

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Predicting the Aesthetic Outcome of Breast Cancer Treatment using Deep Learning

Joana Inês Gonçalves dos Santos



Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Jaime dos Santos Cardoso

Second Supervisor: Maria Helena Sampaio de Mendonça Montenegro e Almeida

July 21, 2025

Predicting the Aesthetic Outcome of Breast Cancer Treatment using Deep Learning

Joana Inês Gonçalves dos Santos

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Ricardo Pereira de Magalhães Cruz

Referee: Prof. Jaime dos Santos Cardoso

Referee: Prof. Wilson José Santos Silva

July 21, 2025

Resumo

Os tratamentos para o cancro da mama conduzem frequentemente a alterações estéticas significativas e altamente individualizadas, com impacto no bem-estar psicológico das pacientes e na sua satisfação com os cuidados recebidos. Prever a aparência pós-operatória continua a ser um desafio complexo devido à variabilidade nos resultados cirúrgicos. Esta dissertação apresenta uma abordagem baseada em *deep learning* para gerar previsões visuais personalizadas do resultado estético pós-cirúrgico, utilizando apenas fotografias pré-operatórias e metadados clínicos.

O estudo começou por avaliar os modelos de transformação de imagem conhecidos, Pix2Pix, BicycleGAN e StarGAN. Embora estes modelos tenham fornecido uma base sólida, tinham limitações significativas em termos de fidelidade de cor, realismo anatómico e relevância clínica. Para ultrapassar estes desafios, foi explorada uma variedade de configurações, funções de *loss* e estratégias de supervisão. O objetivo era criar imagens que fossem simultaneamente realistas do ponto de vista visual e representativas da intervenção cirúrgica, mantendo uma forte semelhança biométrica com a aparência pré-operatória da paciente.

Esta exploração culminou no desenvolvimento de uma arquitetura personalizada com uma separação clara de tarefas: um *encoder* dedicado é responsável pela classificação semântica do tipo de cirurgia e da lateralidade do tumor, enquanto o *discriminator* se foca exclusivamente na avaliação do realismo visual. Esta divisão evita enviesamentos semânticos nas imagens geradas e melhora tanto a qualidade perceptiva como a precisão anatómica.

O modelo suporta duas variantes: uma com skip connections, que melhora a consistência cromática, e outra sem, que reforça o detalhe anatómico. Ambas as variantes incorporam funções de *loss* sensíveis à segmentação, condicionamento específico ao tipo de cirurgia e técnicas de regularização. Os resultados quantitativos revelam melhorias significativas na similaridade estrutural e na precisão da segmentação, demonstrando o potencial do modelo para apoiar decisões clínicas informadas e personalizadas no planeamento do tratamento do cancro da mama.

Palavras-chave: Deep Learning, Tradução de Imagem para Imagem, Cancro da Mama, Previsão de Resultados Pós-operatórios, Síntese de Imagens Médicas, GANs Condicionais.

Abstract

Breast cancer treatments often lead to significant and highly individualised aesthetic changes, which can impact patients' psychological well-being and satisfaction with their care. Anticipating postoperative appearance poses a complex challenge due to variability in surgical outcomes. This dissertation presents a deep learning-based framework that generates personalised visual predictions of post-surgical outcomes using only preoperative photographs and minimal clinical meta-data.

The study began by evaluating state-of-the-art image-to-image translation models, Pix2Pix, BicycleGAN, and StarGAN. While these models provided a strong foundation, they had significant limitations in terms of colour fidelity, anatomical realism, and clinical relevance. To overcome these challenges, a variety of architectural configurations, loss functions, and supervision strategies were explored. The goal was to create images that were both visually realistic and accurately represented the surgical intervention while maintaining a strong biometric similarity to the patient's pre-operative appearance.

This exploration led to the development of a custom architecture with a clear separation of tasks: a dedicated encoder is responsible for the semantic classification of surgery type and tumour laterality, while the discriminator focuses solely on evaluating visual realism. This design helps prevent semantic bias in the generated images and improves both perceptual and anatomical accuracy.

The model supports two variants: one with skip connections for improved colour consistency and one without for sharper anatomical detail. Both variants incorporate segmentation-aware loss functions, surgery-specific conditioning, and regularisation techniques. Quantitative results indicate significant improvements in structural similarity and segmentation accuracy, demonstrating the model's potential to support personalised and informed decision-making in breast cancer treatment planning.

Keywords: Deep Learning, Image-to-Image Translation, Breast Cancer, Post-operative Outcome Prediction, Medical Image Synthesis, Conditional GANs.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework to achieve a better and more sustainable future for all. It includes 17 goals to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

This dissertation contributes to several of these goals by developing an AI-based framework to support personalised and informed decision-making in breast cancer treatment. The project promotes inclusive healthcare, empowers women in clinical decisions, and fosters technological innovation in medical imaging.

The specific Sustainable Development Goals addressed in this work are:

SDG 3 Ensure healthy lives and promote well-being for all at all ages

SDG 5 Achieve gender equality and empower all women and girls

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialisation and foster innovation

SDG 10 Reduce inequality within and among countries

SDG 16 Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions

SDG	Target	Contribution	Performance Indicators and Metrics
3	3.8	Achieve universal health coverage by improving patient access to high-quality predictive tools that support surgical planning and shared decision-making.	Developed a supervised deep learning model for post-operative appearance prediction. Quantitative evaluation using SSIM, IoU, and colour consistency.
5	5.6	Ensure universal access to sexual and reproductive health care services by empowering women with predictive insights on post-operative outcomes in breast cancer treatment.	Created personalised visualisations of post-surgical outcomes. Model conditioned on surgery type and tumour laterality to enhance decision-making.
9	9.5	Enhance research and upgrade technological capabilities by developing AI-driven medical imaging tools for personalised breast cancer surgery planning.	Proposed novel architectures (e.g., encoder-based GAN with multi-loss supervision). Contributed to two scientific publications.
10	10.2	Promote inclusion and equity in women's health by enabling access to predictive tools that support informed surgical decisions, especially for underrepresented or underserved groups.	Ensured generalisation across surgery types, including rare cases like mastectomies. Evaluation included patients with diverse clinical conditions.
16	16.7	Promote participatory and inclusive healthcare decision-making by integrating AI-based outcome visualisations.	Generated clinically interpretable visual outputs. Designed the system to support patient–clinician communication in shared decision-making.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professor Maria Helena Montenegro and Professor Jaime Cardoso, for their guidance, encouragement, and valuable insights throughout this dissertation. Their support was essential at every stage of this work.

To my parents and my brother, thank you for your unconditional love, patience, and unwavering belief in me. Your support has been the foundation of this journey.

A special thank you to Fábio and Marco, whose friendship, motivation, and honest feedback helped me stay focused and motivated.

I am also deeply grateful to the Cinderella Project for the opportunity to work on a topic of real-world impact and for providing access to the dataset that made this research possible, and to INESC TEC for supporting the development of this work through a Research Initiation Grant (BII), reference 11597/BII-E_B4/2025, within the scope of the HfPT project funded by IAPMEI.

Joana Santos

“There is no such thing as a failed experiment, only experiments with unexpected outcomes.”

R. Buckminster Fuller

Contents

1	Introduction	1
1.1	Context and Motivation	1
1.2	Problem	2
1.3	Research question	2
1.4	Contributions	3
1.5	Document structure	3
2	Background	5
2.1	Bibliographic search	5
2.1.1	Databases and Platforms	5
2.1.2	Search Keywords	6
2.2	Introduction to AI	6
2.3	Deep Learning	7
2.3.1	Neural Networks	7
2.3.2	Convolutional neural networks	11
2.3.3	Autoencoders	15
2.3.4	U-Net	15
2.3.5	Transformers	17
2.4	Deep Generative models	18
2.4.1	Variational Autoencoders	20
2.4.2	Generative Adversarial Networks	22
2.4.3	Normalizing Flows	24
2.4.4	Diffusion Models	25
2.5	Discussion	27
3	Literature Review: Image-to-Image translation	28
3.1	Introduction	28
3.1.1	Image-to-Image Translation Methods	30
3.2	Discussion	37
4	Literature Review: Generation of Post-Operative images	39
4.1	Introduction	39
4.2	Objective Aesthetic Evaluation Metrics	40
4.3	Image transformation in breast cancer scenarios	41
4.3.1	Regression-Based Deformation Prediction	41
4.3.2	Keypoint Reallocation and Affine Transformations	42
4.3.3	GAN-Based Image Transformation for Aesthetic Prediction	43
4.4	Image transformation in surgery	46

4.4.1	Conditional GANs for Orbital Decompression Surgery	47
4.4.2	Pix2Pix GANs for Eye Surgery	47
4.4.3	3D-Guided Face Manipulation Using CycleGANs	47
4.4.4	PlasticGAN for Facial Surgery Simulation	48
4.5	Datasets for Aesthetic Evaluation in Breast Cancer Treatment	48
4.6	Discussion	49
5	Preliminary Analysis and Data Preparation	50
5.1	Data Analysis	51
5.2	Image Preprocessing	54
5.2.1	Bounding Box Cropping with Keypoints	54
5.2.2	Flipping, Scaling and Normalisation	54
5.3	Evaluation Metrics	55
6	Baseline Experiments	57
6.1	Pix2Pix	57
6.1.1	Architecture and Training Procedure	57
6.1.2	Results	59
6.2	BicycleGAN	59
6.2.1	Architecture and Training Procedure	60
6.2.2	Results	62
6.3	StarGAN	63
6.3.1	Architecture and Training Procedure	63
6.3.2	Results	64
6.4	Conditioning with Clinical Information	66
6.4.1	Pix2Pix with Conditioning	66
6.4.2	BicycleGAN with Conditioning	68
6.5	Discussion	69
7	Improving Generalisation and Colour Fidelity with Targeted Losses	72
7.1	Overfitting	72
7.1.1	Dropout	72
7.1.2	Discriminator Augmentation	73
7.1.3	Results	73
7.2	Colour Consistency	74
7.2.1	Grayscale L1 Loss	75
7.2.2	Global Colour Loss	75
7.2.3	Histogram Loss	76
7.2.4	Patch-Based Average Colour Loss	77
7.2.5	Bounding Box Masked Colour Loss	78
7.2.6	Results	79
7.3	Discussion	80
8	Architectural Exploration	82
8.1	Loss Components	82
8.2	Architectural Variants Explored	83
8.2.1	Two Discriminators with Local Classification and Global Realism	84
8.2.2	Encoder for Classification	84
8.2.3	Single Discriminator with Post-Only Classification Supervision	87

8.2.4	Encoder with Local and Global Discriminators	88
8.2.5	Results	89
8.3	Bounding Box and Masking Strategies	90
8.3.1	Bounding Box Strategy: Union vs. Centre Expansion	91
8.3.2	Soft-edged Breast Mask for Supervision Loss	92
8.3.3	Effect of Removing Segmentation Loss	93
8.3.4	Results	95
8.4	Discussion	95
9	Refinements in Generator Supervision and Training	96
9.1	Impact of Generator Connectivity and Training Supervision	96
9.1.1	With and Without Skip Connections in Generator	96
9.1.2	Continue Training Without the Supervision Loss	99
9.1.3	Training Without Supervision Loss from Scratch	100
9.1.4	Results	101
9.2	StarGAN and BicycleGAN with the Encoder for Classification	102
9.2.1	StarGAN	102
9.2.2	BicycleGAN	104
9.3	Discussion	106
9.3.1	Testing with Mastectomies	108
10	Conclusion and Future Work	111
10.1	Future Work	112
	References	113
A	Further Use of AI Tools	119

List of Figures

2.1	Neural Networks Structure	8
2.2	Illustration of gradient descent optimisation, showing the iterative process of adjusting weights to minimise the cost function. Each incremental step moves closer to the minimum cost	10
2.3	General structure of a Convolutional Neural Network (CNN), showing the flow from input through convolutional and pooling layers to fully connected layers for final output.	12
2.4	Illustration of the convolution operation in a Convolutional Neural Network (CNN). A convolutional kernel slides across the input matrix, applying element-wise multiplication and summation to produce the corresponding output feature map. . . .	13
2.5	Illustration of max pooling and average pooling operations in a CNN. Max pooling selects the highest value in each region to emphasize the most significant features, while average pooling computes the mean value to capture broader patterns. Both methods use a kernel of size 2×2 to reduce spatial dimensions and improve computational efficiency.	14
2.6	Architecture of an Autoencoder, depicting the encoder, bottleneck (latent space), and decoder components.	15
2.7	U-Net architecture: An encoder-decoder network with skip connections that transfer feature maps from the contracting path to the expanding path [66].	16
2.8	Architecture of the Transformer model, displaying the encoder and decoder layers, including multi-head attention, feed-forward networks, and positional encoding, used for processing sequential input data [79].	18
2.9	Taxonomy of deep generative models, dividing models into explicit and implicit density estimation approaches, with further subdivisions into tractable, approximate, and hybrid models, including GANs, VAEs, Diffusion Models, and Normalizing Flows [50].	19
2.10	Architecture of a Variational Autoencoder (VAE), showing the encoder that generates a mean vector (μ) and a standard deviation vector (σ) to create the latent representation (z), followed by the decoder to reconstruct the input.	21
2.11	Example of latent space manipulation in VAEs: A generated face image shows variations achieved by perturbing the latent space, adjusting features such as head tilt and introducing a subtle smile [39].	22
2.12	Architecture of a Generative Adversarial Network (GAN). The generator creates synthetic images from random input, while the discriminator evaluates whether the input is real or fake. Backpropagation minimizes the error, training both networks to improve their respective outputs.	23

2.13	Illustration of the forward/diffusion and reverse/denoising processes. The forward process involves adding noise to an image step-by-step, while the reverse process reconstructs the original image by removing noise in reverse order [9]	26
3.1	Results of the model BicycleGAN. The left column shows the input. The second shows the ground truth output. The final four columns show randomly generated samples [85].	33
3.2	Examples of image-to-image translation tasks across four domains [82].	33
3.3	Examples of image-to-image translation in various applications. Left: how an image of a different class in ImageNet preserves the salient shape. Right: transformations of a tiger image into other animals (husky, lion, bear).	34
3.4	Image translation examples of the AGG method [43]	35
3.5	VQBB results on face2comic dataset [46].	35
3.6	Comparison of StarGAN with other methods in facial expression synthesis results on the RaFD dataset. [18]	36
4.1	A comparison of surgical scars: (a) Mastectomy [8] and (b) Breast-conserving surgery [7].	40
4.2	The post-surgery Pointcloud is shown with red points, while the predicted Pointcloud is visualised with black points. The displacements between corresponding points are coloured in blue. (a) Poor prediction, $D_{pred \rightarrow post} = 1.624$ mm; (b) Fair prediction, $D_{pred \rightarrow post} = 1.044$ mm; (c) Good prediction, $D_{pred \rightarrow post} = 0.827$ mm. [87]	42
4.3	Workflow of Freitas et al.'s traditional method. This approach involves annotating key points in pre-and post-operative images, extracting asymmetry features, reallocating key points, and performing image morphing to produce a morphed post-operative image [22].	42
4.4	Comparison of pre-operative images (left column) with results from the template-based methods: template images (second column), traditional computer vision (CV) approach (third column), and deep learning (DL) approach (fourth column). The DL method demonstrates superior nipple positioning accuracy and texture preservation, while the CV method excels in contour modification [52].	43
4.5	Workflow of Freitas et al.'s disentangled representation learning approach. This method separates medical, identity, and other features using an encoder-decoder architecture, enabling precise manipulation of asymmetry metrics while preserving patient-specific attributes like skin texture and tone [22].	44
4.6	Montenegro et al.'s improved disentangled representation learning framework [52].	44
4.7	Workflow of Montenegro et al.'s inpainting approach. This method involves covering the breast, calculating asymmetry metrics, and using annotations for the nipple and lowest point to guide the inpainting process [53].	45
4.8	Results of the inpainting approach of Montenegro et al. by incorporating asymmetries from post-operative templates into the original pre-operative photos of the CP dataset [53]	46
4.9	Example results from BreastGAN [16], showing (left to right): preoperative image, GAN-generated image, and true post-operative image. While results are visually plausible, colour and body shape inconsistencies are noticeable.	46

4.10	Comparison of pre-operative (TAO) and post-operative (ground truth) facial appearances with results generated by different conditional GAN (cGAN) methods, including initial latent vector optimisation using MSE + SSIM, MAE, NCC, and models like Pix2Pix and Cycle-GAN [83].	47
4.11	Comparison of pre-operative, actual post-operative, and synthesised post-operative OCT images generated using Pix2Pix GAN for epiretinal membrane (ERM) removal. [37].	48
4.12	Example outputs from a CycleGAN-based method by Andlauer et al. [3], showcasing 3D-guided facial manipulations for predicting postoperative outcomes. The method fuses 2D pre-operative images with 3D simulations of post-operative soft-tissue deformations, enabling changes to specific facial features, such as the chin (a) and nose (b).	48
4.13	Architecture of PlasticGAN, a Conditional GAN framework proposed by Chandaliya et al. [15] for simulating post-operative outcomes in facial plastic surgery. The model employs a combination of adversarial, perceptual, and reconstruction losses during training to generate realistic post-operative images, targeting tasks such as face ageing, rejuvenation, and reconstruction.	49
5.1	(a) Distribution of cancer side (left, right, bilateral). (b) Distribution of patients who received radiotherapy.	51
5.2	Distribution of patients by age range at the time of surgery.	51
5.3	(a) Number of pre-operative images per patient. (b) Number of post-operative images per patient.	52
5.4	Distribution of time (in years) between surgery and post-operative image acquisition.	52
5.5	Scatter plot of image width vs. height.	53
5.6	Examples of patient variation. (a) Pre-operative image. (b) Post-operative image. Differences in body shape, posture, and arm position illustrate the challenge of learning consistent transformations.	54
5.7	Effect of bounding box cropping with keypoints and zoom-out factors. Top row: original pre- and post-operative images. Middle and bottom rows: zoomed crops with factors of 1.1 and 1.4, respectively.	55
5.8	Evaluation metrics used to compare model performance. From left to right: SSIM with the pre-operative image, SSIM with the post-operative image (breast region only), IoU between segmentations (breast and nipple), and colour preservation metrics (outside-mask and patch-based L1, both computed against the pre-operative image).	56
6.1	Pix2Pix architecture: the generator learns a mapping from pre-operative to post-operative images. The discriminator distinguishes real vs. fake pairs.	58
6.2	Visual comparison of generated and real post-operative images in (a) the testing set and (b) the training set.	60
6.3	Architecture and training modes of BicycleGAN. Training alternates between the cVAE-GAN branch (left) and cLR-GAN branch (right), enabling both latent code recovery and output diversity [85].	61
6.4	Visual results from BicycleGAN. Each row shows a pre-operative image, a generated prediction, and the ground truth post-operative image. (a): Test set, (b): Training set. The model struggles with skin tone preservation and generates only minor shape variations.	62

6.5	Adapted StarGAN training process. The original architecture was modified by adding an L1 supervision loss to support paired image translation. The generator performs domain-to-domain transformations, guided by both adversarial and cycle-consistency losses [18].	63
6.6	Visual results from StarGAN. Each row displays a pre-operative image, the predicted image, and the actual post-operative image. (a) Test set, (b) Training set.	65
6.7	Effect of strong reconstruction loss in StarGAN. From left to right: pre-operative input, predicted image, and real post-operative outcome. The predicted image closely mirrors the input, showing minimal surgical changes.	65
6.8	Visual comparison between Pix2Pix outputs without (left) and with (right) clinical conditions. The addition of surgery type and cancer side improves the realism and clinical relevance of the generated post-operative images.	68
6.9	Comparison of results from BicycleGAN models trained with (right) and without (left) clinical conditions. The conditioned model more accurately captures expected surgical outcomes.	69
7.1	Evolution of the adaptive augmentation probability p during training. The value quickly saturates to near 1.0, suggesting that a fixed high value would suffice for regularisation.	74
7.2	Qualitative results showing the effect of dropout and discriminator augmentation. Left: pre-operative image. Right: ground truth post-operative image. Middle columns: predictions from (a) baseline, (b) with dropout, and (c) with dropout and augmentation.	75
7.3	Left: pre-operative image; Middle: binary skin mask using RGB heuristic; Right: visualisation of average colour extracted from masked region.	76
7.4	Effect of patch size on average colour in images of 256x256. (a) 16×16 patches preserve some shape; (b) 32×32 patches smooth out structural details.	77
7.5	Pre-operative and post-operative images may differ in pose (e.g., arm position), so penalising pixel differences globally may hurt learning. The bounding box strategy prevents this by ignoring non-surgical regions.	78
7.6	Right: the shape of the breast is supervised using grayscale losses inside the bounding box; Left: the colour outside is preserved from the pre-operative image.	79
7.7	Visual comparison of different colour supervision strategies across a fixed set of examples.	80
8.1	Visual examples of the four main loss components: Colour Loss 1 (patch-based MSE), Colour Loss 2 (outside-mask MSE), Structural Supervision (L1 + SSIM in grayscale), and segmentation loss. Each pair of images shows the inputs used in the respective loss calculation.	82
8.2	Architecture with two discriminators: D2 (global) performs real/fake discrimination using pre/post pairs; D1 (local) classifies and discriminates based on real or fake post-operative images.	84
8.3	Architecture and results. The discriminator evaluates realism from both the pre- and post-operative images. The encoder classifies surgical conditions. The generator is conditioned on surgery type and cancer side.	85

8.4	The encoder input comparison is illustrated with purple rectangular boxes highlighting the differences. (a): cropped breast region (masked version). (b): full-body image (unmasked version). The masked setup isolates surgical features, while the unmasked version incorporates background and posture information. . .	86
8.5	Comparison of generated images using encoder inputs with and without breast-region masking. Columns: Pre-operative, Predicted, Real Post-operative. Rows show different patients. Masked versions sometimes show clearer nipples or surgical scars.	87
8.6	Single-discriminator architecture. The discriminator learns real vs. fake using pre- and post-operative images. For classification, it is only trained on real post-operative images. The generator is penalised if the discriminator misclassifies its generated output.	88
8.7	Architecture with an encoder and two discriminators. Classification is handled independently by the encoder, while adversarial realism is enforced globally and locally.	89
8.8	Qualitative comparison of architectural variants across five representative patients. Columns: pre-operative image, predictions from each model variant, and the ground-truth post-operative image.	90
8.9	Bounding box overlays for both strategies. Each row shows a pre-operative image (left) and a post-operative image (right). In each image, the black box denotes the bounding box derived from the pre-operative mask, the purple box from the post-operative mask, and the dark purple box corresponds to the final region used for spatial supervision. In (a), the dark purple box is the union of both individual boxes and is applied equally to both images. In (b), the dark purple box is the result of an expansion around each mask, ensuring better anatomical centring for each domain.	91
8.10	Qualitative comparison of bounding box strategies. Left: Union strategy includes excess background and induces spatial misalignment. Right: The expansion method centres on the region of interest and aligns anatomical features more accurately. .	92
8.11	Comparison of models trained with hard-edged (a) and soft-edged (b) supervision masks. Soft-edged masks reduce visible border artefacts and lead to smoother transitions between the breast region and surrounding anatomy. Panel (c) illustrates the effect of the soft spatial falloff applied to the mask, transitioning gradually from full supervision (centre) to none (edges).	93
8.12	Visual comparison of the encoder model with (a) and without (b) segmentation loss. Both columns show pre-operative, predicted, and ground-truth post-operative images. While overall colour and body shape are preserved in both cases, the version without segmentation loss exhibits reduced anatomical precision, especially in the nipple and breast contours.	94
9.1	Comparison of models trained with (left) and without (right) skip connections. Each row shows a pre-operative input, the predicted image, and the ground truth post-operative result. Removing skip connections improves structural realism by forcing the model to rely on encoded representations, but also leads to increased colour variability. Models with skip connections preserve colour more accurately but often produce blurrier and less defined anatomical changes.	97

9.2	Effect of tuning loss weights on models without skip connections. Columns vary λ_{sim} , λ_{colour} , and λ_{seg} . Rows show different patients. Models with higher colour weights preserve tone better, while lower structural weights allow more diverse transformations.	98
9.3	Comparison between the supervised encoder model (a) and the same model after continued training without structural supervision (b). While some outputs appear more detailed, others show semantic drift due to the absence of pixel-level loss.	99
9.4	Comparison of no supervision loss models trained with (left) and without (right) skip connections.	100
9.5	Qualitative comparison of architectural variants across five representative patients. Columns: pre-operative image, predictions from each model variant, and the ground-truth post-operative image.	102
9.6	Qualitative results of StarGAN variants. From left to right: pre-operative image, StarGAN baseline, StarGAN with encoder, StarGAN with encoder and high reconstruction weight, and ground truth post-operative image.	103
9.7	Qualitative comparison of BicycleGAN variants. From left to right: pre-operative image, BicycleGAN with conditioning, BicycleGAN with encoder, BicycleGAN with encoder and stronger KL and latent regression loss, and the ground truth post-operative image.	105
9.8	Qualitative comparison of architectural variants across five representative patients. Columns: pre-operative image, predictions from each model variant, and the ground-truth post-operative image.	107
9.9	Predictions, using the encoder model without skip connection, for different surgery types and cancer sides using the same pre-operative image.	108
9.10	Qualitative comparison of model performance on mastectomy cases (grouped). Each triplet shows pre-operative, predicted, and ground truth post-operative images.	110

List of Tables

2.1	Common Activation Functions with Plots	9
2.2	Comparison of Probabilistic Models in Machine Learning	20
2.3	Types of Variational Autoencoders (VAEs) and Their Characteristics	22
2.4	Variants of Generative Adversarial Networks (GANs)	24
3.1	Image-to-Image Translation Techniques and Applications	29
3.2	Two-Domain Image-to-Image Translation Techniques	31
3.3	Multi-Domain Image-to-Image Translation Techniques	32
5.1	Types of surgery and their frequencies in the dataset.	53
6.1	Quantitative results across all evaluated models.	70
7.1	Test set performance of Pix2Pix models with increasing levels of regularisation. .	74
7.2	Quantitative comparison of colour supervision strategies,	80
8.1	Comparison of full-image and masked encoder inputs across visual similarity (SSIM), anatomical accuracy (IOU), and colour preservation losses.	87
8.2	Quantitative evaluation of architectural variants	90
8.3	Performance comparison of bounding box strategies. The Expansion method provides improved anatomical alignment and slightly better segmentation metrics. .	92
8.4	Comparison of encoder model performance with and without segmentation loss. .	94
9.1	Quantitative evaluation of BicycleGAN variants. The encoder-based version achieves higher SSIM and colour consistency. Excessively strong KL and latent regression constraints lead to instability in visual results.	101
9.2		104
9.3	Quantitative evaluation of BicycleGAN variants. The encoder-based version achieves higher SSIM and colour consistency. Excessively strong KL and latent regression constraints lead to instability in visual results.	106
9.4	Quantitative evaluation of BicycleGAN variants. The encoder-based version achieves higher SSIM and colour consistency. Excessively strong KL and latent regression constraints lead to instability in visual results.	107
9.5	Per-surgery breakdown of results for the two encoder variants.	108
9.6	Types of mastectomies and their frequencies in the dataset.	109
9.7	Quantitative performance of the model across different surgery types, including mastectomies (grouped).	109

List of Acronyms

AE	Autoencoder
AGG	Asymmetric Gradient Guidance
AGUIT	Attribute-guided unpaired Image-to-Image Translation
AI	Artificial Intelligence
AIM	Artificial Intelligence in Medicine
ANN	Neural Networks in Medicine
BMI	Body Mass Index
cGANs	Conditional Generative Adversarial Networks
cLR-GAN	Conditional Latent Regressor GAN
CNN	Convolutional Neural Network
CVAE	Conditional VAE
cVAE-GAN	Variational Autoencoder GAN
DBMS	Deep Boltzmann Machines
DBNS	Deep Belief Networks
DCGAN	Deep Convolutional GAN
DDIB	Dual Diffusion Implicit Bridges
DDPMs	Denoising Diffusion Probabilistic Models
DDS	Decomposed Diffusion Sampling
DGMs	Deep Generative Models
DL	Deep Learning
DMT	Diffusion Model Translator
FNNs	Feedforward Neural Networks
GANS	Generative Adversarial Networks
GMMs	Gaussian Mixture Models
HMMs	Hidden Markov Models
IoU	Intersection over Union
KL	Kullback-Leibler
MAE	Mean Absolute Error
MBs	Boltzmann Machines
MCG	Manifold Constraint Gradient
ML	Machine Learning
NFs	Normalizing Flows
NN	Neural Networks
RBM	Restricted Boltzmann Machines
RNNs	Recurrent Neural Networks
SSIM	Structural Similarity Index Measure
SSVAE	Semi-Supervised VAE
VAE	Variational Autoencoder
VQBB	Vector Quantised Brownian Bridge
WGAN	Wasserstein GAN

Chapter 1

Introduction

1.1 Context and Motivation

In recent years, survival rates for breast cancer have improved significantly due to advancements in screening methods and treatments [10]. These treatments are typically multimodal, involving a combination of surgery, systemic therapy, and radiation. They are also personalised, taking into account individual risk factors, tumour characteristics, and physician recommendations [35]. Among surgical options, patients can choose between breast-conserving surgery, which removes only the cancerous tissue, and mastectomy, where the entire breast is removed, followed by reconstruction if desired. Since multiple treatment protocols can offer similar oncological outcomes, effectively communicating all available options, along with their benefits and drawbacks, can be a complex but essential part of clinical practice.

In this context, shared decision-making plays a vital role in patient-centred care. This approach incentivises collaboration between healthcare providers and patients, ensuring that all risks, treatment options, and patient preferences are discussed in depth [70]. To help patients make informed decisions, health centres can provide diagrams, brochures, and images of post-operative results from other patients. However, this raises concerns about patient privacy and does not fully account for individual variations in healing and aesthetic outcomes.

Despite significant advancements in breast cancer treatments, patients often face challenges related to the cosmetic effects of locoregional therapies, particularly in terms of their physical appearance. Surgical interventions can result in asymmetry, scarring, and deformities that may negatively impact a patient's well-being [38]. Furthermore, a lack of awareness about these potential aesthetic outcomes may result in unrealistic expectations, leading to disappointment with their post-treatment appearance. A study done by Brands-Appeldoorn et al. [6] highlights that patients may assess outcomes differently when compared to their initial expectations. Factors like asymmetry, breast sensitivity, and bra wearability significantly influence patient satisfaction, underlining the importance of providing clear information to help patients better manage their expectations.

1.2 Problem

The lack of understanding of aesthetic outcomes raises the need to give patients better information about the results of the different treatment options, enabling them to make informed decisions. However, despite the availability of resources to support patient decision-making, there remains a significant gap in tools that allow for individualised predictions of post-operative aesthetics.

The rise of deep learning technology, specifically models capable of producing realistic synthetic images, opens up possibilities for tackling these issues. By predicting post-operative outcomes based on pre-operative images and treatment variables, these models could provide patients with personalised views of possible surgical outcomes, increasing satisfaction and facilitating shared decision-making.

The existing literature focuses mainly on creating visual representations from post-operative photos of other patients as templates. However, these methods fail to account for individual differences in anatomy, healing patterns, and surgical responses. Freitas et al. [22] used traditional and deep-learning techniques to detect critical breast asymmetry; however, these algorithms require manual annotations.

Recent work by Montenegro et al. [53] uses inpainting techniques and autoencoder-based models to predict post-operative breast asymmetries from pre-operative images. This approach addresses some limitations of earlier methods, enabling a more realistic manipulation of breast asymmetry. However, as it primarily focuses on asymmetries, it often fails to capture other important aesthetic features such as scarring, skin discolouration, and texture changes.

To address these limitations, a model could be trained using pre-operative images with additional information (e.g., type of surgery and use of radiotherapy) to predict post-operative outcomes. This approach could enable predictions that account for asymmetries and other relevant features such as skin texture changes, scars, and colour variations.

1.3 Research question

To explore the different ways to generate accurate representations of post-surgery images, the following questions can guide the investigation:

1. How can deep learning models be trained to accurately predict post-operative aesthetic outcomes of breast cancer surgeries using pre-operative images and treatment information?
2. What additional information (e.g., type of surgery, use of radiotherapy, surgeon expertise) can be integrated into these models to improve the accuracy and realism of predicted outcomes?
3. What are the limitations of current prediction models in accurately simulating aesthetic outcomes like scarring, skin discolouration, and asymmetry in breast cancer treatments?
4. Can automated models replace the need for manual annotations in predicting breast surgery outcomes while maintaining or improving the realism and precision of the results?

1.4 Contributions

The main contributions of this dissertation are:

- **Technical Contributions:** This thesis introduces a custom deep learning architecture tailored to predict post-operative aesthetic outcomes in breast cancer treatment. The model was designed to improve both visual realism and clinical relevance by explicitly incorporating patient-specific clinical information, such as surgery type and tumour location. Two architectural variants were explored: one aimed at preserving colour consistency and another focused on enhancing anatomical detail. Additionally, the training process was guided by targeted loss functions that account for structural alignment, colour fidelity, and region-specific supervision, enabling the generation of more realistic and personalised post-surgical visualisations. The article has been submitted to "Deep-Brea3th 2025: 2nd Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care", part of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2025), and is currently under review.
- **Empirical Contributions:** This work includes a comprehensive evaluation of leading image-to-image translation models, Pix2Pix, BicycleGAN, and StarGAN, applied to the task of predicting aesthetic outcomes in breast cancer surgery. These models were benchmarked both quantitatively and qualitatively, revealing limitations in visual accuracy and clinical relevance. The proposed approach showed consistent improvements across objective evaluation metrics, such as Structural Similarity Index (SSIM), segmentation accuracy (IoU), and colour preservation. Importantly, the model demonstrated the ability to generalise across a wide range of surgical procedures, including rare and complex cases like mastectomies, highlighting its robustness and potential for broader clinical application.
- **Literature and Survey Contributions:** The state-of-the-art analysis conducted in the early chapters of this thesis informed the writing of a survey paper focused on aesthetic evaluation and outcome prediction after surgery. The contribution to this paper centred specifically on reviewing and analysing predictive approaches for modelling post-operative outcomes.

1.5 Document structure

This thesis is structured to guide the reader through a progressive development of the proposed methodology. It begins with Chapter 2, which provides a background on artificial intelligence and deep learning, ensuring that readers with varying levels of familiarity are brought to a common level of understanding. Then, Chapter 3 surveys the state-of-the-art image-to-image translation methods, focusing on deep generative models. Chapter 4 narrows this scope to models previously applied in breast cancer and surgical image generation. From there, Chapter 5 presents the dataset used in this work, detailing its characteristics and preprocessing steps. Chapter 6 introduces baseline experiments using Pix2Pix, BicycleGAN, and StarGAN, then explores model

conditioning using clinical metadata to improve semantic relevance. In Chapter 7, targeted loss functions are introduced to address colour consistency and anatomical fidelity. Building on these improvements, Chapter 8 explores architectural modifications, including encoder-based designs and multi-discriminator configurations. Finally, Chapter 9 presents improvements in generator and supervision strategies and reintroduces enhanced versions of prior models. The work concludes with Chapter 10, summarising the findings and outlining potential future research directions.

Chapter 2

Background

Predicting the aesthetic outcomes of breast cancer surgeries using deep generative models involves knowing concepts from artificial intelligence (AI) and deep learning. This section aims to provide a clear and accessible overview of these concepts, ensuring a common ground for readers with varying levels of familiarity with the subject. By doing this, the groundwork is laid to discuss the methodologies employed in addressing the challenges of simulating aesthetic outcomes in breast cancer treatments.

2.1 Bibliographic search

A literature search ensured a comprehensive grasp of the research area. It aimed to collect varied studies on deep generative models and their applications, providing a basis for identifying core themes, trends, and research gaps within the field.

2.1.1 Databases and Platforms

The search was performed in several academic databases to capture wide-ranging publications. These databases included:

- **Google Scholar:** For a broad overview and access to supplementary materials.
- **PubMed:** For biomedical and clinical research articles.
- **IEEE Xplore:** For technical and engineering-focused studies.
- **Elsevier (via ScienceDirect):** For peer-reviewed scientific publications spanning medicine and engineering.
- **SpringerLink:** For interdisciplinary research in computer science, medicine, and artificial intelligence.
- **Scopus:** For comprehensive searches across abstracts and citations in scientific journals.

- **ACM Digital Library:** For research articles and computing and information technology proceedings.
- **arXiv:** For preprints and emerging research in AI and computational methodologies.

2.1.2 Search Keywords

A list of keywords was used to identify relevant studies. Each class of keywords was carefully selected to target specific research aspects and emphasise the relationships between deep generative models, medical imaging, and breast cancer surgery outcomes. The main keyword classes, along with their underlying rationales and examples of queries, are listed below:

- **Deep Generative Models:** Added terms "Deep Generative Models," "Generative AI," "Deep Learning," "GANs," "VAEs," "Diffusion Models", and "Autoencoders" to consider state-of-the-art methodologies in generative modelling. These models are a basic building block in image synthesis, transformation, and prediction, all key components of this research.
- **Medical Imaging and Applications:** The keywords "Breast Cancer Outcomes Prediction," "Image Transformation," "AI in Medicine," "Deep Generative Models in Medicine," and "Image-to-Image Translation" were chosen to include research that combines medical imaging with artificial intelligence.
- **Domain-Related Keywords:** Added terms "Breast Surgery," "Aesthetic Outcome Prediction," "Post-operative Image Simulation," "Patient-Specific Outcome Prediction," and "Breast Reconstruction" allow focusing on the subspecialty of breast cancer surgery. Those keywords ensure that the results are patient-centred and clinically relevant.

Each of the studies identified by this initial search represented a large pool of articles, further filtered by looking at the title and abstract. Selected articles were then subjected to full-text reviews.

2.2 Introduction to AI

One of the many definitions of Artificial Intelligence (AI) is using machines and computers to emulate human behaviour and cognitive processes. Alan Turing - a foundational figure in AI - laid the foundation for computer science and AI in his 1950 paper "Computing Machinery and Intelligence" [63]. He proposed that a machine is intelligent if it could perform cognitive tasks similar to humans, an idea that later became known as the "Turing Test" [76].

From then on, AI grew at a rapid rate. What started as simple systems based on "if-then" rules has progressed to sophisticated algorithms that mimic the human brain and, in some ways, surpass it since these algorithms can perceive patterns and insights beyond a human's perception.

Artificial intelligence has various subdisciplines, each contributing uniquely to technological advancement.

- **Machine Learning (ML):** The use of statistical methods to find patterns in data and make predictions. In medicine, ML predicts patient outcomes, optimises treatment, and detects anomalies in medical records.
- **Deep Learning (DL):** Employs multi-layered neural networks for learning representations. DL algorithms can analyse complex visual information within the medical imaging domain, detecting patterns that assist radiologists in their diagnostic workflows.
- **Natural Language Processing (NLP):** Empowers machines capable of understanding and interpreting human language. NLP finds applications in machine translation, sentiment analysis, and speech recognition.
- **Computer Vision (CV):** Empowers computers to draw knowledge from and make inferences based on visual information, such as images and videos. In medical imaging, this technology is instrumental in the detection of abnormalities and provides support in diagnostics.

The focus of the problem at hand will be on explaining and exploring deep learning, specifically generative models.

2.3 Deep Learning

Deep learning is a form of machine learning that builds neural networks that mimic those found in the human brain. These networks are composed of interconnected layers of artificial neurons, each processing information and passing it on to the next.

2.3.1 Neural Networks

A traditional artificial neural network consists of several simple units called neurons, arranged into three kinds of layers (Figure 2.1) with every neuron in one layer connected to every neuron in the next. The first is the input layer, which receives input signals and passes them on to the following layer. Then, there are the hidden layers, which perform various calculations and feature extractions. Often, the network has multiple hidden layers, where the dimensionality determines the model's width. Finally, the output layer produces the network's final prediction or classification based on the processed data from the hidden layers.

Another way to visualise a NN is as a chain of functions, such as $f^{(1)}$, $f^{(2)}$, and $f^{(3)}$, connected sequentially to form $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$. These chain structures correspond to the layers of the network, where $f^{(1)}$ represents the first layer, $f^{(2)}$ the second layer, and so forth [25].

Besides ordinary neural networks, many special architectures were developed to cope with different kinds of data and various tasks. Two of the most salient architectures are:

- **Feedforward Neural Networks (FNNs):** FNNs are the simplest form of neural networks in which connections among neurons do not form a cycle. In addition, there is only one direction for information to flow: from the input layer through any hidden layers to the output

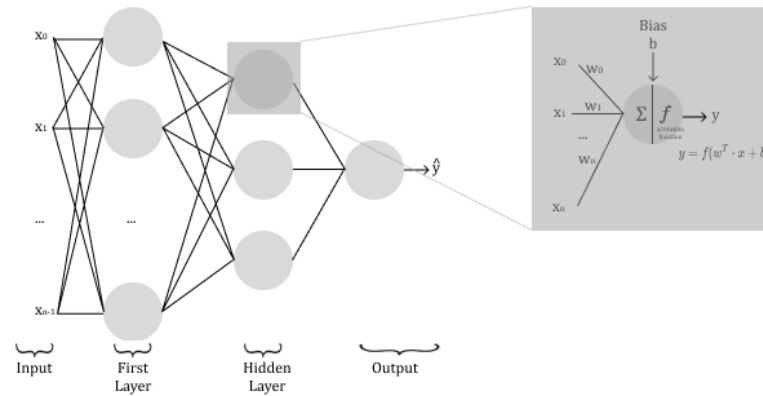


Figure 2.1: Neural Networks Structure

layer. Generally speaking, FNNs classify images and texts, where every input instance is independent of others.

- **Recurrent Neural Networks (RNNs):** RNNs are designed for processing sequence data, where each input in the sequence is related or connected with the previous inputs. Unlike FNNs, RNNs have looping connections that may allow them to keep track of information from past steps. These looping connections inside the network enable it to learn patterns across time or sequence. RNNs help perform language modelling, speech recognition, and time series prediction.

