Università degli Studi di Modena e Reggio Emilia

Department of Engineering Enzo Ferrari

Master's Degree in Artificial Intelligence Engineering

Enhancing Testicular Ultrasound Image Classification Through Synthetic Data and Pretraining Strategies

---

Supervisor
Prof. Costantino Grana

Candidate
Nicola Morelli

Co-supervisors
Prof. Federico Bolelli
Dott. Kevin Marchesini

Academic Year 2024-2025

Abstract

Enhancing Testicular Ultrasound Image Classification Through Synthetic Data and Pretraining Strategies

Male infertility is a widespread health concern, with testicular ultrasound imaging playing a key role in its assessment. In particular, parenchymal inhomogeneity has been proposed as a biomarker, but its reliable evaluation is challenged by subjective interpretation, artifacts, and the limited availability of annotated datasets. This thesis addresses these challenges by investigating strategies that combine pretraining and synthetic data generation to enhance automated classification of testicular ultrasound images.

A ResNet-18 backbone is employed, and both supervised and self-supervised pretraining strategies are evaluated to improve feature representation in data-scarce conditions. To reduce the impact of noisy labels, a heuristic filtering method is proposed, identifying and correcting mislabeled samples. Furthermore, synthetic ultrasound data are generated using diffusion models, followed by a filtering procedure to ensure fidelity and clinical relevance.

Experimental results demonstrate that pretraining consistently improves classification performance compared to training from scratch, while synthetic data can effectively support pretraining, partially overcoming data scarcity and privacy limitations. These findings highlight the potential of integrating synthetic data and pretraining strategies to build robust and reliable diagnostic tools, ultimately contributing to the advancement of automated ultrasound analysis for male infertility.

Keywords: Ultrasound, Medical Imaging, Synthetic, Diffusion Models.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to Prof. Costantino Grana and Prof. Federico Bolelli for their invaluable guidance and for giving me the opportunity to join the AImageLab research group. Being part of this team has been a truly formative experience, both academically and personally. Within the group, I had the pleasure of meeting many extraordinary people, from whom I have learned a great deal and with whom I have shared many enjoyable moments. Among them, I want to especially thank Kevin Marchesini and Luca Lumetti, from whom I have learned so much. I am also grateful to all the other members of the lab for their warm welcome, their company, and all the moments of laughter, collaboration, and play we shared.

I would then like to thank my dearest friends, Diego, Tommaso, Pule, and Giuseppe, some of the oldest companions I have, who have been by my side through countless joyful moments, nights out, and shared experiences over the years. Their presence has always reminded me not to take life too seriously and to face every challenge with lightness and a smile. A heartfelt thank you also goes to my university friends, Omar, Pietro, Davide, Matteo, and Marta, with whom I have shared this long academic journey. Together we faced lessons, exams, and challenges, but also shared meals, laughs, and endless conversations that made this path so much more meaningful.

A very special thank you goes to Giulia, who, despite the short time we've been together, has already shown immense patience, understanding, and support. Thank you for bearing with me even when my cynicism might have hurt you. Your kindness and steadiness have made this period lighter and more peaceful.

Last but certainly not least, I wish to thank my family, who have always been my foundation. To my parents, Fiorenza and Luigi, thank you for supporting me in every possible way since the very beginning. My gratitude also goes to my brothers, Lorenzo and Francesco, who help me feel a little less lonely and more understood in our home.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Problem of Male Infertility

Infertility is a global health issue affecting many couples, and male factors are a major contributor. According to the World Health Organization[1] [3], approximately one in six people of reproductive age experience infertility in their lifetime. In about one-third of infertile couples, male reproductive problems are the primary cause; in another one-third, they contribute jointly with female factors.

The causes of male infertility are varied. Some of the major contributing factors include [4]:

- Problems in sperm production (spermatogenesis): including low count (oligospermia), absent sperm (azoospermia), and defects in sperm morphology or motility.
- Hormonal imbalances: disorders of the endocrine system, such as low testosterone, pituitary or hypothalamic dysfunction, which interfere with the normal stimulation of the testes.
- Genetic causes: including chromosomal abnormalities (e.g., Klinefelter's syndrome), mutations, and congenital defects.
- Anatomical or structural issues: blockages in the sperm transport tract, undescended testicles (cryptorchidism), varicocele (enlarged veins around the testicle), or injury.

[1] https://www.who.int/

- Infections, inflammation, injury: prior infections (including sexually transmitted infections), testicular inflammation or orchitis, or trauma can damage spermatogenic tissue.
- Lifestyle, environmental, and other risk factors: obesity, exposure to toxins, heat, chemical pollutants, age, use of certain medications or steroids, and other health issues.

Male infertility not only affects the ability to conceive but has social, psychological and financial implications for couples. Because male infertility can be partially or sometimes fully treatable (depending on the cause), early detection and accurate diagnosis are critical. However, many causes remain idiopathic (unknown) even after standard evaluation.

1.2 Background, Problem Statement, and Motivation

In discussions with clinicians, the need emerged for tools that could assist in the early detection and prevention of male infertility and related disorders, supporting doctors in diagnostics and patient counselling. Given the available in-house data, a promising starting point was identified in the detection of tissue inhomogeneity in testicular ultrasound images, a feature proposed as a biomarker for male infertility.

Testicular ultrasound (TUS) is a key non-invasive imaging modality for assessing testicular structure and function, playing a crucial role in detecting tissue characteristics such as parenchymal inhomogeneity, an emerging biomarker for male infertility [5]. While medical imaging, including ultrasound, is essential for diagnostics due to its safety, accessibility, and real-time capabilities, its interpretation remains challenging.

A significant challenge arises from subjective image interpretation and complex tissue patterns in TUS, which hinder reliable and standardized assessment, thus highlighting a critical need for automated classification tools. Progress is severely hampered by the lack of large, publicly available datasets. Ethical and privacy concerns limit data sharing, leading to small, institution-specific datasets that constrain deep learning model development and generalization.

Ultrasound imaging presents unique difficulties, including artifacts, noise, low contrast, and operator dependency. Furthermore, limited annotated datasets (due to time-consuming, expert-dependent labeling) and heterogeneity across devices and operators complicate model training

and generalization. A particular challenge is the inherent noisiness in image-label pairing: clinicians rely on real-time video evaluation, but only static screenshots are saved, which may not accurately reflect the tissue's homogeneity, introducing noise into the dataset. Given these obstacles, there is a critical need for automated classification tools to improve reliability and standardization in TUS assessment.

1.3 Addressing the Challenges and Key Contributions

To tackle the intertwined challenges of data scarcity and subjective interpretation in testicular ultrasound (TUS) analysis, this work proposes an integrated approach that leverages pretraining strategies, heuristic noise reduction, and diffusion-based synthetic data generation. These components work synergistically to enhance model robustness, improve generalization, and ensure data fidelity, collectively advancing the automation of TUS interpretation.

Our proposed approach involves the following key steps:

- **Evaluating pretraining strategies:** We systematically compare supervised and self-supervised pretraining methods for testicular inhomogeneity classification using a ResNet-18 [6] backbone. Preliminary experiments revealed that more complex architectures such as ResNet-50 and Vision Transformers [7] tend to overfit given the limited size of our Labeled Dataset (LD). Hence, we focus on architectures that balance representational power with generalization capability.
- **Heuristic label noise reduction:** Recognizing that subjective interpretation can introduce inconsistencies in labeling, we propose a heuristic filtering method that detects "suspicious" samples by tracking their loss trajectories during initial training. Samples identified as unreliable can either be discarded or relabeled, which empirically improves classification accuracy and label consistency.
- **Diffusion-based synthetic data generation:** To further mitigate data scarcity, we employ Denoising Diffusion Probabilistic Models (DDPMs) to synthesize realistic ultrasound images. These models outperform traditional GAN-based approaches in stability and fine-detail preservation, enabling the generation of high-quality, clinically relevant synthetic data that mirrors the distribution of the real LD.

- **Synthetic data filtering via precision-based selection:** To ensure the fidelity of generated images, we compute a "real data manifold" from feature embeddings of the real LD and retain only synthetic samples whose embeddings lie within this manifold. This filtering process reduces the initial synthetic dataset from approximately 20K to around 9K high-quality images, providing a diverse and reliable dataset for pretraining.

Through this cohesive pipeline, the study not only addresses the fundamental limitations of current TUS analysis but also establishes a reproducible framework for applying generative modeling and pretraining in data-limited medical imaging domains.

In summary, this work makes the following key contributions:

(i) A systematic evaluation of supervised and self-supervised pretraining strategies tailored to testicular ultrasound analysis;

(ii) A heuristic label noise reduction method that enhances annotation quality and model reliability;

(iii) The introduction of diffusion-based generative modeling for realistic synthetic TUS data, coupled with a precision-based filtering mechanism to ensure high fidelity.

1.4 Thesis Structure

The remainder of this thesis is structured as follows:

- **Chapter 2: Context Overview** will provide a comprehensive overview of existing literature on deep learning in ultrasound image analysis and generative models for synthetic medical image generation;
- **Chapter 3: Dataset Curation and Noisy Label Filtering** will detail the acquisition and preparation of the in-house dataset, including the Unlabeled Dataset (UD) and Labeled Dataset (LD), as well as the proposed heuristic filtering procedure for noisy labels;

- **Chapter 4: The Generated Dataset** will describe the generation of a synthetic dataset using Denoising Diffusion Probabilistic Models (DDPMs) and the filtering algorithm designed to ensure the quality of the generated images;
- **Chapter 5: Training Methodologies and Evaluation Metrics** will describe the semi-supervised pretraining strategies. It will also outline the neural network architectures, fine-tuning procedures, and evaluation protocols;
- **Chapter 6: Experiments and Results** will present the experimental setup, implementation details, and the results validating the effectiveness of the proposed pretraining and synthetic data integration strategies;
- **Chapter 7: Conclusion and Future Research Directions** will summarize the findings of this study, discuss its implications, and propose directions for future work, including the incorporation of dynamic ultrasound videos and label-conditioned synthetic image generation;
- **Chapter 8: Other Works — MICCAI 2025 UUSIC Challenge** will present an additional study conducted during the final stage of this thesis, which explores a complementary research direction beyond the main scope of the work.

Chapter 2

Context Overview

This chapter provides a comprehensive overview of the essential background for this thesis, beginning with an introduction to ultrasound imaging, followed by a discussion of its application in detecting testicular pathologies. It then delves into the current state of deep learning in medical image analysis, culminating in a detailed explanation of Denoising Diffusion Probabilistic Models (DDPMs) and their role in addressing data scarcity and quality challenges in this sensitive domain.

2.1 The Ultrasound Image Modality

Ultrasound (US) imaging is a key non-invasive tool in medical diagnostics, widely recognized for its safety, accessibility, and real-time capabilities [8]. Unlike other imaging modalities, US utilizes sound waves to create images, avoiding ionizing radiation and making it particularly suitable for repeated examinations and sensitive anatomical regions.

Principles of Ultrasound Imaging

Ultrasound imaging is based on the emission of high-frequency sound waves (typically 2–15 MHz) from a transducer placed on the patient's skin. These sound waves propagate through tissues and are partially reflected whenever they encounter boundaries with different acoustic impedances (e.g., between soft tissue and bone, or tissue and fluid). The transducer

then detects the returning echoes, and the system reconstructs these signals into a grayscale image where brightness corresponds to the amplitude of the reflected echoes. This process enables real-time visualization of anatomy and physiology, such as blood flow or organ motion. However, because ultrasound relies on the acoustic properties of tissues, factors like reflection, refraction, attenuation, and scattering introduce limitations and artifacts.

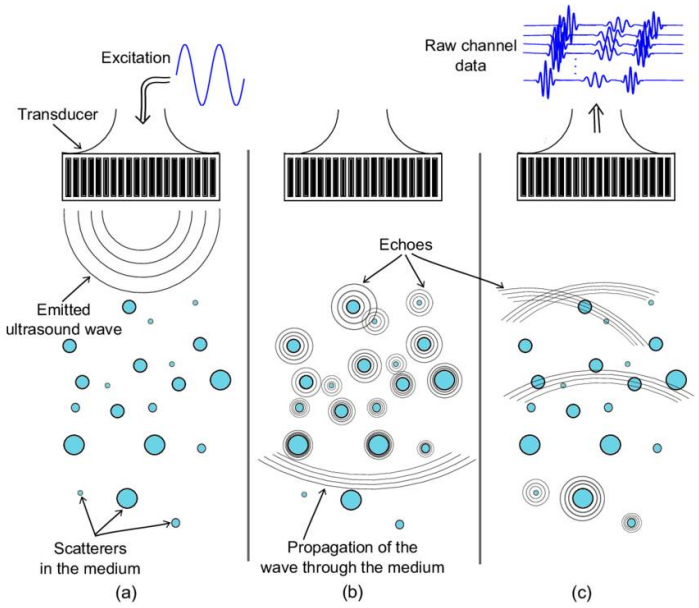


FIGURE 2.1: Basic principle of ultrasound imaging: the transducer emits high-frequency sound waves, receives reflected echoes from tissue boundaries, and reconstructs the measured signal into an image.

Ultrasound Artifacts

However, the interpretation of ultrasound images presents significant challenges, many of which are inherent to the underlying physics. Artifacts represent a particularly important source of difficulty, as they can alter healthy anatomy, mimic pathological findings, or in some cases provide useful diagnostic clues. The most common types of artifacts are:

- **Acoustic shadowing**: Occurs when highly reflective or attenuating structures (e.g., bone, calcifications, gallstones) absorb or reflect nearly all incident sound waves, leading to dark regions distal to the reflector. This may hinder visualization of deeper tissues, but can also be diagnostically useful (e.g., identifying gallstones). Examples of this phenomenon are shown in Fig. 2.2;

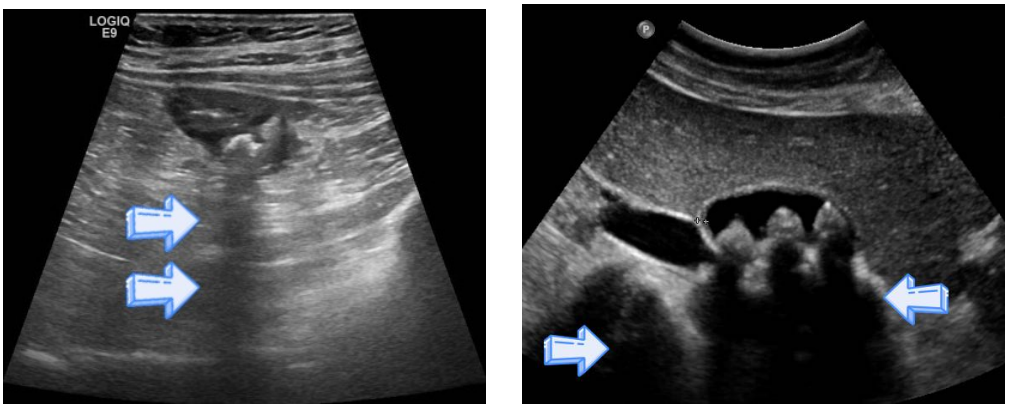


FIGURE 2.2: Examples of acoustic shadowing artifacts.

- **Acoustic enhancement**: Seen as zones of increased echogenicity beneath fluid-filled structures (e.g., cysts, bladder), since such media allow enhanced transmission of ultrasound waves. While sometimes misleading, enhancement can also aid in detecting underlying fluid. Representative examples are shown in Fig. 2.3;

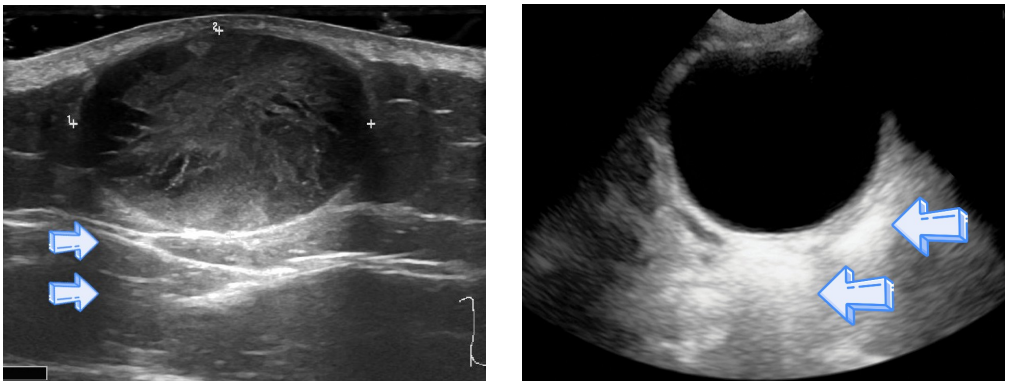


FIGURE 2.3: Examples of acoustic enhancement artifacts.

- **Reverberation artifact**: Results from repeated reflections between strong reflectors or between a reflector and the transducer. Appearing as multiple parallel echoes, these can mimic structures at increasing depths but can be recognized by their regular spacing and diminishing intensity. Illustrative cases are provided in Fig. 2.4;

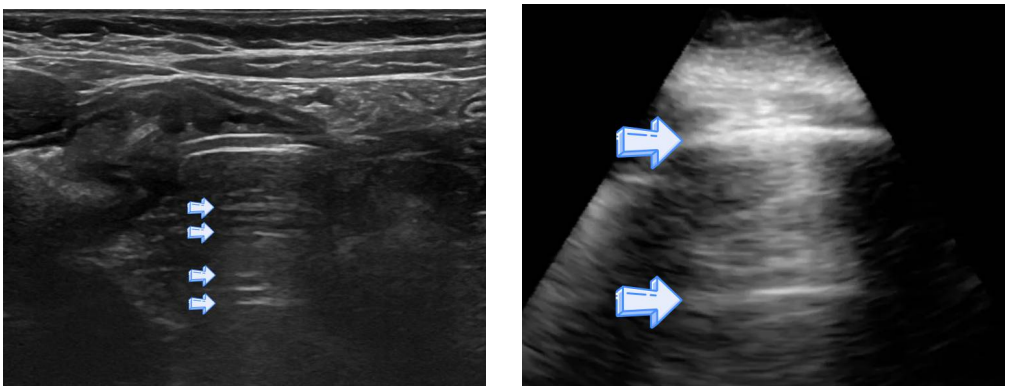


FIGURE 2.4: Examples of reverberation artifacts.

- **Mirror image artifact**: A special form of reverberation where a strong reflector (e.g., diaphragm) acts like a mirror, producing duplicated structures beyond the reflector. Such artifacts often disappear when the transducer angle is changed, distinguishing them from true anatomy (see Fig. 2.5);

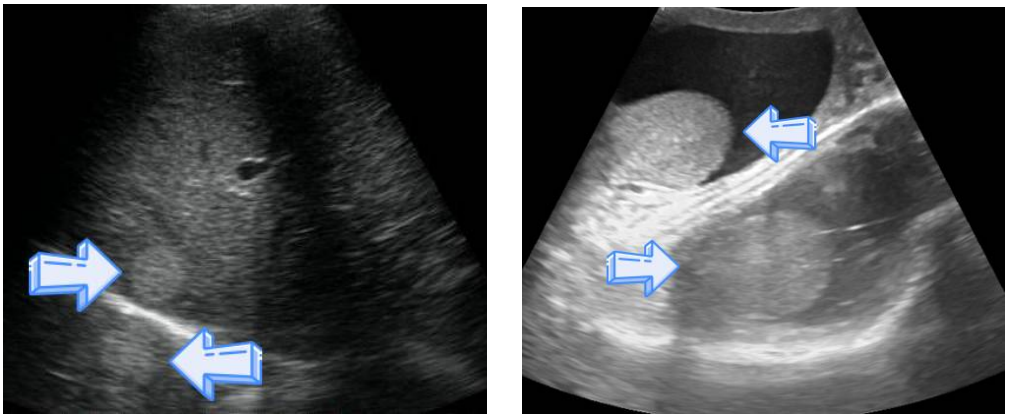


FIGURE 2.5: Examples of mirror image artifacts.

- **Speckle artifact**: Arises from the constructive and destructive interference of scattered ultrasound waves within heterogeneous tissues. This produces a grainy texture that can obscure fine anatomical details and reduce image contrast. While generally undesirable, speckle can also carry information about tissue microstructure and is sometimes exploited in quantitative imaging. Examples are shown in Fig. 2.6.

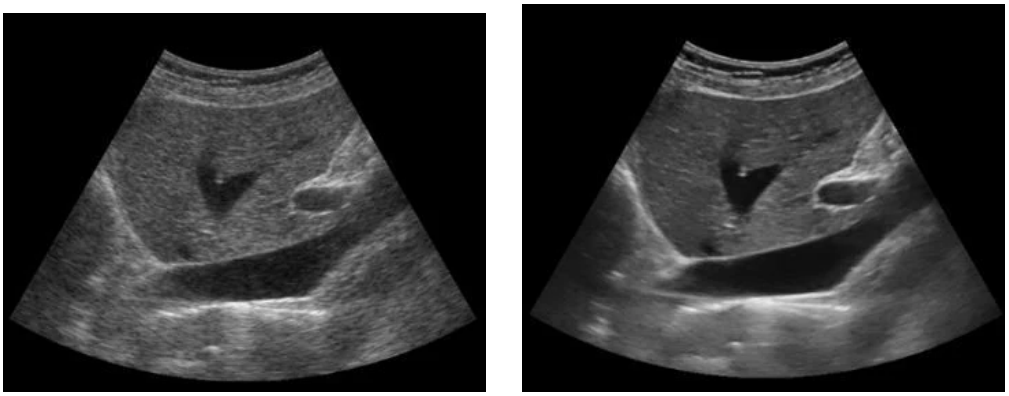


FIGURE 2.6: Examples of speckle artifacts.

These artifacts add to other challenges such as dependence on operator skills and variations across devices, making consistent and automated ultrasound analysis more difficult. For Testicular Ultrasound (TUS) in particular, the lack of annotated data, variability in acquisition conditions, and stringent privacy requirements further hinder the development of reliable deep learning models. Although artifacts may occasionally offer diagnostic value, they more often introduce unpredictability and interfere with standard image processing methods.

2.2 State of the Art in Deep Learning for Ultrasound Imaging

In the context of male reproductive health, Testicular Ultrasound (TUS) imaging is vital for evaluating male infertility. A specific and emerging biomarker for male infertility is testicular parenchymal inhomogeneity, which refers to variations in the tissue's texture and density. The accurate assessment of this biomarker is crucial for diagnosis and patient treatment.

Despite its importance, the reliable, standardized assessment of testicular inhomogeneity is hindered by subjective image interpretation and complex tissue patterns. Clinicians often rely on real-time video evaluation during examinations to assess these anatomical properties comprehensively. However, in routine clinical practice, only static screenshots are typically saved, which may not always accurately reflect the true homogeneity characteristics of the tissue. This discrepancy introduces inherent noisiness in the pairing of images and labels within datasets, posing a significant challenge for automated classification systems. This situation highlights a pressing need for automated classification tools that can provide objective and standardized assessments.

Deep Learning, particularly Convolutional Neural Networks (CNNs) [9], has demonstrated significant promise in automating US image analysis and extracting quantitative information [10], which is essential for overcoming subjective interpretation challenges. These advancements have been crucial in various medical diagnostic applications. However, a major barrier to progress in this field is the lack of large, publicly available ultrasound datasets. Ethical and privacy constraints significantly restrict data sharing, leading most deep learning models to be developed and tested on small, institution-specific datasets. Furthermore, in many clinical settings, medical imaging protocols do not routinely preserve digital versions of ultrasound scans, resulting in additional data loss. This scarcity of accessible and standardized data hampers robust model training and limits the generalizability of deep learning systems across different institutions and patient populations.

