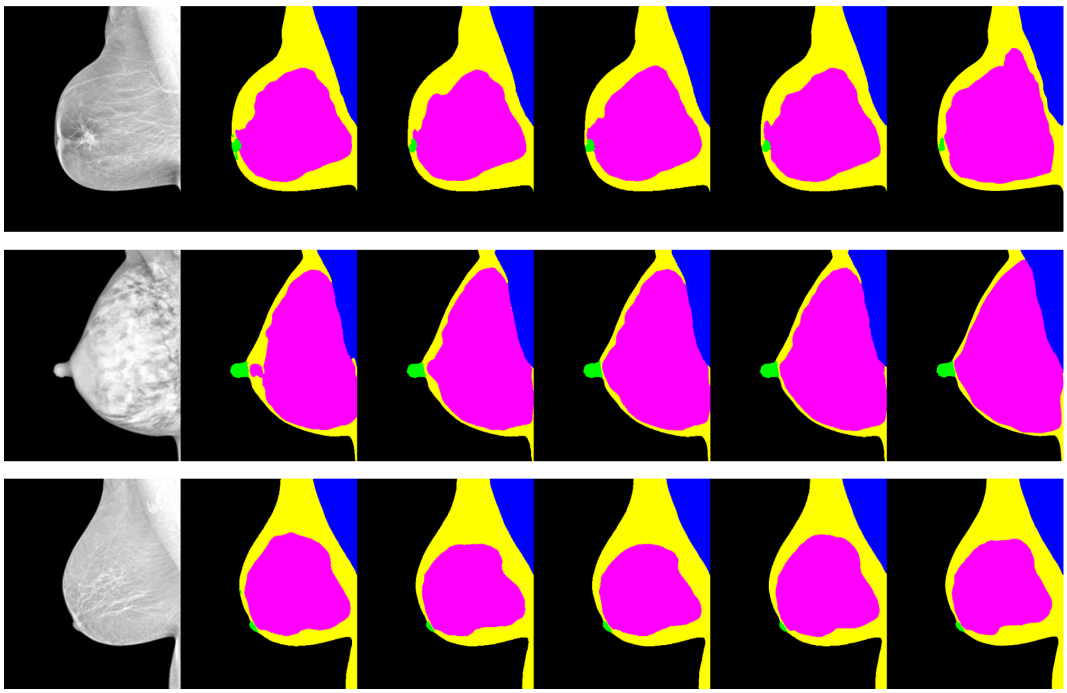

Figure 6.1: Visual results on the GE dataset (test). Each row represents a different case. First column: input image. Second column: baseline result. Third column: image manipulation result. Fourth column: style transfer result. Fifth column: image manipulation and style transfer combination result. Sixth column: ground-truth annotation.

Table 6.2: Numerical results on the IMS dataset

Metric	Method	Nipple	Pec-toral	Fib. Tissue	Fat. Tissue	Mean
Precision	Baseline	**0.9139**	**0.9971**	0.7207	0.9149	0.8867
	Image manipulation	0.8220	0.9853	0.8021	0.8941	0.8759
	Style transfer	0.8859	0.9779	**0.8768**	0.9159	**0.9142**
	Combination	0.8981	0.9853	0.8226	**0.9294**	0.9088
Recall	Baseline	0.7979	0.9191	**0.9855**	0.6411	0.8359
	Image manipulation	0.7871	0.9771	0.9837	0.7805	0.8821
	Style transfer	**0.8844**	**0.9824**	0.9537	**0.8646**	**0.9213**
	Combination	0.8290	0.9765	0.9815	0.8043	0.8978
Accuracy	Baseline	0.9995	0.9943	0.9692	0.9638	0.9817
	Image manipulation	0.9994	0.9975	0.9802	0.9738	0.9877
	Style transfer	0.9996	0.9974	**0.9860**	**0.9818**	**0.9912**
	Combination	**0.9996**	**0.9975**	0.9824	0.9781	0.9894
Dice	Baseline	0.8473	0.9540	0.8287	0.7506	0.8451
	Image manipulation	0.7925	**0.9806**	0.8807	0.8313	0.8713
	Style transfer	**0.8794**	0.9797	**0.9103**	**0.8876**	**0.9143**
	Combination	0.8545	0.9803	0.8916	0.8598	0.8965
IoU	Baseline	0.7401	0.9165	0.7120	0.6070	0.7439
	Image manipulation	0.6679	**0.9628**	0.7902	0.7150	0.7840
	Style transfer	**0.7904**	0.9608	**0.8378**	**0.8001**	**0.8473**
	Combination	0.7545	0.9621	0.8080	0.7573	0.8205
Hausdorff	Baseline	0.0123	0.0148	0.0346	0.0253	0.0218
	Image manipulation	**0.0025**	**0.0068**	0.0156	0.0210	**0.0115**
	Style transfer	0.0227	0.0095	**0.0125**	**0.0190**	0.0159
	Combination	0.0030	0.0078	0.0157	0.0209	0.0118

added during training are helpful in classifying the HOLOGIC image labels as background instead of breast structures.

The style transfer method presents superior performance on the IMS and PLANMED datasets, achieving the best IoU and Dice values for the nipple, fibroglandular tissue, and fatty tissue. For the pectoral muscle, the method presents results similar to the combination of the proposed methods. The Hausdorff distance for the nipple is not the minimum because this method tends to create noisy nipple regions in challenging cases, such as the example shown in the last row of Figure 6.3. These noisy regions do not represent an extensive area and can be easily removed in post-processing. While the style transfer method demonstrates superior performance on the IMS and PLANMED datasets, it exhibits inferior results compared to the image manipulation method on the HOLOGIC dataset. We can see significantly lower IoU values and higher Hausdorff distances for the nipple structure, potentially affecting applications where accurate nipple localization is critical. Figure 6.4 corroborates the latter, revealing misclassified regions like the image label, which was erroneously classified as nipple or fatty tissue instead of the background.

The combination method offers a balance between the two approaches. Upon examining the complete metrics across all four test datasets, we observe that this method consistently achieves the best or near-best numerical results. It leverages the strong generalization capabilities of the image manipulation

