VILNIUS TECH
Department of Electronic Systems

Giga Shubitidze

Research on Generative Artificial Intelligence Algorithms for Building the Dataset of Eyeglass Frame Images

Generatyvinių dirbtinio intelekto algoritmų akinių rėmelių vaizdų duomenų rinkinio sudarymui tyrimas

Master Thesis

Informatics Engineering, State Code 6211BX015
Artificial Intelligence Systems Specialisation
Informatics Engineering Study Area

Vilnius, 2026

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
FACULTY OF ELECTRONICS
DEPARTMENT OF ELECTRONIC SYSTEMS

CONFIRMS
Head of Department

_____

(signature)
Prof. Dr Artūras Serackis
_____ _____, 2026

Giga Shubitidze

RESEARCH ON GENERATIVE ARTIFICIAL INTELLIGENCE ALGORITHMS FOR BUILDING THE DATASET OF EYEGLASS FRAME IMAGES

GENERATYVINIŲ DIRBTINIO INTELEKTO ALGORITMŲ AKINIŲ RĖMELIŲ VAIZDŲ DUOMENŲ RINKINIO SUDARYMUI TYRIMAS

Master Thesis

Informatics Engineering, State Code 6211BX015
Artificial Intelligence Systems Specialisation
Informatics Engineering Study Area

Supervisor Prof. Dr Dalius Matuzevičius _____
(scientific title, name, surname) *(signature) (date)*

Vilnius, 2026

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
FACULTY OF ELECTRONICS
DEPARTMENT OF ELECTRONIC SYSTEMS

Informatics Engineering study field
Information Electronics Systems study programme, state code 6211BX015
Artificial Intelligence Systems specialisation

APPROVED BY
Head of Department
Artūras Serackis
2026-05-02

OBJECTIVES FOR MASTER THESIS

No. DISfmu-24–12628

Vilnius

For student Giga Shubitidze

Master Thesis title: Research on Generative Artificial Intelligence Algorithms for Building the Dataset of Eyeglass Frame Images

Deadline for completion of the final work according to the planned study schedule.

THE OBJECTIVES:

Description of the thesis. Aim: Investigate the application of generative artificial intelligence methods for building a dataset of eyeglass frame images.

Description of the thesis. Objectives:
1. Conduct an analytical review of recent generative artificial intelligence methods for image synthesis and dataset generation.
2. Select deep learning methods that are promising for building a dataset of eyeglass frame images.
3. Prepare and curate a dataset for the experimental investigation of generative methods for eyeglass frame image synthesis.
4. Implement and experimentally investigate the selected generative methods for building a dataset of eyeglass frame images, and to present the research results.

Table of contents. MSc:

Academic Supervisor Professor Dalius Matuzevičius

DECLARATION OF AUTHORSHIP OF THE FINAL THESIS

I, Giga Shubitidze, hereby declare that this Master's thesis, titled *Research on Generative Artificial Intelligence Algorithms for Building the Dataset of Eyeglass Frame Images*, is the result of my own work. All sources used in this thesis have been acknowledged and properly cited. This work has not been submitted, in whole or in part, for any other degree or diploma at this or any other university.

I confirm that the digital text and figures attached to this thesis match the printed version. I take full responsibility for the correctness of the data and conclusions presented.

Giga Shubitidze	_____	_____
(Author's name and surname)	(Signature)	(Date)

Shubitidze, G. 2026. Research on Generative Artificial Intelligence Algorithms for Building the Dataset of Eyeglass Frame Images. Master's Thesis. Vilnius: Vilnius Gediminas Technical University, Faculty of Electronics, Department of Electronic Systems. 78 p., 12 figures, 11 tables, 50 bibliographic references, 4 appendices.

Keywords: generative artificial intelligence, generative adversarial networks, StyleGAN2, image synthesis, segmentation mask generation, dataset augmentation, eyeglass frames, conditional GAN, downstream evaluation.

Annotation

The thesis investigates how generative artificial intelligence can be used to build a labelled dataset of eyeglass frame images, where each generated image comes with a pixel-aligned binary segmentation mask. The motivation is the small size of public eyewear datasets, which limits the accuracy of segmentation, virtual try-on and recommendation models in the eyewear industry.

After reviewing variational autoencoders, generative adversarial networks and diffusion models, the thesis selects a conditional StyleGAN2 architecture and adapts it for the eyewear domain. A U-Net mask decoder is attached to the synthesis network through skip connections so that the image and the mask are produced in a single forward pass and stay pixel-aligned by construction. The model is trained on about 6,500 paired images from a proprietary partner dataset and used to generate a synthetic dataset of 10,000 paired samples.

The synthetic dataset is evaluated against eight metrics, including three domain-specific metrics introduced in this thesis (Edge Coherence, Mask Regularity, Frame Symmetry). The achieved Fréchet Inception Distance is 6.07, comparable to published low-data StyleGAN2 results. The newly-introduced metrics report Edge Coherence 0.626, Mask Regularity 0.984 and Frame Symmetry 0.877, indicating high image–mask consistency, single-piece masks and learned bilateral symmetry. The downstream usefulness of the dataset is then verified by training a U-Net segmentation model on GAN-generated frames composited onto face images. The model reaches an intersection-over-union of 0.867 on a held-out CelebAMask-HQ test set, surpassing the same model trained on real catalogue frames composited the same way (IoU 0.831) by 3.6 points. For the segmentation task tested, the dataset and its compositing pipeline can therefore serve as a practical alternative to manual data collection, at near-zero marginal annotation cost per sample.

Shubitidze, G. 2026. Generatyvinių dirbtinio intelekto algoritmų akinių rėmelių vaizdų duomenų rinkinio sudarymui tyrimas. Magistro baigiamasis darbas. Vilnius: Vilniaus Gedimino technikos universitetas, Elektronikos fakultetas, Elektroninių sistemų katedra. 78 p., 12 paveikslų, 11 lentelių, 50 bibliografiniai šaltiniai, 4 priedai.

Reikšminiai žodžiai: generatyvinis dirbtinis intelektas, generatyviniai priešpriešų tinklai, StyleGAN2, vaizdų sintezė, segmentacijos kaukių generavimas, duomenų rinkinių praplėtimas, akinių rėmeliai, sąlyginis GAN, pasrovinis vertinimas.

Anotacija

Šiame magistro baigiamajame darbe nagrinėjama, kaip generatyvinis dirbtinis intelektas gali būti naudojamas akinių rėmelių vaizdų duomenų rinkiniui sudaryti, kur kiekvienas sugeneruotas vaizdas turi ir su juo pikselio tikslumu sutampančią dvejetainę segmentacijos kaukę. Darbo motyvacija: nedideli viešai prieinami akinių duomenų rinkiniai, ribojantys segmentacijos, virtualaus matavimosi ir rekomendacijų modelių tikslumą akinių pramonėje.

Apžvelgus variacinius autokoderius, generatyvinius priešpriešų tinklus ir difuzinius modelius, darbe pasirenkamas sąlyginis StyleGAN2 modelis ir pritaikomas akinių sričiai. Prie sintezės tinklo per praleidimo jungtis prijungiamas U-Net tipo kaukės dekoderis, todėl vaizdas ir kaukė generuojami per vieną pirmyninį žingsnį ir lieka pikselio tikslumu sutampantys. Modelis apmokomas naudojant apie 6,500 suporuotų partnerio duomenų rinkinio vaizdų ir naudojamas 10,000 suporuotų pavyzdžių sintetiniam duomenų rinkiniui sukurti.

Sintetinis rinkinys įvertinamas pagal aštuonias metrikas, įskaitant tris šiame darbe pasiūlytas srities specifines metrikas (Edge Coherence, Mask Regularity, Frame Symmetry). Pasiekta Fréchet Inception Distance reikšmė yra 6,07, palyginama su publikuotais StyleGAN2 rezultatais mažiems duomenų rinkiniams. Šiame darbe pasiūlytos metrikos parodė reikšmes 0,626 (Edge Coherence), 0,984 (Mask Regularity) ir 0,877 (Frame Symmetry); jos liudija aukštą vaizdo ir kaukės atitikimą, vientisas kaukes ir išmoktą dvipusę simetriją. Po to praktinis naudingumas patikrinamas apmokant U-Net segmentacijos modelį, naudojant GAN sugeneruotus rėmelius, sukomponuotus su veidų nuotraukomis. Modelis pasiekia 0,867 IoU rezultatą atskirtoje CelebAMask-HQ testų aibėje, 3,6 IoU taško daugiau nei tas pats modelis, apmokytas naudojant tikrus katalogo rėmelius (0,831). Tirtos segmentacijos užduoties atveju sukurtas duomenų rinkinys ir kompozicijos algoritmas gali būti praktinė alternatyva realiam duomenų rinkimui, beveik be papildomų anotavimo kaštų vienam pavyzdžiui.

Contents

Notation and Abbreviations

Acronyms

ADA	Adaptive Discriminator Augmentation
BCE	Binary Cross-Entropy
CNN	Convolutional Neural Network
DDPM	Denoising Diffusion Probabilistic Model
EMA	Exponential Moving Average
FID	Fréchet Inception Distance
GAN	Generative Adversarial Network
IoU	Intersection over Union
IS	Inception Score
LDM	Latent Diffusion Model
LPIPS	Learned Perceptual Image Patch Similarity
MLP	Multi-Layer Perceptron
PnP	Perspective-n-Point (camera pose estimation)
ReLU	Rectified Linear Unit
RGB	Red–Green–Blue (colour image)
RGBA	Red–Green–Blue–Alpha (colour image with transparency)
SPADE	Spatially-Adaptive Denormalisation
SVG	Scalable Vector Graphics
VAE	Variational Autoencoder
VGG	Visual Geometry Group (a CNN architecture)
VQ-VAE	Vector-Quantised Variational Autoencoder
WGAN	Wasserstein Generative Adversarial Network

Mathematical Symbols

$\mathbf{x}$	Input image (colour, RGB)
m	Binary segmentation mask
$\hat{\mathbf{x}}, \hat{m}$	Generated image and mask
$\mathbf{z}$	Latent noise vector
$\mathbf{w}$	Style vector in StyleGAN $\mathscr{W}$-space
$\mathbf{e}_s$	Style embedding for training image s
G, D	Generator and discriminator networks
$\mathscr{L}$	Loss function
λ_m, λ_p	Mask-loss and perceptual-loss weights
ψ	StyleGAN truncation parameter

ρ	EMA decay rate
β_1, β_2	Adam optimiser hyper-parameters
γ	R1 gradient-penalty weight
p_{data}	True data distribution
p_{model}	Distribution learned by the model

Introduction

Topicality of the Problem and Aim of the Work

Online eyewear retail is part of a market valued at over USD 200 billion globally in 2024 [47], and most of its customer-facing technology (virtual try-on, frame detection, segmentation, style recommendation, 3-D reconstruction) depends on supervised computer-vision models trained on labelled images of eyeglass frames. The accuracy of these models is bounded by the size and the diversity of the training data available to them. In practice, public datasets of eyeglass frames are small, often a few thousand images, frequently come without segmentation masks, and almost never cover the full assortment of shapes, materials and colours found in modern catalogues. Manual collection and annotation of additional images is slow and expensive, and standard data augmentation techniques such as horizontal flipping or colour jitter cannot create new frame shapes that were not in the original data. Synthetic data generated by deep generative models has been shown to alleviate similar shortages in other narrow product domains [31, 32].

Generative artificial intelligence offers a different way: once a generative model has learned the distribution of real frame photographs, it can produce arbitrarily many new realistic samples that, taken together, broaden the effective coverage of the training data. Two families of generative models are particularly relevant for image synthesis: Generative Adversarial Networks (GANs) [1] and Variational Autoencoders (VAEs) [2]. Diffusion models [3] have appeared more recently and now lead on large general-purpose benchmarks. Each family balances these factors differently. For the eyewear domain (a few thousand training images, 2:1 aspect ratio, strong bilateral symmetry, and a hard requirement for paired segmentation masks) the right architecture is not obvious and has to be chosen with care.

The thesis builds on an active line of work at Vilnius Gediminas Technical University on computer vision for the eyewear industry: 3-D textured frame reconstruction from single images [44] and supervised eyewear detection and segmentation networks [45]. Each of these projects is limited by the same data shortage. The current thesis attacks that shortage directly, by replacing manual data collection with a generative model whose outputs are a drop-in substitute for real catalogue images in the downstream pipeline.

The aim of the thesis is to investigate the application of generative artificial intelligence methods for building a labelled dataset of eyeglass frame images, to design a generative architecture that produces an image and its segmentation mask in a single forward pass, to construct such a dataset experimentally, and to verify its usefulness on a downstream segmentation task using a real-world test corpus.

Tasks of the Work

To reach the stated aim, four tasks are addressed:

1. Conduct an analytical review of recent generative AI methods for image synthesis and dataset generation, covering VAEs, GANs and diffusion models, and select the most promising one for the eyewear domain.
2. Design and implement a conditional GAN architecture that generates a 256×512 RGB frame image and its pixel-aligned binary segmentation mask in a single forward pass.
3. Prepare a curated training corpus of paired image–mask samples, build a perspective-correct compositing pipeline that embeds frames into face contexts, and assemble a held-out real-world test corpus.
4. Train the proposed generator, generate a synthetic dataset of 10,000 paired samples, evaluate the dataset against intrinsic quality metrics, and verify its downstream usefulness by training a segmentation model on the generated data and testing on real face images.

Applied Methods of Investigation and Analysis

The investigation combines analytical and experimental methods.

Analytical methods are used in Chapter 1 to review the literature on generative modelling and to identify the architectures and regularisation techniques best suited to the eyewear domain.

Experimental methods are used in Chapters 2 to 4. A conditional StyleGAN2 model adapted for joint image–mask synthesis is implemented in PyTorch and trained for 350 epochs on approximately 6,500 paired catalogue samples. A five-mode style-interpolation strategy generates 10,000 paired samples. A perspective-correct face compositing pipeline bridges the catalogue-to-real domain gap. Six U-Net segmentation models are then trained from scratch on different subsets of the data and evaluated on a real held-out CelebAMask-HQ test corpus.

Quantitative evaluation uses a combination of generic metrics (Fréchet Inception Distance [20], Inception Score [21], LPIPS [23]), three domain-specific metrics introduced in this thesis (Edge Coherence, Mask Regularity, Frame Symmetry), two diversity metrics (Colour Diversity, Mask Shape Diversity) and standard segmentation metrics (IoU, Dice, precision, recall) on the downstream task.

Qualitative analysis of representative samples and failure cases complements the quantitative evaluation in Chapter 4.

Novelty of the Work and Its Practical Value

The main contributions of the thesis are:

1. A comparison of conditional StyleGAN2 against VAE and diffusion baselines using a decision matrix that maps each candidate against five design requirements.
2. A conditional StyleGAN2 architecture adapted to the eyewear domain with three specific modifications: (i) a non-square synthesis network targeted at the 256×512 aspect ratio

of frontal frame photographs; (ii) a U-Net mask decoder fed by skip connections from the synthesis network so that the image and the mask are produced in one forward pass and stay pixel-aligned by construction; (iii) a 5-channel conditional discriminator that jointly evaluates image realism, mask plausibility and image–mask consistency.

3. A five-mode style-interpolation generation strategy that combines pure styles, two- to four-way blends and intra-style noise variation in fixed proportions, to expand the diversity of the generated dataset beyond the training distribution.
4. Three domain-specific metrics (Edge Coherence, Mask Regularity and Frame Symmetry) that measure the physical realism of synthetic eyeglass frames in a way that generic metrics like FID cannot.
5. A perspective-correct face compositing pipeline that uses MediaPipe Face Landmarker landmarks and OpenCV `solvePnP` to place catalogue frames onto frontal face photographs at the correct head pose, producing realistic image–mask training pairs.
6. A downstream experiment confirming that the synthetic dataset improves segmentation over a real-data baseline on a held-out CelebAMask-HQ test set (details in Chapter 4).
7. A finding that increasing GAN data volume improved downstream performance on the primary test set, with the uncapped condition outperforming both the size-matched and real-data baselines.

The achieved segmentation IoU of 0.867 confirms that the dataset and compositing pipeline are operationally useful as a replacement for manual data collection in the eyewear domain. The work has been published in a conference paper [46] (reproduced in Appendix D).

Scope of the Work

The thesis covers (i) an analytical review of contemporary generative AI methods for moderate-resolution image synthesis on small training corpora; (ii) the design and implementation of a single conditional StyleGAN2 variant for joint image–mask synthesis at 256×512; (iii) the preparation of a proprietary partner dataset of around 6,500 paired image–mask samples; (iv) the training of the proposed model and the generation of a 10,000-sample synthetic dataset; (v) the evaluation of the generated dataset against eight intrinsic metrics; (vi) the construction of a perspective-correct compositing pipeline; and (vii) the downstream task evaluation of six U-Net segmentation models on a real-world test corpus.

The following topics are not addressed: (a) a practical, head-to-head implementation of competing VAE or diffusion baselines on the same corpus, as the comparison is conducted analytically; (b) higher-resolution synthesis at 1024×2048 or above; (c) synthesis of three-quarter or profile views; (d) virtual try-on as a final user-facing application; (e) real-time interactive generation; (f) multi-seed re-runs of the downstream segmentation experiments to obtain seed-level confidence intervals on the IoU differences. Selected directions are discussed as future work in the Conclusions.

The thesis is organised as follows. Chapter 1 presents the analytical review of generative AI methods. Chapter 2 develops the proposed conditional StyleGAN2 architecture and training procedure. Chapter 3 documents the data preparation and the compositing pipeline. Chapter 4 reports the experimental results: GAN training dynamics, intrinsic dataset evaluation and downstream segmentation evaluation. The Conclusions chapter draws the overall findings and identifies directions for future work. Four appendices contain the experimental research artefacts (Appendix A), the source code and prepared dataset (Appendix B), the conference poster (Appendix C) and the conference paper (Appendix D).

1. Analytical Review of Generative AI Methods for Image and Dataset Synthesis

This chapter reviews the generative AI methods relevant to building a labelled dataset of eyeglass frame images. Three families of generative models are examined: Variational Autoencoders, Generative Adversarial Networks and diffusion models, together with the metrics commonly used to evaluate them. The chapter ends by explaining, on the basis of this review, why a conditional StyleGAN2 is the best fit for the task and what gap in the eyewear-specific literature the present thesis addresses.

1.1. Generative Modelling: Problem Statement

Generative modelling is the task of learning the underlying distribution $p_{\text{data}}(\mathbf{x})$ of a dataset $\mathscr{D} = \{\mathbf{x}_i\}_{i=1}^{n}$ from a finite number of samples, so that new samples $\hat{\mathbf{x}} \sim p_{\text{model}}(\mathbf{x})$ can be drawn from the learned distribution. In our case $\mathbf{x}$ is a 256×512 image of an eyeglass frame, and the goal is a p_{model} that captures the diversity of frame shapes, materials and colours found in real catalogues. Three paradigms have emerged for solving this problem at scale: likelihood-based latent-variable models (VAEs), adversarial models (GANs), and score-based / diffusion models. Each one has different strengths and weaknesses.

1.2. Variational Autoencoders

A Variational Autoencoder [2] models the data as generated from an unobserved latent variable $\mathbf{z}$ drawn from a prior $p(\mathbf{z}) = \mathscr{N}(\mathbf{0}, \mathbf{I})$, with the data sample produced by a decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$. Because the marginal likelihood $p_\theta(\mathbf{x})$ is intractable, training is performed by maximising the Evidence Lower Bound:

$$\mathscr{L}_{\text{ELBO}}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - D_{\text{KL}}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\big)\,, \tag{1.1}$$

where the recognition network $q_\phi(\mathbf{z} \mid \mathbf{x})$ is the encoder.

β-VAE [30] trades reconstruction fidelity for disentanglement; VQ-VAE [28] uses a discrete latent codebook; recent hierarchical models such as NVAE [29] approach GAN-level perceptual quality at significant computational cost.

The strengths of VAEs are stable training, an explicit and structured latent space, an exact lower bound on the data likelihood, and direct inference of latent codes via the encoder.

Their main limitation is that vanilla VAEs produce visibly blurry samples, and hierarchical variants narrow the gap to GANs only at large memory and training-time cost. The structural reason for the blurring and its domain-specific consequences for eyeglass frames are discussed in Section 1.6.

1.3. Generative Adversarial Networks

A GAN [1] formulates generative modelling as a two-player game between a generator G that maps noise $\mathbf{z} \sim p(\mathbf{z})$ to an image $\hat{\mathbf{x}} = G(\mathbf{z})$ and a discriminator D that tries to tell real from generated samples. The original objective is the saddle-point problem

$$\min_G \max_D \; \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]. \tag{1.2}$$

Unlike a VAE, a GAN imposes no pixel-wise reconstruction loss; the generator is shaped only by gradients backpropagated through the discriminator. This is why GAN samples are typically sharper.

GANs are famously hard to train. Three common failure modes are mode collapse (the generator covers only a few output modes), vanishing gradients (the discriminator becomes too strong), and non-convergence (loss curves oscillate). Several remedies have become standard. Wasserstein GAN [13] replaces the Jensen–Shannon divergence with a Wasserstein distance for smoother gradients; spectral normalisation [14] constrains the Lipschitz constant of D at low cost; the hinge loss [15] yields more stable training than the original cross-entropy formulation; the R1 gradient penalty [16] prevents D from memorising sharp decision boundaries; Adaptive Discriminator Augmentation (ADA) [6] substantially improves GAN training when data is limited.

The most relevant architectural milestones for high-resolution synthesis are DCGAN [9], Progressive GAN [8], StyleGAN [4], StyleGAN2 [5], StyleGAN2-ADA [6] and StyleGAN3 [7]. The StyleGAN family introduced two ideas that matter directly for our task: (i) a mapping network that produces a disentangled intermediate latent space $\mathscr{W}$, and (ii) weight-demodulated convolutions that fold style modulation into the convolution kernel itself, producing crisper outputs than AdaIN-based variants.