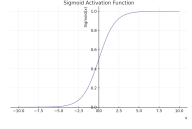
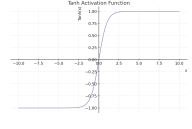
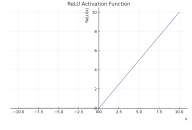
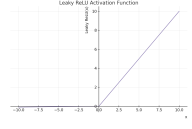
2.3.1.1 Forward Propagation

Forward propagation is the process by which input data is transmitted through a neural network. At each layer, neurons compute a weighted sum of their inputs, apply a non-linear activation function, and pass the result to subsequent layers. This chain of computations continues until the network produces an output. The goal is to approximate a target function, f^* [25].

Each connection between neurons has a weight, determining how much influence one neuron's output has on another's input. These weights are gradually adjusted during training to help the network make better predictions. Additionally, each neuron has a bias, an offset that allows the shift of the neuron's activation. This offset makes the network more flexible, allowing it to capture complex patterns better, especially when data is not perfectly centred.

To enable neural networks to capture complex and non-linear relationships in data, activation functions are applied at each neuron. These functions are crucial because, without them, the entire network would behave like a linear system, limiting its expressive power. Different activation functions influence how quickly a model trains, how well it converges, and its overall effectiveness. The Table 2.1 shows a few of the possible activation functions.

Table 2.1: Common Activation Functions with Plots

Activation Function	Formula	Range	Plot
Sigmoid	$\sigma(x) = \frac{1}{1+e^{-x}}$	$(0, 1)$	
Tanh	$\tanh(x) = \frac{2}{1+e^{-2x}} - 1$	$(-1, 1)$	
ReLU	$ReLU(x) = \max(0, x)$	$[0, \infty)$	
Leaky ReLU	$LeakyReLU(x) = \max(0.01x, x)$	$(-\infty, \infty)$	

2.3.1.2 Loss Function

The loss function measures the difference between the predicted output and the target value. During training, the network adjusts its weights to minimise this loss. Common loss functions include Mean Squared Error (MSE) for regression tasks and Cross-Entropy Loss for classification tasks. The loss function L can be written as:

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (2.1)$$

for MSE, where y_i is the true value, $\hat{y}_i$ is the predicted value, and N is the number of samples. Alternatively, for cross-entropy, it is defined as:

$$L(y, \hat{y}) = - \sum_{i=1}^N y_i \log(\hat{y}_i) \quad (2.2)$$

where y_i is the true class label (one-hot encoded) and $\hat{y}_i$ is the predicted probability for that class [25].

2.3.1.3 Back Propagation

Backpropagation allows the network to update its weights and biases iteratively, reducing the loss and making the output predictions closer to the actual target values. Weights are typically initialised randomly, while hyperparameters such as the number of layers, nodes per layer, and activation functions are defined beforehand.

Training a neural network using backpropagation typically involves the following steps:

1. **Forward Pass:** Input data flows through the network layers, where each layer performs calculations to generate a predicted output.
2. **Loss Computation:** A loss function provides the model with a numerical value of how wrong the prediction is. For example, the network can compare the actual target with the prediction it came up with when the ground-truth targets are available.
3. **Backpropagation:** The error gradient is back-propagated, and the gradients of each weight are computed from the output layer down to the input layer. These gradients at each layer help dictate how the weights can be adjusted to minimise the error.
4. **Weight Adjustment (Gradient Descent):** The gradients are used to update the weights and biases to minimise the loss at each iteration.

2.3.1.4 Gradient Descent

Gradient descent iteratively minimises the loss functions by adjusting model parameters, weights, and biases. This decreases the error and increases the accuracy of the results. Gradient descent follows a cost function's "downhill" slope towards its minimum value, as Figure 2.2 shows.

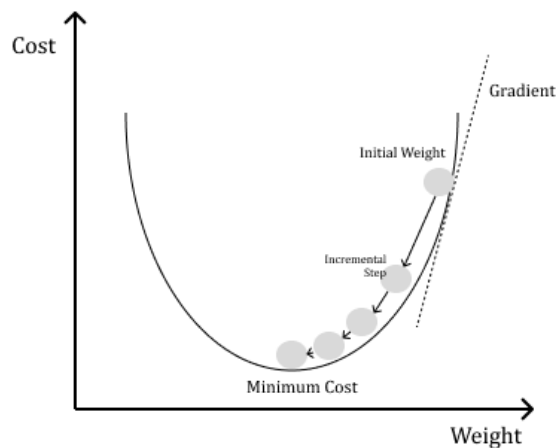


Figure 2.2: Illustration of gradient descent optimisation, showing the iterative process of adjusting weights to minimise the cost function. Each incremental step moves closer to the minimum cost

The effectiveness of gradient descent relies on two factors: the *direction*, obtained from the gradient of the cost function, and the *step size*, obtained from the learning rate η . A proper choice of the learning rate is essential; if it is too high, the gradient descent process might pass the minimum value, while if it is too low, the convergence might never happen or be too slow.

Gradient descent involves three main steps. First, the partial derivative of the cost function for each parameter is calculated, showing how a minor adjustment will affect the loss. Then,

the parameters are adjusted by moving in the direction that reduces the loss using the following equation:

$$\theta := \theta - \eta \frac{\partial L}{\partial \theta} \quad (2.3)$$

Finally, these steps are repeated until convergence or minimal change in the loss function.

There are three main variants of gradient descent:

- **Batch Gradient Descent:** Uses the entire dataset to compute gradients, ensuring stability but requiring significant computation.
- **Stochastic Gradient Descent (SGD):** Calculates gradients for each data point individually, increasing efficiency but adding noise.
- **Mini-Batch Gradient Descent:** Combines stability and efficiency by computing gradients over small batches of data.

2.3.1.5 Supervised, Unsupervised, and Semi-Supervised Learning

Machine learning methods can be classified into supervised, unsupervised, and semi-supervised learning, each suited to different data availability, scenarios and task requirements.

Supervised learning uses labelled datasets, where the model learns the mapping between inputs and outputs by minimising prediction errors using a loss function. These form the basis of several applications, including image classification, segmentation, and anomaly detection, such as identifying anomalies in medical scans or predicting patient prognosis. However, the need for large labelled datasets can pose difficulties, depending on the problem at hand.

Unsupervised learning deals with unlabelled data, discovering patterns or structures. Techniques of clustering and dimensionality reduction are applied to model the underlying distribution of data. In medical imaging, this paradigm helps discover latent features, group similar cases, or assist in exploratory analysis where annotations are unavailable.

Semi-supervised learning combines the best of both worlds. It uses a small labelled dataset together with a large unlabeled one. This is especially applicable to medical imaging, where datasets with labels are often limited in quantity. Pseudo-label techniques can also help models generalise better and improve performances on tasks such as segmentation and outcome prediction, thereby reducing the dependency on extensive annotations.

2.3.2 Convolutional neural networks

Convolutional Neural Networks (CNNs) are a specific type of neural network designed to process structured grid data, such as images. They are characterised by using convolutional layers instead of neuron layers. In contrast to fully connected neural networks, CNNs utilise the same kernel parameters across the entire image, decreasing the number of trainable parameters and enhancing the efficiency of the training process. Depending on the dimensionality of the input and output data, one-, two-, or three-dimensional convolutional kernels may be employed, contributing to the

computational efficiency of CNNs when handling large inputs like images [13]. The architecture of a Convolutional Neural Network (CNN), as shown in Figure 2.3, has three main components: the convolutional, the pooling, and fully connected layers.

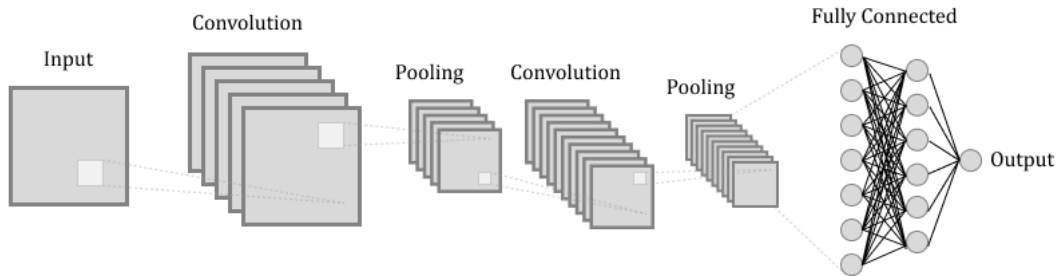


Figure 2.3: General structure of a Convolutional Neural Network (CNN), showing the flow from input through convolutional and pooling layers to fully connected layers for final output.

2.3.2.1 Convolution

The convolution operation gives the CNNs their name and particular functionality. This involves applying a set of filters, also called kernels, that are dimensionally smaller than the input data. These filters slide across the input to produce feature maps highlighting different characteristics in the data, including edges, texture, and patterns. As shown in Figure 2.4, the convolution operation uses a kernel to extract spatial features from the input by applying element-wise multiplication and summation, producing a feature map that retains the most relevant characteristics of the input data. This architecture allows the network to learn spatial hierarchies by detecting simple features in early layers and more intricate structures in deeper ones.

The convolution operation is based on three key ideas: sparse connectivity, parameter sharing, and translational equivariance. [25]

- **Sparse connectivity:** Each filter only interacts with a small local region of the input, the so-called receptive field, and not with the entirety of the picture. This reduces the computational cost and the number of parameters to optimise, which improves the training efficiency. As the network depth increases, these receptive fields expand, enabling the model to capture information from increasingly larger parts of the input and thus enabling the recognition of complex patterns [25].
- **Parameter sharing:** Parameter sharing refers to sharing the same filter at different places in the input data. This allows a feature detected in a particular region of an image to be detected in other areas. This also decreases the total number of parameters, increasing the efficiency [25].

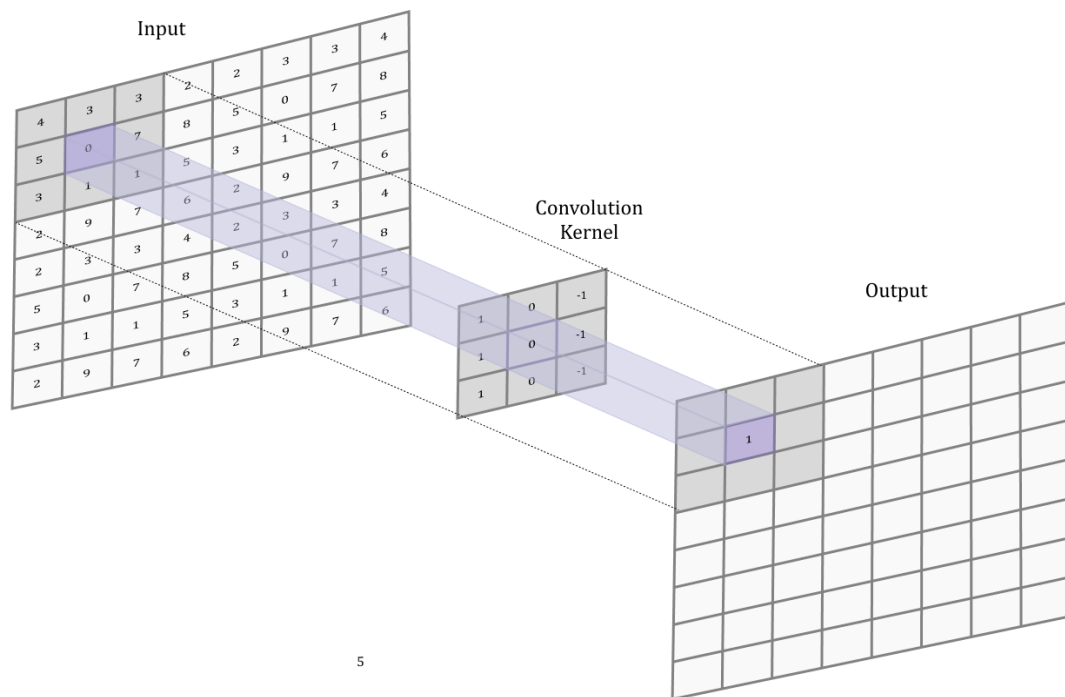


Figure 2.4: Illustration of the convolution operation in a Convolutional Neural Network (CNN). A convolutional kernel slides across the input matrix, applying element-wise multiplication and summation to produce the corresponding output feature map.

- **Equivariance to translation:** if the input changes, the output changes correspondingly. For example, suppose an object in the image moves to the right. In that case, its representation in the feature maps moves to the right so that CNNs can achieve consistent detection over different spatial locations. Still, while convolution provides translation equivariance, it does not naturally allow for other transformations, such as rotation or scaling, sometimes requiring additional tricks or data augmentation [25].

There are different ways to apply the convolution operation: stridden convolution, where the filter shifts by more than one pixel each time; zero-padding, where additional pixels are filled around the border of the input so that the input's spatial dimensions can be retained. In this way, CNNs can control the output size while maintaining computational efficiency and without affecting the network's capability to capture key features.

2.3.2.2 Pooling

Pooling is a common operation performed after convolutional layers to reduce the spatial size of feature maps. Its main purpose is to downsample the input representation and retain the most important features. This helps reduce the sensitivity to small translations and distortions in the input. Therefore, the model increases its robustness to variations and reduces the computational cost on the subsequent layers [25].

Max pooling and average pooling are the most common pooling techniques. Max pooling selects the maximum value in a given region, and average pooling calculates the average value. Max pooling is particularly effective in cases where the presence of a feature is more important than its exact location, as it captures the strongest activation, usually corresponding to the most prominent features. On the other hand, average pooling allows for more uniform representations by considering the mean value, which may be useful for tasks where overall patterns are more relevant than precise peaks. As illustrated in Figure 2.5, pooling operations reduce the spatial dimensions of feature maps.

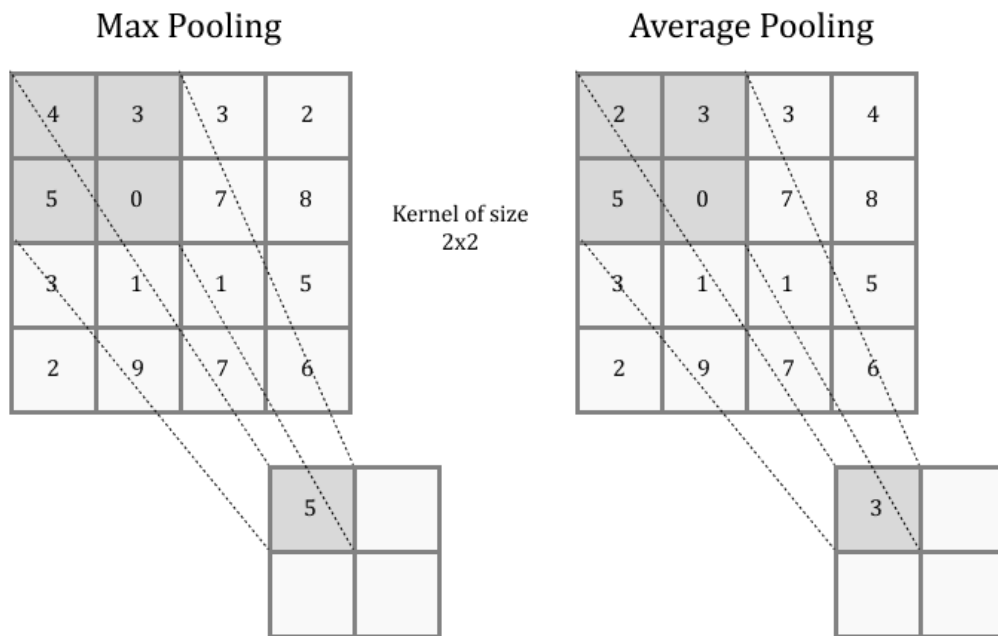


Figure 2.5: Illustration of max pooling and average pooling operations in a CNN. Max pooling selects the highest value in each region to emphasize the most significant features, while average pooling computes the mean value to capture broader patterns. Both methods use a kernel of size 2×2 to reduce spatial dimensions and improve computational efficiency.

Pooling operations can also include a stride, where the pooling window moves by a set number of pixels, allowing for a greater reduction in dimensionality. While this strategy increases efficiency, it can also lead to a loss of spatial resolution. It should be noted that specific architectures use overlapping pooling [69], in which the pooling regions slightly overlap, to balance the decrease in feature map dimensions and the conservation of detailed information.

2.3.2.3 Fully connected Layers

Fully connected layers usually appear at the end of a CNN. Once the features have been extracted from the input data by the convolutional and pooling layers, these layers are part of the neural network that ultimately combines these features to make a final prediction of classification or

regression. They let the network learn complex patterns by incorporating information from all the features extracted earlier in the network.

2.3.3 Autoencoders

Autoencoders (AEs) are a type of artificial neural network that work to learn compressed representations of input data. Their goal is to be able to reconstruct the input at the output stage. AE are feedforward, non-recurrent and depend on backpropagation to capture sparse and essential features from the training data. Thus, the dimensionality of the output vectors is the same as the input. They can be trained in both end-to-end and layer-by-layer manners. This grants extreme flexibility in the training methodology [49].

The major components of an autoencoder (Figure 2.6) are:

- Encoder: Reduces the dimensionality of the input data to a lower-dimensional representation, often called the bottleneck or latent space. Therefore, it learns to find the critical features from the input data while minimising information loss.
- Decoder: The decoder reconstructs the input data from the compressed representation. Its goal is to produce an output value close to the input value.

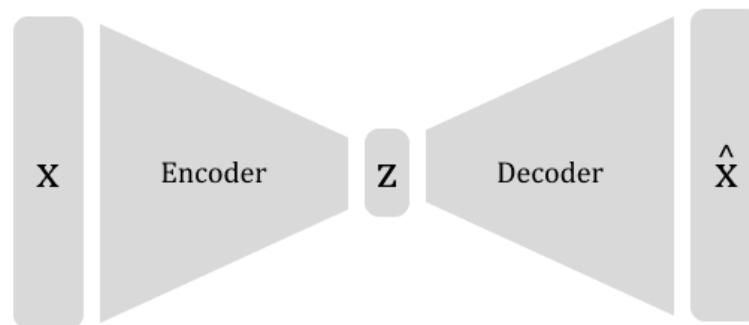


Figure 2.6: Architecture of an Autoencoder, depicting the encoder, bottleneck (latent space), and decoder components.

Training of the autoencoder usually minimises reconstruction loss. Applications of autoencoders range from dimensionality reduction and noise reduction to data generation. However, the autoencoder's disadvantage lies in generating perfect reconstructions only for data similar to the training dataset. There is also reduced fidelity of the reconstructed outputs concerning the input. Autoencoders do not have a generative interpretation either since the model performs only a reconstruction [49].

2.3.4 U-Net

U-Net is a convolutional neural network architecture introduced by Ronneberger et al. in 2015 [66]. It was tailored for biomedical image segmentation and became an important cornerstone in

medical imaging due to its capability of providing exact and accurate segmentations even with scarce training data.

2.3.4.1 Architecture

U-Net uses an encoder-decoder architecture with skip connections (Figure 2.7) between the encoder and decoder layers. This architecture is specified by:

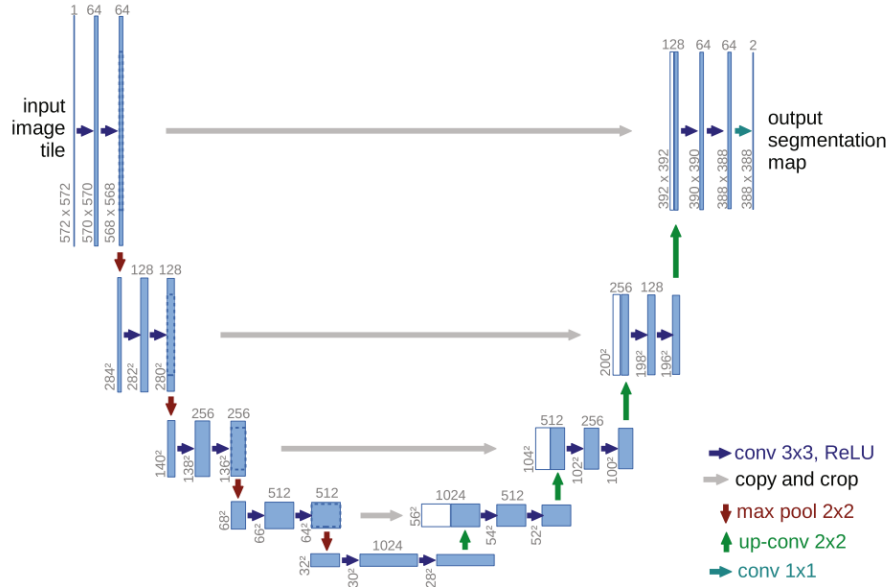


Figure 2.7: U-Net architecture: An encoder-decoder network with skip connections that transfer feature maps from the contracting path to the expanding path [66].

- **Encoder (Contracting Path):** The encoder extracts hierarchical features using a series of convolutional and max-pooling layers, which successively reduces the spatial size while preserving semantic information. Each layer consists of two 3×3 convolutions followed by ReLU activation functions and a 2×2 max-pooling process, which reduces the spatial resolution by half.
- **Decoder (Expanding Path):** The decoder reconstructs the spatial resolution of the image by up-convolutions in a sequence, which are then concatenated with features from the skip connections of the corresponding encoder layers.
- **Skip Connections:** The skip connections directly transfer feature maps from the encoder to the decoder at corresponding levels. This makes sure that the fine-grained spatial information is preserved and reduces the loss of details during downsampling.

The architecture ends with a 1×1 convolutional layer to map the feature maps to the desired number of output classes, typically followed by a softmax or sigmoid activation function, depending on the task.

Since its introduction, U-Net has been extensively used in various medical imaging tasks. In the context of breast cancer, U-Net can assist in segmenting preoperative and postoperative images to simulate aesthetic outcomes, aiding in patient-specific surgical planning.

2.3.5 Transformers

Transformers were first introduced by Vaswani et al. in 2017 [79]. The transformer model radically eliminated recurrent layers in favour of a self-attention mechanism. Thus, it is highly efficient for parallelised processing over large datasets [79].

2.3.5.1 Self-Attention and Multi-Head Attention Mechanism

The self-attention mechanism forms the core of transformers, enabling each part of the input to dynamically "attend" to other parts, weighting their relevance based on context. This mechanism uses three vectors, query (Q), key (K), and value (V), to calculate relevance scores represented as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.4)$$

where d_k is a scaling factor stabilising large matrices' training [79]. Using multiple attention heads, transformers can learn different aspects of the data, enhancing the model's ability to capture diverse relationships within the input.

2.3.5.2 Encoder-Decoder Structure

Encoder-decoder structures are a common architectural pattern in deep learning, also used in models such as autoencoders (see Section 2.3.3). While those models typically aim at reconstruction tasks, the transformer adopts a similar structure for sequence-to-sequence prediction, with key differences in how information is encoded and attended.

The transformer model is based on the two-part encoder-decoder structure (Figure 2.8), consisting of N stacked layers. The encoder maps an input sequence into a series of continuous representations, and the decoder interprets these representations to produce output sequences one step at a time. Each of the layers in the encoder contains two major sub-layers: one is multi-head self-attention, and the other is the position-wise feedforward network; in both cases, it has residual connections followed by layer normalisation [79].

On top of the encoder-decoder structure, decoder layers include a third sub-layer performing multi-head attention over the encoder's output. This design will let the decoder focus on different parts of the input sequence when generating each output token. The important part of the decoder is the masked self-attention mechanism: it ensures that a prediction depends only on already generated outputs, thus keeping the model autoregressive.

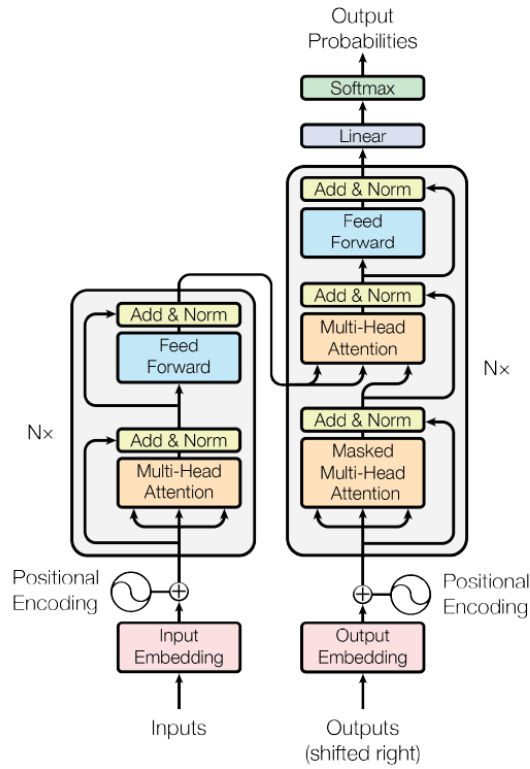


Figure 2.8: Architecture of the Transformer model, displaying the encoder and decoder layers, including multi-head attention, feed-forward networks, and positional encoding, used for processing sequential input data [79].

2.4 Deep Generative models

Generative models attempt to model the underlying data distribution from which the observed data is assumed to be generated. Unlike discriminative models, which predict outcomes by calculating the conditional probability $P(Y|X)$, generative models focus on modelling the joint probability distribution $P(X, Y)$. From this point, one can extract the marginal distribution of inputs $P(X)$ and the conditional distribution of outputs given inputs $P(Y|X)$ [24].

These models assume that the data is sampled from some unknown probability distribution and try to learn an approximation of it. After training, the model can generate new samples resembling the original dataset and estimate the likelihood of different outcomes. Their capability of understanding the whole data distribution makes them useful for applications that are more than just classification, like density estimation, data generation, and imputation of missing values [24].

In the context of breast cancer postoperative outcome prediction, deep generative models offer promising avenues for simulating patient-specific aesthetic results. Models like Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models have shown significant potential in medical imaging applications.

Deep generative models can be broadly categorised based on how they estimate the data distribution, as shown in Figure 2.9. These categories include Explicit Density Models, where the

probability distributions are directly modelled, and Implicit Density Models, where the distribution is learned indirectly.

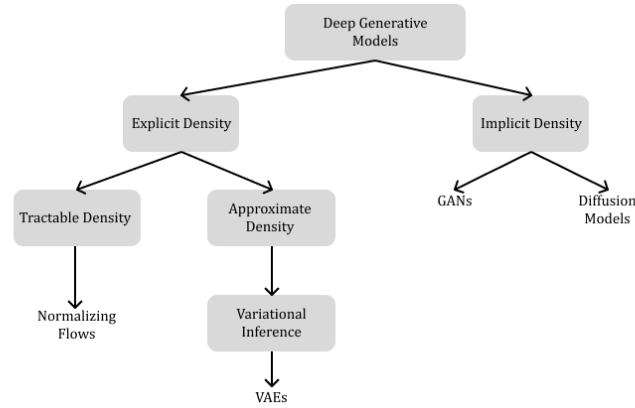


Figure 2.9: Taxonomy of deep generative models, dividing models into explicit and implicit density estimation approaches, with further subdivisions into tractable, approximate, and hybrid models, including GANs, VAEs, Diffusion Models, and Normalizing Flows [50].

However, this research will focus on the most relevant models, VAEs, GANs, and Diffusion Models, due to their applicability in generating realistic medical images and their potential for predicting postoperative outcomes.

- **Explicit Density Models:** Model the probability distribution of data directly. They attempt to compute and understand the underlying patterns in data explicitly. They can be further classified into:
 - **Tractable Density Models:** These models represent the data distribution in a much simpler manner, as their probability is easy to compute, and the probability is "tractable" or manageable.
 - **Approximate Density Models:** If data is too complex to represent accurately, these models estimate an approximation of the probability distribution of the data to learn the patterns in that data. They make smart estimations of the distribution without calculating it directly.
- **Implicit Density Models:** These models do not explicitly calculate the distribution of data; instead, they learn to create samples that look real without ever explicitly computing any probabilities. They do not represent or estimate the probability distribution directly.

The history of generative models reflects the shift from traditional statistical methods to more sophisticated deep-learning techniques. Early methods, as shown in Table 2.2, focused on modelling distributions based on simple probabilistic principles. Modern approaches like VAEs, GANs, and diffusion models have further pushed the boundaries, enabling the generation of realistic images, text, and even 3D structures.

Table 2.2: Comparison of Probabilistic Models in Machine Learning

Model	Description
Hidden Markov Models (HMMs)	Hidden Markov Models represent stochastic processes that proceed through sequences of hidden states, each corresponding to an observable output. The Markov property in HMMs states that the probability of transitioning to the next state depends solely on the current state. Primary components include the State Transition Matrix, which describes the probability of moving between states, and the Emission Matrix, which defines the probability of observing an output given the current state. Applications span fields like speech recognition and bioinformatics [24].
Gaussian Mixture Models (GMMs)	Gaussian Mixture Models cluster data probabilistically by modelling it as a combination of Gaussian distributions, each representing a distinct cluster. GMMs consider the covariance and mean of each Gaussian component, offering flexibility in capturing clusters with different shapes and densities [24].
Boltzmann Machines (BM)	Boltzmann Machines are generative models based on an energy function over binary units, assigning higher probabilities to lower-energy states. Training optimises weights to minimise the energy of observed data, enabling Boltzmann Machines to capture complex data distributions. Unobserved latent variables in BMs further enhance their modelling capability [25].
Restricted Boltzmann Machines (RBMs)	A variant of Boltzmann Machines with connections limited between visible and hidden units only, which simplifies training. RBMs are popular in tasks like feature learning and dimensionality reduction.
Deep Belief Networks (DBNs)	Deep Belief Networks consist of a stack of Restricted Boltzmann Machines, each trained one layer at a time, followed by fine-tuning the full network. DBNs are effective for hierarchical data modelling and are widely used in deep learning applications.
Deep Boltzmann Machines (DBMs)	An extension of RBMs with additional hidden layers to capture complex and hierarchical data structures. DBMs are trained with a layer-wise pre-training process and optimisation through stochastic gradient descent [25].

2.4.1 Variational Autoencoders

Variational Autoencoders (VAEs) are deep generative models that extend the functionality of autoencoders through Bayesian inference, enabling unsupervised learning of complex data distributions. VAEs differ from traditional autoencoders by representing a probability distribution rather than a deterministic point within the latent space, thereby capturing data variability and allowing for controlled data generation.

VAEs consist of several core components - Figure 2.10:

- **Encoder (Recognition Model):** Maps input data to a probability distribution rather than a fixed latent code, capturing both the mean and variance. Moreover, it uses principles of

Bayesian inference to arrive at a posterior distribution, typically Gaussian, that represents the input data in latent space.

- **Reparameterization Trick:** To allow Backpropagation through training, the encoder employs a reparameterisation trick where a sample is drawn from the Gaussian latent-space distribution so that it permits gradients to pass through the network.
- **Decoder (Generative Model):** Takes samples from the latent distribution to reconstruct data similar to the original input. It generates new data samples by sampling from the learned distribution, which results in variations of the original data.
- **Latent Space Representation and Perturbations:** The latent space in a VAE encodes the mean and variance rather than fixed values, creating a continuous, structured representation that facilitates smooth interpolations between samples. VAEs can modify specific attributes of generated outputs by perturbing the latent space vectors. For instance, slight perturbations in specific latent dimensions can modify head orientation in a facial generation application. A small change to a latent vector might result in the generated image shifting from a frontal head position to a slight turn, showcasing the VAE's ability to control feature variations through latent adjustments, as Figure 2.11 shows.

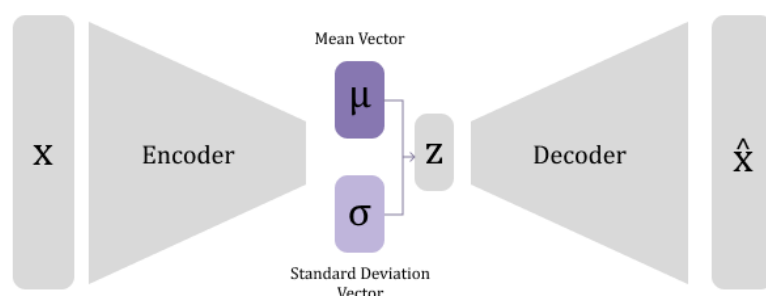


Figure 2.10: Architecture of a Variational Autoencoder (VAE), showing the encoder that generates a mean vector (μ) and a standard deviation vector (σ) to create the latent representation (z), followed by the decoder to reconstruct the input.

The VAE loss function comprises two main terms: reconstruction loss, which guides the decoder in replicating input data accurately, and Kullback-Leibler (KL) divergence, which regularises the latent space by encouraging it to approximate a standard Gaussian. This regularisation improves sampling consistency and latent structure, making the model efficient for downstream tasks such as zero-shot learning and data synthesis.

Although VAEs sometimes generate images with blurred details and may struggle with large, complex datasets, they offer robust applications in image generation, data imputation, semi-supervised learning, and representation learning. By integrating autoencoder architecture with Bayesian inference, VAEs have become valuable tools for generating structured latent embeddings and synthesising realistic, diverse data samples.



Figure 2.11: Example of latent space manipulation in VAEs: A generated face image shows variations achieved by perturbing the latent space, adjusting features such as head tilt and introducing a subtle smile [39].

Table 2.3: Types of Variational Autoencoders (VAEs) and Their Characteristics

Type of VAE	Description and Key Features
Conditional VAE (CVAE)	Extends the VAE by conditioning the latent space on auxiliary variables (e.g., class labels) to control the generation process. Useful for generating data with specific attributes [71].
Beta-VAE	Introduces a weighting parameter (β) to balance the trade-off between reconstruction loss and KL divergence, encouraging disentangled latent representations [29].
Semi-Supervised VAE (SSVAE)	Leverages both labelled and unlabelled data for better generalisation in supervised tasks. [42]
Diffusion VAE	Enhances the generative process using noise diffusion techniques, capturing complex data distributions. [56]

Recent advancements in VAEs, such as hierarchical VAEs [77] and VAE-Gan hybrids [44], have improved their capacity to generate high-resolution images with better detail and less blurriness. VAEs can help simulate different surgical outcomes by manipulating latent variables corresponding to anatomical features, aiding in personalised patient consultations.

2.4.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a powerful class of deep generative models with the ability to generate highly realistic synthetic data resembling the distribution of real-world data. The success of GANs, introduced in 2014 by Ian Goodfellow et al. [26], is due to the unique architecture - Figure 2.12, which consists of two neural networks that engage dynamically in an adversarial training process:

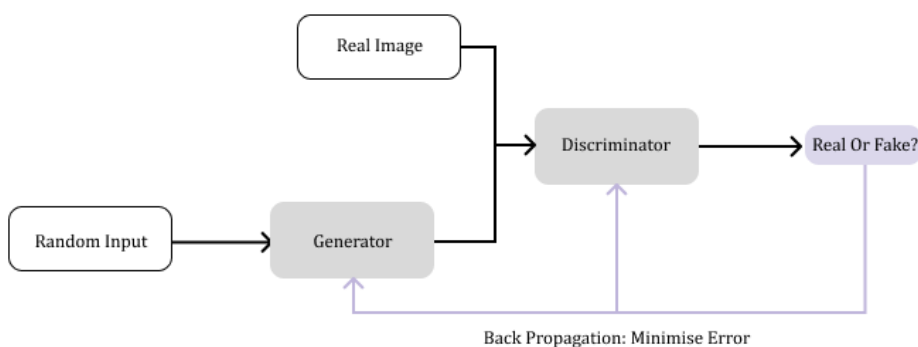


Figure 2.12: Architecture of a Generative Adversarial Network (GAN). The generator creates synthetic images from random input, while the discriminator evaluates whether the input is real or fake. Backpropagation minimizes the error, training both networks to improve their respective outputs.

- **Generator:** This part of the network generates synthetic data. It receives an input of random noise, usually sampled from a simple distribution, such as Gaussian or uniform, and transforms the noise into data samples that could masquerade as the target dataset.
- **Discriminator:** This network functions like a "critic", rating the output from the generator. On any given sample of data, the discriminator will decide whether it is real within the actual dataset or generated through the generator. Although, at first, the discriminator can tell a real sample from a generated one, over time, this becomes increasingly difficult as the generator, through training, produces better and better outputs.

This interaction creates a competitive "game" in which the generator learns incrementally to "fool" the discriminator by producing increasingly valid data. In contrast, the discriminator learns simultaneously to fine-tune its capability to make the call between real and fake. During training, the generator and discriminator optimise different objectives. The generator wishes to minimise the discriminator's accuracy while the discriminator wants to maximise the same. Mathematically, this is captured by the minimax loss function:

$$\theta := \theta - \eta \frac{\partial L}{\partial \theta} \quad (2.5)$$

Here, $D(x)$ is the probability that the discriminator thinks a real sample is authentic, and $D(G(z))$ is the probability that it thinks a generated sample is fake [14]. Convergence implies that the generator produces samples indistinguishable from real data; in this case, the discriminator has no reason to believe that one might be more real than any other probability of saying either is 0.5, which means it has no grounds for discrimination.

In Table 2.4, we summarize key variants of GANs and their unique contributions to advancing generative modelling. These variants demonstrate how the core GAN architecture has evolved to address specific challenges and extend its applicability to various tasks.

Table 2.4: Variants of Generative Adversarial Networks (GANs)

GAN Variant	Key Features and Description
Conditional GAN (CGAN)	Introduces additional conditional information (e.g., class labels) into both the generator and discriminator to control the data generation process. This enables more structured and targeted outputs. [51]
Deep Convolutional GAN (DCGAN)	Utilizes convolutional and transposed convolutional layers in the generator and discriminator, improving training stability and generating higher-quality images. [62]
Wasserstein GAN (WGAN)	Replaces the traditional loss function with the Wasserstein distance to improve convergence and training stability. It helps mitigate mode collapse. [4]
CycleGAN	Enables unpaired image-to-image translation by introducing a cycle consistency loss, ensuring the mapping can be reversed without losing essential features. [86]
Attention GANs	Attention mechanisms have been incorporated into GANs for better quality outputs, particularly in image synthesis tasks requiring fine-grained detail. With embedded attention layers in GAN architectures, Attention GANs allow a generator to pay extra attention to certain regions of the input or previously generated output to enhance coherence and detail. These attention layers help the model focus on essential features and ignore irrelevant parts, enabling more realistic images and complex scenes [49].

2.4.3 Normalizing Flows

Normalizing Flows (NFs) are a class of deep generative models that allow the construction of complex probability distributions by transforming a base distribution, usually a standard Gaussian, through a sequence of reversible, differentiable transformations. This approach simultaneously provides exact density estimation and efficient sampling, making the NF particularly useful when exact probabilistic modelling is a requirement [64].

The underlying concept of Normalizing Flows (NFs) is based on the formula for variable transformations in probability distributions. Given an invertible transformation f that maps a latent variable z from the base distribution $p_Z(z)$ to the observed variable $x = f(z)$, the density of x can be expressed as:

$$p_X(x) = p_Z(f^{-1}(x)) \left| \det \frac{\partial f^{-1}(x)}{\partial x} \right| \quad (2.6)$$

Here, $\det \frac{\partial f^{-1}(x)}{\partial x}$ is the determinant of the Jacobian matrix for the inverse transformation, capturing the change in volume induced by f . Through the composition of a sequence of transformations $f_1, f_2, \dots, f_K$, where each is described as invertible and differentiable, NFs can represent distributions with considerable flexibility. The overall transformation is written as $f = f_K \circ f_{K-1} \circ \dots \circ f_1$, and the density resulting from this operation is:

$$p_X(x) = p_Z(f^{-1}(x)) \prod_{i=1}^K \left| \det \frac{\partial f_i^{-1}(x)}{\partial x} \right| \quad (2.7)$$

Training NFs involves maximizing the likelihood of the observed data under the model. This is equivalent to setting the parameters of the transformations such that the transformed base distribution is aligned with the target data distribution.

Several architectures have been proposed:

- **Planar and Radial Flows:** Introduced by Rezende et al. [64], these flows achieve simple yet powerful transformations in the latent space, enhancing the flexibility of variational inference models.
- **Real NVP (Real-valued Non-Volume Preserving):** Introduced by Dinh et al. [21], this architecture utilizes coupling layers that enable the efficient calculation of the Jacobian determinant, thereby supporting scalable density estimation.
- **Glow:** Proposed by Kingma et al. [41], this model builds upon Real NVP and incorporates invertible 1×1 convolutions for high-quality image synthesis.
- **Masked Autoregressive Flows (MAF):** Introduced by Papamakarios et al. [58], this architecture combines autoregressive models with NFs to retain tractable likelihood estimation and model complex dependencies.

NFs have been used successfully in many fields, including density estimation, variational inference, and generative modelling of complex data such as images and audio. The ability to provide exact likelihoods and efficient sampling render them a powerful tool in probabilistic modelling. However, challenges still exist, mainly related to the computational cost of maintaining invertibility and computing the Jacobian determinants, which is worse in high-dimensional settings.

2.4.4 Diffusion Models

Diffusion, introduced by Ho et al. [30], was developed to overcome the limitations of adversarial networks and autoencoders. They allow for stable and robust training in a process that changes data from simple distributions, for example, Gaussian noise, into complex data distributions; this process is called denoising [9].

A diffusion model generally consists of two main processes: a forward process and a reverse process.

- The forward process does not require a deep learning model because it involves a deterministic Markov chain that progressively adds Gaussian noise to the data. This chain is parameterized by predefined schedules for variance (β_t) at each step. These schedules can be linear or cosine-optimized to achieve better-quality samples [30].
- Conversely, the reverse process denoises this representation step by step and reconstructs data samples that closely approximate the original distribution. This process relies on a neural network, typically a U-Net architecture, to predict the added noise ϵ_t at each timestep. At each step, the U-Net takes as input a noisy image x_t and the corresponding timestep t , then predicts the noise component that was added during the forward process [30].