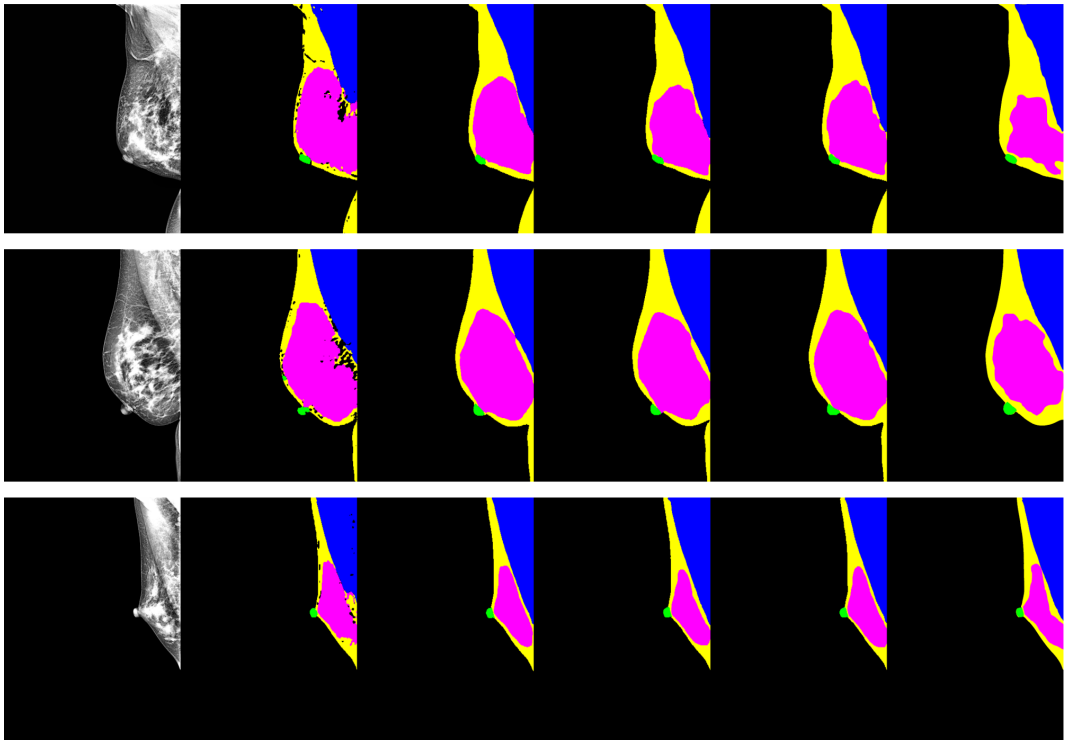

Figure 6.2: Visual results on the IMS dataset. Each row represents a different case. First column: input image. Second column: baseline result. Third column: image manipulation result. Fourth column: style transfer result. Fifth column: image manipulation and style transfer combination result. Sixth column: ground-truth annotation.

method on HOLOGIC images while benefiting from the effective generalization of the style transfer method on IMS and PLANMED images. Further, from Figures 6.1, 6.2, 6.3, and 6.4, we can notice consistent results with less noise, making this method the best choice for integration in the clinical practice.

6.4 Uncertainty analysis

Figure 6.5 shows the uncertainty maps of the three proposed methods and the baseline method on four images from the HOLOGIC dataset. Notice how the proposed strategies minimize the uncertainty regions compared to the baseline, concentrating them at the prediction boundaries. This indicates that the trained models are more confident and reliable when processing this kind of image. A similar behavior occurs when processing the IMS and PLANMED images.

Table 6.3: Numerical results on the PLANMED dataset

Metric	Method	Nipple	Pectoral	Fib. Tissue	Fat. Tissue	Mean
Precision	Baseline	**0.9463**	**0.9826**	0.7816	**0.9688**	0.9198
	Image manipulation	0.7940	0.9691	0.8525	0.9143	0.8825
	Style transfer	0.9204	0.9705	**0.9001**	0.9116	0.9256
	Combination	0.9291	0.9790	0.8661	0.9329	**0.9268**
Recall	Baseline	0.7352	0.9597	**0.9878**	0.6091	0.8229
	Image manipulation	0.5990	0.9901	0.9715	0.7863	0.8367
	Style transfer	**0.8119**	**0.9903**	0.9357	**0.8495**	**0.8969**
	Combination	0.7834	0.9866	0.9656	0.8027	0.8846
Accuracy	Baseline	0.9989	0.9946	0.9493	0.9422	0.9713
	Image manipulation	0.9982	0.9959	0.9668	0.9607	0.9804
	Style transfer	**0.9991**	0.9963	**0.9713**	**0.9674**	**0.9835**
	Combination	0.9990	**0.9965**	0.9689	0.9648	0.9823
Dice	Baseline	0.8132	0.9700	0.8693	0.7395	0.8480
	Image manipulation	0.6549	0.9789	0.9061	0.8420	0.8455
	Style transfer	**0.8506**	0.9798	**0.9151**	**0.8753**	**0.9052**
	Combination	0.8327	**0.9824**	0.9113	0.8592	0.8964
IoU	Baseline	0.7015	0.9432	0.7736	0.5962	0.7536
	Image manipulation	0.5198	0.9596	0.8308	0.7306	0.7602
	Style transfer	**0.7521**	0.9611	**0.8455**	**0.7809**	**0.8349**
	Combination	0.7325	**0.9659**	0.8389	0.7560	0.8233
Hausdorff	Baseline	0.0196	0.0170	0.0292	0.0214	0.0218
	Image manipulation	0.0059	0.0089	0.0169	0.0184	0.0126
	Style transfer	0.0073	0.0075	**0.0147**	**0.0172**	0.0117
	Combination	**0.0025**	**0.0071**	0.0167	0.0185	**0.0112**

Table 6.4: Numerical results on the HOLOGIC dataset

Metric	Method	Nipple	Pectoral	Fib. Tissue	Fat. Tissue	Mean
Precision	Baseline	0.4805	**0.9963**	0.8664	0.7193	0.7656
	Image manipulation	0.6903	0.9786	0.8986	0.8425	0.8525
	Style transfer	0.7126	0.9816	**0.9230**	0.8079	0.8563
	Combination	**0.7138**	0.9846	0.8984	**0.8488**	**0.8614**
Recall	Baseline	0.1545	0.7705	0.7212	0.5058	0.5380
	Image manipulation	0.8535	**0.9584**	0.9272	0.8306	0.8924
	Style transfer	0.7060	0.9477	0.9062	**0.8510**	0.8527
	Combination	**0.8620**	0.9545	**0.9308**	0.8258	**0.8933**
Accuracy	Baseline	0.9981	0.9785	0.9169	0.8994	0.9482
	Image manipulation	0.9990	0.9946	0.9628	0.9529	0.9773
	Style transfer	0.9988	0.9933	**0.9633**	0.9487	0.9760
	Combination	**0.9991**	**0.9947**	0.9631	**0.9532**	**0.9775**
Dice	Baseline	0.2245	0.8467	0.7826	0.5850	0.6097
	Image manipulation	0.7295	0.9673	0.9085	0.8283	0.8584
	Style transfer	0.6647	0.9621	**0.9107**	0.8212	0.8397
	Combination	**0.7501**	**0.9685**	0.9103	**0.8287**	**0.8644**
IoU	Baseline	0.1463	0.7677	0.6487	0.4192	0.4955
	Image manipulation	0.5901	0.9377	0.8349	0.7105	0.7683
	Style transfer	0.5246	0.9302	**0.8376**	0.7007	0.7483
	Combination	**0.6174**	**0.9398**	0.8375	**0.7111**	**0.7764**
Hausdorff	Baseline	0.0454	0.0298	0.0326	0.0314	0.0347
	Image manipulation	**0.0049**	0.0168	0.0180	**0.0211**	0.0152
	Style transfer	0.0398	0.0181	**0.0174**	0.0348	0.0275
	Combination	0.0050	**0.0163**	0.0175	0.0213	**0.0150**

6.5 Screen-film mammography results

The presented approaches focus on digital mammography images, which represent the prevailing technology in contemporary practice. However, screen-