For our task we need the generator to produce not just a frame image but also its segmentation mask. The conditional GAN framework [10] extends both G and D to accept an auxiliary input. Pix2Pix [11] performs paired image-to-image translation with a U-Net generator and a PatchGAN discriminator; SPADE [12] uses spatially-adaptive normalisation to condition on a segmentation map. The variant we adopt, joint image and mask synthesis from a shared generator, combines these ideas: the image and the mask are two coupled outputs of the same network, sharing internal representations.

The strengths of GANs are sharp samples at moderate resolution, fast single-pass inference, well-tested regularisation, and a disentangled $\mathscr{W}$-space in style-based models. The disentangled latent space is particularly relevant for eyeglass generation, where frame shape and material appearance must be independently controllable.

The limitations are training instability, residual susceptibility to mode collapse, no exact likelihood, and no built-in inverse map from images to latents.

1.4. Diffusion Models

Denoising Diffusion Probabilistic Models [3] define a forward Markov chain that progressively corrupts a sample with Gaussian noise, and learn a reverse process that denoises back to the data distribution. Recent variants include classifier-free guidance [24], latent diffusion models such as Stable Diffusion [25], the deterministic DDIM sampler [26] and consistency models [27] that distil multi-step diffusion into a single-step student.

Diffusion models offer state-of-the-art sample fidelity on large-scale benchmarks, very stable training, no mode-collapse failure mode, and better mode coverage than GANs.

Their limitations are high computational cost for both training and sampling; awkward pixel-aligned auxiliary outputs; under-resourced on corpora of a few thousand images. The quantitative implications for the present task are detailed in Section 1.6.

1.5. Evaluation Metrics

Evaluating generative image models is hard because there is no single agreed measure of perceptual quality. The standard metrics are summarised below.

The Fréchet Inception Distance (FID) [20] models Inception-v3 [35] feature activations of real and generated images as multivariate Gaussians and computes the Fréchet distance:

$$\mathrm{FID} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \mathrm{tr}\Big(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{1/2}\Big). \tag{1.3}$$

Lower is better; for reference, Karras et al. [5] report FID 2.8–3.2 on FFHQ at 1024^2 with $\sim$70k training images.

The Inception Score (IS) [21] combines per-image classifier sharpness with marginal class-distribution diversity. Barratt and Sharma [22] showed that IS is unreliable for domains absent from ImageNet (such as ours), and we report it only for completeness.

LPIPS [23] is the ℓ_2 distance between deep VGG features of two images, and correlates well with human perceptual similarity. The mean pairwise LPIPS distance among generated samples measures diversity; the LPIPS distance between generated and nearest-neighbour real samples measures the absence of memorisation.

Generic metrics measure global distributional similarity but do not check whether a generated image is *physically plausible* in a specialised domain. For eyeglass frames, three additional metrics are introduced in this thesis: Edge Coherence, Mask Regularity and Frame Symmetry, described in detail in Section 4.3.

1.6. Analytical Evaluation of Candidate Architectures Against Domain Requirements

This subsection evaluates three candidate architectures against the specific requirements of the present task: generating 10,000 paired 256×512 eyeglass frame images with pixel-aligned binary segmentation masks, trained on a corpus of approximately 6,500 annotated samples, with generation time low enough for practical dataset construction.

Variational Autoencoders

The fundamental limitation of a VAE for this task lies in the interaction between its training objective and the visual structure of eyeglass frames. Frame boundaries are bimodal: each pixel in the boundary region either belongs to the frame or to the background, with a sharp transition between the two states. The Gaussian decoder in a standard VAE [2] is trained by maximising a pixel-wise log-likelihood (equivalently minimising MSE or L1 reconstruction loss). When the latent code maps to a region of the image where the true pixel value is bimodal, the optimal decoder output under MSE is the expected value of that bimodal distribution, which falls between the two modes. The result is a soft, blurred boundary rather than a crisp edge.

This blurring is not merely an aesthetic defect. The paired segmentation mask must align precisely with the frame boundary; a soft boundary in the generated image makes the corresponding mask ambiguous. If the mask is generated jointly (sharing the same decoder), the softness propagates into the mask channel as well, producing non-binary or misaligned masks that are unusable for downstream segmentation training without post-processing. Hierarchical VAEs such as NVAE [29] partially mitigate blurring through deep latent hierarchies, but at computational cost that approaches or exceeds GAN training while still not fully eliminating the averaging artefact on high-frequency boundaries.

Denoising Diffusion Probabilistic Models

DDPM [3] achieves state-of-the-art sample fidelity on large-scale benchmarks but presents two structural mismatches with the present task.

A standard DDPM with $T = 1000$ denoising steps requires approximately 2 s per 256×512 sample on a single consumer GPU. Generating the target 10,000 paired samples therefore requires on the order of 5.5 hours of GPU time. A conditional StyleGAN2 generator produces a paired image and mask in a single forward pass at approximately 0.05 s per sample, completing the same 10,000-sample generation in under 9 minutes. Accelerated samplers such as DDIM [26] reduce the gap but still require 50–200 function evaluations per sample, remaining one to two orders of magnitude slower than a single-pass generator.

Producing a pixel-aligned binary mask alongside the RGB image requires one of two approaches: (a) running a separate segmentation network on the generated image, which intro-

duces alignment error because the segmentation model was not trained on synthetic artefacts, or (b) jointly denoising a 4-channel (RGB + mask) tensor, which doubles the effective channel count and increases training cost while requiring the diffusion process to learn a joint distribution over visually dissimilar modalities (continuous colour vs. binary mask). Neither option is straightforward, and published paired-output diffusion pipelines typically rely on large-scale pre-training that is unavailable in our setting.

Latent Diffusion Models

Latent diffusion [25] addresses the computational cost of pixel-space diffusion by operating in a compressed latent space learned by a separate autoencoder. However, the quality of this latent space depends critically on the richness of the autoencoder training data. The original Stable Diffusion model was trained on LAION-5B, a corpus of several billion image-text pairs. With only 6,500 training images of a single object category, the autoencoder cannot learn a latent manifold that captures the full variation of frame geometry, material and colour. The resulting latent space would be poorly structured, leading to artefacts in the decoded outputs. Furthermore, the two-stage training pipeline (autoencoder then diffusion model) doubles the hyperparameter-tuning burden on a limited dataset.

Requirements Compliance Matrix

Table 1.1 summarises the analytical evaluation of the three candidate architectures against the five key requirements of the task.

Table 1.1. Analytical compliance of candidate architectures with task requirements.

Requirement	**VAE**	**Diffusion / Latent Diffusion**	**Conditional StyleGAN2**
Sharp frame edges at 256×512	Blurred by MSE averaging over bimodal boundaries	Sharp after many denoising steps	Sharp; adversarial loss penalises any softness
Pixel-aligned paired mask	Joint decoder feasible but softness propagates to mask channel	Requires 4-channel joint denoising or separate segmentor	Native: shared generator outputs RGB + mask in one pass
Training on $\sim$6,500 images	Trains stably but under-fits fine detail at this scale	Latent-space autoencoder under-fits; pixel-space DDPM trains slowly	StyleGAN2-ADA designed for limited data; documented on 5k–10k corpora
Generation of 10,000 samples in practical time	Fast single-pass ($\sim$0.05 s/sample)	Slow: $\sim$2 s/sample (DDPM) to $\sim$0.3 s (50-step DDIM)	Fast single-pass ($\sim$0.05 s/sample)
Controllable diversity via latent manipulation	Structured latent space; limited sharpness of outputs	Guidance mechanisms exist but require classifier or text encoder	Disentangled $\mathcal{W}$-space supports style mixing and truncation

We therefore proceed with conditional StyleGAN2 in the methodology of Chapter 2; the full justification is consolidated in the chapter generalisation in Section 1.8.

1.7. Synthetic Data and Generative Models in the Eyewear Domain

Synthetic data has been used to extend or replace real training corpora in many vision applications. Frid-Adar et al. [31] showed that DCGAN-generated liver-lesion images improved CNN classifier accuracy. Wood et al. [32] trained competitive face-landmark detectors on entirely synthetic faces. Karras et al. [6] demonstrated that StyleGAN2-ADA trains useful generators on as few as 5,000 images per domain.

The eyewear-specific generative-modelling literature is sparser and mostly concentrates on virtual try-on rather than dataset construction. GlassesGAN [43] performs personalised try-on through StyleGAN inversion but does not produce stand-alone frame images or segmentation masks. At Vilnius Tech, Gisleris and Serackis [44] reconstruct 3-D textured frames from single RGB views; Giedra et al. [45] train supervised eyewear detection and segmentation networks. To the best of our knowledge, no prior published work has addressed the joint synthesis of front-facing eyeglass frame images with paired binary segmentation masks at production resolution, which is what the present thesis attempts.

1.8. A Brief Generalisation of the Chapter

The reviewed literature supports five conclusions that drive the methodological choices made in the rest of this thesis.

1. VAEs are not suitable as the primary dataset generator for our resolution and domain, because their inherent softness loses the high-frequency edge detail that downstream segmentation and try-on training need. Hierarchical VAEs narrow the gap to GANs but at substantial cost. We exclude VAEs from the experimental investigation on this analytical basis.
2. Diffusion models are mature but not the right fit for this task. Their advantages (mode coverage and stable training) show up mainly at large data scales. Their disadvantages (expensive inference and awkward production of pixel-aligned auxiliary outputs) are particularly costly here, where the target is 10,000 paired image–mask samples on a corpus of about 6,500 training images.
3. Conditional StyleGAN2 is the right starting point for the experimental investigation. Its style-based generator produces sharp samples at our target resolution; the weight-demodulated convolutions and path-length regularisation eliminate the most common artefacts of earlier style-based generators; the disentangled $\mathcal{W}$-space supports interpolation-based diversification of the generated dataset; and stable training is well documented when the architecture is paired with hinge loss, R1 penalty and (where ap-

plicable) ADA.

4. Joint image–mask synthesis from a single forward pass is the right conditional formulation. It enforces geometric consistency through shared representations and eliminates a separate mask-annotation stage.
5. Evaluation must combine generic and domain-specific metrics, because generic distributional measures do not capture the physical plausibility of generated eyeglass frames.

These conclusions define the methodology developed in Chapter 2 and applied experimentally in Chapters 3 and 4.

2. Application of Generative AI Methods for Building a Dataset of Eyeglass Frame Images

This chapter describes the proposed solution: a conditional StyleGAN2 architecture adapted to the eyewear domain that produces a 256×512 RGB image of an eyeglass frame and a pixel-aligned binary segmentation mask from a single forward pass. The chapter formalises the problem, lists the design requirements, justifies the architectural choice, describes generator and discriminator in detail, defines the composite loss function, gives the training procedure, and explains the five-mode generation strategy used to build the final synthetic dataset.

2.1. Problem Formulation and Design Requirements

Let $\mathscr{X} = \{(\mathbf{x}_i, m_i)\}_{i=1}^{n}$ be a dataset of n paired samples, with $\mathbf{x}_i \in \mathbb{R}^{3\times H\times W}$ a colour image of an eyeglass frame and $m_i \in \{0,1\}^{1\times H\times W}$ the matching binary mask, where $H = 256$ and $W = 512$. We want to train a generator G such that, given a noise vector $\mathbf{z} \in \mathbb{R}^{d_z}$ and a style index $s \in \{1,\ldots,n\}$,

$$G(\mathbf{z}, s) = (\hat{\mathbf{x}}, \hat{m}) \tag{2.1}$$

is drawn from a distribution close to the joint $p(\mathbf{x}, m)$. Once G is trained, we sample from it to build a synthetic dataset $\mathscr{X}_{\text{syn}} = \{(\hat{\mathbf{x}}_k, \hat{m}_k)\}_{k=1}^{N}$ with $N = 10{,}000 \gg n$ that is usable directly for downstream segmentation, virtual try-on and recommendation training.

The analytical review of Chapter 1 identified five design requirements:

1. Native 256×512 output without distortion, matching the 2:1 ratio of frontal catalogue photographs.
2. A single forward pass must emit both the RGB image and a pixel-aligned binary mask, so we never need to annotate generated images by hand.
3. The architecture and loss must be robust on $\sim$6,500 training samples, with explicit safeguards against mode collapse and over-fitting.
4. The final dataset must cover the full range of frame shapes, materials and colours present in the training data, with low intra-class repetition.
5. Training and inference must run on a single GPU within reasonable wall-clock time.

2.2. Justification of the Selected Methodology

Three candidate architectures were considered: a hierarchical VAE, a conditional latent diffusion model, and a conditional StyleGAN2. The detailed analytical evaluation of Section 1.6 compared each candidate against five design requirements and concluded that conditional StyleGAN2 is the only architecture that is excellent on sharp output, joint mask generation and inference cost, while at least good on training stability and diversity with appropriate regularisation

and a designed generation strategy. We therefore proceed with conditional StyleGAN2.

2.3. Generator Architecture

The generator combines a learned style-embedding table, a four-layer mapping network, a style-modulated synthesis network with self- attention, and a U-Net mask decoder. Figure 2.1 gives a high-level diagram; the components are described below.

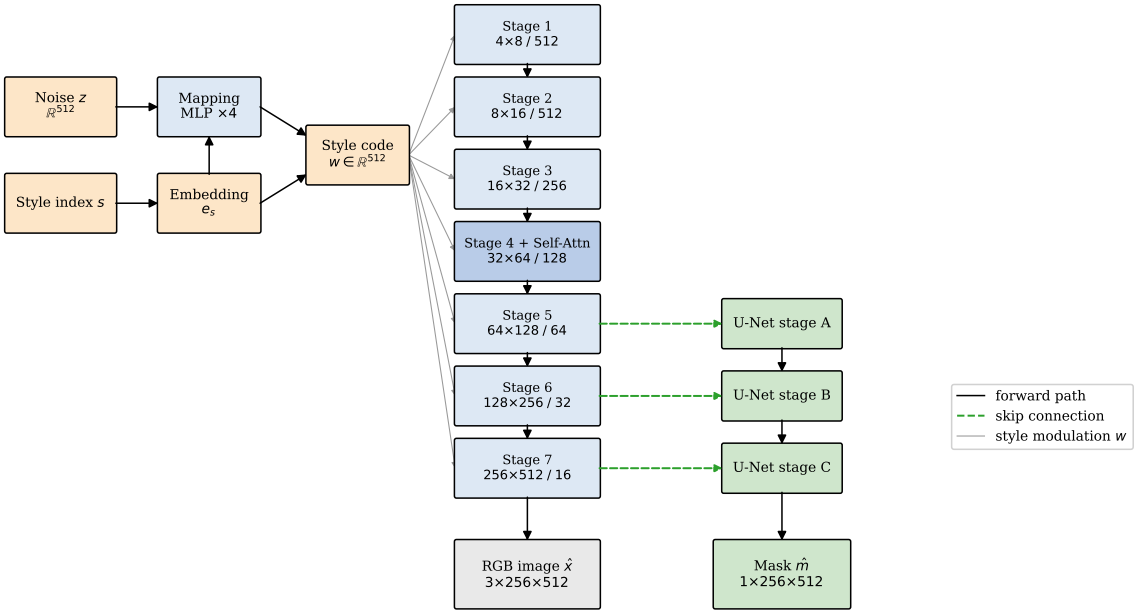

Fig. 2.1. Block diagram of the proposed generator. The mapping network takes the noise vector $\mathbf{z}$ and the style embedding $\mathbf{e}_s$ and produces a style code $\mathbf{w}$ that is fed into all synthesis stages. The synthesis network produces the RGB image; the U-Net mask decoder reads intermediate synthesis features through skip connections (dashed lines) and produces the binary segmentation mask.

2.3.1. Style Embedding and Mapping Network

A learned embedding table $\mathbf{E} \in \mathbb{R}^{n \times d_w}$ holds one row per training image, with $d_w = 512$. During training, the i-th row is the style code of the i-th training image, indexed by s. The table is initialised from $\mathcal{N}(0, 0.5^2)$ and learned end-to-end with the rest of the network.

The mapping network is a four-layer MLP with LeakyReLU activations and an input pixel-norm (the effect of depth is evaluated in the ablation study of Section 4.9). It takes the concatenation of the ℓ_2-normalised noise vector $\mathbf{z}$ and the style embedding $\mathbf{e}_s = \mathbf{E}[s, :]$ and outputs a style code $\mathbf{w} \in \mathbb{R}^{512}$:

$$\mathbf{w} = f_{\text{MLP}}(\text{concat}(\text{norm}(\mathbf{z}), \mathbf{e}_s)) \,. \tag{2.2}$$

2.3.2. Synthesis Network

Synthesis starts from a learned $1 \times 512 \times 4 \times 8$ constant tensor and applies six progressive upsampling stages. Each stage doubles the spatial resolution and halves the channel count after the first few stages, with channel progression $512 \to 256 \to 256 \to 128 \to 64 \to 32 \to 16$.

Each stage contains two *styled convolution blocks*. A styled block performs a 3×3 convolution whose filter weights are modulated per-sample by an affine transform of $\mathbf{w}$ and then demodulated to unit RMS:

$$W'_{ijk} = \frac{s_i \cdot W_{ijk}}{\sqrt{\sum_{i',k'} (s_{i'} \cdot W_{i'jk'})^2 + \varepsilon}}, \tag{2.3}$$

where $s = A\mathbf{w} + b$ is the per-channel modulation. After the convolution, a single-channel Gaussian noise map is added with a per-channel learned weight (initialised to zero, the standard StyleGAN noise injection). A LeakyReLU activation with negative slope 0.2 follows.

Every two styled blocks the current feature map is projected to a 3-channel image via a modulated 1×1 convolution, and added to the bilinearly upsampled previous-stage output (the StyleGAN2 *skip-RGB* accumulation [5]).

A self-attention module [17] is inserted after the third synthesis stage at the 32×64 resolution. This is critical here because eyeglass frames are bilaterally symmetric: the left lens region must match the right, and the bridge that connects them is a global feature. Attention is applied at 32×64 rather than at full resolution because the cost of the attention map scales as $O((HW)^2)$.

The six stages progress from coarse spatial layout (Stages 1–2) through material and bridge detail (Stage 4) to pixel-level edge sharpness (Stages 5–6). The self-attention layer at Stage 3 (32×64) is critical: it enforces bilateral symmetry across the full spatial extent, which is the defining geometric property of eyeglass frames. This coarse-to-fine decomposition disentangles the style code: at coarse stages it controls *geometric* attributes (frame shape, lens proportions), while at fine stages it controls *appearance* attributes (material, colour, glossiness).

2.3.3. U-Net Mask Decoder

The mask decoder is a U-Net branch [33] attached to the synthesis network. It receives skip connections from the four shallowest synthesis stages with channel widths $128, 64, 32, 16$. At each decoder step the current feature map is bilinearly upsampled, concatenated with the next skip tensor, and refined by a pair of convolutional layers with instance normalisation and LeakyReLU. A final 1×1 projection followed by a sigmoid produces the predicted mask $\hat{m} \in [0,1]^{1 \times 256 \times 512}$.

Because image and mask share internal feature representations, the mask naturally aligns with the generated frame without an explicit alignment loss, and no manual annotation is needed at generation time.

2.4. Discriminator Architecture

The discriminator (Figure 2.2) must judge image realism, mask plausibility and image–mask consistency together. To do this it receives a 5-channel tensor:

$$\mathbf{u} = \text{concat}(\mathbf{x}, m, \boldsymbol{\sigma}_s) \in \mathbb{R}^{5\times256\times512}, \tag{2.4}$$

where $\boldsymbol{\sigma}_s$ is a 1-channel *spatial style map* produced by projecting $\mathbf{e}_s$ through a small MLP onto a low-resolution map and bilinearly upsampling it. The spatial style map gives the discriminator information about the requested style, allowing it to detect mismatches between style and content.

The backbone is five residual downsampling blocks, each with two 3×3 convolutions, LeakyReLU activations and a learned skip projection. Spectral normalisation [14] is applied to all convolutions and linear layers, providing a low-cost Lipschitz constraint that stabilises training. Self-attention is inserted at the 16×32 resolution. After the five blocks, a minibatch standard-deviation layer appends a single feature encoding the per-channel standard deviation across the batch, a global signal about diversity that helps detect mode collapse. A final 4×8 convolution flattens the feature map and a fully-connected head outputs a single real-valued score per sample.

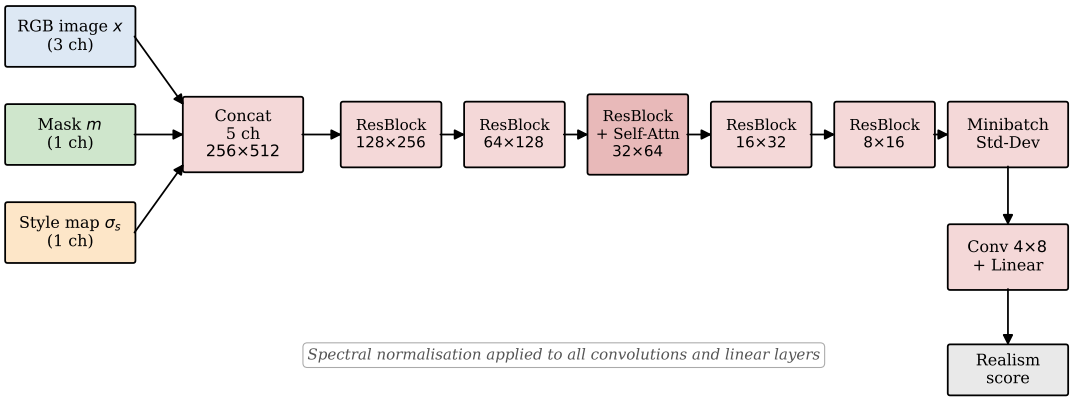

Fig. 2.2. Block diagram of the conditional discriminator. The 5-channel input concatenates the RGB image, the binary mask and a 1-channel spatial style map; five residual downsampling blocks (the third with self-attention) reduce spatial resolution while a minibatch standard-deviation feature is appended before the final score head.

2.5. Composite Loss Function

The generator and discriminator are trained with a composite objective combining four terms.