To address the pervasive issue of data scarcity, medical imaging research increasingly leverages two key strategies:

- **Pretraining:** Training on large, diverse datasets (either general or medical-specific) can significantly enhance feature extraction capabilities and improve the performance of deep learning models when fine-tuned on smaller target datasets;
- **Synthetic data generation:** Generative models were initially explored to augment datasets, enable cross-modality synthesis, and facilitate anonymization.

Early approaches to synthetic data generation primarily relied on Generative Adversarial Networks (GANs), which, despite their success, often exhibited unstable training dynamics and struggled to produce sufficiently diverse samples. In contrast, the advent of Denoising Diffusion Probabilistic Models (DDPMs) [11] has marked a significant advancement in generating high-quality and realistic medical images. Compared to GANs [12], DDPMs demonstrate superior performance in producing diverse and lifelike samples, particularly in imaging modalities such as MRI and CT. This progress provides an effective means to address data-sharing restrictions and the persistent issue of data scarcity in sensitive domains like TUS classification.

2.3 Introduction to DDPMs

Denoising Diffusion Probabilistic Models (DDPMs) [11] are powerful generative models that have demonstrated superior performance over GANs in terms of stability and detail modeling for high-quality medical image synthesis. They operate through a two-phase process to generate realistic images (as depicted in Figure 2.7):

1. **Forward Diffusion Phase:** A Markov chain gradually adds Gaussian noise to a clean image over multiple time steps. Each step introduces a small amount of noise, following a predefined variance schedule, until the original image is completely transformed into pure Gaussian noise;

2. **Reverse Denoising Phase:** A neural network, typically a U-Net architecture, is trained to reconstruct the original image by progressively removing the added noise. The U-Net learns to predict the noise component present in the image at each time step, effectively learning to reverse the diffusion process.

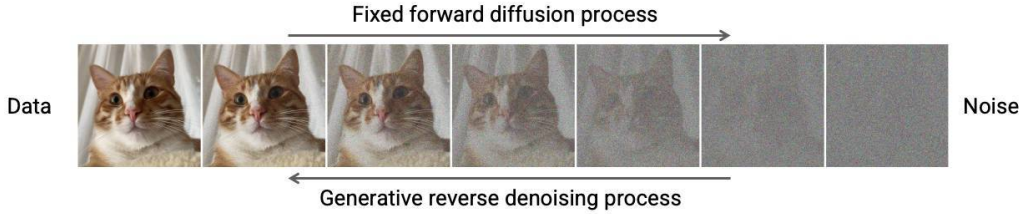


FIGURE 2.7: A conceptual overview of the two-phase process in a Denoising Diffusion Probabilistic Model (DDPM). The forward diffusion phase (top arrow) shows the gradual addition of Gaussian noise to a clean image over multiple time steps until it becomes pure noise. The generative reverse denoising phase (bottom arrow) illustrates how a trained neural network iteratively removes noise to reconstruct a clean image from pure noise [1].

During inference, the process starts with a sample of pure Gaussian noise. The trained DDPM then iteratively refines this noise through a series of denoising steps, progressively removing the estimated noise component. This iterative refinement continues until the final step, resulting in a synthetic image that closely resembles the distribution of real ultrasound images. Furthermore, Conditional DDPMs offer the capability for controlled image generation based on specific clinical attributes or segmentation maps, providing a powerful tool for targeted data synthesis.

In medical imaging applications, DDPMs can be trained on real datasets to generate synthetic samples that mirror clinical data distributions. To ensure the fidelity and practical utility of these synthetic images, filtering strategies are often applied to remove out-of-distribution samples and retain only those consistent with real data characteristics. This results in a refined synthetic dataset that can be used to augment training, improve data diversity, and address challenges of data scarcity while preserving clinical relevance. Such curated synthetic datasets are particularly valuable for pretraining in domains where annotated medical images are limited.

Chapter 3

Dataset Curation and Noisy Label Filtering

This chapter details the acquisition, characteristics, and preprocessing of the dataset used in this study, specifically focusing on the challenges of data scarcity and label noise inherent in testicular ultrasound imaging. It outlines how images were collected, the structure and distribution of the data, and the heuristic algorithm developed to filter noisy labels to enhance training quality.

3.1 Dataset Acquisition and Characteristics

To address the lack of publicly available testicular ultrasound image datasets, all experiments were conducted using an in-house dataset. This dataset was collected at the Antonio Nalin Center of the Baggiovara Hospital in Modena, Italy, utilizing two different ultrasound acquisition systems: *Esaote® MyLab25 Gold* and *Esaote® MyLab XPro80*.

The dataset comprises image pairs, with each pair containing static views of the same testicle captured from both transverse and sagittal planes. These pairs were then cropped to remove metadata and isolate single views as shown in 3.1, with each view being treated independently and inheriting the original label.

The overall dataset is structured into two main components:

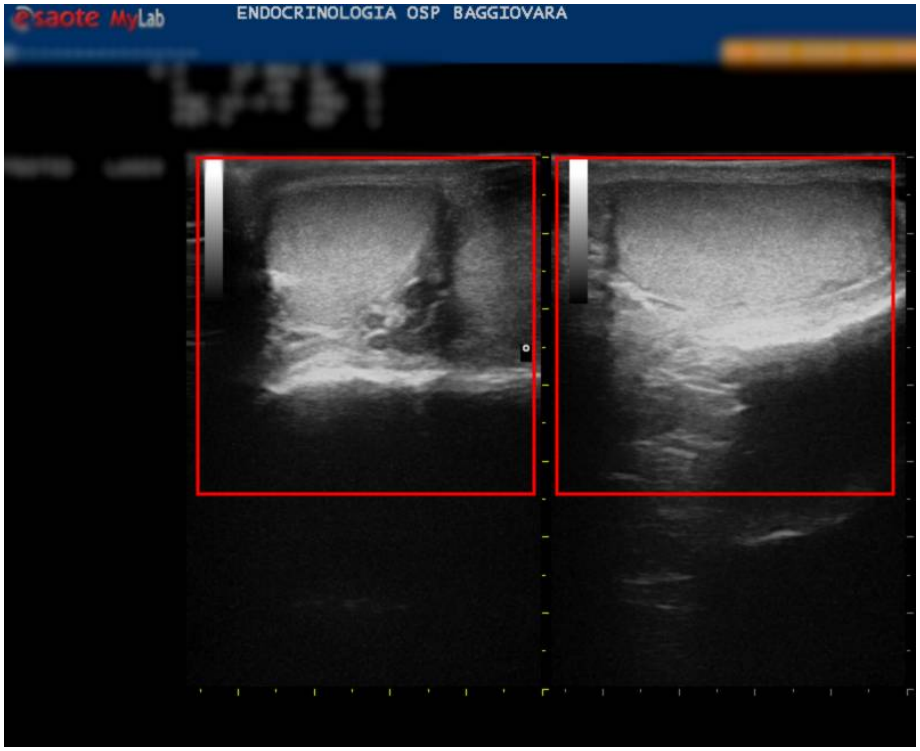


FIGURE 3.1: Visualization of the preprocessing pipeline, showing the transformation from raw images to cropped representations in two different views.

- **Unlabeled Dataset (UD):** This dataset is primarily used for pretraining and contains a mix of testicular and thyroid ultrasound images. It consists of 25,792 images in total, with 1,666 testicular scans and 24,126 thyroid scans, not necessarily from the same patients;
- **Labeled Dataset (LD):** This subset is crucial for the classification task, focusing on predicting testicular tissue homogeneity versus inhomogeneity. It includes 880 testicular images from 220 patients, with homogeneity/inhomogeneity labels available. The class distribution within this labeled dataset is notably uneven, approximately 80–20% respectively.

A significant challenge encountered was the inherent noisiness in the pairing of images and labels. Clinicians typically rely on real-time video evaluation during ultrasound examinations to assess anatomical properties comprehensively. However, only static screenshots are typically saved in routine clinical practice, and these images may not always accurately reflect the

actual homogeneity characteristics of the tissue, thereby introducing noise into the dataset for automated systems.

3.2 Noisy Label Filtering Method

The labeled dataset (LD) contains 880 static ultrasound images annotated as *homogeneous* or *inhomogeneous*. However, due to the intrinsic nature of ultrasound imaging, these annotations can be noisy. During clinical examinations, radiologists evaluate the testicular tissue in real time, whereas only single frames (screenshots) are saved in routine practice. These static images may not always represent the true tissue homogeneity observed during the full video assessment, introducing inconsistencies in the labels. To mitigate this problem, we developed a heuristic filtering procedure aimed at identifying and discarding mislabeled or unreliable samples prior to training.

We first trained a simple ResNet-18 [6] classifier to predict tissue homogeneity using a three-fold cross-validation scheme and a standard cross-entropy loss. To avoid data leakage, all images belonging to the same patient were kept in the same fold. The model exhibited limited generalization and clear signs of overfitting, suggesting that some samples were perturbing the learning process due to inconsistent labeling.

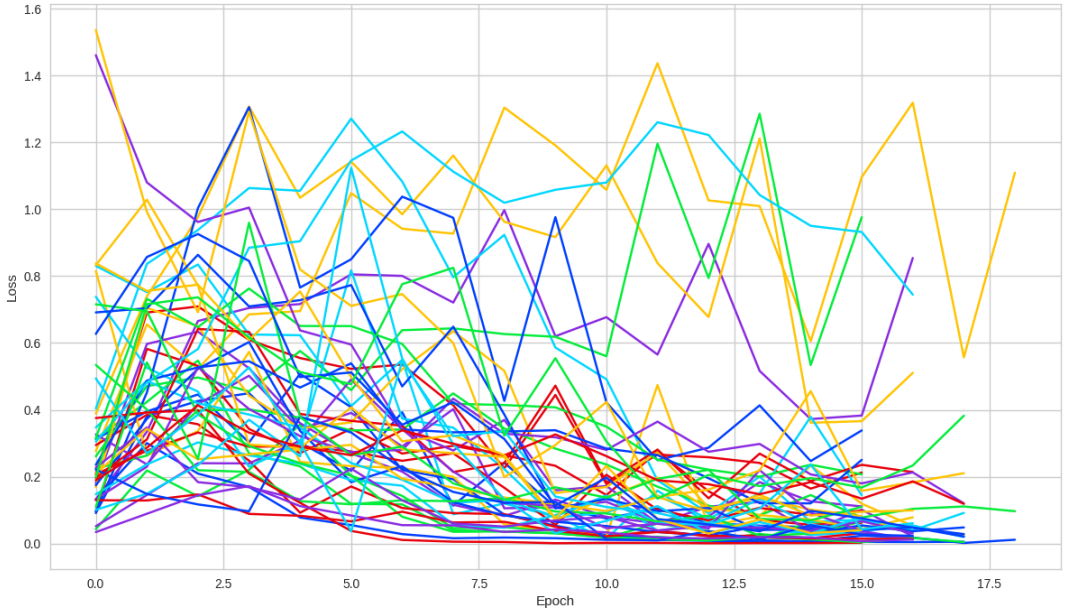


FIGURE 3.2: Example of per-sample training loss trajectories across epochs. Most samples show a smooth, decreasing trend, while a few exhibit irregular spikes where losses repeatedly exceed a value of 1, indicating potentially mislabeled cases. Note that samples may span different numbers of epochs due to the use of a weighted sampler.

To systematically detect these unreliable samples, we analyzed the per-sample loss evolution throughout training(see Fig. 3.2). The intuition is that correctly labeled samples should show a consistent reduction in loss as training progresses, while mislabeled ones tend to display erratic behavior with repeated loss spikes.

A sample was flagged as "suspicious" if its training loss exceeded a threshold of 1 at least three times during the learning process. This criterion was determined empirically: smaller thresholds produced too many false positives, while larger ones failed to capture evident outliers.

To enhance robustness, the filtering process was repeated using two different pretrained initializations of the ResNet-18 [6] backbone, one from ImageNet and one from our custom ultrasound pretraining. For each initialization, we performed four independent training runs with different random seeds under a three-fold cross-validation setup, resulting in a total of 24 runs. Since each sample appears in multiple folds and seeds, it could be flagged as suspicious up to 16 times in total.

Dataset	**Accuracy (↑)**	**F1-Score (↑)**
Complete	81.51 ± 2.78	55.72 ± 4.37
Filtered	$\mathbf{88.15 \pm 1.94}$	68.59 ± 3.30
Flipped	86.78 ± 2.21	$\mathbf{73.17 \pm 1.55}$

TABLE 3.1: Performance of a ResNet pretrained on ImageNet and fine-tuned on the complete LD dataset, compared with the filtered and label-flipped versions after noisy sample removal.

Figure 3.3 compares the results obtained using thresholds of 1, 3, and 5. A value of 1 was overly aggressive, marking many valid samples, whereas 5 was too conservative. Therefore, the final filtering procedure adopted a threshold of 3. Samples consistently flagged as suspicious in all 16 evaluations (72 images in total) were subsequently either removed from the dataset or had their labels flipped after manual inspection.

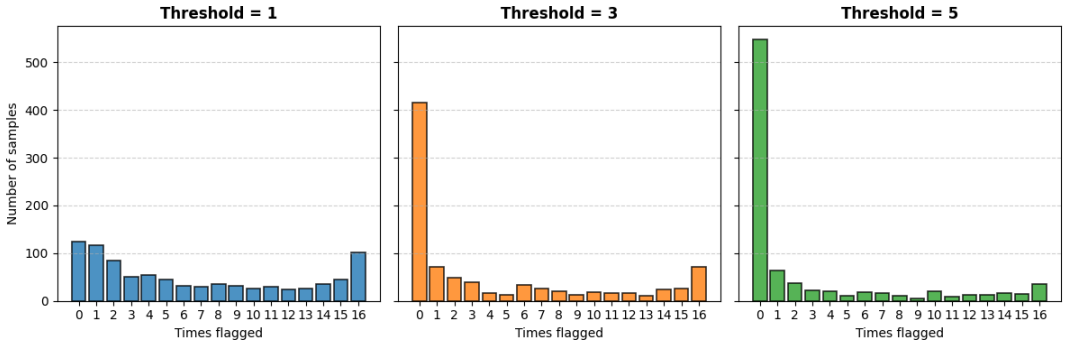


FIGURE 3.3: Distribution of the number of times each sample was flagged as suspicious across 16 evaluations under thresholds of 1, 3, and 5. The threshold of 3 provided the best balance between conservativeness and aggressiveness, effectively separating genuinely mislabeled cases from the majority of stable samples.

Table 3.1 reports the classification performance before and after filtering. Removing or correcting mislabeled samples led to a clear improvement in both accuracy and F1-score, confirming that label noise was a major source of performance degradation.

To validate these findings, clinicians re-examined the suspicious cases using only the available static images (since the dynamic examination videos were not accessible). The re-evaluation revealed several discrepancies with the original annotations, reinforcing the importance of using complete examination videos for accurate labeling in future studies.

All subsequent experiments in this thesis were conducted using the refined labeled dataset produced by this filtering and label-correction process.

Chapter 4

The Generated Dataset

This chapter details the methodology employed for generating a synthetic dataset of testicular ultrasound images using Denoising Diffusion Probabilistic Models (DDPMs) and describes the subsequent filtering algorithm designed to ensure the quality and fidelity of these generated images. This approach directly addresses the significant challenges of data scarcity and privacy concerns prevalent in medical imaging research, particularly in the sensitive domain of testicular ultrasound (TUS).

4.1 Synthetic Data Generation with Denoising Diffusion Probabilistic Models (DDPMs)

As introduced in Chapter 2, Denoising Diffusion Probabilistic Models (DDPMs) [11] are powerful generative models that have demonstrated superior performance over Generative Adversarial Networks (GANs) [12] in terms of training stability and detail modeling for high-quality, realistic image synthesis. This approach is particularly effective in addressing the limitations of small, institution-specific datasets and ethical restrictions on data sharing prevalent in medical imaging. A DDPM operates through a two-phase process: a fixed forward diffusion phase and a learned reverse denoising phase. The forward process incrementally adds Gaussian noise to an image, while the reverse process learns to remove it.

Mathematical Formulation

The diffusion process begins with a clean image $x_0 \sim p_{\text{data}}(x)$, sampled from the real data distribution. The goal of the forward process is to gradually corrupt x_0 by adding Gaussian noise step by step, until after T iterations the image becomes indistinguishable from pure noise. This is achieved by defining a Markov chain where, at each step t, we generate a noisier version x_t from the previous step x_{t-1}:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t;\ \sqrt{\alpha_t}\, x_{t-1},\ (1-\alpha_t)\mathbf{I}\big), \qquad t = 1, \ldots, T. \tag{4.1}$$

Here, α_t represents the fraction of the original signal that is preserved after timestep t, while its complement β_t controls the amount of noise added:

$$\beta_t := 1 - \alpha_t. \tag{4.2}$$

Hence, a smaller β_t implies that less noise is injected at that step, preserving more image information, whereas a larger β_t leads to faster degradation of the image. The entire noise schedule is therefore governed by the sequence $\{\beta_t\}_{t=1}^{T}$, commonly referred to as the *variance schedule*, which determines how the diffusion process evolves over time.

It is often more convenient to work with the cumulative product of the α_t terms, denoted as

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \tag{4.3}$$

which expresses the total amount of signal remaining after t diffusion steps. By marginalizing the forward process, we can directly relate any noised sample x_t to the original clean image x_0:

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t;\ \sqrt{\bar{\alpha}_t}\, x_0,\ (1-\bar{\alpha}_t)\mathbf{I}\big). \tag{4.4}$$

This closed-form expression reveals an important property: instead of applying noise step-by-step, we can sample x_t at any arbitrary timestep t directly. Using the reparameterization trick, we can write

$$x_t = \sqrt{\bar{\alpha}_t}, x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{4.5}$$

which is crucial for training efficiency. In practice, during training, a timestep t is drawn uniformly, noise ε is sampled, and the network is trained to predict ε given (x_t, t). This allows the model to learn how to denoise across all noise levels simultaneously.

The reverse process aims to invert this corruption by learning a denoising distribution of the form

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1};\ \mu_\theta(x_t, t),\ \sigma_t^2 \mathbf{I}\big), \tag{4.6}$$

where the mean $\mu_\theta(x_t, t)$ depends on the current noisy input x_t and the timestep t, and σ_t^2 represents the variance of the reverse transition. Under the ε-parameterization, the mean can be expressed as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \, \varepsilon_\theta(x_t, t) \right). \tag{4.7}$$

Originally, the variance σ_t^2 was fixed to the closed-form posterior variance

$$\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \, \beta_t, \tag{4.8}$$

or simply set to β_t during inference. Later improvements to DDPMs introduced learning of the diagonal variance, interpolating between $\tilde{\beta}_t$ and β_t in log-space, which allowed for more accurate modeling and improved image fidelity.

While the full variational lower bound (VLB) can be optimized to learn the reverse process, in practice it was found that a much simpler surrogate loss works better. In this formulation, the network $\varepsilon_\theta(x_t, t)$ is trained to directly predict the noise ε used in generating x_t. The corresponding *simple loss* is defined as

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}([1,T]),\, x_0 \sim p_{\text{data}},\, \varepsilon \sim \mathcal{N}(0,\mathbf{I})} \left[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2 \right]. \tag{4.9}$$

This formulation avoids the complexities of directly optimizing the VLB while still yielding excellent generative performance. Intuitively, the model learns to remove the injected Gaussian noise step by step, gradually reconstructing the original image.

A critical design choice lies in how the variance schedule $\{\beta_t\}$ is defined. In the original work [11], a linear schedule was proposed, but this quickly injected a large portion of the noise in the very first steps. As a result, the signal-to-noise ratio (SNR) collapsed early, making the denoising task unnecessarily difficult. To mitigate this, the improved [13] introduced a cosine

noise schedule defined as:

$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, \qquad f(t) = \cos^2\Big(\tfrac{\pi}{2} \cdot \tfrac{t/T+s}{1+s}\Big), \quad s \approx 0.008, \tag{4.10}$$

which distributes the noise more evenly across timesteps. This schedule preserves meaningful structure for a longer fraction of the diffusion chain, allowing the model to denoise more effectively (see Fig. 4.1).

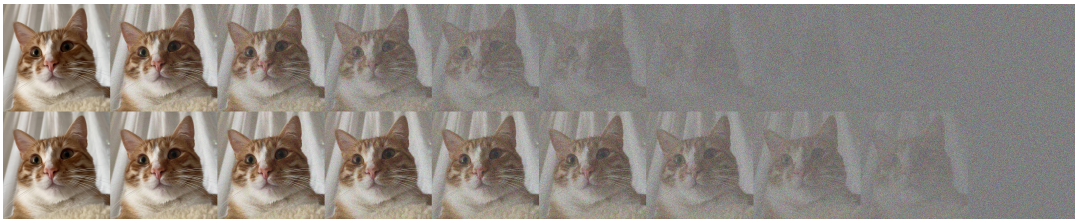


FIGURE 4.1: Conceptual comparison of linear (up) and cosine (down) noise schedules. Linear scheduling injects most of the noise in the first steps, whereas cosine distributes noise more evenly and preserves signal longer.[1]

Finally, state-of-the-art extensions such as Latent Diffusion Models (LDMs) apply the diffusion process in a compressed latent space rather than directly on pixels. An autoencoder $\mathcal{E}, \mathcal{D}$ is used to map images to a latent representation $z_0 = \mathcal{E}(x_0)$, the forward diffusion process is performed on z_t, and after denoising, the image is reconstructed as $\hat{x} = \mathcal{D}(z_0^{\text{sample}})$. This approach significantly reduces computational cost while retaining high fidelity, which is particularly advantageous for high-resolution ultrasound data.

The U-Net Architecture

The denoising network $\varepsilon_\theta(x_t, t)$ follows a U-Net [14] backbone adapted from the implementation of OpenAI's improved diffusion models [13] (see Fig. 4.2). Its design is symmetric, consisting of an *encoder* (downsampling path), a *bottleneck*, and a *decoder* (upsampling path), with skip connections linking corresponding layers. In addition, residual connections, timestep conditioning, and self-attention layers are integrated throughout the network.

The encoder is organized into stages, each composed of two residual blocks followed by downsampling (except the last stage).

A residual block (ResBlock) itself follows a "two-convolution" structure. Each block receives not only the noisy image representation x_t but also the corresponding timestep t, which encodes how much noise has been applied in the diffusion process. To integrate this temporal information, the scalar timestep t is transformed into a high-dimensional embedding $\gamma(t)$ through a sinusoidal positional encoding, similar to that used in the Transformer architecture [15]. This embedding provides the network with a continuous representation of diffusion time, allowing it to adapt its denoising behavior according to the current noise level. A small multilayer perceptron (MLP) projects $\gamma(t)$ to match the dimensionality of the feature maps before it is injected into each residual block.