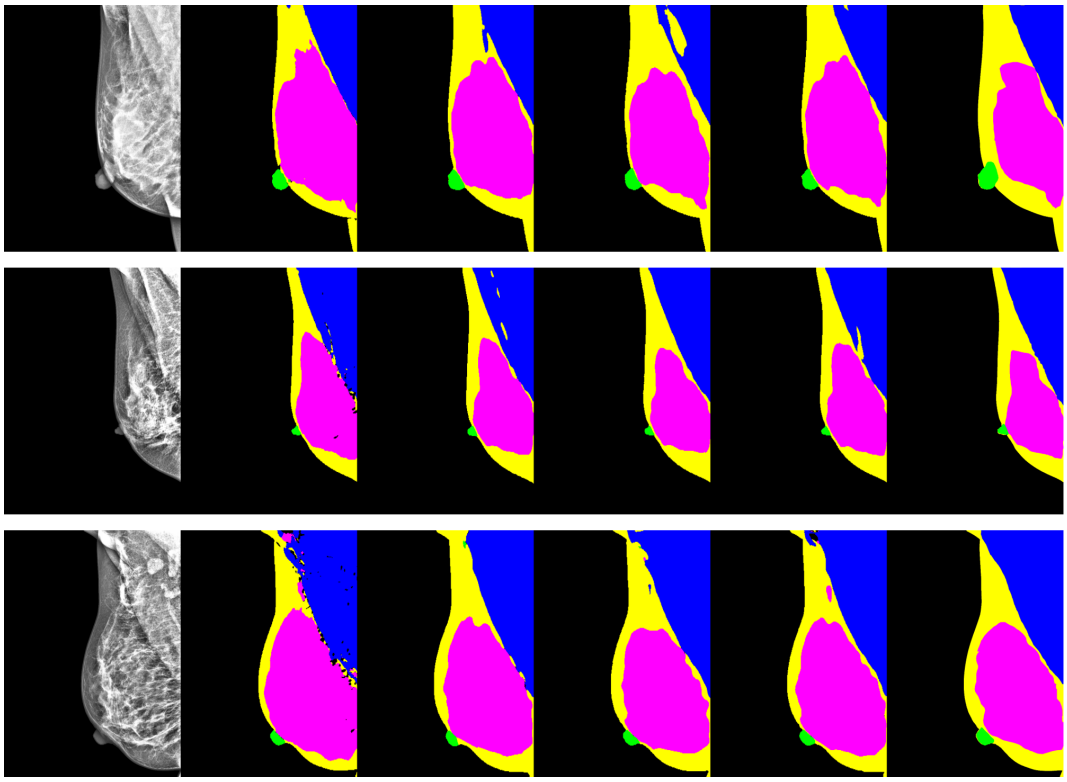

Figure 6.3: Visual results on the PLANMED dataset. Each row represents a different case. First column: input image. Second column: baseline result. Third column: image manipulation result. Fourth column: style transfer result. Fifth column: image manipulation and style transfer combination result. Sixth column: ground-truth annotation.

film mammography remains in use across numerous medical centers, with extensive repositories established around this technology. The screen-film mammography images exhibit significant differences compared to digital mammography images, presenting challenges in their processing with the proposed models. Figure 6.6 presents the predictions of the baseline and the combination methods on screen-film mammography from the DDSM dataset [54]. We can see how both methods present limitations when processing these images. Nevertheless, the combination method seems to be more stable and robust in this data.

6.6 Comparison for pectoral muscle segmentation

In [1], a deep learning-based approach is introduced for segmenting the pectoral muscle and breast in mammography images. The authors utilize a diverse dataset comprising mammography images from various vendor equipment, coupled with an aggressive augmentation procedure, to enhance gen-

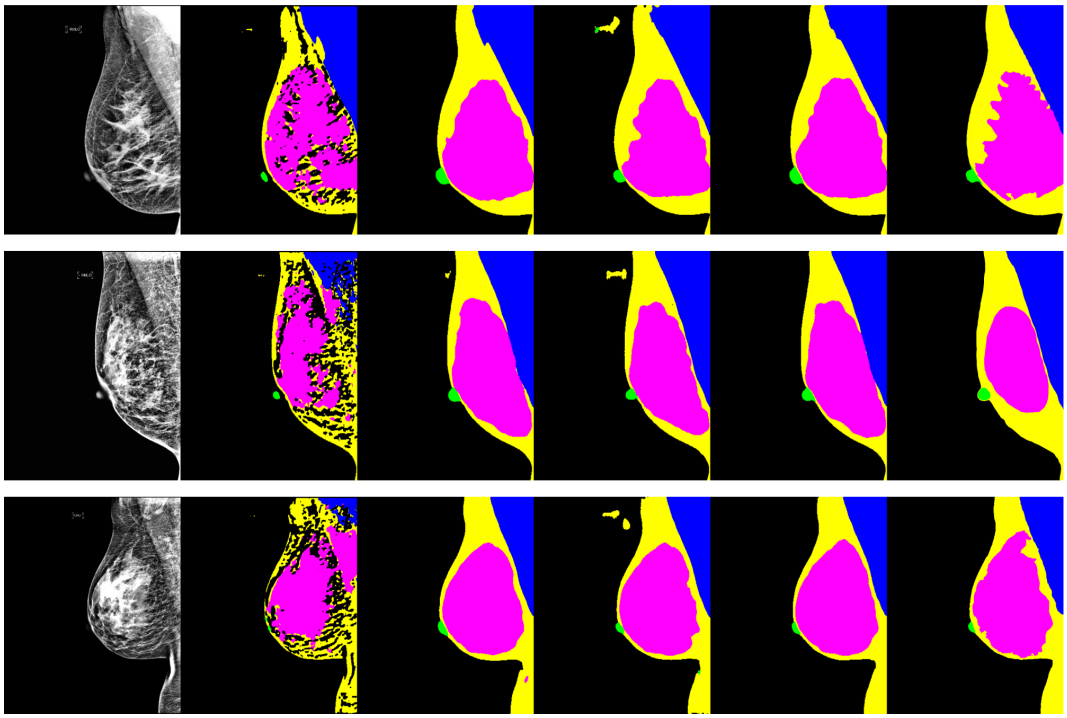

Figure 6.4: Visual results on the HOLOGIC dataset. Each row represents a different case. First column: input image. Second column: baseline result. Third column: image manipulation result. Fourth column: style transfer result. Fifth column: image manipulation and style transfer combination result. Sixth column: ground-truth annotation.

eralization performance. In contrast to this approach, we focus on a dataset exclusively comprising GE images and extend the segmentation task to include additional structures of interest. Nonetheless, a comparative analysis can be conducted for the segmentation of the pectoral muscle on an unseen dataset, such as the HOLOGIC dataset employed in our experiments. Table 6.5 presents the numerical results on the pectoral muscle segmentation task, comparing the performance of both [1] and our combination method. Notice how our method achieves significantly superior metric values, demonstrating that it is more robust and confident for this task. Further, Figure 6.7 shows some visual results, where we can see that our method is more consistent in predicting a single compact shape for the pectoral muscle, while the other presents noisy and incomplete predictions.