2.5.1. Hinge adversarial loss.

We use the hinge formulation [15], which is more stable than the original cross-entropy:

$$\mathscr{L}_D = \mathbb{E}_{(\mathbf{x},m,s)}[\mathrm{ReLU}(1 - D(\mathbf{x},m,s))] + \mathbb{E}_{\mathbf{z},s}[\mathrm{ReLU}(1 + D(\hat{\mathbf{x}},\hat{m},s))], \tag{2.5}$$

$$\mathscr{L}_G^{\mathrm{adv}} = -\mathbb{E}_{\mathbf{z},s}[D(\hat{\mathbf{x}},\hat{m},s)]. \tag{2.6}$$

2.5.2. R1 gradient penalty.

The R1 penalty [16] stabilises early training:

$$\mathscr{L}_{R_1} = \frac{\gamma}{2}\,\mathbb{E}_{\mathbf{x}}[\|\nabla_{\mathbf{x}} D(\mathbf{x},m,s)\|_2^2], \quad \gamma = 10. \tag{2.7}$$

Both $\gamma = 10$ and the lazy-R1 schedule (applying the penalty once every 16 discriminator updates) follow the StyleGAN2 defaults [5].

2.5.3. Mask supervision.

Mask quality is supervised by a sum of focal binary cross-entropy [18] and soft Dice loss:

$$\mathscr{L}_{\mathrm{mask}} = \mathscr{L}_{\mathrm{focal}}(\hat{m},m) + \mathscr{L}_{\mathrm{dice}}(\hat{m},m), \tag{2.8}$$

$$\mathscr{L}_{\mathrm{dice}}(\hat{m},m) = 1 - \frac{2\sum \hat{m}m + \varepsilon}{\sum \hat{m} + \sum m + \varepsilon}. \tag{2.9}$$

Focal loss focuses gradient on hard pixels (the frame boundary); soft Dice handles the strong class imbalance between frame and background.

2.5.4. Perceptual loss.

A frozen VGG-16 [19] provides feature-matching supervision in `relu1_2` and `relu2_2`:

$$\mathscr{L}_{\mathrm{perc}} = \sum_{\ell} \|\phi_\ell(\hat{\mathbf{x}}) - \phi_\ell(\mathbf{x})\|_1. \tag{2.10}$$

Operating in feature space captures texture and edge similarity in a way that is invariant to small misalignments and correlates better with human perception than pixel MSE.

2.5.5. Total generator objective.

The total generator loss is

$$\mathscr{L}_G^{\mathrm{total}} = \mathscr{L}_G^{\mathrm{adv}} + \lambda_m \mathscr{L}_{\mathrm{mask}} + \lambda_p \mathscr{L}_{\mathrm{perc}}, \quad \lambda_m = 15,\ \lambda_p = 1.5. \tag{2.11}$$

The high $\lambda_m = 15$ prioritises mask boundary geometry early in training; 84% of the total mask-loss reduction occurs in the first 50 epochs (Section 4.2), confirming that this weighting has the intended effect.

2.6. Training Procedure

Both networks are optimised with Adam [37] using $\beta_1 = 0$, $\beta_2 = 0.99$, generator learning rate 2×10^{-4} and discriminator learning rate 1×10^{-4}, following the StyleGAN2 configuration [5].

Batch size is 32, which fits in 24 GB of VRAM at 256×512. Gradient norms are clipped to 1.0 before each optimiser step, eliminating rare but catastrophic spikes.

An exponential moving average (EMA) of generator weights [5]

$$\bar{\theta}_t = \rho \bar{\theta}_{t-1} + (1-\rho)\theta_t, \quad \rho = 0.999 \tag{2.12}$$

is maintained throughout training. The EMA weights are used for all inference (sample generation, dataset construction, qualitative evaluation), while the raw weights continue to be updated by the optimiser. The EMA acts as a form of post-hoc smoothing and consistently produces sharper, less artefact-laden outputs than the raw weights.

Training runs for 350 epochs at batch size 32 on the ∼6,500-image corpus. Checkpoints of both networks, the EMA, and the optimiser states are saved every 10 epochs. Training is monitored on the discriminator's hinge loss on real and fake samples separately, the generator's adversarial loss, the mask focal+Dice loss, the perceptual VGG loss, and proxy IoU and Dice scores. Training is stopped when both the adversarial loss and the reconstruction losses have plateaued for at least 80 epochs and visual inspection of generated samples confirms high and consistent quality.

The model is implemented in PyTorch 2.x [38]. Training runs on a single NVIDIA A100 (40 GB VRAM). The dataset loader uses 2 worker processes and pinned host memory to keep the GPU saturated. The empirical wall-clock cost is reported in Chapter 4.

2.7. Reused Components and New Contributions

The following components are reused: PyTorch [38] as the deep learning framework; pretrained Inception-v3 from `torchvision` for FID feature extraction; pretrained VGG-16 for the perceptual loss; the LPIPS implementation from the official package [23]; OpenCV's `Canny` [36] for edge detection used in the Edge Coherence metric; and the standard Adam optimiser [37].

The following components are new, built for this thesis: the non-square synthesis network adapted for 256×512 output; the U-Net mask decoder fed by skip connections from the synthesis network; the 5-channel discriminator that includes a spatial style map; the composite loss formulation in Section 2.5; the five-mode generation strategy in Section 2.8 below; and the three

domain-specific evaluation metrics introduced in Section 4.3 of Chapter 4.

2.8. Dataset Generation Strategy

A trained GAN can in principle be sampled by drawing fresh noise vectors and style indices, applying G, and saving the outputs. Naive sampling, however, tends to produce a dataset whose effective diversity is bounded by the diversity of the training set, because each style index reproduces a specific training style. For dataset construction this is wasteful: the goal is precisely to expand variety beyond what is in the training corpus.

We use a five-mode style-interpolation strategy with a budget of $N = 10{,}000$ samples. Pure mode (25%) uses a single training style, with styles sampled uniformly at random from the training set. Interp2 mode (40%) blends two styles with weights $(\alpha, 1-\alpha)$ where $\alpha \sim \text{Beta}(0.4, 0.4)$; the Beta distribution with concentration below 1 is bimodal, favouring blends close to a pure style rather than uniform 50/50 mixes, which empirically produces more visually distinct hybrids. Interp3 (20%) and Interp4 (10%) blend three and four styles respectively using Dirichlet-distributed weights with concentration parameters 0.7 and 0.5. Finally, Intra mode (5%) applies the same style with multiple independent noise vectors, testing the intra-style diversity contributed by the noise injection layers.

A truncation trick [4] with $\psi = 0.7$ is applied to every sample:

$$\mathbf{w}' = \bar{\mathbf{w}} + \psi(\mathbf{w} - \bar{\mathbf{w}}), \tag{2.13}$$

where $\bar{\mathbf{w}}$ is the mean of the $\mathcal{W}$-space. Truncation pulls each sample toward the population mean and trades a small amount of diversity for substantial improvement in the quality of the most extreme samples. The value $\psi = 0.7$ is the StyleGAN2 default for high-quality production sampling.

2.9. A Brief Generalisation of the Chapter

The chapter described the conditional StyleGAN2 architecture proposed for joint synthesis of eyeglass frame images and segmentation masks. Joint image–mask generation through a U-Net decoder attached to the synthesis network ensures pixel alignment without an explicit alignment loss and removes the manual annotation step. Style-embedding-conditioned generation with a learned table and a mapping network produces a disentangled $\mathcal{W}$-space that supports the dataset-diversification strategy, while self-attention at 32×64 provides a global receptive field at manageable spatial cost. The composite loss (hinge, R1, focal+Dice mask and VGG perceptual terms) enforces realism, training stability, mask accuracy and texture fidelity, with the high mask weight $\lambda_m = 15$ reflecting the diagnostic role of the frame boundary. The spectral-normalised 5-channel discriminator with self-attention and minibatch standard-deviation evaluates realism, mask plausibility and image–mask consistency jointly. Finally, the five-mode

style-interpolation strategy with truncation $\psi = 0.7$ expands diversity beyond the training distribution while preserving quality.

The next chapter describes the data preparation and the compositing pipeline used to build the training and evaluation corpora for the experimental investigation.

3. Data Preparation and Experimental Setup

This chapter describes how the data used in the experimental investigation was assembled. It covers the GAN training corpus, the auxiliary data used for the downstream task evaluation (CelebAMask-HQ face subsets), and the face compositing pipeline that connects them. The chapter ends with a description of the segmentation training and evaluation setup used in the downstream task.

3.1. Dataset Uniqueness and Collection Methodology

A central premise of this work is that no existing public dataset satisfies the requirements of a conditional GAN trained to produce isolated eyeglass frame images with pixel-aligned segmentation masks.

3.1.1. Absence of a suitable public dataset.

Several face-centric datasets include eyeglasses as an incidental attribute, but none provides the combination of properties required for our task.

CelebAMask-HQ [40] annotates 30,000 face images at 1024×1024 resolution with 19 semantic classes, one of which is "eyeglasses." However, those masks delineate the glasses region within a face photograph rather than an isolated frame. The glasses occupy a small fraction of the total image area, so the effective resolution of the frame region is low. More critically, the images are not catalogue-style: frames appear under variable lighting, at diverse head poses, partially occluded by hair or hats, and against cluttered backgrounds. No background-free, front-facing frame image can be extracted from CelebAMask-HQ without heavy manual editing.

The Labeled Faces in the Wild (LFW) dataset and the Flickr Faces HQ (FFHQ) dataset both contain face photographs in which some subjects wear glasses, but neither provides segmentation masks for the frame specifically. FFHQ images are high-resolution (1024×1024), yet the glasses are still embedded in a face context and cannot be isolated without a separate segmentation step that would itself require training data.

To train a conditional GAN that outputs an isolated frame image together with its binary mask, the training data must satisfy four properties simultaneously: isolated, front-facing frame photographs with no face visible; pixel-aligned binary segmentation masks of the frame outline; a uniform or near-uniform background; and a sample count exceeding approximately 5,000 to provide sufficient diversity for GAN convergence. No publicly available dataset meets all four criteria. This gap motivates the use of a proprietary partner dataset of catalogue product photographs, which satisfies every requirement by construction.

3.1.2. Collection methodology.

The proprietary partner dataset consists of catalogue product photographs of eyeglass frames, originally captured for e-commerce use and supplied to the project together with the corresponding binary segmentation masks. The preparation pipeline that produced the supplied dataset, as documented by the partner, comprised several stages.

Raw images varied in resolution, aspect ratio and background tone. Each image was centre-cropped to a 2:1 aspect ratio (width to height), then resized to 256×512 pixels using bicubic interpolation for the RGB channel and nearest-neighbour interpolation for the mask channel to preserve hard label boundaries.

The supplied segmentation masks delineate the frame outline, including temples (arms), hinges, bridge and rims. Because eyeglass frames are hollow objects, the mask represents only the opaque material of the frame, not the transparent lens area. For frames with opaque or tinted lenses, the lens region was annotated as a separate class and merged with the frame mask via morphological closing to produce a single connected component per frame.

We rejected images exhibiting any of the following: photographic watermarks or retailer overlays superimposed on the image (brand markings engraved or printed on the temple arms as part of the physical frame were retained, since they are intrinsic to the product); extreme camera angles departing more than approximately 15° from the frontal plane; non-standard framing such as close-ups of a single temple or hinge; and images with colour casts or heavy post-processing artefacts that would confuse the GAN's colour distribution.

The most persistent residual issue concerns thin wire-frame and rimless styles. Their frame material occupies only a few pixels in width at the target resolution, making consistent mask annotation difficult. Small annotation errors that would be negligible on a thick acetate frame become proportionally significant on a wire frame, where the mask may be only two to four pixels wide. This issue is partially mitigated by the morphological post-processing applied during mask preparation, but residual boundary noise remains concentrated on these thin styles.

A second challenge was ensuring diversity across frame categories. Commercial catalogues are biased toward popular styles (rectangular acetate frames in dark colours), which would produce an imbalanced training set if sampled uniformly. Manual balancing was applied to ensure representation across style categories including rectangular, round, oval, cat-eye, aviator, browline and rimless designs, as well as across material types and colour palettes.

3.1.3. Possible dataset modifications and extensions.

The current dataset is restricted to frontal catalogue images of optical frames (prescription eyeglasses). Several extensions could broaden its applicability in future work.

Adding three-quarter-view or profile photographs would enable a pose-aware GAN that generates frames at multiple orientations. This would reduce the reliance on the compositing pipeline

for pose matching, as the GAN could directly produce a frame at the target head pose. However, multi-view catalogue images are less standardised and would require view-angle annotation as an additional conditioning label.

The current dataset excludes sunglasses because tinted and mirrored lenses create reflections and colour distributions that differ substantially from optical frames. Including sunglasses would broaden the domain but would likely require a lens-type conditioning label to prevent mode collapse between transparent and opaque lens styles.

Other possible extensions include stratified sampling by frame thickness to reduce annotation-noise bias, and the inclusion of on-face product photographs directly in the training set.

3.2. GAN Training Corpus on Disk and at Training Time

The approximately 6,500 paired samples retained after the collection methodology described above form the GAN training corpus. On disk, all masks are stored as 1-channel PNG files with pixel values in $\{0, 255\}$, spatially aligned with their corresponding 256×512 RGB images.

One on-the-fly augmentation is applied to the RGB image only: colour jitter with brightness $\pm 15\%$, contrast $\pm 15\%$, saturation $\pm 15\%$ and hue $\pm 5\%$, applied independently per sample. The mask is not colour-jittered. Geometric augmentations such as horizontal flip, rotation and elastic deformation are deliberately not used, as they would produce orientations and shapes outside the natural distribution of front-facing catalogue frames.

3.3. Auxiliary Data Sources for the Downstream Task

The GAN training and intrinsic evaluation use only the catalogue corpus above. The downstream segmentation evaluation, however, uses three additional data sources.

A curated subset of approximately 11,500 frontal face photographs from CelebAMask-HQ [40] is used as the face-compositing substrate. The subset is obtained in two filtering stages: (i) the dataset's attribute annotation file is used to retain only subjects not wearing eyeglasses ($\texttt{Eyeglasses} = -1$); (ii) the MediaPipe Face Landmarker [41] is then run on each remaining image, and only images for which all six canonical pose-anchor landmarks (eye outer corners, nose bridge, nose tip, mouth corners) are detected reliably are kept, so that the compositing pipeline of Section 3.4 is guaranteed to find a valid head pose for every substrate image.

The held-out test corpus for the downstream evaluation consists of 1,010 1024×1024 face photographs from CelebAMask-HQ [40] in which the subject is wearing eyeglasses, each accompanied by a hand-annotated ground-truth glasses mask. Bounding boxes around the glasses region were additionally annotated using the labelme tool [50] to support the region-restricted "Case 2" evaluation. None of the 1,010 test images is used in any training condition. CelebAMask-HQ is publicly available, so results are directly comparable to any future work in this area.

There is no image overlap between the CelebAMask-HQ face substrate (no glasses) and the CelebAMask-HQ test corpus (with glasses), as the two subsets are defined by mutually exclusive attribute labels.

A second held-out test set is used for cross-dataset validation: a 500-image eyeglasses segmentation dataset published on Roboflow [48]. The images originate from a different source than CelebAMask-HQ and were annotated in JSON polygon format by an independent team; the polygons are converted to binary masks for evaluation and the images are resized to 640×640. This dataset is never seen during training and is used in Section 4.7 to verify that the downstream findings generalise beyond CelebAMask-HQ.

3.4. Compositing Pipeline

Catalogue images of eyeglass frames are taken against uniform light backgrounds. A segmentation model trained directly on such images learns to associate "frame" pixels with sharp colour transitions on a near-white background, a feature distribution that does not exist in real face photographs. The result, as confirmed empirically in Section 4.6, is catastrophic failure at test time. To bridge this catalogue-to-real domain gap we composite each catalogue image (whether real or GAN-generated) into a more realistic background.

3.4.1. Perspective-correct face compositing.

The compositing pipeline (Figure 3.1) places the frame on a frontal face photograph at the correct head pose. Head pose is recovered from the face photograph using the MediaPipe Face Landmarker model [41], which extracts 478 facial landmarks. Six canonical points (left and right eye outer corners, nose bridge, nose tip, left and right mouth corners) serve as 2-D anchors for pose estimation. These six pixel coordinates, paired with their canonical 3-D positions in a face-centric millimetre coordinate system, are passed to OpenCV's `solvePnP` [42], which returns the rotation vector $\mathbf{r}$, the translation $\mathbf{t}$ and an estimated camera intrinsic K. A sanity check rejects pose solutions with negative depth or implausible projection error.

Given the recovered pose, the frame's four corners are placed in the same face-centric coordinate system: width proportional to the inter-pupillary distance, height locked to the frame's original aspect ratio (so frames are never stretched) and a forward Z offset of $\sim$1 mm so the frame sits slightly in front of the eye plane. These 3-D corners are projected to pixel coordinates with `projectPoints`, producing a four-point quadrilateral that represents the frame's foreshortened shape on the target face.

A homography is fitted between the source frame's rectangular boundary and the projected quadrilateral, and `warpPerspective` maps the frame image onto the face photograph; the same homography is applied to the mask with nearest-neighbour interpolation so that image and mask remain pixel-aligned. The warped frame is then alpha-blended onto the face with Gaussian edge softening to reduce cut-paste boundary artefacts.

Because the face substrate has already been pre-filtered for reliable MediaPipe landmark detection (Section 3.3), the `solvePnP` sanity checks are essentially never triggered at compositing time, and every catalogue frame is composited successfully. The pipeline yields one composite per frame: approximately 6,500 for the real source and 9,695 for the full GAN set. In conditions A_2/B_2 the GAN composites are capped to match the source count; in A_3/B_3 the full set is used.

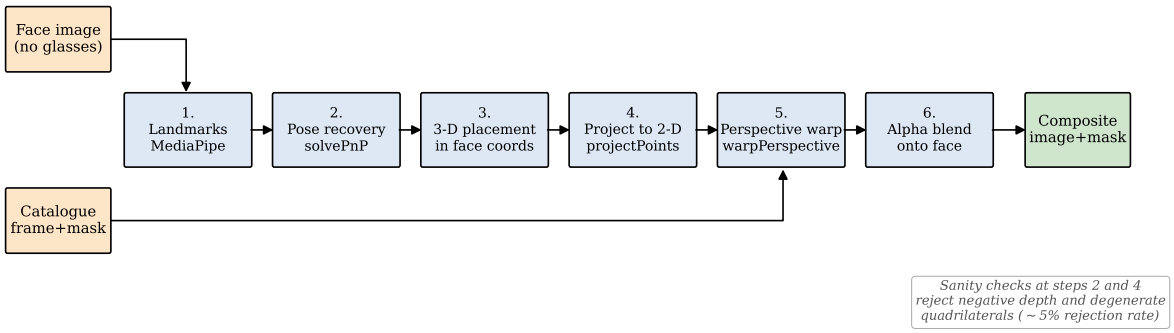

Fig. 3.1. Perspective-correct face compositing pipeline. Inputs (left) are a face image without glasses and a catalogue frame with mask. The pipeline recovers the head pose with MediaPipe landmarks and OpenCV `solvePnP`, projects the frame back to 2-D, and applies a homography warp to image and mask in lockstep before alpha-blending onto the face.

3.4.2. Software reuse and contributions.

The compositing pipeline is built almost entirely from existing high-quality components: MediaPipe Face Landmarker for landmark detection; OpenCV's `solvePnP`, `projectPoints` and `warpPerspective` for pose recovery and image warping; PIL/Pillow for alpha-blend composition. The code contributions of this thesis are: the choice of canonical 3-D anchor points; the inter-pupillary-distance-based width scaling and the forward Z offset; the sanity checks on the recovered pose; and the orchestration into a batch pipeline that produces matched image–mask pairs ready for downstream training.

3.5. Six Training Conditions for the Downstream Task

The downstream evaluation uses six training conditions that vary the image source (real vs. GAN-generated catalogue images) and compositing strategy (none vs. face composites), with an additional pair testing whether more GAN data beyond the size-matched cap improves performance. Table 3.1 summarises the conditions. Each condition trains a fresh U-Net segmentation model, with all hyper-parameters held constant. Every data source is split into training, validation and test subsets (10% validation, 10% test, capped at 300 each); the training sizes in Table 3.1 reflect this split. For example, the the 6,511 source catalogue images yield 5,911 training, 300 validation and 300 test pairs.

Table 3.1. Six training conditions for the downstream evaluation. Approximate training set size shown; all conditions use the same architecture, optimiser, schedule and augmentations.

Condition	Image source	Compositing strategy	Train size
A	Real (Source)	— (catalogue only)	5,911
B	Real (Source)	+ Faces	11,822
A_2	GAN (capped)	— (catalogue only)	5,911
B_2	GAN (capped)	+ Faces	11,822
A_3	GAN (full)	— (catalogue only)	9,095
B_3	GAN (full)	+ Faces	18,190

The six conditions are designed to answer three questions through paired comparisons. Comparing A vs. A_2 and B vs. B_2 isolates the effect of *image source* (real vs. GAN) with the GAN set capped to the same size as the real set, ensuring any difference reflects data quality rather than quantity. Comparing B_2 vs. B_3 tests whether additional GAN volume beyond the size-matched baseline provides further benefit. Comparing within each source group (A vs. B, A_2 vs. B_2, A_3 vs. B_3) isolates the effect of face compositing for a fixed image source.

3.6. Segmentation Architecture and Training Procedure

The segmentation architecture is a U-Net with an ImageNet-pretrained ResNet-34 [34] encoder, implemented through `segmentation_models.pytorch` [39]. The architecture is deliberately a well-established off-the-shelf network: the experimental question concerns the *training data*, not the segmentation architecture, and a known good encoder removes architectural confounders that would otherwise obscure the comparison between conditions.

The loss is an equally-weighted sum of soft binary cross-entropy and Dice loss applied to model logits.

Optimisation uses Adam with learning rate 1×10^{-4} and default β parameters, with cosine annealing from 1×10^{-4} to 1×10^{-6} over the full run.