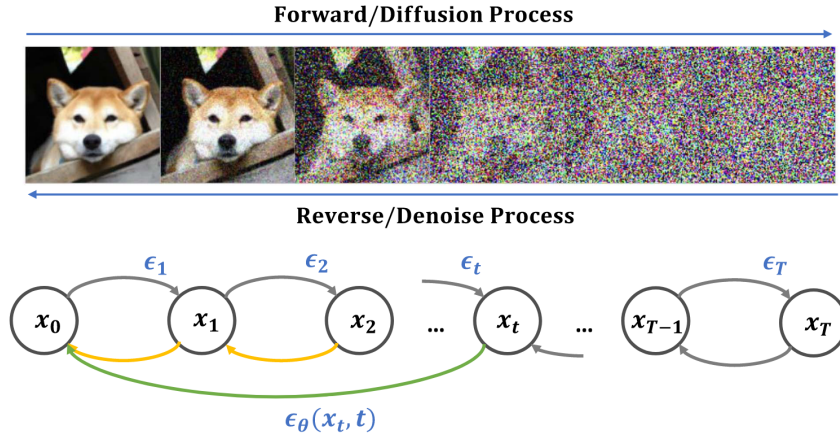


Figure 2.13: Illustration of the forward/diffusion and reverse/denoising processes. The forward process involves adding noise to an image step-by-step, while the reverse process reconstructs the original image by removing noise in reverse order [9]

In the forward process, the step from one noisy state to the next is defined by predefined noise schedules, which are linear or cosine-optimised for better-quality samples. That would be notated as,

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (2.8)$$

where β_t is the noise coefficient at step t .

The sampling process in the reverse direction is as follows:

1. The process begins with an image that is entirely composed of Gaussian noise that is then passed to the model (typically a U-Net) along with the corresponding timestep $t = T$.
2. The U-Net predicts the noise ϵ_t present in the noisy image at the given timestep, which is subtracted from the noisy image, resulting in a partially denoised version x_{t-1} .
3. To ensure stability, a small amount of Gaussian noise is reintroduced to x_{t-1} after the denoising step.
4. Repeat the denoising and noise reintroduction steps iteratively for all timesteps, moving backwards from $t = T$ to $t = 0$.
5. At $t = 0$, the process produces the final denoised image x_0 , which is a sample from the original data distribution.

The training objective typically minimises the variational bound on negative log-likelihood, simplified through mean squared error losses computed between the model's predicted and actual noise levels.

One variant of diffusion models is the Conditional Diffusion Model. This model incorporates additional conditioning during samplings, such as class labels or text embeddings, to guide sample generation. Some conditioning methods include classifier-guided [20] and classifier-free guidance [31]; both guide the sampling process with learned condition representations.

Diffusion models have been successfully applied across various fields, demonstrating their versatility:

- **Image and Video Generation:** Diffusion models excel in generating high-quality, photorealistic images and videos, supporting applications like image synthesis, super-resolution, and image-to-image translation.
- **Medical imaging:** Diffusion models can generate synthetic images to train on and form high-quality reconstructions of noisy and incomplete data. Applications such as MRI reconstruction and medical image translation are well-matched for this technology [19].

2.5 Discussion

Generative models have the potential to address the challenges of predicting the aesthetic outcomes of breast cancer surgery. If they can produce realistic, personalised post-operative images, they can help patients through increased understanding and decision-making. Still, training generative models involves many challenges, many of which relate to the complexity and sensitivity of medical applications.

The first major limitation is access to high-quality annotated datasets. Without diverse data, models can be challenged to adapt to the wide range of anatomies and responses to treatments seen in clinics.

Another common problem is mode collapse, especially in GAN-based techniques. In this case, a model may not produce diverse results, seriously deteriorating its ability to portray various surgical possibilities. Moreover, overfitting is likely to occur while training the network on small datasets because a model that learns specific patterns within the training data is inclined to acquire features not generally applicable.

Reconciling realism with specificity is challenging. For generative models to be successful, the generated images must be visually persuasive and faithful to the characteristics of patients and surgical outcomes.

Despite these challenges, advances in deep learning techniques, including transfer learning, semi-supervised methods, and improved loss functions, increasingly offset these limitations. Iterative improvements in the range of datasets and optimisation in generative model architectures will further increase their applicability. Ultimately, these models hold the potential to redefine patient-centred care by providing personalised insights that place individuals in the driver's seat and improve the shared decision-making process.

Chapter 3

Literature Review: Image-to-Image translation

Now that the background on artificial intelligence, deep learning, and generative models has been established, let's focus on a specialised application: image-to-image translation.

3.1 Introduction

The concept of image-to-image translation has significantly evolved, starting from the pioneering work by Hertzmann et al. [28], who introduced the idea of "image analogies". This approach could translate images using non-parametric models based on paired images, focusing on replicating visual features and textures between similar photos.

Image-to-image translation converts an input image X from a source domain $\mathcal{X}$ to produce a corresponding output image Y in a target domain $\mathcal{Y}$. This transformation must preserve essential content, such as structure and geometry, while altering characteristics to reflect the target domain's characteristics. This task has many applications - Table 3.1- including style transfer, image colourisation, super-resolution, and data augmentation.

Before discussing the different techniques involved, it is essential to define some key concepts of image-to-image translation [2]:

- **Unimodal Image-to-Image Translation:** Unimodal translation refers to cases where only one output can be generated based on a given input. For example, in super-resolution, the input is a low-resolution image, and the model generates a single high-resolution version of that specific image. In this case, the mapping from the source to the target is deterministic and provides only one expected outcome.
- **Multimodal Image-to-Image Translation:** In multimodal translation, several ways of producing outputs from one given input could exist due to differences in style or variations.

Table 3.1: Image-to-Image Translation Techniques and Applications

Technique	Description	Applications
Inpainting	Filling in missing or corrupted parts of an image using contextual information.	Photo restoration, object removal.
Sketch-to-Image	Translating hand-drawn sketches into photorealistic images by learning structural and textural mappings.	Design prototyping, digital art.
Image Colourisation	Converting grayscale images to plausible colour representations by learning colour distributions.	Restoring historical photos, stylisation.
Style Transfer	Changing the artistic style of an image while preserving its content using domain-specific features.	Photo editing, artistic generation.
Super-Resolution	Enhancing the resolution of low-resolution images while preserving details.	Surveillance, medical imaging.
Domain Adaptation	Translating images between domains to adapt features for tasks like classification or detection.	Medical imaging, autonomous driving.
Semantic Image Synthesis	Generating photorealistic images from semantic maps or labels, maintaining spatial structure.	Urban planning, game asset creation.
Photorealistic Translation	Transforming synthetic images into real-world images for realistic visualization.	Virtual reality, training data generation.
Face Attribute Editing	Modifying specific facial attributes (e.g., age, hair colour) while preserving identity features.	Entertainment, virtual try-on systems.
Pose-to-Image Translation	Generating images of humans in specific poses given skeletal or pose key-point inputs.	Animation, motion analysis.
Image Decomposition	Splitting images into components such as shape and texture for independent manipulation.	Material modelling, content generation.
Cross-View Translation	Generating images from different viewpoints (e.g., aerial to street-level views).	Mapping, navigation systems.

For instance, a summer landscape translated to a winter scene could have multiple representations based on the intensity of the snow colour temperature, among other atmospheric effects.

- **Domain Independent Features:** Domain-independent features are shared properties between domains that do not change during translation. These include edges, contours, and spatial relationships within an image that should remain constant across changes in style or other visual properties. In this instance, translating a photograph of a house from a summer scene into that of a winter scene keeps the shape and structure of the house constant while its colours and textures change.
- **Domain-specific features:** Domain-specific features are only specific to the source/target domain that are altered during translation. These include the brushstrokes or colour palettes that translate from photographs to artistic paintings that characterise the art style. These

features define what sets the translated image apart from the original and place it into the target domain.

- **Two-Domain Translation:** Two-domain translation refers to tasks where transformations occur within a single domain or involve mapping between two specific domains (e.g., grayscale to colour, summer-to-winter landscapes). Models optimised for these tasks generally have simpler architectures tailored for particular mappings.
- **Multi-Domain Translation:** Multi-domain translation encompasses tasks where transformations occur across multiple domains using a unified model, such as translating between multiple weather conditions or artistic styles. Multi-domain techniques are more versatile and scalable, enabling simultaneous learning across various domains while preserving content consistency.
- **Paired Image-to-Image Translation:** Paired image translation involves the training data available as pairs of corresponding images from source and target domains. Unfortunately, paired datasets are highly expensive and challenging to acquire for many real-world applications.
- **Unpaired Image-to-Image Translation:** Translation is said to be unpaired if the source and target images do not have one-to-one correspondence while training. Instead, the model learns to transform using two independent sets of images from each domain. This approach significantly broadens the applicability of image-to-image translation to domains where paired data is unavailable.

3.1.1 Image-to-Image Translation Methods

Image-to-image translation methodologies have quickly diversified into many strategies for single-domain, multi-domain, and multimodal tasks. A very recent comprehensive survey by Pang et al. [57] categorises these methods, emphasising their respective advantages, application scope, and the type of supervision required. Tables 3.2 and 3.3 summarise these methods by classifying them into two-domain and multi-domain approaches and detailing their properties, including the necessity of paired data and their capability to deal with multimodal outputs.

3.1.1.1 Pix2Pix

The *Pix2Pix* model, introduced by Isola et al. [33], was a significant advance in supervised image-to-image translation. This method leverages cGANs for automatically learning cross-domain relations of paired images. Most traditional image processing-based methods rely on a specific loss function for a particular image processing task. On the other hand, Pix2Pix established a generic system that enhances the loss function with adversarial training. It makes it versatile enough to do things like image decomposition or object changes by directly manipulating pixel information from place to place.

Table 3.2: Two-Domain Image-to-Image Translation Techniques

Supervision Type	Technique	Description and Reference	Multimodal?
Supervised	Pix2Pix	Maps paired domains using conditional GANs and L1 loss [33].	No
	Pix2PixHD	High-resolution extension of Pix2Pix for photorealistic translations [81].	No
	SPADE	Uses spatially-adaptive normalisation for semantic synthesis, enabling precise control of spatial details [59].	No
	BicycleGAN	Combines cVAE-GAN and cLR-GAN for generating diverse multimodal outputs conditioned on paired data [85].	Yes
	PixelNN	Translates using nearest-neighbour matching for localised control of visual details [5].	Yes
	DMT	Lightweight module integrating pre-trained DDPMs for efficient domain-specific feature transfer [82].	No
	VQBB	Employs Brownian Bridge processes and VQGAN latent spaces for domain mapping, improving accuracy and diversity [46].	Yes
	PITI	Combines pre-training and diffusion models to adapt pre-trained latent spaces for diverse image-to-image tasks [80].	Yes
	Palette	A generalist image-to-image diffusion model achieving state-of-the-art results on tasks like colourisation, inpainting, uncropping, and JPEG restoration [67].	No
Semi-Supervised	TCR	Combines labelled and unlabeled data with regularisation to improve generalisation in low-data scenarios [54].	No
	Asymmetric Gradient Guidance (AGG)	Guides diffusion processes with asymmetric updates, balancing structural preservation and style transfer [43].	Yes
Unsupervised	CycleGAN	Uses cycle-consistency loss for unpaired translation, ensuring content preservation [86].	No
	CycleGAN	Uses cycle-consistency loss for unpaired translation, ensuring content preservation [86].	No
	UNIT	Models shared latent space for unpaired domains, enabling smooth translations [48].	No
	DRIT	Disentangles content and style to enable diverse and multimodal outputs [45].	Yes
	pix2pix-zero	Enables zero-shot translation by learning latent correspondences without manual prompting [60].	No
	UVCAN v2	Combines modern GAN advancements with cycle-consistency for robust unpaired translation [74].	No
	Latent Diffusion Models (LDMs)	Operate in a compressed latent space to reduce computational costs while maintaining high-quality outputs [65].	Yes
	UNIT-DDPM	Combines DDPMs with a shared latent space for domain-agnostic translation without adversarial training [68].	Yes
	EGSDE	Employs energy-guided SDEs to balance realism and faithfulness in unpaired translations using domain-specific energy functions [84].	Yes

Pix2Pix has two main components: a *generator* and a *discriminator*. The generator is designed as a U-Net, and the discriminator uses a *PatchGAN* architecture, where each little patch of the image is classified as either "real" or "fake". Unlike traditional GAN discriminators that analyse the entirety of an image in one take, PatchGAN looks at local patches (e.g. 50x50 pixels) with an advantage in capturing fine-grained textures and patterns.

The Pix2Pix offers a flexible framework for transforming images across domains without requiring custom features or task-specific losses. In the breast cancer domain, this model may learn complex mappings between pre- and post-operative photos, capturing essential details of the

Table 3.3: Multi-Domain Image-to-Image Translation Techniques

Supervision Type	Technique	Description and Reference	Multimodal?
Semi-Supervised	AGUIT	Handles multi-domain/multimodal translation using attribute labels and shared latent spaces [47].	Yes
Unsupervised	StarGAN	Performs multi-domain translation using domain labels for control [18].	No
	StarGAN v2	Introduces multimodal domain translations [17].	Yes
	DRIT++	Extends DRIT with improved disentanglement and diversity for cross-domain translation [45].	Yes
	SCONE-GAN	Combines semantic contrastive learning with GANs for highly realistic multi-domain outputs [1].	No
	Stable Diffusion	Pretrained diffusion models operate on latent representations, enabling multi-domain adaptability and efficient fine-tuning [65].	Yes
	Dual Diffusion Implicit Bridges (DDIB)	Applies independent diffusion models for each domain, utilising optimal transport principles for privacy-preserving and high-quality translations [73].	Yes

anatomy to provide personalised visual predictions.

3.1.1.2 BicycleGAN

The *BicycleGAN* model, proposed by Zhu et al. [85], introduced a multimodal approach to image-to-image translation, addressing the limitation of deterministic output in traditional models like Pix2Pix. It combines Conditional Variational Autoencoder GAN (cVAE-GAN) and Conditional Latent Regressor GAN (cLR-GAN). This hybrid model enables multimodal output generation by encouraging a bijective mapping between the input, latent space, and output, thus reducing mode collapse and avoiding deterministic outputs.

BicycleGAN comprises a generator, a discriminator, and an encoder. The generator uses a U-Net architecture and takes both the input image and a latent code to produce multimodal outputs. The encoder maps generated images back to the latent space while the discriminator learns to classify real and generated images across variations. This architecture effectively integrates two complementary objectives from cVAE-GAN and cLR-GAN, enhancing diversity and realism.

Unlike Pix2Pix, which produces a single deterministic output, BicycleGAN generates diverse, plausible images for a given input, thus overcoming mode collapse - Figure 3.1.

3.1.1.3 Diffusion Model Translator

The Diffusion Model Translator (DMT), as introduced by Xia et al. [82], adapts Denoising Diffusion Probabilistic Models (DDPMs) for image-to-image translation and shows improved efficiency. Unlike traditional methods that inject information at every denoising step, DMT performs domain translation at a single, optimised time step, which reduces computational complexity.

DMT is implemented based on a simplified translation module mapping source and target domains, maintaining the main content while adjusting the characteristics of the target domain, as Figure 3.2 shows.



Figure 3.1: Results of the model BicycleGAN. The left column shows the input. The second shows the ground truth output. The final four columns show randomly generated samples [85].

Although it works very well in tasks such as image stylisation, colourisation, and segmentation-to-image translation, relying on pre-trained diffusion models may not generalise well in highly specific or completely new datasets, such as those from medical imaging applications requiring customised training data.

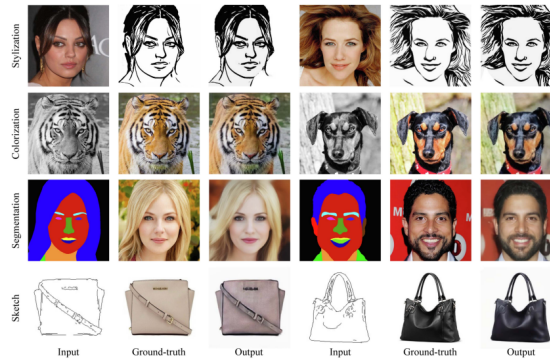


Figure 3.2: Examples of image-to-image translation tasks across four domains [82].

3.1.1.4 Dual Diffusion Implicit Bridges

The Dual Diffusion Implicit Bridges (DDIB) method, proposed by Su et al. [73], is a new image-to-image translation approach using independent diffusion models for source and target domains. Unlike traditional methods that require paired training on source-target datasets, DDIB disentangles this dependence, enabling privacy-preserving and modular domain adaptation. Figure 3.3 shows some of the results.

Some key features are:



Figure 3.3: Examples of image-to-image translation in various applications. Left: how an image of a different class in ImageNet preserves the salient shape. Right: transformations of a tiger image into other animals (husky, lion, bear).

- **Independent Training:** DDIB uses different diffusion models for the source and target domains; hence, jointly training on paired datasets is unnecessary.
- **Cycle Consistency:** The method guarantees near-perfect cycle consistency; translating from source to target and then back again will almost return the original image with very small errors.
- **Privacy Protection:** Because training is conducted individually, DDIB is able to maintain the privacy of sensitive data for each area. This makes it suitable for applications requiring privacy protection, such as medical imaging.

It prevents problems such as mode collapse and unstable training, which may occur with GAN-based methods. Its modular design makes it easy to use pre-trained models for different tasks, thus saving considerable unnecessary computing. This approach is also grounded in solid optimal transport and entropy regularisation theories.

In applications involving predictions of surgical outcomes, DDIB provides privacy-preserving domain adaptation and precise structural preservation. Working without paired data makes it a feasible option when obtaining paired medical datasets is impractical.

3.1.1.5 Asymmetric Gradient Guidance

The Asymmetric Gradient Guidance (AGG) approach proposed by Kwon and Ye [43] takes the reverse process of diffusion sampling and enhances this with a two-step guidance mechanism that combines an initial manifold constraint gradient (MCG) step with a fine-tuning step using lightweight decomposed diffusion sampling (DDS).

AGG works in a manner that first applies MCG for alignment with the noisy manifold of the source image; then, it uses Tweedie's formula to estimate the clean samples, which DDS further refines to optimise target structural and style attributes. This hybrid approach reduces

computational overhead while improving stability. Some of its applications can be seen in Figure 3.4.

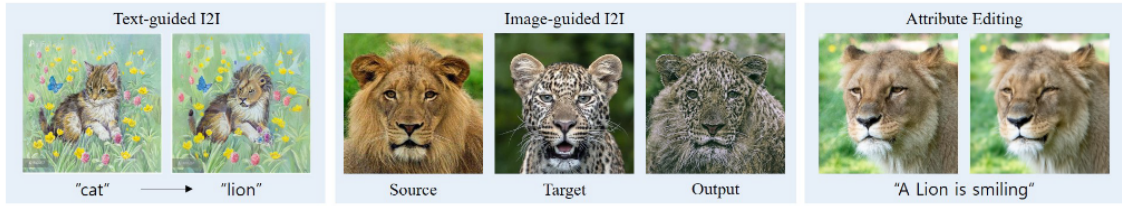


Figure 3.4: Image translation examples of the AGG method [43]

Given its ability to balance structure preservation and domain adaptation, AGG could be particularly useful in medical imaging, where maintaining anatomical accuracy while adapting to different modalities or post-operative appearances is critical.

3.1.1.6 Vector Quantised Brownian Bridge

Vector Quantised Brownian Bridge (VQBB) was presented by Li et al. [46] as an image-to-image translation approach that incorporates the process of Brownian Bridge diffusion in the latent space of a pre-trained VQGAN. Operating in a quantised latent space offers better generalisation and efficiency, with significantly reduced computational overhead compared to conventional diffusion methods.

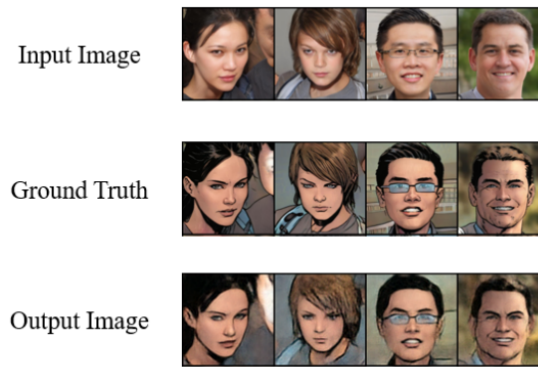


Figure 3.5: VQBB results on face2comic dataset [46].

The proposed VQBB leverages the Brownian Bridge to map between source and target domain latent representations (Figure 3.5). Concretely, this approach first encodes images into the corresponding latent spaces using separate domain-specific VQGANs for both domains. Then, using the Brownian Bridge, these latent spaces are transformed smoothly and accurately into the target domain and decoded into that target domain.

3.1.1.7 AGUIT

Attribute-guided unpaired Image-to-Image Translation (AGUIT) was proposed by Li et al. [47] and uses semi-supervised learning with AGUIT that combines labelled data with unlabeled ones, hence allowing robust training and diverse image outputs.

The generator network G takes the content code C and style code S as input, combining them through Adaptive Instance Normalisation (AdaIN) layers to produce the output image. The style code is further decomposed into two parts:

1. *Noise Part*: A random vector drawn from a Gaussian distribution, introducing diversity in generated outputs.
2. *Attribute Part*: A vector guided by domain labels to enforce specific stylistic attributes.

AGUIT supports multimodal translation by changing style codes to get diverse outputs. It also promotes multi-domain translation by handling multiple domains with a single model. It also generalises exceptionally well, allowing the attributes learned to be applied to previously unseen domains.. This approach decreases the dependence on vast amounts of labelled datasets and adds more interpretability and control to the translation tasks. As AGUIT introduces training complexity due to its disentangled representation, this is a very flexible and powerful solution for unpaired image-to-image translation.

3.1.1.8 StarGAN

The *StarGAN* model, introduced by Choi et al. [18], can perform transformations over multiple domains using only one generator and discriminator network (Figure 3.6). StarGAN allows parallel training with several datasets with different domain labels to increase its flexibility in complex image translation tasks.

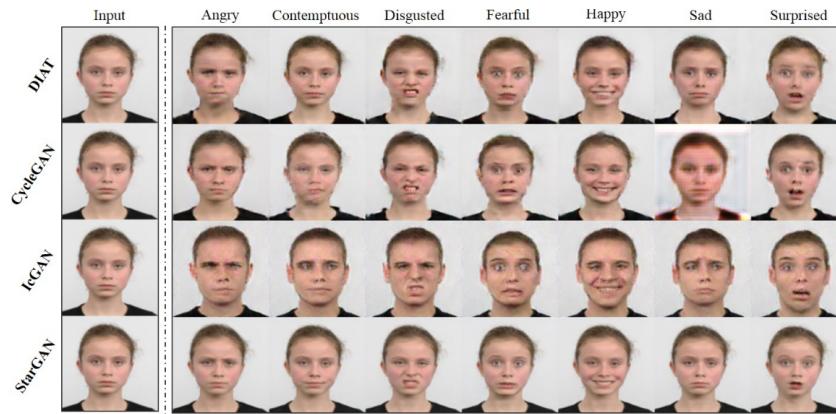


Figure 5. Facial expression synthesis results on the RaFD dataset.

Figure 3.6: Comparison of StarGAN with other methods in facial expression synthesis results on the RaFD dataset. [18]

StarGAN works in a conditional GAN setup, with only one generator G and one discriminator D for multi-domain image-to-image translation. The generator's input is an image x and the target

domain label c , a one-hot or binary vector. During translation, G takes the target domain label and translates the image to the target domain, preserving the fundamental content of the original input image. The discriminator D , armed with an auxiliary classifier, has a twofold task: it not only discriminates between real and fake images but also makes correct domain predictions for real images.

3.2 Discussion

Predicting aesthetic outcomes in breast cancer surgery is best approached using paired image training, which ensures accurate correspondences between pre-operative and post-operative images and allows for the capture of finer details such as scars or skin discolouration. However, when paired datasets are scarce, semi-supervised or unpaired methods provide feasible alternatives. Domain adaptation techniques are particularly relevant for adjusting pre-operative images to approximate post-operative outcomes, while semantic image synthesis maintains anatomical and spatial fidelity during the transformation. Additionally, methods inspired by face attribute editing, which alter specific features while preserving overall identity, can be adapted to construct personalised patient predictions.

Various methodologies can be used to generate post-operative results, each with unique advantages. For instance, unimodal approaches produce a single post-operative prediction, offering simplicity and clarity. Conversely, multimodal methods generate multiple plausible outcomes, embracing the variability inherent in surgical results and presenting patients with diverse scenarios.

From a domain perspective, the problem of generating post-operative images for breast cancer treatments can be framed as either single-domain or multi-domain. Single-domain approaches rely on conditioning variables, such as surgical details and patient clinical history, either as direct inputs or encoded in a latent space. These methods are efficient when the variability in outcomes is limited. In contrast, multi-domain approaches account for significant divergences between surgical contexts by treating them as distinct domains, allowing for greater specificity in adaptations. However, this added complexity increases data requirements and the need for careful domain control.

A key challenge in paired image-to-image translation lies in acquiring paired datasets, particularly in medical imaging, where data availability and privacy concerns limit collection. Consequently, most methods are developed for unpaired data, utilising domain labels and cycle consistency constraints to approximate transformations between domains. Semi-supervised approaches bridge this gap by integrating labelled and unlabelled data, enhancing model generalisation and robustness.

Numerous image-to-image translation techniques offer valuable tools, each suited to different aspects of the problem. For example, StarGAN, initially designed for unpaired datasets, can be adapted to paired scenarios by incorporating image consistency into its loss functions, achieving realistic and controlled transformations. Similarly, Pix2Pix can be extended with domain labels to improve flexibility and accuracy in medical applications. Generating post-operative results

involves a careful balance of methodologies tailored to meet the specific requirements of each task.

Chapter 4

Literature Review: Generation of Post-Operative images

Building on the background of deep learning, generative models, and image-to-image translations, this chapter reviews advancements in creating post-operative images, from regression-based techniques to deep learning approaches like GANs and disentangled representation learning. It analyses each method's strengths and limitations, highlighting their clinical relevance and computational challenges.

4.1 Introduction

Breast cancer treatment is a multifaceted decision-making process influenced by several factors, such as the type and stage of the tumour, along with the age and overall health condition of the patient. More advanced stages of breast cancer are usually treated more aggressively. Treatment options can be divided into three main classes: surgical treatment, radiation therapy, and systemic therapies, each targeting different aspects of this disease [12].

Surgical options, Figure 4.1, include mastectomies, where the whole breast is removed, and breast-conserving surgery (BCS), otherwise known as lumpectomy, where the tumour and a margin of normal tissue surrounding it are excised.

Radiation therapy is often an adjunct to surgery, much more so in breast-conserving treatment. This treatment involves the use of high-energy beams that target and destroy any residual cancer cells in the breast, chest wall, or axillary area.

Systemic therapies, including chemotherapy, hormone therapy, targeted therapy, and immunotherapy, provide a systemic approach by treating the entire body. Chemotherapy targets cells growing more rapidly, while hormone therapy reduces estrogen levels to inhibit the growth of hormone-sensitive cancers.

Significant advancements have been made in the treatment of breast cancer; however, the psychosocial impacts of surgery remain a critical concern. Patients who undergo a mastectomy without reconstruction are at a higher risk of developing depression compared to those who have

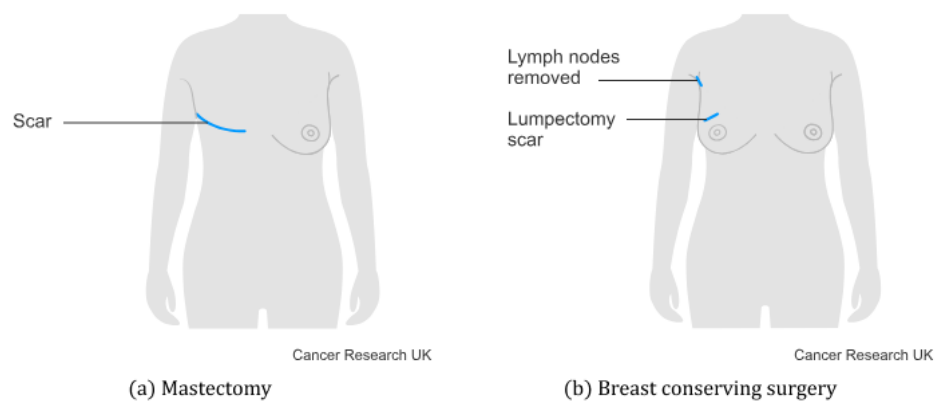


Figure 4.1: A comparison of surgical scars: (a) Mastectomy [8] and (b) Breast-conserving surgery [7].

reconstruction [55]. Additionally, even after breast conservation therapy, many patients experience aesthetic dissatisfaction, which can negatively affect their quality of life and mental well-being [72].

Therefore, patients need to be fully informed about the various possible outcomes of their treatment. One potential solution is to create personalised and realistic visualisations of these outcomes using generative models. These models can simulate the effects of different surgical options based on individual patient characteristics. By providing tailored visualisations, managing their expectations, and making informed decisions, patients can better understand potential aesthetic results.

This has been the subject of great research, ranging from trying to find an objective way to classify aesthetic surgical outcomes, since most traditional methods rely on subjective evaluations by patients or clinicians, to being able to recreate a viable post-operation image.

4.2 Objective Aesthetic Evaluation Metrics

Aesthetic evaluation of breast cancer surgery outcomes has been the subject of investigation. Many objective metrics have been developed to quantify breast asymmetry and standardise the assessments. Key metrics focus primarily on measures for the relative position and contour of the breasts, with particular emphasis on the nipple position and the shape of the breast. Standard metrics used are:

- **Breast Retraction Assessment (BRA):** Pezner et al. [61] described this measure as the horizontal and vertical displacement of the nipples relative to the sternal notch.
- **Breast Compliance Evaluation (BCE):** Introduced by Tsouskas et al. [75] BCE calculates the vertical displacement of each nipple with respect to the lowest point of the breast contour.

- Horizontal Nipple Retraction (HNR): Freitas et al. [22] metric calculates the distance from each nipple to the external contour of the breast, thereby isolating horizontal displacements from other spatial changes.
- Upwards Nipple Retraction (UNR): Similarly to HNR, Van Limbergen et al. [78] proposed UNR as a metric to measure vertical retraction by calculating the vertical distance between the nipples, measuring the degree of upward displacement.
- Lower Breast Contour (LBC): Also proposed by Van Limbergen et al. [78], the LBC refers to the vertical distance between the lowest points of each breast's contour and reflects asymmetry at the base of the breast.
- Breast Contour Difference (BCD): Cardoso et al. [11] developed BCD to evaluate the contour length difference between the two breasts, a measure sensitive to overall breast shape changes.
- Breast Area Difference (BAD): This is another measure suggested by Cardoso et al. [11], which calculates the difference in area between the two breasts, thus giving a volumetric asymmetry measure.
- Breast Overlap Difference (BOD): To further capture asymmetrical characteristics, BOD measures the extent of overlap in contour between the two breasts, offering a metric sensitive to surface-level differences in breast shape [11].

4.3 Image transformation in breast cancer scenarios

The transformation of breast images for aesthetic evaluation and surgical planning has evolved significantly over the years. This section reviews key contributions in the field and critically analyses methodologies across traditional computer vision, deep learning, and regression-based approaches.

4.3.1 Regression-Based Deformation Prediction

Zolfagharnasab et al. [87] proposed a regression-based method for predicting breast deformations after Breast Conserving Surgery (BCS). Their approach was based on a 3D finite element model of the breast obtained from MRI data to simulate surgical deformations. They employed machine learning techniques, Random Forests (RF) and Gradient Boosting Regression (GBR), to model the spatial displacement of points on the 3D breast surface. Figure 4.2 shows one of the research results.

While this study enabled a computationally efficient way to model deformations, its reliance on synthetic data during training limits its generalizability to real-world clinical cases. Additionally, the study does not address broader aesthetic outcomes such as scarring, skin texture, or changes in the nipple, as it focuses solely on geometric deformations.

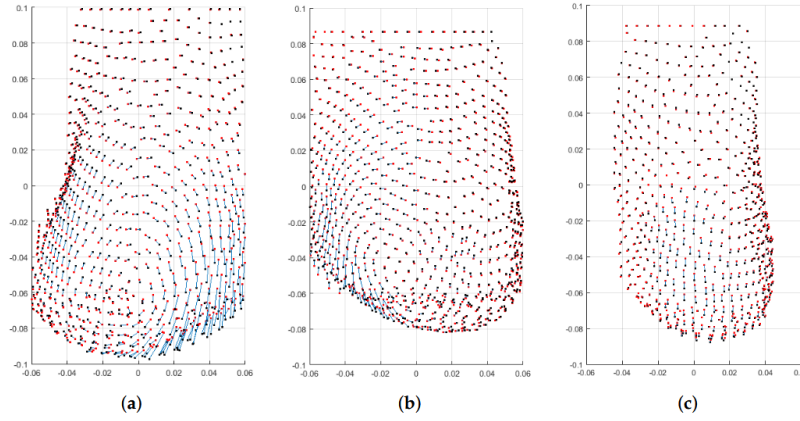


Figure 4.2: The post-surgery Pointcloud is shown with red points, while the predicted Pointcloud is visualised with black points. The displacements between corresponding points are coloured in blue. (a) Poor prediction, $D_{pred \rightarrow post} = 1.624$ mm; (b) Fair prediction, $D_{pred \rightarrow post} = 1.044$ mm; (c) Good prediction, $D_{pred \rightarrow post} = 0.827$ mm. [87]

4.3.2 Keypoint Reallocation and Affine Transformations

Freitas et al. [22] proposed a classic computer vision approach based on keypoint reallocation combined with affine transformations to recreate post-operative breast asymmetries on pre-operative images. The technique depends on the annotated key points responsible for defining the contour of the breast and the nipple position, together with a set of asymmetry metrics. This technique uses templates, that is, post-operative images of other patients with similar anatomical characteristics, to extract asymmetry metrics. Driven by the metrics, the key points on the target image are reallocated, followed by affine transformations to morph the pre-operative breast image to the desired post-operative asymmetry - Figure 4.3.

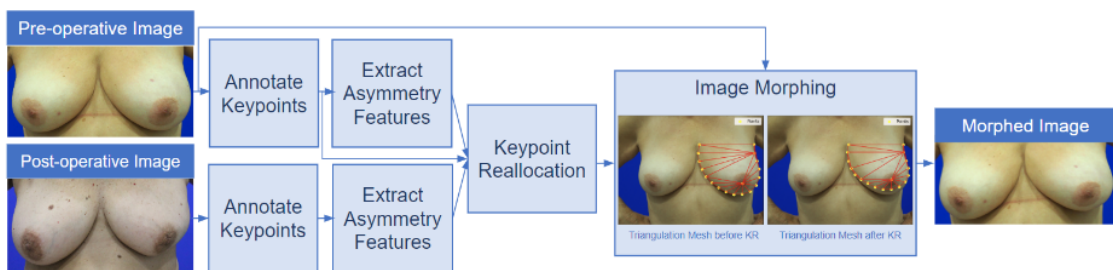


Figure 4.3: Workflow of Freitas et al.'s traditional method. This approach involves annotating key points in pre-and post-operative images, extracting asymmetry features, reallocating key points, and performing image morphing to produce a morphed post-operative image [22].

While this yielded a good result in modifying the breast's contour and general shape, it relied on manual keypoint annotation and struggled with realigning the nipple. Besides, while templates indicate asymmetry reproduction, the method does not address finer details related to texture or scarring, which are essential characteristics of realistic aesthetic evaluations. An example of the

resulting transformations using this approach, along with a deep learning approach, is shown in Figure 4.4.



Figure 4.4: Comparison of pre-operative images (left column) with results from the template-based methods: template images (second column), traditional computer vision (CV) approach (third column), and deep learning (DL) approach (fourth column). The DL method demonstrates superior nipple positioning accuracy and texture preservation, while the CV method excels in contour modification [52].

4.3.3 GAN-Based Image Transformation for Aesthetic Prediction

Carvalho et al. [12] investigated the application of Generative Adversarial Networks (GANs) for aesthetics prediction after breast cancer surgery. Pix2Pix and CycleGANs were adopted to learn the mappings from aesthetic categories labelled "Excellent" and "Poor" outcomes.

But the results were not ideal. Pix2Pix could not deal with artefacts in the presence of significant asymmetry, and CycleGANs generated unpredictable transformations when given diverse or noisy data. This study highlights GANs' aesthetic transformation potential while emphasising their reliance on high-quality data and computational resources.

4.3.3.1 Disentangled Representation Learning

To overcome the limitation of traditional techniques, Freitas et al. [22] also proposed a deep generative model-based disentangled approach. This method also uses templates, extracting asymmetry metrics from post-operative images of other patients to guide transformations. The approach disentangles medical features and asymmetry metrics from identity features. Therefore, the separation allows for manipulating asymmetries alone while keeping intact patient-specific attributes, such as texture or skin tone - Figure 4.5.

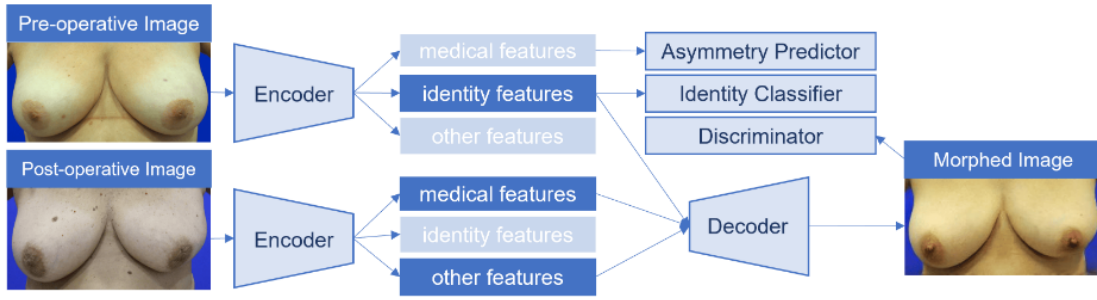


Figure 4.5: Workflow of Freitas et al.'s disentangled representation learning approach. This method separates medical, identity, and other features using an encoder-decoder architecture, enabling precise manipulation of asymmetry metrics while preserving patient-specific attributes like skin texture and tone [22].

This approach effectively reproduced nipple asymmetry and variations in the shape of the breast by exploiting templates while keeping visual outputs realistic. Nevertheless, templates with extreme asymmetries or poor quality still reveal challenges in artefact-free results. Moreover, this model is focused on asymmetry and does not capture other aesthetic aspects, such as scarring and skin discolouration. The results of this method, compared to the traditional computer vision approach, are also illustrated in Figure 4.4

Building on these foundations, Montenegro et al. [52] introduced improvements to the disentangled representation learning framework, addressing identity preservation and feature control limitations. Their model removes the identity feature vector used by Freitas et al. and instead exploits the natural symmetry of the breasts to separate clinical features, such as breast asymmetry and nipple position, from non-clinical ones, like skin tone and texture. This separation allows for more precise and independent manipulation of aesthetic features.

The updated framework - Figure 4.6 - uses a symmetry-based loss function to preserve patient-specific traits better while disentangling asymmetries into breast shape and nipple position. This refinement enables targeted transformations, improving realism and flexibility compared to Freitas et al.'s model.

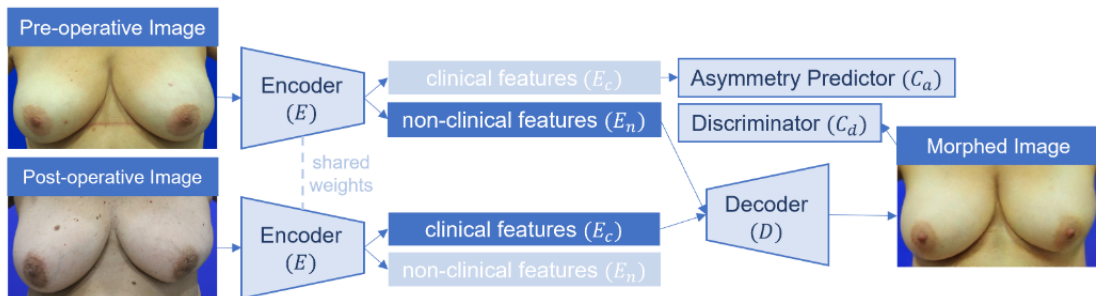


Figure 4.6: Montenegro et al.'s improved disentangled representation learning framework [52].

Montenegro et al.'s approach achieves higher accuracy in metrics like nipple retraction and better maintains the resemblance to the original pre-operative image. However, it still relies on

post-operative templates for feature extraction, which may limit its generalisability when template quality or diversity is insufficient. By overcoming some of the challenges in Freitas et al.'s method, this work further advances the potential for personalised surgical outcome predictions.

4.3.3.2 Inpainting Networks

Montenegro et al. [53] proposed a strategy based on inpainting. This approach utilises sophisticated models, such as U-Nets and GAN-enhanced networks, to reconstruct the masked regions of the breast. The process starts by overlaying a mask over the target breast in a pre-operative image, which is then annotated by pointing out the desired nipple position, followed by the lowest point of the breast contour. The provided inputs then drive the inpainting network to produce a complete breast image reflecting the changes. This workflow is shown in Figure 4.7

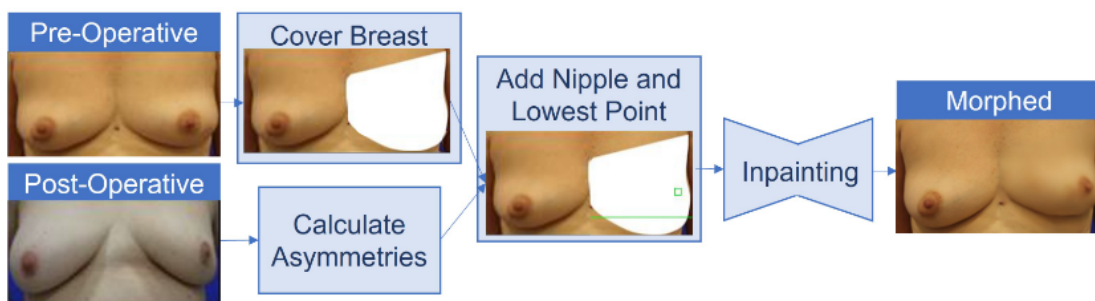


Figure 4.7: Workflow of Montenegro et al.'s inpainting approach. This method involves covering the breast, calculating asymmetry metrics, and using annotations for the nipple and lowest point to guide the inpainting process [53].