6.7 Extension to the CC view

Although the presented experiments focus on the MLO view, our methods can be easily extended for the CC view. To show this adaptability, we selected the CC view mammography segmentation dataset introduced in [6] for training

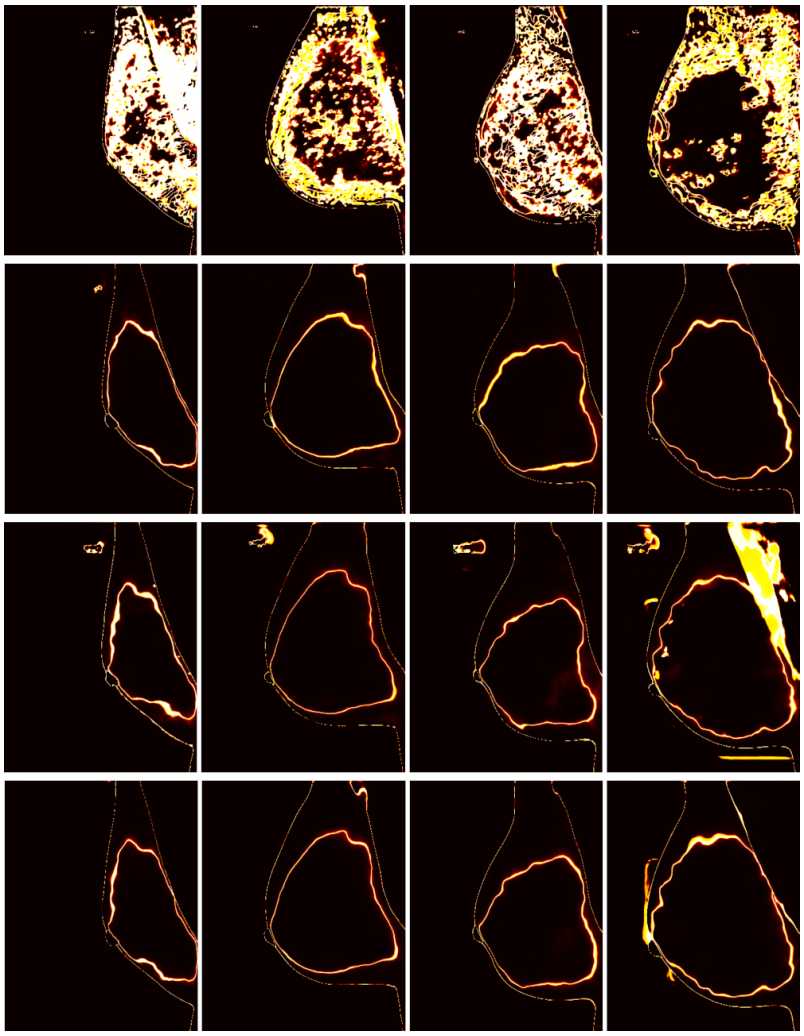

Figure 6.5: Uncertainty maps. Each column represents a different case. First row: baseline model. Second row: image manipulation. Third row: style transfer. Fourth row: image manipulation and style transfer combination.

Table 6.5: Numerical results for pectoral muscle segmentation on the HOLOGIC dataset

Method	**Dice**	**IoU**	**Haus-dorff**
[1]	0.8822	0.8145	0.0301
Ours	**0.9685**	**0.9398**	**0.0163**

and evaluation. This dataset consists of 5137 fully annotated GE images, where 3737, 943, and 457 images are considered for the training, validation, and test sets, respectively. Additionally, for the generalization evaluation, we consider a test set that consists a set of 34 fully-annotated HOLOGIC images. In both datasets, the same structures of interest presented in the previous sections are considered.

We train a baseline model on the GE images using the same settings

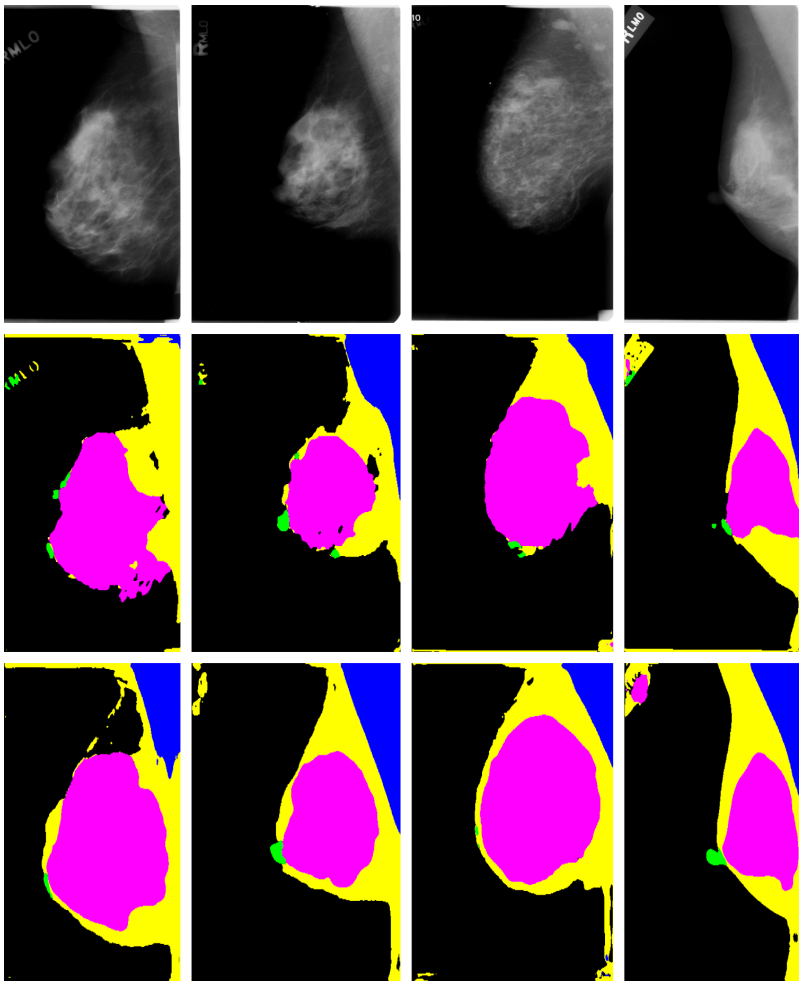

Figure 6.6: Results on screen-film mammography images from the DDSM dataset. Each column represents a different case. First row: input image. Second row: baseline. Third row: image manipulation and style transfer combination.

considered for the MLO view segmentation training. Then, leveraging our image manipulation method, we train another model using the same training settings as the CC view baseline model. Tables 6.6 and 6.7 present the numerical results on the GE and HOLOGIC test sets, respectively. Similarly to the behavior noticed for the MLO view, our method presents better results on the CC view HOLOGIC images while preserving the performance on the CC view GE images. Figure 6.8 shows some visual results on HOLOGIC images, confirming the superior performance of our method on the processing of CC view images, even in the challenging segmentation of the pectoral muscle [59].