The internal structure of a residual block is as follows:

- **Input layers:** normalization $\rightarrow$ SiLU $\rightarrow$ 3×3 convolution;
- **Timestep embedding:** $\gamma(t)$ is projected by an MLP and injected after the first convolution; If scale-shift normalization is enabled, the embedding produces affine parameters $(\text{scale}, \text{shift})$ that modulate the normalized activations; otherwise it is simply added;
- **Output layers:** normalization $\rightarrow$ SiLU $\rightarrow$ dropout $\rightarrow$ 3×3 convolution (with zero-initialization);
- **Skip connection:** identity if the number of channels is preserved, or a 1×1 (or optional 3×3) convolution if channel dimensions change.

The block output is the sum of the transformed path and the skip connection. Self-attention layers are inserted right after residual blocks whenever the resolution matches a predefined set (in this configuration, 16×16 and 8×8).

Stacking two such ResBlocks defines one stage. With a base channel size of 128 and multipliers $(1, 1, 2, 2, 4, 4)$, the encoder has channel widths

$$128 \rightarrow 128 \rightarrow 256 \rightarrow 256 \rightarrow 512 \rightarrow 512$$

at resolutions

$$256 \times 256 \rightarrow 128 \times 128 \rightarrow 64 \times 64 \rightarrow 32 \times 32 \rightarrow 16 \times 16 \rightarrow 8 \times 8.$$

Downsampling between stages is performed by strided convolution, halving spatial resolution and increasing channel count according to the multiplier.

The bottleneck at the lowest resolution combines residual and attention layers in the sequence: ResBlock $\rightarrow$ AttentionBlock $\rightarrow$ ResBlock. This allows the model to process features both locally (via convolution) and globally (via attention).

The decoder mirrors the encoder. At each stage, features are upsampled with nearest-neighbor interpolation followed by convolution, halving the channel count while doubling spatial resolution. Residual blocks conditioned on the timestep embedding refine the upsampled features. Attention layers are again inserted right after ResBlocks at 16×16 and 32×32, enhancing global consistency. Skip connections concatenate encoder features with decoder features at the same resolution, ensuring that spatial detail lost in downsampling is preserved. A final convolution maps the reconstructed features to the image domain, predicting either the clean image $\hat{x}_{t-1}$ or, in the ε-parameterization, the added noise $\varepsilon_\theta(x_t, t)$.

4.2 Evaluation Metrics for Synthetic Images

To assess the quality and practical utility of the generated synthetic images, four established metrics were employed: improved Precision (P), and Recall (R) [16], Fréchet Inception Distance (FID) [17] and Inception Score (IS) [18]. These metrics help quantify how well the synthetic data mimics real data and covers its distribution.

- **Precision (P):** This metric assesses the fidelity of the generated images, quantifying the distributional similarity between the real and generated data. It is computed as the fraction of generated samples whose embeddings fall within the real data manifold. The real data manifold, $\mathcal{M}_r$, is constructed by first extracting feature embeddings for all real images using a pretrained network (e.g., Inception). Then, for each real image embedding, a hypersphere is defined. The radius of each hypersphere is determined by the distance to its k-th nearest neighbor within the set of real embeddings. The manifold is the union of all these hyperspheres;
- **Recall (R):** This metric measures the diversity of the synthetic data, indicating how much of the real data distribution is covered by the synthetic samples. It is computed as

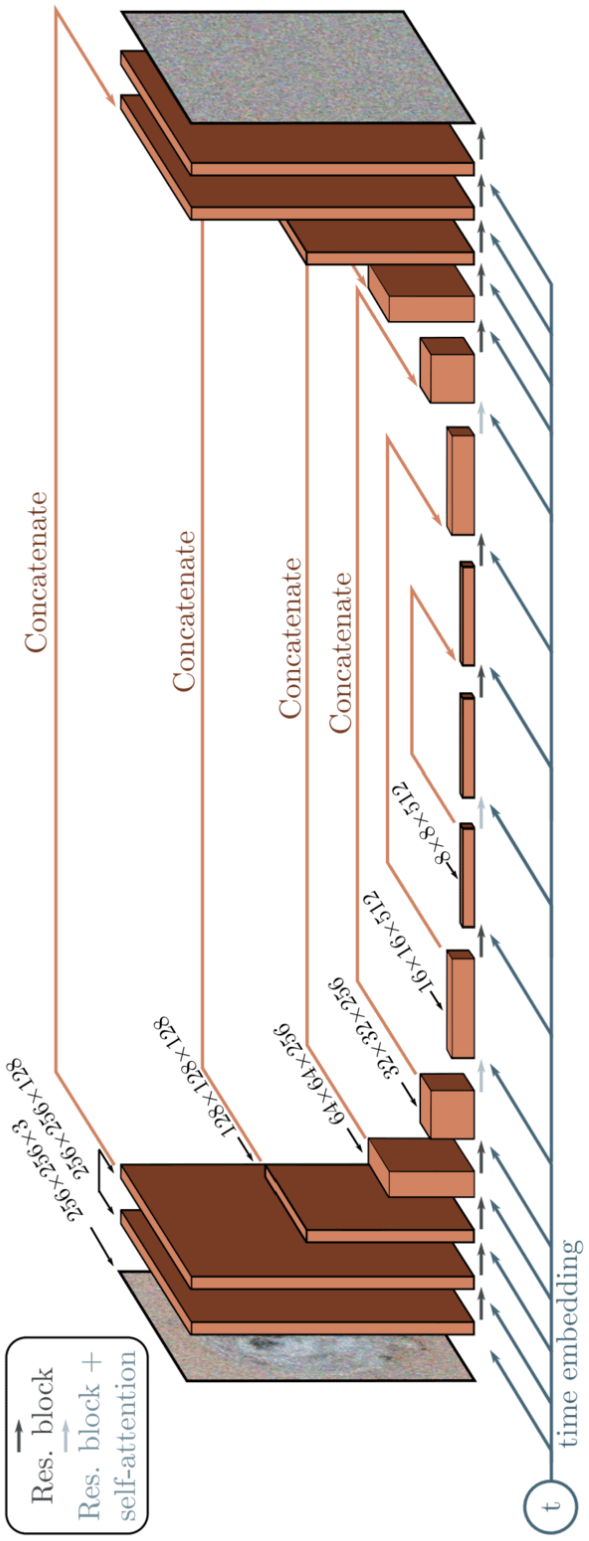


FIGURE 4.2: U-Net architecture used in the DDPM denoising process. Encoder applies residual blocks with timestep conditioning and attention at selected resolutions, bottleneck integrates residual and attention layers, decoder upsamples features while using skip connections to recover spatial detail.

the fraction of real samples that fall inside the generated data, $\mathcal{M}_g$, which is defined symmetrically to the precision metric but with respect to the generated data;

- **Fréchet Inception Distance (FID):** This metric captures both fidelity and diversity by comparing the mean and covariance of real and generated feature distributions. A lower FID value indicates that the generated images are statistically more similar to the real ones. The metric is computed as in equation 4.11 where μ_r and Σ_r are the mean and covariance of the real image features, and μ_g and Σ_g are those of the generated image features:

$$\text{FID} = ||\mu_r - \mu_g||_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right) \tag{4.11}$$

- **Inception Score (IS):** This metric evaluates both the quality and diversity of generated images based on the predictions of a pretrained Inception network. High-quality images are expected to yield low-entropy conditional label distributions $p(y|x)$, meaning the model is confident in its prediction. At the same time, a diverse set of images should yield a marginal distribution $p(y)$ that is close to uniform. The IS is defined as the exponential of the Kullback–Leibler divergence between these two distributions:

$$\text{IS} = \exp\left(\mathbb{E}_{x\sim p_g}\left[D_{\text{KL}}\big(p(y|x)\,||\,p(y)\big)\right]\right) \tag{4.12}$$

  A higher IS indicates that the generated images are both high-quality and diverse.

4.3 Filtering Method for Synthetic Data

To ensure the quality and fidelity of the generated synthetic ultrasound images, a dedicated filtering method was developed based on the precision metric. This process computes a real data manifold $\mathcal{M}_r$ from the feature embeddings of the real labeled dataset (LD) and selects only synthetic images whose embeddings fall within this manifold.

Algorithm 1 Filtering Algorithm for Synthetic Images

Input: $\mathcal{X}_g$, Φ_r, Φ_g, # of neighbors k.
Initialize $\mathcal{X}_g^{\text{filtered}} \leftarrow \emptyset$
for all ϕ_i^r in Φ_r **do**
  Compute $r_i \leftarrow$ Euclidean distance to the k-th nearest neighbor of ϕ_i^r in Φ_r
  $B(\phi_i^r, r_i) \leftarrow$ hypersphere centered at ϕ_i^r with radius r_i
end for
$\mathcal{M}_r \leftarrow \bigcup_{i=1}^{N} B(\phi_i^r, r_i)$
for all ϕ_j^g in Φ_g **do**
  if $\phi_j^g \in \mathcal{M}_r$ **then**
    Add x_j^g to $\mathcal{X}_g^{\text{filtered}}$
  end if
end for
return $\mathcal{X}_g^{\text{filtered}}$

The filtering procedure involves:

1. **Real Data Manifold Computation:** The real data manifold, $\mathcal{M}_r$, is computed from the feature embeddings of the real labeled dataset (LD). It is defined as the union of hyperspheres around each real image feature vector;

2. **Synthetic Image Selection:** Given a set of generated images, their feature representations are computed, and only those whose embeddings lie inside the real data manifold, $\mathcal{M}_r$, are selected.

4.4 Generation Results and Dataset Characteristics

The generation and subsequent filtering of synthetic data are crucial steps in this study, aiming to produce high-quality images that effectively address the challenge of data scarcity. An overview of the entire pipeline, from noise input to the final filtered dataset, is shown in Figure 4.3. The process leverages a Denoising Diffusion Probabilistic Model (DDPM) to generate an initial pool of approximately 20,000 synthetic images, which are then refined through a filtering method to ensure their fidelity and diversity.

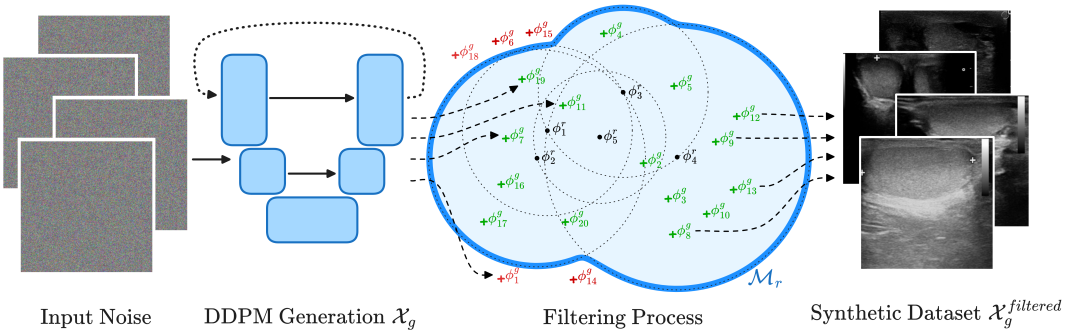


FIGURE 4.3: Overview of the pipeline for synthetic data generation and filtering.

The filtering strategy relies on a key parameter, k, which defines the radius of the hyperspheres used to construct the real data manifold. This parameter governs the trade-off between precision and recall. As shown in Figure 4.4, smaller k values produce tighter hyperspheres, yielding higher precision but lower recall since only samples very close to real embeddings are retained. Conversely, larger k values create a looser manifold that includes more generated samples, improving recall. However, because precision is evaluated with a fixed neighborhood size ($k = 3$), the apparent precision decreases as the manifold expands, since more distant samples are still considered within the real-data region. A value of $k = 50$ was found to provide the best balance between fidelity and diversity in the filtered dataset.

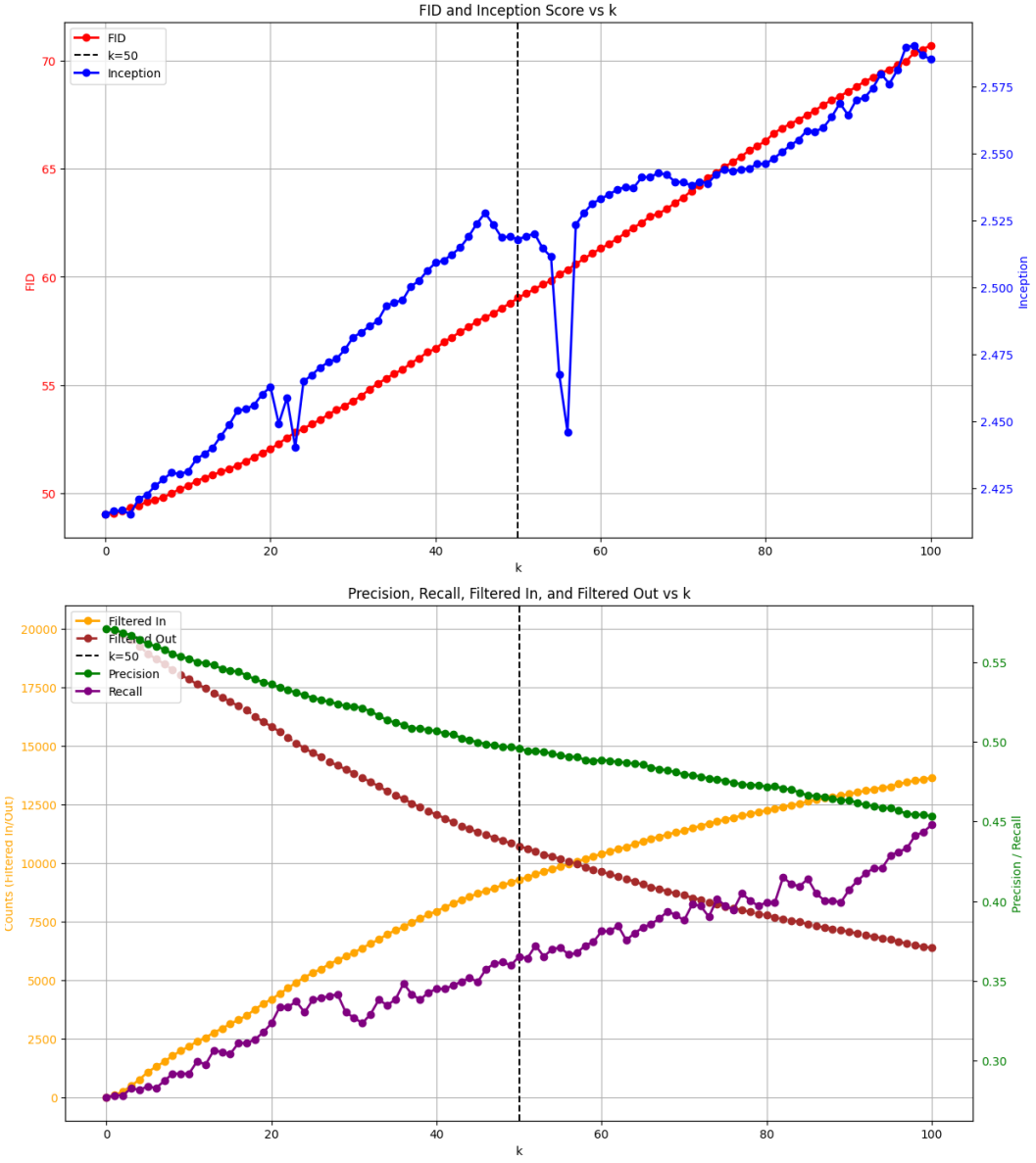


FIGURE 4.4: Evaluation of filtering metrics computed on the sampled 20,000 synthetic images. The plots show how FID, Inception Score, precision, recall, and the number of filtered-in and filtered-out samples evolve as the neighborhood size parameter k increases, while the k used for computing metrics is fixed at 3. The threshold of $k = 50$ (dashed line) is highlighted as found to be the optimal trade-off between fidelity and diversity.

The application of the filtering strategy produced the expected trade-offs across the evaluation metrics. Precision decreased from 57.10 to 49.57, indicating that the retained samples were

more tightly constrained within the real-data manifold. At the same time, recall increased from 27.60 to 36.53, showing that the filtering procedure improved coverage of the real data distribution. The variation in inception values remained limited, while the Fréchet Inception Distance (FID) rose from 49.03 to 59.03. Overall, these results confirm that the filtering method achieved a reasonable balance between fidelity to the real data and diversity across the feature space. After filtering, the initial pool of approximately 20,000 synthetic samples was reduced to 9,289 testicular ultrasound images. The final dataset, released in PNG format and corresponding to the Ultrasound/Mode-B modality, has been made publicly accessible online [1]. Representative examples of images retained (green boxes) and discarded (red boxes) by the filtering process are shown in Figure 4.5.

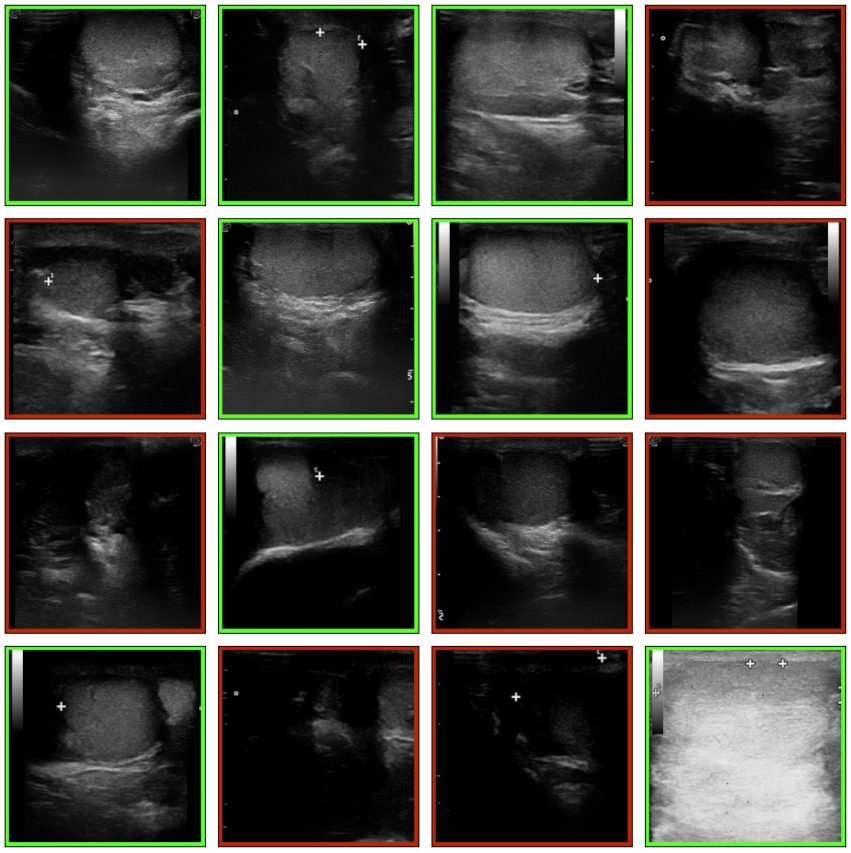


FIGURE 4.5: Examples of filtered synthetic images. Images retained in the dataset are highlighted with green boxes, while discarded images are marked with red boxes.

[1] Filtered synthetic data are publicly released at `https://ditto.ing.unimore.it/testiculus`.

Chapter 5

Training Methodologies and Evaluation Metrics

This chapter presents the methodologies employed to train and evaluate the proposed classification model. It first introduces the network architecture, including the choice of backbone and classification head. It then describes the semi-supervised pretraining strategy, which leverages both real and synthetic data to improve feature representations. Afterwards, the fine-tuning procedure on the labeled dataset (LD) is outlined, together with the data augmentation and regularization strategies designed to improve generalization. Finally, the evaluation protocols and metrics used to assess the model's performance are presented.

5.1 Classification Model Architecture

The backbone model selected for this study is ResNet-18 [6], a convolutional neural network architecture widely recognized for its efficiency and strong performance in data-limited scenarios. Preliminary experiments revealed that deeper networks such as ResNet-50 [6], or more recent transformer-based [7] architectures, tended to overfit when trained on the Labeled Dataset (LD). This was expected given the relatively small dataset size, where model complexity can become detrimental to generalization. By contrast, ResNet-18 [6] provided a suitable compromise, being expressive enough to capture relevant visual features while remaining robust against overfitting.

On top of this backbone, a lightweight classification head was added. The head consists of a single fully connected linear layer, whose output dimension corresponds to the binary classification task of predicting testicular homogeneity versus inhomogeneity. This minimal design places the representational burden primarily on the pretrained backbone, while ensuring that the fine-tuning stage remains stable and efficient.

5.2 Semi-Supervised Pretraining Strategy

In the context of ultrasound analysis, pretraining plays a central role in addressing the scarcity of annotated data. Instead of relying exclusively on supervised training from scratch, a strategy that leverages both labeled and unlabeled data in a semi-supervised framework was adopted. The goal is to train a backbone encoder $f(\cdot)$ that extracts domain-specific features from ultrasound images, which can later be transferred and specialized for the binary classification of testicular homogeneity. An overview of the complete pretraining and fine-tuning pipeline is illustrated in Figure 5.1.

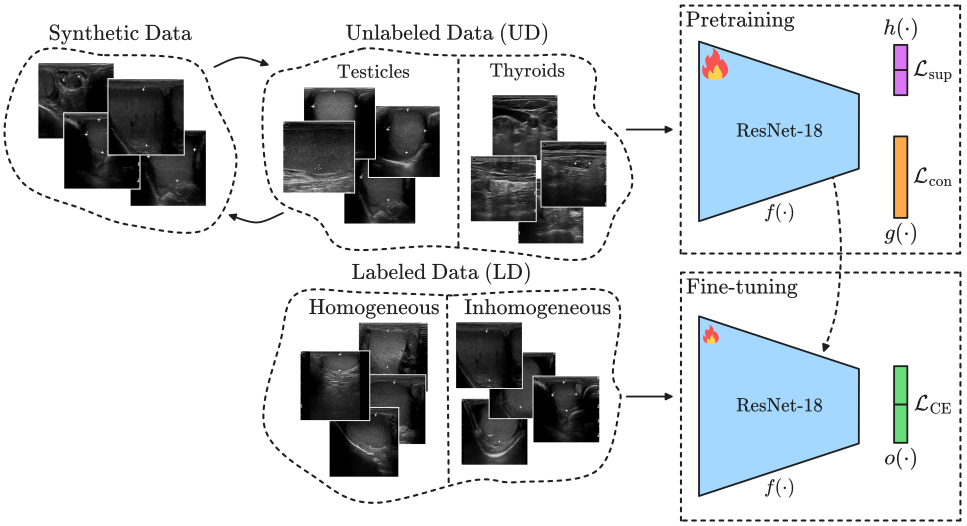


FIGURE 5.1: Overview of the pretraining and fine-tuning pipeline. The encoder $f(\cdot)$ is pretrained with a semi-supervised strategy combining contrastive loss L_{con} (via projection head $g(\cdot)$) and supervised loss L_{sup} (via classification head $h(\cdot)$). Pretraining is performed on the UD or synthetic dataset, followed by fine-tuning on the LD.

Data Sources

Two data sources were employed during pretraining:

- The **Unlabeled Dataset (UD)**, composed of real ultrasound scans of both testicles and thyroids, providing variability in anatomy and acquisition conditions;
- The **Filtered Synthetic Dataset**, generated by a diffusion model trained on real testicular scans and curated through a filtering strategy to remove unrealistic or out-of-distribution samples (see Chapter 4 for more details); This dataset provided additional training samples while mitigating the risks of data scarcity.