Other tested techniques include invertible architectures, such as Glow [41] and i-RevNet [34]. These models perform segmentation and inpainting in one step without requiring manual annotation of the contours and nipple positions on the breasts. Although such networks achieve high realism and automate the transformation process, significant deformations remain challenging to handle without introducing visual artefacts -Figure 4.8.

4.3.3.3 BreastGAN

Chartier et al. proposed BreastGAN [16], a Pix2Pix-based GAN trained on frontal pre- and post-operative image pairs to simulate breast augmentation outcomes.

BreastGAN validates the potential of supervised GANs in medical image simulation, despite being purely augmentation-focused and lacking conditioning (e.g., implant size). Notably, the model introduces visual inconsistencies by inadvertently changing non-breast regions like the arms, neck, and skin tone. Nevertheless, as Figure 4.9 illustrates, it produces a variety of realistic results.

While this work addresses a different problem than simulating the outcomes of breast cancer surgeries, it shares the high-level goal of breast aesthetic prediction. The results from BreastGAN

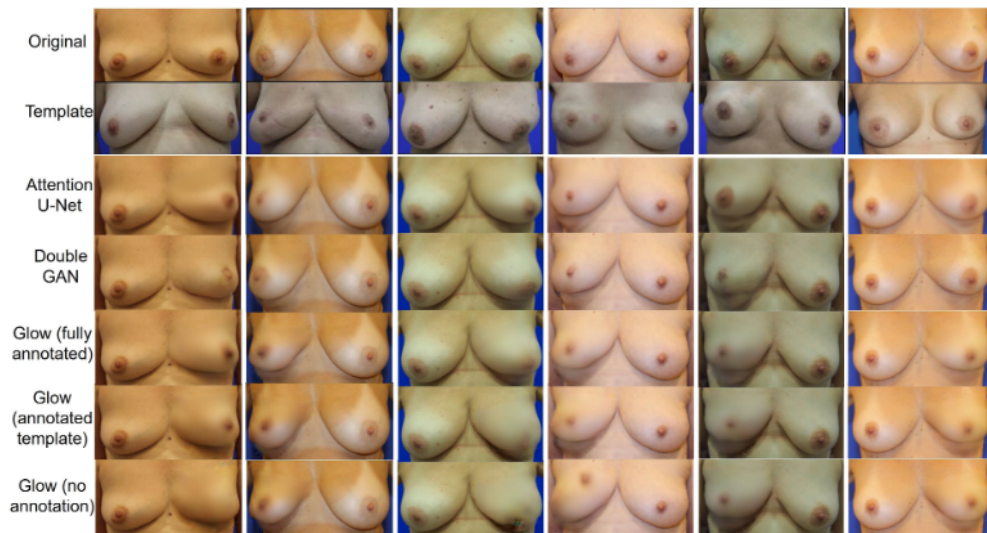


Figure 4.8: Results of the inpainting approach of Montenegro et al. by incorporating asymmetries from post-operative templates into the original pre-operative photos of the CP dataset [53]



Figure 4.9: Example results from BreastGAN [16], showing (left to right): preoperative image, GAN-generated image, and true post-operative image. While results are visually plausible, colour and body shape inconsistencies are noticeable.

reinforce that such generative models are a promising avenue for clinical simulation, supporting the feasibility of a more complex and condition-driven approach.

4.4 Image transformation in surgery

Besides the work that exists around the generation of post-operative images of breast cancer treatments, it is important to point out that other fields have also explored the generation of these images.

4.4.1 Conditional GANs for Orbital Decompression Surgery

Yoo et al. [83] developed a cGAN to predict post-operative facial appearance after orbital decompression surgery for thyroid-associated ophthalmopathy (TAO). The model generated post-operative images from pre-operative inputs by encoding the surgical condition into the latent vector.

The method demonstrated improved alignment between generated and actual post-operative images, as measured by the Structural Similarity Index (SSIM). However, the generated images (Figure 4.10) lacked sufficient realism, highlighting the ongoing challenge of producing high-fidelity outputs in medical imaging. Despite these limitations, the cGAN framework offers a fast and automated alternative for surgical planning, enhancing patients' ability to visually anticipate potential outcomes.

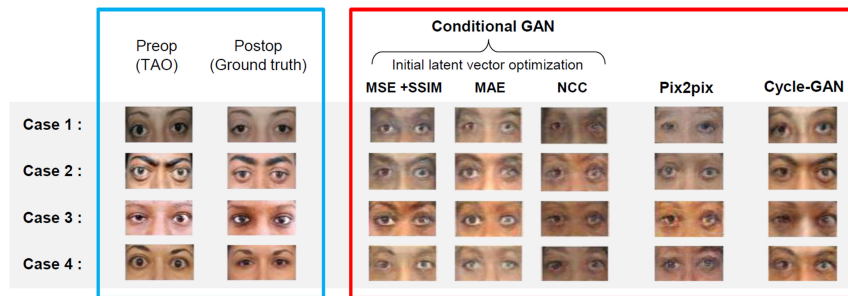


Figure 4.10: Comparison of pre-operative (TAO) and post-operative (ground truth) facial appearances with results generated by different conditional GAN (cGAN) methods, including initial latent vector optimisation using MSE + SSIM, MAE, NCC, and models like Pix2Pix and CycleGAN [83].

4.4.2 Pix2Pix GANs for Eye Surgery

Kim and Chin [37] used Pix2Pix, a GAN-based image-to-image translation model, for the synthesis of post-operative optical coherence tomography (OCT) images after epiretinal membrane (ERM) removal. The network generated predictive post-operative OCT images with high structural similarity from pre-operative inputs. Figure 4.11 illustrates a comparison between pre-operative, actual post-operative, and synthesized post-operative OCT images, demonstrating the model's effectiveness in visual prediction.

4.4.3 3D-Guided Face Manipulation Using CycleGANs

Andlauer et al. [3] presented a CycleGAN-based method that predicted the post-operative facial appearance after craniomaxillofacial surgery. That method fused 2D pre-operative photos with 3D simulations of post-operative soft tissue deformations. The model used semi-supervised learning and 3D simulations to change parts of the face, like the chin and nose (Figure 4.12).

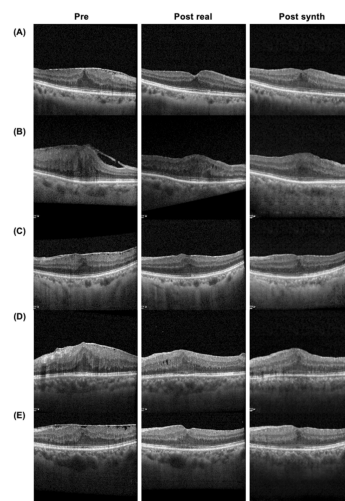


Figure 4.11: Comparison of pre-operative, actual post-operative, and synthesised post-operative OCT images generated using Pix2Pix GAN for epiretinal membrane (ERM) removal. [37].



Figure 4.12: Example outputs from a CycleGAN-based method by Andlauer et al. [3], showcasing 3D-guided facial manipulations for predicting postoperative outcomes. The method fuses 2D pre-operative images with 3D simulations of post-operative soft-tissue deformations, enabling changes to specific facial features, such as the chin (a) and nose (b).

4.4.4 PlasticGAN for Facial Surgery Simulation

Chandaliya et al. [15] proposed PlasticGAN, a Conditional GAN framework (Figure 4.13) for simulating post-operative outcomes in facial plastic surgery. PlasticGAN integrates adversarial, perceptual, and reconstruction losses to generate photorealistic images of surgical results, such as face ageing, rejuvenation, and reconstruction.

4.5 Datasets for Aesthetic Evaluation in Breast Cancer Treatment

Developing and evaluating generative models for predicting post-operative outcomes relies heavily on access to high-quality datasets. Several datasets are available to support aesthetic evaluation in breast cancer treatments, each with unique characteristics and contributions to the field:

- The **CINDERELLA** dataset, developed as part of the CINDERELLA project, is a private dataset comprising pre- and post-surgical images. It includes over 2,000 annotated post-surgical images specifically for aesthetic evaluation [35]. As this work contains pre-/post-

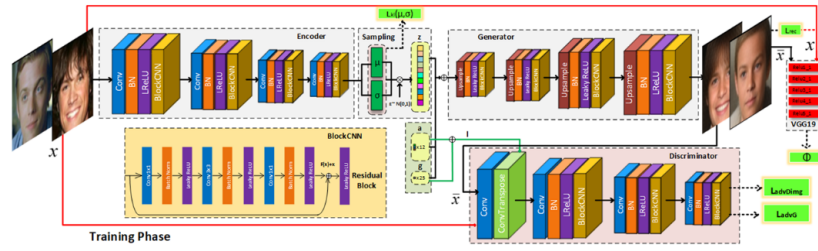


Figure 4.13: Architecture of PlasticGAN, a Conditional GAN framework proposed by Chandaliya et al. [15] for simulating post-operative outcomes in facial plastic surgery. The model employs a combination of adversarial, perceptual, and reconstruction losses during training to generate realistic post-operative images, targeting tasks such as face ageing, rejuvenation, and reconstruction.

surgical pairs, it can be applied to post-operative outcome prediction..

- The **Breast Cosmesis** dataset, a public dataset introduced by Guo et al. [27], includes over 3,000 annotated post-surgical images. However, this dataset does not include pre-operative images, making it unsuitable for tasks involving outcome prediction from pre-surgical data.
- The **PORTO** dataset is a smaller dataset consisting of 120 annotated post-surgical images for aesthetic evaluation [11]. Like the Breast Cosmesis dataset, it lacks pre-operative images and is therefore not applicable to models focused on pre-to-post outcome prediction.

4.6 Discussion

Various methodologies have been employed to generate post-operative images in the treatment of breast cancer. Nevertheless, generative models exhibit considerable promise. While traditional techniques, such as regression-based schemes and template-driven strategies, provide fundamental insights, they frequently fall short regarding realism and personalisation.

Generative models, such as GANs, disentangled representation learning, and inpainting, are promising solutions. They simulate realistic outcomes based on pre-operative images and treatment data. These models can overcome the significant limitations by incorporating clinical features such as asymmetry and scarring while preserving individual attributes such as skin tone and texture.

Generative models can potentially create personalised outcomes through automated predictions and enhanced realism. This capability will enable patients to comprehend the results more effectively, supporting shared decision-making. By incorporating additional information, such as surgery type and radiotherapy application, these models can take into account factors like scarring, skin discolouration, and texture changes, which are vital elements for achieving realistic visualisations. Such advancements would contribute to a deeper understanding of patient outcomes and enable adjustments to expectations as necessary, fostering shared decision-making and enhancing patient satisfaction.

Chapter 5

Preliminary Analysis and Data Preparation

Breast cancer surgery leads to physical changes that can impact a patient’s body image and confidence. As a result, the ability to visualise post-operative outcomes before surgery has become increasingly important in supporting shared decision-making. However, accurately predicting how a specific patient’s body will look after surgery is a highly challenging task as it requires models that can preserve the unique characteristics of each patient while realistically simulating the surgical effects.

This thesis addresses this challenge by proposing a generative modelling framework that predicts post-operative breast appearances from pre-operative images and associated clinical metadata. Unlike prior approaches, which often rely on unsupervised learning or retrieval of example cases from other patients, this work focuses on patient-specific prediction using paired before-and-after surgery data. The aim is to go beyond generic outcomes and offer personalised, realistic visualisations.

To achieve this, the methodology begins by adapting and evaluating existing image-to-image translation architectures, including Pix2Pix, BicycleGAN, and StarGAN. These models are extended to incorporate structured conditioning based on treatment details (e.g., surgery type and cancer side) and guided using a set of carefully designed loss functions that supervise structural, anatomical, and visual realism.

The expected outcome is a system capable of generating post-operative predictions that (1) are realistic, (2) remain faithful to the patient shown in the pre-operative image, preserving skin tone, body shape, lighting, and background, and (3) contain traits associated with the surgical procedure, such as scars and asymmetries.

This chapter presents the dataset and its clinical characteristics, outlines the preprocessing steps used to standardise input images, and defines the evaluation strategy that will be followed throughout the experiments.

5.1 Data Analysis

To better understand the characteristics of the dataset, several demographic, clinical, and imaging-related variables were examined. Among the findings, it shows that the distribution of breast cancer cases between the left and right sides is relatively balanced, with only a small fraction of bilateral cases, as shown in Figure 5.1 (a). In terms of treatment, most patients received radiotherapy, as illustrated in Figure 5.1 (b), reflecting standard clinical practices. Moreover, the majority of patients are aged between 41 and 70 years (Figure 5.2), aligning with the typical demographic profile for breast cancer incidence.

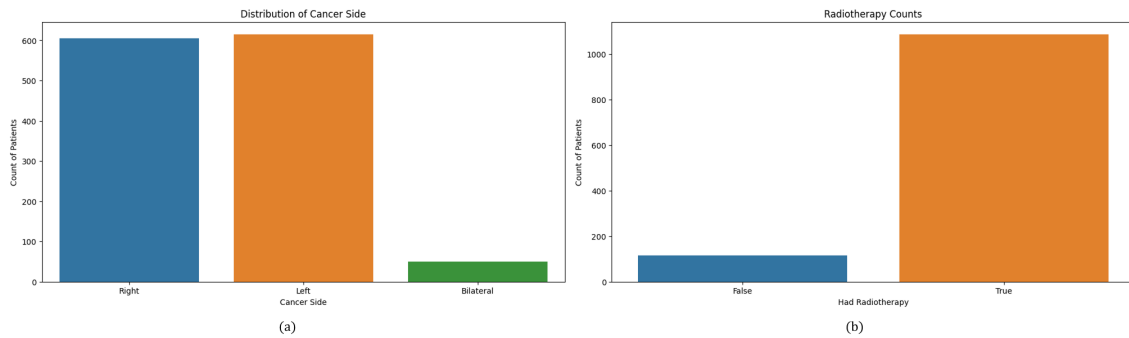


Figure 5.1: (a) Distribution of cancer side (left, right, bilateral). (b) Distribution of patients who received radiotherapy.

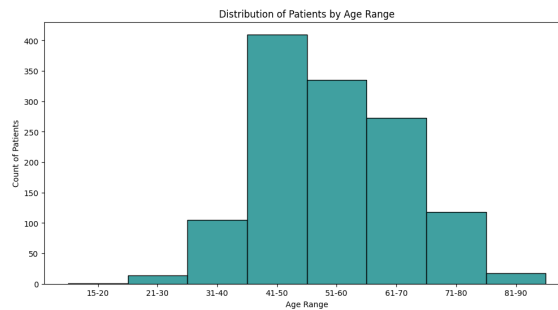


Figure 5.2: Distribution of patients by age range at the time of surgery.

Regarding image availability, most patients have only one pre-operative image, although some have more than one. In such cases, the image closest to the surgery date was selected for analysis (Figure 5.3 (a)). In contrast, post-operative images varied more widely in number. While some patients had none, others had up to three post-operative images (Figure 5.3 (b)).

The time interval between surgery and the acquisition of post-operative images varies significantly across the dataset (see Figure 5.4). This variability complicates fair comparisons of post-operative results, as there is no standardised time point for capturing outcomes. One possible solution would be to limit the dataset to images taken within a narrow window (for example, 1–2 years); however, this would significantly reduce the sample size. Instead, a reference time interval was computed based on the most common delay observed among patients with only one

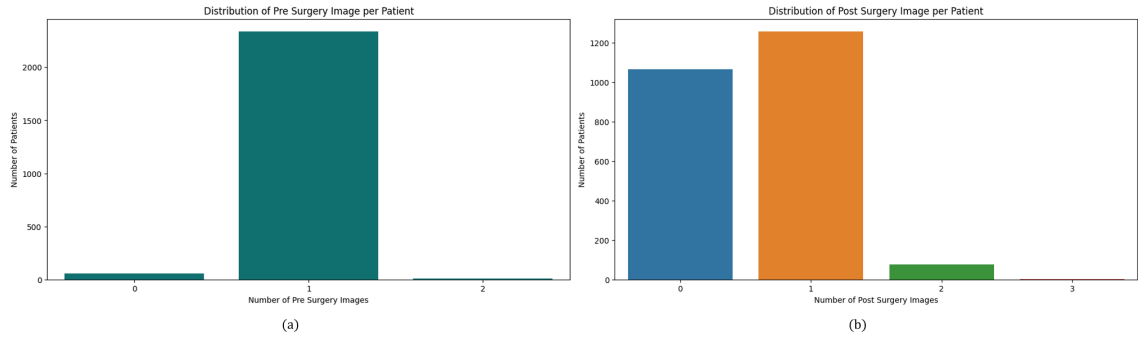


Figure 5.3: (a) Number of pre-operative images per patient. (b) Number of post-operative images per patient.

post-operative image. For patients with multiple post-operative images, the image closest to this reference interval was selected. This approach allowed for the standardisation of image timing while retaining as much data as possible for training.

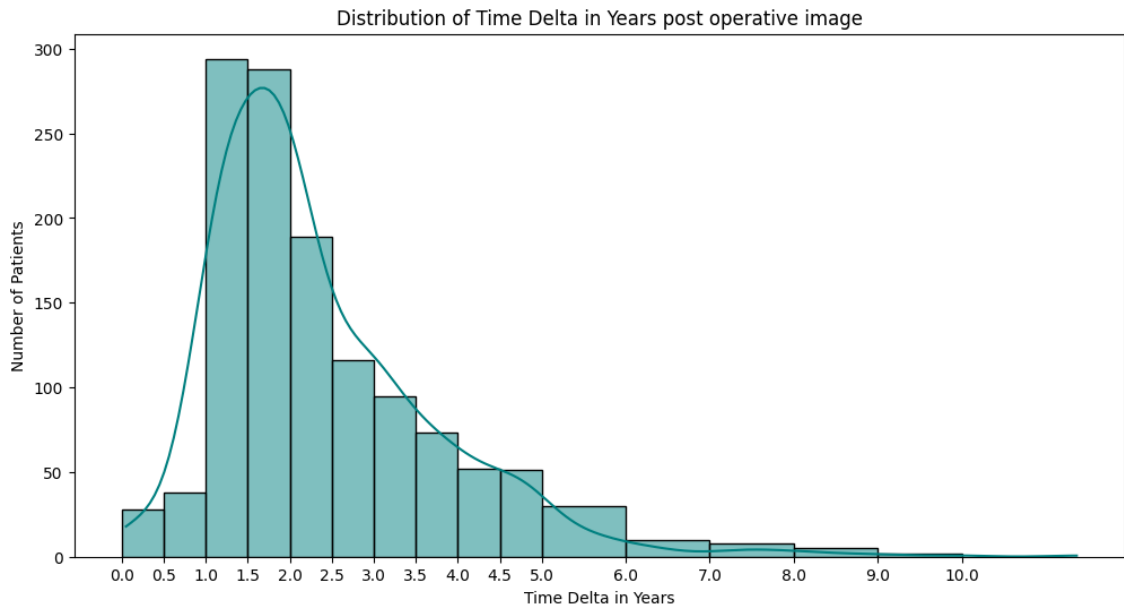


Figure 5.4: Distribution of time (in years) between surgery and post-operative image acquisition.

Surgical procedures in the dataset range from conservative treatments to complex mastectomies involving reconstruction and contralateral symmetrisation. Table 5.1 provides a summary of the different surgery types and their frequencies. The most common procedure was conservative surgery (C1), followed by conservative surgery with bilateral reduction (C2). It is noteworthy that in some cases, the clinical records indicate cancer on only one side, yet the surgery affected both breasts (e.g. C2 procedure). This situation can introduce ambiguity when aligning clinical intent with image features, particularly for models designed to learn asymmetry or side-specific variations.

Code	Description	Count
C1	Conservative Surgery	679
C2	Conservative Surgery with bilateral reduction	286
M5	Mastectomy with unilateral reconstruction with implant and contralateral symmetrisation with implant (augmentation)	58
M3	Mastectomy with bilateral reconstruction with implants	51
C3	Conservative Surgery with LD or LICAC/TDAP flaps	46
M1	Mastectomy with unilateral reconstruction with implant	40
M6	Mastectomy with unilateral reconstruction with implant and contralateral symmetrisation with reduction	31
C4	Conservative Surgery with bilateral implants	27
M2	Mastectomy with unilateral reconstruction with autologous flap	25
M7	Mastectomy with unilateral reconstruction with autologous flap and contralateral symmetrisation with reduction	22
M4	Mastectomy with bilateral reconstruction with autologous flaps	4

Table 5.1: Types of surgery and their frequencies in the dataset.

Finally, the dataset exhibits significant inconsistency in image resolution, as shown in Figure 5.5, but retains a similar aspect ratio. This observation further justifies the need for standardisation in preprocessing.

A total of 1269 image pairs were collected from the dataset, each consisting of a pre-operative and a post-operative image; however, for early experiments, only the conservative surgeries were taken into account, which reduced the number to 1,038 pairs. These were divided into training (85%), validation (5%), and test (10%) sets. To ensure diversity and clinical representativeness across splits, patients were grouped based on surgery type and cancer side, and the split was performed using stratified sampling at the patient level. This ensured that all subsets contained a balanced distribution of conservative surgeries, as well as left/right/bilateral cancer sides.

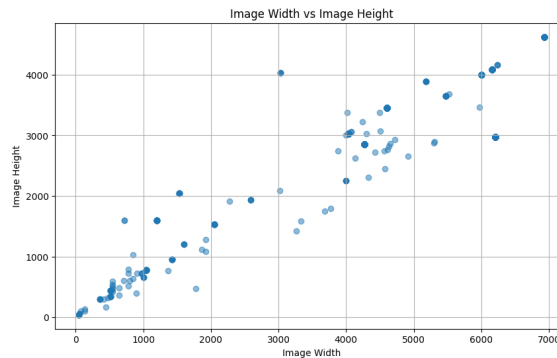


Figure 5.5: Scatter plot of image width vs. height.

5.2 Image Preprocessing

A preprocessing pipeline was applied to standardise inputs to prepare the dataset for training. Differences in image resolution, patient pose, and body shape can introduce noise that affects model performance. For example, pre- and post-operative images may differ in posture or framing, as shown in Figure 5.6. These inconsistencies make it harder for the model to focus on surgical changes alone.

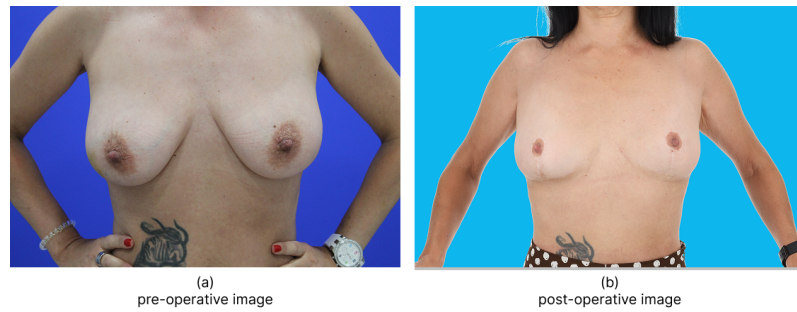


Figure 5.6: Examples of patient variation. (a) Pre-operative image. (b) Post-operative image. Differences in body shape, posture, and arm position illustrate the challenge of learning consistent transformations.

5.2.1 Bounding Box Cropping with Keypoints

To reduce irrelevant variation due to differing zoom levels, background content, and patient pose, each image was cropped using anatomical keypoints annotated in the dataset. These keypoints correspond to both breast contours.

The cropping procedure was as follows:

1. Identify the full breast region using all keypoints and calculate the bounding box.
2. Apply a random zoom-out factor between 1.1 and 1.4 for augmentation. This variability was crucial, as using a fixed zoom led to rapid overfitting during training.
3. Crop the image using the expanded bounding box and centre the breast region within the new frame.

This preprocessing step is illustrated in Figure 5.7, which shows the same pair of pre- and post-operative images after applying different zoom levels to the keypoint-based bounding box. This approach improves spatial alignment across samples and reduces the risk of overfitting to background artefacts or inconsistent framing.

5.2.2 Flipping, Scaling and Normalisation

To improve data variability, random horizontal flipping was applied as an augmentation technique. The images were resized to a width of 256 pixels while maintaining their aspect ratio, which

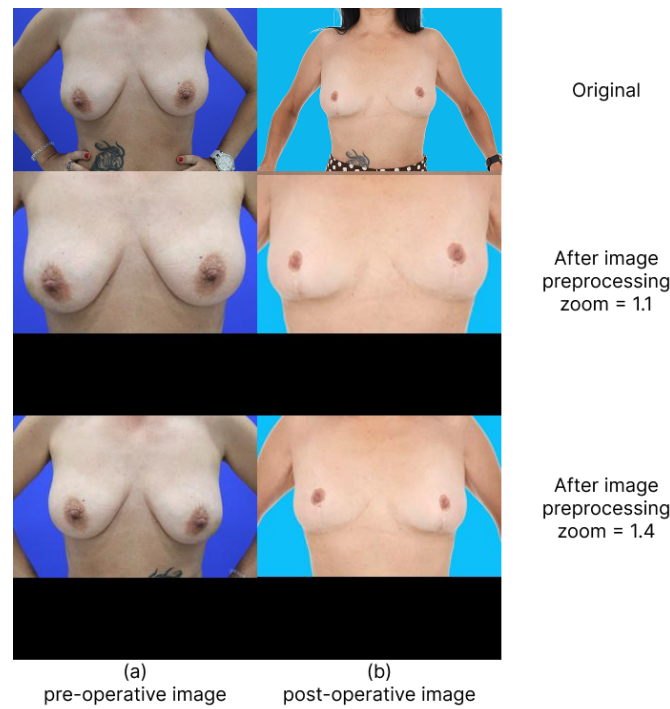


Figure 5.7: Effect of bounding box cropping with keypoints and zoom-out factors. Top row: original pre- and post-operative images. Middle and bottom rows: zoomed crops with factors of 1.1 and 1.4, respectively.

ensured compatibility with batch processing and model input requirements. Finally, pixel values were normalised to the range of $[-1, 1]$ to ensure consistent input scaling during training.

5.3 Evaluation Metrics

To assess and compare the performance of different generative models, a set of tailored evaluation metrics was defined, each targeting a specific aspect of image quality and clinical relevance. These include metrics for global consistency, local fidelity, and anatomical accuracy, as illustrated in Figure 5.8.

- **SSIM with Pre-operative Image:** This metric calculates the Structural Similarity Index Measure (SSIM) between the generated image and the original pre-operative image. It is used to assess how well the model preserves global features such as body pose, lighting, and skin tone. Ideally, SSIM should be moderately high, indicating consistency with the patient's original appearance, while still allowing enough deviation to reflect the surgical changes. Very high SSIM values may suggest the model is simply copying the input without learning meaningful transformations. For reference, the average SSIM between their pre- and post-operative images of the same patient is 0.66.
- **SSIM with Post-operative Image (Breast Region Only):** This metric calculates SSIM between generated and post-operative images, focusing on the breast region and evaluated

in grayscale. While grayscale reduces the influence of colour, it does not eliminate it entirely (skin tone can still affect the score). The goal of this metric is to assess the model's ability to replicate surgical changes such as shape, scarring, and breast structure. Ideally, SSIM should be high enough to indicate structural similarity, but not so high that it suggests the model is copying post-operative appearance too closely, including undesired tone shifts.

- **IoU with Breast Mask:** Measures the Intersection over Union between the predicted and ground truth breast segmentation masks. This captures alignment and consistency in breast shape.
- **IoU with Nipple Mask:** Similar to the above, but focused specifically on nipple segmentation. It is a more sensitive and localised anatomical metric, especially valuable for aesthetic realism.
- **Colour 1 (Patch-based MAE):** This metric computes the mean absolute error (MAE) between the average colour values across spatial patches of the generated and pre-operative images. By enforcing colour similarity at the patch level, it discourages desaturation, tinting, or shifts in illumination, ensuring the body's overall appearance remains true to the original.
- **Colour 2 (Outside-mask MAE):** This metric computes the MAE between the generated image and the pre-operative image, restricted to the region *outside* the breast area. It ensures that the non-surgical parts of the image, such as the arms, torso, and background, remain consistent in both shape and colour. This is especially important for preserving patient identity and preventing the generator from generating unrealistic lighting or body deformation in unaffected regions.

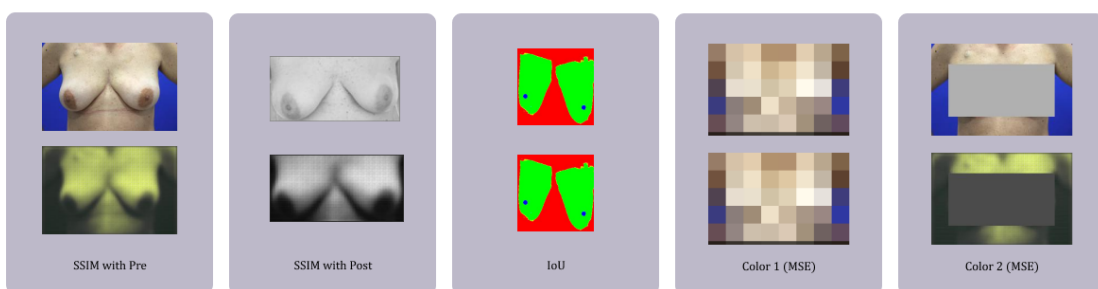


Figure 5.8: Evaluation metrics used to compare model performance. From left to right: SSIM with the pre-operative image, SSIM with the post-operative image (breast region only), IoU between segmentations (breast and nipple), and colour preservation metrics (outside-mask and patch-based L1, both computed against the pre-operative image).

Together, these metrics provide a multi-faceted evaluation framework, allowing models to be assessed not only on visual fidelity but also on their ability to preserve patient-specific features and replicate realistic surgical outcomes.

Chapter 6

Baseline Experiments

The first experiment focuses on training and testing three advanced generative models for image-to-image translation: Pix2Pix, BicycleGAN and Stargan. The goal was to evaluate how effectively these standard architectures could learn meaningful post-operative changes by training each model with paired pre- and post-operative images.

Additionally, this chapter also presents the addition of conditioning information to investigate whether surgical metadata could guide the generation process toward more accurate and diverse outcomes.

6.1 Pix2Pix

6.1.1 Architecture and Training Procedure

Pix2Pix was developed using an existing implementation¹, following the original structure introduced by Isola et al. [33]. This adopts a conditional GAN (cGAN) framework, where the generator G is trained to map an input image A (in our case, the pre-operative image) to an output image B (the post-operative result). The discriminator D learns to classify whether the pair (A, B) is real or if the pair $(A, G(A))$ is fake.

The generator architecture uses a U-Net structure with skip connections, which allows low-level information to be passed directly from the encoding to the decoding layers, which helps preserve structural details like body pose. The discriminator follows a PatchGAN design, where the image is evaluated in overlapping patches rather than as a whole.

Figure 6.1 shows the forward process for Pix2Pix. The generator receives the pre-operative image A and generates a synthetic post-operative image $\hat{B} = G(A)$. This output is concatenated with the original input and passed to the discriminator D for realism evaluation. The training process is guided by two main objectives:

¹<https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix>

- **Adversarial Loss ($\mathcal{L}_{\text{GAN}}$):** Encourages the generator to produce realistic images that can fool the discriminator. It is defined as:

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{A, B}[\log D(A, B)] + \mathbb{E}_A[\log(1 - D(A, G(A)))] \quad (6.1)$$

where A is the pre-operative image, B is the real post-operative image, and $G(A)$ is the generated image.

- **L1 Loss ($\mathcal{L}_{\text{L1}}$):** Promotes pixel-level similarity between the generated image and the real target. It is defined as:

$$\mathcal{L}_{\text{L1}}(G) = \mathbb{E}_{A, B}[\|B - G(A)\|_1] \quad (6.2)$$

where $\|\cdot\|_1$ denotes the pixel-wise L1 norm between the real and generated post-operative images.

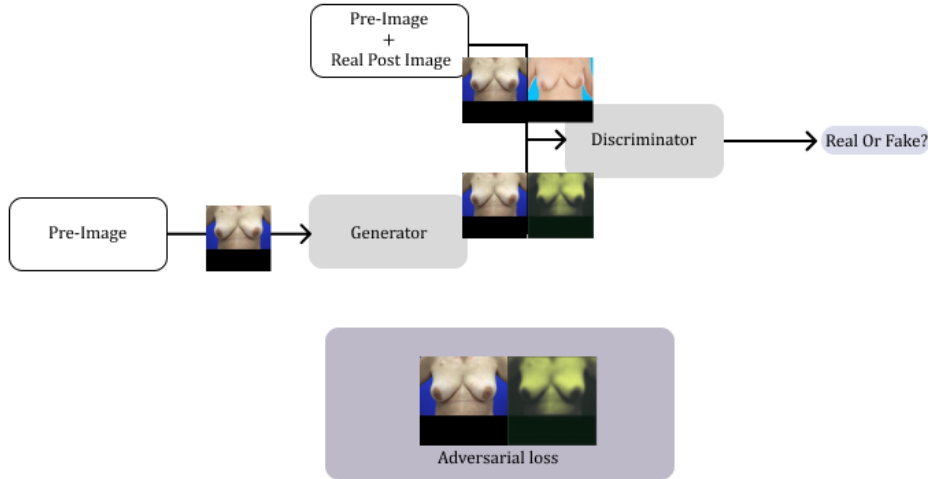


Figure 6.1: Pix2Pix architecture: the generator learns a mapping from pre-operative to post-operative images. The discriminator distinguishes real vs. fake pairs.

The full generator objective is

$$\mathcal{L}_G = \mathcal{L}_{\text{GAN}}(G, D) + \lambda_{\text{L1}} \cdot \mathcal{L}_{\text{L1}}(G(A), B) \quad (6.3)$$

where λ_{L1} is typically set to 100, as in the original paper, to balance realism and fidelity.

The model was trained using the Adam optimiser [40], an adaptive optimisation algorithm that combines momentum and per-parameter learning rate scaling based on first and second moment estimates of the gradients. A learning rate of $lr = 0.0002$ and $\beta_1 = 0.5$ (the exponential decay rate for the first moment estimate) were used, following common practice in GAN training. The training procedure alternates between optimising the generator and the discriminator at each iteration:

- The generator is optimised to minimise the combined GAN and L1 losses.
- The discriminator is optimised using real and fake image pairs to maximise its classification performance.

During the second half of training, a linear learning rate decay was used, which made the model converge more smoothly. For instance, during a 400-epoch training run, the learning rate stayed the same for the first 200 epochs and then slowly dropped to zero over the next 200 epochs. This simple strategy proved effective in producing more stable results.

Xavier initialisation [23], which is intended to preserve a constant variance of activations across layers, was used to initialise the network weights. This approach is recommended for deep convolutional networks using ReLU activations, as it helps prevent problems such as vanishing or exploding gradients during training.

In order to stabilise and speed up training, the input images were resized to 256×256 and normalised using batch normalisation [32], which standardises activations within each mini-batch. In certain model variations, dropout was also used to enhance generalisation and lessen overfitting by randomly deactivating a portion of neurons during training. Strict spatial correspondence between pre- and post-operative images was ensured by training all models with paired image data.

6.1.2 Results

The initial results showed that, while the model could learn general image-to-image transformations, the lack of clinical information, such as surgery type, often led to generated images that did not correspond to the surgical outcome. For example, the model might produce a conservative surgery outcome when the expected outcome corresponds to breast reduction.

Moreover, the model struggled to preserve important visual characteristics such as skin tone and body pose. As seen in Figure 6.2, the generated outputs often differed from the pre-operative images in lighting and colouration. This effect likely results from inconsistencies in imaging conditions between pre- and post-operative datasets, including differences in equipment, angles, and lighting.

6.2 BicycleGAN

BicycleGAN² [85] was adopted to overcome a key limitation of deterministic models like Pix2Pix, their inability to represent multiple plausible post-operative outcomes for the same pre-operative image. BicycleGAN extends the conditional GAN framework by incorporating a low-dimensional latent vector z to model this ambiguity and enable diversity in output generation.

²<https://github.com/junyanz/BicycleGAN/tree/master>

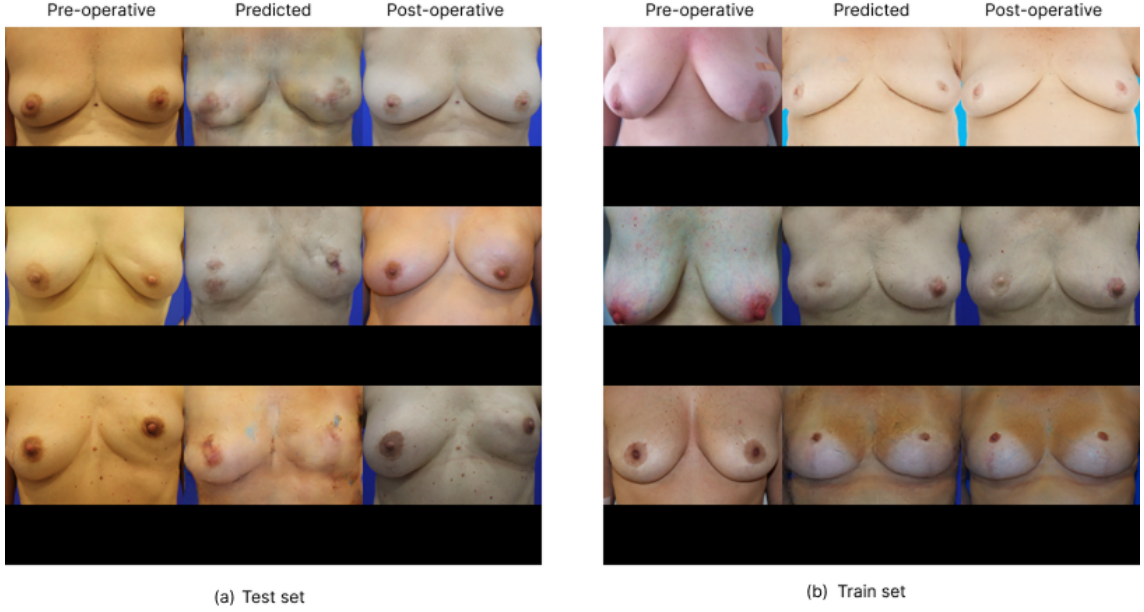


Figure 6.2: Visual comparison of generated and real post-operative images in (a) the testing set and (b) the training set.

6.2.1 Architecture and Training Procedure

BicycleGAN extends the conditional GAN framework by incorporating a latent variable to model the one-to-many nature of the post-operative outcome prediction task. Instead of generating a single deterministic output, the model learns a distribution of plausible outputs for a given pre-operative image. This is achieved by combining two training strategies: a conditional Variational Autoencoder GAN (cVAE-GAN) and a conditional Latent Regressor GAN (cLR-GAN).

The model consists of three main components: a generator G , an encoder E , and two discriminators D and $D2$. The generator is a U-Net, with skip connections, that receives a pre-operative image A and a latent vector z as input and outputs a predicted post-operative image $\hat{B} = G(A, z)$. The encoder infers a latent code from real post-operative images, producing a mean μ and log-variance $\log \sigma^2$, from which z is sampled. The two discriminators evaluate the realism of generated outputs under two scenarios: when using encoded latent vectors (cVAE branch) and when using randomly sampled ones (cLR branch). The training alternates between these two modes (Figure 6.3).

In the cVAE-GAN path, the ground truth post-operative image B is encoded into a latent vector $z = E(B)$, and the generator attempts to reconstruct B from A and z . This path is guided by three loss terms: an adversarial loss that encourages realism, an L1 reconstruction loss between the output and B , and a KL divergence loss that regularises the encoder's output to match a standard normal distribution. In the cLR-GAN path, a random latent vector $z \sim \mathcal{N}(0, I)$ is sampled, used to generate $\hat{B} = G(A, z)$, and then passed through the encoder to recover an estimate of z . A latent regression loss ensures that the encoded output matches the original sampled vector.