6.8 Using transformer-based models

Our approaches to improving the generalization of our models have primarily utilized U-Nets. To enhance the scope of our comparisons, we will

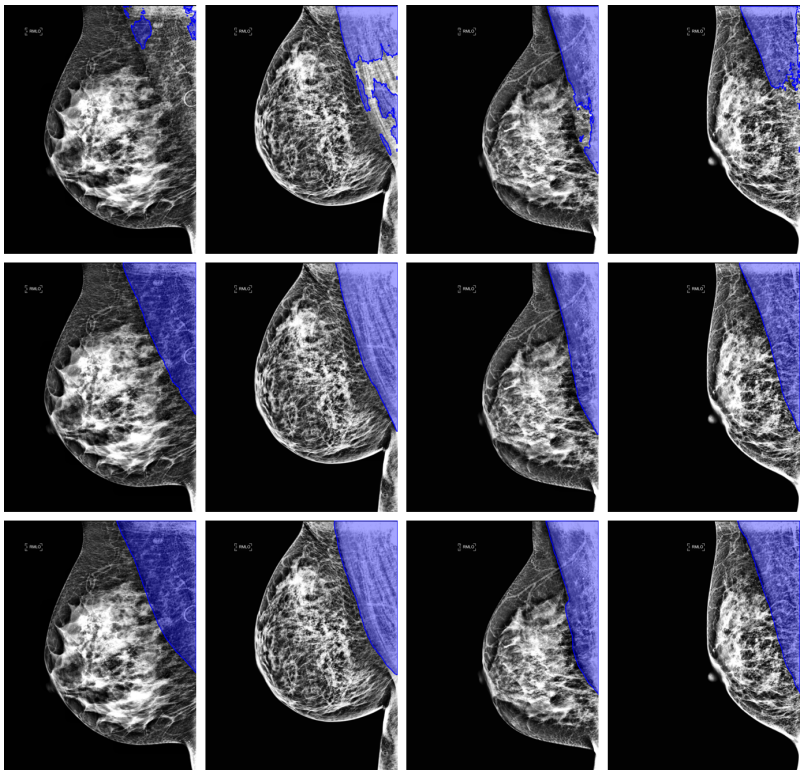

Figure 6.7: Pectoral segmentation comparison on the HOLOGIC dataset. Each column represents a different case. First row: results of the trained model introduced in [1]. Second row: results of the proposed combination method. Third row: ground-truth annotation.

Table 6.6: Numerical results on CC view GE images

Metric	Method	Nipple	Pectoral	Fib. Tissue	Fat. Tissue	Mean
Dice	Baseline	**0.8610**	**0.3998**	**0.9569**	**0.9103**	**0.7820**
	Image manipulation	0.8473	0.3870	0.9543	0.9083	0.7742
IoU	Baseline	**0.7731**	0.8289	**0.9194**	**0.8417**	**0.8408**
	Image manipulation	0.7567	**0.8456**	0.9151	0.8380	0.8389
Hausdorff	Baseline	**0.0003**	**0.0008**	**0.0022**	**0.0023**	**0.0015**
	Image manipulation	0.0003	0.0009	0.0022	0.0024	0.0016

Table 6.7: Numerical results on CC view HOLOGIC images

Metric	Method	Nipple	Pectoral	Fib. Tissue	Fat. Tissue	Mean
Dice	Baseline	0.1632	0.3059	0.8259	0.6608	0.4889
	Image manipulation	**0.6990**	**0.4136**	**0.8767**	**0.8108**	**0.7000**
IoU	Baseline	0.1011	0.5331	0.7070	0.4949	0.4590
	Image manipulation	**0.5613**	**0.7126**	**0.7865**	**0.6850**	**0.6863**
Hausdorff	Baseline	0.0500	0.0117	0.0262	0.0356	0.0341
	Image manipulation	**0.0032**	**0.0053**	**0.0250**	**0.0247**	**0.0167**

now explore segmentation models based on Transformer networks, diverging from the neural network architectures we used before.

Transformer networks have recently gained popularity, particularly in

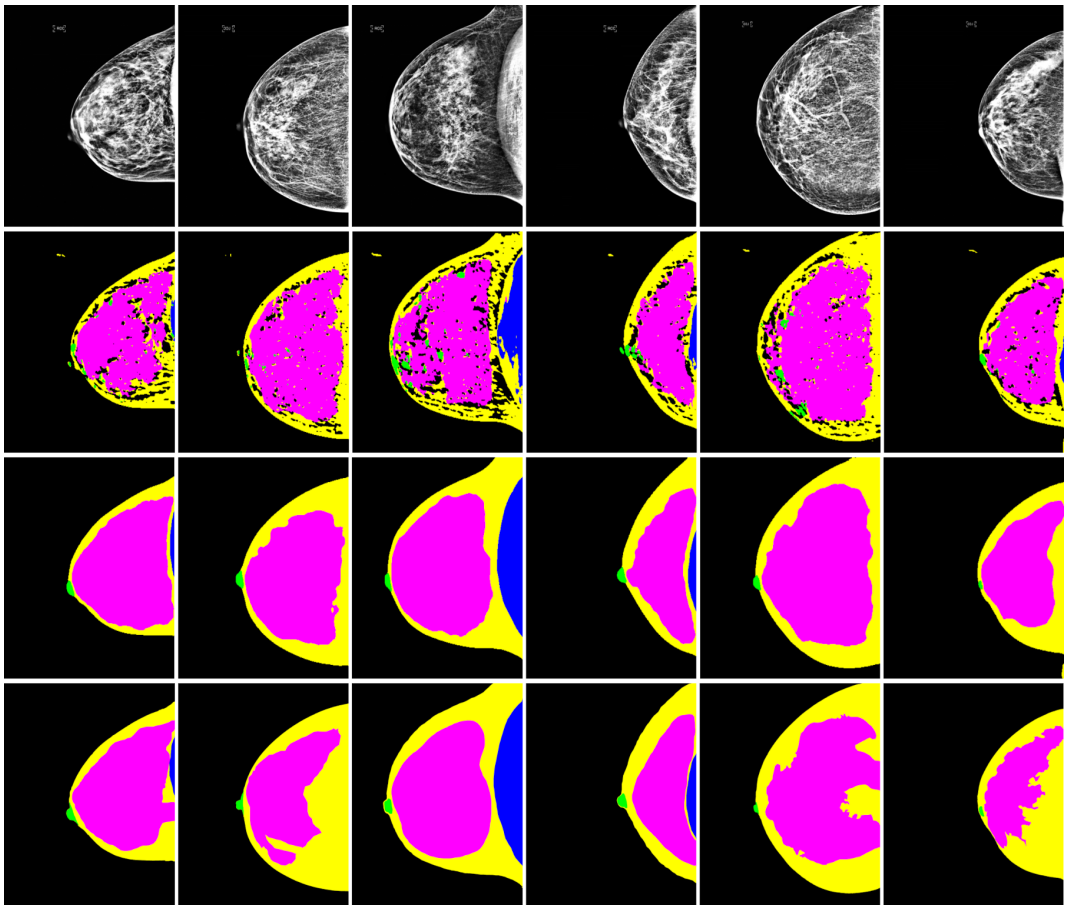

Figure 6.8: Visual results on CC view HOLOGIC images. Each column represents a different case. First row: input image. Second row: baseline result. Third row: image manipulation result. Fourth row: ground-truth annotation.