Contrastive Learning with SimCLR

The unsupervised component of the pretraining procedure is based on the SimCLR [19] framework, a contrastive learning approach that has shown strong results in visual representation learning. The key idea is to learn an embedding function $f(\cdot)$ that maps images into a feature space where different augmented views of the same image are brought closer together, while views of different images are pushed apart. This encourages the encoder to focus on intrinsic, semantically meaningful features that remain stable across transformations, rather than superficial pixel-level patterns.

Formally, each input image x_i is augmented twice through a stochastic pipeline of transformations such as cropping, flipping, color jittering, and Gaussian noise. These produce two correlated views, $\tilde{x}_{2i-1}$ and $\tilde{x}_{2i}$, which form a *positive pair*. Conversely, augmented views derived from different images constitute *negative pairs*. All $2N$ augmented views in a batch are processed by the backbone encoder $f(\cdot)$, followed by a projection head $g(\cdot)$ that maps embeddings into a latent space suited for contrastive training. The resulting vectors are denoted as:

$$z_k = g(f(\tilde{x}_k)), \quad k \in \{1, \ldots, 2N\}. \tag{5.1}$$

The similarity between two vectors is measured via cosine similarity:

$$s(z_j, z_k) = \frac{z_j^\top z_k}{\|z_j\| \|z_k\|}. \tag{5.2}$$

For each positive pair (j, k), SimCLR [19] defines the contrastive loss as:

$$\ell(j,k) = -\log \frac{\exp(s(z_j, z_k)/\tau)}{\sum_{m=1}^{2N} \mathbb{1}_{[m \neq j]} \exp(s(z_j, z_m)/\tau)}, \tag{5.3}$$

where τ is a temperature parameter that controls the sharpness of the similarity distribution. This loss penalizes cases where positive pairs are not significantly closer than negatives in the latent space.

The overall objective is the average of all such terms across the batch:

$$L_{con} = \frac{1}{2N} \sum_{i=1}^{N} \Big[\ell(2i-1, 2i) + \ell(2i, 2i-1) \Big]. \tag{5.4}$$

Intuitively, minimizing L_{con} forces the encoder to learn invariances to the chosen augmentations, a property that is particularly desirable in ultrasound imaging, where appearance can vary substantially due to acquisition settings, probe position, or patient anatomy. By learning to associate different views of the same image, the model becomes robust to such sources of variability and captures domain-relevant features that generalize across cases.

Supervised Auxiliary Classification Task

To complement the unsupervised contrastive component, a supervised classification task was integrated during pretraining. Specifically, each ultrasound image was labeled according to its anatomical organ (testicle or thyroid). On top of the encoder $f(\cdot)$, a lightweight classification head $h(\cdot)$ was trained to predict these organ labels. For an augmented view $\tilde{x}_k$, the predicted logits are given by:

$$c_k = h(f(\tilde{x}_k)). \tag{5.5}$$

The *supervised loss* was defined as the standard cross-entropy:

$$L_{sup} = \frac{1}{2N} \sum_{i=1}^{N} \Big[CE(c_{2i-1}, y_i) + CE(c_{2i}, y_i) \Big], \tag{5.6}$$

where $y_i \in \{\text{testicle}, \text{thyroid}\}$ denotes the organ label.

This auxiliary task provided high-level semantic guidance, encouraging the encoder to organize its representation space around anatomical categories, which was expected to improve the transferability of the learned features.

Relation to USCL and Final Objective

The supervised auxiliary loss employed during pretraining is directly inspired by the Ultrasound Self-Supervised Contrastive Learning (USCL) framework [20], which originally introduced the idea of enriching contrastive learning with an organ-classification branch. In the original USCL design, positive pairs were drawn from distinct frames of ultrasound videos, thereby exploiting temporal redundancy as a source of semantic consistency. This setup was well suited to sequential data but not directly applicable to our image-only setting.

Our contribution was to adapt this strategy from videos to still images by integrating it with the SimCLR paradigm. Instead of sampling frames across time, we generate positive pairs through stochastic augmentation of the same image, following the SimCLR [19] framework. In this way, our approach preserves the organ-level semantic supervision of USCL while grounding it in an image-based contrastive formulation. The encoder therefore learns representations that are both invariant to augmentation-induced variability and organized according to anatomical distinctions.

The overall pretraining objective combines the contrastive and supervised components:

$$L = L_{con} + \lambda L_{sup}, \tag{5.7}$$

where λ controls the balance between unsupervised contrastive learning and supervised anatomical classification. Unlike in generic frameworks where λ is chosen heuristically, its impact in our setting was systematically analyzed through an ablation study, of which results presented in Section 6.2. This evaluation confirmed that the combined objective achieves a more effective representation space for transfer to the downstream classification task of testicular homogeneity.

5.3 Fine-tuning Procedure

After pretraining, the encoder $f(\cdot)$ was adapted to the downstream classification task of testicular homogeneity versus inhomogeneity. During this stage, the pretraining heads $g(\cdot)$ and $h(\cdot)$ were discarded and replaced with a new binary classification head $o(\cdot)$, trained on top of the encoder (see Figure 5.1). Given an input image x, the final prediction is obtained as:

$$\hat{y} = o(f(x)). \tag{5.8}$$

The fine-tuning objective was the standard binary cross-entropy loss computed on the Labeled Dataset (LD):

$$L_{FT} = CE(\hat{y}, y), \tag{5.9}$$

where $y \in \{\text{homogeneous}, \text{inhomogeneous}\}$ is the ground-truth label.

Several regularization and augmentation strategies were adopted to improve robustness and prevent overfitting. First, standard spatial transformations such as random rotations, horizontal flips, and small shifts were applied. More importantly, a novel marker insertion strategy was introduced: synthetic markers were randomly overlaid on the training images, mimicking the annotations and measurement artifacts frequently present in clinical scans. This forced the network to disregard superficial cues and instead extract features from the parenchymal tissue, leading to more clinically relevant predictions (see Figure 5.3). The complete set of synthetic markers employed for this augmentation is illustrated in Figure 5.2.

Finally, the class distribution in the LD was highly imbalanced, with approximately 20% inhomogeneous and 80% homogeneous samples. To mitigate this, a weighted sampling strategy was employed, ensuring that both classes contributed equally during training. This adjustment stabilized the optimization process and improved sensitivity to the minority class.

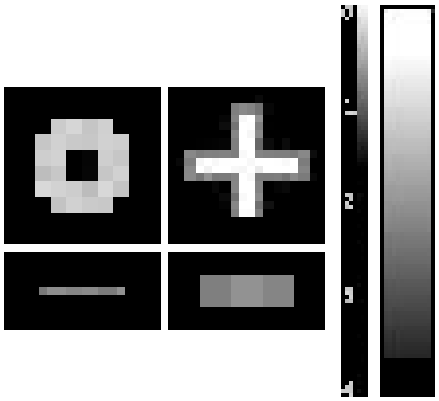


FIGURE 5.2: Synthetic markers used for augmentation. The figure shows six representative marker types that were randomly placed on training images to mimic measurement annotations commonly found in ultrasound scans.

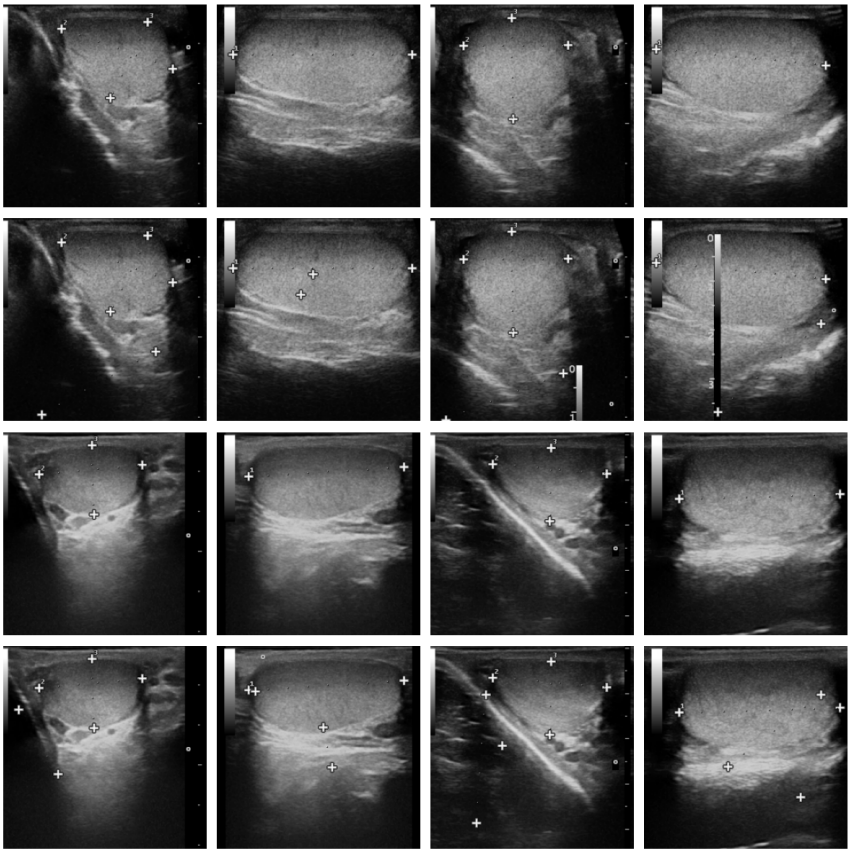


FIGURE 5.3: Illustration of the marker insertion augmentation. The first and third rows show the original images, while the second and fourth rows present their corresponding augmented versions with inserted markers.

5.4 Evaluation Protocols

The evaluation of the trained models relied on standard classification metrics: *Accuracy* (5.10), *Precision* (5.11), *Recall* (5.12), and *F1-Score* (5.13). Let TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively. The metrics are formally defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{5.10}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.12}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.13}$$

Each of these measures highlights different aspects of classification performance. Accuracy captures the overall rate of correct predictions, while Precision focuses on the reliability of positive predictions and Recall reflects the model's sensitivity to positive cases. The F1-Score provides a balanced measure that is particularly useful when classes are imbalanced, as it harmonically combines Precision and Recall.

Thresholds for Random Guessing and Class Imbalance

In the specific case of our dataset, the class distribution is approximately 80% negative (homogeneous) and 20% positive (inhomogeneous). This imbalance strongly affects the interpretation of the metrics. For instance, a trivial classifier that always predicts the negative class would reach an Accuracy of 80% without learning any meaningful representation. However, its Recall on the positive class would be zero, and consequently its F1-Score would also be zero.

Similarly, a random guesser that assigns labels according to the class proportions (80% negative, 20% positive) would on average achieve:

$$\text{Accuracy} \approx 0.68, \quad \text{Precision} \approx 0.20, \quad \text{Recall} \approx 0.20, \quad \text{F1-Score} \approx 0.20. \tag{5.14}$$

These values represent the *non-informative baseline* in our setting. Therefore, for a model to demonstrate genuine predictive ability, its performance must significantly exceed both the trivial 80% Accuracy baseline and the random guessing scores, with particular emphasis on Recall and F1-Score, which better capture the model's capability to identify the minority positive class.

Cross-Validation Strategy

To ensure robust and unbiased evaluation, a three-fold cross-validation schema was adopted. Images from the same patient were always assigned to the same fold, thus preventing data leakage and ensuring a fair estimate of generalization ability. The final reported performance corresponds to the average of the metrics across the three folds, together with their standard deviations.

Chapter 6

Experiments and Results

This chapter presents the experimental evaluation of the proposed methodologies. Both quantitative results and qualitative analyses are reported, highlighting the impact of pretraining strategies, the role of synthetic data, and the effectiveness of marker augmentation. The aim is to provide a thorough validation of the contributions, complementing the methodological description of Chapter 5.

6.1 Experimental Setup and Implementation Details

Hardware and Software

All experiments were conducted on two types of GPUs: NVIDIA L40S and NVIDIA RTX 2080 Ti. The training pipeline was implemented in the PyTorch framework (v2.7), leveraging mixed-precision training to reduce memory consumption and training time. Data preprocessing, augmentation, and evaluation routines were implemented using TorchVision and custom scripts.

Hyperparameters

To ensure reproducibility, here are detailed the most relevant hyperparameters (Table 6.1):

- **Pretraining:** Optimizer LARS, initial learning rate 0.3, cosine decay schedule, weight decay 10^{-6}, momentum 0.9, batch size of 1024;
- **Fine-tuning:** Three-fold cross-validation. Batch size of 64, learning rate of 10^{-5} for the backbone and 10^{-4} for the classification head. Weighted sampling was applied to mitigate class imbalance;
- **Other settings:** All experiments used input images resized to 224×224, random horizontal flips, random rotations, and color jittering where appropriate.

TABLE 6.1: Summary of hyperparameters used for pretraining and fine-tuning.

Stage	Parameter	Value
Pretraining	Optimizer	LARS
Pretraining	Initial Learning Rate	10^{-3}
Pretraining	LR Schedule	Cosine Decay
Pretraining	Batch Size	1024
Pretraining	Momentum	0.9
Pretraining	Weight Decay	10^{-6}
Fine-tuning	Batch Size	64
Fine-tuning	LR (Backbone)	10^{-5}
Fine-tuning	LR (Head)	10^{-4}
Fine-tuning	Cross-validation	3 folds
Fine-tuning	Class Balancing	Weighted Sampler
Fine-tuning	Epochs	30

6.2 The Role and Impact of Pretraining Strategies

The experiments highlight the important role of pretraining for downstream classification. As shown in Table 6.2, models trained from scratch struggle to learn effective representations, with both accuracy and F1-Score far below desirable levels. This confirms the difficulty of relying solely on the limited labeled dataset without prior knowledge.

Pretraining on ImageNet provides substantial improvements, establishing a strong baseline with notable gains in accuracy and recall. In contrast, USCL pretraining does not transfer as effectively, despite being ultrasound-specific. This gap highlights the strong influence of acquisition domain shift: representations learned from one ultrasound machine or setting

TABLE 6.2: Three-fold cross-validation results on the homogeneous and inhomogeneous downstream task, starting from different pretraining strategies.

Pretraining	**# Samples**	**Accuracy (↑)**	**F1-Score (↑)**	**Precision (↑)**	**Recall (↑)**
$\sim$	$\sim$	73.93 ± 6.80	56.05 ± 7.11	45.48 ± 9.91	75.12 ± 3.66
ImageNet [6]	1.28M	83.89 ± 2.15	67.89 ± 1.85	61.72 ± 3.08	76.08 ± 4.11
USCL	23.00K	75.46 ± 2.64	57.43 ± 1.63	47.37 ± 2.16	73.86 ± 4.76
UD (only testicles)	1.66K	81.09 ± 2.95	65.44 ± 2.32	55.86 ± 2.74	79.55 ± 2.79
UD (only thyroids)	24.13K	80.13 ± 2.83	63.92 ± 2.46	54.30 ± 3.87	78.88 ± 2.06
UD	25.79K	$\mathbf{86.78 \pm 2.21}$	$\mathbf{73.17 \pm 1.55}$	$\mathbf{67.84 \pm 1.67}$	$\mathbf{80.27 \pm 2.13}$

may not generalize well to others. In practice, USCL shows promise when pretraining and downstream data come from the same acquisition source, but broader generalization likely requires orders of magnitude more data to cover the variability across devices and protocols.

An ablation on the composition of the Unlabeled Dataset (UD) further illustrates the importance of data diversity. Testicle-only pretraining favors recall but lowers balance between precision and recall, while thyroid-only pretraining achieves comparable gains. The best results, however, arise from combining both organs, with multi-organ pretraining delivering the highest accuracy, F1-Score, and precision. This confirms that heterogeneous ultrasound corpora support richer and more transferable feature learning.

Overall, these findings demonstrate that effective pretraining requires both scale and diversity. While ImageNet pretraining leverages size and variability, ultrasound-specific strategies must also overcome the acquisition domain shift. To generalize across machines and clinical settings, very large and diverse ultrasound datasets are likely necessary. Multi-organ pretraining appears to be a step in this direction, offering the most effective initialization for classifying testicular inhomogeneity.

Ablation Study on Pretraining Loss Composition

As introduced in Chapter 5.2, the pretraining objective combines a contrastive loss $\mathcal{L}_{\text{con}}$ with a supervised auxiliary classification loss $\mathcal{L}_{\text{sup}}$, balanced by the parameter λ (see Eq. 5.7). To better understand the contribution of each component, we performed an ablation study by selectively activating or removing losses and varying λ. The results are summarized in Table 6.3.

The analysis highlights three key findings. First, training with only the contrastive component already yields solid performance, confirming that invariance to augmentations provides robust ultrasound representations. Second, adding the supervised auxiliary loss consistently improves results, as it encourages the encoder to organize the representation space around anatomical categories, leading to higher accuracy and F1-score. Third, the weighting factor λ plays a crucial role: too low or too high values reduce the benefit of combining the two terms, whereas $\lambda = 0.2$ achieves the best trade-off, yielding the highest downstream accuracy (86.78%) and F1-score (73.17%). Finally, relying solely on the supervised loss ($\lambda = 1.0$) results in inferior performance compared to the combined formulation, demonstrating that contrastive learning remains essential for capturing domain-invariant features.

Overall, this ablation confirms the intuition from Chapter 5 that the synergy between unsupervised and supervised techniques is critical, and that a carefully balanced composition of losses produces the most transferable feature representations for the downstream classification of testicular inhomogeneity.

TABLE 6.3: Ablation study on using different loss components and varying λ.

$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{sup}}$	λ	**Accuracy** (↑)	**F1-Score** (↑)
✓	✗	0.0	84.77 ± 2.56	69.41 ± 0.51
✓	✓	0.1	85.38 ± 2.71	71.40 ± 0.93
✓	✓	0.2	$\mathbf{86.78 \pm 2.21}$	$\mathbf{73.17 \pm 1.55}$
✓	✓	0.3	85.29 ± 2.68	71.36 ± 0.70
✓	✓	0.4	85.03 ± 2.43	71.22 ± 1.26
✗	✓	1.0	83.88 ± 2.56	70.11 ± 1.56

6.3 The Impact of Synthetic Data in Pretraining

Given the scarcity of real medical images, it is crucial to assess whether synthetic data can serve as a viable substitute for pretraining. To this end, we evaluated pretraining on synthetic data, both with and without the proposed filtering procedure, and compared its effectiveness to real data pretraining. The results of this comparison are reported in Table 6.4.

TABLE 6.4: Three-fold results on the downstream task when pretraining with different combinations of real and synthetic data with ($\mathcal{X}_g^f$) or without ($\mathcal{X}_g$) applying the proposed filtering procedure.

Real Testicle	**Real Thyroids**	**Synthetic Testicle**	**Accuracy (↑)**	**F1-Score (↑)**	**Precision (↑)**	**Recall (↑)**
✗	✗	$\mathcal{X}_g$	80.58 ± 3.58	64.24 ± 2.31	55.46 ± 3.08	77.13 ± 1.70
✗	✗	$\mathcal{X}_g^f$	80.42 ± 2.71	64.77 ± 1.62	54.69 ± 3.24	80.12 ± 2.02
✓	✗	$\sim$	81.09 ± 2.95	65.44 ± 2.32	55.86 ± 2.74	79.55 ± 2.79
✗	✓	$\mathcal{X}_g$	82.88 ± 2.11	67.61 ± 1.56	58.87 ± 1.71	79.59 ± 3.63
✗	✓	$\mathcal{X}_g^f$	84.60 ± 2.08	70.63 ± 1.20	62.19 ± 1.76	$\mathbf{82.17 \pm 0.97}$
✓	✓	$\sim$	$\mathbf{86.78 \pm 2.21}$	$\mathbf{73.17 \pm 1.55}$	$\mathbf{67.84 \pm 1.67}$	80.27 ± 2.13

The results highlight several important findings:

- **Synthetic-only pretraining:** When relying exclusively on synthetic data, the classification performance is lower than that obtained from ImageNet initialization. This gap suggests that, although synthetic images can provide a useful training signal, their quality and diversity are not yet sufficient to fully replace large-scale real datasets. Substantial progress in generation fidelity and representativeness is still needed before synthetic-only pretraining can rival established baselines;
- **Effect of filtering:** Applying the proposed filtering procedure consistently improves results across all settings. This indicates that filtering effectively removes unrealistic or out-of-distribution samples, improving the quality of the learned representations;
- **Comparison with real data:** Pretraining on real thyroid images yields stronger results than synthetic-only pretraining. However, when filtering is applied, synthetic pretraining narrows the gap, especially in recall, where it nearly matches real data performance;
- **Hybrid pretraining:** Combining real datasets (testicle + thyroid) leads to the best overall performance, achieving the highest accuracy (86.78%) and F1-score (73.17%). This suggests that while synthetic data is useful, real data remains essential to have better performance.

In summary, synthetic data represents a promising alternative in low-data regimes, particularly when enhanced by filtering. While it cannot yet fully replace real data, it provides a robust pretraining signal that helps close gaps left by limited datasets, making it especially valuable in domains where access to real data is scarce or restricted.

6.4 Qualitative Evaluation and the Challenge of Markers

While quantitative results are essential, qualitative evidence provides further insights into the strengths and weaknesses of the proposed strategies. A common challenge in ultrasound analysis is the presence of superficial cues such as on-screen text and measurement markers. If not handled, models often overfit to these spurious signals rather than the anatomical features of interest.

As shown in Figure 6.1, visual explanations highlight how models behave in the presence of markers. The baseline model, trained without any augmentation, initially concentrates on clinically meaningful areas of the testicular tissue. However, when artificial markers are introduced, its attention shifts almost entirely to these superficial cues. This confirms that, without countermeasures, models are highly sensitive to such artifacts and may learn shortcuts that compromise generalization.

The proposed marker augmentation strategy in Section 5.3 addresses this issue by exposing the model to synthetic annotations during training. As a result, the network learns to disregard markers and instead consistently focus on the anatomical regions of interest. The qualitative evidence therefore complements the quantitative improvements reported earlier, demonstrating that marker augmentation may not always be beneficial for numerical performance but is essential to ensure clinically reliable predictions.

In summary, this analysis underlines the importance of explicitly handling superficial artifacts in ultrasound data. Without targeted augmentation, models risk misinterpreting irrelevant signals, whereas with marker-aware training they become more robust, interpretable, and better aligned with clinical needs.