During training, the model receives a batch of paired images (A, B) , split in half to alternate

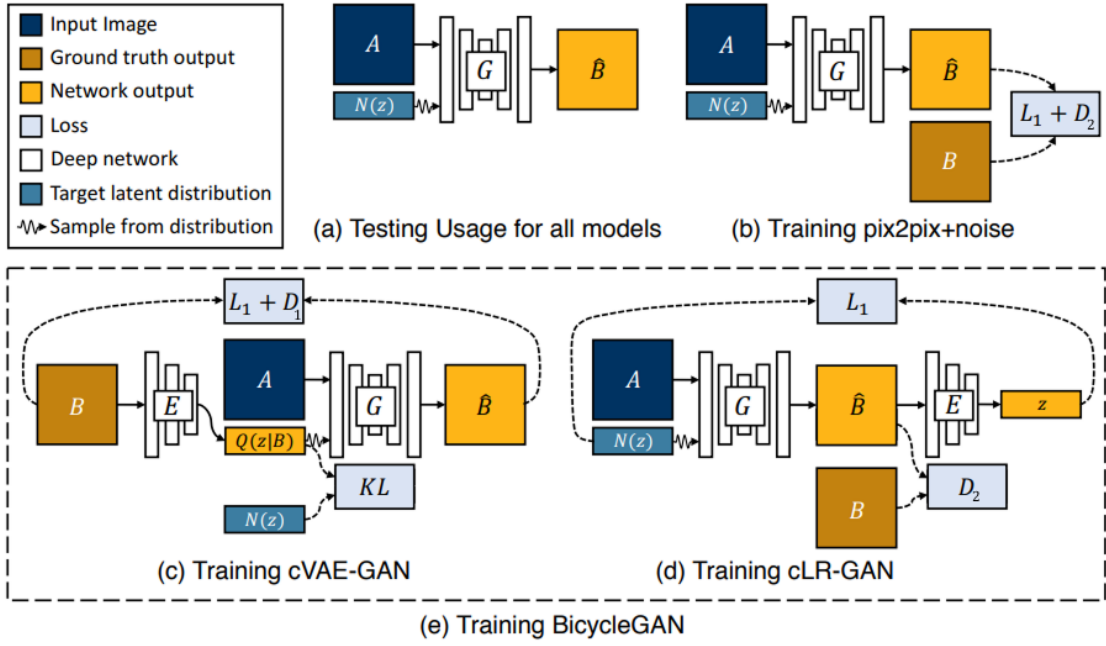


Figure 6.3: Architecture and training modes of BicycleGAN. Training alternates between the cVAE-GAN branch (left) and cLR-GAN branch (right), enabling both latent code recovery and output diversity [85].

between the two branches. The encoder, generator, and both discriminators are optimised using Adam. The total loss function used to train the model includes the following components:

- **Adversarial Loss** ($\mathcal{L}_{\text{GAN}}$): Encourages the generator to produce images that are indistinguishable from real post-operative images. For the discriminator D , it is defined as:

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{A, B} [\log D(A, B)] + \mathbb{E}_{A, z} [\log (1 - D(A, G(A, z)))] \quad (6.4)$$

- **Reconstruction Loss** ($\mathcal{L}_{L1}$): Ensures that when the generator is provided with a latent code encoded from the real image, the output closely matches the ground truth.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{A, B, z} [\|B - G(A, z)\|_1] \quad (6.5)$$

- **KL Divergence** ($\mathcal{L}_{\text{KL}}$): Regularises the distribution of the latent codes encoded from real images to approximate a standard normal distribution.

$$\mathcal{L}_{\text{KL}}(E) = \mathbb{E}_B \left[\frac{1}{2} \sum_{j=1}^d (\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1) \right] \quad (6.6)$$

- **Latent Regression Loss** ($\mathcal{L}_{\text{latent}}$): Encourages the encoder to recover the originally sampled

latent vector from the generated output.

$$\mathcal{L}_{\text{latent}}(G, E) = \mathbb{E}_{A, z} [\|z - E(G(A, z))\|_1] \quad (6.7)$$

The complete objective function to minimise is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN}} + \lambda_{L1} \cdot \mathcal{L}_{L1} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}} + \lambda_{\text{latent}} \cdot \mathcal{L}_{\text{latent}} \quad (6.8)$$

where λ_{GAN} , λ_{L1} , λ_{KL} , and λ_{latent} are scalar weights that control the relative importance of each term during training.

This hybrid training strategy encourages the generator to produce outputs that are both realistic and diverse, while ensuring the latent space remains meaningful and controllable. A summary of the training flows is shown in Figure 6.3, illustrating how latent codes are used both to reconstruct and to regularise the output distribution.

6.2.2 Results

BicycleGAN was used to generate diverse post-operative outcomes from a single pre-operative image, leveraging its latent space sampling to produce multiple plausible outputs. These variations mostly reflected subtle shifts in colour and slight changes in breast shape. However, like Pix2Pix, the model also suffered from key limitations. As shown in Figure 6.4, many predictions exhibited unnatural colour deviations, failing to preserve the original skin tone and body shape.

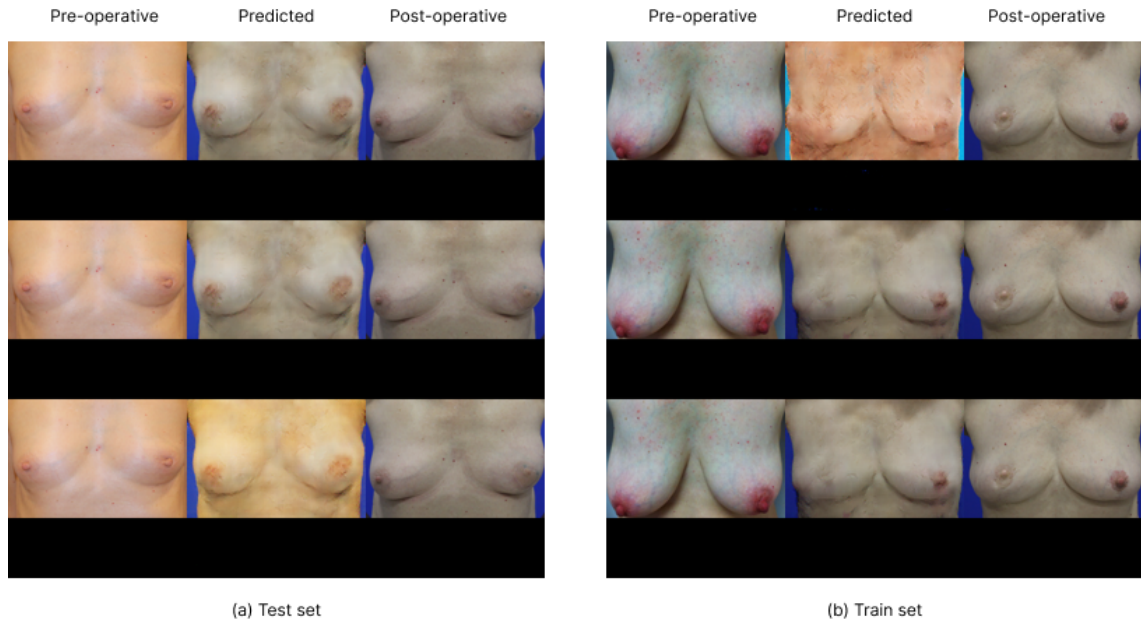


Figure 6.4: Visual results from BicycleGAN. Each row shows a pre-operative image, a generated prediction, and the ground truth post-operative image. (a): Test set, (b): Training set. The model struggles with skin tone preservation and generates only minor shape variations.

6.3 StarGAN

6.3.1 Architecture and Training Procedure

StarGAN³ [18] was also evaluated as a baseline model due to its ability to handle multi-domain image-to-image translation using a single generator–discriminator pair. In its original design, StarGAN was trained to translate images from one domain to another and then back again using a cycle-consistency loss. However, since this project aims to explore a supervised setup with paired pre- and post-operative images, the architecture was adapted accordingly.

To better align with the available paired data, we added a Structural Similarity Index (SSIM) supervision loss between the generated image and the true post-operative target. This loss complements the original StarGAN training objective by directly encouraging structural fidelity in the output. In our modified setup, the model still performs the domain translation–reconstruction cycle (original domain $\rightarrow$ target domain $\rightarrow$ original), but now it also learns from the ground truth post-operative image through direct supervision. The training process is illustrated in Figure 6.5, which shows both the original-to-target and target-to-original translation flows, as well as the domain classification components.

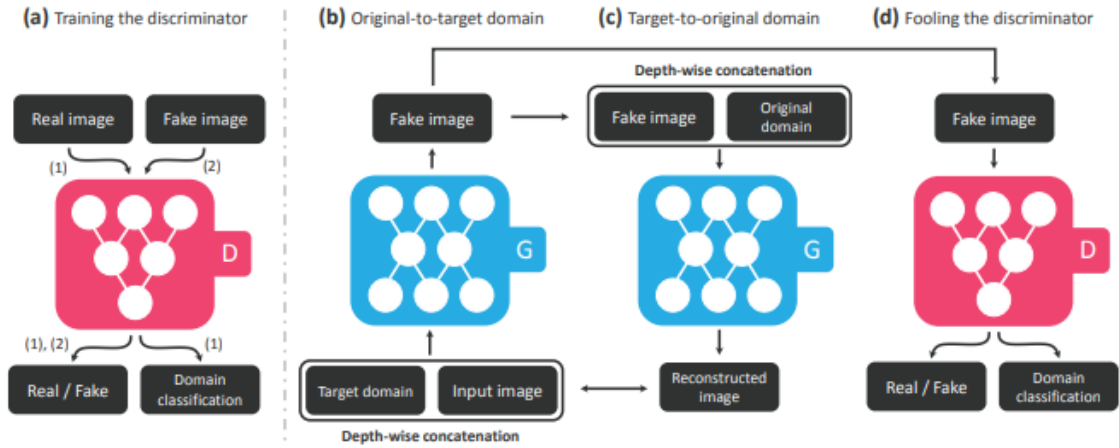


Figure 6.5: Adapted StarGAN training process. The original architecture was modified by adding an L1 supervision loss to support paired image translation. The generator performs domain-to-domain transformations, guided by both adversarial and cycle-consistency losses [18].

The generator G receives the input image x and a target domain label c , such as surgery type or cancer side. The label c is spatially replicated and concatenated channel-wise x to form the input $[x, c]$. The generator outputs a synthetic image $\hat{x} = G(x, c)$ intended to reflect the desired surgical transformation. This generator follows an encoder–residual–decoder structure, and instance normalisation is used throughout, except at the output.

The discriminator D takes an image and outputs two predictions: (1) D_{src} , which determines whether the image is real or fake, and (2) D_{cls} , which predicts the domain label of the image.

³<https://github.com/yunjey/stargan/tree/master>

The discriminator uses a PatchGAN design with several convolutional layers and Leaky ReLU activations.

Training uses a combination of adversarial, classification, reconstruction, and SSIM losses:

- **Adversarial Loss** ($\mathcal{L}_{adv}$): Encourages the generator to produce images that the discriminator cannot distinguish from real ones.

$$\mathcal{L}_{adv} = \mathbb{E}_x[\log D_{src}(x)] + \mathbb{E}_{x,c}[\log(1 - D_{src}(G(x, c)))] \quad (6.9)$$

- **Domain Classification Loss** ($\mathcal{L}_{cls}$): Ensures correct domain prediction for real and generated images.

$$\mathcal{L}_{cls}^{real} = \mathbb{E}_{x,c'}[-\log D_{cls}(c'|x)] \quad (6.10)$$

$$\mathcal{L}_{cls}^{fake} = \mathbb{E}_{x,c}[-\log D_{cls}(c|G(x, c))] \quad (6.11)$$

- **Reconstruction Loss** ($\mathcal{L}_{rec}$): Enforces cycle-consistency by reconstructing the original image using the inverse domain label.

$$\mathcal{L}_{rec} = \mathbb{E}_{x,c,c'}[\|x - G(G(x, c), c')\|_1] \quad (6.12)$$

- **SSIM Loss** ($\mathcal{L}_{ssim}$): Measures perceptual similarity between the generated image and the paired ground truth image.

$$\mathcal{L}_{ssim} = 1 - \text{SSIM}(G(x, c), y) \quad (6.13)$$

The full objectives to minimise for the discriminator and generator are:

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls} \cdot \mathcal{L}_{cls}^{real} + \lambda_{gp} \cdot \mathcal{L}_{gp} \quad (6.14)$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls} \cdot \mathcal{L}_{cls}^{fake} + \lambda_{rec} \cdot \mathcal{L}_{rec} + \lambda_{ssim} \cdot \mathcal{L}_{ssim} \quad (6.15)$$

where λ_{cls} , λ_{rec} , λ_{ssim} , and λ_{gp} are scalar weights balancing each term.

6.3.2 Results

Despite the addition of supervision, the model exhibited problems similar to those of the other baselines. As shown in Figure 6.6, the generated images often failed to preserve the skin tone. In many cases, the predicted images resembled filtered versions of the input, lacking the structural and aesthetic changes present in the post-operative ground truth.

This behaviour suggests that the model overemphasised reconstruction and domain consistency rather than focusing on task-specific transformations (i.e., learning the surgical effect). Although the SSIM loss helped enforce some fidelity, the global training objective diluted this effect due to the presence of adversarial and domain classification losses.

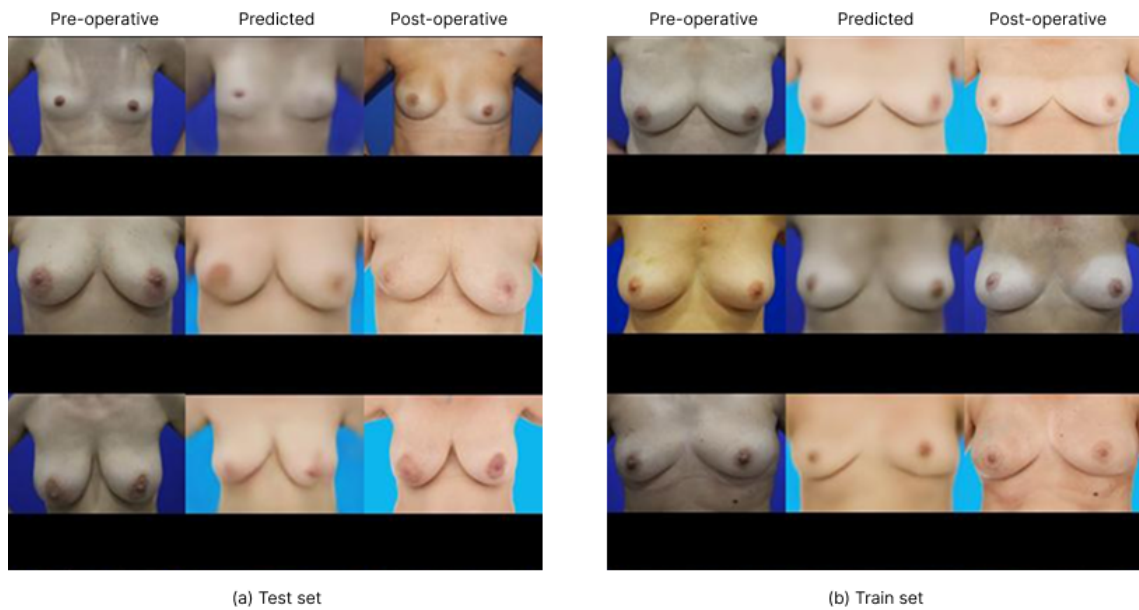


Figure 6.6: Visual results from StarGAN. Each row displays a pre-operative image, the predicted image, and the actual post-operative image. (a) Test set, (b) Training set.

When the reconstruction (or SSIM) loss was too strong, the generator tended to produce outputs that closely resembled the input image (Figure 6.7), suppressing visible surgical changes. Reducing the reconstruction weight improved visual differentiation but at the cost of skin tone fidelity and made the model behave more like Pix2Pix, since the reconstruction loss is the primary architectural difference between the two.

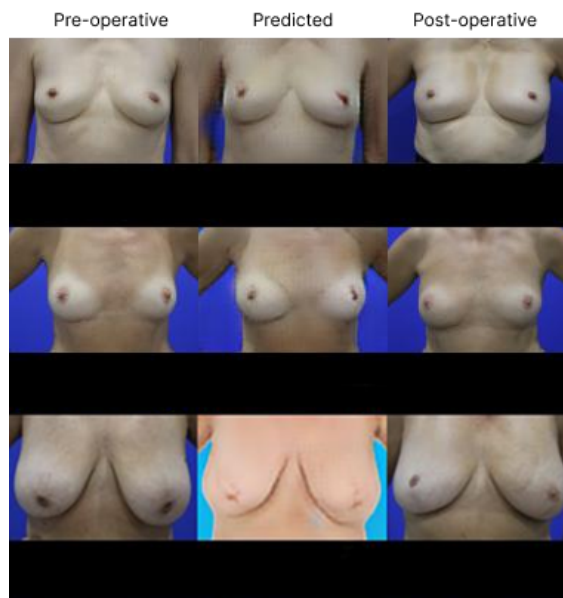


Figure 6.7: Effect of strong reconstruction loss in StarGAN. From left to right: pre-operative input, predicted image, and real post-operative outcome. The predicted image closely mirrors the input, showing minimal surgical changes.

6.4 Conditioning with Clinical Information

The previously presented findings showed that although deep generative models are capable of learning to convert pre-operative images into post-operative images, their outputs frequently lack anatomical and clinical precision without extra conditioning. This is mostly because some of these models do not have access to the rich contextual data that influences surgical outcomes; instead, they only use image-based features.

To produce more clinically meaningful and personalised results, the models must be guided by additional cues beyond the image itself. In particular, the visual outcome is specifically influenced by the type of surgical procedure and the side of the body that is affected. By explicitly encoding relevant variables into the model, the goal is to bridge the gap between appearance-based translation and clinically accurate image synthesis.

In this context, Pix2Pix and BicycleGAN were the target of these experiments, as these models do not inherently support conditioning and thus require explicit integration of clinical metadata. StarGAN, on the other hand, is already designed to handle multi-domain conditional image translation.

With this in mind, two types of clinical metadata were selected as conditioning variables: the type of surgery and the side on which the cancer was located (left, right, or bilateral). These variables were encoded using one-hot vectors and integrated into the generative models to better guide image synthesis in accordance with relevant clinical factors.

Table 5.1 lists the different types of surgery present in the dataset, along with their respective frequencies. For the purpose of training and validating the models, only conservative surgery cases were included in the experiments—namely, the most frequent classes: **C1**, **C2**, **C3**, and **C4**. Mastectomy-based procedures (e.g., with implants or autologous flaps) were excluded during training to reduce variability and due to the small amount of available data. However, these classes are analysed and discussed in the final evaluation section for qualitative insight.

6.4.1 Pix2Pix with Conditioning

To add clinical information into the Pix2Pix framework, the model was modified to condition both the generator and the discriminator on surgery-specific data. Through this conditioning process, the network should be able to produce post-operative images that not only accurately depict the intended clinical context, but also the input’s anatomical structure.

With this in mind, the type of surgery and the laterality of the cancer were one-hot encoded and concatenated into a single vector, which was added to the generator’s input. To inject this information, the one-hot vector was spatially repeated to match the height and width of the input image. Specifically, for an input image tensor of shape $[B, 3, H, W]$ and a condition vector $c \in \mathbb{R}^{B \times C}$, the vector was reshaped to $[B, C, 1, 1]$ and then expanded to $[B, C, H, W]$, which was then concatenated to the image, resulting in a combined input of shape $[B, 3 + C, H, W]$. This allowed the generator to access the clinical context from the first convolutional layer.

To further ensure that the generated images were semantically aligned with the conditioning variables, a classification loss was introduced. An auxiliary classifier was added to the discriminator to predict the conditioning labels (surgery type and cancer side) from both real and generated images. Cross-entropy loss was computed between the predicted and true class labels, and this loss was backpropagated not only through the discriminator but also through the generator, encouraging it to produce images consistent with the given conditions.

The full generator loss was therefore composed of three terms:

$$\mathcal{L}_G = \mathcal{L}_{\text{GAN}} + \lambda_{\text{L1}} \mathcal{L}_{\text{L1}} + \lambda_{\text{class}} \mathcal{L}_{\text{class}}, \quad (6.16)$$

Where:

- $\mathcal{L}_{\text{GAN}}$ is the adversarial loss used to encourage realistic image,
- $\mathcal{L}_{\text{L1}}$ is the pixel-wise reconstruction loss between the generated and ground-truth post-operative images,
- $\mathcal{L}_{\text{class}}$ is the auxiliary classification loss promoting semantic consistency,
- $\lambda_{\text{L1}}, \lambda_{\text{class}}$ are hyperparameters controlling their respective weights.

On the discriminator side, the architecture was adapted to perform multi-task learning. In addition to the standard real/fake classification, the discriminator included an auxiliary head responsible for predicting the conditioning labels. The discriminator's final layer outputs two branches: one sigmoid output for adversarial loss, and another softmax output for classifying the surgical conditions. The total discriminator loss was defined as:

$$\mathcal{L}_D = \mathcal{L}_{\text{adv}} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}}, \quad (6.17)$$

where $\mathcal{L}_{\text{adv}}$ is the standard binary adversarial loss, and $\mathcal{L}_{\text{aux}}$ is the classification loss on the conditioning labels. This setup ensured that the discriminator not only judged visual realism but also evaluated whether the image matched the provided clinical context.

6.4.1.1 Results

Together, these modifications enabled the Pix2Pix model to generate more clinically relevant outputs by embedding structured medical information directly into the generative process via early conditioning and auxiliary classification. As shown in Figure 6.8, the results generated with clinical conditions were consistently more aligned with the expected surgical outcomes, producing breast shapes and transformations that accurately reflected the type of procedure performed. In contrast, the baseline Pix2Pix model without conditioning often produced anatomically plausible changes, but these were not always coherent with the clinical context, occasionally introducing transformations inconsistent with the actual surgery.

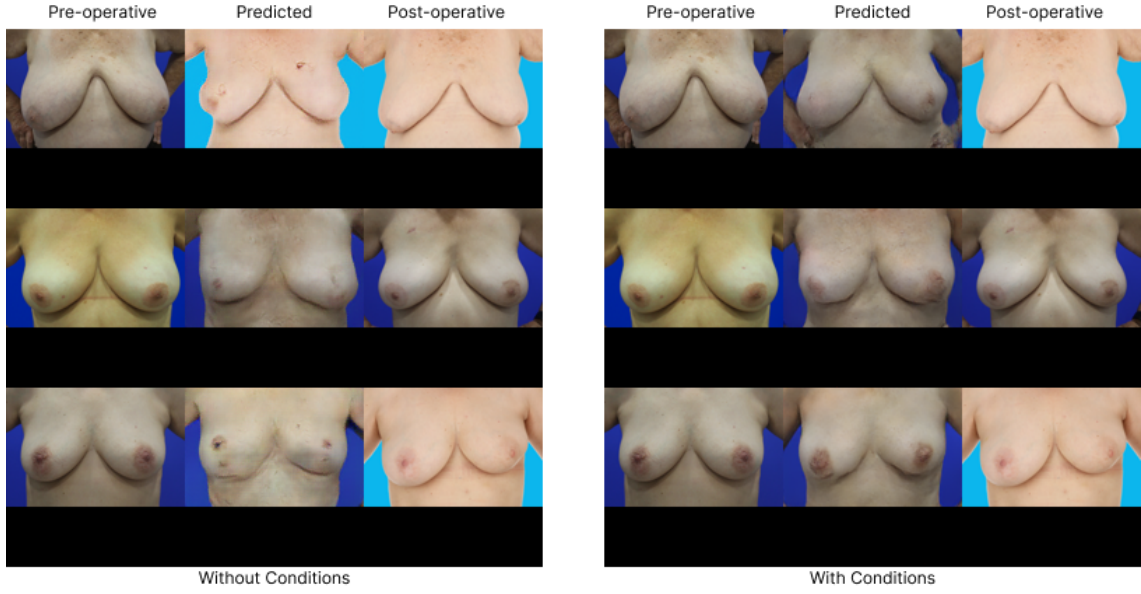


Figure 6.8: Visual comparison between Pix2Pix outputs without (left) and with (right) clinical conditions. The addition of surgery type and cancer side improves the realism and clinical relevance of the generated post-operative images.

6.4.2 BicycleGAN with Conditioning

Unlike the Pix2Pix model, which conditions the generator on the clinical vector alone, the BicycleGAN generator receives both a latent noise vector z and a clinical condition vector c concatenated together. The resulting vector $[c | z]$ is then spatially replicated and concatenated to the input image along the channel dimension. This enables the generator to produce diverse outputs for the same clinical condition by varying z , while still respecting the structure imposed by c .

As with the Pix2Pix model, the discriminator was modified to include auxiliary classification heads that predict the surgery type and cancer laterality from both real and generated images. The corresponding classification loss was computed using cross-entropy and backpropagated to both the generator and discriminator, reinforcing semantic consistency between generated content and clinical metadata.

The generator loss for BicycleGAN included the following terms:

$$\mathcal{L}_G = \mathcal{L}_{GAN}^{(1)} + \mathcal{L}_{GAN}^{(2)} + \lambda_{L1} \mathcal{L}_{L1} + \lambda_{KL} \mathcal{L}_{KL} + \lambda_{cls} \mathcal{L}_{class} \quad (6.18)$$

Where:

- $\mathcal{L}_{GAN}^{(1)}$ and $\mathcal{L}_{GAN}^{(2)}$ are adversarial losses for generated samples using encoded and random z values, respectively,
- $\mathcal{L}_{L1}$ and $\mathcal{L}_{KL}$ are standard reconstruction and KL divergence losses from the original BicycleGAN formulation,
- $\mathcal{L}_{class}$ is the classification loss from the discriminator's auxiliary heads.

6.4.2.1 Results

As shown in Figure 6.9, conditioning clearly improved the semantic alignment between input, prediction, and ground truth. The generated breasts better reflected the expected surgical outcomes when the model was guided by clinical metadata. In contrast, the unconditioned variant often applied transformations that were anatomically valid but not necessarily consistent with the intended procedure.

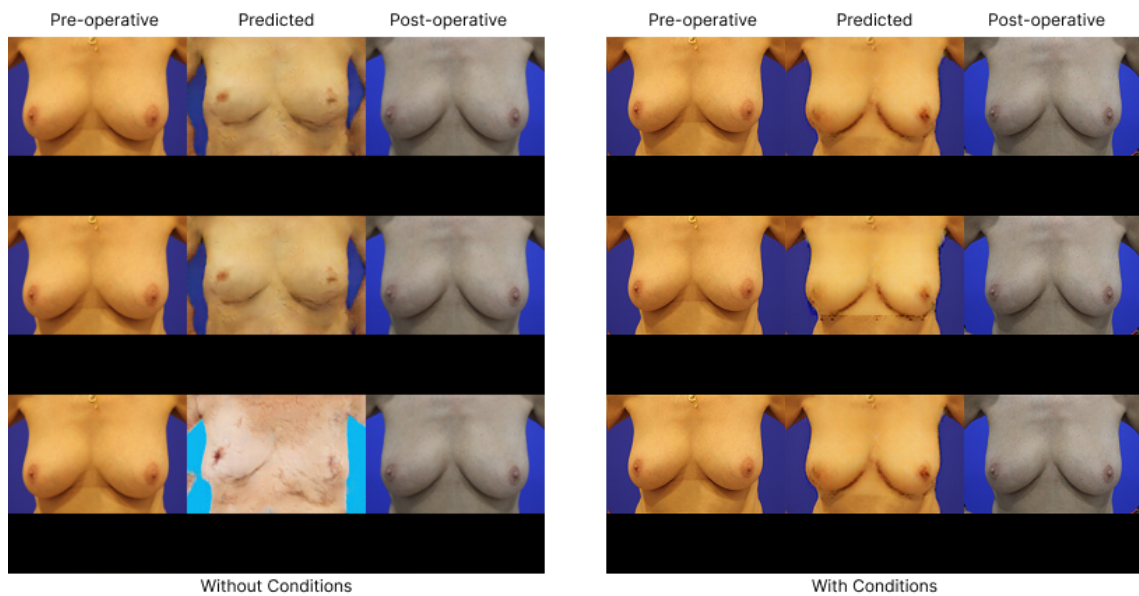


Figure 6.9: Comparison of results from BicycleGAN models trained with (right) and without (left) clinical conditions. The conditioned model more accurately captures expected surgical outcomes.

However, this conditioning came with a cost. While the model successfully prioritised the deterministic clinical variables, it often suppressed the latent noise vector, reducing the diversity of outputs across multiple samples for the same input. The generator learned to "ignore" stochastic variation in favour of maximising class fidelity. This highlights a fundamental challenge in balancing realism, semantic control, and diversity, particularly when conditioning information dominates the latent space.

6.5 Discussion

The baseline experiments indicated that although generative models such as Pix2Pix, BicycleGAN, and StarGAN can create realistic post-operative appearances, their unconditioned versions frequently struggle to generate transformations that are both anatomically and clinically accurate. Specifically, when Pix2Pix and BicycleGAN were trained without any metadata, they relied solely on visual cues. As a result, they often produced generic surgical outcomes, overlooking the specifics of the procedure or the patient context.

Introducing conditioning based on clinical metadata has enhanced the realism and relevance of the generated outcomes. This conditioning allows the models to produce outputs that align more closely with expected anatomical changes, especially in conservative surgeries where the differences between types (such as reduction versus implant) are subtle yet clinically significant. As illustrated in Figures 6.8 and 6.9, incorporating this guidance has helped the models learn surgery-specific transformations, including changes in breast volume, contour, and nipple location.

StarGAN naturally supports conditional generation and benefits from its domain label input structure. However, its effectiveness was highly sensitive to the balance between the reconstruction loss and adversarial loss. When the reconstruction loss is too heavy, the model defaults to reproducing the input image with minimal changes, sacrificing anatomical transformations in favour of structural consistency (Figure 6.7). When the reconstruction loss was reduced, the outputs diverged more noticeably, but at the cost of skin tone fidelity and finer image details.

Quantitative metrics for all models, as shown in Table 6.1, support these findings. Unconditioned models like Pix2Pix and BicycleGAN achieved moderate results in terms of SSIM and IoU. However, conditioning consistently enhanced structural accuracy, particularly for specific features such as breast and nipple IoU, while also preserving fidelity to the input body shape. Notably, Pix2Pix with conditioning attained a nipple IoU of 0.10, which is twice that of the unconditioned version, indicating a sharper localisation of important landmarks. Additionally, BicycleGAN with conditioning surpassed all baseline models across most metrics, suggesting it effectively balances realism, structural integrity, and clinical consistency when guided by supervision.

Table 6.1: Quantitative results across all evaluated models.

Model	SSIM (post) $\uparrow$	SSIM (pre) $\uparrow$	IOU (breast) $\uparrow$	IOU (nipple) $\uparrow$	Colour 1 $\downarrow$	Colour 2 $\downarrow$
Pix2Pix	0.75	0.73	0.77	0.05	0.07	0.07
BicycleGAN	0.76	0.74	0.78	0.07	0.07	0.07
Pix2Pix with conditions	0.76	0.75	0.78	0.10	0.07	0.07
BicycleGAN with conditions	0.77	0.77	0.79	0.08	0.07	0.06
StarGAN	0.76	0.77	0.75	0.05	0.06	0.06

Several limitations remained common across all approaches:

- **Skin tone inconsistency:** Due to differences in lighting, background, and camera settings between pre- and post-operative images, models frequently introduced colour shifts or unnatural lighting, even when trained with conditioning.
- **Overfitting:** Especially in low-data regimes or with models like Pix2Pix, overfitting to the training set led to less generalisable transformations and more artefacts on the testing set.
- **Lack of anatomical localisation:** Since most loss functions operated globally over the entire image, the network often failed to focus on the surgical area (i.e., the breasts), leading to diluted learning signals and less precise structural changes.

Despite these challenges, the conditioning experiments clearly demonstrated that embedding clinical metadata is critical for generating clinically meaningful, personalised post-operative predictions. By encoding procedure-specific and patient-specific context into the generative pipeline, the models became more capable of modelling real surgical effects, reducing ambiguity, and producing anatomically plausible outputs, especially when the surgical transformation was subtle.

Chapter 7

Improving Generalisation and Colour Fidelity with Targeted Losses

The generative models portrayed in earlier sections demonstrated potential in transforming pre-operative images into realistic post-operative images. However, they also exhibited significant limitations, especially in two areas: colour consistency (the generated images often contained skin tones that differed from the original) and overfitting.

This chapter describes different methods for addressing these issues. Many forms of augmentation and adaptive training measures were tested to address overfitting, and to address colour consistency, different targeted loss functions were explored.

The Pix2Pix model was initially employed to develop and evaluate all of the enhancements and provided a helpful starting point for experimentation, making it easier to iterate quickly and facilitating better evaluation of separate changes.

7.1 Overfitting

The generative models tested in the previous chapter exhibited a clear tendency to overfit, since the small dataset available led the models to quickly memorise the training data. The generator produced plausible outputs on the training set but had weak generalisation ability to unseen examples, often leading to artefacts. Two techniques were put in place to address these limitations: dropout in the generator and augmentation in the discriminator.

7.1.1 Dropout

Dropout layers were built into the intermediate layers of the generator's U-Net architecture. Dropout encourages the network to make less use of certain neuron activations, adding robustness and preventing memorisation. As a side effect, it also introduces some variability in the outputs given the same input, adding flexibility to its generative behaviour.

7.1.2 Discriminator Augmentation

While the generator exhibited signs of overfitting, regularising the discriminator also played a key role in improving training dynamics. Following the approach by Karras et al. [36], stochastic augmentations were applied to real and fake images before passing them to the discriminator. These augmentations prevent the discriminator from over-relying on low-level visual cues, making its task more challenging and encouraging it to learn more generalisable features. This, in turn, gives the generator more room to improve and reduces the risk of early mode collapse.

Therefore, various augmentations are applied dynamically to image pairs passed to the discriminator. Specifically, for each training step, a probability-controlled pipeline performs random geometric and colour-based transformations:

- **Affine rotation** sampled from $[0^\circ, 360^\circ]$ in 15° increments
- **Contrast and brightness jittering**, randomly scaled per channel
- **Hue and saturation shifts**, with global consistency between paired images
- **Grayscale conversion**, applied with reduced probability

Initially, discriminator augmentations were explored using an adaptive probability scheme, particularly in follow-up experiments on the simpler models presented in Section 6.4. The idea was to adjust the augmentation probability p dynamically based on the discriminator's average confidence on real images. If the discriminator became too confident (i.e., $\mathbb{E}[D(x_{\text{real}})] > t$), the augmentation strength was increased to make its task harder. Conversely, if confidence dropped, p was reduced to avoid excessive regularisation. The update rule was defined as:

$$p \leftarrow \begin{cases} \min(1.0, p + \delta) & \text{if } \mathbb{E}[D(x_{\text{real}})] > t \\ \max(0.0, p - \delta) & \text{otherwise} \end{cases} \quad (7.1)$$

where t is a target threshold (e.g., 0.6) and δ is a small adjustment step (e.g., 0.005).

In practice, however, p quickly saturated to its maximum value, as shown in Figure 7.1. This indicated that the discriminator rapidly required strong regularisation, making adaptiveness unnecessary beyond the first few iterations. To simplify training and reduce overhead, subsequent experiments used a fixed augmentation probability of $p = 0.8$, offering consistent and effective regularisation across all training stages.

7.1.3 Results

To evaluate the effect of regularisation on overfitting and artefact suppression, several Pix2Pix configurations were tested: (a) the baseline with clinical conditions, (b) with dropout added to the generator, and (c) with both generator dropout and adaptive discriminator augmentation. In this context, overfitting often manifests as visible artefacts, such as overly sharp textures, unnatural colour transitions, or noisy details. These issues are likely caused by the generator memorising

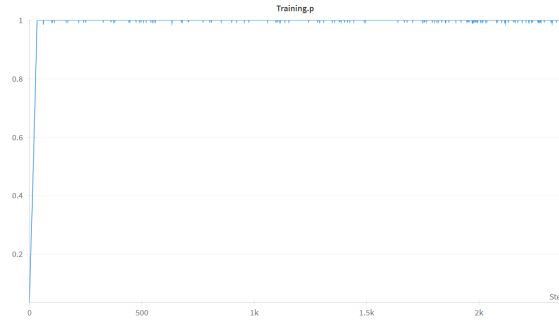


Figure 7.1: Evolution of the adaptive augmentation probability p during training. The value quickly saturates to near 1.0, suggesting that a fixed high value would suffice for regularisation.

specific outputs from the training set. When the model is evaluated on unseen data, it tries to apply these memorised transformations to new inputs that differ in appearance, resulting in degraded visual quality.

As shown in Figure 7.2, the visual quality improves from configuration (a) to (c). The baseline model (a) often produces outputs with visible artefacts and inconsistent colour tones. Adding dropout in (b) helps reduce these issues slightly by introducing variation during training. The most notable improvement comes from setup (c), where the combination of dropout and discriminator augmentation leads to cleaner, more consistent images with fewer artefacts. This suggests that the generator benefits from a more balanced training dynamic and learns to rely less on memorised details. Quantitatively, configuration (c) achieves the best performance in SSIM and IOU (Table 7.1), confirming that regularisation strategies help the model generalise more effectively.

Table 7.1: Test set performance of Pix2Pix models with increasing levels of regularisation.

Metric	SSIM (post) $\uparrow$	SSIM (pre) $\uparrow$	IOU (breast) $\uparrow$	IOU (nipple) $\uparrow$	Colour 1 $\downarrow$	Colour2 $\downarrow$
(a)	0.73	0.76	0.78	0.09	0.07	0.07
(b)	0.73	0.74	0.77	0.09	0.07	0.08
(c)	0.74	0.76	0.78	0.1	0.07	0.07

7.2 Colour Consistency

Recent results have highlighted a significant issue: the generated images frequently altered the patient’s skin tone and, in some instances, even changed their overall body shape. This undermines the system’s objective of providing patients with realistic, personalised predictions. The generated image should maintain the patient’s appearance as closely as possible, apart from the intended surgical changes, so patients can still recognise themselves in the final result.

To address this, several loss functions were implemented to improve the consistency of skin tone and background appearance across generated images. Each loss targets a different aspect of

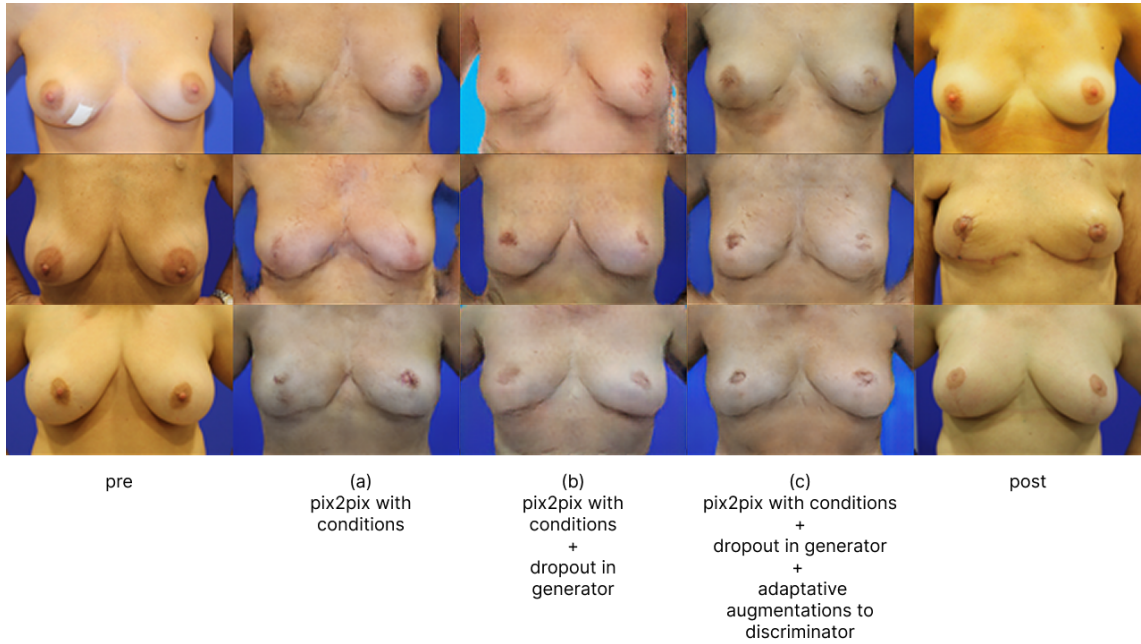


Figure 7.2: Qualitative results showing the effect of dropout and discriminator augmentation. Left: pre-operative image. Right: ground truth post-operative image. Middle columns: predictions from (a) baseline, (b) with dropout, and (c) with dropout and augmentation.

colour stability, helping to ensure that only the breast region changes while the rest of the body remains faithful to the pre-operative image.

7.2.1 Grayscale L1 Loss

A simple baseline approach to mitigate colour inconsistency was to discard colour entirely during training supervision. Both the generated image and the ground-truth post-operative image were converted to grayscale, and an L1 loss was applied:

$$\mathcal{L}_{\text{gray-L1}} = \|\text{Gray}(G(x)) - \text{Gray}(y)\|_1 \quad (7.2)$$

The model, however, adjusted the skin tone despite the absence of direct colour supervision. This suggests that the adversarial training gradually compensated for the missing RGB signal. Compared to the baseline with standard RGB L1 loss, the grayscale strategy required more training steps to reach visually coherent colour, but it eventually produced comparable results, although with slightly more desaturation and reduced colour vibrancy in some cases.

7.2.2 Global Colour Loss

To encourage skin tone consistency, the next step was to isolate the patient's body in the image and measure colour similarity only within those regions between the pre-operative image and the generated image. Since all clinical photographs were taken against a blue background, identifying

the body region was relatively straightforward: pixels where the red channel dominates over green and blue, and exceeds a brightness threshold, can be reliably interpreted as skin.

Formally, given a pre-operative image x_{real} , a binary mask $M_{\text{skin}} \in \{0, 1\}^{H \times W}$ is computed as:

$$M_{\text{skin}}(i, j) = \begin{cases} 1 & \text{if } R_{i,j} > 0.4 \text{ and } R_{i,j} > G_{i,j} \text{ and } R_{i,j} > B_{i,j} \\ 0 & \text{otherwise} \end{cases} \quad (7.3)$$

A first thought was applying a pixel-wise L1 loss over the masked region:

$$\mathcal{L}_{\text{masked-L1}} = \|M_{\text{skin}} \odot (G(x) - x)\|_1 \quad (7.4)$$

However, this approach would have produced undesired side effects. Since the mask corresponds to the actual body shape, the model was encouraged not only to match the skin colour, but also to align its generated breast shape with the pre-image, unintentionally limiting its ability to simulate realistic post-operative transformations.

As a workaround, the comparison happens only for the average colour inside the skin regions. For each image, the mean RGB vector is computed within the mask (Figure 7.3):

$$\bar{c}_{\text{fake}} = \frac{\sum_{i,j} M_{i,j} \cdot G(x)_{i,j}}{\sum_{i,j} M_{i,j} + \varepsilon}, \quad \bar{c}_{\text{real}} = \frac{\sum_{i,j} M_{i,j} \cdot x_{i,j}}{\sum_{i,j} M_{i,j} + \varepsilon} \quad (7.5)$$

where ε is a small constant to avoid division by zero. The final colour consistency loss is then:

$$\mathcal{L}_{\text{colour}}^{\text{global}} = \|\bar{c}_{\text{fake}} - \bar{c}_{\text{real}}\|_1 \quad (7.6)$$

This approach prevents the model from learning specific anatomical shapes or details, while still encouraging overall colour alignment between the generated and reference images.