Natural Language Processing (NLP) tasks, such as text generation using large language models. Unlike traditional methods, Transformers leverage the self-attention mechanism to process data correlations in parallel, rather than sequentially. This approach allows the model to identify and focus on the most pertinent parts of the data over extended sequences.

While Transformers were originally developed for language-related applications, recent advancements have expanded their use to image processing. Vision Transformers (ViTs) and SegFormers are notable adaptations of Transformer architecture for visual data and semantic segmentation, respectively. In this section, we will compare the performance of our current approach (U-Net) against that of a SegFormer architecture.

To train the SegFormer model, we utilized a stylized dataset created using the style transfer technique. Additionally, we applied basic transformations such as cropping and rotation to augment the data. The model was trained on top of the pre-trained ADE20K dataset at a resolution of 512x512, provided

by NVIDIA on Hugging Face [60].

6.8.1 SegFormer tranning

This section aims to provide a succinct comparison between transformer-based models and our current approach.

To train the SegFormer model, we employed a stylized dataset created through style transfer techniques. Additionally, we applied basic data augmentation methods such as cropping and rotation. The model was fine-tuned on a pre-trained SegFormer base model, initially trained on the ADE20K dataset at a resolution of 512x512, as provided by NVIDIA on Hugging Face [60].

Our dataset was processed using the default feature extractor from the aforementioned dataset, specifically designed for 512x512 images. However, due to technical limitations in our development environment and potential high computational costs, we resized the images to 128x128.

Unlike our U-Net models, which were trained with the entire dataset at once, we trained the SegFormer model in batches using data from different styles. Given our technical constraints, we trained the model over 25 epochs for each batch. The results of these training sessions are presented in the following section.

6.8.2 SegFormer Results

Table 6.8: SegFormer evaluation over different metrics

Metric	Dataset	Nipple	Pectoral	Fib. Tissue	Fat. Tissue	Mean
Precision	GE	0.69	0.93	0.97	0.70	0.82
	HOLOGIC	0.37	0.93	0.93	0.74	0.74
	PLANMED	0.85	0.97	0.93	0.82	0.89
	IMS	0.77	0.95	0.95	0.80	0.87
Recall	GE	0.61	0.99	0.79	0.90	0.82
	HOLOGIC	0.15	0.97	0.81	0.83	0.69
	PLANMED	0.53	0.97	0.87	0.84	0.80
	IMS	0.63	0.97	0.81	0.83	0.81
Accuracy	GE	0.61	0.99	0.79	0.90	0.82
	HOLOGIC	0.15	0.97	0.81	0.82	0.64
	PLANMED	0.53	0.97	0.87	0.84	0.80
	IMS	0.63	0.97	0.81	0.83	0.81
Dice	GE	0.72	0.50	0.32	0.25	0.45
	HOLOGIC	0.32	0.49	0.32	0.25	0.35
	PLANMED	0.65	0.49	0.33	0.25	0.43
	IMS	0.73	0.49	0.32	0.25	0.45
IoU	GE	0.50	0.91	0.77	0.65	0.71
	HOLOGIC	0.12	0.90	0.76	0.63	0.60
	PLANMED	0.50	0.94	0.81	0.70	0.74
	IMS	0.51	0.92	0.77	0.69	0.72

As observed during the experiments across different metrics and datasets, the performance of SegFormers was inferior to the best methods using U-Net. However, it is important to note that this does not imply that SegFormers cannot perform better in our dataset. It is crucial to remember that the SegFormer model was trained for significantly fewer epochs, approximately 10 percent of the training performed on the U-Net models.

Overall, with more time and resources, further exploration of SegFormers may lead to the development of a superior solution and a more generalized model.

7
Conclusion

We address the challenge of segmenting landmark structures in mammography images, which is crucial for breast cancer assessment. Our approach considers data-centric strategies to enrich training data for deep learning-based segmentation. This involves augmenting training samples through annotation-guided image intensity manipulation and style transfer to improve generalization beyond conventional training methods.

Our findings demonstrate the effectiveness of the proposed methods in achieving improved generalization across various vendor equipment, even when considering training data from a single vendor. This approach avoids the need to generate new training images and manual annotations, thus reducing labor costs and saving time in clinical settings.

Although our evaluations are based on a limited number of different vendor equipment, the corresponding images represent the most diverse samples compared to those used for training, i.e. GE images. We expect the trained models to perform even better on images that closely resemble the GE images, such as those generated by Siemens or Fujifilm equipment.

We present visual results on screen-film mammography images, demonstrating a subtle enhancement achieved by the proposed method compared to the baseline. However, the predictions include noisy structures, requiring further post-processing operations to achieve a reliable segmentation. Further exploration of generalization across this domain and other image settings remains an open problem that we plan to address in future work.

Our assessment of CC view images demonstrates the applicability of the image intensity manipulation method to this domain. We expect that the style transfer method and the combination of both will exhibit similar efficacy. However, further investigation is needed, including an evaluation of pectoral muscle detection, which is challenging in the CC view.

Regarding the Segformer, we still have a long way to go. The preliminary results have shown promise, especially considering how little training was necessary to produce average results. Moving forward, we might want to train this model on more powerful machines for a longer period.

While we highlighted the importance of segmenting landmark structures

for assessing cancer risk and image acquisition adequacy, our experiments do not directly evaluate the efficacy of the proposed methods for these tasks. In future work, we aim to explore these applications and others using them.

Bibliography

[1] VERBOOM, S. D.; CABALLO, M.; PETERS, J.; GOMMERS, J.; VAN DEN OEVER, D.; BROEDERS, M. J.; TEUWEN, J. ; SECHOPOULOS, I.. **Deep learning-based breast region segmentation in raw and processed digital mammograms: generalization across views and vendors**. Journal of Medical Imaging, 11(1):014001–014001, 2024.

[2] OF NORTH AMERICA (RSNA), R. S.; OF RADIOLOGY (ACR), A. C.. **Mammography — radiologyinfo.org**. `https://www.radiologyinfo.org/en/info/mammo`. [Accessed 30-09-2023].

[3] **Mammograms — cancer.gov**. `https://www.cancer.gov/types/breast/mammograms-fact-sheet`. [Accessed 30-09-2023].

[4] MICHAEL, E.; MA, H.; LI, H.; KULWA, F. ; LI, J.. **Breast cancer segmentation methods: Current status and future potentials**. 2021:1–29.

[5] MA, J.; HE, Y.; LI, F.; HAN, L.; YOU, C. ; WANG, B.. **Segment anything in medical images**, 2023.

[6] SIERRA-FRANCO, C. A.; HURTADO, J.; THOMAZ, V. D. A.; DA CRUZ, L. C.; SILVA, S. V. ; RAPOSO, A. B.. **Towards automated semantic segmentation in mammography images**. arXiv preprint arXiv:2307.10296, 2023.

[7] MICHAEL, E.; MA, H.; LI, H.; KULWA, F. ; LI, J.. **Breast cancer segmentation methods: current status and future potentials**. BioMed Research International, 2021:1–29, 2021.

[8] MUSTRA, M.; GRGIC, M.. **Robust automatic breast and pectoral muscle segmentation from scanned mammograms**. Signal processing, 93(10):2817–2827, 2013.