Training runs for 20 epochs at batch size 32, image resolution 512×512, letterbox padding for non-square inputs. Best validation IoU checkpoint is retained for evaluation.

Augmentations include horizontal flip ($p = 0.5$); small-angle rotation ($\pm 15^\circ$, $p = 0.5$); random scale-and-crop within $[0.80, 1.00]$ of the input ($p = 0.5$); colour jitter (brightness, contrast, saturation, small hue, $p = 0.5$); mild Gaussian blur ($\sigma \in [0.1, 1.5]$, $p = 0.7$). Hue jitter is deliberately small (± 0.05); rotation is restricted to $\pm 15^\circ$ to avoid physically unrealistic poses.

The model has approximately 24 M trainable parameters. Each condition trains in roughly 30 minutes on a single consumer GPU (about 1.5 minutes per epoch); the exact hardware is reported in Section 4.1.

3.7. Downstream Evaluation Procedure

All six trained models are evaluated on the same fixed CelebAMask-HQ test corpus (1,010 images). Two metrics are reported per condition.

In Case 1 (full-image evaluation), the trained model is run on each 1024×1024 photograph (letterbox-padded and resized to 512×512 for inference, with the prediction mask resized back to ground-truth resolution). IoU, Dice, precision and recall are computed against the ground-truth mask over the entire image.

In Case 2 (region-restricted evaluation), both the full-image prediction and the ground-truth mask are cropped to the labelme-annotated bounding box around the glasses (with a 25% margin on each side), and the same metrics are computed within the cropped region. Case 2 isolates the model's performance on the part of the image that actually matters.

Case 2 is implemented as a region-restricted view of the Case 1 prediction, not as a separate inference on the cropped image. We discuss why in Section 4.8 of the next chapter.

3.8. Known Limitations of the Prepared Data

The prepared data has several limitations. All GAN training images come from contemporary commercial catalogues, so the generator can reproduce only the styles present in these collections and older, vintage or custom-made frames are under-represented. All training backgrounds are approximately uniform, which is why the compositing pipeline of Section 3.4 is needed to bridge the domain gap. The training set contains only frontal poses, with no three-quarter or profile views. Finally, residual annotation error in the training masks is concentrated at the frame boundary, especially for thin wire frames.

3.9. A Brief Generalisation of the Chapter

The chapter described how the data used in the experimental investigation was assembled. The GAN training corpus is $\sim$6,500 paired image–mask samples from a proprietary partner dataset, normalised to 256×512 at 2:1 aspect ratio, with binary masks provided as part of the curated dataset. Two auxiliary data sources support the downstream evaluation: CelebAMask-HQ faces without glasses, and the CelebAMask-HQ glasses test corpus. A perspective-correct face compositing pipeline bridges the catalogue-to-real domain gap by producing $\sim$6,500 real-source and $\sim$9,695 GAN-source face composites. Six U-Net segmentation training conditions cross image source against compositing strategy, all evaluated on the same held-out CelebAMask-HQ test corpus. The next chapter reports the experimental results.

4. Research Results

This chapter reports the experimental investigation in two parts. The first part (Sections 4.1 to 4.4) characterises the GAN training and the intrinsic quality of the synthetic dataset. The second part (Section 4.6) reports the downstream segmentation evaluation on the CelebAMask-HQ test corpus. The chapter ends with a discussion of trustworthiness and accuracy, and a brief generalisation of the chapter's findings.

4.1. Experimental Setup and Plan

All experiments were run on a single NVIDIA A100 (40 GB VRAM) for GAN training and an NVIDIA RTX 5090 (32 GB VRAM) for segmentation, under PyTorch 2.x with CUDA 12. Dataset loaders used pinned memory with 2 worker processes for GAN training and 8 for segmentation training.

The GAN training configuration is summarised in Table 4.1. All hyperparameters were fixed at the values reported in Chapter 2 and were not retuned on the validation results; this section therefore reports a single, complete training run rather than a hyperparameter search.

Table 4.1. GAN training configuration.

Parameter	Value
Resolution	256×512
Latent dimension d_z	512
Style embedding dimension d_w	512
Mapping network depth	4
Batch size	32
Optimiser	Adam ($\beta_1 = 0$, $\beta_2 = 0.99$)
Generator learning rate	2×10^{-4}
Discriminator learning rate	1×10^{-4}
R1 weight γ	10
R1 application interval	every 16 D updates
Mask loss weight λ_m	15
Perceptual loss weight λ_p	1.5
EMA decay ρ	0.999
Training epochs	350
Gradient norm clip	1.0

The investigation proceeds in four steps:

1. Train the GAN, monitor convergence by tracking six training-time quantities, and stop when both adversarial and reconstruction losses have plateaued.
2. Evaluate the synthetic dataset intrinsically against eight metrics: three generic, three domain-specific, two diversity.
3. Train six U-Net segmentation models on the conditions defined in Section 3.5, on the same hardware.

4. Evaluate the six models on the CelebAMask-HQ test corpus using both Case 1 (full image) and Case 2 (bounding-box restricted) variants.

These checks test whether the GAN converged and whether the resulting data is useful for downstream training.

4.2. GAN Training Dynamics

Mean epoch time was 182.5 seconds, giving a total wall-clock training time of approximately 17.7 hours over 350 epochs. Table 4.2 reports the values of the six tracked quantities at six representative epochs. Figure 4.1 plots the full training curves with their 20-epoch moving averages.

Table 4.2. Training metrics at representative epochs.

Epoch	G adv	D loss	Mask	Perc	D(real)	D(fake)
1	1.213	0.670	0.783	0.774	+1.700	−1.213
50	0.214	1.874	0.197	0.602	+0.108	−0.214
100	1.192	1.012	0.140	0.575	+0.859	−1.192
200	1.681	0.597	0.107	0.531	+1.336	−1.681
270	1.817	0.491	0.096	0.510	+1.504	−1.817
350	1.833	0.450	0.088	0.497	+1.571	−1.833

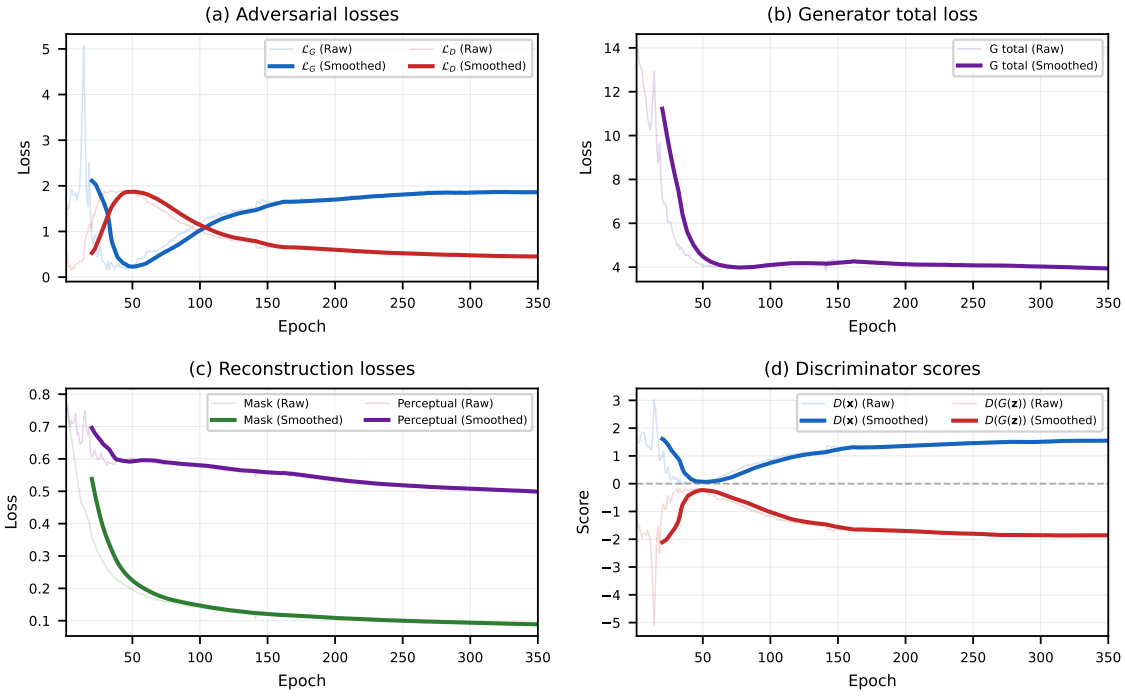

Fig. 4.1. Training dynamics over 350 epochs. Light grey curves are per-epoch values; bold colour curves are 20-epoch moving averages. (a) adversarial losses; (b) generator total loss; (c) reconstruction losses (mask and perceptual); (d) discriminator scores on real and fake samples.

The two reconstruction losses follow distinct trajectories. Mask loss decreases monotonically and substantially, from 0.783 at epoch 1 to 0.088 at epoch 350 (an 88.8% reduction), with 84%

of the total reduction occurring in the first 50 epochs. This rapid early convergence is consistent with the high mask weight $\lambda_m = 15$, which encourages the model to learn the frame boundary geometry before the synthesis branch wastes capacity on textures whose spatial support is wrong. Perceptual loss, by contrast, decreases gradually from 0.774 to 0.497 (a 35.8% reduction), reflecting its softer role: it does not impose a hard constraint but guides the generator toward textures and edges that match the real distribution in feature space.

The adversarial losses settle into a stable equilibrium. The discriminator's score on real samples stabilises near $+1.57$ and on fake samples near -1.83 at epoch 350. The hinge loss enforces a margin of 2 between real and fake scores; the empirical margin of $|+1.57| + |-1.83| = 3.40$ is comfortably above this threshold, which means the discriminator has not collapsed but is also not overpowering the generator. The brief dip around epoch 50 self-corrects by epoch 100, consistent with expected behaviour under hinge loss with R1 regularisation.

The mask IoU oscillates between 0.33 and 0.36 for the second half of training, and Dice between 0.48 and 0.51. These are intentionally treated as *proxy* metrics because the GAN does not produce masks spatially aligned with any specific real mask: each generated mask corresponds to a fresh sample drawn from the joint distribution. The IoU is computed against the mask of the sample whose style index was used, but the noise vector $\mathbf{z}$ is independent, so even a perfect generator would not yield IoU near 1.0. Stability around a fixed level is the meaningful signal here, and that stability is present.

Training was stopped at epoch 350 based on three criteria: (i) the perceptual loss changed by less than 3% between epochs 270 and 350; (ii) $D(\text{real})$ and $D(\text{fake})$ tracked their epoch-350 values for the preceding 80 epochs; (iii) visual inspection of sample grids from each checkpoint after epoch 270 showed no further improvement.

4.3. Intrinsic Dataset Evaluation: Metrics

After training, we used the EMA generator to produce a synthetic dataset of 10,000 paired image–mask samples at 256×512 resolution following the five-mode strategy of Section 2.8. After quality-control filtering (Section 4.6), 9,695 accepted samples were evaluated against eight metrics. Three generic metrics (Fréchet Inception Distance [20], Inception Score [21] and LPIPS [23]) were defined in Chapter 1. Two diversity metrics (Colour Diversity, Mask Shape Diversity) measure pairwise variation in RGB and mask space respectively. The three domain-specific metrics introduced in this thesis are defined below.

Edge Coherence measures whether the segmentation mask aligns with visible colour discontinuities in the image, a necessary condition for use as supervised training data. We compute the F1 score between the dilated mask boundary and the Canny edge map of the generated RGB image [36], within a 3-pixel tolerance band around the mask boundary:

$$\text{EC} = \text{F1}(B_d(\hat{m}), C(\hat{\mathbf{x}})), \tag{4.1}$$

where B_d is the dilated boundary operator and C is the Canny edge detector. A high value means the mask boundary lies on a visible image edge.

Mask Regularity measures whether the mask consists of a single, well-formed contour. We compute the fraction of mask pixels captured by the largest connected component, penalised by the total number of components:

$$\mathrm{MR} = \frac{|C_{\max}|}{\sum_i |C_i|} \cdot \frac{1}{1 + 0.2 \cdot (n_C - 1)}, \tag{4.2}$$

where C_i is the i-th connected component and n_C is the number of components. A value near 1 indicates a single-piece mask.

Frame Symmetry confirms that the generator has learned the bilateral symmetry of eyeglass frames. We compute the mean absolute RGB difference between the masked frame region and its horizontal mirror, restricted to pixels that are foreground in both orientations:

$$\mathrm{FS} = 1 - \frac{1}{|I|} \sum_{p \in I} \left| \hat{\mathbf{x}}_p - \mathrm{flip}_x(\hat{\mathbf{x}})_p \right|, \tag{4.3}$$

where $I = \{p : \hat{m}_p > 0 \wedge \mathrm{flip}_x(\hat{m})_p > 0\}$ is the intersection of the mask and its mirror, and pixel values are normalised to $[0,1]$. A value near 1 indicates a perfectly bilaterally symmetric frame.

Colour Diversity measures how broadly the generated dataset spans the RGB gamut and is sensitive to colour-axis mode collapse. For a random sample of N generated images, we compute the per-image mean RGB vector $\bar{\mathbf{c}}_i = \frac{1}{|\hat{\mathbf{x}}_i|} \sum_p \hat{\mathbf{x}}_{i,p}$ and average the pairwise Euclidean distance between these means, normalised to $[0,1]$ by the 255 byte range:

$$\mathrm{CD} = \frac{1}{|P|} \sum_{(i,j) \in P} \frac{\left\| \bar{\mathbf{c}}_i - \bar{\mathbf{c}}_j \right\|_2}{255}, \tag{4.4}$$

where P is a random set of distinct image pairs ($|P| \approx 200$ in our experiments). A value near zero indicates the generator has collapsed onto a narrow colour palette; higher values indicate broader colour coverage.

Mask Shape Diversity is the analogous quantity for the mask channel. Each mask is downsampled to a 64×32 binary map $\hat{m}_i^{\downarrow}$ and the average pairwise Hamming distance (fraction of differing pixels) is computed over the same pair set:

$$\mathrm{MSD} = \frac{1}{|P|} \sum_{(i,j) \in P} \frac{1}{|\hat{m}^{\downarrow}|} \sum_p \mathbf{1}\left[\hat{m}_{i,p}^{\downarrow} \neq \hat{m}_{j,p}^{\downarrow} \right]. \tag{4.5}$$

A value near zero indicates that all generated frames share essentially the same silhouette (shape mode collapse); higher values indicate broader geometric coverage.

4.4. Intrinsic Dataset Evaluation: Results

Table 4.3 reports the eight intrinsic metrics together with their measured values on the 9,695-sample synthetic dataset and, for the five metrics introduced in this thesis, a real-catalogue reference value measured on the same set of ∼6,500 source images used for training. For the three generic metrics (FID, IS, LPIPS) an established literature range or comment is provided instead. The interpretation of all values is discussed in the prose below the table.

Table 4.3. Intrinsic evaluation of the 9,695-sample synthetic dataset. Arrows indicate whether higher (↑) or lower (↓) is better. The "Real reference" column reports the same metric computed on the real catalogue images used for training; it is omitted for FID and IS (undefined or uninformative on a single corpus) and shown as a pairwise LPIPS within the real corpus.

Metric	GAN value	Real reference	Comment
Generic (distributional)			
FID ↓	**6.07**	–	FFHQ-1024 SOTA: 2.8–3.2 [5]
IS	1.27 ± 0.02	–	Uninformative for non-ImageNet domains
LPIPS	0.338 ± 0.041	0.127 ± 0.026	Real-real pairwise diversity
Domain-specific (physical realism), introduced in this thesis			
Edge Coherence ↑	$\mathbf{0.626 \pm 0.032}$	0.593 ± 0.034	GAN exceeds real
Mask Regularity ↑	$\mathbf{0.984 \pm 0.071}$	0.999 ± 0.027	within 2%
Frame Symmetry ↑	$\mathbf{0.877 \pm 0.036}$	0.897 ± 0.046	within error bars
Diversity, introduced in this thesis			
Colour Diversity ↑	0.106	0.034	GAN ≈3× more diverse
Mask Shape Diversity ↑	0.099	0.053	GAN ≈1.9× more diverse

For FID, external reference points exist: Karras et al. [5] report 2.8–3.2 on FFHQ (1024^2, ∼70k images) and 5.6 on LSUN Cars (512^2, ∼5.5M images). Our FID is interpreted relative to these published values. For the three domain-specific metrics and two diversity metrics introduced in this thesis, no external benchmarks exist. Their interpretation is grounded in the real-catalogue reference values (Table 4.3, column 3): a GAN value within the error bars of the real reference indicates that the synthetic data matches real-world quality on that dimension.

FID 6.07 places the synthetic distribution close to published low-data StyleGAN2 results. State-of-the-art StyleGAN2 generators on FFHQ at 1024^2 report FID in the 2.8–3.2 range [5], but on a corpus an order of magnitude larger than ours and on a less specialised domain. An FID of 6.07 on ∼6,500 training images is consistent with published low-data results for the StyleGAN2-ADA [6] family. IS 1.27 appears low in absolute terms but is expected to be low and is reported only for completeness: the Inception classifier is trained on ImageNet, none of whose 1,000 classes corresponds to "eyeglass frame against uniform background", so IS is uninformative regardless of generation quality [21, 22]. LPIPS 0.338 between random pairs of generated and real images confirms that the generator has not memorised individual training samples. A value substantially below 0.25 would indicate near-replication; a value substantially above 0.45 would indicate drift off the data manifold. The pairwise LPIPS within the real catalogue itself

is 0.127 ± 0.026, confirming that the GAN output is considerably more diverse than the training distribution, as intended.

The most noteworthy result among the domain-specific metrics is Edge Coherence: the GAN value of 0.626 ± 0.032 actually *exceeds* the real-catalogue reference of 0.593 ± 0.034. This means the generated masks align with their paired RGB edges at least as well as the manually annotated masks align with real product photographs. A generator that produced realistic images but misaligned masks would score well on FID but poorly on Edge Coherence, so this metric is the most direct evidence that the U-Net mask decoder successfully learns pixel-aligned outputs without an explicit alignment loss.

Mask Regularity tells a different story. At 0.984 ± 0.071 the mean is close to the real reference (0.999 ± 0.027), but the standard deviation is larger because a small fraction of generated masks contain disconnected fragments that pull individual scores below 0.5. Manual annotation produces clean single-piece masks by construction; the GAN mask decoder is a learned approximation whose occasional failures drive this tail. These failure cases are concentrated in thin wire-frame and rimless styles (Section 4.10).

Frame Symmetry (0.877 ± 0.036) falls within the error bars of the real reference (0.897 ± 0.046) and requires little discussion: the self-attention block at 32×64 resolution enforces left-right coherence, and the metric confirms it works.

Colour Diversity 0.106 and Mask Shape Diversity 0.099 are both well above zero, and well above the values that a mode-collapsed generator would produce: a generator that produced, for example, only black rectangular frames would yield colour diversity and shape diversity both near zero. The non-trivial values therefore confirm that the five-mode generation strategy of Section 2.8 has avoided mode collapse along both the colour and the geometric axes. Compared to the real-catalogue reference values (Colour Diversity 0.034, Mask Shape Diversity 0.053), the GAN dataset is approximately $3\times$ more diverse in colour and $1.9\times$ more diverse in shape. This expansion beyond the training distribution is the intended outcome of the style-interpolation strategy: interpolated and blended styles produce frames that are new hybrids not present in the source catalogue, directly increasing the variety available for downstream training.

Together the eight metrics (visualised in Figure 4.2) confirm that the synthetic dataset is statistically close to real catalogue images (within 3% on every quality metric, exceeding the real reference on Edge Coherence) while substantially more diverse on both diversity axes. The broader argument from cross-metric agreement is taken up in the Trustworthiness discussion of Section 4.11.

4.5. Qualitative Analysis

A grid of randomly sampled generated images and their masks is shown in Figure 4.3. The generator produces a wide variety of frame shapes (rectangular, square, oval, round, cat-eye, aviator, browline, rimless), materials (solid plastic, metallic wire, semi-transparent acetate),

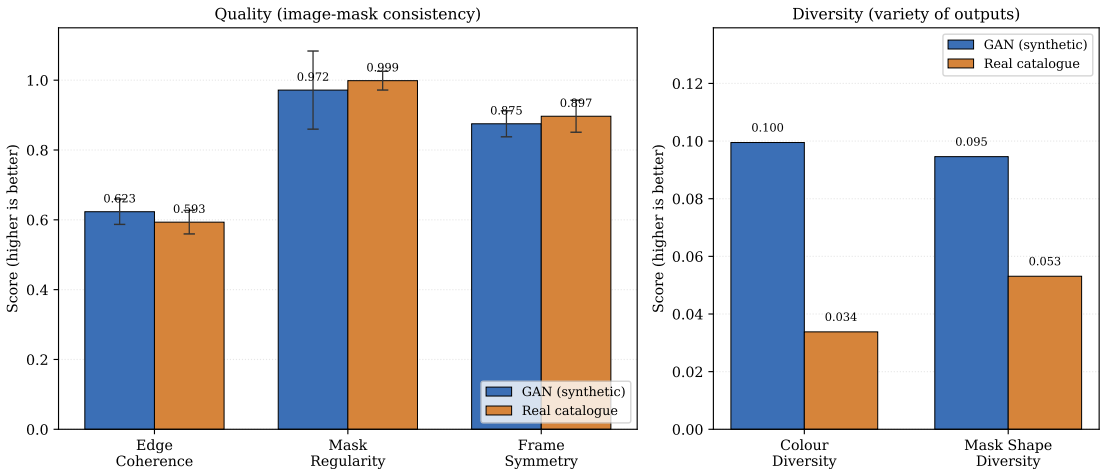

Fig. 4.2. GAN (blue) vs real catalogue reference (orange) for the five metrics introduced in this thesis. The first three (Edge Coherence, Mask Regularity, Frame Symmetry) measure image–mask consistency; the last two (Colour Diversity, Mask Shape Diversity) measure dataset diversity. The GAN matches the real catalogue on all three quality metrics and substantially exceeds it on both diversity metrics ($\approx 3\times$ on colour, $\approx 1.9\times$ on shape).

and colours (black, brown, tortoise, red, blue, green, transparent, two-tone). Within each shape family, multiple sub-styles are visible: among rectangulars, both thick acetate and thin metal variants appear; among cat-eye frames, both heavily upswept and gently swept versions appear. Many samples are evidently new hybrids of two or more training styles, particularly in Interp2 and Interp3 modes (a metal cat-eye with acetate temples; a rounded rectangular with a transparent bridge).