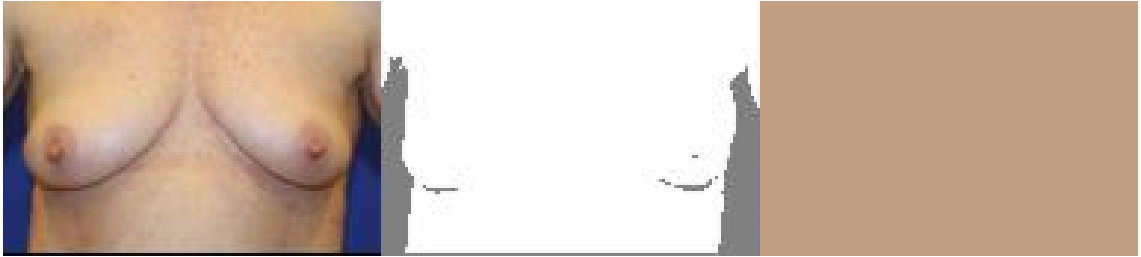


Figure 7.3: Left: pre-operative image; Middle: binary skin mask using RGB heuristic; Right: visualisation of average colour extracted from masked region.

7.2.3 Histogram Loss

Another strategy explored for colour alignment was to match the colour distributions of the generated and pre-operative images using histogram comparison. The initial implementation computed a histogram of all the pixels in the image. However, this led to poor results, as the blue background,

which dominates many of the images, disproportionately influenced the loss, often leading to unnatural skin tones.

To address this, the histogram was later restricted to only the skin regions. As described previously, a binary skin mask M_{skin} is computed using a simple RGB heuristic. Then, instead of comparing pixel values directly, histograms are constructed over pixel intensities in each RGB channel, limited to the skin area.

For each colour channel, histograms are computed over the non-zero masked pixels. Let h_c^{fake} and h_c^{real} be the histograms for the channel $c \in \{R, G, B\}$. These are then normalised into probability distributions and compared using the Kullback–Leibler divergence:

$$\mathcal{L}_{\text{colour}}^{\text{hist}} = \frac{1}{3} \sum_{c \in \{R, G, B\}} \sum_{i=1}^N h_{c,i}^{\text{fake}} \log \left(\frac{h_{c,i}^{\text{fake}} + \varepsilon}{h_{c,i}^{\text{real}} + \varepsilon} \right) \quad (7.7)$$

where N is the number of bins (typically 32) and ε is a small constant to ensure numerical stability.

This formulation ensures that the generator learns to match the overall colour distribution of the skin, rather than individual pixels. Compared to global L1-based colour losses, this approach allows for more tolerance in local variations while enforcing global appearance realism.

7.2.4 Patch-Based Average Colour Loss

To preserve colour consistency without enforcing structural alignment, a patch-based average colour loss was implemented. The idea is to divide both the generated and reference images into non-overlapping square patches and compare their average RGB values. This is computed using average pooling followed by L1 distance:

$$\mathcal{L}_{\text{colour}}^{\text{patch}} = \|\text{AvgPool}(G(x)) - \text{AvgPool}(x)\|_1 \quad (7.8)$$

Patch size is a critical factor: it must be large enough to avoid encoding structural information such as the breast outline. As illustrated in Figure 7.4, smaller patches (e.g., 16×16) may begin to reveal anatomical patterns, while larger ones (e.g., 32×32) abstract only regional tone.

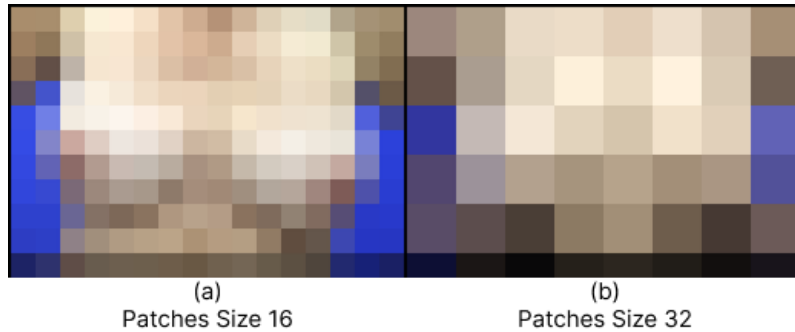


Figure 7.4: Effect of patch size on average colour in images of 256x256. (a) 16×16 patches preserve some shape; (b) 32×32 patches smooth out structural details.

7.2.5 Bounding Box Masked Colour Loss

A key challenge in post-operative image synthesis is localising the supervision. Direct pixel-wise losses over the entire image can be counterproductive, as regions outside the breast area (e.g., arms, torso, background) are not intended to change. Moreover, inconsistencies in pose arm position between pre- and post-operative images can introduce artefacts if those areas are penalised (Figure 7.5).

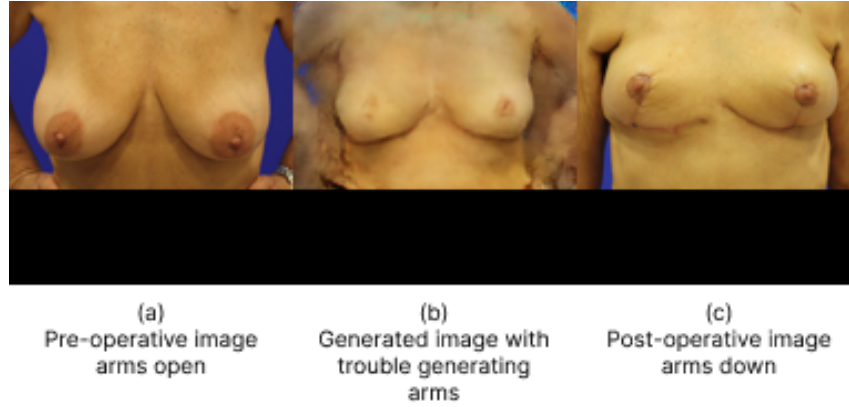


Figure 7.5: Pre-operative and post-operative images may differ in pose (e.g., arm position), so penalising pixel differences globally may hurt learning. The bounding box strategy prevents this by ignoring non-surgical regions.

To address this, a bounding box strategy was introduced to spatially separate supervision into two distinct zones:

- **Inside the bounding box** (corresponding to the breast region), shape-focused supervision is applied using L1 and SSIM losses on grayscale images.
- **Outside the bounding box**, a colour consistency constraint is enforced using L1 loss to preserve the appearance of the original body, ensuring the generator does not change unrelated anatomy.

Formally, let M_{in} and M_{out} be binary masks for the inside and outside of the bounding box, respectively. Let x_{fake} and x_{real} be the generated and post-operative images, and x_{pre} the original pre-operative image. The loss terms are defined as:

Inside (breast region):

$$\mathcal{L}_{\text{bbox}}^{\text{L1}} = \|M_{\text{in}} \odot (G(x_{\text{pre}}) - x_{\text{real}})\|_1 \quad (7.9)$$

$$\mathcal{L}_{\text{bbox}}^{\text{SSIM}} = 1 - \text{SSIM}(M_{\text{in}} \odot G(x_{\text{pre}}), M_{\text{in}} \odot x_{\text{real}}) \quad (7.10)$$

Outside (preserve original anatomy):

$$\mathcal{L}_{\text{colour}}^{\text{out}} = \|M_{\text{out}} \odot (G(x_{\text{pre}}) - x_{\text{pre}})\|_1 \quad (7.11)$$

This separation allows the model to focus learning on surgical outcomes in the breast area while preserving anatomical realism elsewhere. The adversarial loss helps reconcile tone differences across regions, producing globally coherent images.

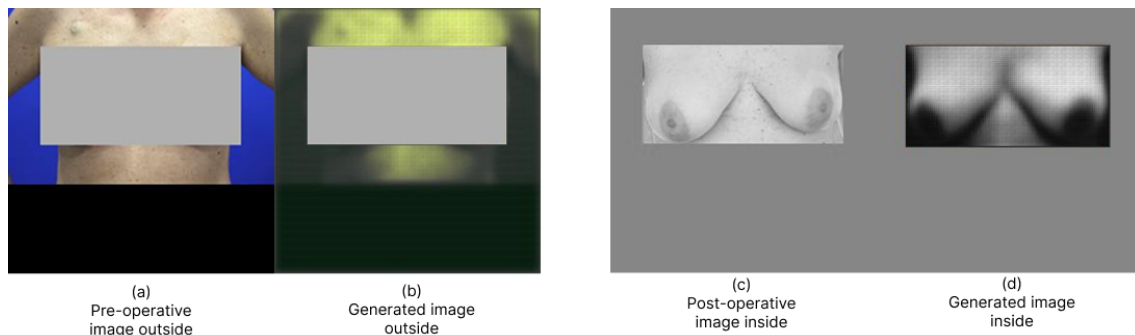


Figure 7.6: Right: the shape of the breast is supervised using grayscale losses inside the bounding box; Left: the colour outside is preserved from the pre-operative image.

7.2.6 Results

Despite extensive experimentation, achieving consistent and realistic skin tone across generated post-operative images remains a challenging problem. Each loss function introduced in this section improves upon specific aspects, but none fully resolves the issue of colour inconsistency.

The **global colour loss** aligns the overall hue by comparing average RGB values within skin regions. While this avoids learning structural details, it does not capture local variations, often resulting in random strength of colours in the shoulders and arms (see the "colour" column in Figure 7.7).

Histogram-based losses aim to match the full distribution of colour values. However, early implementations that computed histograms over the entire image were heavily biased by the blue background, producing unnatural and erratic results. This can be seen in the "histogram" column (Figure 7.7), where many predictions exhibit unrealistic or mismatched tones. Even after masking the skin region, this approach only partially corrects the issue.

The **patch-based colour loss** does not solve the problem either. Since it enforces similarity on coarse spatial regions, it avoids learning the shape of the pre-image, but it can still generate patchy or desaturated skin colour, failing to deliver the desired realism.

Bounding-box constraints also help isolate shape-related and colour-related supervision, preserving the patient's body while allowing accurate breast modification. This improves anatomical realism but still struggles with subtle colour transitions across the boundary region.

Table 7.2 shows that patch-based and bbox-masked models perform better quantitatively but still do not ensure colour stability in all cases.

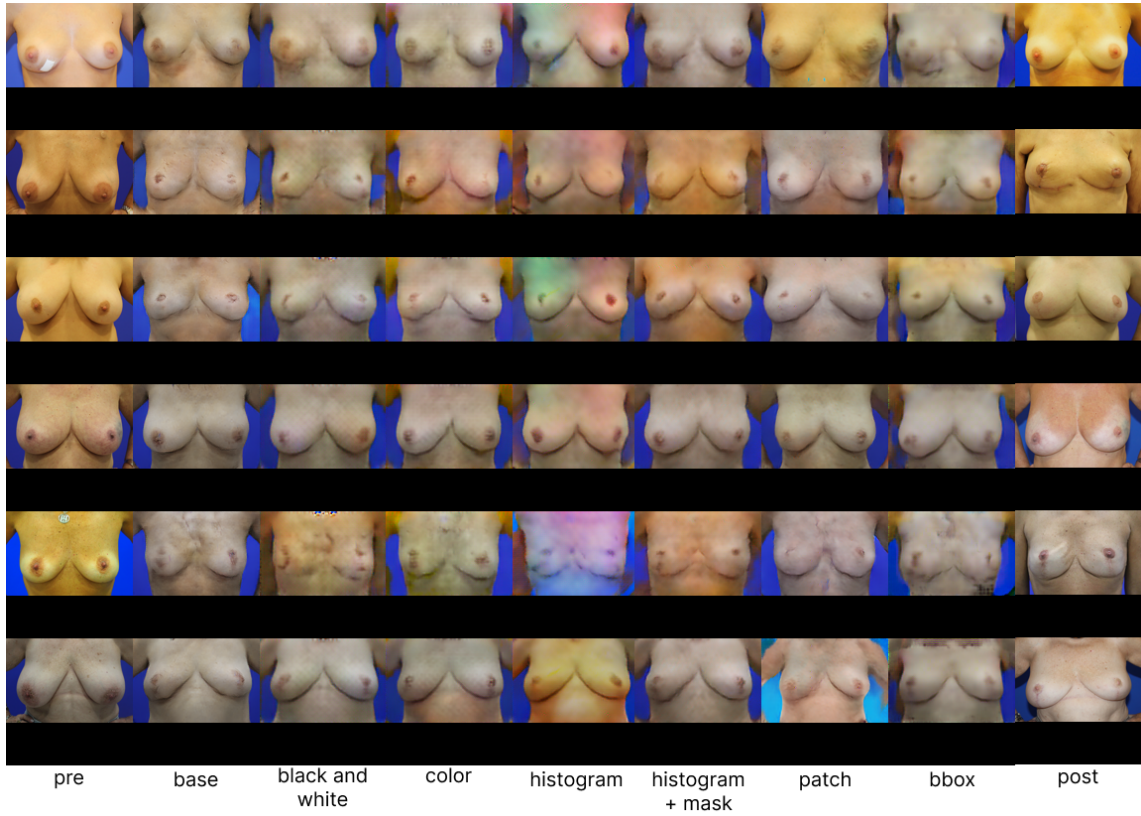


Figure 7.7: Visual comparison of different colour supervision strategies across a fixed set of examples.

Table 7.2: Quantitative comparison of colour supervision strategies,

Metric	SSIM (post) $\uparrow$	SSIM (pre) $\uparrow$	IOU (breast) $\uparrow$	IOU (nipple) $\uparrow$	Colour 1 $\downarrow$	Colour 2 $\downarrow$
black and white	0.76	0.74	0.79	0.09	0.07	0.07
colour	0.77	0.73	0.79	0.08	0.07	0.07
histogram	0.76	0.73	0.79	0.10	0.07	0.07
histogram + mask	0.76	0.75	0.78	0.08	0.06	0.06
patch	0.75	0.75	0.78	0.08	0.06	0.06
bbox	0.77	0.74	0.78	0.07	0.06	0.06

7.3 Discussion

This chapter tackled two major limitations observed in baseline generative models: overfitting and poor colour consistency. Both issues hindered the clinical reliability of the generated post-operative predictions and motivated the exploration of targeted strategies to mitigate them.

To address overfitting, dropout was added to the generator, and discriminator augmentation was applied during training. These modifications successfully reduced artefacts and improved generalisation. The combination of both techniques resulted in slightly cleaner and more stable outputs, as evidenced by qualitative and quantitative metrics.

Colour consistency proved to be an even more persistent challenge. The objective in this context was not to match the colour of the ground-truth post-operative image, but to preserve the skin tone, lighting, and body appearance of the pre-operative image, ensuring that the patient could still recognise themselves in the generated result. Unfortunately, the adversarial loss consistently pulled the generator toward the post-operative distribution, where lighting conditions, imaging artefacts, and colour tones were often inconsistent or unrepresentative. Despite extensive experimentation with grayscale supervision, histogram matching, patch-based colour losses, and bounding box masking, the generator frequently altered skin tone, desaturated colours, or introduced unintended variations across the image.

These findings suggest that loss-level interventions alone are insufficient. Even though the proposed constraints targeted the correct regions and features, they could not fully overcome the influence of the adversarial signal. The generator was still optimising toward realism from a global dataset perspective, rather than maintaining fidelity to the patient’s original appearance. Altogether, this motivates a shift from generic loss engineering to architectural redesign. Rather than relying on indirect supervision to enforce colour or shape consistency, the next chapter proposes a new architecture that explicitly incorporates clinical context, region-aware discriminators, and spatial constraints. These changes aim to align the model’s learning objectives more directly with the requirements of personalised, anatomically faithful outcome prediction.

Chapter 8

Architectural Exploration

The previous tests struggled to preserve the patients' skin tone in the generated image, leading to realistic but unfaithful results regarding what the patient looked like at the moment of inference. The models often learnt the lighting or pose that the post-operative images usually had, leading to tone shift and patchy textures. This chapter explores different enhancements to the generator and discriminator to create realistic and semantically correct post-operative images.

The chapter begins by outlining the loss components and then presents a series of experimental architecture variants.

8.1 Loss Components

The model uses a multi-term loss function to supervise both visual consistency and clinical realism. Each term targets a specific aspect of image generation, and its visual illustrations are shown in Figure 8.1.

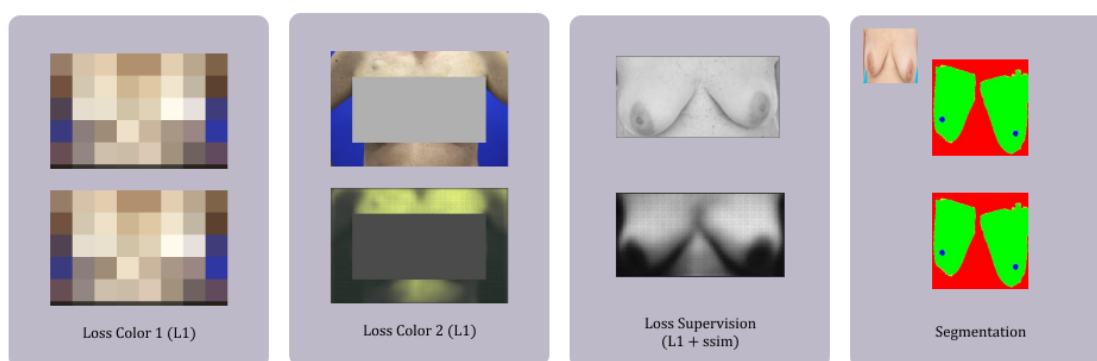


Figure 8.1: Visual examples of the four main loss components: Colour Loss 1 (patch-based MSE), Colour Loss 2 (outside-mask MSE), Structural Supervision (L1 + SSIM in grayscale), and segmentation loss. Each pair of images shows the inputs used in the respective loss calculation.

- **Loss Colour 1 (Patch Colour L1):** As discussed in Section 7.2.4, this loss compares average colours across spatial patches in the generated and input images to promote a uniform tone across the full body. It supports colour consistency without enforcing rigid spatial constraints.
- **Loss Colour 2 (Outside Mask L1):** Introduced earlier in Section 7.2.5, this term helps preserve the patient's body shape and original skin tone by applying an L1 loss between the generated image $\hat{B}$ and the pre-operative image A , excluding the breast region.
- **Loss Supervision (L1 + SSIM in Grayscale):** Also described in Section 7.2.5, this term enforces perceptual similarity between the generated and real post-operative images within the breast region. It combines grayscale L1 and SSIM losses to emphasise structural accuracy.
- **Segmentation Loss:** A pretrained segmentation model predicts anatomical masks for both real and generated post-operative images, and a class-weighted cross-entropy loss is used to align them. This encourages the generator to produce anatomically realistic transformations, particularly in the breast and nipple regions. This loss term aims to promote the alteration to the shape of the breasts according to the post-operative image, and its effectiveness is demonstrated in Section 8.3.3.

8.2 Architectural Variants Explored

This Section explores a set of architectural configurations aimed at improving image quality, anatomical realism, and semantic consistency. These variants were designed to test different combinations of generator structure, discriminator input, and classification strategy. In particular, the goal was to resolve two core challenges: maintaining patient-specific colour and appearance, and enforcing semantic correctness (e.g., surgery type, cancer side) without compromising visual quality.

A key insight that emerged from early experiments was that when the discriminator is trained only on post-operative images, it learns not only to detect realistic breast transformations but also to associate "realism" with the specific colour and lighting characteristics that dominate the post-operative dataset. This includes tone, background, and brightness changes that are not intrinsic to the surgery itself. This presents a critical problem in clinical settings. When showing a patient a possible post-surgery outcome, the generated image should reflect how her body might change, not alter aspects of her identity, such as skin tone or overall appearance. If the model produces an image that is realistic in shape but fails to resemble the patient's natural colour, it undermines trust and interpretability.

To address this, all architectural variants adopted a common modification: instead of providing the discriminator with only the post-operative image B , it also receives the corresponding pre-operative image A . This input change removes the implicit bias that a post-operative image must

"look" a certain way regarding colour. Instead, the discriminator learns just to evaluate whether $\hat{B}$ appears real.

Each architecture was implemented, trained, and evaluated using the same dataset and loss functions.

8.2.1 Two Discriminators with Local Classification and Global Realism

This variant incorporates two discriminators: one local and one global. The local discriminator D_1 is trained on image pairs, either (A, B) or $(A, \hat{B})$, and is responsible for both real/fake discrimination and classification of surgery type and cancer side.

The global discriminator D_2 receives single images, either A , B , or $\hat{B}$, and focuses purely on adversarial supervision. Its objective is to enforce global realism by encouraging structural consistency, plausible lighting, and natural background across generated samples.

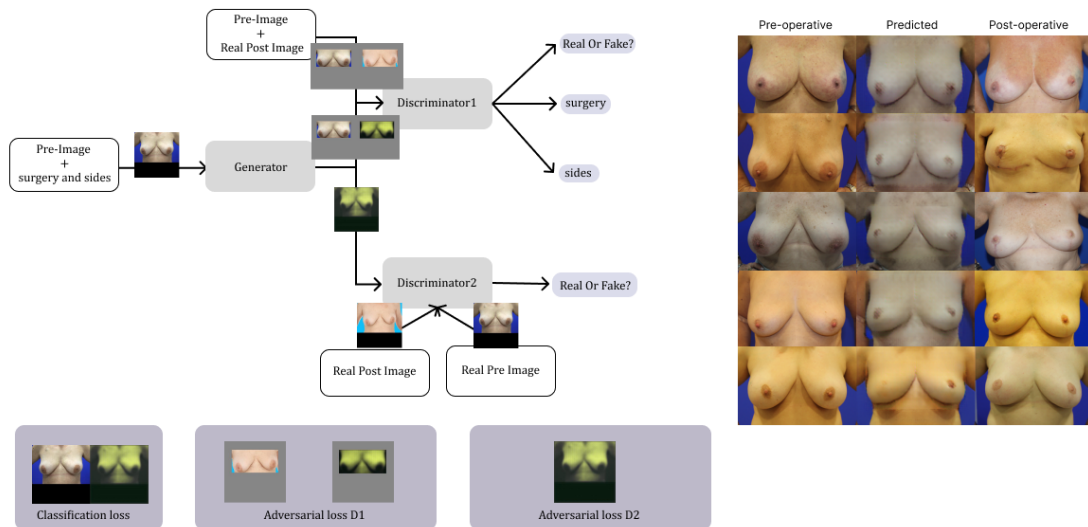


Figure 8.2: Architecture with two discriminators: D_2 (global) performs real/fake discrimination using pre/post pairs; D_1 (local) classifies and discriminates based on real or fake post-operative images.

This structure, however, introduced a significant conflict between the colour loss and the breast-focused discriminator, since D_1 , being both adversarial and semantic, learned strong priors on typical post-operative colour distributions. This often pushed the generator toward a biased hue, especially in the breast region. As Figure 8.2 shows, the arms often preserved the correct skin tone, but the central chest and breast areas, where surgical changes occurred and the local adversarial loss was applied, showed shifted hues.

8.2.2 Encoder for Classification

The previous configuration, which combined local adversarial supervision and semantic classification within a single discriminator D_1 , highlighted a crucial trade-off. By training D_1 to handle both

real/fake discrimination and surgery classification, the model began to associate post-operative realism with specific colour distributions present in the training set. This led to a shift in skin tone within the breast region of generated images, changing the overall appearance of the patient.

To address this issue, the adversarial component of D_1 was removed, and the classification task was reassigned to a dedicated encoder. This adjustment allowed the architecture to separate the two competing objectives:

- The discriminator D focuses exclusively on visual realism. It receives single images, either pre-operative A , real post-operative B , or generated $\hat{B}$, and is trained only with an adversarial loss.
- The encoder E receives the image pair $(A, \hat{B})$ and is trained to predict the surgery type and cancer side. By isolating the semantic task, the encoder can focus solely on learning meaningful clinical patterns, without introducing colour shifts or adversarial gradients into the generator.
- The generator continues to be conditioned on the surgery metadata and learns to balance realism, structure, semantic alignment, and colour fidelity under supervision from both the discriminator and the encoder.

This architectural separation ensures that each component is optimised for a single objective, reducing cross-task interference and improving the quality and interpretability of the generated images.

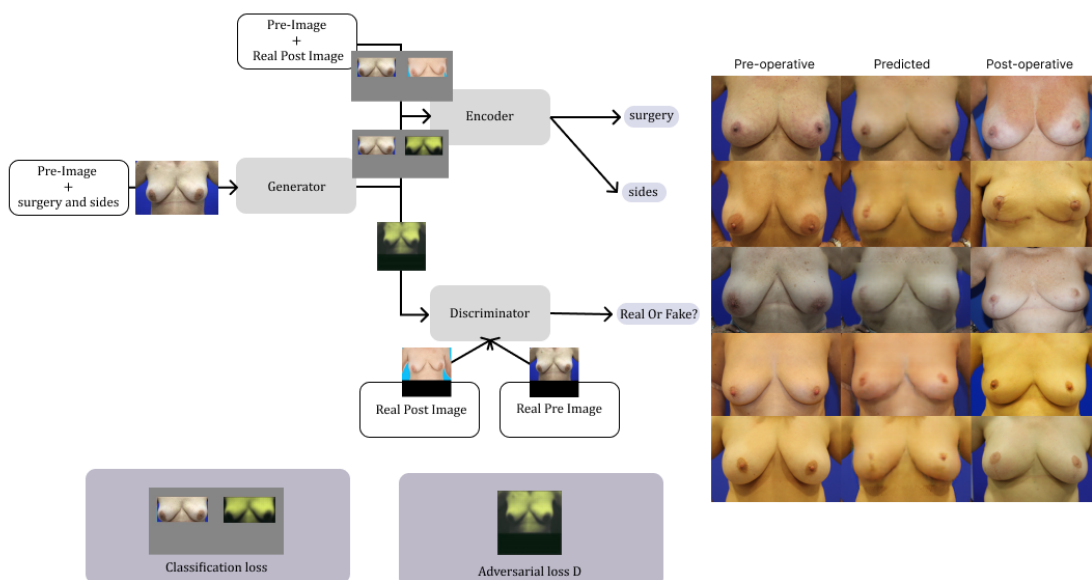


Figure 8.3: Architecture and results. The discriminator evaluates realism from both the pre- and post-operative images. The encoder classifies surgical conditions. The generator is conditioned on surgery type and cancer side.

The generator loss integrates all objectives as follows:

$$\mathcal{L}_G = \underbrace{\mathcal{L}_{\text{GAN}}}_{\text{Adversarial}} + \lambda_{\text{L1}} \cdot \underbrace{\mathcal{L}_{\text{structure}}}_{\text{Grayscale SSIM + L1}} + \lambda_{\text{class}} \cdot \underbrace{\mathcal{L}_{\text{class}}}_{\text{Surgery + Side Classification}} + \lambda_{\text{colour}} \cdot \left(\underbrace{\mathcal{L}_{\text{colour1}}}_{\text{Patch Colour Loss}} + \underbrace{\mathcal{L}_{\text{colour2}}}_{\text{Outside Mask L1}} \right) + \lambda_{\text{seg}} \cdot \underbrace{\mathcal{L}_{\text{seg}}}_{\text{Segmentation Consistency}} \quad (8.1)$$

8.2.2.1 Encoder Input: Masked (Breast Only) vs. Full Image

This experiment compares two strategies for the input to the encoder network responsible for classifying surgery type and cancer side:

- **Masked version:** The encoder receives only the breast region, obtained by masking the image using bounding boxes derived from breast keypoints. This removes distractions such as arms, torso, and background, focusing classification purely on surgical alterations.
- **Unmasked version:** The encoder receives the full image that provides additional context, such as posture, but risks introducing visual noise unrelated to surgery.

Figure 8.4 shows examples of both input variants.

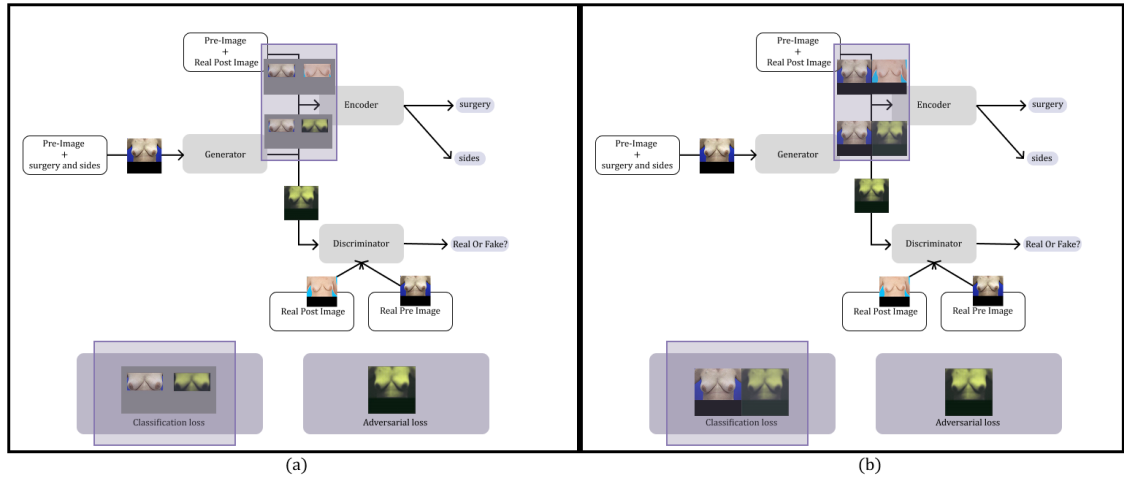


Figure 8.4: The encoder input comparison is illustrated with purple rectangular boxes highlighting the differences. (a): cropped breast region (masked version). (b): full-body image (unmasked version). The masked setup isolates surgical features, while the unmasked version incorporates background and posture information.

The quantitative results are presented in Table 8.1. Both configurations perform similarly across all metrics, with the masked version showing slightly higher SSIM values and slightly better colour consistency (Colour Loss 2). Qualitatively, as shown in Figure 8.5, the masked encoder appears to yield slightly more detailed and sharper reconstructions in the breast region

for some patients, particularly in the nipple area. Overall, while the masked setup may provide a slight perceptual advantage by narrowing the encoder’s attention to surgically relevant structures, the performance differences are minor.

Table 8.1: Comparison of full-image and masked encoder inputs across visual similarity (SSIM), anatomical accuracy (IOU), and colour preservation losses.

	SSIM (post) ↑	SSIM (pre) ↑	IOU (breast) ↑	IOU (nipple) ↑	Colour 1 ↓	Colour 2 ↓
No Mask	0.78	0.85	0.79	0.08	0.02	0.01
Masked	0.78	0.86	0.79	0.08	0.01	0.01

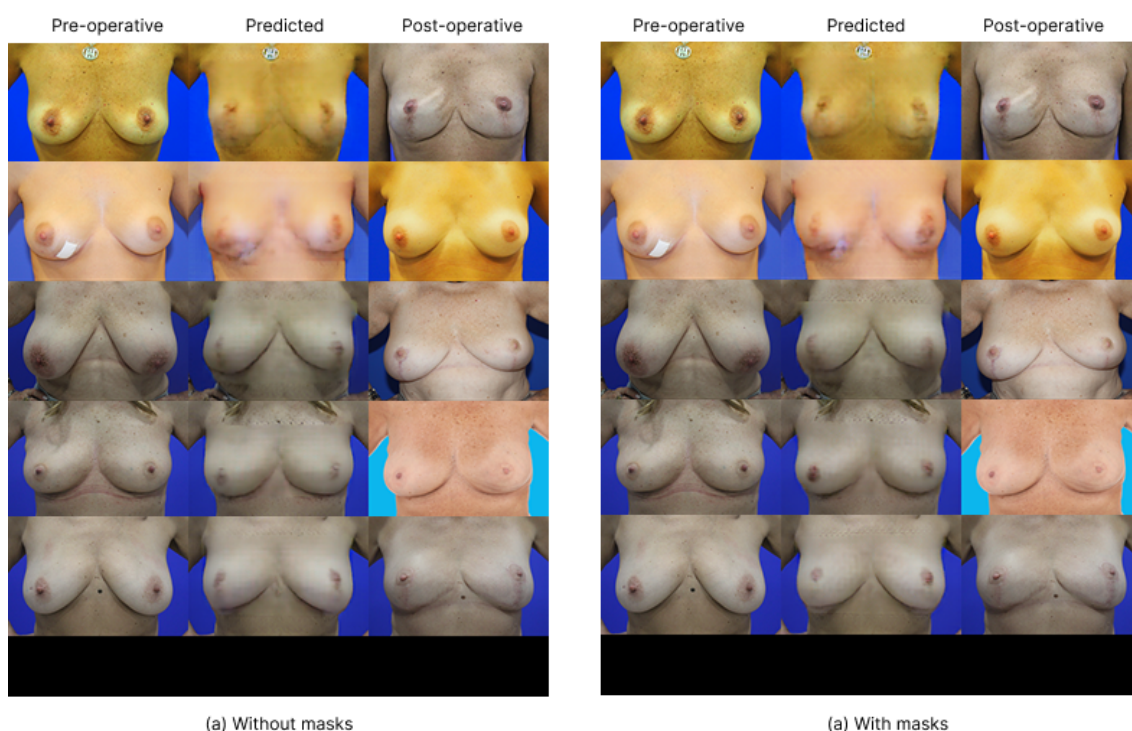


Figure 8.5: Comparison of generated images using encoder inputs with and without breast-region masking. Columns: Pre-operative, Predicted, Real Post-operative. Rows show different patients. Masked versions sometimes show clearer nipples or surgical scars.

8.2.3 Single Discriminator with Post-Only Classification Supervision

This architectural variant uses a single discriminator to perform both adversarial and classification tasks. Here, instead of introducing a separate encoder for semantic classification, the goal was to consolidate all tasks into a unified discriminator. The training strategy is the following:

- The discriminator receives triplets consisting of a pre-operative image A , a post-operative image B and a generated image $\hat{B}$ and learns to distinguish real from fake.

- For the classification head, the discriminator is trained to predict surgery type and cancer side **only on real post-operative images B** .

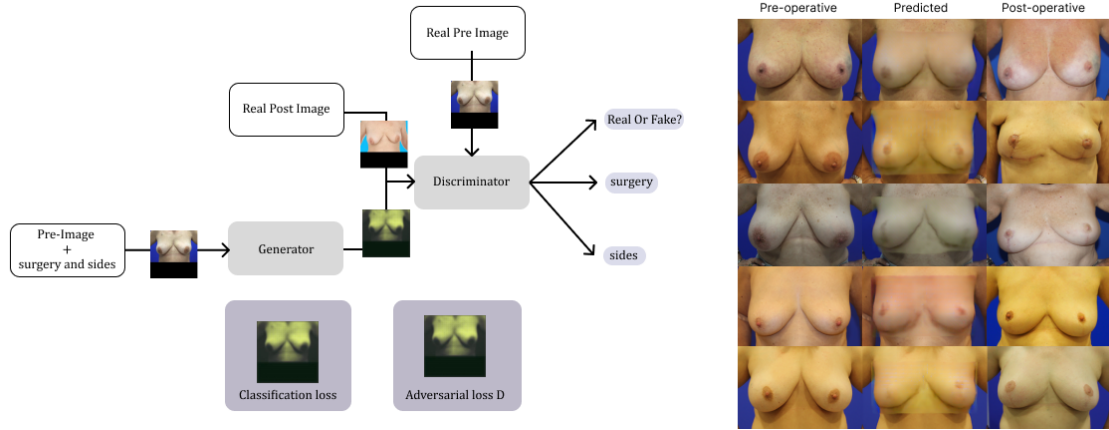


Figure 8.6: Single-discriminator architecture. The discriminator learns real vs. fake using pre- and post-operative images. For classification, it is only trained on real post-operative images. The generator is penalised if the discriminator misclassifies its generated output.

This architecture aimed to maintain the identity-preserving benefits of using the pre-image in the discriminator while simplifying the pipeline by removing the encoder. However, in practice, it often underperformed due to conflicting optimisation pressures: the discriminator had to jointly learn realism and clinical semantics, but from different inputs and objectives. This entanglement made the classification less reliable and led to blurrier generations compared to the architecture with a dedicated encoder. In addition, the predicted images frequently exhibited a faint but visible rectangular boundary around the breast region, likely caused by the bounding box supervision not blending naturally with the rest of the image. This artefact disrupted the visual continuity of the skin and made the outputs appear less realistic overall, as seen in Figure 8.6.

8.2.4 Encoder with Local and Global Discriminators

This architecture introduces two discriminators, one operating globally and one locally, alongside a dedicated encoder for semantic classification. The motivation for this design was to further enhance the quality of the breast region, which in earlier models was a bit blurred.

Previous experiments showed that training the discriminator with both pre- and post-operative images helped preserve patient-specific features like skin tone and shape. However, the architecture that unified classification and adversarial tasks in a single network showed trade-offs in both realism and semantic clarity.

To decouple these objectives, this version splits the architecture into three cooperating branches:

- A standalone **encoder** receives both the pre-operative image A and the generated post-operative image $\hat{B}$, and is trained to classify surgery type and cancer side.

- A **global discriminator** sees the full image and learns to distinguish real from fake, focusing on overall structure, lighting, and body alignment.
- A **local discriminator** operates on the cropped breast region (as defined by bounding boxes) and focuses on fine-scale texture and shape consistency, especially around the nipple and surgical area.

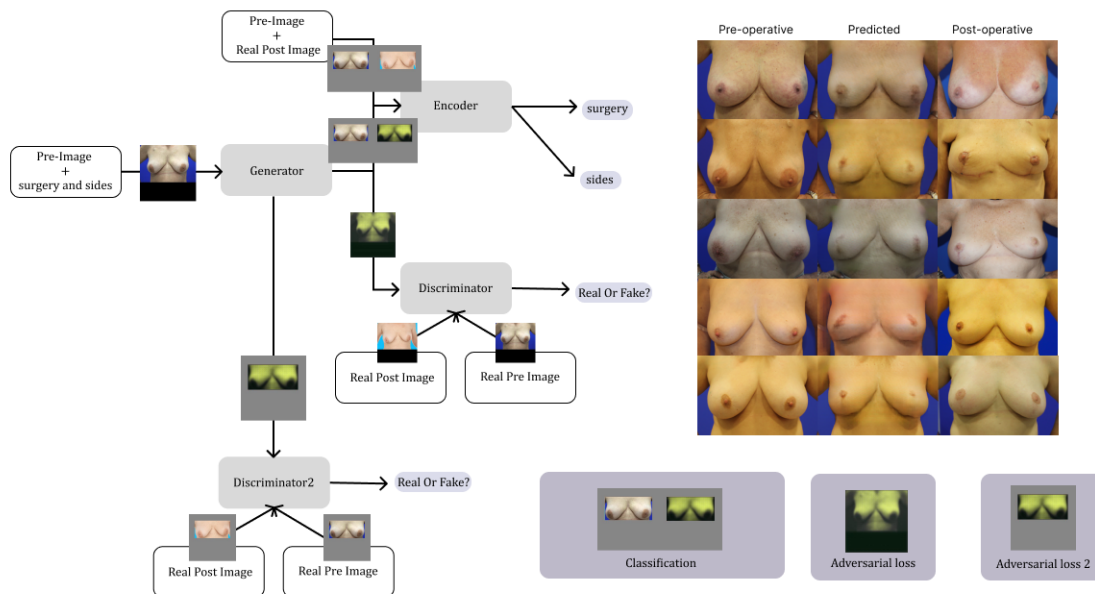


Figure 8.7: Architecture with an encoder and two discriminators. Classification is handled independently by the encoder, while adversarial realism is enforced globally and locally.

While the separation of tasks led to more stable training and sharper outputs in some cases, the predicted images still suffered from certain visual artefacts. In particular, a faint bounding box boundary around the breast region was often visible, suggesting that the region-specific supervision was not fully integrated with the surrounding anatomy. Additionally, some generations appeared slightly blurred, lacking the texture, sharpness and fine detail needed for highly realistic outcomes, Figure 8.7.

8.2.5 Results

Figure 8.8 presents a qualitative comparison of five model variants explored in this chapter. The encoder-based model (Section 8.2.2) produces the most visually consistent results in terms of colour tone, while also preserving realistic and coherent anatomical structures. In contrast, the two-discriminator model experiences colour shifts, and the single-discriminator model often shows sharp, square-shaped visual artefacts around the breast area, which are conflicts in the adversarial and classification tasks.

Meanwhile, the hybrid model, which combines an encoder with two discriminators, tends to generate slightly blurrier outputs, along with occasional rectangular artefacts near the supervised

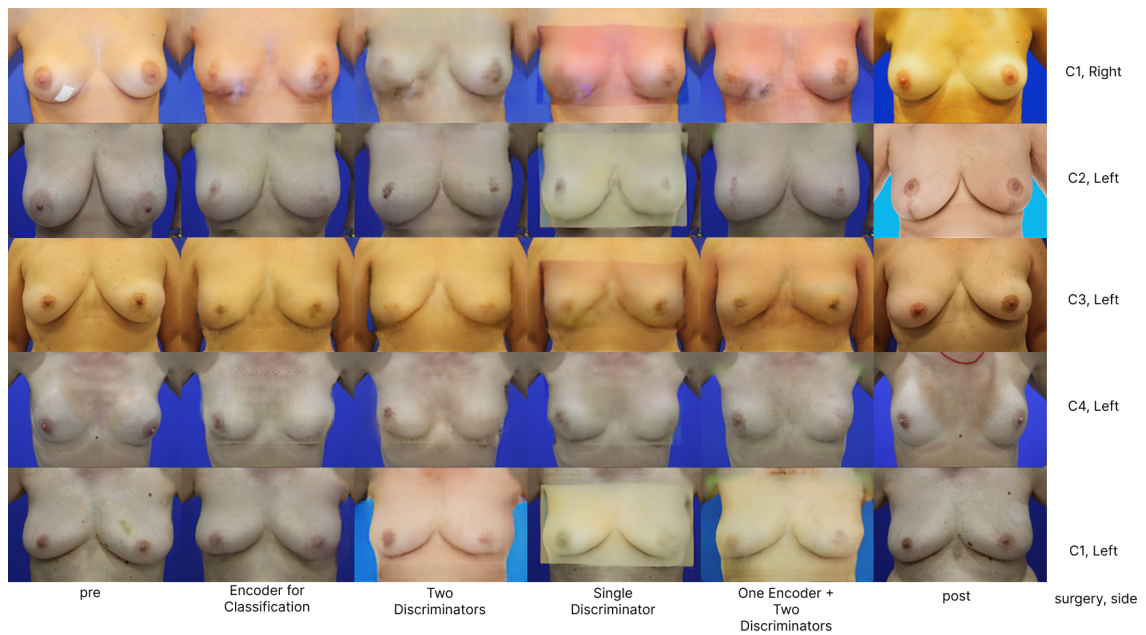


Figure 8.8: Qualitative comparison of architectural variants across five representative patients. Columns: pre-operative image, predictions from each model variant, and the ground-truth post-operative image.

region. These issues underscore the importance of careful supervision and design, as well as the advantages of decoupling realism from classification in architectural choices. These visual findings support the quantitative metrics presented in Table 8.2, where the encoder model outperforms other variants in both SSIM and colour preservation.