[9] LIU, L.; LIU, Q. ; LU, W.. **Pectoral muscle detection in mammograms using local statistical features**. Journal of digital imaging, 27:633–641, 2014.

[10] OLIVER, A.; LLADÓ, X.; TORRENT, A. ; MARTÍ, J.. **One-shot segmentation of breast, pectoral muscle, and background in digitised mammograms**. In: 2014 IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP), p. 912–916. IEEE, 2014.

[11] SREEDEVI, S.; SHERLY, E.. **A novel approach for removal of pectoral muscles in digital mammogram**. Procedia Computer Science, 46:1724–1731, 2015.

[12] TAGHANAKI, S. A.; LIU, Y.; MILES, B. ; HAMARNEH, G.. **Geometry-based pectoral muscle segmentation from mlo mammogram views**. IEEE Transactions on Biomedical Engineering, 64(11):2662–2671, 2017.

[13] VIKHE, P.; THOOL, V.. **Detection and segmentation of pectoral muscle on mlo-view mammogram using enhancement filter**. Journal of medical systems, 41:1–13, 2017.

[14] RAMPUN, A.; MORROW, P. J.; SCOTNEY, B. W. ; WINDER, J.. **Fully automated breast boundary and pectoral muscle segmentation in mammograms**. Artificial intelligence in medicine, 79:28–41, 2017.

[15] HAZARIKA, M.; MAHANTA, L. B.. **A novel region growing based method to remove pectoral muscle from mlo mammogram images**. In: ADVANCES IN ELECTRONICS, COMMUNICATION AND COMPUTING: ETAEERE-2016, p. 307–316. Springer, 2018.

[16] TOZ, G.; ERDOGMUS, P.. **A single sided edge marking method for detecting pectoral muscle in digital mammograms**. Engineering, Technology and Applied Science Research, 8(1):2367–2373, 2018.

[17] AHMED, L.; IQBAL, M. M.; ALDABBAS, H.; KHALID, S.; SALEEM, Y. ; SAEED, S.. **Images data practices for semantic segmentation of breast cancer using deep neural network**. Journal of Ambient Intelligence and Humanized Computing, p. 1–17, 2020.

[18] DIVYASHREE, B.; AMARNATH, R.; NAVEEN, M. ; KUMAR, H.. **Segmentation of pectoral muscle in mammograms using granular computing**. Journal of Information Technology Research (JITR), 15(1):1–14, 2022.

[19] RAMPUN, A.; LÓPEZ-LINARES, K.; MORROW, P. J.; SCOTNEY, B. W.; WANG, H.; OCAÑA, I. G.; MACLAIR, G.; ZWIGGELAAR, R.; BALLESTER,

M. A. G. ; MACÍA, I.. **Breast pectoral muscle segmentation in mammograms using a modified holistically-nested edge detection network**. Medical image analysis, 57:1–17, 2019.

[20] SOLEIMANI, H.; MICHAILOVICH, O. V.. **On segmentation of pectoral muscle in digital mammograms by means of deep learning**. IEEE Access, 8:204173–204182, 2020.

[21] ALI, M. J.; RAZA, B.; SHAHID, A. R.; MAHMOOD, F.; YOUSUF, M. A.; DAR, A. H. ; IQBAL, U.. **Enhancing breast pectoral muscle segmentation performance by using skip connections in fully convolutional network**. International Journal of Imaging Systems and Technology, 30(4):1108–1118, 2020.

[22] GUO, Y.; ZHAO, W.; LI, S.; ZHANG, Y. ; LU, Y.. **Automatic segmentation of the pectoral muscle based on boundary identification and shape prediction**. Physics in Medicine & Biology, 65(4):045016, 2020.

[23] RUBIO, Y.; MONTIEL, O.. **Multicriteria evaluation of deep neural networks for semantic segmentation of mammographies**. Axioms, 10(3):180, 2021.

[24] YIN, F.-F.; GIGER, M. L.; DOI, K.; VYBORNY, C. J. ; SCHMIDT, R. A.. **Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique**. Medical Physics, 21(3):445–452, 1994.

[25] MÉNDEZ, A. J.; TAHOCES, P. G.; LADO, M. J.; SOUTO, M.; CORREA, J. ; VIDAL, J. J.. **Automatic detection of breast border and nipple in digital mammograms**. Computer methods and programs in biomedicine, 49(3):253–262, 1996.

[26] CHANDRASEKHAR, R.; ATTIKIOUZEL, Y.. **A simple method for automatically locating the nipple on mammograms**. IEEE transactions on medical imaging, 16(5):483–494, 1997.

[27] MUSTRA, M.; BOZEK, J. ; GRGIC, M.. **Nipple detection in craniocaudal digital mammograms**. In: 2009 INTERNATIONAL SYMPOSIUM ELMAR, p. 15–18. IEEE, 2009.

[28] ZHOU, C.; CHAN, H.-P.; PARAMAGUL, C.; ROUBIDOUX, M. A.; SAHINER, B.; HADJIISKI, L. M. ; PETRICK, N.. **Computerized nipple identification for multiple image analysis in computer-aided**

diagnosis: Computerized nipple identification on mammograms. Medical Physics, 31(10):2871–2882, 2004.

[29] KINOSHITA, S. K.; AZEVEDO-MARQUES, P. M.; PEREIRA, R. R.; RODRIGUES, J. A. H. ; RANGAYYAN, R. M.. **Radon-domain detection of the nipple and the pectoral muscle in mammograms**. Journal of digital imaging, 21:37–49, 2008.

[30] CASTI, P.; MENCATTINI, A.; SALMERI, M.; ANCONA, A.; MANGIERI, F. F.; PEPE, M. L. ; RANGAYYAN, R. M.. **Automatic detection of the nipple in screen-film and full-field digital mammograms using a novel hessian-based method**. Journal of digital imaging, 26:948–957, 2013.

[31] JIANG, J.; ZHANG, Y.; LU, Y.; GUO, Y. ; CHEN, H.. **A radiomic feature–based nipple detection algorithm on digital mammography**. Medical physics, 46(10):4381–4391, 2019.

[32] LIN, Y.; LI, M.; CHEN, S.; YU, L. ; MA, F.. **Nipple detection in mammogram using a new convolutional neural network architecture**. In: 2019 12TH INTERNATIONAL CONGRESS ON IMAGE AND SIGNAL PROCESSING, BIOMEDICAL ENGINEERING AND INFORMATICS (CISP-BMEI), p. 1–6. IEEE, 2019.

[33] HE, W.; JUETTE, A.; DENTON, E. R.; OLIVER, A.; MARTÍ, R.; ZWIGGELAAR, R. ; OTHERS. **A review on automatic mammographic density and parenchymal segmentation**. International journal of breast cancer, 2015, 2015.

[34] MATSUBARA, T.; YAMAZAKI, D.; KATO, M.; HARA, T.; FUJITA, H.; IWASE, T. ; ENDO, T.. **An automated classification scheme for mammograms based on amount and distribution of fibroglandular breast tissue density**. In: INTERNATIONAL CONGRESS SERIES, volumen 1230, p. 545–552. Elsevier, 2001.