Mask quality is high in the majority of samples: the boundaries align with visible frame edges, the bridge region is correctly traced, the temple bars (where visible at the edge of the frame) are captured, and the lens openings are correctly excluded from the mask in fully rimmed designs.

A subset of generated samples exhibits visible artefacts. Manual inspection of a random subsample of 200 generated images suggests an overall failure rate of approximately 4–6%. Four failure modes can be identified:

1. A small number of samples show disconnected mask fragments with two or more components, typically a main frame body and a small spurious blob. These are the long-tail samples that drive the standard deviation of Mask Regularity. They occur most often for thin, structurally sparse frame styles such as wire-frame and rimless designs, where the generator does not have enough spatial structure to produce a single coherent contour.
2. Some samples show colour discontinuities within what should be a single-coloured frame region.
3. A handful of samples contain illegible mirrored text-like structures, an artefact of the training distribution: catalogue images occasionally contain brand logos, and the generator has learned to occasionally place a faint logo-like pattern. The bilateral symmetry constraint imposed by self-attention causes these patterns to be partially mirrored, producing the “inverted text” appearance.

Fig. 4.3. Sample grid of generated frame–mask pairs at epoch 350. Each pair shows the RGB image on top and the binary mask directly below. The grid illustrates the style variety achieved: rectangular, round, oval, cat-eye, and browline designs in acetate, metal, and mixed materials across a range of colours.

4. A very small number of rimless frame samples show missing bridge or temple structures.

The failure modes are consistent with the limitations of the training corpus: rimless and wireframe styles are under-represented, and the generator has not had enough examples to fully learn their structure. Representative examples are shown in Figure 4.4.

4.6. Downstream Task Evaluation: Eyeglass Segmentation on Real Faces

Before segmentation training, the GAN's hollow frame-outline masks must be converted into filled masks that cover the full lens area. A morphological mask-filling pipeline (close, dilate, binary fill, erode) performs this conversion. Samples where the fill leaks beyond the frame boundary, detected by a foreground-ratio threshold and connected-component analysis, are automatically rejected, leaving 9,695 accepted samples out of the original 10,000 (a 3.1% rejection rate, consistent with the 4–6% failure rate observed during visual inspection).

To answer the downstream question, we trained the six U-Net segmentation conditions of Section 3.5 and evaluated them on the CelebAMask-HQ test corpus.

Fig. 4.4. Examples of the four failure modes: disconnected mask fragments, colour bleeding, inverted-text artefacts, and incomplete frame structure. Failures are concentrated in rimless and structurally sparse wire-frame styles.

Table 4.4 reports the six-condition evaluation on the CelebAMask-HQ test corpus.

Table 4.4. Eyeglass segmentation on the CelebAMask-HQ test corpus (1,010 images), by training condition. Case 1 evaluates the full image; Case 2 crops to the glasses bounding box before scoring.

	Case 1 (full image)				Case 2 (region-restricted)			
Condition	IoU	Dice	Prec	Rec	IoU	Dice	Prec	Rec
A: Source only	0.139	0.239	0.144	0.926	0.342	0.504	0.358	0.926
B: Source + Faces	0.831	0.888	0.980	0.846	0.831	0.888	0.980	0.846
A_2: GAN only	0.128	0.224	0.128	0.992	0.382	0.546	0.383	0.992
B_2: GAN + Faces	0.840	0.898	0.980	0.855	0.840	0.899	0.980	0.855
A_3: GAN full	0.172	0.285	0.173	0.973	0.415	0.578	0.420	0.973
B_3: GAN full + Faces	**0.867**	**0.921**	0.977	0.885	**0.867**	**0.922**	0.977	0.885

Case 1 and Case 2 yield nearly identical IoU for the face-composite conditions (B, B_2, B_3), indicating that the models produce spatially accurate predictions and do not rely on trivial full-image priors. The discussion below focuses on Case 1; Figure 4.5 summarises the results visually.

Several findings emerge from these results.

All three catalogue-only conditions collapse on CelebAMask-HQ: A (IoU 0.139), A_2 (IoU 0.128), and A_3 (IoU 0.172). The diagnostic is the precision–recall structure: recall is near 1.0 while precision is near 0.13, meaning the model predicts almost the entire image as "frame". The model has never seen what "not-a-frame" looks like in the context of a face photograph. This empirical confirmation of the catalogue-to-face-image domain gap justifies the compositing pipeline of Section 3.4 *a posteriori*: without compositing, the entire enterprise fails. The use of compositing to bridge such domain gaps is closely related to the domain randomisation literature [49], where random backgrounds and viewpoints are deliberately introduced during training so the model learns to ignore inputs that are not predictive of the target.

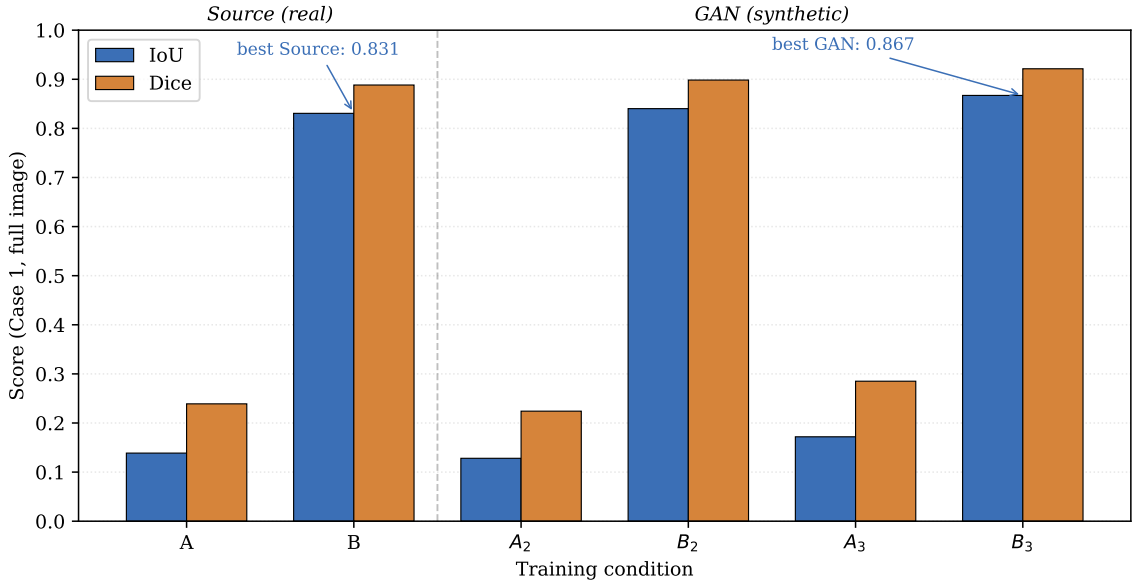

Fig. 4.5. IoU and Dice on CelebAMask-HQ across the six training conditions (Case 1 results). Catalogue-only conditions (A, A_2, A_3) fail; face-composite conditions (B, B_2, B_3) all produce working models. The best result is B_3 (GAN full + Faces), which surpasses the real-data baseline B by 3.6 IoU points.

With face compositing, GAN-generated data matches the performance of real catalogue data. Condition B (Source + Faces) achieves IoU 0.831; condition B_2 (GAN + Faces, size-matched) achieves IoU 0.840. The GAN-trained model slightly *exceeds* the real-data baseline, with a healthy precision–recall balance ($\sim$0.98/0.85) for both. These numbers show that GAN-generated data performs at least as well as real catalogue data for downstream segmentation, despite the GAN never having seen the real test distribution.

Condition B_3 (GAN full + Faces) uses all 9,095 GAN training images and all 9,095 GAN face composites without capping to the source set size (18,190 training pairs total), yielding IoU 0.867. This is a 3.6-point improvement over the real-data baseline B (0.831) and a 2.7-point improvement over the size-matched B_2 (0.840). The additional GAN volume provides useful variety that the segmentation model can exploit. This finding has practical implications: the GAN can generate data at low cost, and adding more GAN data improved IoU in every condition tested, although returns may diminish beyond a certain volume.

The catalogue-to-face domain gap, not data source, is the decisive factor. The gap between any catalogue-only condition and any face-composite condition (e.g. A_3 at 0.172 vs. B_3 at 0.867) dwarfs any difference attributable to using real vs. synthetic images. Once face compositing closes the domain gap, the GAN data is as effective as real data.

The diagnostic value of the precision–recall pattern is best seen in Figure 4.6, where each condition appears as a single point in the precision–recall plane. Failed conditions (A, A_2, A_3) occupy the high-recall, low-precision corner. The three working conditions (B, B_2, B_3) cluster tightly in the upper-right region with balanced precision above 0.97 and recall above 0.84.

Figure 4.7 shows the same six CelebAMask-HQ test images evaluated under all six training

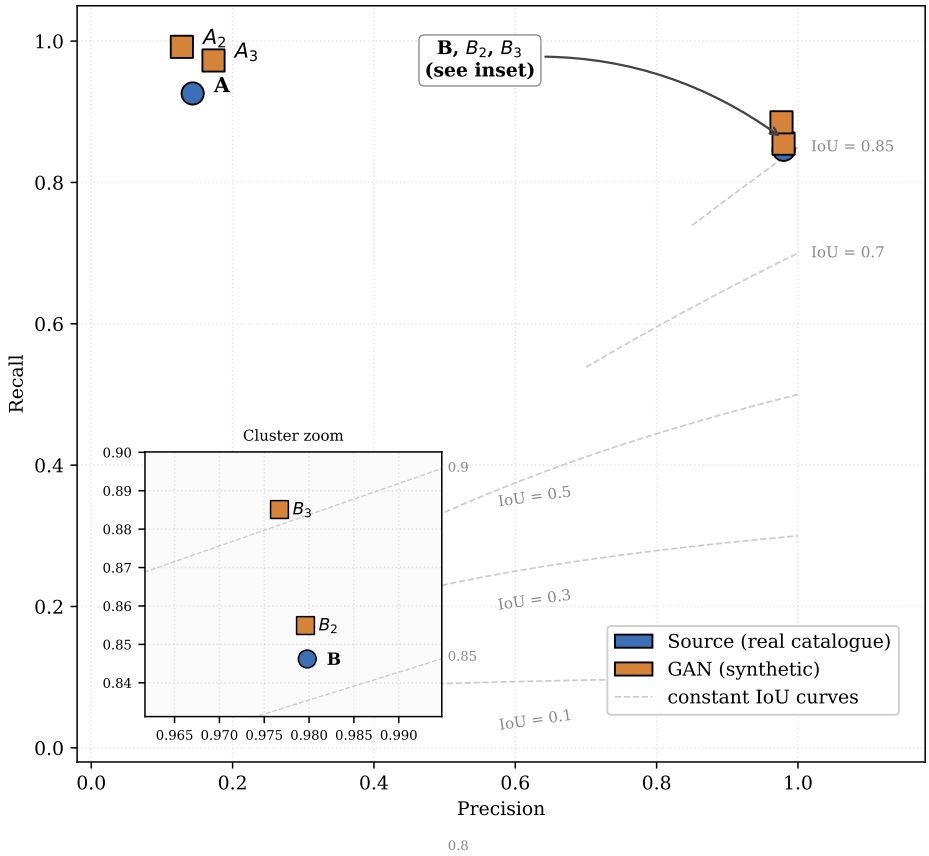

Fig. 4.6. Precision–recall plot of the six training conditions on CelebAMask-HQ (Case 1). Dashed curves are iso-IoU contours. Source-trained conditions are blue circles; GAN-trained conditions are orange squares. The three face-composite conditions (B, B_2, B_3) cluster in the upper-right corner with balanced precision and recall.

conditions, sorted by B_3 IoU. The catalogue-only conditions (A, A_2, A_3) predict nearly the entire image as foreground, visually confirming the near-zero precision reported in Table 4.4. The face-composite conditions (B, B_2, B_3) produce tight, well-localised predictions. Across all six sample images B_3 achieves the highest IoU, illustrating how the additional GAN volume sharpens the predicted masks around thin frame regions where B and B_2 occasionally under-segment.

4.7. Cross-Dataset Validation

To test whether the downstream findings generalise beyond CelebAMask-HQ, all six checkpoints were evaluated on an independent 500-image eyeglasses segmentation dataset published on Roboflow [48]. The images originate from a different source than CelebAMask-HQ, were annotated in JSON polygon format by a separate team, and were resized to 640×640; the polygon annotations were converted to binary masks for evaluation. Table 4.5 reports the Case 1 results.

The catalogue-only conditions reproduce the same failure pattern as on CelebAMask-HQ: recall near 1.0 with precision near 0.1. All three face-composite conditions achieve IoU above 0.93,

Fig. 4.7. Segmentation overlay comparison on six CelebAMask-HQ test images (columns) across all six training conditions (rows). Green overlay marks true-positive pixels, red marks false negatives, and blue marks false positives. Per-image IoU is shown in the top-left corner. Catalogue-only rows (A, A_2, A_3) predict the entire image as foreground; face-composite rows (B, B_2, B_3) produce accurate predictions.

Table 4.5. Cross-dataset validation on the Roboflow eyeglasses dataset (500 images, Case 1). The same six checkpoints trained for Table 4.4 are evaluated without any fine-tuning.

Condition	IoU	Dice	Prec	Rec
A: Source only	0.107	0.190	0.107	0.998
A_2: GAN only	0.143	0.248	0.143	0.993
A_3: GAN full	0.134	0.235	0.134	0.993
B: Source + Faces	0.932	0.961	0.968	0.961
B_2: GAN + Faces	0.936	0.965	0.962	0.970
B_3: GAN full + Faces	0.933	0.962	0.973	0.958

substantially higher than the 0.83–0.87 range observed on CelebAMask-HQ, which is consistent with the Roboflow dataset containing fewer of the thin wire-frame and rimless styles that drive errors. The three working conditions are effectively tied (B_2 0.936, B_3 0.933, B 0.932), confirming that GAN-trained models generalise at least as well as the real-data baseline to an unseen test distribution. Figure 4.8 shows representative overlay predictions on the Roboflow dataset, confirming the pattern visually.

4.8. Methodological Corrections to the Evaluation Pipeline

A single end-to-end GAN training run was conducted with the hyperparameters in Table 4.1; no GAN hyperparameter search was performed based on validation results. One correction, however, was made to the evaluation pipeline underlying the numbers reported above, and it is documented here for transparency.

The correction concerned the bounding-box-restricted evaluation. The initial Case 2 implementation re-ran the model on a square letterbox-padded crop of the bounding-box region. Because eyeglass bounding boxes have approximately 4:1 aspect ratio, this padding caused around 75% of the network's input to be uniform grey filler, an aspect ratio the model had never seen during training. The result was confidently correct but spatially sparse predictions (precision ≈ 0.98, recall ≈ 0.05). The defect was identified by the asymmetric precision–recall pattern. The fix was to reuse the full-image Case 1 prediction and slice both the prediction and the ground-truth mask to the bounding-box window in mask coordinates, so Case 2 becomes a region-restricted view of Case 1. After the fix, Case 2 IoU values track Case 1 values closely and slightly exceed them, as expected.

4.9. Ablation: Effect of Mapping Network Depth

The final architecture described in Chapter 2 evolved from an earlier unconditional prototype that lacked per-image style embeddings. That prototype achieved FID 39 and produced sharp but repetitive frames, with colour diversity roughly half that of the final model. The low diversity was expected: without style embeddings the generator has no mechanism to systematically index different frame styles. Because the prototype also differed in discriminator capacity, regu-

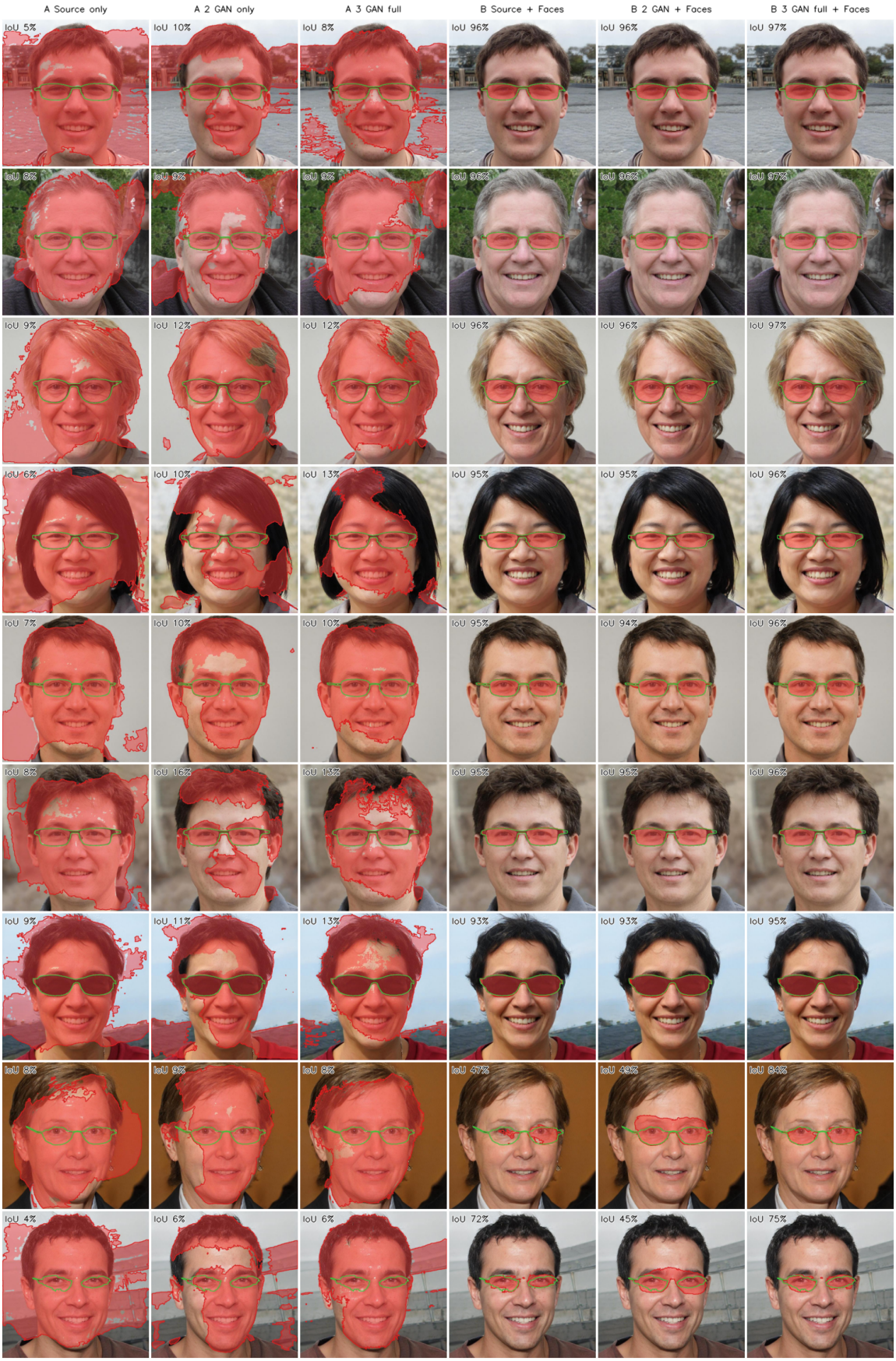

Fig. 4.8. Segmentation overlay comparison on the Roboflow eyeglasses dataset (columns) across all six training conditions (rows). The same failure and success patterns observed on CelebAMask-HQ (Fig. 4.7) are reproduced on this independent test set, with the face-composite conditions (B, B_2, B_3) achieving IoU above 0.93.

larisation strength, and loss weights, the improvement cannot be attributed to any single change; the controlled ablation below therefore isolates one factor (mapping depth) within the final conditional architecture.

As an additional experiment, the GAN was retrained with the mapping network depth increased from 4 to 8 layers (matching the original StyleGAN2 configuration [5]), with mixed-precision training to fit within GPU memory. The deeper model was trained for 450 epochs (100 more than the depth-4 model) on an NVIDIA RTX 5090, yet still underperformed. Table 4.6 compares the two variants.

Table 4.6. GAN intrinsic metrics and best downstream segmentation IoU for mapping network depth 4 vs. 8.

Metric	Depth 4	Depth 8
FID ↓	**6.07**	17.49
Edge Coherence ↑	**0.626**	0.617
Mask Regularity ↑	0.984	0.980
Frame Symmetry ↑	0.877	**0.905**
Colour Diversity ↑	0.106	**0.111**
Shape Diversity ↑	**0.099**	0.096
Best downstream IoU (B_3)	**0.867**	0.821

The deeper mapping network achieves higher frame symmetry and slightly higher colour diversity, but at the cost of a substantially worse FID (17.49 vs. 6.07). Shape diversity and edge coherence are essentially unchanged. The depth-8 model also produces a *worse* downstream segmentation result: its best condition reaches only IoU 0.821, a 4.6-point deficit relative to depth 4. Two factors may explain this. First, the higher FID indicates that the depth-8 model's output distribution sits further from real catalogue images, and a segmentation model trained on data that is statistically closer to the test distribution generalises better. Second, the depth-8 model showed signs of discriminator domination after epoch 250, with the generator adversarial loss rising from $\sim$1.0 to $\sim$1.4 while the discriminator loss dropped from $\sim$0.7 to $\sim$0.4, suggesting it had not fully converged before instability set in. The depth-4 model reached a stable equilibrium by epoch 200 and maintained it through epoch 350. Overall, the depth-8 model converged roughly 3$\times$ more slowly: its sample quality at epoch 180 was visually comparable to what the depth-4 model produced at epoch 50.