Table 8.2: Quantitative evaluation of architectural variants

	SSIM (post) $\uparrow$	SSIM (pre) $\uparrow$	IOU (breast) $\uparrow$	IOU (nipple) $\uparrow$	Colour 1 $\downarrow$	Colour 2 $\downarrow$
Encoder for Classification	0.78	0.86	0.79	0.08	0.01	0.01
Two Discriminators	0.78	0.80	0.78	0.09	0.04	0.03
Single Discriminator	0.78	0.84	0.79	0.08	0.03	0.01
One Encoder + Two Discriminators	0.78	0.84	0.80	0.09	0.02	0.01

8.3 Bounding Box and Masking Strategies

Even with strong architectures and losses, the way spatial supervision is applied can greatly affect output quality. This section evaluates how bounding box strategies and soft-edged masks impact anatomical alignment and realism on the top-performing model with a separate discriminator and encoder (Section 8.2.2). These refinements aim to improve local detail without compromising identity or clinical accuracy.

8.3.1 Bounding Box Strategy: Union vs. Centre Expansion

To constrain spatial losses such as SSIM, L1, and region-specific metrics (e.g., colour or segmentation loss), two strategies were evaluated for defining the bounding box over the breast region.

The **union** method computes a bounding box that encompasses both the pre-operative and post-operative regions. Specifically, it takes the minimum and maximum coordinates from both bounding boxes and defines the smallest rectangle that fully contains them. Although this ensures shared spatial supervision across timepoints, it often includes a larger portion of the image, capturing irrelevant background or pose-related variation. Because the original pre- and post-operative masks may differ in position and size, due to changes in anatomy, pose, or imaging conditions, this unified box can introduce anatomical misalignment. As illustrated in the left panels of Figures 8.9 and 8.10, this misalignment may lead to spatial drift in the predicted outputs and reduced focus on the actual region of interest.

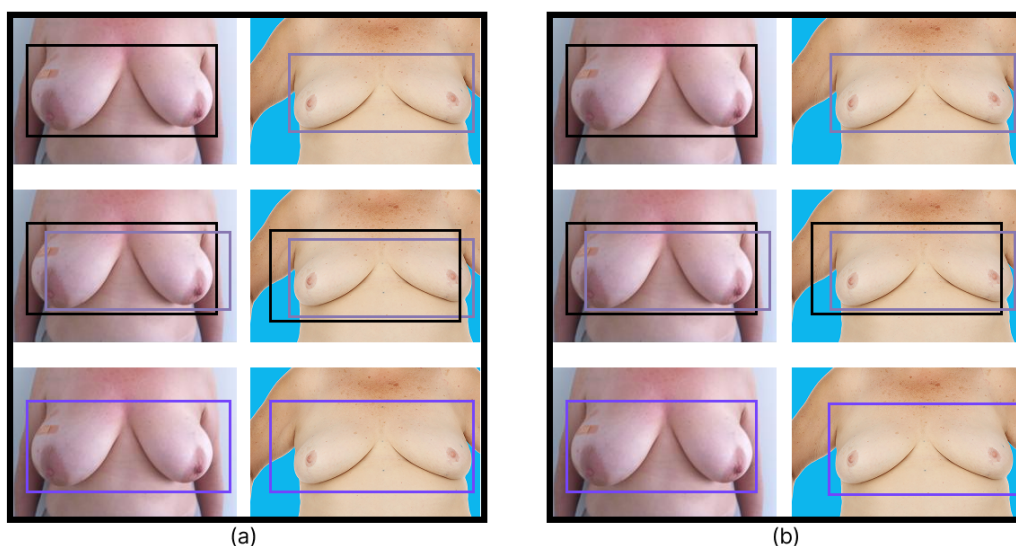


Figure 8.9: Bounding box overlays for both strategies. Each row shows a pre-operative image (left) and a post-operative image (right). In each image, the **black box** denotes the bounding box derived from the pre-operative mask, the **purple box** from the post-operative mask, and the **dark purple box** corresponds to the final region used for spatial supervision. In (a), the dark purple box is the **union** of both individual boxes and is applied equally to both images. In (b), the dark purple box is the result of an **expansion** around each mask, ensuring better anatomical centring for each domain.

Alternatively, the **expansion** strategy computes bounding boxes independently for pre- and post-operative images and then expands both boxes symmetrically around their centres so that they match in size. Specifically, the maximum width and height across both bounding boxes are used as the target dimensions. This ensures consistent spatial alignment between pre- and post-operative crops, reducing variability due to pose, zoom, or breast deformation. The right panels of Figure 8.10 and Figure 8.9 show improved alignment and centring, particularly in the nipple and breast contour regions.

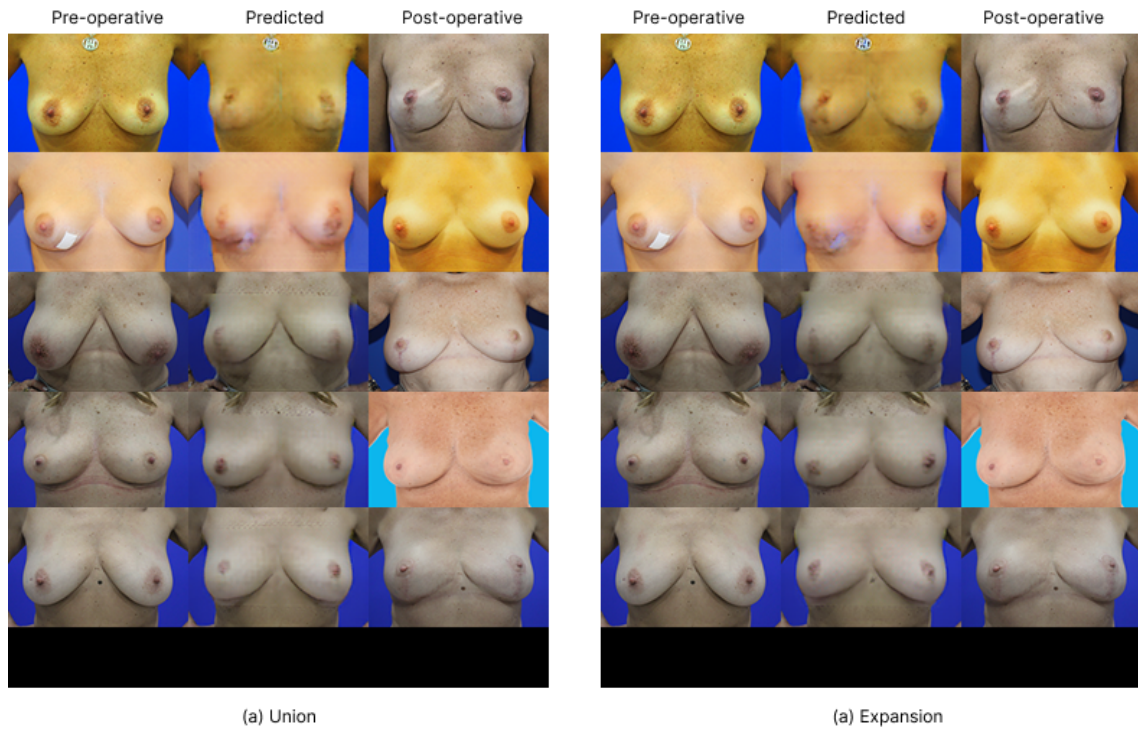


Figure 8.10: Qualitative comparison of bounding box strategies. Left: Union strategy includes excess background and induces spatial misalignment. Right: The expansion method centres on the region of interest and aligns anatomical features more accurately.

Quantitative results presented in Table 8.3 further support the advantage of the Expansion strategy. Both methods yield similar SSIM and colour consistency scores, but expansion results in an increase in nipple IoU (from 0.08 to 0.10). Despite the small change, the improvement in nipple segmentation is notable given the region's small size and high sensitivity to misalignment.

Table 8.3: Performance comparison of bounding box strategies. The Expansion method provides improved anatomical alignment and slightly better segmentation metrics.

	SSIM (post) ↑	SSIM (pre) ↑	IOU (breast) ↑	IOU (nipple) ↑	Colour 1 ↓	Colour 2 ↓
Union	0.78	0.86	0.79	0.08	0.01	0.01
Expansion	0.79	0.87	0.79	0.10	0.02	0.01

8.3.2 Soft-edged Breast Mask for Supervision Loss

In early experiments, applying structural losses (L1 and SSIM) strictly within the hard-edged bounding box of the breast region often led to visible artefacts along the mask boundaries. These artefacts emerged due to the abrupt transition between strongly supervised regions (inside the mask) and unsupervised areas (outside), producing visually noticeable edges in the generated outputs.

To address this issue, a soft-edged mask was introduced in the encoder-based model (Section 8.2.2) using a bounding box expansion strategy (Section 8.3.1). This mask employs a custom distance-based falloff function rather than a binary boundary. Specifically, a radial falloff is calculated from each pixel to the edge of the bounding box, and the loss is gradually reduced over a defined margin (for instance, 15 pixels). Two falloff modes were explored: one with a linear decay and another with a smoother, cosine-based decay, which gently diminishes the influence of the loss near the boundary. As a result, the mask has a value of one at the centre of the region and gradually tapers to zero near the boundary, helping to avoid abrupt supervision discontinuities (as shown in panel (c) of Figure 8.11).

As shown in the Figure 8.11, this soft-masked supervision produces outputs with more natural transitions around the breast contour. Compared to the hard-edged mask (a), the soft-edged version (b) leads to smoother blending into the surrounding anatomy and fewer visible edge artefacts.



Figure 8.11: Comparison of models trained with hard-edged (a) and soft-edged (b) supervision masks. Soft-edged masks reduce visible border artefacts and lead to smoother transitions between the breast region and surrounding anatomy. Panel (c) illustrates the effect of the soft spatial falloff applied to the mask, transitioning gradually from full supervision (centre) to none (edges).

This soft spatial weighting improves the quality of the predicted post-operative region while maintaining focused supervision in anatomically relevant areas.

8.3.3 Effect of Removing Segmentation Loss

To better understand the impact of segmentation supervision, the encoder-based architecture with the soft-edged mask (Section 8.3.2) was trained twice: once with the segmentation loss $\mathcal{L}_{\text{seg}}$ included, and once without it. All other loss components, architectural choices, and training conditions were held constant. The comparison results are shown in Table 8.4.

Table 8.4: Comparison of encoder model performance with and without segmentation loss.

	SSIM (post) ↑	SSIM (pre) ↑	IOU (breast) ↑	IOU (nipple) ↑	Colour 1 ↓	Colour 2 ↓
Encoder for Classification	0.78	0.89	0.80	0.10	0.005	0.003
Encoder for Classification Without Segmentation Loss	0.77	0.86	0.77	0.05	0.02	0.01

Removing the segmentation loss led to a noticeable drop in anatomical accuracy. The nipple IoU fell from 0.10 to 0.05, and the breast IoU dropped from 0.80 to 0.77, indicating that the generator struggled to reconstruct key anatomical regions without segmentation guidance. SSIM scores for the pre-operative image comparison also decreased from 0.89 to 0.86. While colour metrics remained relatively high in both cases, the loss of spatial precision highlights the importance of segmentation as a targeted supervisory signal for guiding anatomically faithful generation. Visual results (Figure 8.12) also confirm that the version without segmentation loss produced blurrier breast contours and less defined nipples, with overall transformations appearing more faded and less realistic.

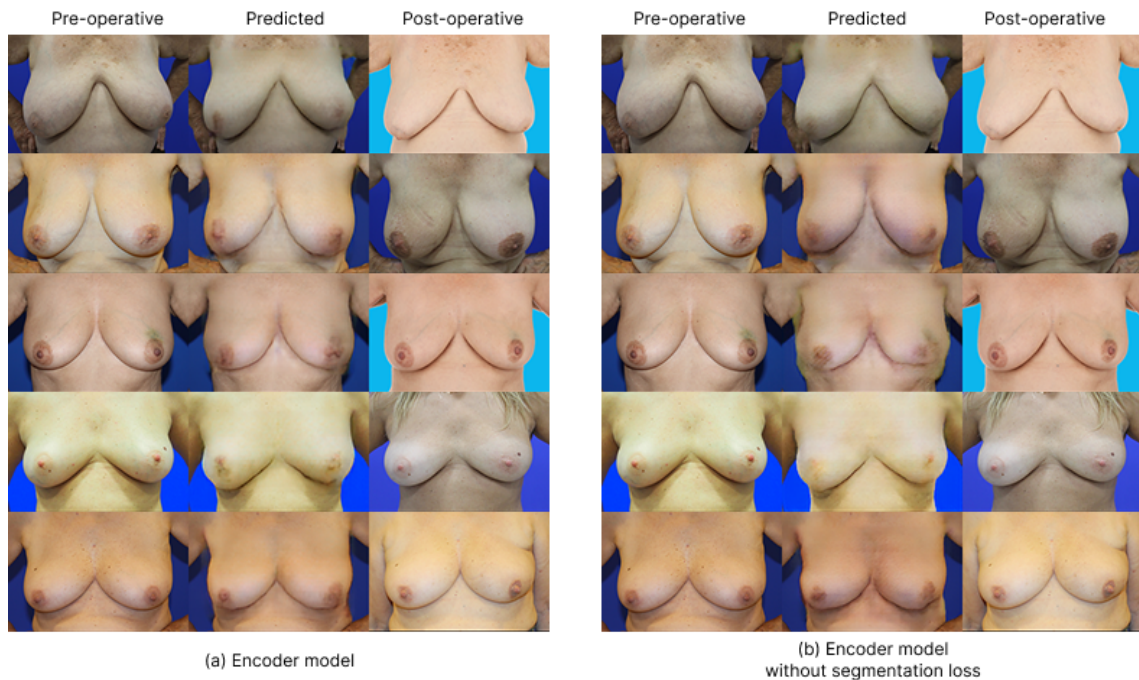


Figure 8.12: Visual comparison of the encoder model with (a) and without (b) segmentation loss. Both columns show pre-operative, predicted, and ground-truth post-operative images. While overall colour and body shape are preserved in both cases, the version without segmentation loss exhibits reduced anatomical precision, especially in the nipple and breast contours.

8.3.4 Results

Spatial supervision strategies have a significant impact on the anatomical accuracy and visual realism of generated post-operative images. The use of centre-based bounding box expansion improved the alignment of the nipple region and reduced spatial drift compared to the union strategy, all while maintaining colour consistency. Furthermore, the application of soft-edged masks in the loss computation effectively eliminated harsh boundary artefacts, resulting in smoother transitions between the supervised breast region and the surrounding tissue.

To further evaluate the role of spatial supervision, an ablation study was conducted to assess the effects of removing the segmentation loss. Although overall SSIM and colour metrics remained relatively stable, the absence of this targeted signal resulted in a noticeable drop in anatomical accuracy, particularly in the nipple area and breast IoU scores. This confirms that segmentation loss is crucial for promoting anatomically accurate reconstructions, especially in clinically sensitive and structurally complex regions.

8.4 Discussion

The experiments in this chapter demonstrate that separating classification and adversarial tasks into dedicated modules leads to more stable training and higher-quality outputs. In particular, models that attempt to unify classification and realism (e.g., the two-discriminator and single-discriminator variants) often suffer from colour bias, oversmoothing, and unrealistic textures.

The encoder-based architecture stands out for three main reasons:

- It preserves patient-specific colour tones, avoiding the typical yellow or grey shifts caused by post-op dataset priors.
- It generalises better across different body types and surgery types, producing diverse but accurate outcomes.
- It avoids overfitting to dataset biases by grounding generation in pre-operative appearance, clinical metadata, and supervised anatomical constraints.

Further refinements to the supervision mask, such as replacing hard-edged boxes with soft-edged falloff masks, helped mitigate visible boundary artefacts in generated images.

The best overall results were achieved by combining the encoder-based architecture with centre-aligned bounding box expansion and soft-edged masks, which together produced the most anatomically realistic and colour-consistent predictions.

Despite these improvements, the generated images still lack the desired level of clarity and natural appearance. Subtle artefacts, such as local blurriness and under-detailed textures, detract from the overall realism, especially in clinically sensitive areas like the nipple or surgical scars.

Chapter 9

Refinements in Generator Supervision and Training

After identifying the most effective architecture for generating clinically accurate post-operative images, which combines an encoder for semantic supervision (Section 8.2.2), soft-edged masking (Section 8.3.2), and a center-aligned bounding box expansion strategy (Section 8.3.1), this chapter explores a series of refinements aimed at enhancing visual quality, anatomical precision, and the model’s generalizability.

While the previous chapter focused on architectural configurations and loss formulations, the strategies presented here build on the selected base model to address remaining challenges such as regional blurriness, visual artefacts, and performance on underrepresented surgical cases. In particular, this chapter investigates how spatial supervision methods, generator connectivity (e.g., skip connections), and architectural variants (e.g., StarGAN and BicycleGAN) influence the final output quality.

9.1 Impact of Generator Connectivity and Training Supervision

Building on the encoder-based architecture (Section 8.3.2), this section explores how skip connections and the level of training supervision influence the model’s capacity to perform meaningful transformations. The goal is to shed light on how architectural and training constraints interact to influence the final image quality by comparing models with and without skip connections in both supervised and unsupervised environments.

9.1.1 With and Without Skip Connections in Generator

This experiment evaluated the importance of skip connections in the generator. By removing skip connections from the architecture, the model’s ability to directly copy low-level features (e.g. background textures and patient-specific colour distributions and body shape) was reduced, forcing the model to rely more on learned transformations.

Figure 9.1 shows that removing skip connections led to outputs that tended to be more anatomically realistic. The anatomical shapes of the breast were better defined, and the nipples and scars were more visible. This likely happens because, without skip connections, the generator cannot directly reuse input features and must rely entirely on encoded representations. This pushes the model to focus on semantic structure and learn the transformation more explicitly, resulting in clearer anatomical changes.

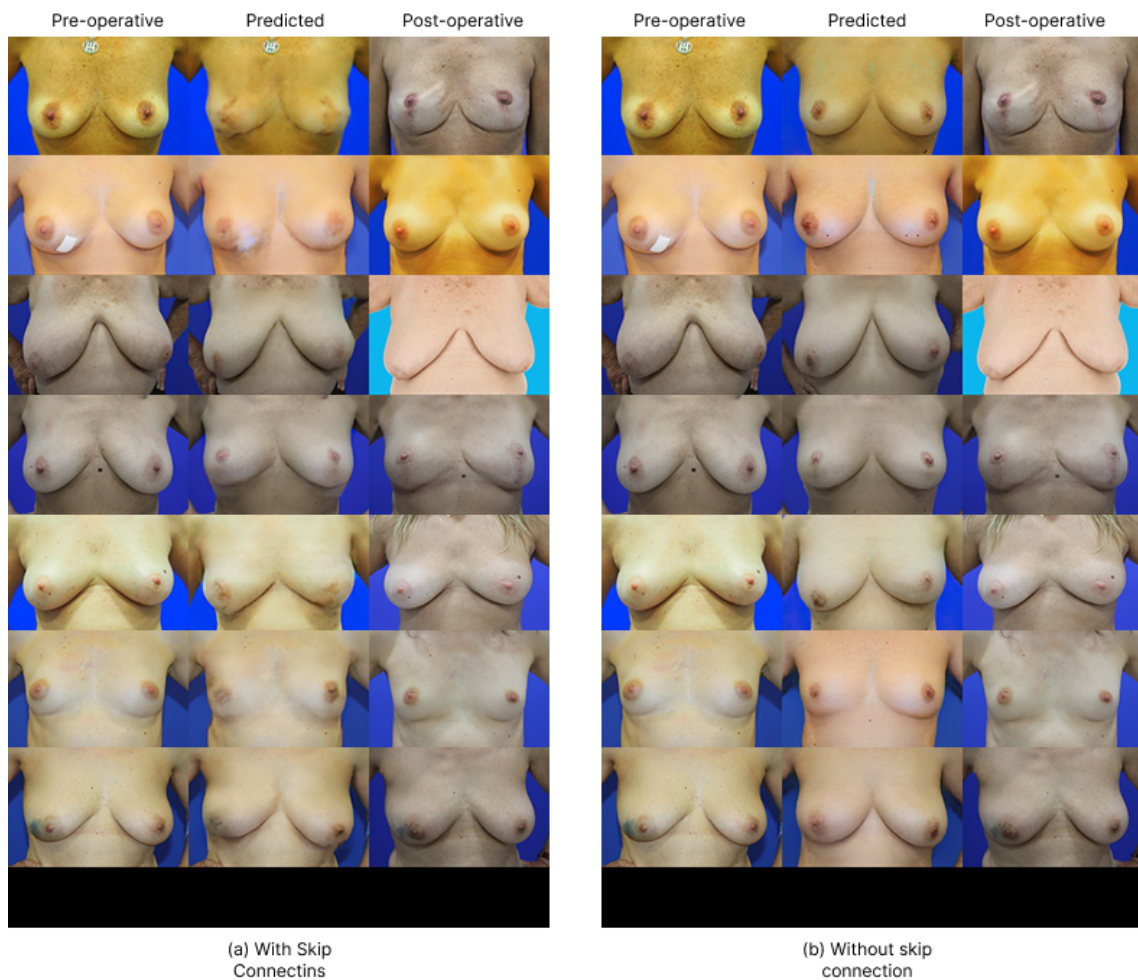


Figure 9.1: Comparison of models trained with (left) and without (right) skip connections. Each row shows a pre-operative input, the predicted image, and the ground truth post-operative result. Removing skip connections improves structural realism by forcing the model to rely on encoded representations, but also leads to increased colour variability. Models with skip connections preserve colour more accurately but often produce blurrier and less defined anatomical changes.

Skip connections pass low-level features such as skin tone, edges, and textures from the encoder to the decoder. While this helps preserve aspects of the input's colour, lighting, and body shape, it may also make it harder for the model to decide which features should be retained versus altered. As a result, the model that used skip connections was better at maintaining colour and body fidelity, but produced outputs that were more blurry and less confident shape transformations.

To further examine this effect, Figure 9.2 presents examples from models *without* skip connections, trained under different combinations of loss weights: λ_{sim} (structural loss: L1 + SSIM), λ_{colour} (colour consistency), and λ_{seg} (segmentation alignment). Varying these weights had clear visual consequences:

- Higher values of λ_{colour} led to better preservation of skin tone and overall colour consistency.
- Increasing λ_{seg} helped maintain anatomical accuracy but sometimes introduced smoother or flatter textures in the breast region.

In the no-skip model without the supervision loss, these issues were more pronounced. While it produced structural changes, the lack of supervision led to generic outputs, such as nipples that looked the same across patients, and lower overall realism.

These results underscore a central trade-off in the generator design: skip connections help retain appearance features from the input, while removing them encourages clearer, learned transformations. The interaction between skip connections and loss balancing is, therefore, a critical design choice for achieving the desired clinical realism.

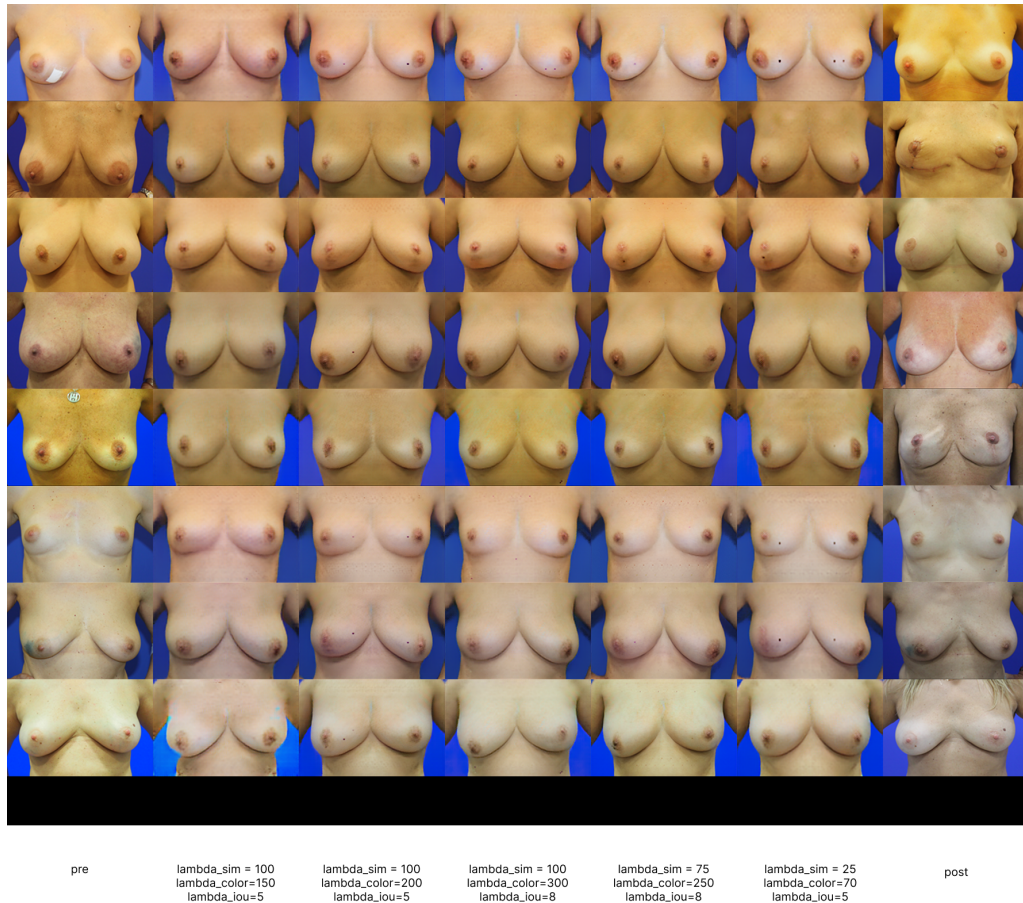


Figure 9.2: Effect of tuning loss weights on models without skip connections. Columns vary λ_{sim} , λ_{colour} , and λ_{seg} . Rows show different patients. Models with higher colour weights preserve tone better, while lower structural weights allow more diverse transformations.

9.1.2 Continue Training Without the Supervision Loss

In this experiment, the encoder-based model from Section 8.3.2 was further trained without the supervision loss ($L1 + SSIM$) computed between the generated and ground-truth post-operative images. The motivation was to enhance visual quality in the breast region by reducing blurriness and improving fine detail. Although direct pixel-level supervision was removed, segmentation loss was still retained, providing indirect anatomical guidance. To preserve the learned semantic conditioning, the encoder component was frozen, and only the decoder continued updating. This setup allowed the generator to refine its outputs through self-supervision while still being constrained by segmentation accuracy.

As shown in Figure 9.3, the results were mixed. While the continuation led to slightly sharper textures in some cases, it also introduced undesirable changes, such as shading artefacts and unnatural contours. Without pixel-level guidance, the model exhibited signs of semantic drift, reinforcing visual biases instead of faithfully reconstructing realistic transformations. Although segmentation helped preserve general structure, the absence of direct supervision limited control over visual fidelity. These findings highlight the importance of combining pixel-wise, structural and anatomical supervision for consistent, high-quality outcomes.

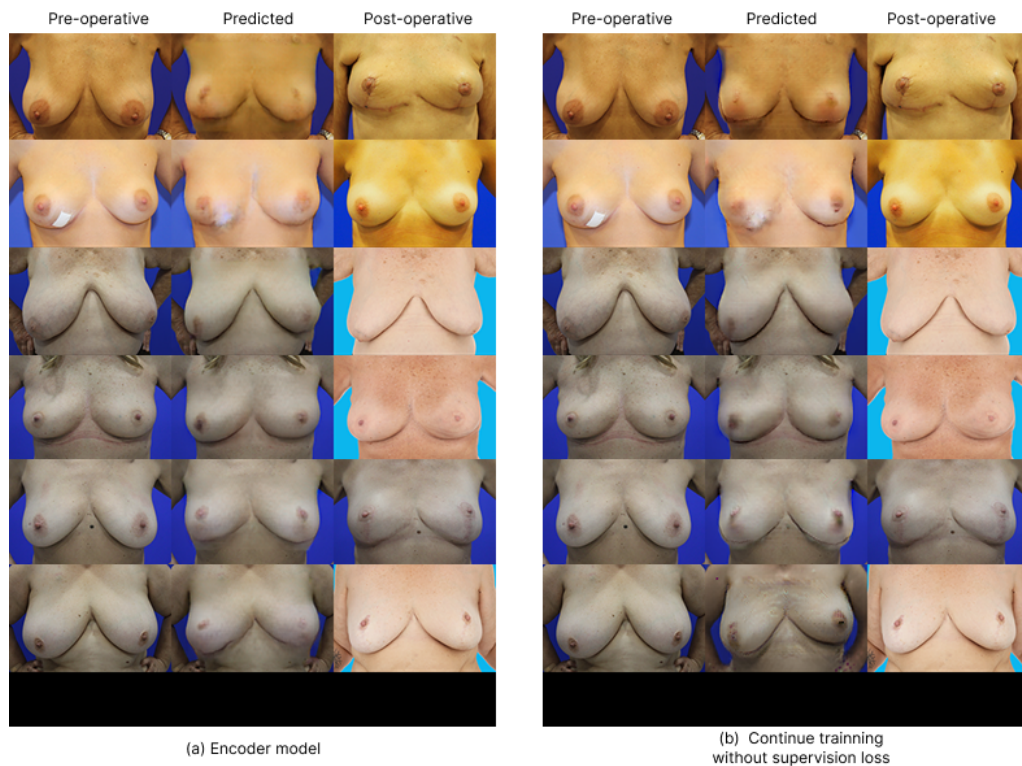


Figure 9.3: Comparison between the supervised encoder model (a) and the same model after continued training without structural supervision (b). While some outputs appear more detailed, others show semantic drift due to the absence of pixel-level loss.

9.1.3 Training Without Supervision Loss from Scratch

Finally, a model was trained entirely without using direct supervision losses (L1 or SSIM) between the generated and post-operative images. The training objectives included adversarial loss, colour consistency loss (calculated with respect to the pre-operative image), and segmentation loss.

As shown in Figure 9.4, the model with skip connections (left) produced outputs nearly identical to the pre-operative input, indicating that it had mostly learned to copy the image rather than perform any real transformation. In several cases, the model appeared to "trick" the segmentation network by drawing faint horizontal or curved lines across the chest, mimicking a new breast contour without actually changing the anatomy. This visual illusion allowed it to score well on the segmentation-based loss while leaving the original breast shape untouched. Without proper supervision, the model prioritised deceiving the loss functions instead of learning true post-operative transformations.

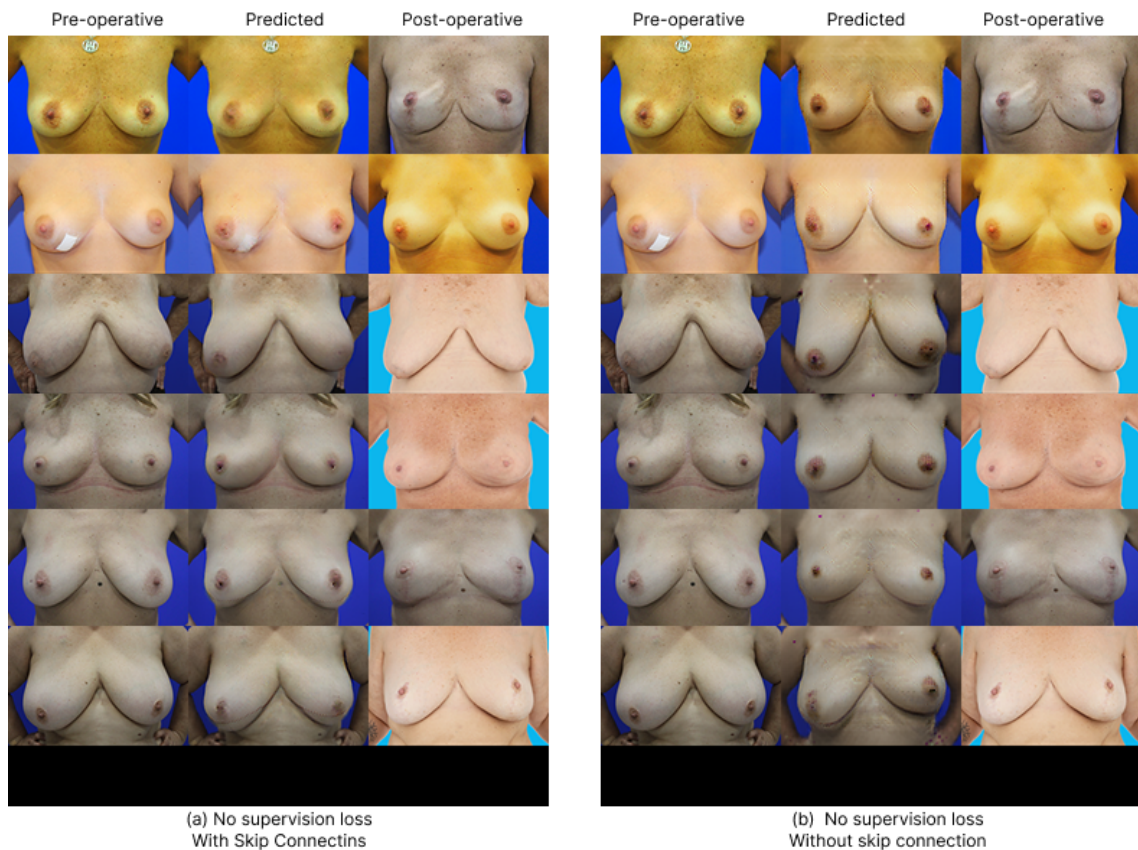


Figure 9.4: Comparison of no supervision loss models trained with (left) and without (right) skip connections.

In contrast, the version without skip connections (right) was forced to rely on encoded features, as it no longer had direct access to the input's low-level details. As a result, it began making more visible structural changes in the breast area, attempting to simulate post-operative outcomes. However, the lack of direct supervision led to some artefacts, and some anatomical details, such as

nipples that had a uniform colour across all patients. This shows that while removing skip connections encourages transformation, without strong appearance-level guidance, it also introduces new challenges in texture realism and identity preservation. This highlights the importance of combining semantic conditioning with both spatial and structural supervision for reliable post-operative prediction.

9.1.4 Results

As shown in Table 9.1 and Figure 9.5, the model with skip connections produced anatomically realistic results while preserving patient-specific features like skin tone and body shape. However, fine details such as nipples and scars appeared slightly blurred, likely due to the model struggling to balance which features to transform versus retain.

Table 9.1: Quantitative evaluation of BicycleGAN variants. The encoder-based version achieves higher SSIM and colour consistency. Excessively strong KL and latent regression constraints lead to instability in visual results.

	SSIM (post) $\uparrow$	SSIM (pre) $\uparrow$	IOU (breast) $\uparrow$	IOU (nipple) $\uparrow$	Colour 1 $\downarrow$	Colour 2 $\downarrow$
Encoder for Classification with skip connections	0.78	0.89	0.80	0.10	0.005	0.003
Encoder for Classification without skip connections	0.77	0.81	0.79	0.09	0.02	0.02
Encoder for Classification continue training	0.76	0.92	0.78	0.05	0.002	0.004
Encoder for Classification without skip connections no supervision loss	0.75	0.79	0.79	0.08	0.02	0.02

The model without skip connections generated sharper and more detailed outputs, with clearer anatomical changes. By relying solely on encoded features, it avoided copying input textures, resulting in more distinct transformations. However, it struggled slightly with patient identity, sometimes producing less personalised features.

In the no-skip model without the supervision loss, these issues were more pronounced. While it produced structural changes, the lack of supervision led to generic outputs, such as nipples that looked the same across patients, and lower overall realism. Additionally, the model that continued training without the supervision loss did not improve performance. Despite using a frozen encoder, its outputs increasingly resembled those of the model trained from scratch without supervision using skip connections. Notably, its SSIM (pre) score rose to 0.92, indicating excessive similarity to the pre-operative input. As shown in Figure 9.5, the model appeared to rely on copying rather than generating realistic transformations, occasionally drawing artificial lines to deceive the segmentation loss.

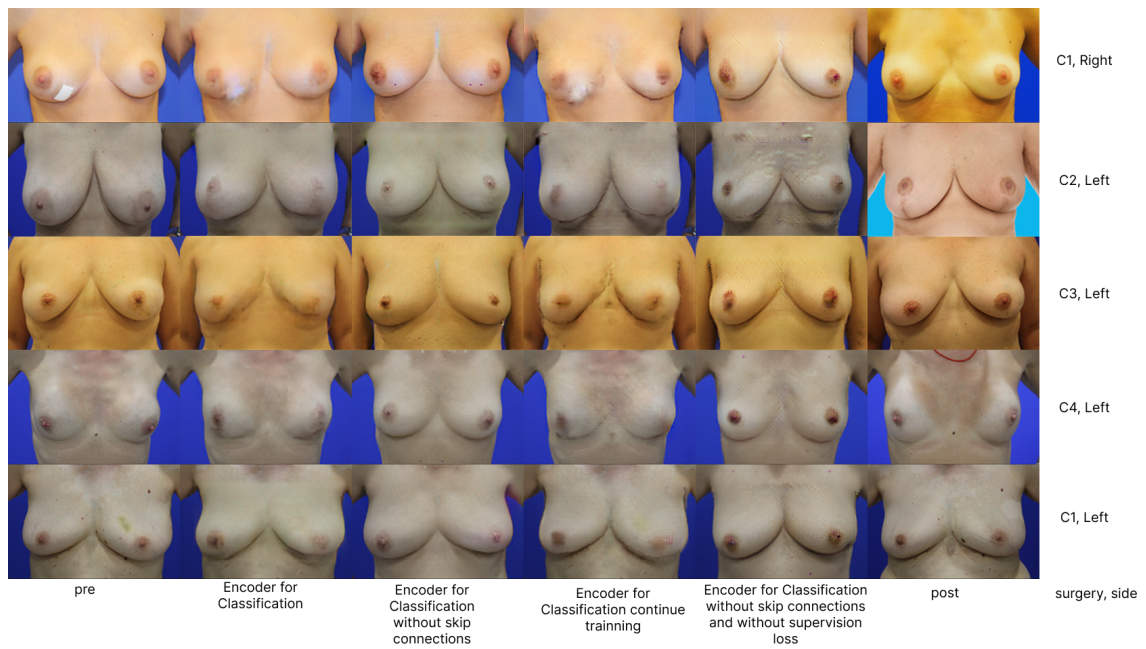


Figure 9.5: Qualitative comparison of architectural variants across five representative patients. Columns: pre-operative image, predictions from each model variant, and the ground-truth post-operative image.

These results highlight a trade-off: skip connections support identity preservation but soften transformed details, while removing them improves structural clarity at the cost of personalisation.

9.2 StarGAN and BicycleGAN with the Encoder for Classification

This section investigates how well StarGAN and BicycleGAN perform when adapted to use the same encoder-based classification strategy described in Section 8.3.2. By integrating these models into the unified loss and classification framework, the aim is to assess whether their architectural advantages (e.g., multi-domain flexibility or latent diversity) can complement the supervised, patient-specific prediction task.

9.2.1 StarGAN

This model extends the original StarGAN architecture introduced in Section 6.3, which was designed for multi-domain image translation using a single adversarial and classification discriminator. In its baseline form, StarGAN shared the classification and realism objectives within the same discriminator, often leading to convergence instability and unwanted shifts in skin tone or lighting.

To address these issues, this modified version introduces a cleaner architectural decomposition and additional supervision losses based on prior findings from the Pix2Pix. The architecture is the same as the encoder model without the skip connection, with the added reconstruction loss for

cycle-consistency losses:

$$\mathcal{L}_G = \lambda_{\text{GAN}} \mathcal{L}_{\text{GAN}} + \lambda_{\text{L1}} \mathcal{L}_{\text{struct}} + \lambda_{\text{colour}} (\mathcal{L}_{\text{colour1}} + \mathcal{L}_{\text{colour2}}) + \lambda_{\text{seg}} \mathcal{L}_{\text{seg}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} \quad (9.1)$$

9.2.1.1 Results

Figure 9.6 shows a qualitative comparison across three variants of StarGAN: the baseline model, the new version with the encoder, and a further variant with a stronger reconstruction (cycle-consistency) loss. These are compared against the original pre-operative and real post-operative images.