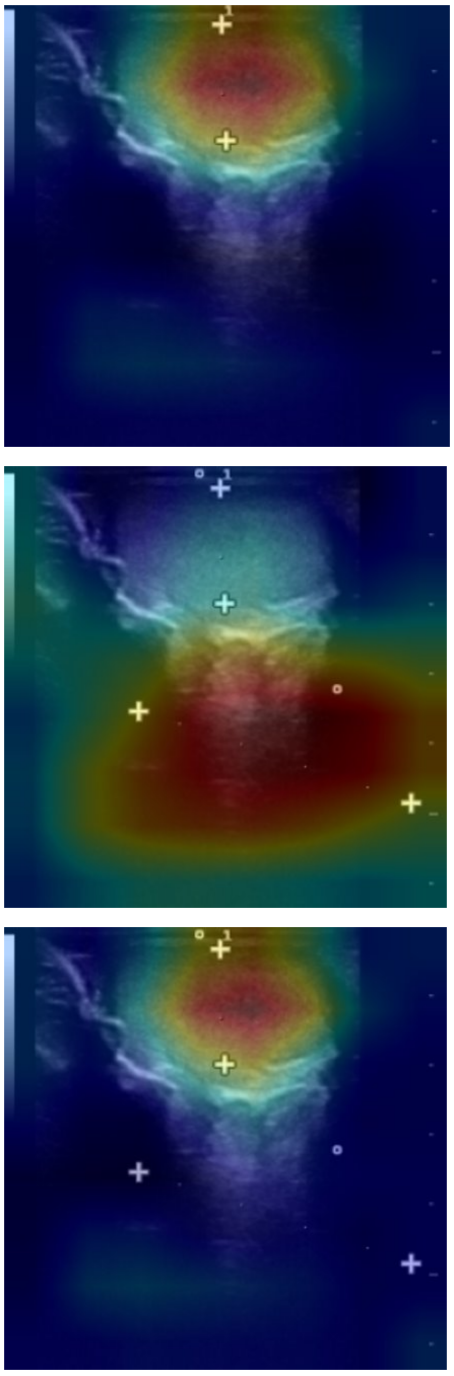


FIGURE 6.1: Grad-CAM++ [2] visualizations illustrating the impact of marker augmentation on model attention. **Top:** baseline model trained without marker-specific augmentation, focusing correctly on relevant regions of the image. **Middle:** when artificial markers are inserted at test time, the same model is misled, shifting its attention toward superficial annotations. **Bottom:** after training with marker augmentation to enforce marker invariance, the model restores its focus on the relevant regions despite the presence of markers.

Chapter 7

Conclusion and Future Research Directions

Male infertility is a pressing health issue worldwide, and Testicular Ultrasound (TUS) is a key non-invasive modality for its diagnosis. In particular, the evaluation of testicular parenchymal inhomogeneity has been suggested as a promising biomarker. Yet, the development of automated tools to support clinicians is hindered by several obstacles, including scarce annotated data, noisy labels, and the intrinsic variability of ultrasound imaging. The research presented in this thesis set out to address these challenges by combining pretraining strategies with synthetic data generation, aiming to enhance classification performance and bring automated TUS analysis closer to clinical practice.

7.1 Summary of Contributions and Key Findings

The work unfolded across several complementary directions. First, in Chapter 3 we addressed the issue of noisy labels, proposing a heuristic filtering method that successfully identified and corrected problematic samples. This step proved crucial in stabilizing model training and improving downstream performance. In Chapter 4 we then explored the potential of Denoising Diffusion Probabilistic Models for generating synthetic testicular ultrasound images. The introduction of a filtering strategy based on the precision metric ensured that the generated

data faithfully reflected the real distribution, ultimately yielding a curated synthetic dataset suitable for training.

Building on these resources, Chapter 5 investigated pretraining strategies designed to exploit both real and synthetic data. By combining supervised and contrastive self-supervised components, the proposed semi-supervised pretraining scheme produced representations that transferred effectively to the downstream classification of testicular inhomogeneity. The experiments presented in chapter 6 confirmed the impact of these approaches: pretraining consistently improved accuracy and F1-score over training from scratch, and the integration of filtered synthetic data further reinforced the model's robustness. Moreover, the introduction of augmentation techniques, such as the insertion of synthetic markers, enhanced the resilience of the classifier to common clinical artifacts.

7.2 Limitations

Although these results are encouraging, some limitations must be acknowledged. The labeled dataset, comprising only 880 images, inevitably restricts the model's ability to generalize. Furthermore, the reliance on static screenshots does not fully capture the dynamic evaluation process on which clinicians base their diagnoses. Finally, the experiments were conducted using a single-center dataset, and thus the findings remain to be validated across different institutions and imaging devices.

7.3 Future Research Directions

These limitations naturally open the way for future research. A first direction concerns the integration of temporal information through dynamic ultrasound video analysis, which would bring automated predictions closer to real-world diagnostic practice. Another promising avenue is the development of label-conditioned diffusion models, capable of producing synthetic data tailored to specific clinical categories. Such an approach could enhance both data augmentation and the interpretability of generative pipelines. Equally important is the extension of these experiments to multi-center datasets, which would provide a more rigorous test of the robustness and generalizability of the proposed methods. Finally, the strategies

developed here hold potential far beyond testicular ultrasound, and could be applied to other ultrasound modalities where data scarcity is a persistent obstacle.

7.4 Closing Remarks

In conclusion, this thesis has demonstrated that carefully designed pretraining strategies, when combined with synthetic data generation and noise-aware dataset curation, can substantially advance automated testicular ultrasound classification. While challenges remain, the contributions of this work offer a foundation upon which future research can build, ultimately moving closer to clinically reliable and widely applicable AI-assisted tools in medical imaging.

Chapter 8

Other Works — MICCAI 2025 UUSIC Challenge

8.1 Introduction & Background

The *Universal UltraSound Image Challenge* (UUSIC), organized within MICCAI 2025, was conceived as a large-scale benchmark for ultrasound image analysis across multiple anatomical regions. Unlike many prior challenge settings restricted to a single organ, UUSIC provides a unified dataset spanning several organs, each accompanied by classification labels and pixel-level segmentation masks. This design enables participants to develop methods that are consistent across anatomies, while still allowing conditioning on organ identity.

The competition defines two main tasks: classification of clinical pathologies and segmentation of anatomical regions of interest. In practice, the organ label is assumed to be known at inference time, so models may explicitly condition on it. The classification task requires identifying the presence or absence of a pathology within the given organ, and the segmentation task focuses on generating accurate masks of the anatomical region affected by the pathology. This joint setup emphasizes both predictive robustness and spatial consistency across different organs.

Performance is measured on a held-out test set to ensure fairness. For classification, primary metrics include accuracy and Area Under Curve(AUC) score, meanwhile for segmentation,

Dice Similarity Coefficient(DSC) is the main metric, complemented by Normalized Surface Distance (NSD). By combining multiple organs, shared benchmarks, and a multi-task protocol, UUSIC 2025 offers a rigorous environment to assess how well algorithms generalize across tasks, anatomies, and acquisition conditions.

Because we have been developing methods for ultrasound analysis as part of this thesis, participating in UUSIC represents a natural extension of our work into a more diverse and challenging setting. To contextualize our approach, it is instructive to review the prior art that directly inspired the challenge, particularly the *UniUSNet* [21] framework. Notably, UniUSNet was developed by the same research group that later organized the UUSIC benchmark, serving as its conceptual and methodological precursor. Understanding its architectural design and practical limitations is therefore essential to motivate the enhancements introduced in our proposed approach.

The UniUSNet framework proposes a unified architecture for joint disease classification and tissue segmentation across multiple ultrasound anatomies. Its key contributions and design choices can be summarized along three axes: data scale, architectural design, and adaptation strategy.

UniUSNet is trained on a dataset referred to as BroadUS-9.7K, consisting of around 6.9 K ultrasound images and 9.7 K associated annotations from seven anatomical positions (breast, cardiac, liver, thyroid, head, kidney, appendix). Although this is a substantial collection compared to very small medical sets, within each anatomical subset the number of images and masks remains constrained, especially for segmentation tasks in less frequently represented organs. Consequently, the model must cope with class imbalance, domain shift across organs, and limited examples per organ–task combination; in a sense, UniUSNet works in a moderate-data regime. In contrast, the UUSIC challenge dataset is more extensive and heterogeneous, pushing the limits of model generalization and scalability beyond what UniUSNet was originally evaluated on.

At its core, UniUSNet adopts a modified Swin-UNet backbone: a shared encoder with two parallel decoders, one for segmentation and one for classification. The segmentation decoder applies upsampling and skip-connections to recover spatial details; the classification branch is shallower, omitting upsampling to preserve computational efficiency. To allow a single network to handle multiple organs, tasks, and input types, UniUSNet introduces a prompting mechanism: four categories of prompts (nature of the input, anatomical position, task, and

input format) embedded into each transformer layer. These prompt embeddings are injected into the model to condition the feature computations dynamically, instead of having entirely separate heads or architectures per organ/task.

UniUSNet training follows a curriculum learning schedule: the segmentation task is introduced first to allow the model to learn strong structural representations of anatomical shapes, and only later the classification supervision is introduced. To mitigate imbalances across organs, the authors apply sample reweighting or balancing strategies across anatomical subsets.

In summary, UniUSNet establishes a strong baseline for universal ultrasound modeling by combining shared architectures, conditional prompting, and a hybrid training schedule. However, its constraints in data scale, organ coverage, and adaptability provide clear motivation to explore alternative architectures, regularization strategies, and training protocols for the more demanding UUSIC benchmark. In the next chapter sections, we will present our own methodological innovations, designed precisely to address these gaps and push further in generalization across organs and tasks.

8.2 Data

The public UUSIC dataset (as released via Zenodo)[1] comprises 7 anatomical organs, with a total of 8,736 ultrasound images of which 6,614 with segmentation masks and 4,945 with classification labels. Each image comes with the organ label (which is provided to the model as a conditioning signal), and may also have a binary (or multi-class) pathological label for classification, together with a pixelwise segmentation mask delineating the anatomical region of interest. From these segmentation masks, we additionally derived bounding boxes tightly enclosing the annotated regions. These bounding boxes served as auxiliary annotations, enabling later experiments where models were encouraged to attend more strongly to the anatomical ROI during training and evaluation.

Because the dataset covers multiple organs, i.e., liver, breast, thyroid, kidney, cardiac, head, and appendix, the number of samples per organ and per task varies. Some organs are better represented in terms of image count and mask availability, while others have fewer annotated examples or more challenging imaging conditions (e.g. low signal-to-noise ratio, greater

[1] `https://zenodo.org/records/15094669`

variation in shape). The multi-organ aspect introduces domain shift between organs: texture, echo patterns, noise characteristics, and anatomical geometry differ substantially across organ types, as illustrated in Figure 8.1.

Within the dataset, segmentation labels define a region of interest representing a whole pathology, and the classification labels correspond to presence/absence (or categories) of clinical pathology in that organ.

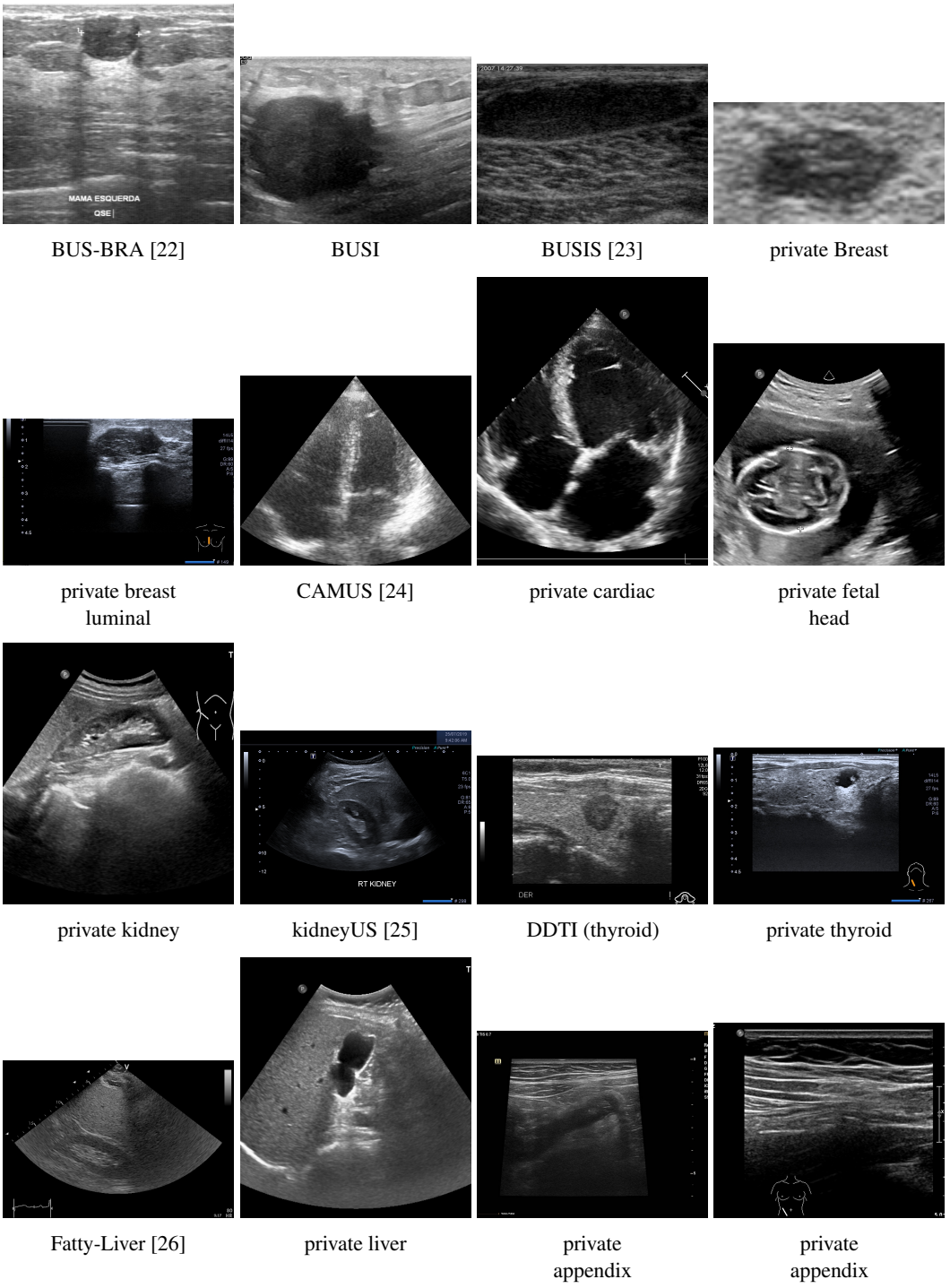
The hidden test set used for evaluation is withheld by the challenge organizers, and participants submit predictions via the Codabench platform[2] under standard evaluation metrics (DSC, NSD for segmentation; accuracy, AUC for classification). This setup ensures a fair comparison among submissions.

Preprocessing

In order to standardize the input data and remove irrelevant contextual information, we designed a deterministic preprocessing pipeline based on connected component analysis. The raw ultrasound images often include black borders, scale markers, or other acquisition-specific artifacts that can bias the training process. To address this, the preprocessing consists of the following steps:

1. **Binarization:** Each image is thresholded to obtain a rough mask of the ultrasound field of view;
2. **Connected component extraction:** The largest connected component is identified, corresponding to the actual ultrasound echogenic region;
3. **Cropping:** A tight bounding box around this component is computed and used to crop the image, removing unnecessary black borders and metadata;
4. **Resizing:** The cropped image is resized depending on the downstream task:
    - For the **segmentation task**, images are resized to a fixed resolution of 512×512 pixels, preserving fine spatial details needed for pixel-level predictions;

[2] `https://www.codabench.org/competitions/9106/`



BUS-BRA [22] | BUSI | BUSIS [23] | private Breast

private breast luminal | CAMUS [24] | private cardiac | private fetal head

private kidney | kidneyUS [25] | DDTI (thyroid) | private thyroid

Fatty-Liver [26] | private liver | private appendix | private appendix

FIGURE 8.1: Examples of ultrasound images from the UUSIC dataset, collected from multiple public and private sources. The grid illustrates the diversity of anatomical regions and acquisition conditions across breast, cardiac, fetal head, kidney, thyroid, liver, and appendix datasets.

- For the **classification task**, images are resized to 256×256 pixels, which balances computational efficiency with the ability to capture global parenchymal patterns.

This procedure ensures a deterministic and reproducible preprocessing step that retains only the clinically relevant regions of the scan. An example of the cropping pipeline is illustrated in Figure 8.2.

Postprocessing

After model inference, a lightweight postprocessing step was applied to further refine the predictions, an example is reported in Figure 8.3. Depending on the specific challenge task, this step included:

- **Probability thresholding:** Model outputs were calibrated into binary predictions by applying an empirically chosen threshold on the sigmoid probabilities. The optimal threshold was selected on the validation set to balance sensitivity and specificity;
- **Sanity checks:** A connected component labeling (CCL) analysis was performed on the predicted masks, and minor components were filtered out using a threshold proportional to the size of the largest connected component to ensure anatomical plausibility.

This postprocessing stage ensured that the final outputs were reliable and consistent, improving their suitability for downstream clinical applications.

8.3 Methodology

Our solution to the UUSIC challenge is a multi-task framework that combines a FiLM [27]-UNet for segmentation with a CBAM [28]-ResNet for classification. The two branches are coupled through a controlled feature fusion mechanism, as summarized in Figures 8.4, 8.5, and 8.6.

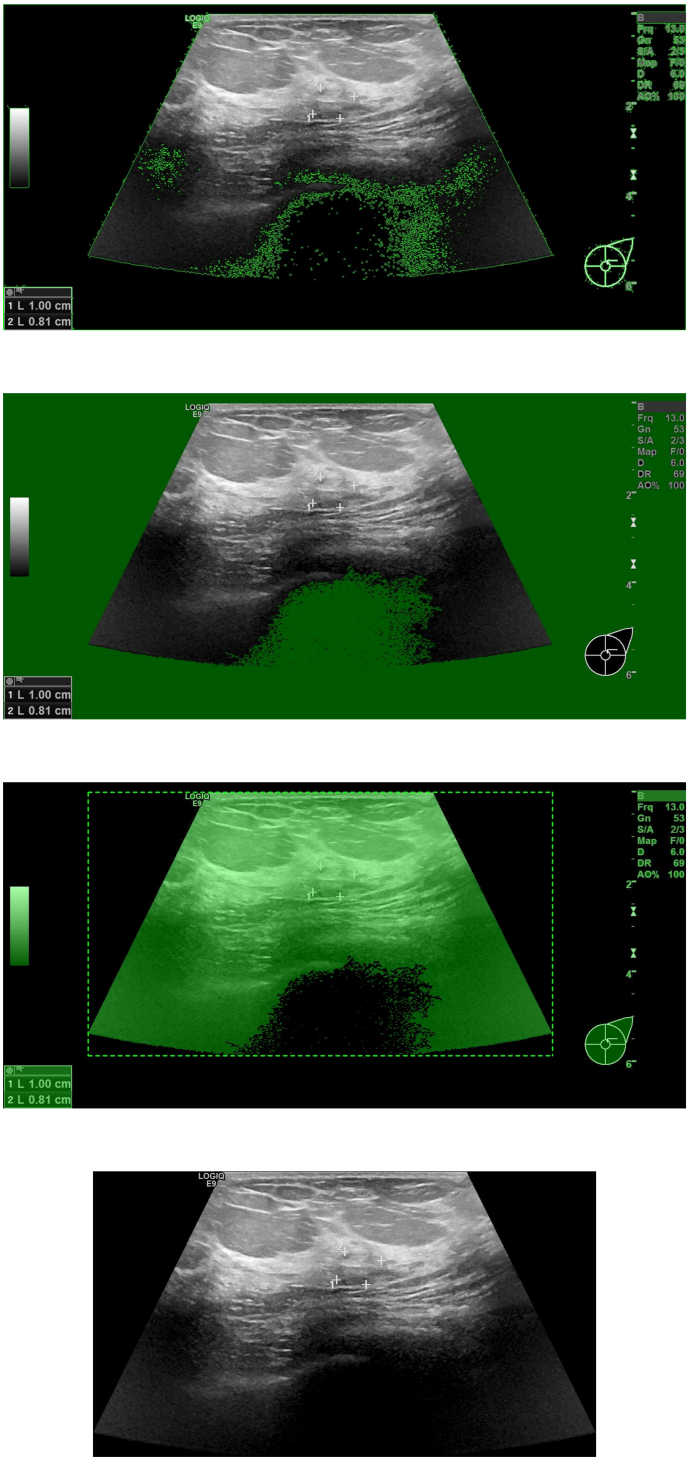


FIGURE 8.2: Deterministic cropping pipeline based on connected component analysis. Steps: (1) raw ultrasound image, (2) connected component extraction, (3) bounding box computation, (4) final cropped image.

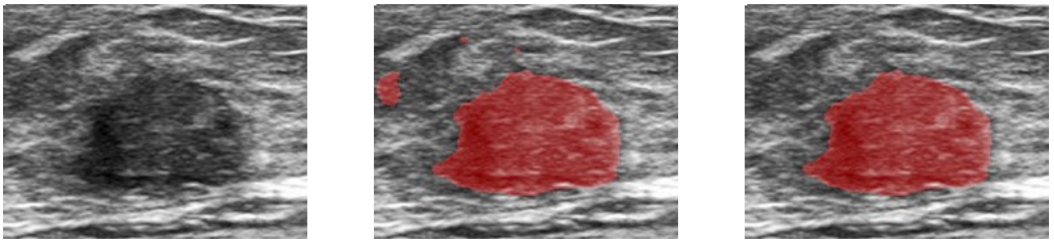


FIGURE 8.3: Illustration of the application of the postprocessing algorithm to a generated mask.

Model Architecture

The segmentation backbone is a five-scale 2D U-Net operating on 512×512 inputs. Each encoder block is composed of two 3×3 convolutions with stride 1 and padding 1, each followed by instance normalization, LeakyReLU activation, and Feature-wise Linear Modulation (FiLM) [27] conditioning. Downsampling is performed with 2×2 max pooling, while the decoder mirrors this structure with transposed convolutions and skip connections. A final 1×1 convolution produces the segmentation logits. Organ-aware conditioning through FiLM layers enables the model to adapt to different target organs by modulating intermediate features with learned scale and shift parameters. This design is particularly relevant in the UUSIC context, where the dataset covers multiple organs with diverse textures and geometries. Without conditioning, a standard U-Net would risk learning a compromise representation that fails to generalize across domains, whereas FiLM conditioning allows the model to specialize feature extraction depending on the organ identity while keeping the parameter overhead low.

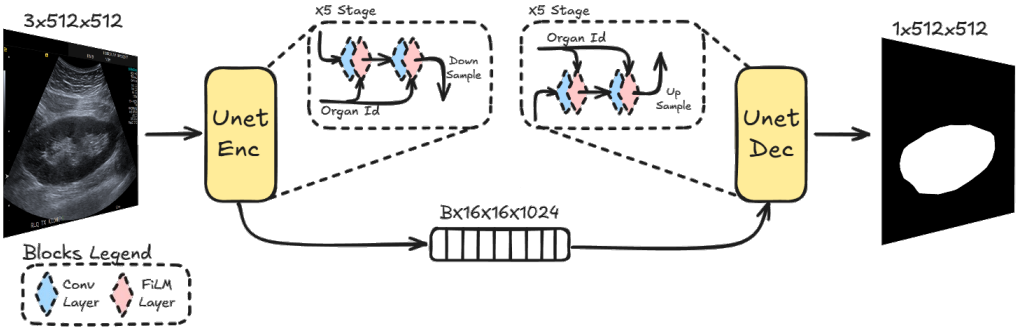


FIGURE 8.4: FiLM-UNet architecture for segmentation. Each encoder block consists of two 3×3 convolutions, each followed by instance normalization, LeakyReLU, and FiLM conditioning.

For classification, we employ a lightweight ResNet-18 backbone enhanced with Convolutional Block Attention Modules (CBAM) [28] added after every ResNet layer. The CBAM modules sequentially apply channel and spatial attention, guiding the model to focus on discriminative anatomical structures while suppressing acquisition-specific artifacts such as markers or overlays. This choice is motivated by the well-known difficulty of ultrasound classification, where models can easily overfit to irrelevant cues instead of capturing pathology-related patterns. By explicitly reweighting features in both the channel and spatial domains, CBAM encourages the network to attend to subtle tissue characteristics and suppress spurious correlations, thereby improving robustness and interpretability. Moreover, CBAM adds minimal computational cost, making it an attractive option for a challenge setting where efficiency is important.