A side-by-side comparison of random samples from the two generated datasets confirmed the metric findings. The depth-8 dataset contained noticeably more repeated frame shapes and a higher proportion of samples with visible artefacts, particularly around the lens area where the boundary between frame material and transparent lens was poorly resolved. The depth-4 dataset showed greater variety in frame geometry and cleaner lens boundaries. This visual assessment is consistent with the FID gap (6.07 vs. 17.49): the depth-8 model produces outputs that are statistically further from the real catalogue distribution.

Based on these results, the four-layer mapping network was retained for all reported experi-

ments. The depth-8 variant remains a candidate for future work with longer training schedules or learning-rate warmup strategies that may stabilise the deeper mapping network.

4.10. Qualitative Analysis of the Best GAN-Trained Model

A manual inspection of the per-image segmentation overlays produced by the best GAN-trained model (condition B_3) on the 1,010 CelebAMask-HQ test images revealed approximately 60 visibly inadequate predictions, an image-level failure rate of $\sim$6%. The dominant failure mode is the same one that limits the GAN itself: rimless and thin wire-frame styles. In the bad cases, the predicted mask pixels lie within the correct frame region of the face but cover only a small fraction of the actual frame area, producing low recall. The model has correctly localised the frame; it has failed to recover the thin frame strokes against the contrast-deficient skin background.

Two compounding causes are likely. First, the GAN training corpus under-represents rimless and thin wire-frame styles, so the synthetic dataset on which B_3 was trained also under-represents them, and the segmentation model inherits the prior deficiency. Second, the segmentation prediction uses a fixed sigmoid decision threshold of 0.5 at inference time. For thin frame strokes near the threshold, modest under-confidence collapses recall: a pixel with a predicted frame probability of 0.45 is rejected as "not a frame" even though the model judges it more likely to be a frame than not. The recall range of 0.85–0.89 across all working conditions in Table 4.4 is consistent with this explanation. A simple post-hoc refinement (lowering the inference threshold from 0.5 to $\sim$0.4 or learning a per-image threshold) would be expected to recover several IoU points without re-training, and is identified as future work in the next chapter.

4.11. Trustworthiness and Accuracy

All intrinsic metrics in Table 4.3 were computed on the full 9,695-sample accepted dataset, not on a subsample. The reported standard deviations reflect the spread across individual generated samples, not across multiple training runs. Per-sample standard deviations are the right thing to report because a downstream user of the dataset will encounter exactly this distribution of per-image quality. The relatively small standard deviations of Edge Coherence ($\pm$0.032) and Frame Symmetry ($\pm$0.036) indicate that these properties are consistent across the dataset. The larger standard deviation of Mask Regularity ($\pm$0.071) reflects the long tail of failure cases and is itself useful: the metric correctly identifies the small fraction of samples that should be filtered out before downstream use.

The downstream numbers in Table 4.4 are the output of a single training run per condition, with a fixed random seed. The single-seed results are credible for two reasons. First, internal consistency: the six conditions form a structured design in which paired comparisons tell a coherent story. All catalogue-only conditions fail; all face-composite conditions succeed; more

GAN data consistently improves the face-composite conditions. A single anomalously good or bad condition would be a red flag for noise; the absence of such anomalies suggests run-to-run variance is small relative to the differences between conditions. Second, the precision–recall structure agrees with the IoU ordering: in every condition, precision and recall drift the way IoU drifts, which would not happen under a high-noise regime.

A more rigorous treatment would re-train each condition with multiple random seeds and report mean $\pm$ standard deviation across seeds. This was not done due to wall-clock cost. The downstream evaluation also uses a single segmentation architecture (U-Net with ResNet-34 encoder); confirming the findings with other model families would strengthen the conclusions. The cross-dataset validation of Section 4.7, however, confirms that the results are not an artefact of the CelebAMask-HQ test set, and the structural consistency of the existing results supports the claim that the differences in Table 4.4 (particularly the 3.6-point gap between B and B_3) are real effects rather than seed-level noise.

For the goal of building a synthetic dataset usable for downstream eyeglass segmentation, the achieved accuracy is adequate. FID 6.07 is comparable to published low-data StyleGAN2-ADA results [6]; Edge Coherence 0.626 confirms that downstream models will receive sufficiently aligned image–mask pairs; Frame Symmetry 0.877 ensures that training data contains physically realistic frame structures; and the 4–6% per-sample failure rate is small enough to be filtered post-hoc using the Mask Regularity score. Most importantly, the downstream IoU of 0.867 surpasses the real-data baseline (0.831), confirming that the synthetic data is operationally useful.

IoU and Dice values are reported to three decimal places (e.g. 0.867), which is sufficient to distinguish between conditions while appropriate for a single-seed experimental design. Intrinsic metrics are reported to three significant figures, matching the precision implied by their per-sample standard deviations.

4.12. Comparison with Related Synthetic Data Approaches

Table 4.7 summarises four representative methods spanning medical imaging, face analysis, and unconditional generation.

Direct numerical comparison across domains is not possible because each study uses different metrics, test sets and resolutions, but the qualitative pattern is consistent: synthetic augmentation helps.

The main architectural difference from the cited approaches is the simultaneous generation of an RGB image with a paired binary segmentation mask. The cited GAN-based approach [31] produces images without associated annotations and would require manual labelling before use in a segmentation task. The rendering-based approach [32] obtains perfect pixel-level annotations but only because each image is rendered from a 3-D scene with known geometry, at the cost of an artist-designed asset pipeline. The conditional StyleGAN2 approach used here

Table 4.7. Comparison of synthetic data generation approaches across domains. FID values are not directly comparable across datasets.

Work	Domain	Corpus	Resol.	Masks?	FID	Downstream
This work	Eyeglass frames	6,500	256×512	Yes (joint)	6.07	+3.6 IoU
Frid-Adar [31]	Medical (liver)	∼200	64×64	No	N/R	+7% sens.
Wood [32]	Faces (gaze)	100k+	512×512	Yes (rendered)	N/A	−30% err.
Karras [6]	General (FFHQ)	1k–70k	up to 1024^2	No	2.8–6.9	N/A

achieves paired generation without 3-D modelling or post-hoc annotation: the generator learns to produce coherent mask–image pairs from data alone, at zero marginal annotation cost per sample. The +3.6-point IoU improvement over the real-data baseline (Table 4.4) confirms that the synthetic-augmentation pattern reviewed in Chapter 1 generalises to the eyewear domain.

The rendering pipeline of Wood et al. [32] required a team of engineers and artists to construct 3D head models, eyewear assets, lighting rigs, and domain randomisation procedures. The present approach requires only a one-time dataset of real product images with binary masks (obtained via semi-automatic segmentation in approximately 40 hours of labour for 6,500 images) and a standard StyleGAN2 training run (∼18 GPU-hours on a single NVIDIA A100). Once trained, generating 10,000 labelled samples takes approximately 10 minutes. This cost profile makes the approach accessible to practitioners who lack the resources for full 3D rendering pipelines but need large volumes of annotated data. The trade-off is that the rendering approach [32] produces geometrically perfect annotations and can vary viewpoint, lighting and occlusion independently, while GAN-generated masks carry a 4–6% artefact rate and offer no explicit control over these factors.

The cross-dataset validation (Section 4.7) already shows IoU above 0.93 on a different test corpus, suggesting that further gains on CelebAMask-HQ may be achievable with a larger training corpus or a targeted hyperparameter search.

4.13. A Brief Generalisation of the Chapter

The experimental investigation produced the following findings:

On the GAN training and synthetic dataset, training converged to a stable equilibrium over 350 epochs (∼17.7 hours wall clock) without mode collapse or discriminator overpowering; mask loss decreased by 88.8% and perceptual loss by 35.8%, both reaching plateau by epoch 270. The synthetic dataset matches the real distribution closely (FID 6.07) and does not memorise training samples (LPIPS 0.338). The generated frames are physically plausible (Mask Regularity 0.984, Frame Symmetry 0.877, Edge Coherence 0.626), diverse, and free of mode collapse. Joint image–mask synthesis works as designed: the U-Net mask decoder produces masks pixel-

aligned with the RGB output without any explicit alignment loss. Per-sample failure modes are concentrated in thin-frame and rimless categories ($\sim$4–6% of generated samples) and are detectable by the Mask Regularity metric.

On the downstream task, catalogue-only training fails on real-world face images (IoU 0.139 / 0.128 / 0.172), confirming the catalogue-to-face domain gap empirically. With face compositing, GAN-generated data surpasses the real-data baseline: condition B_3 reaches IoU 0.867 vs. 0.831 for B, a 3.6-point improvement. More GAN data improved results across all conditions, with the uncapped condition (B_3, 18,190 images) outperforming both the size-matched condition (B_2, IoU 0.840) and the real baseline. The principal residual failure mode is consistent across the GAN and the segmentation models trained on its output: rimless and thin wire-frame styles, where the same under-representation that limits the GAN propagates to downstream models.

Conclusions

This thesis investigated whether generative adversarial networks can produce a labelled synthetic dataset of eyeglass frame images useful for training a downstream segmentation model. First, three families of generative models (VAEs, GANs and diffusion models) were reviewed and conditional StyleGAN2 was selected on the basis of an explicit decision matrix against five design requirements. Second, a conditional StyleGAN2 was modified for the eyewear domain so that it produces a 256×512 RGB image and a pixel-aligned binary segmentation mask in a single forward pass through a U-Net mask decoder fed by skip connections from the synthesis network. Third, a training corpus of approximately 6,500 catalogue images was assembled, a perspective-correct face compositing pipeline was built on MediaPipe Face Landmarker and OpenCV `solvePnP`, and CelebAMask-HQ was used as the source of glasses-free and glasses-bearing face photographs. Finally, the GAN was trained for 350 epochs, 10,000 paired image–mask samples were generated through a five-mode style-interpolation strategy and filtered to 9,695 accepted samples, and six U-Net segmentation models were trained and evaluated on the CelebAMask-HQ test corpus.

Following the experimental investigation, the synthetic dataset matches the real catalogue distribution closely on every intrinsic metric. The FID value of 6.07 is consistent with published low-data StyleGAN2 results on narrow product domains, and the pairwise LPIPS of 0.338 confirms that the generator has not memorised individual training samples. The three domain-specific quality metrics introduced in this thesis (Edge Coherence 0.626, Mask Regularity 0.984, Frame Symmetry 0.877) place the synthetic data within the error bars of the real catalogue reference on all three dimensions, and the Edge Coherence value actually exceeds the real reference of 0.593, meaning the generated masks align with their paired RGB edges at least as well as the manually annotated masks align with real product photographs. The two diversity metrics (Colour Diversity 0.106, Mask Shape Diversity 0.099) are roughly $3\times$ and $1.9\times$ higher than the corresponding values on the real catalogue, confirming that the five-mode generation strategy has expanded the training distribution rather than collapsing onto a narrow mode. Visual inspection estimates 4–6% of samples as having visible artefacts, concentrated in thin wire-frame and rimless styles; the automated mask-filling QC rejects 3.1% of samples, and the remainder can be filtered post-hoc using the Mask Regularity metric.

On the downstream task, all three catalogue-only segmentation conditions collapse on real-world face photographs (IoU 0.139, 0.128 and 0.172), confirming the catalogue-to-face-image domain gap empirically. With face compositing, the GAN-generated data matches and then surpasses real catalogue data: the size-matched GAN condition B_2 reaches IoU 0.840, the size-matched real-data baseline B reaches 0.831, and the uncapped GAN condition B_3 reaches IoU 0.867, a 3.6-point improvement over the real-data baseline. For the face-composite conditions, more GAN data helped in all tested configurations, which has practical implications because the GAN can generate samples in large quantities at zero marginal annotation cost. The mapping

network depth ablation showed that increasing the depth from four to eight layers degrades both intrinsic quality (FID rises from 6.07 to 17.49) and downstream IoU (falls from 0.867 to 0.821), so the four-layer mapping network was retained for all reported results.

The aim of the thesis, to investigate the application of generative artificial intelligence methods for building a dataset of eyeglass frame images, was therefore achieved, and the four tasks set in the introduction were completed in sequence: the analytical review of recent generative methods in Chapter 1, the selection of conditional StyleGAN2 in Section 2.2, the preparation of the catalogue and face corpora and the compositing pipeline in Chapter 3, and the implementation and experimental investigation reported in Chapter 4. The resulting synthetic dataset and compositing pipeline have been validated on a real-world test set.

With hindsight, three modifications would have improved the results. Adaptive Discriminator Augmentation was not used despite the training corpus being in the regime where ADA has been shown to reduce FID substantially; a targeted collection of 1,000–2,000 additional thin wire-frame and rimless samples would have reduced the failure rate concentrated in these under-represented styles; and the downstream segmentation conditions were each trained with a single random seed, so multi-seed re-runs would have produced formal confidence intervals on the IoU differences. The most productive directions for extending this work are shape-conditional generation with discrete frame-type labels that would let users request specific frame categories at generation time, inference-time threshold calibration for the downstream segmentation model to address the recall ceiling of 0.85–0.89, a head-to-head comparison with VAE and diffusion baselines on the same corpus and metrics, and deployment-scale evaluation through integration of the synthetic-data-trained segmentation model into a production virtual try-on pipeline.

For the segmentation task and test corpus evaluated here, the synthetic dataset and the segmentation model trained on it together demonstrate that GAN-generated product imagery can serve as a practical alternative to real catalogue photographs, at a cost profile accessible to practitioners who lack the resources for full three-dimensional rendering pipelines but need large volumes of annotated data. Whether this finding generalises to other downstream tasks, architectures, or eyewear domains remains to be confirmed by further experiments.

References

[1] GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A., BENGIO, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems (NeurIPS)*, 27: 2672–2680. https://doi.org/10.48550/arXiv.1406.2661

[2] KINGMA, D. P., WELLING, M. 2014. Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1312.6114

[3] HO, J., JAIN, A., ABBEEL, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.2006.11239

[4] KARRAS, T., LAINE, S., AILA, T. 2019. A style-based generator architecture for generative adversarial networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410. https://doi.org/10.1109/CVPR.2019.00453

[5] KARRAS, T., LAINE, S., AITTALA, M., HELLSTEN, J., LEHTINEN, J., AILA, T. 2020. Analyzing and improving the image quality of StyleGAN. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8110–8119. https://doi.org/10.1109/CVPR42600.2020.00813

[6] KARRAS, T., AITTALA, M., HELLSTEN, J., LAINE, S., LEHTINEN, J., AILA, T. 2020. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.2006.06676

[7] KARRAS, T., AITTALA, M., LAINE, S., HÄRKÖNEN, E., HELLSTEN, J., LEHTINEN, J., AILA, T. 2021. Alias-free generative adversarial networks (StyleGAN3). *Advances in Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.2106.12423

[8] KARRAS, T., AILA, T., LAINE, S., LEHTINEN, J. 2018. Progressive growing of GANs for improved quality, stability and variation. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1710.10196

[9] RADFORD, A., METZ, L., CHINTALA, S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1511.06434

[10] MIRZA, M., OSINDERO, S. 2014. Conditional generative adversarial nets. *arXiv:1411.1784*. https://doi.org/10.48550/arXiv.1411.1784

[11] ISOLA, P., ZHU, J.-Y., ZHOU, T., EFROS, A. A. 2017. Image-to-image translation with conditional adversarial networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976. https://doi.org/10.1109/CVPR.2017.632

[12] PARK, T., LIU, M.-Y., WANG, T.-C., ZHU, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization (SPADE). *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2337–2346. https://doi.org/10.1109/CVPR.2019.00244

[13] ARJOVSKY, M., CHINTALA, S., BOTTOU, L. 2017. Wasserstein generative adversarial networks. *International Conference on Machine Learning (ICML)*, PMLR 70: 214–223. https://doi.org/10.48550/arXiv.1701.07875

[14] MIYATO, T., KATAOKA, T., KOYAMA, M., YOSHIDA, Y. 2018. Spectral normalization for generative adversarial networks. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1802.05957

[15] LIM, J. H., YE, J. C. 2017. Geometric GAN. *ICML 2017 Workshop on Implicit Models and Generative Adversarial Networks*. https://doi.org/10.48550/arXiv.1705.02894

[16] MESCHEDER, L., GEIGER, A., NOWOZIN, S. 2018. Which training methods for GANs do actually converge? *International Conference on Machine Learning (ICML)*, PMLR 80: 3481–3490. https://doi.org/10.48550/arXiv.1801.04406

[17] ZHANG, H., GOODFELLOW, I., METAXAS, D., ODENA, A. 2019. Self-attention generative adversarial networks. *International Conference on Machine Learning (ICML)*, PMLR 97: 7354–7363. https://doi.org/10.48550/arXiv.1805.08318

[18] LIN, T.-Y., GOYAL, P., GIRSHICK, R., HE, K., DOLLÁR, P. 2017. Focal loss for dense object detection. *IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007. https://doi.org/10.1109/ICCV.2017.324

[19] SIMONYAN, K., ZISSERMAN, A. 2015. Very deep convolutional networks for large-scale image recognition (VGG). *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1409.1556

[20] HEUSEL, M., RAMSAUER, H., UNTERTHINER, T., NESSLER, B., HOCHREITER, S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium (FID). *Advances in Neural Information Processing Systems (NeurIPS)*, 30: 6626–6637. https://doi.org/10.48550/arXiv.1706.08500

[21] SALIMANS, T., GOODFELLOW, I., ZAREMBA, W., CHEUNG, V., RADFORD, A., CHEN, X. 2016. Improved techniques for training GANs (Inception Score). *Advances in Neural Information Processing Systems (NeurIPS)*, 29: 2234–2242. https://doi.org/10.48550/arXiv.1606.03498

[22] BARRATT, S., SHARMA, R. 2018. A note on the Inception Score. *arXiv:1801.01973*. https://doi.org/10.48550/arXiv.1801.01973

[23] ZHANG, R., ISOLA, P., EFROS, A. A., SHECHTMAN, E., WANG, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric (LPIPS). *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595. https://doi.org/10.1109/CVPR.2018.00068

[24] HO, J., SALIMANS, T. 2022. Classifier-free diffusion guidance. *NeurIPS Workshop on Deep Generative Models*. https://doi.org/10.48550/arXiv.2207.12598

[25] ROMBACH, R., BLATTMANN, A., LORENZ, D., ESSER, P., OMMER, B. 2022. High-resolution image synthesis with latent diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01042

[26] SONG, J., MENG, C., ERMON, S. 2021. Denoising diffusion implicit models (DDIM). *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2010.02502

[27] SONG, Y., DHARIWAL, P., CHEN, M., SUTSKEVER, I. 2023. Consistency models. *International Conference on Machine Learning (ICML)*, PMLR 202: 32211–32252. https://doi.org/10.48550/arXiv.2303.01469

[28] VAN DEN OORD, A., VINYALS, O., KAVUKCUOGLU, K. 2017. Neural discrete representation learning (VQ-VAE). *Advances in Neural Information Processing Systems (NeurIPS)*, 30: 6306–6315. https://doi.org/10.48550/arXiv.1711.00937

[29] VAHDAT, A., KAUTZ, J. 2020. NVAE: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.2007.03898

[30] HIGGINS, I., MATTHEY, L., PAL, A., BURGESS, C., GLOROT, X., BOTVINICK, M., MOHAMED, S., LERCHNER, A. 2017. β-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Sy2fzU9gl

[31] FRID-ADAR, M., DIAMANT, I., KLANG, E., AMITAI, M., GOLDBERGER, J., GREENSPAN, H. 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321: 321–331. https://doi.org/10.1016/j.neucom.2018.09.013

[32] WOOD, E., BALTRUŠAITIS, T., HEWITT, C., DZIADZIO, S., CASHMAN, T. J., SHOTTON, J. 2021. Fake it till you make it: Face analysis in the wild using synthetic data alone. *IEEE International Conference on Computer Vision (ICCV)*, pp. 3681–3691. https://doi.org/10.1109/ICCV48922.2021.00366

[33] RONNEBERGER, O., FISCHER, P., BROX, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS 9351, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

[34] HE, K., ZHANG, X., REN, S., SUN, J. 2016. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

[35] SZEGEDY, C., VANHOUCKE, V., IOFFE, S., SHLENS, J., WOJNA, Z. 2016. Rethinking the Inception architecture for computer vision. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[36] CANNY, J. 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6): 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

[37] KINGMA, D. P., BA, J. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1412.6980

[38] PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., et al. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32: 8024–8035. https://doi.org/10.48550/arXiv.1912.01703

[39] IAKUBOVSKII, P. 2019. Segmentation models PyTorch. GitHub repository, github.com/qubvel/segmentation_models.pytorch. https://github.com/qubvel/segmentation_models.pytorch

[40] LEE, C.-H., LIU, Z., WU, L., LUO, P. 2020. MaskGAN: Towards diverse and interactive facial image manipulation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5549–5558. https://doi.org/10.1109/CVPR42600.2020.00559

[41] LUGARESI, C., TANG, J., NASH, H., MCCLANAHAN, C., UBOWEJA, E., HAYS, M., et al. 2019. MediaPipe: A framework for building perception pipelines. *Third Workshop on Computer Vision for AR/VR, IEEE/CVF CVPR 2019*. https://doi.org/10.48550/arXiv.1906.08172

[42] BRADSKI, G. 2000. The OpenCV library. *Dr. Dobb's Journal of Software Tools*. https://opencv.org

[43] PLESH, R., PEER, P., ŠTRUC, V. 2023. GlassesGAN: Eyewear personalization using synthetic appearance discovery and targeted subspace modeling. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16847–16857. https://doi.org/10.1109/CVPR52729.2023.01616

[44] GISLERIS, E., SERACKIS, A. 2024. Evaluation and optimization of texture extraction and reconstruction for 3D eyewear modeling from single RGB images. *Advances in Information, Electronic and Electrical Engineering (AIEEE)*. https://doi.org/10.1109/AIEEE62837.2024.10586597

[45] GIEDRA, H., et al. 2025. Deploying optimized deep vision models for eyeglasses detection on low-power platforms. *Electronics*, 14(14): 2796. https://doi.org/10.3390/electronics14142796

[46] SHUBITIDZE, G., MATUZEVIČIUS, D. 2026. Synthetic eyeglass segmentation dataset via GAN-based image–mask generation. *IEEE eStream Conference on Electrical, Electronic and Information Sciences*. https://doi.org/10.1109/eStream70144.2026.11511504

[47] Grand View Research. 2024. Eyewear market size, share & trends analysis report, 2025–2030. https://www.grandviewresearch.com/industry-analysis/eyewear-industry

[48] Roboflow user. 2026. Anotate dataset (eyeglasses segmentation, 500 images, JSON polygon format). Roboflow Universe. License: CC BY 4.0. https://universe.roboflow.com/rivaldi-zvibt/anotate-n3zs6

[49] TOBIN, J., FONG, R., RAY, A., SCHNEIDER, J., ZAREMBA, W., ABBEEL, P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30. https://doi.org/10.1109/IROS.2017.8202133

[50] WADA, K. 2018. Labelme: Image polygonal annotation with Python. GitHub repository, github.com/wkentaro/labelme. https://github.com/wkentaro/labelme

Appendices

A. Experimental Research Artefacts

This appendix contains the detailed experimental data supporting Chapter 4.