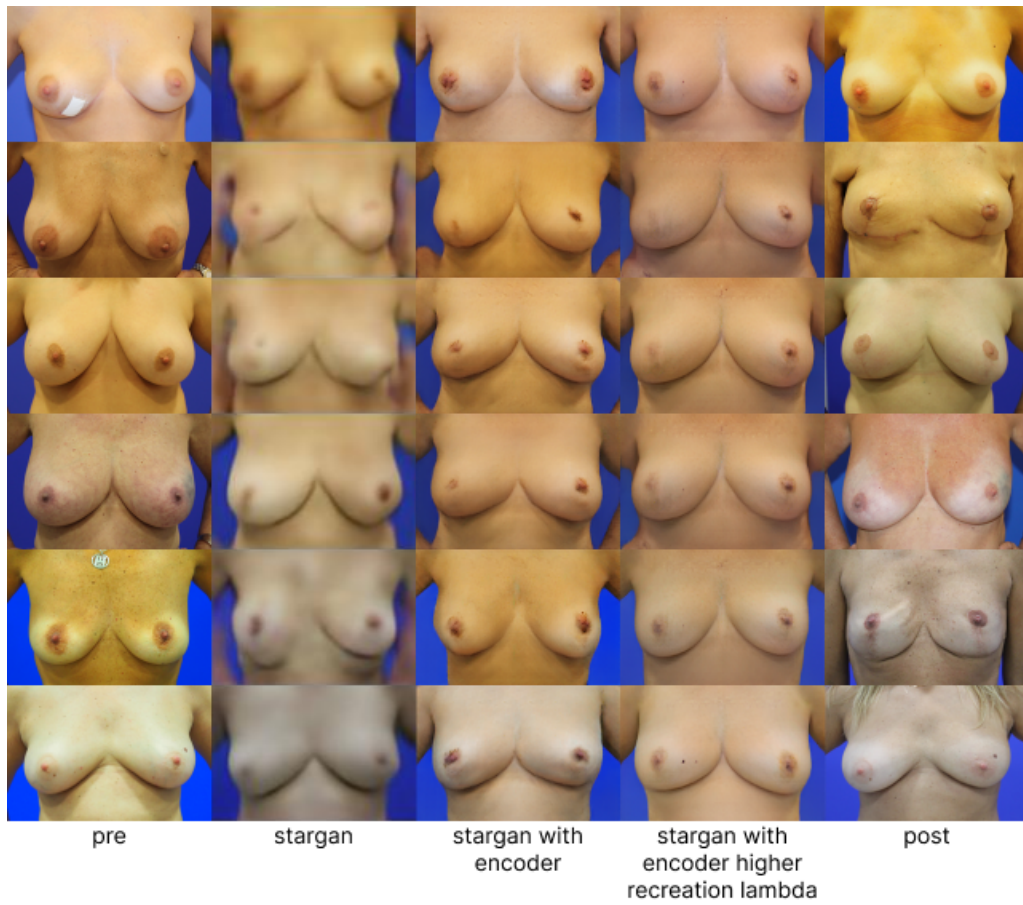


Figure 9.6: Qualitative results of StarGAN variants. From left to right: pre-operative image, StarGAN baseline, StarGAN with encoder, StarGAN with encoder and high reconstruction weight, and ground truth post-operative image.

Introducing the encoder slightly improves semantic consistency, as seen in better colour preservation and more accurate nipple structure. However, increasing the weight of the reconstruction loss excessively causes a collapse in diversity. The model begins to ignore individual patient features and converges toward a narrow set of generic-looking outputs, reducing clinical utility.

This trade-off is also reflected quantitatively in Table 9.2, where the strong reconstruction variant shows a slight drop in SSIM and nipple IoU, despite high colour alignment. The results suggest that moderate reconstruction loss may be helpful, but excessive weight reduces the specificity of the generated outcome.

Table 9.2

	SSIM (post) $\uparrow$	SSIM (pre) $\uparrow$	IOU (breast) $\uparrow$	IOU (nipple) $\uparrow$	Colour 1 $\downarrow$	Colour 2 $\downarrow$
StarGAN	0.76	0.77	0.75	0.05	0.06	0.06
StarGAN with Encoder	0.76	0.76	0.78	0.08	0.03	0.03
StarGAN with Encoder $\mathcal{L}_{\text{rec}}$	0.75	0.72	0.76	0.06	0.05	0.05

9.2.2 BicycleGAN

This architecture builds upon the previously discussed BicycleGAN with conditioning (Section 6.2), maintaining the use of latent vector sampling and clinical metadata to support diverse and guided post-operative image generation.

In the original conditioned BicycleGAN, classification of surgery type and cancer side was handled by auxiliary heads in the discriminator. This approach often biases the model toward post-operative dataset priors (e.g., lighting and skin tone). In this improved version, following the insights from the Pix2Pix experiments, classification is handled by a dedicated encoder network, which receives the generated image and independently predicts the clinical labels.

The architecture is now composed of the following four core components:

- **Discriminator D:** This discriminator trains on pre-operative and post-operative images (either real or generated). It outputs a real/fake prediction and provides an adversarial signal for the generator. D focuses on assessing the global realism and coherence of the full-body image, and is used to supervise the cVAE-GAN branch (which uses encoded latent vectors).
- **Discriminator D2:** D2 has the same structure as D but is used to supervise the cLR-GAN branch, where the generator is conditioned on randomly sampled latent vectors. Its role is to ensure the generator maintains high realism and structural integrity under stochastic sampling, supporting output diversity and regularisation of the latent space.
- **Encoder for Classification:** The encoder receives pairs of real and generated post-operative images and outputs surgery type and cancer side predictions. It is responsible for enforcing semantic correctness. The encoder's classification loss is shared with the generator to ensure that generated outputs match the intended clinical metadata.
- **Generator (G):** The generator receives a pre-operative image, a latent vector (sampled or encoded), and a one-hot encoded clinical condition vector. These inputs are concatenated and passed through an AE generator without the skip connections. The generator is trained

with a combination of adversarial, structural, colour, segmentation, and classification losses:

$$\begin{aligned} \mathcal{L}_G = & \lambda_{\text{GAN}}(\mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{GAN}2}) + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_z\mathcal{L}_z + \lambda_{\text{L1}}\mathcal{L}_{\text{struct}} \\ & + \lambda_{\text{colour}}(\mathcal{L}_{\text{colour1}} + \mathcal{L}_{\text{colour2}}) + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} \end{aligned} \quad (9.2)$$

9.2.2.1 Results

Figure 9.7 presents qualitative results comparing the standard BicycleGAN with conditioning, a variant with the classification encoder, and a third version where the KL and latent regression loss terms were given increased weights.

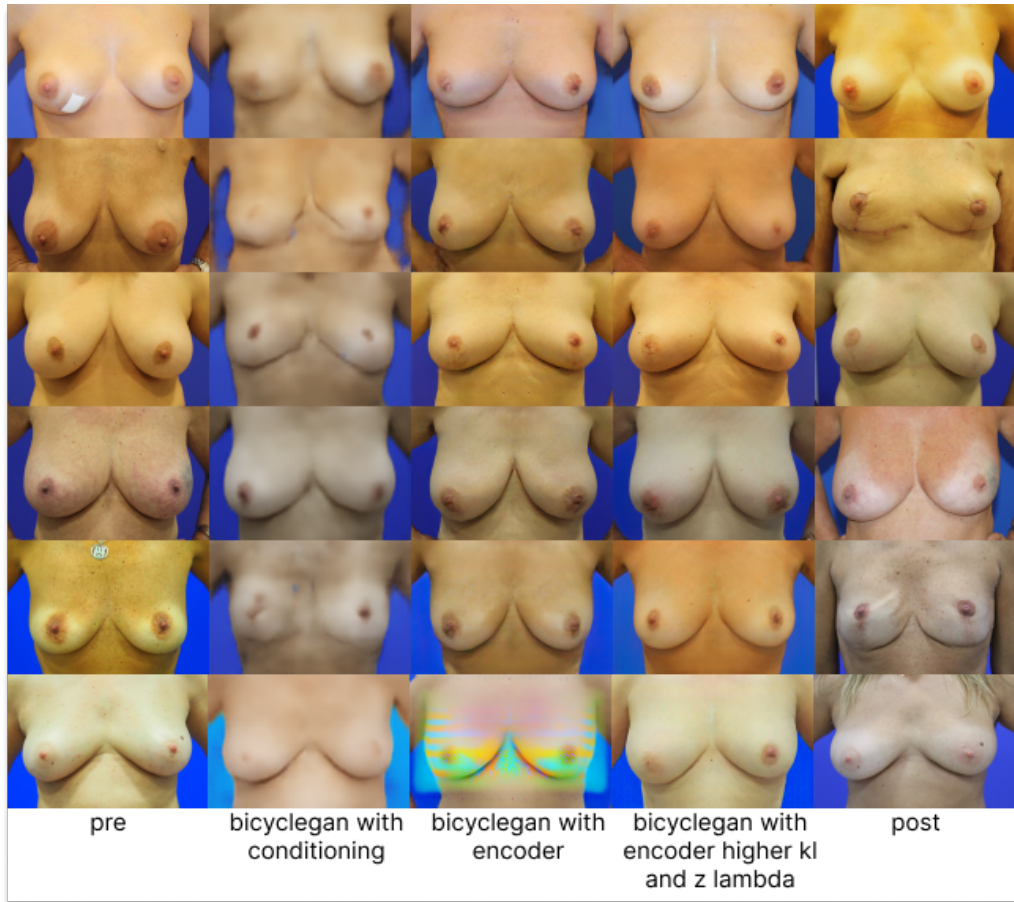


Figure 9.7: Qualitative comparison of BicycleGAN variants. From left to right: pre-operative image, BicycleGAN with conditioning, BicycleGAN with encoder, BicycleGAN with encoder and stronger KL and latent regression loss, and the ground truth post-operative image.

Adding the encoder improves semantic alignment and structural realism, particularly in skin tone and nipple prediction. This is corroborated by increased SSIM and colour metrics in Table 9.3. However, the variant with higher latent consistency losses ($\mathcal{L}_{\text{KL}}, \mathcal{L}_z$) exhibits visual artefacts, such as random-coloured boxes or blurs, especially in the breast region. These are likely

caused by over-constraining the latent space, which limits the generator’s capacity to adjust to spatially variant patient-specific details.

Table 9.3: Quantitative evaluation of BicycleGAN variants. The encoder-based version achieves higher SSIM and colour consistency. Excessively strong KL and latent regression constraints lead to instability in visual results.

	SSIM (post) $\uparrow$	SSIM (pre) $\uparrow$	IOU (breast) $\uparrow$	IOU (nipple) $\uparrow$	Colour 1 $\downarrow$	Colour 2 $\downarrow$
BicycleGAN with conditions	0.77	0.77	0.79	0.08	0.07	0.06
BicycleGAN with Encoder	0.76	0.76	0.78	0.09	0.03	0.03
BicycleGAN with Encoder higher $\mathcal{L}_{KL}$ and $\mathcal{L}_z$	0.75	0.79	0.78	0.08	0.02	0.02

9.3 Discussion

Despite integrating the encoder and supervision strategies into more complex architectures like StarGAN and BicycleGAN, these models did not outperform the simpler encoder-based baseline. As shown in Table 9.4 and Figure 9.8, both alternatives maintained acceptable structural and colour consistency, but introduced softness and instability in finer anatomical details. StarGAN, in particular, began generating repetitive features across patients as the reconstruction loss increased, reducing diversity and realism. BicycleGAN performed better overall, but introduced artefacts, like boxes with unnatural colours.

A clear trade-off also emerged between models with and without skip connections. While skip connections helped preserve patient identity and colour fidelity, they often led to blurrier outputs, especially in regions like nipples and scars. Removing skip connections encouraged sharper, more realistic anatomical transformations, though with slightly reduced texture consistency and personalisation.

These results suggest that increased architectural complexity does not necessarily translate to better outcomes. The most reliable and clinically realistic results came from the simplest supervised model, where architectural choices and loss objectives were tightly aligned. Overcomplicating the network often led to less detailed predictions.

Despite the strong overall results, breaking performance down by surgery type reveals important limitations. While quantitative results (Table 9.5) suggested that the model generalised well across surgery types, including less common procedures like C3 and C4. However, qualitative inspection of the outputs in Figure 9.9 paints a more nuanced picture. Despite being fed different surgery types and cancer sides, the predicted images for C3 and C4 appear visually similar to those of C2 (conservative surgery with bilateral reduction). The changes introduced are minimal, and the resulting images often lack the distinctive characteristics expected from those procedures.

This behaviour is likely due to a combination of two factors: (1) limited training data for C3 and C4 surgeries, and (2) the observation that, in the test set, C3 and C4 cases are often associated

Table 9.4: Quantitative evaluation of BicycleGAN variants. The encoder-based version achieves higher SSIM and colour consistency. Excessively strong KL and latent regression constraints lead to instability in visual results.

	SSIM (post) $\uparrow$	SSIM (pre) $\uparrow$	IOU (breast) $\uparrow$	IOU (nipple) $\uparrow$	Colour 1 $\downarrow$	Colour 2 $\downarrow$
Encoder for Classification with skip connections	0.78	0.89	0.80	0.10	0.005	0.003
Encoder for Classification without skip connections	0.77	0.81	0.79	0.09	0.02	0.02
BicycleGAN with Encoder	0.76	0.76	0.78	0.09	0.03	0.03
StarGAN with Encoder	0.76	0.76	0.78	0.08	0.03	0.03

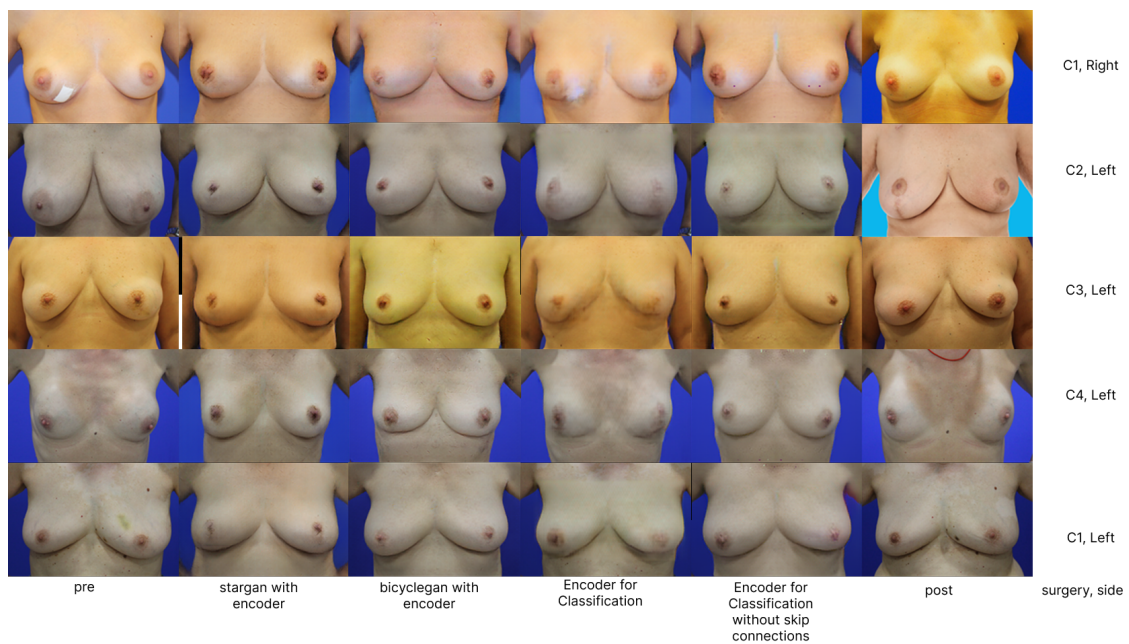


Figure 9.8: Qualitative comparison of architectural variants across five representative patients. Columns: pre-operative image, predictions from each model variant, and the ground-truth post-operative image.

with patients who have smaller breasts. As a result, when these surgery labels are applied to pre-operative images of patients with larger breast volumes, it may default not only to transformation patterns learned from more frequent surgeries like C2, but also to generating smaller post-operative breast shapes, implicitly replicating anatomical traits it has seen during training.

One possible solution would be to assign a higher weight to the classification loss for under-represented surgery types during training. This might encourage the generator to better model the distinctive characteristics of C3 and C4 procedures; however, it risks introducing instability: by over-emphasising rarely seen patterns, the model may amplify noise or produce unrealistic transformations in an attempt to satisfy the loss signals.

	Surgery, Count	SSIM (post) ↑	SSIM (pre) ↑	IOU (breast) ↑	IOU (nipple) ↑	Colour 1 ↓	Colour 2 ↓
Encoder with skip connections	C1, 75	0.79	0.90	0.79	0.12	0.005	0.003
	C2, 24	0.74	0.87	0.80	0.03	0.006	0.002
	C3, 4	0.79	0.87	0.79	0.01	0.006	0.003
	C4, 4	0.82	0.89	0.78	0.13	0.005	0.003
Encoder without skip connections	C1, 75	0.78	0.81	0.79	0.09	0.02	0.02
	C2, 24	0.74	0.80	0.79	0.04	0.01	0.02
	C3, 4	0.81	0.78	0.80	0.16	0.02	0.02
	C4, 4	0.82	0.80	0.77	0.08	0.02	0.02

Table 9.5: Per-surgery breakdown of results for the two encoder variants.

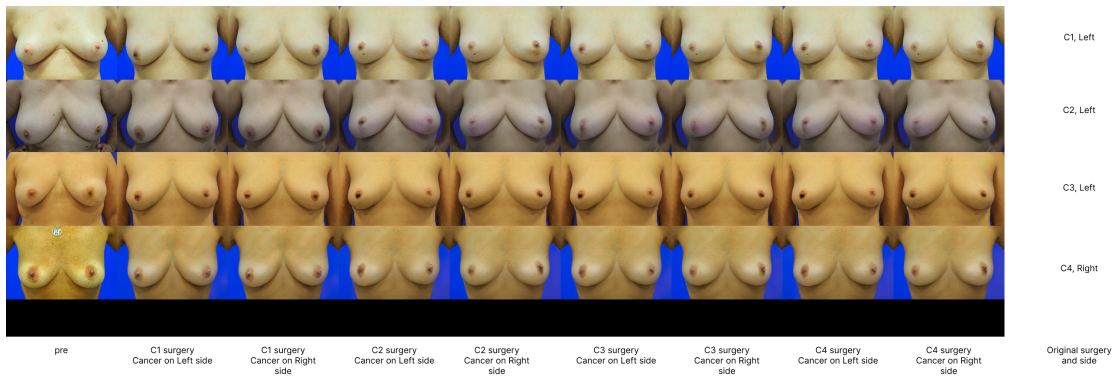


Figure 9.9: Predictions, using the encoder model without skip connection, for different surgery types and cancer sides using the same pre-operative image.

9.3.1 Testing with Mastectomies

Now that there is a stable model using conservative surgery data, the mastectomy cases were reintroduced into both the training and testing pipeline to evaluate generalisability. Mastectomies represent a wide range of surgical procedures, as summarised in Table 9.6. However, each individual subtype had very limited representation in the dataset, typically under 60 cases, which made it impractical to model them separately. As a result, all mastectomy types were grouped together and treated as a single class during training and evaluation.

Despite this simplification, the model produced surprisingly good structural metrics on mastectomy cases, achieving comparable SSIM and colour scores to conservative surgeries, as shown in Table 9.7. However, qualitative inspection (Figure 9.10) reveals a notable limitation: the predicted images fail to exhibit characteristic surgical details such as scarring, tissue reconstruction, or asymmetries typically associated with mastectomies.

This issue likely stems from the combination of high intra-class variation and low data volume. Grouping all mastectomy types under a single label may have diluted the model’s ability to learn distinct visual outcomes, especially since implant-based and flap-based reconstructions produce

Code	Description	Count
M5	Mastectomy with unilateral reconstruction with implant and contralateral symmetrisation with implant (augmentation)	58
M3	Mastectomy with bilateral reconstruction with implants	51
M1	Mastectomy with unilateral reconstruction with implant	40
M6	Mastectomy with unilateral reconstruction with implant and contralateral symmetrisation with reduction	31
M2	Mastectomy with unilateral reconstruction with autologous flap	25
M7	Mastectomy with unilateral reconstruction with autologous flap and contralateral symmetrisation with reduction	22
M4	Mastectomy with bilateral reconstruction with autologous flaps	4

Table 9.6: Types of mastectomies and their frequencies in the dataset.

	Surgery, Count	SSIM (post) ↑	SSIM (pre) ↑	IoU (breast) ↑	IoU (nipple) ↑	Colour 1 ↓	Colour 2 ↓
Encoder without skip connections	C1, 75	0.78	0.81	0.78	0.04	0.02	0.02
	C2, 24	0.74	0.81	0.77	0.03	0.01	0.02
	C3, 4	0.80	0.80	0.79	0.04	0.02	0.02
	C4, 4	0.82	0.80	0.79	0.08	0.02	0.02
	M, 27	0.78	0.82	0.76	0.04	0.02	0.02

Table 9.7: Quantitative performance of the model across different surgery types, including mastectomies (grouped).

very different results. The lack of realistic scarring in the predictions may also be due to the subtle and variable nature of these features in the data. Scars vary in position, shape, and colour, and are not consistently visible in all training examples. Without explicit supervision or sufficient samples, the model may ignore these cues entirely.

While the model generalises reasonably well to mastectomy anatomy in terms of global structure and skin tone, the lack of high-frequency surgical details suggests that either finer-grained labels or focused data augmentation may be required to improve realism in these complex cases.

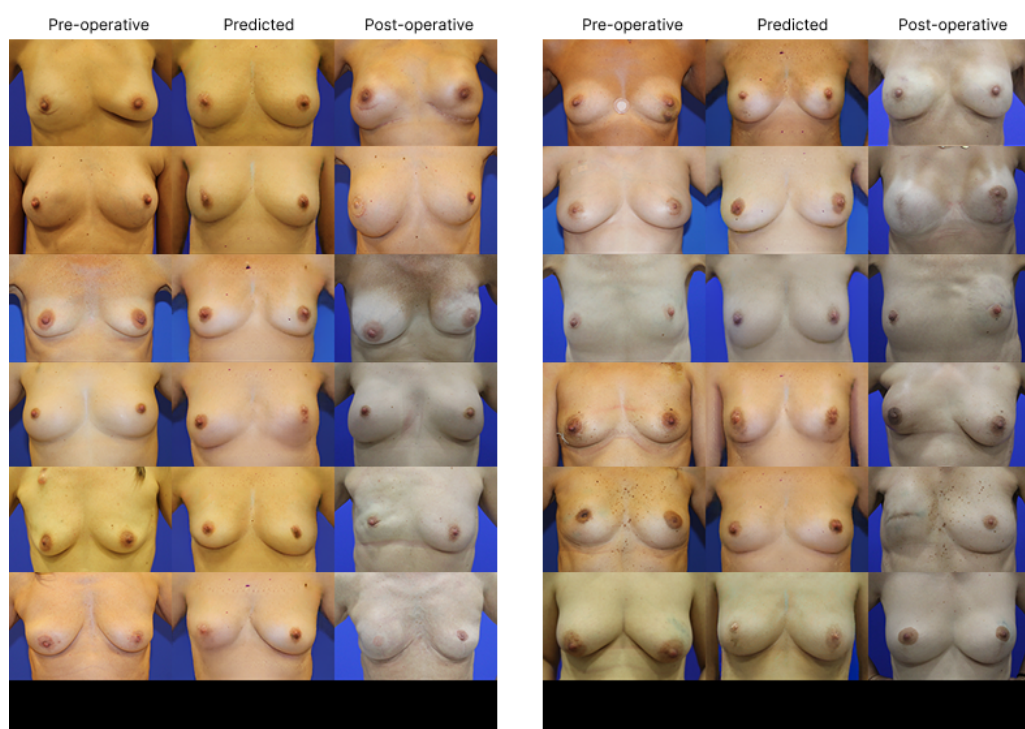


Figure 9.10: Qualitative comparison of model performance on mastectomy cases (grouped). Each triplet shows pre-operative, predicted, and ground truth post-operative images.

Chapter 10

Conclusion and Future Work

This dissertation presented a deep learning-based framework for generating personalised visual predictions of post-operative appearances following breast cancer surgery. By leveraging only pre-operative images and minimal clinical metadata, the proposed method aimed to provide a realistic and faithful forecast of aesthetic outcomes, addressing a critical psychological and clinical need for patients undergoing breast cancer treatment.

The research process began with a systematic assessment of the most recent models of image-to-image translation with state-of-the-art, such as Pix2Pix, StarGAN, and BicycleGAN. These models showed promise, but ultimately presented significant limitations: lack of colour consistency, anatomical fidelity, and clinical relevance.

Conditioned versions of Pix2Pix and BicycleGAN were explored to incorporate surgical metadata and patient-specific context. These modifications enhanced the models' adaptability to different domains and produced outputs that were more aligned with clinical scenarios. However, they still faced challenges related to colour fidelity. Various loss-based strategies were assessed to tackle this issue, including histogram matching and direct colour correction, but these methods were largely ineffective due to the dominant influence of adversarial training. The most effective approach for preserving the original colour distribution involved explicitly maintaining the areas outside the breasts, along with a patch-based mechanism that enforced global colour consistency. These ideas were implemented in the final proposed model.

The final proposed model incorporated several novel contributions: the use of bounding-box-based spatial supervision with soft-edged masks, a segmentation-aware loss strategy, and clinical conditioning through metadata such as surgery type and tumour laterality. The architecture was further enhanced by decoupling semantic classification from adversarial learning, using a dedicated encoder and a discriminator grounded in the patient's pre-operative appearance. This design choice improved colour fidelity and anatomical consistency by reducing post-operative bias in the learning signal.

Empirical results demonstrated that the model outperformed baseline approaches across multiple metrics, including SSIM, IOU, and colour consistency. Architectures without skip connections

showed stronger anatomical realism but at the cost of tonal variation, while skip-connected variants preserved colour better but occasionally blurred structural transformations.

Overall, this dissertation demonstrated that deep generative models can be powerful tools for supporting personalised surgical outcome prediction. By addressing challenges such as anatomical accuracy, colour preservation, and clinical conditioning, this work takes an important step toward making AI-generated post-operative previews more trustworthy and useful in real-world medical settings. While the journey from research to clinical practice remains ongoing, the results presented here lay a strong foundation for future advances in patient-centred decision support and personalised aesthetic outcomes.

10.1 Future Work

Based on the findings and limitations of this work, several promising directions can be explored to improve the system's performance and clinical applicability:

- **Skip Connection Optimisation:** This work showed a clear trade-off between structural realism and colour preservation when using skip connections. Future studies could explore varying the number, depth, and type of skip connections (e.g., partial, gated) to find more balanced configurations that preserve patient identity while allowing meaningful transformations.
- **Learning from Low-Resource Settings:** With more annotated post-operative images becoming available in the future, it will be possible to train better models on surgery types with currently low representation (e.g., mastectomies). Until then, transfer learning or few-shot learning approaches could be explored to improve performance in these underrepresented cases.
- **Patient-Specific Conditioning:** In addition to surgery type and tumour laterality, integrating other clinical factors (e.g., age, skin tone, BMI) as conditioning signals could improve the specificity of the predicted outcomes and better reflect realistic variation.
- **Evaluation via Clinical and Patient Feedback:** While quantitative metrics like SSIM and IOU provide useful benchmarks, ultimately, these predictions' clinical value depends on clinicians and patients' perceived realism. Conducting user studies and integrating perceptual feedback could help fine-tune the system toward real-world usability.

References

- [1] Iman Abbasnejad et al. “SCONE-GAN: Semantic Contrastive learning-based Generative Adversarial Network for an end-to-end image translation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 1111–1120.
- [2] Aziz Alotaibi. “Deep Generative Adversarial Networks for Image-to-Image Translation: A Review”. en. In: *Symmetry* 12.10 (Oct. 2020), p. 1705. ISSN: 2073-8994. DOI: [10.3390/sym12101705](https://doi.org/10.3390/sym12101705).
- [3] Robin Andlauer et al. “3D-Guided Face Manipulation of 2D Images for the Prediction of Post-Operative Outcome After Cranio-Maxillofacial Surgery”. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 7349–7363. DOI: [10.1109/TIP.2021.3096081](https://doi.org/10.1109/TIP.2021.3096081).
- [4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein generative adversarial networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.
- [5] Aayush Bansal, Yaser Sheikh, and Deva Ramanan. “Pixelnn: Example-based image synthesis”. In: *arXiv preprint arXiv:1708.05349* (2017).
- [6] ATPM Brands-Appeldoorn et al. “Breast cancer patient-reported outcome of factors influencing cosmetic satisfaction after breast-conserving therapy”. In: *Breast Cancer* 29.1 (2022), pp. 114–120. DOI: [10.1007/s12282-021-01287-0](https://doi.org/10.1007/s12282-021-01287-0).
- [7] Cancer Research UK. *Breast-Conserving Surgery (Lumpectomy)*. Accessed: 2024-12-01. 2024. URL: <https://www.cancerresearchuk.org/about-cancer/breast-cancer/treatment/surgery/breast-conserving-surgery-lumpectomy>.
- [8] Cancer Research UK. *Mastectomy: Surgery to Remove the Whole Breast*. Accessed: 2024-12-01. 2024. URL: <https://www.cancerresearchuk.org/about-cancer/breast-cancer/treatment/surgery/remove-whole-breast-mastectomy>.
- [9] Hanqun Cao et al. “A survey on generative diffusion models”. In: *IEEE Transactions on Knowledge and Data Engineering* (2024).
- [10] J.S. Cardoso, W. Silva, and M.J. Cardoso. “Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment”. In: *Breast* 49 (2020), pp. 123–130. DOI: [10.1016/j.breast.2019.11.006](https://doi.org/10.1016/j.breast.2019.11.006).
- [11] Jaime S Cardoso and Maria J Cardoso. “Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment”. In: *Artificial intelligence in medicine* 40.2 (2007), pp. 115–126.
- [12] Maria Manuel Domingos Carvalho. “Towards Biometrically-Morphed Medical Case-based Explanations”. In: (2022).

- [13] Isabella Castiglioni et al. “AI applications to medical images: From machine learning to deep learning”. In: *Physica Medica* 83 (2021), pp. 9–24. ISSN: 1120-1797. DOI: <https://doi.org/10.1016/j.ejmp.2021.02.006>. URL: <https://www.sciencedirect.com/science/article/pii/S1120179721000946>.
- [14] Tanujit Chakraborty et al. “Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art”. In: *Machine Learning: Science and Technology* 5.1 (2024), p. 011001.
- [15] Praveen Kumar Chandaliya and Neeta Nain. “PlasticGAN: Holistic generative adversarial network on face plastic and aesthetic surgery”. In: *Multimedia Tools and Applications* 81.22 (2022), pp. 32139–32160.
- [16] Christian Chartier et al. “BreastGAN: artificial intelligence-enabled breast augmentation simulation”. In: *Aesthetic Surgery Journal Open Forum*. Vol. 4. Oxford University Press US. 2022, ojab052.
- [17] Yunjey Choi et al. “Stargan v2: Diverse image synthesis for multiple domains”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8188–8197.
- [18] Yunjey Choi et al. “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8789–8797.
- [19] Hyungjin Chung and Jong Chul Ye. “Score-based diffusion models for accelerated MRI”. In: *Medical Image Analysis* 80 (2022), p. 102479. ISSN: 1361-8415. DOI: <https://doi.org/10.1016/j.media.2022.102479>. URL: <https://www.sciencedirect.com/science/article/pii/S1361841522001268>.
- [20] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat gans on image synthesis”. In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.
- [21] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. “Density estimation using real nvp”. In: *arXiv preprint arXiv:1605.08803* (2016).
- [22] Nuno Freitas et al. “Reproducing Asymmetries Caused by Breast Cancer Treatment in Pre-Operative Breast Images”. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. 2024, pp. 1–5. DOI: [10.1109/ISBI56570.2024.10635739](https://doi.org/10.1109/ISBI56570.2024.10635739).
- [23] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feed-forward neural networks”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.
- [24] Harshvardhan GM et al. “A comprehensive survey and analysis of generative models in machine learning”. In: *Computer Science Review* 38 (2020), p. 100285. ISSN: 1574-0137. DOI: <https://doi.org/10.1016/j.cosrev.2020.100285>. URL: <https://www.sciencedirect.com/science/article/pii/S1574013720303853>.
- [25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016.
- [26] Ian Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

- [27] Chenqi Guo et al. “A fully automatic framework for evaluating cosmetic results of breast conserving therapy”. In: *Machine learning with applications* 10 (2022), p. 100430.
- [28] Aaron Hertzmann et al. “Image analogies”. en. In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, Aug. 2001, pp. 327–340. ISBN: 978-1-58113-374-5. DOI: [10.1145/383259.383295](https://doi.org/10.1145/383259.383295). URL: <https://dl.acm.org/doi/10.1145/383259.383295>.
- [29] Irina Higgins et al. “beta-vae: Learning basic visual concepts with a constrained variational framework.” In: *ICLR (Poster)* 3 (2017).
- [30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.
- [31] Jonathan Ho and Tim Salimans. “Classifier-free diffusion guidance”. In: *arXiv preprint arXiv:2207.12598* (2022).
- [32] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.
- [33] Phillip Isola et al. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.
- [34] Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. “i-revnet: Deep invertible networks”. In: *arXiv preprint arXiv:1802.07088* (2018).
- [35] Orit Kaidar-Person et al. “Evaluating the ability of an artificial-intelligence cloud-based platform designed to provide information prior to locoregional therapy for breast cancer in improving patient’s satisfaction with therapy: The CINDERELLA trial”. In: *PLoS One* 18.8 (2023), e0289365.
- [36] Tero Karras et al. “Training generative adversarial networks with limited data”. In: *Advances in neural information processing systems* 33 (2020), pp. 12104–12114.
- [37] Joseph Kim and Hee Seung Chin. “Deep learning-based prediction of the retinal structural alterations after epiretinal membrane surgery”. In: *Scientific Reports* 13.1 (2023), p. 19275.
- [38] M Kim Kim et al. “Effect of cosmetic outcome on quality of life after breast cancer surgery”. In: *European Journal of Surgical Oncology (EJSO)* 41.3 (2015), pp. 426–432.
- [39] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: [1312.6114](https://arxiv.org/abs/1312.6114) [stat.ML].
- [40] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: [1412.6980](https://arxiv.org/abs/1412.6980) [cs.LG]. URL: <https://arxiv.org/abs/1412.6980>.
- [41] Durk P Kingma and Prafulla Dhariwal. “Glow: Generative flow with invertible 1x1 convolutions”. In: *Advances in neural information processing systems* 31 (2018).
- [42] Durk P Kingma et al. “Semi-supervised learning with deep generative models”. In: *Advances in neural information processing systems* 27 (2014).
- [43] Gihyun Kwon and Jong Chul Ye. “Improving Diffusion-based Image Translation using Asymmetric Gradient Guidance”. In: *arXiv preprint arXiv:2306.04396* (2023).

- [44] Anders Boesen Lindbo Larsen et al. “Autoencoding beyond pixels using a learned similarity metric”. In: *International conference on machine learning*. PMLR. 2016, pp. 1558–1566.
- [45] Hsin-Ying Lee et al. “Diverse image-to-image translation via disentangled representations”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 35–51.
- [46] Bo Li et al. “Vqbb: Image-to-image translation with vector quantized brownian bridge”. In: *arXiv preprint arXiv:2205.07680* 2 (2022).
- [47] Xinyang Li et al. “Attribute guided unpaired image-to-image translation with semi-supervised learning”. In: *arXiv preprint arXiv:1904.12428* (2019).
- [48] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. “Unsupervised image-to-image translation networks”. In: *Advances in neural information processing systems* 30 (2017).
- [49] Rayeesa Mehmood, Rumaan Bashir, and Kaiser Giri. “Deep Generative Models: A Review”. In: *Indian Journal Of Science And Technology* 16 (Feb. 2023), pp. 460–467. DOI: [10.17485/IJST/v16i7.2296](https://doi.org/10.17485/IJST/v16i7.2296).
- [50] Rayeesa Mehmood, Rumaan Bashir, and Kaiser Giri. “Deep Generative Models: A Review”. In: *Indian Journal Of Science And Technology* 16 (Feb. 2023), pp. 460–467. DOI: [10.17485/IJST/v16i7.2296](https://doi.org/10.17485/IJST/v16i7.2296).
- [51] Mehdi Mirza and Simon Osindero. “Conditional Generative Adversarial Nets”. en. In: *arXiv:1411.1784* (Nov. 2014). *arXiv:1411.1784* [cs]. URL: <http://arxiv.org/abs/1411.1784>.
- [52] Helena Montenegro, Maria J Cardoso, and Jaime S Cardoso. “A disentangled approach to predict the aesthetic outcomes of breast cancer treatment”. In: *European Conference on Computer Vision*. Springer. 2025, pp. 311–327.
- [53] Helena Montenegro, Maria J Cardoso, and Jaime S Cardoso. “An inpainting approach to manipulate asymmetry in pre-operative breast images”. In: *arXiv preprint arXiv:2502.05652* (2025).
- [54] Aamir Mustafa and Rafał K Mantiuk. “Transformation consistency regularization—a semi-supervised paradigm for image-to-image translation”. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer. 2020, pp. 599–615.
- [55] Sriyani Padmalatha et al. “Higher risk of depression after total mastectomy versus breast reconstruction among adult women with breast cancer: A systematic review and metaregression”. In: *Clinical breast cancer* 21.5 (2021), e526–e538.
- [56] Kushagra Pandey et al. “Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents”. In: *arXiv preprint arXiv:2201.00308* (2022).
- [57] Yingxue Pang et al. “Image-to-image translation: Methods and applications”. In: *IEEE Transactions on Multimedia* 24 (2021), pp. 3859–3881.
- [58] George Papamakarios, Theo Pavlakou, and Iain Murray. “Masked autoregressive flow for density estimation”. In: *Advances in neural information processing systems* 30 (2017).
- [59] Taesung Park et al. “Semantic image synthesis with spatially-adaptive normalization”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2337–2346.
- [60] Gaurav Parmar et al. “Zero-shot image-to-image translation”. In: *ACM SIGGRAPH 2023 Conference Proceedings*. 2023, pp. 1–11.

- [61] Richard D Pezner et al. “Breast retraction assessment: an objective evaluation of cosmetic results of patients treated conservatively for breast cancer”. In: *International Journal of Radiation Oncology* Biology* Physics* 11.3 (1985), pp. 575–578.
- [62] Alec Radford. “Unsupervised representation learning with deep convolutional generative adversarial networks”. In: *arXiv preprint arXiv:1511.06434* (2015).
- [63] AN Ramesh et al. “Artificial Intelligence in Medicine”. In: *Annals of the Royal College of Surgeons of England* 86 (2004), pp. 334–338. DOI: [10.1308/147870804290](https://doi.org/10.1308/147870804290).
- [64] Danilo Rezende and Shakir Mohamed. “Variational inference with normalizing flows”. In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.
- [65] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.
- [66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional networks for biomedical image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2015, pp. 234–241.
- [67] Chitwan Saharia et al. “Palette: Image-to-image diffusion models”. In: *ACM SIGGRAPH 2022 conference proceedings*. 2022, pp. 1–10.
- [68] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. “Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models”. In: *arXiv preprint arXiv:2104.05358* (2021).
- [69] Shallu Sharma and Rajesh Mehra. “Implications of pooling strategies in convolutional neural networks: A deep insight”. In: *Foundations of Computing and Decision Sciences* 44.3 (2019), pp. 303–330.
- [70] K. Slim and J.-E. Bazin. “From informed consent to shared decision-making in surgery”. In: *Journal of Visceral Surgery* 156.3 (2019), pp. 181–184. ISSN: 1878-7886. DOI: <https://doi.org/10.1016/j.jviscsurg.2019.04.014>. URL: <https://www.sciencedirect.com/science/article/pii/S1878788619300608>.
- [71] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. “Learning Structured Output Representation using Deep Conditional Generative Models”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.
- [72] Ilona Stolpner et al. “Long-term patient satisfaction and quality of life after breast-conserving therapy: a prospective study using the BREAST-Q”. In: *Annals of Surgical Oncology* 28 (2021), pp. 8742–8751.
- [73] Xuan Su et al. “Dual diffusion implicit bridges for image-to-image translation”. In: *arXiv preprint arXiv:2203.08382* (2022).
- [74] Dmitrii Torbunov et al. “Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation”. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2023, pp. 702–712.
- [75] LI Tsouskas and IS Fentiman. “Breast compliance: a new method for evaluation of cosmetic outcome after conservative treatment of early breast cancer”. In: *Breast cancer research and treatment* 15 (1990), pp. 185–190.
- [76] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009, pp. 433–460.

- [77] Arash Vahdat and Jan Kautz. “NVAE: A deep hierarchical variational autoencoder”. In: *Advances in neural information processing systems* 33 (2020), pp. 19667–19679.
- [78] E Van Limbergen, E Van der Schueren, and K Van Tongelen. “Cosmetic evaluation of breast conserving treatment for mammary cancer. 1. Proposal of a quantitative scoring system”. In: *Radiotherapy and oncology* 16.3 (1989), pp. 159–167.
- [79] A Vaswani. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* (2017).
- [80] Tengfei Wang et al. “Pretraining is all you need for image-to-image translation”. In: *arXiv preprint arXiv:2205.12952* (2022).
- [81] Ting-Chun Wang et al. “High-resolution image synthesis and semantic manipulation with conditional gans”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8798–8807.
- [82] Mengfei Xia et al. “A diffusion model translator for efficient image-to-image translation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [83] Tae Keun Yoo, Joon Yul Choi, and Hong Kyu Kim. “A generative adversarial network approach to predicting postoperative appearance after orbital decompression surgery for thyroid eye disease”. In: *Computers in biology and medicine* 118 (2020), p. 103628.
- [84] Min Zhao et al. “Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 3609–3623.
- [85] Jun-Yan Zhu et al. “Toward multimodal image-to-image translation”. In: *Advances in neural information processing systems* 30 (2017).
- [86] Jun-Yan Zhu et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.
- [87] Hooshiar Zolfagharnasab et al. “A regression model for predicting shape deformation after breast conserving surgery”. In: *Sensors* 18.1 (2018), p. 167.

Appendix A

Further Use of AI Tools

AI tools, specifically Grammarly and ChatGPT, were used during the preparation of this thesis to assist with grammar, spelling, and stylistic improvements. Their use was limited to enhancing the clarity and readability of the text. All AI-generated suggestions, even those related to language and style, were carefully reviewed and revised by me. The content, ideas, analysis, implementation, and conclusions presented in this thesis are entirely my own work.