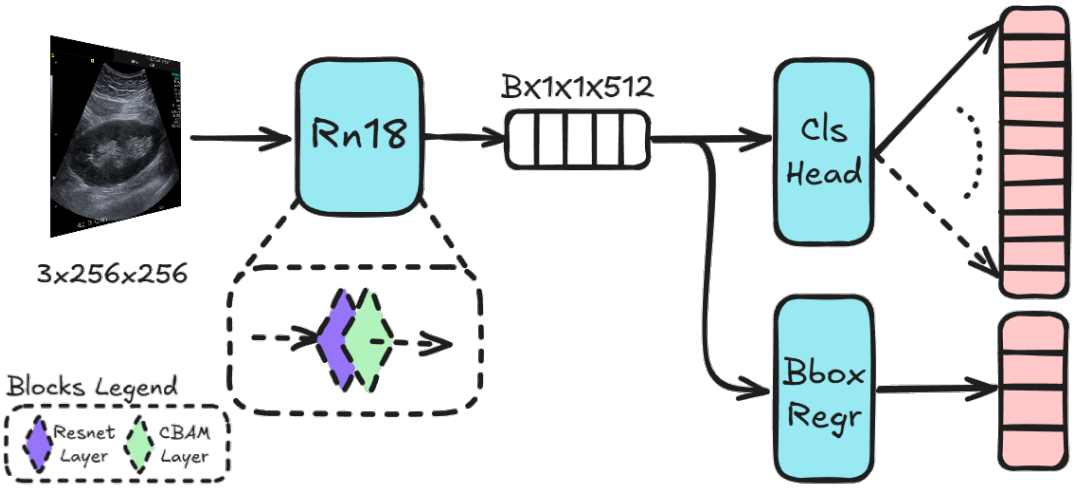


FIGURE 8.5: CBAM-ResNet18 classification architecture. CBAM modules refine intermediate feature maps by applying channel and spatial attention.

The segmentation and classification branches are integrated through a tanh-gating fusion mechanism. Features extracted by the ResNet-18 encoder are linearly projected to match the dimensionality of the U-Net bottleneck and fused as:

$$F_{\text{fused}} = F_{\text{UNet}} + \tanh(\alpha) \odot F_{\text{ResNet}},$$

where α is a learnable gating parameter initialized at zero. This mechanism enables the segmentation decoder to benefit from high-level semantic cues provided by the classifier without degrading its ability to localize fine structures. In practice, this controlled fusion

allows the segmentation branch to remain precise while exploiting classification-derived context that reinforces the delineation of pathological regions.

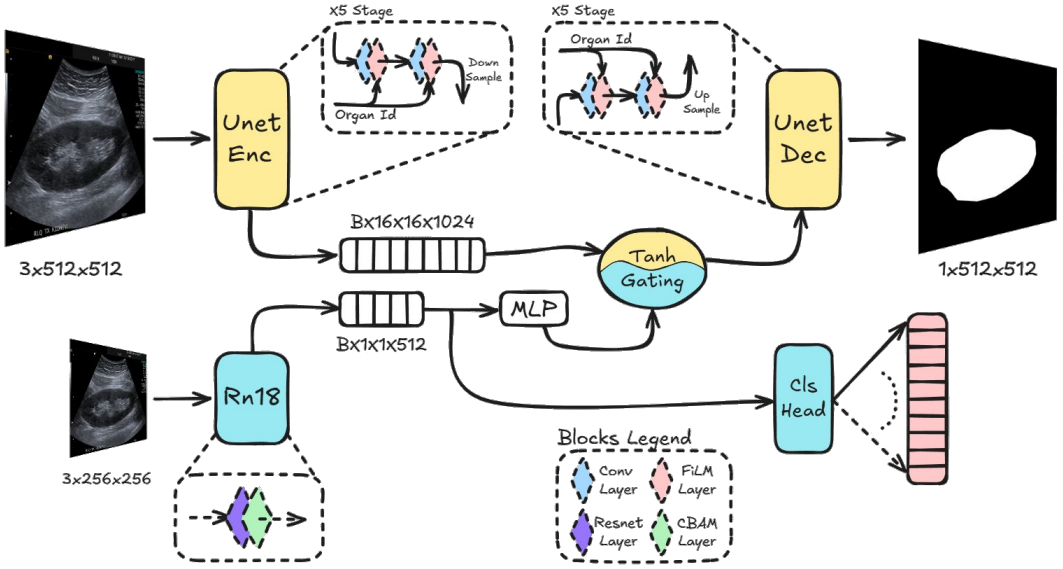


FIGURE 8.6: Fusion strategy with tanh gating. CBAM-ResNet18 features are projected and fused with the FiLM-UNet bottleneck to inject semantic context into the segmentation branch.

8.4 Experiments & Results

To evaluate the effectiveness of the proposed multi–task framework, we conducted a series of experiments on the UUSIC dataset, encompassing both the classification and segmentation tasks across multiple organs. Our main objectives were to (i) quantify the contribution of each architectural component (FiLM conditioning, CBAM attention, cropping, and postprocessing), (ii) assess cross–organ robustness under heterogeneous acquisition conditions, and (iii) validate the efficiency of the proposed lightweight design compared to more complex foundation models.

Experimental Setup

We followed the official UUSIC challenge protocol, splitting the public subset into training and validation partitions while keeping the private portion unseen for model selection. Evaluation

on the hidden test set was performed through the Codabench platform under the standard metrics specified by the organizers: Accuracy / AUC for classification and Dice Similarity Coefficient (DSC) / Normalized Surface Distance (NSD) for segmentation.

Training was performed sequentially for the two branches.

- **Classification phase**: The CBAM–ResNet18 branch was first pretrained for 20 epochs on all public data using a multi–objective loss combining Focal Loss, Cross–Entropy, and Smooth L1 Loss (for auxiliary bounding–box regression). Fine-tuning on the private subset was then carried out for 5 epochs with a reduced starting learning rate (1×10^{-5}) and cosine scheduling, ensuring smooth convergence and preservation of pretrained features;
- **Segmentation phase**: The FiLM–UNet was trained from scratch while freezing the classification branch; the bottleneck features were fused through the tanh-gating mechanism (Fig. 8.6). Optimization relied on a composite loss $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}$ using stochastic gradient descent (initial learning rate 2×10^{-2}, momentum 0.9) and a LambdaLR scheduler.

During training, images were normalized in intensity and augmented with random flips, rotations, elastic deformations, and gamma variations.

Classification Results

Table 8.1 reports quantitative results for the main ablation study. U-Net encoder features alone proved sub-optimal for classification, as their local receptive fields lack sufficient global semantic context. By contrast, the dedicated ResNet-18 classifier substantially improved performance, confirming that segmentation-oriented encoders are ill-suited to global decision tasks.

The inclusion of CBAM attention further enhanced the overall accuracy to 60.7%, the highest among all tested variants. CBAM's dual channel-and-spatial attention effectively redirected the network toward discriminative parenchymal structures while suppressing spurious cues such as on-screen markers or acquisition labels.

Model	CBAM	FiLM	Appendix	Breast	Breast Lum.	Liver	Overall
U-Net [14]	✗	✗	47.3%	51.2%	17.3%	46.7%	40.6%
	✗	✓	48.8%	46.5%	26.9%	42.4%	41.2%
ResNet-18 [6]	✗	✗	61.2%	62.0%	9.3%	52.4%	46.2%
	✗	✓	42.7%	67.8%	61.9%	51.8%	56.1%
	✓	✗	**61.2%**	**68.7%**	**59.3%**	**53.4%**	**60.7%**
	✓	✓	61.2%	65.3%	65.7%	47.1%	59.8%

TABLE 8.1: Classification performance of different strategies on the UUSIC dataset.

FiLM conditioning yielded moderate improvements in certain organ-specific subsets; however, since some of these (e.g., breast luminal) are not included in the public dataset, the effect is not particularly meaningful. In addition, combining FiLM with CBAM did not lead to further gains, suggesting that explicit FiLM-based modulation is less important for global classification, where a simple multi-class formulation already provides sufficient organ separation.

Qualitative inspection supported these findings: activation maps from CBAM-ResNet highlighted coherent anatomical regions, whereas the plain ResNet often attended to irrelevant peripheral annotations. Overall, the lightweight CBAM-ResNet18 provided a robust trade-off between accuracy, interpretability, and efficiency, forming the backbone for subsequent multi-task fusion.

Segmentation Results

The outcomes of our segmentation experiments are summarized in Table 8.2. What clearly emerges from these results is that each of the three components introduced into the segmentation pipeline, namely FiLM conditioning, deterministic cropping, and the lightweight postprocessing stage, contributes in a complementary way to the overall performance.

Starting from the baseline U-Net without any of these modifications, the network already delivered reasonable masks across all organs, but performance was uneven and often limited by cross–organ confusion or by the presence of acquisition artifacts in the input. Introducing FiLM conditioning immediately brought noticeable improvements in both Dice and surface

distance metrics. By modulating intermediate representations according to organ identity, the model became better at preserving sharp boundaries and avoiding leakage of irrelevant structures, a particularly visible effect in the cardiac and kidney subsets.

Cropping the input images also played a surprisingly important role. Many ultrasound frames contain broad black borders, on–screen annotations, or machine–specific overlays, which provide no diagnostic value but nevertheless bias the learning process. By removing these non-informative regions and presenting the network only with the relevant echogenic sector, segmentation accuracy improved, especially for anatomies such as breast and kidney, where background clutter is more prevalent. In practice, this simple preprocessing step made the learning problem more homogeneous and reduced the variance of predictions across different acquisition conditions.

The addition of a minimal postprocessing stage further refined the masks. Although the gains may look modest when expressed in aggregate metrics, they were consistently observed across organs: small isolated blobs and implausible edge artifacts were removed. This effect was particularly evident in fetal head and thyroid cases, where raw predictions tended to fragment into multiple disconnected components.

Configuration			**DSC/NSD (%)**					**Average**	
FiLM	**Crop**	**Postproc**	Breast	Cardiac	Thyroid	Fetal	Kidney	DSC	NSD
✗	✗	✗	81.1/15.1	81.1/5.2	62.7/6.9	92.0/9.4	81.5/9.4	79.7	9.2
✗	✓	✗	82.4/17.0	81.7/**6.0**	67.1/8.8	91.4/9.7	85.2/10.3	81.6	10.4
✓	✗	✗	83.9/18.0	82.9/5.4	66.2/8.5	93.2/11.2	86.2/11.3	82.5	10.9
✓	✗	✓	84.0/**18.3**	83.0/5.9	67.1/9.0	93.3/11.5	88.6/11.8	83.2	11.3
✓	✓	✗	84.2/18.0	**83.3**/5.8	67.2/9.2	93.0/11.3	88.1/12.0	83.1	11.2
✓	✓	✓	**85.5/18.3**	83.2/5.8	**67.8/9.7**	**93.7/12.3**	**88.7/12.1**	**83.8**	**11.6**

TABLE 8.2: Segmentation performance for different ablations on the UUSIC dataset.

When all three strategies were combined, the system reached its best configuration, with an average Dice of 83.8% and an NSD of 11.6%. Beyond the numerical gains, qualitative inspection confirmed that masks were sharper, cleaner, and closer to clinical expectations. In short, FiLM conditioning improved the representational capacity of the network, cropping simplified the input space by removing confounding artifacts, and postprocessing enforced structural

plausibility. Together, these elements produced a lightweight yet robust segmentation pipeline, well suited to the heterogeneous and multi-organ nature of the UUSIC challenge.

8.5 Discussion & Conclusion

The UUSIC benchmark provided a rigorous testbed for multi-organ and multi-task ultrasound modeling. Within this setting, our proposed pipeline combined a FiLM-UNet for segmentation with a CBAM-ResNet18 for classification, together with a lightweight tanh-gated fusion mechanism. This design proved to be both strong and efficient. On our validation splits, the best classification variant (CBAM-ResNet18) reached an overall accuracy of 60.7%, while the full segmentation stack (FiLM conditioning, deterministic cropping, and minimal post-processing) achieved an average DSC of 83.8% and NSD of 11.6% across organs (Tables 8.1 and 8.2).

Qualitatively, FiLM reduced cross-organ interference and sharpened boundaries. Cropping suppressed scanner overlays and non-diagnostic borders, while post-processing eliminated small implausible blobs, improving surface conformity. These findings are consistent with the accompanying technical report for the challenge, where we also highlighted the role of CBAM in guiding the classifier away from acquisition markers and toward anatomically meaningful patterns.

Large Multimodal LMs for Ultrasound: MedGemma

Motivated by the rapid progress of medical multimodal Large Language Models (LLMs), we evaluated Google's *MedGemma* [29] as a candidate universal backbone. We explored both zero-shot prompting and lightweight fine-tuning (parameter-efficient LoRA adapters) with paired text–image inputs on UUSIC tasks.

In practice, results were consistently underwhelming. Zero-shot prompting produced unstable predictions that were sometimes semantically plausible but frequently misaligned with the ground-truth labels. LoRA fine-tuning did not resolve these issues; instead, it led the model to degenerate behavior where predictions collapsed toward always selecting a single class, severely limiting discriminative ability. Two aspects were particularly problematic. First,

UUSIC requires precise low-level spatial reasoning (such as tight pixel masks and echotexture cues) that current medical LMMs are not yet reliably capturing from single B-mode frames when supervision is limited. Second, inference and memory footprints were substantially larger than for our tailored CNNs, reducing throughput without any compensating accuracy gains. For this benchmark and the available supervision, a compact and well-regularized architecture therefore remained the best trade-off.

Promptable Foundation Segmentation: MedSAM

We also tested *MedSAM* [30] as a promptable segmentation backbone, conditioning the model with bounding boxes derived from ground-truth masks (and, in ablations, detector-produced boxes). Compared to our FiLM-UNet, MedSAM delivered slightly higher mask quality in several organs, improving the DSC by a few points on average in our internal validation.

However, this improvement came with a significant efficiency penalty. End-to-end inference was roughly $\mathbf{4/5}\times$ slower than our deployed architecture on identical hardware and batch size, due to heavier encoder–decoder stacks and prompt conditioning overhead. In the context of the challenge, which required fast, repeatable submissions and broad ablations, this latency gap outweighed the modest metric gains. For this reason, we prioritized the FiLM-UNet, which offered a better balance between accuracy and throughput after adding deterministic cropping and minimal morphological cleanup.

An overview of the MedSAM architecture is reported in Figure 8.7, which illustrates its encoder–decoder backbone and the prompt-based conditioning mechanism that differentiates it from classical U-Net designs.

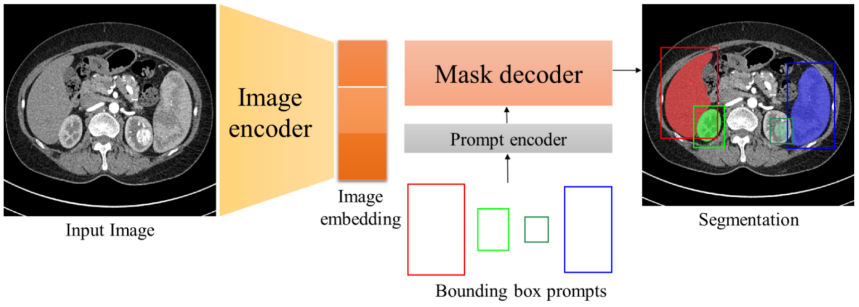


FIGURE 8.7: Overview of the MedSAM architecture. The model extends a heavy encoder–decoder backbone with prompt conditioning modules that inject external box or point information into the segmentation process.

Why the Lightweight Route Wins in this Setting

Three pragmatic considerations justified our final choice:

- **Data regime and supervision.** UUSIC's per-organ supervision is heterogeneous. Under such moderate-data conditions, architectures that enforce strong spatial inductive biases (such as U-Net) and lightweight attention mechanisms (such as CBAM) generalized more reliably than large pretrained models that were not explicitly tuned for B-mode artifacts;
- **Conditioning simplicity.** FiLM provided parameter-efficient organ-aware modulation throughout the hierarchy, mitigating inter-organ domain shifts without fragmenting capacity across many heads;
- **Throughput and maintainability.** The segmentation/classification split with controlled fusion was easy to train, ablate, and deploy. It sustained higher validation throughput and faster iteration cycles than MedSAM or Med-Gemma, which was crucial for fair model selection and robust early stopping in a challenge pipeline.

Final Conclusion

In conclusion, our UUSIC submission underscored a practical message and was validated by a strong empirical outcome, securing the **second position on the official leaderboard**. Under realistic annotation budgets and heterogeneous ultrasound sources, a carefully regularized and organ-conditioned CNN pipeline proved capable of outperforming substantially larger foundation models, both in terms of accuracy–throughput trade-offs and engineering simplicity. MedGemma and MedSAM remain promising research directions, yet in this setting their empirical benefits did not outweigh the added complexity and runtime costs. The adopted architecture therefore constitutes a robust and efficient baseline for universal ultrasound analysis in multi-organ, multi-task scenarios, and a solid platform for the incremental advances outlined above.

Appendix A

Publication

The research activity conducted in the context of this thesis has led to the submission of a paper to ICIAP 2025, the *18th International Conference on Image Analysis and Processing*. The paper was accepted and presented during the conference, which happened in Rome in September 2025:

N. Morelli, K. Marchesini, L. Lumetti, D. Santi, C. Grana, F. Bolelli. “Enhancing Testicular Ultrasound Image Classification Through Synthetic Data and Pretraining Strategies”, *ICIAP 2025*.

Enhancing Testicular Ultrasound Image Classification Through Synthetic Data and Pretraining Strategies

Nicola Morelli, Kevin Marchesini, Luca Lumetti, Daniele Santi, Costantino Grana, and Federico Bolelli ✉

University of Modena and Reggio Emilia, Italy
{*name.surname*}@unimore.it

Abstract. Testicular ultrasound imaging is vital for assessing male infertility, with testicular inhomogeneity serving as a key biomarker. However, subjective interpretation and the scarcity of publicly available datasets pose challenges to automated classification. In this study, we explore supervised and unsupervised pretraining strategies using a ResNet-based architecture, supplemented by diffusion-based generative models to synthesize realistic ultrasound images. Our results demonstrate that pretraining significantly enhances classification performance compared to training from scratch, and synthetic data can effectively substitute real images in the pretraining process, alleviating data-sharing constraints. These methods offer promising advancements toward robust, clinically valuable automated analysis of male infertility. The source code is publicly available at `https://github.com/AImageLab-zip/TesticulUS/`.

Keywords: Ultrasound · Medical Imaging · Synthetic · Diffusion Models

1 Introduction

Testicular UltraSound (TUS) imaging is a key non-invasive tool for evaluating male reproductive health by assessing tissue characteristics, such as parenchymal inhomogeneity, an emerging biomarker for male infertility [28]. However, subjective image interpretation and complex tissue patterns hinder reliable, standardized assessment, highlighting the need for automated classification tools.

Progress in this area is hampered by the lack of large, publicly available datasets. Ethical and privacy concerns limit data sharing, resulting in small, institution-specific datasets that constrain deep learning model development and hinder generalization. To address data scarcity, medical imaging research increasingly leverages pretraining [7] and synthetic data generation [25]. Supervised and self-supervised pretraining on large datasets enhances feature extraction and improves performance on smaller target datasets [18, 23]. Meanwhile, diffusion-based generative models [13] have emerged as powerful tools for synthesizing high-quality, realistic medical images, outperforming Generative Adversarial Networks (GANs) in stability and details modeling [24].

✉ Corresponding author: `federico.bolelli@unimore.it`

In this work, we evaluate supervised and self-supervised pretraining strategies for classifying testicular inhomogeneities using a ResNet-18 backbone. To address label noise common in ultrasound data, we propose a heuristic filtering method to improve training quality. Additionally, we explore diffusion-based synthetic data as a practical alternative to real images [4, 29], aiming to replicate the real data distribution and overcome data scarcity. Together, these strategies target improvements in both data quality and availability.

To summarize, our key contributions are: (*i*) a systematic evaluation of pretraining in testicular ultrasound analysis, (*ii*) a heuristic approach to reduce label noise, and (*iii*) the application of diffusion models for synthetic data generation in this sensitive, data-scarce domain.

2 Related Work

Deep Learning in UltraSound Image Analysis. Ultrasound (US) imaging is essential in medical diagnostics due to its safety, accessibility, and real-time capabilities, but interpretation remains challenging due to artifacts, noise, low contrast, and operator dependency [14]. Deep learning, particularly Convolutional Neural Networks (CNNs), has shown promise in automating analysis and extracting quantitative information [11, 17, 19, 20, 26].

However, US imaging presents unique challenges compared to modalities like MRI [21] or CT [3]. Limited annotated datasets, stemming from time-consuming, expert-dependent labeling, hinder model training, while heterogeneity and variability across devices and operators complicate generalization [30]. Privacy concerns further restrict data availability, impeding the development of robust models for applications like TUS analysis [27].

Generative Models for Synthetic Medical Image Generation. Generative models, particularly Generative Adversarial Networks (GANs) [10], have been used to alleviate data scarcity by augmenting datasets, performing cross-modality synthesis, and enabling anonymization [9, 15, 25], though they often suffer from training instability and limited diversity [2].

Recently, Denoising Diffusion Probabilistic Models (DDPMs) [13] emerged as a more effective alternative, achieving superior performance in generating realistic, diverse samples for MRI [24] and CT [22]. Conditional DDPMs further allow controlled generation based on clinical attributes or segmentation maps [29].

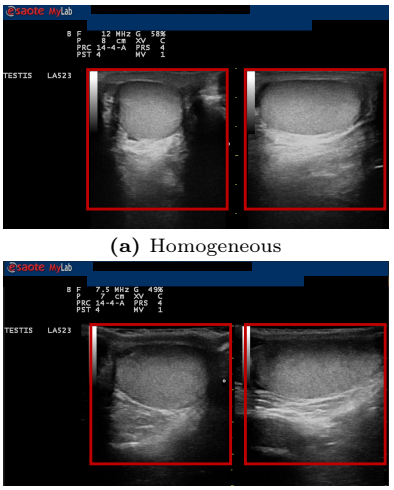
In this work, we address the underexplored domain of TUS classification by introducing the first benchmark targeting testicular pathology classification. To overcome data-sharing constraints, we demonstrate the effectiveness of DDPM-generated synthetic datasets when integrated into our pretraining pipeline.

3 Dataset Curation and Filtering

To the best of our knowledge, there is currently no publicly available dataset of testicular ultrasound images. Existing automatic approaches are primarily focused on testicular segmentation, and they typically rely on private datasets for training and evaluation [1]. For this reason, all the experiments presented

in this paper are conducted on an in-house dataset collected at the Antonio Nalin Center of the Baggiovara Hospital in Modena, Italy, using two different ultrasound acquisition systems: Esaote® MyLab25 Gold and Esaote® MyLab XPro80.[1]

The dataset includes image pairs, as illustrated in Fig. 1. Each pair contains static views of the same testicle, captured from transverse and sagittal planes. Pairs are cropped to remove metadata and isolate single views. Each view is treated separately, inheriting the original label.