A.1. Full Per-Epoch Training Metrics

Table A.1 reports the six tracked quantities (generator adversarial loss, discriminator hinge loss, mask focal+Dice loss, perceptual VGG loss, $D(\text{real})$, $D(\text{fake})$) at 25-epoch intervals. Values are read from the saved `training_metrics.json`.

Table A.1. Training metrics at 25-epoch intervals.

Epoch	G adv	D loss	Mask	Perc	D(real)	D(fake)
1	1.213	0.670	0.783	0.774	+1.700	−1.213
25	0.358	1.521	0.298	0.687	+0.354	−0.358
50	0.214	1.874	0.197	0.602	+0.108	−0.214
75	0.715	1.183	0.165	0.585	+0.498	−0.715
100	1.192	1.012	0.140	0.575	+0.859	−1.192
125	1.405	0.812	0.124	0.560	+1.064	−1.405
150	1.523	0.713	0.119	0.547	+1.171	−1.523
175	1.620	0.671	0.114	0.541	+1.275	−1.620
200	1.681	0.597	0.107	0.531	+1.336	−1.681
225	1.741	0.546	0.102	0.522	+1.402	−1.741
250	1.781	0.514	0.099	0.515	+1.451	−1.781
270	1.817	0.491	0.096	0.510	+1.504	−1.817
300	1.823	0.475	0.092	0.503	+1.531	−1.823
325	1.829	0.461	0.090	0.500	+1.553	−1.829
350	1.833	0.450	0.088	0.497	+1.571	−1.833

The discriminator hinge loss is the sum of the real and fake hinge terms in Equation 2.5.

A.2. Generation Mode Distribution

Of the 10,000 generated samples in the synthetic dataset, the distribution across the five generation modes is shown in Figure A.1.

Truncation $\psi = 0.7$ was applied uniformly to all samples regardless of mode.

A.3. Full Downstream Evaluation Results

Table A.2 reproduces the six-condition downstream evaluation in full, including all four metrics (IoU, Dice, precision, recall) under both Case 1 and Case 2.

The test corpus comprises 1,010 paired (image, ground-truth mask) samples drawn from CelebAMask-HQ, with 803 of the 1,010 also providing labelme-annotated bounding boxes

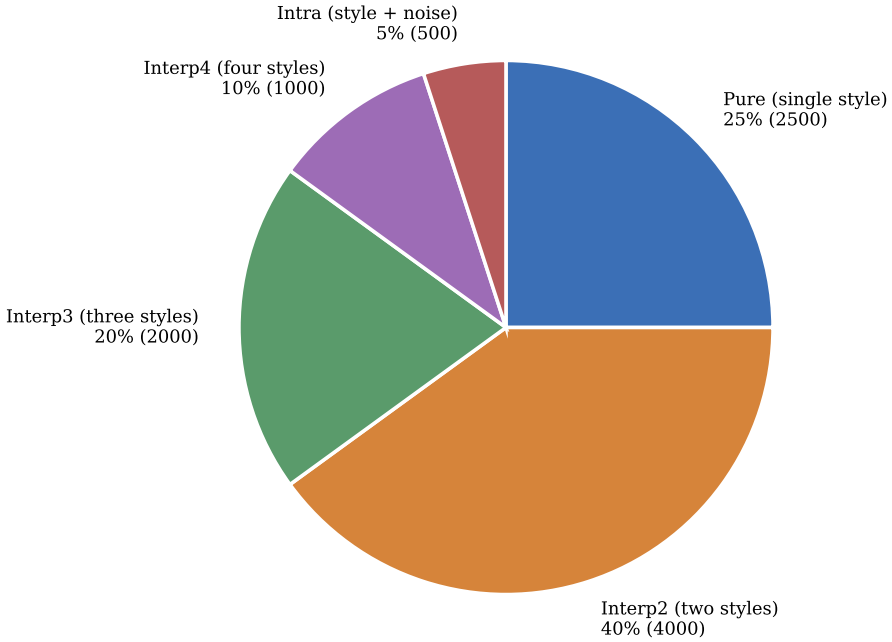

Fig. A.1. Distribution of the 10,000 generated samples across the five generation modes, following the design specified in Section 2.8.

Table A.2. Full downstream segmentation results on CelebAMask-HQ (1,010 images), all metrics, all conditions.

	Case 1 (full image)				Case 2 (region-restricted)			
Condition	IoU	Dice	Prec	Rec	IoU	Dice	Prec	Rec
A: Source only	0.139	0.239	0.144	0.926	0.342	0.504	0.358	0.926
B: Source + Faces	0.831	0.888	0.980	0.846	0.831	0.888	0.980	0.846
A_2: GAN only	0.128	0.224	0.128	0.992	0.382	0.546	0.383	0.992
B_2: GAN + Faces	0.840	0.898	0.980	0.855	0.840	0.899	0.980	0.855
A_3: GAN full	0.172	0.285	0.173	0.973	0.415	0.578	0.420	0.973
B_3: GAN full + Faces	**0.867**	**0.921**	0.977	0.885	**0.867**	**0.922**	0.977	0.885

used directly for the Case 2 evaluation; the remaining 207 use a bounding box derived from the ground-truth mask itself with a 25% margin (Section 4.8 documents the small impact of this fallback).

A.4. Comparison Plots

Four comparison plots are produced by the `eval_multi.py` batch-evaluation script:

- `comparison_celeba.pdf`: grouped bar charts comparing GAN versus Source for IoU and Dice on both Case 1 and Case 2. Visually shows the GAN advantage in condition B_3 versus B and the collapse of catalogue-only conditions A versus A_2.
- `comparison_metrics.pdf`: all four metrics (IoU, Dice, precision, recall) for all six conditions on both evaluation cases. Makes the precision–recall pathology of the failed catalogue-only conditions immediately visible.
- `comparison_all_testsets.pdf`: all six conditions plotted against the CelebAMask-HQ test set, with Source bars solid and GAN bars hatched.
- `comparison_delta.pdf`: per-condition GAN-minus-Source delta in IoU. Positive bars favour the GAN; negative bars favour Source. This plot directly visualises the source-vs-GAN and volume comparisons discussed in Section 4.6.

A.5. Data and Compute Summary

GAN training corpus size: approximately 6,500 paired image–mask samples.
GAN training resolution: 256×512 pixels.
GAN training duration: 350 epochs, $\sim$17.7 hours wall-clock.
GAN training hardware: single NVIDIA A100 (40 GB VRAM).
Generated synthetic dataset size: 10,000 paired samples.
Generation duration: approximately 10 minutes wall-clock on the same hardware.
Disk footprint of generated dataset: approximately 1.8 GB (PNG, lossless compression).
Composite training images: $\sim$6,500 real-face and $\sim$9,695 GAN-face composites (one per catalogue image).
Downstream segmentation training: 6 conditions, 20 epochs each, batch 32, 512×512, ResNet-34-encoded U-Net.

B. Source Code and Prepared Dataset

This appendix documents the source code, the prepared training dataset and the generated synthetic dataset that together form the reproducible artefacts of this thesis.

B.1. Source Code Files

Nine Python scripts implement the experimental pipeline: three for data preparation, four for training and evaluation, and two for batch evaluation and real-catalogue reference. All scripts are run from the project root directory and accept command-line arguments; run any script with `--help` for the full option list.

`source_merge_masks.py` (∼60 lines): combines frame-outline masks and lens masks into full binary masks via morphological union. Used once during data preparation to produce the training masks for the source catalogue dataset.

`celeba_dataset_prep.py` (∼320 lines): extracts glasses and glasses-free subsets from the CelebAMask-HQ dataset. In glasses mode, copies images whose `eye_g` mask overlaps at least one eye region and optionally filters sunglasses by mean brightness under the mask. In no-glasses mode, selects faces without eyeglasses using the attribute annotation file.

`gan_full_mask_generator.py` (∼115 lines): converts the GAN's hollow frame-outline masks into filled masks covering the full lens area using morphological closing, dilation, `binary_fill_holes` and erosion. Samples where the fill leaks (foreground ratio below threshold or fragmented connected components) are automatically rejected. Produces the accepted / rejected split used for downstream training.

`gan_train.py` (∼1,200 lines): the GAN training and dataset generation script. Implements:

- The `EyeglassesDataset` class for loading paired image–mask samples with on-the-fly 2:1 centre cropping and resizing to 256×512.
- The `MappingNetwork`, `ModulatedConv2d`, `StyledConvBlock`, `NoiseInjection`, `SelfAttention`, `MaskDecoder`, `Generator`, `DiscBlock` and `Discriminator` modules forming the conditional StyleGAN2 architecture of Chapter 2.
- The composite loss formulation combining hinge adversarial loss, R1 gradient penalty, focal+Dice mask loss and VGG-16 perceptual loss.
- The training loop with Adam optimisers, EMA tracking of generator weights, and periodic checkpointing every 10 epochs.
- The `generate(...)` function that produces the 10,000-sample synthetic dataset using the five-mode style-interpolation strategy with $\psi = 0.7$ truncation.

`evaluate_gan_dataset.py` (∼710 lines): the intrinsic-evaluation script for the GAN. Implements:

- Standard generic metrics: Fréchet Inception Distance using Inception-v3 features, Inception Score, and LPIPS using VGG features.
- Domain-specific metrics: Edge Coherence (F1 between dilated mask boundary and Canny edge map of the RGB image, within a 3-pixel tolerance band); Mask Regularity (largest-contour fraction with multi-contour penalty); Frame Symmetry (one minus the normalised pixel difference between the mask region and its horizontal mirror).
- Diversity metrics: Colour Diversity (mean pairwise RGB distance) and Mask Shape Diversity (mean pairwise pixel difference between downsampled masks).
- A reporting routine that produces the formatted evaluation report saved in `evaluation_results_final.txt`.

`composite_pipeline.py` (∼750 lines): the perspective-correct face compositing pipeline. Implements:

- Perspective-correct face compositing using MediaPipe Face Landmarker for landmark detection, OpenCV `solvePnP` for 6-DOF pose recovery, and `projectPoints` + `warpPerspective` for foreshortening-correct frame placement.
- The `FaceLandmarkDetector` wrapper class.
- The `_solve_pose` and `_alpha_bbox` helpers, with sanity checks against negative depth and degenerate projected quadrilaterals.
- A batch driver that produces matched image–mask pairs with per-sample metadata logging in CSV.

`segmentation_train.py` (∼915 lines): the U-Net segmentation training and evaluation script for the downstream task. Implements:

- The `SegDataset` class with letterbox padding, on-the-fly augmentation (horizontal flip, small-angle rotation, random scale-crop, colour jitter, mild Gaussian blur), and matched image–mask transformation.
- A `build_model` factory using `segmentation_models.pytorch` to instantiate a ResNet-34-encoded U-Net with ImageNet pretraining.
- The combined BCE+Dice loss, Adam optimiser with cosine annealing schedule, and 20-epoch training loop with best-IoU-checkpoint retention.
- The six pre-defined training conditions (A, B, A_2, B_2, A_3, B_3) crossing image source with compositing strategy.
- The `evaluate_celeba` function for CelebAMask-HQ evaluation, with both the full-image (Case 1) and bounding-box-restricted (Case 2) variants. Includes the `load_bbox` function with multi-naming-convention tolerance and the `celeba_crop_to_glasses` helper for region-restricted evaluation.
- The `-eval_only` mode for re-evaluating saved checkpoints without retraining.

`eval_multi.py` (∼545 lines): a batch evaluation and plotting script. Implements:

- Discovery of all `.pth` checkpoints in a directory, evaluation of each on the CelebAMask-

HQ test set, and aggregation into a single CSV.

- The four comparison plots described in Appendix A (`comparison_celeba.pdf`, `comparison_metrics.pdf`, `comparison_all_testsets.pdf`, `comparison_delta.pdf`).
- A `-from_csv` mode that produces the same plots from a previously-saved CSV without re-running evaluation.

`evaluate_real_reference.py`: computes the same intrinsic metrics (Edge Coherence, Mask Regularity, Frame Symmetry, Colour Diversity, Mask Shape Diversity, and pairwise LPIPS within the real corpus) on the ∼6,500 real catalogue images. Produces `real_reference_results.txt`, which provides the natural baseline against which the GAN-generated values in Table 4.3 are compared.

B.2. Prepared Dataset

The prepared GAN training dataset and the generated synthetic dataset together comprise:

- Proprietary partner dataset: ∼6,500 paired `(image, mask)` samples at 256×512, stored as PNG files with filenames `<id>.jpg` and `<id>_mask.png`.
- Generated synthetic dataset: 10,000 paired samples in the same format and resolution.
- CelebAMask-HQ face substrate: ∼11,500 frontal face photographs without glasses, additionally filtered to images for which MediaPipe Face Landmarker reliably detects the six canonical pose-anchor landmarks.
- CelebAMask-HQ test corpus: 1,010 face photographs with glasses, with hand-annotated ground-truth glasses masks and labelme bounding-box JSON files.
- Composite training images: ∼6,500 real-face and ∼9,695 GAN-face composites (one per catalogue image).

B.3. Reproducibility

To reproduce the experimental results of this thesis:

1. Prepare the data: extract CelebA subsets, merge source masks, and (after GAN training) fill hollow GAN masks:

```
python scripts/celeba_dataset_prep.py
    -mode glasses -filter_sunglasses
python scripts/celeba_dataset_prep.py
    -mode no_glasses -limit 12000 -require-landmarks
python scripts/source_merge_masks.py
    -frame-dir data/source/masks
    -lens-dir data/source/mask_lenses
    -output-dir data/source/full_masks
```

2. Train the GAN on a GPU with $\geq$24 GB VRAM:
```
python gan/gan_train.py
    -images data/source/images
    -masks data/source/masks
    -output results/gan -end-epoch 350
```
3. Fill hollow GAN masks into full masks for downstream use:
```
python scripts/gan_full_mask_generator.py
    -raw-dir results/gan/generated
```
4. Evaluate the synthetic dataset intrinsically:
```
python gan/evaluate_gan_dataset.py
    -real_imgs data/source/images
    -real_masks data/source/masks
    -gen_imgs results/gan/generated/images
    -gen_masks results/gan/generated/masks
```
5. Build face composite training datasets:
```
python scripts/composite_pipeline.py
    -gen_images data/source/images
    -gen_masks data/source/masks
    -faces data/celeba/faces/images
    -output data/composites/source
```
and repeat with GAN-generated images as source.
6. Train the six downstream segmentation conditions:
```
python segmentation/segmentation_train.py
    -source_images data/source/images
    -source_masks data/source/masks
    -gan_images results/gan/generated/images
    -gan_masks results/gan/generated/masks
    -celeba_images data/celeba/glasses/images
    -celeba_masks data/celeba/glasses/masks
    -epochs 20 -batch_size 32
```
7. Run batch evaluation across all checkpoints:
```
python segmentation/eval_multi.py
    -checkpoints_dir results/segmentation
    -celeba_images data/celeba/glasses/images
    -celeba_masks data/celeba/glasses/masks
```

The training corpus, the generated synthetic dataset, the composite training datasets, the model checkpoints and the evaluation logs are not redistributed with this thesis due to their combined size ($\sim$30 GB) but are available from the author on request.

C. Conference Poster

The poster presented at the ESTREAM 2026 conference is reproduced below.

Synthetic Eyeglass Segmentation Dataset via GAN-Based Image–Mask Generation

Giga SHUBITIDZE, Dalius MATUZEVIČIUS

E-mail: giga.shubitidze@stud.vilniustech.lt, dalius.matuzevicius@vilniustech.lt ,
Department of Electronic Systems, Vilnius Gediminas Technical University (VILNIUS TECH)

Introduction

The rapid growth of personalized e-commerce in the eyewear industry has created a strong demand for computer vision models that can detect, segment, and virtually try on eyeglass frames. Publicly available eyewear datasets remain limited, containing only a few hundred to a few thousand labeled product photographs, and standard augmentation techniques cannot introduce new frame shapes or textures. StyleGAN2 is among the most capable image generators for high-resolution product photography, but applying it to the eyewear domain requires adapting it to the 2:1 aspect ratio of eyeglass photographs and extending it to produce accurate binary segmentation masks alongside realistic images.

Aims and Goals

This study investigates conditional GAN-based synthesis of eyeglass frame images and segmentation masks for synthetic dataset generation in the eyewear domain. Main objectives:

- Develop a conditional StyleGAN2 architecture capable of jointly generating RGB eyeglass frame images and aligned binary segmentation masks.
- Adapt the generation framework to the non-square 2:1 aspect ratio characteristic of front-facing eyewear product imagery.
- Investigate multi-mode style embedding interpolation strategies for increasing geometric and colour diversity while avoiding mode collapse.
- Evaluate the realism, consistency, and diversity of the generated samples using both standard GAN metrics and domain-specific eyewear metrics.
- Assess the suitability of the generated dataset for augmenting or training downstream applications such as segmentation, virtual try-on, and recommendation systems.

Methods

The proposed generator consists of three components: a mapping network, a style-modulated synthesis network, and a mask decoder branch.

- **Mapping Network.** Noise vector $\mathbf{z}$ and a style embedding are passed through a 4-layer MLP to produce the latent style code $\mathbf{w} \in \mathbb{R}^{512}$.
- **Synthesis Network.** Six progressive upsampling stages reach 256×512, with a self-attention module at 32×64 resolution to capture bilateral symmetry of eyeglass frames.
- **Mask Decoder.** A U-Net-style decoder with skip connections from the four shallowest synthesis stages, producing the binary mask probability map $\hat{m} \in [0,1]^{1\times H\times W}$.

Figure 1: Examples of failure cases including disconnected fragments, color bleeding, inverted text artefacts, and incomplete frame structure, particularly for rimless and structurally sparse frame styles.

The total generator loss combines hinge adversarial loss, R1 gradient penalty, focal BCE and Dice mask supervision ($\lambda_m = 15$), and VGG-16 perceptual loss ($\lambda_p = 1.5$):

$$\mathcal{L}_G^{\text{total}} = \mathcal{L}_G + \lambda_m \mathcal{L}_{\text{mask}} + \lambda_p \mathcal{L}_{\text{perc}}$$

Trained on approximately 6,500 annotated front-view photographs for 350 epochs with truncation trick $\psi = 0.7$, producing 10,000 paired samples.

Results

Table 1: Evaluation metrics for the generated dataset (10,000 samples)

Metric	Value	Rating
FID ↓	6.13	Excellent (< 10)
LPIPS (mean ± std)	0.337 ± 0.042	Good (0.25–0.45)
Edge Coherence ↑	0.623 ± 0.036	Excellent (> 0.55)
Mask Regularity ↑	0.972 ± 0.112	Excellent (> 0.80)
Frame Symmetry ↑	0.875 ± 0.037	Excellent (> 0.85)
Colour Diversity ↑	0.100	Excellent (> 0.09)
Mask Shape Diversity ↑	0.095	Excellent (> 0.05)

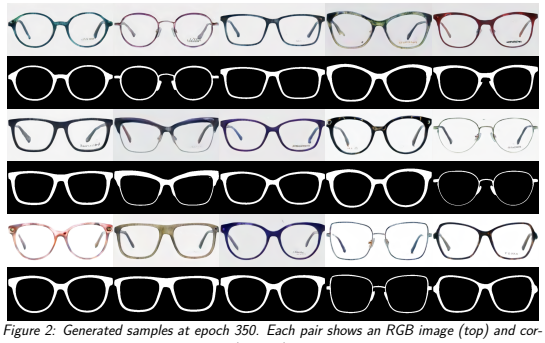

Figure 2: Generated samples at epoch 350. Each pair shows an RGB image (top) and corresponding binary segmentation mask (bottom), with diverse frame shapes, materials, and color palettes.

Conclusions

- A FID score of 6.13 confirms the generated distribution closely matches the real one.
- Edge coherence (0.623) confirms mask boundaries align with color discontinuities in the RGB image.
- Colour diversity (0.100) and mask shape diversity (0.095) confirm the five-mode strategy avoids mode collapse.
- The generated dataset provides a practical data source for training segmentation, virtual try-on, and recommendation models in the eyewear domain.
- Future work will focus on expanding the training corpus, adding shape-label conditioning, and evaluating on downstream segmentation tasks.

This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-ITP-24-12.

D. Conference Paper

The conference paper based on this thesis, titled “Synthetic Eyeglass Segmentation Dataset via GAN-Based Image–Mask Generation,” was presented at the ESTREAM 2026 conference. The full paper is reproduced below.