[35] EL-ZAART, A.. **Expectation–maximization technique for fibroglandular discs detection in mammography images**. Computers in Biology and Medicine, 40(4):392–401, 2010.

[36] HIGHNAM, R.; BRADY, S. M.; YAFFE, M. J.; KARSSEMEIJER, N. ; HARVEY, J.. **Robust breast composition measurement-volpara tm**. In: DIGITAL MAMMOGRAPHY: 10TH INTERNATIONAL WORKSHOP,

IWDM 2010, GIRONA, CATALONIA, SPAIN, JUNE 16-18, 2010. PROCEEDINGS 10, p. 342–349. Springer, 2010.

[37] TORRES, G. F.; SASSI, A.; ARPONEN, O.; HOLLI-HELENIUS, K.; LÄÄPERI, A.-L.; RINTA-KIIKKA, I.; KÄMÄRÄINEN, J. ; PERTUZ, S.. **Morphological area gradient: System-independent dense tissue segmentation in mammography images**. In: 2019 41ST ANNUAL INTERNATIONAL CONFERENCE OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY (EMBC), p. 4855–4858. IEEE, 2019.

[38] KELLER, B. M.; NATHAN, D. L.; WANG, Y.; ZHENG, Y.; GEE, J. C.; CONANT, E. F. ; KONTOS, D.. **Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation**. Medical physics, 39(8):4903–4917, 2012.

[39] KELLER, B. M.; CHEN, J.; DAYE, D.; CONANT, E. F. ; KONTOS, D.. **Preliminary evaluation of the publicly available laboratory for breast radiodensity assessment (libra) software tool: comparison of fully automated area and volumetric density measures in a case–control study with digital mammography**. Breast cancer research, 17:1–17, 2015.

[40] SAFFARI, N.; RASHWAN, H. A.; ABDEL-NASSER, M.; KUMAR SINGH, V.; ARENAS, M.; MANGINA, E.; HERRERA, B. ; PUIG, D.. **Fully automated breast density segmentation and classification using deep learning**. Diagnostics, 10(11):988, 2020.

[41] LARROZA, A.; PÉREZ-BENITO, F. J.; PEREZ-CORTES, J.-C.; ROMÁN, M.; POLLÁN, M.; PÉREZ-GÓMEZ, B.; SALAS-TREJO, D.; CASALS, M. ; LLOBET, R.. **Breast dense tissue segmentation with noisy labels: A hybrid threshold-based and mask-based approach**. Diagnostics, 12(8):1822, 2022.

[42] HU, J.; LIU, Z. ; WANG, Q.. **Breast density segmentation in mammograms based on dual attention mechanism**. In: PROCEEDINGS OF THE 3RD INTERNATIONAL SYMPOSIUM ON ARTIFICIAL INTELLIGENCE FOR MEDICINE SCIENCES, p. 430–435, 2022.

[43] TIRYAKI, V.; KAPLANOĞLU, V.. **Deep learning-based multi-label tissue segmentation and density assessment from mammograms**. IRBM, 43(6):538–548, 2022.

[44] DUBROVINA, A.; KISILEV, P.; GINSBURG, B.; HASHOUL, S. ; KIMMEL, R.. **Computational mammography using deep neural networks**. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 6(3):243–247, 2018.

[45] BOU, A.. **Deep learning models for semantic segmentation of mammography screenings**, 2019.

[46] GARCEA, F.; SERRA, A.; LAMBERTI, F. ; MORRA, L.. **Data augmentation for medical imaging: A systematic literature review**. Computers in Biology and Medicine, 152:106391, 2023.

[47] ZHANG, H.; CISSE, M.; DAUPHIN, Y. N. ; LOPEZ-PAZ, D.. **mixup: Beyond empirical risk minimization**, 2018.

[48] OSUALA, R.; KUSHIBAR, K.; GARRUCHO, L.; LINARDOS, A.; SZAFRANOWSKA, Z.; KLEIN, S.; GLOCKER, B.; DIAZ, O. ; LEKADIR, K.. **A review of generative adversarial networks in cancer imaging: New applications, new solutions**. arXiv preprint arXiv:2107.09543, p. 1–64, 2021.

[49] GATYS, L. A.; ECKER, A. S. ; BETHGE, M.. **A neural algorithm of artistic style**. CoRR, abs/1508.06576, 2015.

[50] ZHENG, X.; CHALASANI, T.; GHOSAL, K.; LUTZ, S. ; SMOLIC, A.. **STaDA: Style transfer as data augmentation**. In: PROCEEDINGS OF THE 14TH INTERNATIONAL JOINT CONFERENCE ON COMPUTER VISION, IMAGING AND COMPUTER GRAPHICS THEORY AND APPLICATIONS. SCITEPRESS - Science and Technology Publications, 2019.

[51] WEI, Y.; LI, C.; LI, H. ; ZHANG, Z.. **Image data augmentation method based on style transfer**. In: PROCEEDINGS OF THE 2022 INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION AND INTELLIGENT SYSTEMS, PRIS '22, p. 1–7, New York, NY, USA, 2022. Association for Computing Machinery.

[52] NGUYEN, H. T.; NGUYEN, H. Q.; PHAM, H. H.; LAM, K.; LE, L. T.; DAO, M. ; VU, V.. **Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography**. Scientific Data, 10(1):277, 2023.

[53] ZUIDERVELD, K.. **Contrast Limited Adaptive Histogram Equalization**, p. 474–485. Academic Press Professional, Inc., USA, 1994.

[54] HEATH, M.; BOWYER, K.; KOPANS, D.; KEGELMEYER JR, P.; MOORE, R.; CHANG, K. ; MUNISHKUMARAN, S.. **Current status of the digital database for screening mammography**. In: DIGITAL MAMMOGRAPHY: NIJMEGEN, 1998, p. 457–460. Springer, 1998.

[55] SAHIN, Ö.; SAHIN, Ö.. **Introduction to apple ml tools**. Develop Intelligent iOS Apps with Swift: Understand Texts, Classify Sentiments, and Autodetect Answers in Text Using NLP, p. 17–39, 2021.

[56] **Classification: Precision and recall | machine learning**, Oct. 2023.

[57] TAHA, A. A.; HANBURY, A.. **Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool**. BMC Med Imaging, 15:29, Aug 2015.

[58] WANG, G.; LI, W.; AERTSEN, M.; DEPREST, J.; OURSELIN, S. ; VERCAUTEREN, T.. **Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks**. Neurocomputing, 338:34–45, 2019.

[59] SILVA, S. V.; SIERRA-FRANCO, C. A.; HURTADO, J.; DA CRUZ, L. C.; THOMAZ, V. D. A.; SILVA-CALPA, G. F. M. ; RAPOSO, A. B.. **A data-centric approach for pectoral muscle deep learning segmentation enhancements in mammography images**. In: INTERNATIONAL SYMPOSIUM ON VISUAL COMPUTING, p. 56–67. Springer, 2023.

[60] XIE, E.; WANG, W.; YU, Z.; ANANDKUMAR, A.; ALVAREZ, J. M. ; LUO, P.. **Segformer: Simple and efficient design for semantic segmentation with transformers**. Advances in neural information processing systems, 34:12077–12090, 2021.