(a) Homogeneous

(b) Inhomogeneous

Fig. 1: Example images provided by the clinical center. Red boxes indicate the regions selected during the cropping process.

Unlabeled Dataset (UD). While the primary focus of this work is on predicting the homogeneity versus inhomogeneity of testicular tissue, the dataset is enriched with thyroid ultrasound images, which are leveraged for pretraining purposes. Among the total 25 792 images, 1 664 correspond to testicular scans and 24 126 to thyroid scans, not necessarily from the same patients.

Labeled Dataset (LD). Additionally, for a subset of 880 testicular images, belonging to 220 patients, the inhomogeneity/homogeneity label is available, with a class distribution of approximately 20-80%. A significant challenge encountered during this project is the inherent noisiness in the pairing of images and labels. Ultrasound examinations are inherently dynamic, with clinicians relying on real-time video evaluation to assess anatomical properties. However, only static screenshots are saved during clinical practice. As a result, these images may not always accurately reflect the actual homogeneity characteristics of the tissue, introducing noise into the dataset.

Filtering Noisy Labels. To address this, we first developed an automatic filtering procedure applied to the 880 labeled images. This step aimed to identify and discard low-quality or misleading samples, thereby improving the reliability of the subsequent analyses and model training.

Leveraging a three-fold cross-validation schema, we train a simple ResNet-18 classifier for homogeneity classification based on the cross-entropy loss. Images from the same patient are always placed in the same fold to avoid data leakage. Results demonstrate poor classification performance and overfitting on training data. Particularly, it was clear that some of the examples were strongly perturbing the loss, indicating possible inconsistent labeling. A simple yet effective filtration schema has been adopted as follows. During model training, the per-sample loss values were recorded across epochs. Upon analyzing the loss trajectories, it was observed that most samples exhibited a near-monotonically

[1] Data will be anonymized and publicly released after receiving approval from the ethical committee. Download link: `https://ditto.ing.unimore.it/testiculus`.

decreasing loss trend. However, a subset of samples displayed highly irregular behavior, with spikes where the loss exceeded the value of 1. Empirically, a sample was flagged as "suspicious" if its training loss exceeded the threshold value of 1 on at least three[2] occasions during training. This evaluation process was repeated two times, leveraging ResNet-18 initialized with distinct pretrained weights, i.e., ImageNet and those provided by Chen *et al*. [7]. For each model, training was conducted using four random seeds and a three-fold cross-validation schema, resulting in a total of 24 runs. Due to the three-fold setup, each sample could be evaluated as "suspicious" between 0 and 16 times (i.e., appearing in multiple folds and seeds). Samples that were consistently flagged as "suspicious" in all 16 evaluations (72 images in total) were either discarded from the dataset or the corresponding label was flipped. Tab. 1 shows an improvement in model classification performance, confirming our hypothesis.

Table 1: ResNet-18 pretrained on ImageNet and fine-tuned on the *complete* LD dataset, *filtering* 72 "suspicious" images, or *flipping* their labels.

Dataset	Accuracy (↑)	F1-Score (↑)
Complete	81.51 ± 2.78	55.72 ± 4.37
Filtered	**88.15 ± 1.94**	68.59 ± 3.30
Flipped	86.78 ± 2.21	**73.17 ± 1.55**

Finally, a clinical evaluation was also performed. Clinicians were tasked to re-evaluate the labels of "suspicious" cases, this time using only the static views provided with the dataset since no video from the medical visit is available. Surprisingly, the evaluation was inconsistent with the original annotation, meaning that further study should rely not only on static images but on the entire video of the examination. All the experiments discussed in the rest of the paper were performed with the dataset resulting from such a polishing operation.

4 Methods

This section describes the strategies we propose to pretrain the classification model in a semi-supervised fashion and the process we leverage for generating and filtering synthetic ultrasound images. We also detail the neural network architectures evaluated, the fine-tuning procedure on the target classification task, and the evaluation protocol adopted to assess the synthetic data generation.

4.1 Pretraining Strategies

For the classification task, we selected a ResNet-18 architecture [11] as the backbone model.[3] It is widely recognized that, in such low-data regimes, models can benefit from pretraining strategies that enable better feature extraction [7]. Therefore, we explored effective approaches for pretraining the network to enhance its performance on the classification task (Fig. 2).

In contrast with classical ImageNet-based pretraining or other existing approaches [7], we investigated two different sources of data for pretraining:

– Real ultrasound images of the thyroid and testicular areas, using our UD dataset described in Sec. 3;

[2] Thresholds of four and five exceedances were also tested, but found to be less effective.

[3] Preliminary experiments showed that more complex architectures, such as ResNet-50 and Vision Transformers, tended to overfit, given the limited size of our LD dataset.

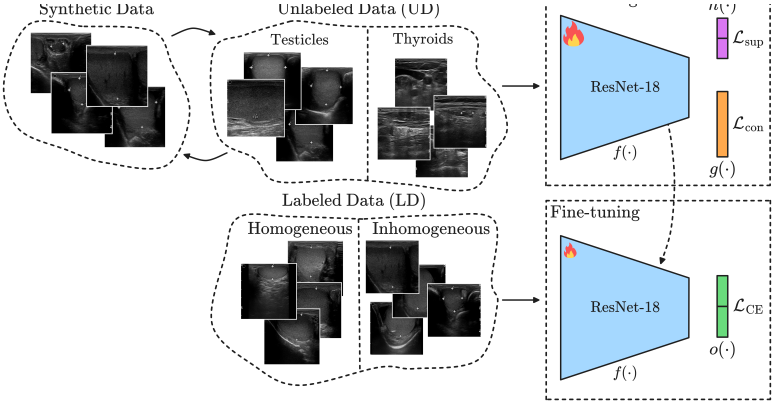


Fig. 2: Proposed pretraining leveraging synthetic or UD, and fine-tuning on the LD.

– Synthetic ultrasound images of testicles, generated using a diffusion model with a procedure to filter out-of-distribution samples, detailed in Sec. 4.2.

For the pretraining task, we employed a semi-supervised approach combining contrastive learning with supervised classification of the type of organ targeted in the ultrasound image (thyroid or testicle in our case). Specifically, we used the SimCLR framework [6] for the unsupervised contrastive component and a cross-entropy loss for the classification of the organ.

Contrastive Pretraining. SimCLR is a contrastive learning framework aimed at training an image encoder $f(\cdot)$ to produce representations that are invariant to image augmentations. This is achieved by maximizing the agreement between differently augmented views of the same image (positive pairs), while minimizing the agreement between views of different images (negative pairs).

Given a batch of N images $\{x_i\}_{i=1}^{N}$, we apply a stochastic augmentation pipeline twice to each image, resulting in two correlated views $(\tilde{x}_{2i-1}, \tilde{x}_{2i})$ per image. This effectively yields a batch of $2N$ augmented examples. Each augmented view $\tilde{x}_k$ is passed through a shared encoder $f(\cdot)$ followed by a projection head $g(\cdot)$, resulting in projected representations $z_k = g(f(\tilde{x}_k))$.

The similarity $\mathrm{s}(z_j, z_k)$ between any pair of representations is measured using cosine similarity, and the contrastive loss $\ell(j,k)$ for a positive pair is defined via a normalized temperature-scaled cross-entropy, as follows:

$$\mathrm{s}(z_j, z_k) = \frac{z_j^\top z_k}{\|z_j\| \, \|z_k\|}, \qquad \ell(j,k) = -\log \frac{\exp\left(\mathrm{s}(z_j, z_k)/\tau\right)}{\sum_{m=1}^{2N} \mathbb{1}_{[m \neq j]} \exp\left(\mathrm{s}(z_j, z_m)/\tau\right)}, \tag{1}$$

where τ is a temperature hyperparameter, and $\mathbb{1}_{[m \neq j]}$ is an indicator function equal to 1 when $m \neq j$, and 0 otherwise. Then the contrastive loss is computed by averaging over all positive pairs in the batch:

$$\mathcal{L}_{\text{con}} = \frac{1}{2N} \sum_{i=1}^{N} \left[\ell(2i-1, 2i) + \ell(2i, 2i-1)\right]. \tag{2}$$

Supervised Classification. To further enhance the learned representations, we incorporate a supervised classification objective during pretraining. For this

purpose, each image is annotated with a label corresponding to its anatomical region, i.e., *thyroid* or *testicle*, and a classification head $h(\cdot)$ is attached to the encoder $f(\cdot)$ to predict these labels. For each augmented view, we predicted the logits $c_k = h(f(\tilde{x}_k))$, and the supervised loss is computed using the cross-entropy across all pairs of augmented views:

$$\mathcal{L}_{\text{sup}} = \frac{1}{2N} \sum_{i=1}^{N} \big(\text{CE}(c_{2i-1}, y_i) + \text{CE}(c_{2i}, y_i) \big). \tag{3}$$

The final pretraining objective is a weighted combination of the contrastive and supervised losses:

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \lambda \, \mathcal{L}_{\text{sup}}, \tag{4}$$

where λ is set to 0.2 to balance the contribution of the supervised loss.

4.2 Synthetic Data Generation and Filtering

To address data scarcity in ultrasound imaging and overcome privacy-related data sharing constraints, we explored synthetic image generation for pretraining. A Denoising Diffusion Probabilistic Model (DDPM) was used to produce high-quality synthetic images as a substitute for real data. Specifically, we used the framework introduced by [8], which has demonstrated superior performance over GANs in image synthesis tasks.

The diffusion model operates through a two-phase process: a forward diffusion phase and a reverse denoising phase. In the forward phase, Gaussian noise is incrementally added to an image over multiple time steps, transforming a clean image into pure noise. This process is defined by a Markov chain, where each step adds a small amount of noise, controlled by a predefined variance schedule. In the reverse phase, a U-Net architecture is trained to reconstruct the original image by progressively removing the added noise. The model learns to predict the noise component and the diagonal covariance matrix of the noise distribution at each time step, allowing it to denoise the image iteratively. During inference, starting from a Gaussian noise sample, the model iteratively refines this noise through a series of denoising steps. At each step t, the model estimates the noise component $\epsilon_\theta(x_t, t)$ present in the current noisy image and computes a less noisy image x_{t-1}. This iterative process continues until the final step $t = 0$, resulting in a synthetic image x_0 that resembles the distribution of real ultrasound images. For our application, we trained the diffusion model on the LD dataset of real testicular ultrasound images.

Evaluation Metrics for Synthetic Images. We employed three established metrics to evaluate the quality of the generated images: improved *precision* and *recall*, both introduced by Kynkäänniemi *et al.* [16], and the *Fréchet Inception Distance* (FID) [12]. The precision assesses the fidelity of the generated images, quantifying the distributional similarity between real and generated data, while the recall measures the diversity of the synthetic data, indicating how much of the real data distribution is covered by the synthetic samples. FID, by comparing

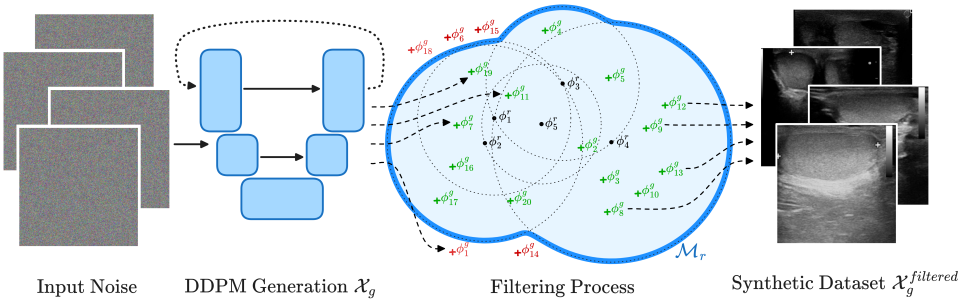


Fig. 3: Overview of the pipeline used for synthetic data generation and filtering.

both the mean and the covariance of the real and generated feature distributions, captures both two aspects.

The process to compute these metrics involves embedding both real and generated images into a high-dimensional feature space using a pretrained network (i.e., Inception). Let $\Phi_r = \{\phi_1^r, \phi_2^r, \ldots, \phi_N^r\}$ denotes the set of feature vectors for real images, and $\Phi_g = \{\phi_1^g, \phi_2^g, \ldots, \phi_M^g\}$ for generated images. For each real image feature vector ϕ_i^r we define an hypersphere $B(\phi_i^r, r_i)$ centered at ϕ_i^r, where the radius is the distance to its k-th nearest neighbor in Φ_r (symmetrically hyperspheres $B(\phi_j^g, r_j)$ are constructed around each generated sample ϕ_j^g using its k-th nearest neighbor in Φ_g).

Defining the real data manifold $\mathcal{M}_r$ (respectively, the generated data manifold $\mathcal{M}_g$) as the union of all the real data (generated data) hyperspheres:

$$\mathcal{M}_r = \bigcup_{i=1}^{N} B(\phi_i^r, r_i), \qquad \left(\mathcal{M}_g = \bigcup_{j=1}^{M} B(\phi_j^g, r_j)\right), \tag{5}$$

the precision P is computed as the fraction of generated samples whose embeddings fall inside the real data manifold $\mathcal{M}_r$, and the recall R is the fraction of real samples falling inside the generated data manifold $\mathcal{M}_g$:

$$P = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_{\mathcal{M}_r}(\phi_j^g), \qquad R = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\mathcal{M}_g}(\phi_i^r), \tag{6}$$

where $\mathbb{1}_{\mathcal{M}_r}(\phi_j^g)$ (respectively $\mathbb{1}_{\mathcal{M}_g}(\phi_i^r)$) is an indicator function which is 1 if $\phi_j^g \in \mathcal{M}_r$ ($\phi_i^r \in \mathcal{M}_g$), 0 otherwise.

The FID assumes that the feature vectors of real and generated data, extracted from the pretrained Inception network, follow multivariate Gaussian distributions. Let μ_r and Σ_r be the mean and covariance of the real image features Φ_r, and μ_g and Σ_g those of the generated image features Φ_g. The FID is defined as the Fréchet distance between these two distributions:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right). \tag{7}$$

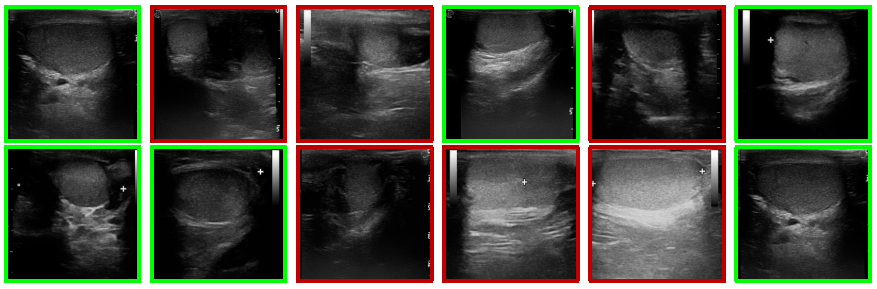


Fig. 4: Sample images from our generated dataset. Green represents high-quality samples, while red identifies those discarded by our filtering method.

A lower FID value indicates that the generated images are statistically more similar to the real ones, reflecting both high fidelity and appropriate variability in the synthetic data distribution.

Algorithm 1 Filtering Algorithm.

Input: $\mathcal{X}_g$, Φ_r, Φ_g, # of neighbors k.
Initialize $\mathcal{X}_g^{\text{filtered}} \leftarrow \emptyset$
for all ϕ_i^r in Φ_r **do**
  Compute $r_i \leftarrow$ Euclidean distance to the k-th nearest neighbor of ϕ_i^r in Φ_r
  $B(\phi_i^r, r_i) \leftarrow$ hypersphere centered at ϕ_i^r with radius r_i
end for
$\mathcal{M}_r \leftarrow \bigcup_{i=1}^{N} B(\phi_i^r, r_i)$
for all ϕ_j^g in Φ_g **do**
  if $\phi_j^g \in \mathcal{M}_r$ **then**
    Add x_j^g to $\mathcal{X}_g^{\text{filtered}}$
  end if
end for
return $\mathcal{X}_g^{\text{filtered}}$

Filtering Method. To ensure the quality of the generated synthetic ultrasound images, we developed a filtering method based on the abovementioned precision metric. Specifically, we computed the real data manifold, $\mathcal{M}_r$, from the feature embeddings of the LD dataset used to train the diffusion model, as specified in Eq. (5). Given the set of generated images $\mathcal{X}_g = \{x_1^g, x_2^g, \ldots, x_M^g\}$, we compute their corresponding feature representations, Φ_g, and filter them by selecting only those whose embeddings lie inside $\mathcal{M}_r$, obtaining the final filtered synthetic dataset $\mathcal{X}_g^{\text{filtered}}$ (Algorithm 1, Fig. 3).

Generation Results. Following [8], to compute precision and recall metrics on generated data, we set the number of neighbors $k = 3$ and leverage three non-overlapping reference batches of 64 real images sampled from the LD dataset. Instead, since the filtering process is based on the entire LD dataset, it leverages a $k = 50$. We found that k should scale approximately linearly with the number of real data employed in the computation of the manifold to maintain hyperspheres of comparable size across different settings.

Applying our filtering, we increased the precision of the synthetic dataset from 79.68 ± 4.81 to 90.1 ± 4.70. As a natural consequence, some generated samples were removed, leading to a reduction of the recall from 25.0 ± 1.56 to 11.9 ± 3.25. However, the variations in recall remained limited, and the FID decreased from 119.84 ± 2.39 to 116.84 ± 4.12, confirming that the filtering strategy achieved a good compromise between maintaining similarity to the real data distribution and preserving adequate coverage of the feature space. After filtering, the original $\sim 20K$ synthetic samples were reduced to $\sim 9K$.[4] Samples of generated images are available in Fig. 4.

[4] Filtered synthetic data are publicly released at `https://ditto.ing.unimore.it/testiculus`.

Table 2: Three-fold cross-validation results for ResNet-18 on the homogeneous and inhomogeneous downstream task, starting from different pretraining strategies.

Pretraining	# Samples	Accuracy (↑)	F1-Score (↑)	Precision (↑)	Recall (↑)
∼	∼	73.93 ± 6.80	56.05 ± 7.11	45.48 ± 9.91	75.12 ± 3.66
ImageNet	1.28M	83.89 ± 2.15	67.89 ± 1.85	61.72 ± 3.08	76.08 ± 4.11
USCL [7]	23.00K	75.46 ± 2.64	57.43 ± 1.63	47.37 ± 2.16	73.86 ± 4.76
UD (only testicles)	1.66K	81.09 ± 2.95	65.44 ± 2.32	55.86 ± 2.74	79.55 ± 2.79
UD (only thyroids)	24.13K	80.13 ± 2.83	63.92 ± 2.46	54.30 ± 3.87	78.88 ± 2.06
UD	25.79K	$\mathbf{86.78 \pm 2.21}$	$\mathbf{73.17 \pm 1.55}$	$\mathbf{67.84 \pm 1.67}$	$\mathbf{80.27 \pm 2.13}$

4.3 Fine-tuning

Fine-tuning was conducted on the LD dataset using a standard transfer learning setup. The pretrained ResNet-18 backbone was used as a feature extractor, and a lightweight classification head $o(\cdot)$ (Fig. 2), consisting of a single linear layer, was appended on top to perform binary classification. To prevent the network from focusing on superficial visual cues, such as spurious patterns or annotations, we added random synthetic markers to each image (Fig. 5) during training. This forced the network to learn more robust and generalizable features. In addition to marker insertion, we applied a series of spatial data augmentations, including random rotations, horizontal flipping, and small random shifts, to further improve generalization and mitigate overfitting. Finally, in order to mitigate class imbalance, we used a weighted sampler.

5 Experiments and Results

Implementation Details. The ResNet-18 backbone was pretrained on synthetic ultrasound data using supervised and unsupervised objectives (batch size 1024) on two NVIDIA L40S GPUs (48 GB), with the LARS optimizer and a polynomial learning rate schedule (initial value 10^{-3}). Inputs were normalized (mean 0.5, std. dev. 0.25). Fine-tuning for binary classification employed three-fold cross-validation over four random seeds, using a batch size of 64 on a single RTX 2080 Ti GPU. The backbone and classification head were fine-tuned with learning rates of 10^{-5} and 10^{-4}, respectively. Synthetic ultrasound images were generated using a diffusion model retrained from scratch on domain-specific data (batch size 16, 256×256 input resolution) with a single L40S GPU, following [8]. Both mean and variance were learned, and sampling was unguided.

On the Role of Pretraining. To validate the effectiveness of the proposed pretraining strategy, we applied it to ResNet using our UD dataset. For comparative analysis, we also considered ResNet pretrained on ImageNet and the ultrasound-specific pretrained weights provided by Chen *et al.* [7], devised for ultrasound data, although focused on lung and liver. In each of the three pretraining scenarios, the model was subsequently fine-tuned on our LD dataset following a three-fold cross-validation schema. All data from the same patient were strictly confined to the same fold to effectively prevent data leakage. To ensure a fair comparison and reliable convergence, the number of training steps

Table 3: Three-fold results on the downstream task when pretraining ResNet-18 with different combinations of real and synthetic data with ($\mathcal{X}_g^f$) or without ($\mathcal{X}_g$) applying the proposed filtering procedure.

Real Testicle	Real Thyroids	Synthetic Testicle	Accuracy (↑)	F1-Score (↑)	Precision (↑)	Recall (↑)
✗	✗	$\mathcal{X}_g$	80.58 ± 3.58	64.24 ± 2.31	55.46 ± 3.08	77.13 ± 1.70
✗	✗	$\mathcal{X}_g^f$	80.42 ± 2.71	64.77 ± 1.62	54.69 ± 3.24	80.12 ± 2.02
✓	✗	∼	81.09 ± 2.95	65.44 ± 2.32	55.86 ± 2.74	79.55 ± 2.79
✗	✓	$\mathcal{X}_g$	82.88 ± 2.11	67.61 ± 1.56	58.87 ± 1.71	79.59 ± 3.63
✗	✓	$\mathcal{X}_g^f$	84.60 ± 2.08	70.63 ± 1.20	62.19 ± 1.76	**82.17 ± 0.97**
✓	✓	∼	**86.78 ± 2.21**	**73.17 ± 1.55**	**67.84 ± 1.67**	80.27 ± 2.13

was held constant across all experiments, regardless of variations in the size of the pretraining dataset. This ensures that each model undergoes the same total number of forward and backward passes. The results are summarized in Tab. 2.

Our initial observation indicates that leveraging ultrasound-specific pretrained weights from Chen *et al.* (third row of Tab. 2) does not necessarily yield optimal performance, particularly when the pretraining data originates from different ultrasound acquisition systems, as in the case of the USCL dataset [7]. Conversely, when images are sourced from the same acquisition system, variations in the anatomical regions, namely testicles and thyroid, did not significantly impact performance. Specifically, downstream performance on testicular imaging was nearly identical regardless of whether testicular or thyroid images were employed during pretraining (fourth and fifth rows of Tab. 2). It is important to highlight that in these instances, pretraining leveraged exclusively the contrastive component of the loss in Eq. (4). Interestingly, the ImageNet-pretrained model achieved the best results, highlighting the superior generalization of features learned through supervised training on diverse natural images, even across domains as different as natural and ultrasound images.