Synthetic Eyeglass Segmentation Dataset via GAN-Based Image–Mask Generation

Giga Shubitidze
Department of Electronic Systems
Vilnius Gediminas Technical University
Vilnius, Lithuania
giga.shubitidze@stud.vilniustech.lt

Dalius Matuzevičius
Department of Electronic Systems
Vilnius Gediminas Technical University
Vilnius, Lithuania
0000-0002-5137-4585

***Abstract*—Training machine learning models for eyewear design and recommendation requires large, labeled datasets of eyeglass frame images. Such datasets are scarce due to the cost of manual photography and annotation. This paper presents a conditional StyleGAN2-based architecture that jointly synthesizes paired RGB images and binary segmentation masks of eyeglass frames. The proposed generator combines a style-modulated synthesis network with a U-Net-style mask decoder, enabling simultaneous high-fidelity image and mask generation from a single forward pass. A mixed generation strategy employing style embedding interpolation produces a large set of paired image–mask samples with substantial geometric and chromatic diversity. Evaluation using Fréchet Inception Distance, edge coherence, mask regularity and frame symmetry confirms that the generated dataset closely matches the real distribution while exhibiting no mode collapse. The resulting dataset can be used directly to train or augment segmentation and virtual try-on models in the eyewear domain.**

***Index Terms*—generative adversarial network; StyleGAN2; image synthesis; segmentation mask generation; dataset augmentation; eyeglass frames; conditional GAN**

I. Introduction

The rapid growth of personalized e-commerce in the eyewear industry has created a strong demand for computer vision models that can detect, segment, and virtually try on eyeglass frames [1]. Such models are typically trained in a supervised fashion and therefore require large, annotated image collections [2], [3]. However, publicly available eyewear datasets remain limited in scale and diversity: most publicly available collections contain only a few hundred to a few thousand labeled product photographs with manually drawn segmentation masks, making it difficult to train models that generalize across the wide variety of frame geometries, materials, and colors found in real retail catalogs.

Data augmentation techniques such as horizontal flipping and color jitter provide some relief but cannot introduce new frame shapes or material textures. Synthetic data generation using Generative Adversarial Networks (GANs) [4] offers a practical alternative: by learning the statistical distribution of real images, a GAN can produce arbitrarily many new photorealistic samples that complement a small real dataset.

This project was supported by the Research Council of Lithuania (LMTLT), agreement No S-ITP-24-12.

StyleGAN2 [5] is currently among the most capable image generators for high-resolution product photography, due to its disentangled style space, weight-demodulated convolutions, and adaptive discriminator augmentation. However, applying StyleGAN2 directly to the eyewear domain poses two challenges. First, the original architecture targets square images and must be adapted for the 2:1 aspect ratio of front-facing eyeglass photographs. Second, downstream segmentation tasks require not only realistic images, but also accurate binary masks marking the frame region, information that a plain image generator does not provide.

This paper addresses both challenges by proposing a conditional StyleGAN2 variant that outputs paired (image, mask) tuples from a single forward pass. The key contributions are as follows.

- A conditional StyleGAN2 generator adapted to the non-square aspect ratio of eyeglass frames, producing paired RGB images and segmentation masks in a single forward pass, eliminating the need for manual mask annotation of generated images.
- A conditional discriminator that jointly supervises both image realism and mask accuracy, ensuring the two outputs are always geometrically consistent with each other.
- A five-mode style-embedding interpolation scheme that generates diverse frame shapes and colors beyond what is present in the training set, without mode collapse.
- A domain-specific evaluation protocol introducing edge coherence, mask regularity, and frame symmetry metrics that measure the physical realism of eyeglass frames, going beyond standard GAN metrics.
- A synthetic dataset of paired image–mask samples that can be directly used to train or augment segmentation and virtual try-on models in the eyewear domain, without any additional annotation effort.

The remainder of the paper is structured as follows. Section II reviews related work. Section III describes the proposed architecture and training procedure. Section IV presents the experimental results. Section V discusses the findings. Section VI concludes the paper.

II. Related Work

A. Generative Adversarial Networks for Image Synthesis

Since the seminal work of Goodfellow et al. [4], GANs have become the dominant paradigm for high-fidelity image synthesis. Conditional GAN (cGAN) [6] extended the framework to class-conditioned generation by feeding the label information to both the generator and the discriminator. Subsequent architectures such as DCGAN [7] demonstrated that deep convolutional networks stabilize adversarial training in real-world imagery, while Progressive GAN [8] showed that gradual resolution growth produces sharper high-resolution outputs.

The introduction of the style-based generator [9] marked a major improvement: by injecting a learned style code via Adaptive Instance Normalization (AdaIN) into each synthesis scale, the model achieved a controllable separation of coarse (pose, shape) and fine (color, texture) attributes. StyleGAN2 [5] refined this design by replacing AdaIN with weight demodulation, a reparameterization that eliminates the characteristic droplet artefacts of the first version, and adding a path-length regularization term that smooths the mapping from latent space to image space.

B. Conditional and Paired Image Synthesis

Pix2Pix [10] established the standard approach for image-to-image translation: a U-Net generator paired with a patch-level discriminator (PatchGAN) learns a mapping between aligned image pairs, supervised by a combination of adversarial and ℓ_1 reconstruction losses. Pix2PixHD [11] scaled the framework to high resolution using multi-scale discriminators and a perceptual feature-matching loss, demonstrating photo-realistic synthesis at 2048×1024 pixels.

For unpaired image translation, CycleGAN [12] introduced cycle-consistency constraints that remove the need for perfectly aligned training pairs. More recently, diffusion-based approaches [13] have achieved state-of-the-art perceptual quality in diverse generation tasks; however, their inference cost and limited controllability of auxiliary outputs (such as segmentation masks) make them less practical for building large-scale labeled datasets than GAN-based methods.

C. Synthetic Data for Visual Recognition

The effectiveness of GAN-generated data for training downstream models has been demonstrated in multiple domains. Wood et al. [14] showed that entirely synthetic face images with automatically rendered annotations could train competitive facial landmark detectors. In the medical imaging domain, Frid-Adar et al. [15] used DCGAN-generated liver lesion images to significantly improve CNN classification performance, demonstrating the practical value of GAN-based augmentation when real annotated data is scarce.

In the eyewear domain, most previous work focuses on virtual try-on [16] or face-based recommendation rather than frame detection or segmentation. The absence of a large-scale labeled dataset of eyeglass frame images motivates the present work.

D. Evaluation of Generative Models

The Fréchet Inception Distance (FID) [17] is the most widely adopted quantitative metric for generative models: it measures the Wasserstein-2 distance between the distributions of Inception-v3 features extracted from real and generated images. The Inception Score (IS) [18] evaluates sharpness and diversity using the conditional label distribution of an ImageNet-pretrained classifier, but is unreliable for domain-specific generators targeting categories absent from ImageNet. The Learned Perceptual Image Patch Similarity (LPIPS) [19] quantifies perceptual diversity using deep VGG features and complements FID as a diversity measure.

III. Methodology

A. Problem Formulation

Let $\mathcal{X} = \{(\mathbf{x}_i, m_i)\}_{i=1}^{n}$ denote a real dataset of n paired samples, where $\mathbf{x}_i \in \mathbb{R}^{3\times H\times W}$ is an RGB image of an eyeglass frame and $m_i \in \{0,1\}^{1\times H\times W}$ is the corresponding binary segmentation mask, with $H = 256$ and $W = 512$. The objective is to train a generator G such that, given a random noise vector $\mathbf{z}$ and a style index s, the pair $G(\mathbf{z}, s) = (\hat{\mathbf{x}}, \hat{m})$ follows the joint distribution $p(\mathbf{x}, m)$.

B. Generator Architecture

The proposed generator consists of three components: a mapping network, a style-modulated synthesis network, and a mask decoder branch.

Mapping Network. A per-image style embedding is concatenated with a pixel-normalized noise vector $\mathbf{z}$ and passed through a 4-layer MLP to produce the latent style code $\mathbf{w} \in \mathbb{R}^{512}$, making $\mathbf{w}$ dependent on both appearance and random fine detail.

Synthesis Network. The synthesis network begins with a learned 4×8 constant tensor and applies six progressive upsampling stages (bilinear $\times 2$ each) to reach the target resolution 256×512. Each stage consists of two style-modulated convolutional blocks (StyledConvBlock), followed by a skip-connected RGB projection. Style modulation uses weight demodulation [5], which modulates each convolution kernel by the style scale and then normalizes the result to unit output variance, eliminating the need for explicit normalization layers. Stochastic noise is injected per-channel via learned scalar weights to model fine texture variation. A self-attention module [20] is inserted in the resolution of the feature 32×64 to capture long-range spatial dependencies, which is particularly important for learning the bilateral symmetry of the eyeglass frames. The channel progression follows $512 \to 256 \to 256 \to 128 \to 64 \to 32 \to 16$.

Mask Decoder. A U-Net-style decoder is attached to the synthesis network and receives skip connections from the four shallowest synthesis stages (channels 128, 64, 32, 16). At each decoder step, the feature map is bilinearly upsampled and concatenated with the corresponding skip tensor before two convolutional layers with instance normalization and LeakyReLU activation. The final layer projects to a single channel followed by Sigmoid activation to produce the binary

mask probability map $\hat{m} \in [0,1]^{1\times H\times W}$. This architecture ensures that the mask is generated from the same internal representations as the image, enforcing geometric consistency between both outputs without any additional alignment loss.

C. Discriminator Architecture

The discriminator D processes a 5-channel input formed by concatenating the 3-channel RGB image, the 1-channel mask, and a 1-channel spatial style map. The style map is produced by projecting the style embedding onto a low-resolution tensor that is bilinearly upsampled to the full image resolution, providing D with explicit knowledge of the style variant to expect. Five residual downsampling blocks (DiscBlock) with spectral normalization [21] reduce the feature map from 256×512 to 4×8, followed by a minibatch standard deviation layer [8], which appends a feature map encoding the variation across the batch to help D detect mode collapse, and a fully connected scoring head. Self-attention is applied at the intermediate resolution of 256-channels. Spectral normalization on all convolutional and linear layers stabilizes the Lipschitz constant of D without requiring weight clipping.

D. Loss Functions

Adversarial Loss. Both generator and discriminator use the hinge adversarial loss [22]. For the discriminator:

$$\begin{aligned}\mathcal{L}_D = &\mathbb{E}\left[\text{ReLU}(1 - D(\mathbf{x}, m, s))\right] \\ &+ \mathbb{E}\left[\text{ReLU}(1 + D(\hat{\mathbf{x}}, \hat{m}, s))\right],\end{aligned} \tag{1}$$

and for the generator:

$$\mathcal{L}_G = -\mathbb{E}\left[D(\hat{\mathbf{x}}, \hat{m}, s)\right]. \tag{2}$$

The hinge formulation enforces a margin of 2 between real and fake scores, improving training stability over the standard GAN loss [5].

R1 Gradient Penalty. The R1 regularizer [23] penalizes large gradients of D with respect to real images:

$$\mathcal{L}_{R_1} = \frac{\gamma}{2}\,\mathbb{E}\left[\|\nabla_{\mathbf{x}} D(\mathbf{x})\|^2\right], \quad \gamma = 10, \tag{3}$$

preventing D from memorizing sharp decision boundaries around real samples. It is applied every 16 steps (amortized R1) to reduce cost.

Mask Supervision. Mask quality is supervised by a combination of Focal Binary Cross-Entropy [24] and soft Dice loss:

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{focal}}(\hat{m}, m) + \mathcal{L}_{\text{dice}}(\hat{m}, m). \tag{4}$$

The focal weight $(1 - p_t)^2$ reduces the influence of easily classified background pixels, focusing the gradient on frame boundaries. Dice loss measures global mask overlap and is not affected by class imbalance. The combined loss is weighted by $\lambda_m = 15$.

Perceptual Loss. A VGG-16 perceptual loss [11] computes ℓ_1 distances in the `relu1_2` and `relu2_2` feature spaces:

$$\mathcal{L}_{\text{perc}} = \sum_{l \in \{\text{relu1\_2},\, \text{relu2\_2}\}} \|\phi_l(\hat{\mathbf{x}}) - \phi_l(\mathbf{x})\|_1, \tag{5}$$

where $\phi_l(\cdot)$ is the feature map at layer l of a frozen VGG-16. Operating in feature space captures texture and edge similarity better than pixel-level comparison. The perceptual weight is $\lambda_p = 1.5$.

Total Generator Objective.

$$\mathcal{L}_G^{\text{total}} = \mathcal{L}_G + \lambda_m\,\mathcal{L}_{\text{mask}} + \lambda_p\,\mathcal{L}_{\text{perc}}, \quad \lambda_m = 15,\ \lambda_p = 1.5. \tag{6}$$

The adversarial term drives realism, the mask term enforces frame boundary accuracy, and the perceptual term ensures texture quality. The high mask weight $\lambda_m = 15$ prioritizes mask accuracy early in training.

E. Training Procedure

The model was trained on approximately 6,500 front-view eyeglass frame photographs with manually annotated masks, center-cropped to 2:1 and resized to 256×512. Both networks used Adam ($\beta_1 = 0$, $\beta_2 = 0.99$) with generator and discriminator learning rates of 2×10^{-4} and 1×10^{-4}, respectively. The batch size was 32, the gradients were clipped to the unit norm, and an EMA of the generator weights (decay 0.999) was used for all inference. Training proceeded for 350 epochs with checkpoints every 10 epochs.

F. Dataset Generation Strategy

To maximize the diversity of the generated dataset while avoiding mode collapse, five generation modes were employed:

- **Pure (25%):** a single training style embedding, covering all styles cyclically.
- **Interp2 (40%):** weighted blend of two style embeddings with Beta(0.4, 0.4) weights, producing bimodal blends biased toward either parent style.
- **Interp3 (20%):** Dirichlet(0.7) blend of three styles, generating smooth interpolations.
- **Interp4 (10%):** Dirichlet(0.5) blend of four styles for more complex combinations.
- **Intra (5%):** same style, different noise vectors, to stress-test intra-style diversity.

A truncation trick with $\psi = 0.7$ was applied by interpolating each $\mathbf{w}$-code toward the population mean, computed from 10,000 forward passes through the average style embedding. This trades a small amount of diversity for improved visual quality of the most extreme samples.

IV. Experimental Results

A. Training Dynamics

Fig. 1 summarizes the training curves over 350 epochs. Panels (a) and (d) show that discriminator and generator adversarial losses reached a stable balance, the discriminator score on real images stabilized near $+1.57$ and on fake images near -1.83 by epoch 350, consistent with the expected hinge loss margin, indicating that neither network collapsed or overpowered by the other. The mask loss panel (c) decreased steadily from 0.783 at epoch 1 to 0.088 at epoch 350, an 88.7% reduction, while the perceptual VGG loss fell from 0.774 to 0.497. The rapid improvement in mask quality during

the first 50 epochs reflects the high mask weight ($\lambda_m = 15$) that forces the model to learn the frame boundary geometry before refining the image texture. Panel (b) shows proxy IoU and Dice scores, which are noisy since masks are not spatially aligned during training. Training was stopped at epoch 350 because both adversarial loss and reconstruction loss had plateaued during the preceding 80 epochs (the perceptual loss changed by less than 1% between epochs 270 and 350), and visual inspection of the generated samples confirmed high and consistent output quality.

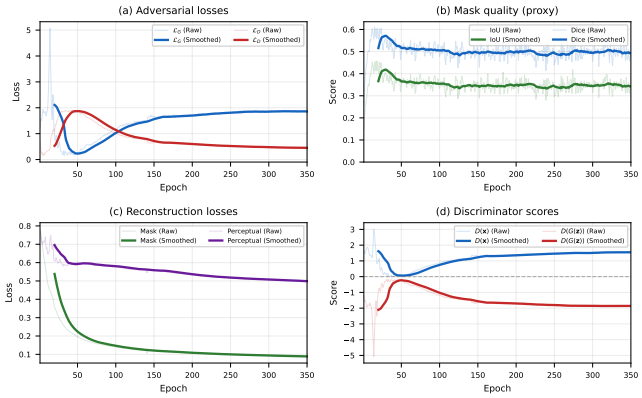

Fig. 1. Training metrics over 350 epochs. Thin light-colored lines show raw per-epoch values; thick colored lines represent 20-epoch moving averages. Panels display adversarial losses, mask quality proxies, reconstruction losses (mask and perceptual), and discriminator scores, respectively. Steady convergence indicates stable training with proper generator-discriminator balance.

B. Quantitative Evaluation

Table I presents the evaluation results computed on the full 10,000-image generated dataset. Real images were center-cropped and resized to 256×512 before feature extraction to ensure a fair comparison. Domain-specific metrics are computed within the masked frame region only and are therefore not affected by the background.

TABLE I
EVALUATION METRICS FOR THE GENERATED DATASET (10,000 SAMPLES)

Metric	Value	Rating
FID ↓	6.13	Excellent (< 10)
IS (mean ± std)*	1.29 ± 0.02	N/A*
LPIPS (mean ± std)	0.337 ± 0.042	Good (0.25–0.45)
Edge Coherence ↑	0.623 ± 0.036	Excellent (> 0.55)
Mask Regularity ↑	0.972 ± 0.112	Excellent (> 0.80)
Frame Symmetry ↑	0.875 ± 0.037	Excellent (> 0.85)
Colour Diversity ↑	0.100	Excellent (> 0.09)
Mask Shape Diversity ↑	0.095	Excellent (> 0.05)

*IS is unreliable for domain-specific generators; see text.

FID. A score of 6.13 is excellent for a narrow product domain [17], confirming that the generated distribution closely matches the real one. FID < 10 is the commonly accepted threshold for practical data augmentation quality.

IS. The score of 1.29 is low but expected, since the ImageNet classifier assigns nearly uniform probabilities to eyeglass frames, which are absent from its 1,000 categories [18]. We include it only for completeness.

LPIPS. A score of 0.337 ± 0.042 within the typical range of 0.25–0.45 confirms that generated images are perceptually dissimilar from real ones, showing no memorization of training samples.

Domain-Specific Metrics. Edge coherence (0.623) confirms that mask boundaries align with visible frame edges in the RGB image. Mask regularity (0.972) shows that almost all masks form a single clean contour with minimal fragmentation. Frame symmetry (0.875) confirms the generator produces physically realistic frame shapes.

Diversity. Color diversity (0.100) and mask shape diversity (0.095) both exceed their excellent thresholds, confirming the five-mode strategy avoids mode collapse across frame geometries and colors.

C. Qualitative Results

Fig. 2 shows generated RGB samples with the corresponding binary masks, sampled at epoch 350. The generated frames show diverse shapes (cat-eye, rectangular, round, oval, rimless), a wide range of materials (solid plastic, metallic wire, translucent acetate), and varied color palettes. The masks consistently outline the frame.

Fig. 2. Generated samples at epoch 350 arranged in alternating rows. Each pair shows RGB image (top row) and corresponding binary segmentation mask (bottom row). Diverse frame shapes, materials, and color palettes are represented.

A subset of generated samples exhibit visible artefacts such as color bleeding, disconnected fragments, inverted text artefacts, and incomplete frame structure (Fig. 3). Failures are most noticeable for structurally sparse styles such as rimless frames, where missing rim structure leaves the generator without sufficient spatial guidance, but also occur in fully-rimmed frames with complex textures or fine details. These issues are consistent with the limited size and style coverage

of the training corpus and could be solved by collecting additional annotated examples of challenging frame types or by introducing shape-label conditioning to guide generation.

Fig. 3. Examples of failure cases including disconnected fragments, color bleeding, inverted text artefacts, and incomplete frame structure, particularly for rimless and structurally sparse frame styles.

V. Discussion

The joint generation approach ensures geometric consistency since both outputs share internal representations. Edge coherence (0.623) confirms mask boundaries align with color discontinuities in the RGB image.

One limitation is that the style embedding table is fixed at the size of the training set, constraining style variety. Style mixing through interpolation mitigates this, but cannot fully substitute for a larger and more diverse training corpus.

VI. Conclusion

This paper presented a conditional StyleGAN2 architecture for the joint synthesis of RGB images and binary segmentation masks of eyeglass frames at 256×512 resolution. The proposed generator combines style-modulated convolutional synthesis, self-attention for long-range symmetry modeling, and a U-Net mask decoder, while the conditional discriminator jointly evaluates visual realism and mask-image consistency. A mixed style-interpolation generation strategy produces a 10,000 sample dataset with excellent FID (6.13), edge coherence (0.623), mask regularity (0.972), frame symmetry (0.875), and diversity scores.

The generated dataset addresses the critical shortage of labeled eyewear imagery and provides a practical data source for training segmentation, virtual try-on, and recommendation models in the eyewear domain.

Future work will focus on expanding the training corpus, adding shape-label conditioning, and evaluating the dataset on downstream segmentation tasks.

References

[1] H. Giedra, T. Sledevič, and D. Matuzevičius, "Deploying optimized deep vision models for eyeglasses detection on low-power platforms," *Electronics*, vol. 14, no. 14, p. 2796, 2025.

[2] T. Sledevič, A. Serackis, and D. Plonis, "FPGA implementation of a convolutional neural network and its application for pollen detection upon entrance to the beehive," *Agriculture*, vol. 12, no. 11, p. 1849, 2022.

[3] G. Vdoviak, T. Sledevič, A. Serackis, D. Plonis, D. Matuzevičius, and V. Abromavičius, "Evaluation of deep learning models for insects detection at the hive entrance for a bee behavior recognition system," *Agriculture*, vol. 15, no. 10, p. 1019, 2025.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 27, 2014.

[5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8110–8119, 2020.

[6] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[9] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, 2017.

[11] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8798–8807, 2018.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2223–2232, 2017.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[14] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, M. Johnson, V. Nair, J. Sheratt, and J. Shotton, "Fake it till you make it: Face analysis in the wild using synthetic data alone," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3681–3691, 2021.

[15] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[16] H. Jia, Y. Zheng, C. Liu, and D. Tao, "Virtual try-on of glasses with augmented reality," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 1–5, 2020.

[17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017.

[18] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 29, 2016.

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

[20] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 7354–7363, 2019.

[21] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[22] J. H. Lim and J. C. Ye, "Geometric GAN," *arXiv preprint arXiv:1705.02894*, 2017.

[23] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3481–3490, 2018.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.