Finally, combining ultrasound data from different anatomical structures, in this case testicular and thyroid images (last row of Tab. 2), enabled the integration of supervised and unsupervised losses, delivering the best overall performance. This strategy resulted in a noticeable improvement, increasing accuracy by approximately 3 points and the F1-score by about 6 points compared to the ImageNet pretrained baseline. An ablation study to compare the impact of loss weightings is also conducted in Tab. 4. Results show both losses affect performance, with $\lambda = 0.2$ yielding the best outcome.

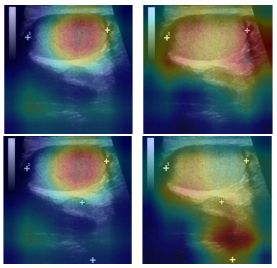
Table 4: Ablation study on using different loss components and varying λ, Eq. (4).

$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{sup}}$	λ	Accuracy (↑)	F1-Score (↑)
✓	✗	0.0	84.77 ± 2.56	69.41 ± 0.51
✓	✓	0.1	85.38 ± 2.71	71.40 ± 0.93
✓	✓	0.2	**86.78 ± 2.21**	**73.17 ± 1.55**
✓	✓	0.3	85.29 ± 2.68	71.36 ± 0.70
✓	✓	0.4	85.03 ± 2.43	71.22 ± 1.26
✗	✓	1.0	83.88 ± 2.56	70.11 ± 1.56

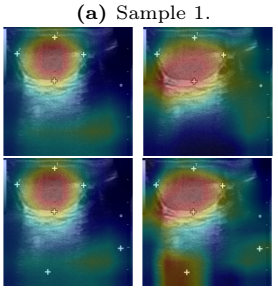
The Impact of Synthetic Data. Tab. 3 demonstrates that using filtered synthetic data ($\mathcal{X}_g^{\text{filtered}}$) consistently yields better performance compared to unfiltered synthetic data ($\mathcal{X}_g$). Although the highest accuracy is achieved when both real datasets are included, the performance of models trained solely on synthetic data remains competitive, exhibiting only a modest degradation. These findings underscore the practical utility of synthetic data,

particularly in scenarios where access to real data is restricted due to privacy concerns or availability limitations.

Qualitative Evaluation. A heatmap visualization generated using Grad-CAM++ [5] on two representative samples from the test set of our LD dataset is reported in Fig. 5. For each sample, four images are shown: outputs of the same model trained either **with** (left column) or **without** (right column) applying the augmentation strategies reported in Sec. 4.3. The second row for each sample differs from the first by the introduction of artificial markers ✣. As can be observed, the use of augmentation strategies helped focus the model's attention more precisely on the inner part of the testicle, the region most closely associated with the inhomogeneity property relevant to our downstream task. Furthermore, the proposed augmentation approach, which introduces synthetic random artifacts during training, proved effective in mitigating the influence of such artifacts on the predictions. This effect is particularly evident when comparing the bottom-left and bottom-right images of each sample.



(a) Sample 1.



(b) Sample 2.

Fig. 5: Grad-CAM++.

6 Conclusion and Future Research Directions

In this study, we addressed key challenges in the automated classification of testicular ultrasound inhomogeneity, a promising biomarker for male infertility. By combining supervised and unsupervised pretraining with diffusion-based synthetic augmentation, we achieved significant improvements over models trained from scratch or using pretraining strategies tailored for ultrasound imaging.

Future work will focus on incorporating dynamic ultrasound videos, which may offer richer contextual information for classification. We also aim to develop label-conditioned synthetic image generation to produce datasets suitable for both pretraining and fine-tuning. Advancing these directions will be essential for the next generation of automated, clinically deployable tools.

Acknowledgements. This project is funded by the University of Modena and Reggio Emilia and Fondazione di Modena through FAR-2024 (E93C24002080007) and FARD-2024, and by the Italian Ministry of Research's NRRP complementary actions "Fit4MedRob – Fit for Medical Robotics" (PNC0000007).

References

1. Abdalla, A.M., et al.: Automatic Segmentation and Detection System for Varicocele Using Ultrasound Images. CMC **72**(1) (2022)
2. Arora, S., et al.: Generalization and Equilibrium in Generative Adversarial Nets (GANs). In: ICML (2017)

3. Bolelli, F., et al.: Segmenting Maxillofacial Structures in CBCT Volumes. In: CVPR (2025)
4. Cartella, G., et al.: Unveiling the Truth: Exploring Human Gaze Patterns in Fake Images. IEEE SPL (2024)
5. Chattopadhay, A., et al.: Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: WACV (2018)
6. Chen, T., et al.: A Simple Framework for Contrastive Learning of Visual Representations. In: ICML (2020)
7. Chen, Y., et al.: USCL: Pretraining Deep Ultrasound Image Diagnosis Model Through Video Contrastive Representation Learning. In: MICCAI (2021)
8. Dhariwal, P., et al.: Diffusion Models Beat Gans on Image Synthesis. In: NeurIPS (2021)
9. Frid-Adar, M., et al.: GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification. Neurocomputing (2018)
10. Goodfellow, I.J., et al.: Generative Adversarial Nets. In: NeurIPS (2014)
11. He, K., et al.: Deep Residual Learning for Image Recognition. In: CVPR (2016)
12. Heusel, M., et al.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: NeurIPS (2017)
13. Ho, J., et al.: Denoising Diffusion Probabilistic Models. In: NeurIPS (2020)
14. Huang, Q., et al.: Machine Learning in Ultrasound Computer-Aided Diagnostic Systems: A Survey. BioMed Research International **2018**(1) (2018)
15. Kazeminia, S., et al.: GANs for Medical Image Analysis. AIME **109** (2020)
16. Kynkäänniemi, et al.: Improved Precision and Recall Metric for Assessing Generative Models. In: NeurIPS (2019)
17. Litjens, G., et al.: A survey on deep learning in medical image analysis. MedIA **42** (2017)
18. Lumetti, L., et al.: U-Net Transplant: The Role of Pre-training for Model Merging in 3D Medical Segmentation. In: MICCAI
19. Lumetti, L., et al.: Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. IEEE Access (2024)
20. Lumetti, L., et al.: Taming Mambas for 3D Medical Image Segmentation. IEEE Access (2025)
21. Menze, B.H., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE TMI **34**(10) (2014)
22. Pan, S., et al.: Synthetic CT Generation from MRI Using 3D Transformer-Based Denoising Diffusion Model. Medical Physics **51**(4) (2024)
23. Panariello, A., et al.: Consistency-Based Self-supervised Learning for Temporal Anomaly Localization. In: ECCV (2022)
24. Pinaya, W.H., et al.: Brain Imaging Generation with Latent Diffusion Models. In: MICCAI Workshop (2022)
25. Pollastri, F., et al.: Augmenting Data with GANs to Segment Melanoma Skin Lesions. MTAP **79**(21-22) (2019)
26. Porrello, A., et al.: Spotting Insects from Satellites: Modeling the Presence of Culicoides Imicola Through Deep CNNs. In: SITIS (2019)
27. Price, W.N., et al.: Privacy in the age of medical big data. Nature Medicine **25**(1) (2019)
28. Spaggiari, G., et al.: Testicular ultrasound inhomogeneity is an informative parameter for fertility evaluation. AJA **22**(3) (2020)
29. Wang, W., et al.: Semantic image Synthesis Via Diffusion Models. ArXiv Preprint arXiv:2207.00050 (2022)
30. Yap, M.H., et al.: Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. IEEE J-BHI **22**(4) (2017)

Appendix B

UUSIC Challenge Technical Report

This appendix contains the technical report prepared for the participation in the Universal UltraSound Image Challenge (UUSIC), hosted within MICCAI 2025. The report describes the methods and experiments carried out during the competition. The results obtained in this challenge led to a **2 placement**.

K. Marchesini, **N. Morelli**, C. Grana, F. Bolelli. "FiLM-UNet Meets CBAM-ResNet: A Lightweight Pipeline for the UUSIC Challenge", *MICCAI 2025*.

FiLM-UNet Meets CBAM-ResNet: A lightweight Pipeline for the UUSIC Challenge

Team Name: `AImageLab-zip`

Kevin Marchesini*, Nicola Morelli, and Federico Bolelli

University of Modena and Reggio Emilia, Italy
{*name.surname*}@unimore.it

1 Introduction

Deep learning has significantly advanced medical image analysis in recent years, particularly in segmentation and classification across several imaging modalities. Among the medical acquisition modalities, ultrasound (US) stands out as one of the most common due to its low cost, and real-time acquisition. At the same time, US poses distinctive challenges for automatic analysis: speckle noise and low signal-to-noise ratios, intensity nonuniformities, weak or ambiguous anatomical boundaries, and operator and scanner dependence that induce domain shifts. Over the past year, the community tried to mitigate these difficulties through the release of larger, more heterogeneous US datasets and the organization of benchmarks and challenges.

Within this context, the *Universal UltraSound Image Challenge* (UUSIC) has been conceived as a multi-task benchmark aimed at developing a single model capable of performing classification, and segmentation across multiple organs and diseases. This technical report presents our solution to UUSIC. Our contributions are threefold. First, we demonstrate that Feature-wise Linear (FiLM) [7] is an effective and parameter-efficient mechanism to adapt a single 2D U-Net [8] to segment multiple pathologies targeting different organs. Second, we investigate a lightweight ResNet-18-based classifier improved with Convolutional Block Attention Module (CBAM) [12] to perform classification, and the effective integration of its extracted features to boost the segmentation. Third, we explore using the UNet encoder's multi-scale features themselves to perform classification, yielding a compact, end-to-end multi-task model.

2 Model: Multi-Task FiLM–UNet with CBAM–ResNet

2.1 FiLM–UNet for Segmentation

The segmentation backbone is a five-scale 2D U-Net operating on 512×512 inputs. Each convolutional block consists of a 3×3 convolution with stride 1 and padding 1, followed by instance normalization and a LeakyReLU activation.

* AImageLab-zip Team Leader.

Downsampling is performed by 2×2 max pooling at each scale and mirrored by symmetric upsampling in the decoder via transposed convolutions with kernel size 2 and stride 2. Skip connections concatenate encoder features with decoder activations at matching resolutions, and a final 1×1 convolution maps the last decoder features to the desired number of segmentation logits.

Organ-aware conditioning is implemented with FiLM [7] layers applied after each convolutional block in both encoder and decoder. Given a feature tensor $x \in \mathbb{R}^{B\times C\times H\times W}$ and an organ identifier $o \in \{0, \ldots, N_{\text{org}}-1\}$, FiLM learns an affine transformation:

$$y \;=\; \gamma(o) \odot x \;+\; \beta(o),$$

where $(\beta, \gamma) \in \mathbb{R}^{B\times C}$ are learned by embedding the organ ID into a 64-dimensional vector and passing it through a small MLP that outputs $2C$ parameters. As suggested by the original paper, we initialize FiLM to the identity mapping (i.e., $\gamma = 1$ and $\beta = 0$), which stabilizes early training.

2.2 Classification Branch

For image-level classification, we tried two different strategies, one leveraging the U-Net encoder's features and a final head which process them, and another using a ResNet-based classification branch, then integrated with the U-Net decoder to improve also the segmentation results.

Leveraging Encoder Features for Classification. For the first strategy, we explored two different variants built on top of the segmentation encoder. In the first variant, we used the deepest encoder features by flattening them into a single dimension. This tensor is passed to a layer normalization and used to predict the class with a stack of two linear layers with GELU activation. In the second version, we tried a multi-scale approach by globally average pooling each channel of the last three encoder stages' features, and concatenating them. This embedding is mapped to classification output with a two-layer MLP with ReLU.

Classification Branch with CBAM Integration. Our second approach uses a dedicated ResNet-18 network for classification. To enhance feature extraction and address the issue of overfitting markers common in ultrasound image classification, we integrate a Convolutional Block Attention Module (CBAM) into the architecture. CBAM sequentially generates attention maps across both channel and spatial dimensions. This dual mechanism guides the network to focus on the most informative features and their optimal spatial locations.

2.3 Controlled Feature Fusion via Tanh Gating

Our multi-task learning approach integrates features from both the classification and segmentation branches to enhance overall performance. The U-Net is improved by incorporating rich, high-level semantic information from the ResNet-18 encoder. The fusion process begins by aligning the output feature maps of the ResNet-18 encoder to the U-Net's bottleneck layer, through a linear projection, which matches the dimensions of the classification features to the channel

dimensions of the U-Net's bottleneck embeddings. A crucial component of this fusion is a learnable gating parameter, denoted as α, which is initialized to a tensor of zeros. A hyperbolic tangent (tanh) function is applied to this parameter to produce the final gate values. As the model learns, the values of α are gradually adjusted, allowing the segmentation model to leverage valuable semantic context without compromising its ability to perform accurate pixel-level localization. This controlled fusion process is defined by the following equation:

$$F_{\text{fused}} = F_{\text{UNet}} + \tanh(\alpha) \odot F_{\text{ResNet}} \tag{1}$$

where F_{fused} is the final fused feature, F_{UNet} represents the U-Net's bottleneck embeddings, F_{ResNet} are the linearly projected classification features, and α is the learnable gating parameter. The operator $\odot$ denotes element-wise multiplication.

3 Methodology

3.1 Dataset and data preprocessing

The ultrasound dataset provided by the challenge combines public and private datasets spanning classification and segmentation tasks, i.e breast nodules classification and segmentation (BUSI, BUSIS [13], and BUS-BRA [2]), thyroid nodules segmentation (DDTI), kidney contour delineation (KidneyUS [10]), fetal head contour delineation (Fetal Head Circumference [11]), cardiac chamber segmentation (CAMUS [4]), Appendix [6] and Fatty-Liver [1] classification. The dataset has a total of 8736 images (7,448 public, and 1,248 private), 6,614 with segmentation masks and 4,945 with classification labels. To standardize field-of-view and remove machine-specific overlays, we applied a deterministic cropping pipeline based on connected component analysis. For each image, we first built a binary mask of "black" pixels (intensity 0, allowing a small tolerance) and extracted the largest component that touches the image border, which captures the non-informative frame/regions outside the ultrasound sector. We then negated this mask and selected the largest remaining component, corresponding to the actual echogenic sector, and used its tight bounding box to crop both the image and, when present, the corresponding annotation masks. This procedure succeeded on all samples except two, which were flagged for manual review. Prior to model input, images underwent intensity normalization and during training, we applied on-the-fly data augmentation.

3.2 Implementation Details

The training of our multi-task model is a sequential process, with each branch trained in a distinct phase. The architecture was implemented using Torch 2.7.1, and all models were trained on an NVIDIA L40S GPU with 48GB of memory.

Training of the Classification Branch. The ResNet-18 classification branch is trained first in a two-phase process: a general pre-training followed by fine-tuning. Throughout both phases, we use the AdamW optimizer with a cosine

annealing learning rate scheduler. The initial phase consists of a general pre-training for 20 epochs on the public datasets, to perform classification and bounding box regression. The optimization is driven by a multi-objective loss function combining Focal Loss and standard Cross-Entropy Loss for classification, and a Smooth L1 loss for bounding box regression with an initial learning rate of 1×10^{-4}. Then the classification branch undergoes a fine-tuning phase for approximately 5 epochs on the smaller, private part of the dataset. The training procedure focuses solely on classification, using the Focal Loss and Cross-Entropy Loss, with an initial learning rate of 1×10^{-5}. This lower learning rate, combined with the cosine scheduler, ensures that the model fine-tunes its parameters without drastically altering the robust features learned during the pre-training.

Training of the Segmentation Branch. We train the UNet for segmentation from scratch, freezing the classification branch, which output features are fused in the bottleneck using the tanh gating mechanism (Sec. 2.3 for details). The U-Net is trained using a multi-objective loss function that includes Binary Cross-Entropy (BCE) Loss and Dice Loss, using the SGD optimizer with an initial learning rate of 2×10^{-2} and a LambdaLR scheduler.

3.3 Postprocessing Strategy

We apply a connected-components post-processing to binary masks. Specifically, we first smooth the mask with a small morphological closing and opening (3×3 elliptical kernels) to suppress speckle and thin artifacts. We then run 8-connected components analysis, retain the largest component and also keep the second largest only if its area is at least 30% of the largest (to preserve plausible bi-lobed shapes), and finally fill small internal holes ($\leq 0.3\%$ of the ROI). Instead, for thyroid, which have more heterogeneous and less precise output masks, we use a simplified rule that removes components smaller than 0.1% area of the whole image, suppressing residual speckle while preserving the main structure.

4 Results and Conclusion

From the results presented in Table 1, it is evident that using a U-Net backbone is not a suitable approach for this classification task. This reinforces the hypothesis that the primary objective of segmentation, for which U-Net was designed, is in direct conflict with the demands of a classification objective. While the results with the ResNet-18 backbone show a significant improvement, demonstrating that a dedicated classification architecture is far more effective. A major challenge in classifying ultrasound images is preventing the model from overfitting to irrelevant markers or artifacts from different acquisition machines, rather than focusing on the anatomical features. A qualitative evaluation suggests that the CBAM [12] module is highly effective at helping the model focus on relevant features, a finding that is also supported by the quantitative results from the table. Furthermore, the data reveals a true correlation in performance only when there is a sufficient amount of public and high-quality data available, as is the case for the breast nodules task.

Model	CBAM	FILM	Appendix	Breast	Breast Lum.	Liver	Overall
U-Net [8]	✗	✗	47.3%	51.2%	17.3%	46.7%	40.6%
	✗	✓	48.8%	46.5%	26.9%	42.4%	41.2%
ResNet-18 [3]	✗	✗	61.2%	62.0%	9.3%	52.4%	46.2%
	✗	✓	42.7%	67.8%	61.9%	51.8%	56.05%
	✓	✗	**61.2%**	**68.7%**	**59.3%**	**53.4%**	**60.7%**
	✓	✓	61.2%	65.3%	65.7%	47.1%	59.8%

Table 1: Performance of different classification strategies, on private and unseen dataset during pre-training (breast luminal performances are negligible since they are not directly trained due to lack of publicly available dataset).

Configuration			DSC/NSD (%)					Average (%)	
FiLM	Crop	Postproc	Breast	Cardiac	Thyroid	Fetal	Kidney	DSC	NSD
✗	✗	✗	81.1/15.1	81.1/5.2	62.7/6.9	92.0/9.4	81.5/9.4	79.7	9.2
✗	✓	✗	82.4/17.0	81.7/**6.0**	67.1/8.8	91.4/9.7	85.2/10.3	81.6	10.4
✓	✗	✗	83.9/18.0	82.9/5.4	66.2/8.5	93.2/11.2	86.2/11.3	82.5	10.9
✓	✗	✓	84.0/**18.3**	83.0/5.9	67.1/9.0	93.3/11.5	88.6/11.8	83.2	11.3
✓	✓	✗	84.2/18.0	**83.3**/5.8	67.2/9.2	93.0/11.3	88.1/12.0	83.1	11.2
✓	✓	✓	**85.5**/**18.3**	83.2/5.8	**67.8**/**9.7**	**93.7**/**12.3**	**88.7**/**12.1**	**83.8**	**11.6**

Table 2: Performance evaluation of different segmentation configurations on the validation dataset.

Regarding the segmentation task, in Table 2, we observe that each of the three complementary choices in the segmentation branch, bring some improvements to the metrics. First, FiLM conditioning improves both DSC and NSD across organs by modulating feature responses to the organ context, which reduces cross-organ confusion and stabilizes boundaries. Second, cropping the input is beneficial, especially for datasets that include broad black borders, on-screen markers, and large extra-anatomical region, because it suppresses machine-specific artifacts and background clutter, allowing the model to focus on the relevant anatomy. Third, the light post-processing step further removes small spurious blobs and enforces anatomically plausible masks, yielding steady NSD improvements by cleaning false positives near edges. The combination of FiLM, cropping, and post-processing delivers the best overall performance in both DSC and NSD, indicating that representation conditioning, input focus, and minimal morphological cleanup are effective.

It is important to note that these findings are particularly valuable for small-sized models. We also explored strategies that involved the use of significantly larger models, such as MedSAM [5] and Med-Gemma [9]. However, we did not find a valuable trade-off between the marginal performance gains and the substantial increase in computational resources, including time and memory. Therefore, the strategies outlined here represent an optimal approach for developing efficient and effective models for the UUSIC challenge's tasks.

References

1. Byra, M., et al.: Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. International Journal of Computer

Assisted Radiology and Surgery **13**(12), 1895–1903 (2018)
2. Gómez-Flores, W., et al.: Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. Medical Physics **51**(4), 3110–3123 (2024)
3. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
4. Leclerc, S., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)
5. Ma, J., et al.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
6. Marcinkevičs, R., et al.: Regensburg pediatric appendicitis dataset. (No Title) (2023)
7. Perez, E., et al.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
8. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
9. Sellergren, A., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
10. Singla, R.C., Rohling, R.: The open kidney ultrasound data set. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 155–164. Springer (2023)
11. Van Den Heuvel, T., et al.: Automated measurement of fetal head circumference using 2d ultrasound images. PloS One **13**(8), e0200412 (2018)
12. Woo, S., et al.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)
13. Zhang, Y., et al.: Busis: a benchmark for breast ultrasound image segmentation. In: Healthcare. vol. 10, p. 729. MDPI (2022)

Bibliography

[1] Boris. My awesome cat!, 2024.

[2] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[3] World Health Organization. *Infertility prevalence estimates, 1990–2021*. World Health Organization, 2023.

[4] Tammy J Lindsay and Kirsten R Vitrikas. Evaluation and treatment of infertility. *American Family Physician*, 91(5):308–314, 2015.

[5] Giorgia Spaggiari, Antonio RM Granata, and Daniele Santi. Testicular ultrasound inhomogeneity is an informative parameter for fertility evaluation. *Asian Journal of Andrology*, 22(3):302–308, 2020.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Carmel M Moran and Adrian JW Thomson. Preclinical ultrasound imaging—a review of techniques and imaging applications. *Frontiers in Physics*, 8:124, 2020.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[10] Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. Deep learning in medical ultrasound analysis: a review. *Engineering*, 5(2):261–275, 2019.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[12] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[13] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[16] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems*, 31, 2018.

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[18] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 29, 2016.

[19] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Llearning*, pages 1597–1607. PmLR, 2020.

[20] Yixiong Chen, Chunhui Zhang, Li Liu, Cheng Feng, Changfeng Dong, Yongfang Luo, and Xiang Wan. Uscl: Pretraining deep ultrasound image diagnosis model through video contrastive representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 627–637. Springer, 2021.

[21] Zehui Lin, Zhuoneng Zhang, Xindi Hu, Zhifan Gao, Xin Yang, Yue Sun, Dong Ni, and Tao Tan. Uniusnet: A promptable framework for universal ultrasound disease prediction and tissue segmentation. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3501–3504. IEEE, 2024.

[22] W. Gómez-Flores et al. Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4):3110–3123, 2024.

[23] Y. Zhang et al. Busis: a benchmark for breast ultrasound image segmentation. In *Healthcare*, volume 10, page 729. MDPI, 2022.

[24] S. Leclerc et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Transactions on Medical Imaging*, 38(9):2198–2210, 2019.

[25] R. Christopher Singla and Robert Rohling. The open kidney ultrasound data set. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 155–164. Springer, 2023.

[26] M. Byra et al. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. *International Journal of Computer Assisted Radiology and Surgery*, 13(12):1895–1903, 2018.

[27] E. Perez et al. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[28] S. Woo et al. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[29] A. Sellergren et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[30] J. Ma et al. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.