

Article

Not peer-reviewed version

Benchmarking Conditional GANs in Industrial Marble Texture Synthesis via a Dual-Evaluation Framework

[António Alves de Campos](#)*, Margarida Figueiredo, [Carlos M. A. Diogo](#), [Gustavo Paneiro](#), [Pedro Amaral](#)

Posted Date: 10 February 2026

doi: 10.20944/preprints202602.0818.v1

Keywords: conditional generative adversarial networks; texture synthesis; deep learning; image processing; computer vision; perceptual quality assessment; industrial quality control; human evaluation



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Benchmarking Conditional GANs in Industrial Marble Texture Synthesis via a Dual-Evaluation Framework

António Alves de Campos ^{1,*}, Margarida Figueiredo ², Carlos M. A. Diogo ¹, Gustavo Paneiro ¹ and Pedro Amaral ³

¹ CERENA, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

² Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

³ LAETA, IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

* Correspondence: antonio.campos@tecnico.ulisboa.pt

Featured Application

This framework enables multiple stakeholder archetypes across the marble industry's value chain to generate photorealistic texture variants for virtual prototyping and architectural visualization without costly manual annotation, thereby reducing design iteration cycles while maintaining human-indistinguishable quality.

Abstract

Generative Adversarial Networks (GANs) have demonstrated remarkable capabilities for synthesizing photorealistic textures, yet deploying conditional GANs (cGANs) in industrial settings faces two barriers: the prohibitive cost of annotating proprietary data and the uncertain alignment between automated metrics and human perception. This study addresses both challenges for marble texture synthesis. We adapt an unsupervised segmentation pipeline combining Simple Linear Iterative Clustering (SLIC) superpixels, Gaussian Mixture Models (GMMs), and Graph Cut optimization to extract vein structures from 289 industrial scans without manual annotation. We then benchmark four cGAN architectures, a baseline cGAN, Pix2Pix, BicycleGAN, and GauGAN, using a dual-evaluation protocol contrasting automated assessment via pixel-based metrics, structural metrics, statistical metrics, and learned distributional metrics with human-centered assessment. Results reveal a significant metric-perception discrepancy: Pix2Pix achieved the best FID yet received the lowest human ratings due to checkerboard artifacts, whereas GauGAN produced textures statistically indistinguishable from real marble (Visual Turing Pass Rate, VTPR: 0.533; Mean Opinion Score on Marble Authenticity, MOS-MA: 2.89) despite inferior FID (87.3). These findings establish three contributions: (1) an unsupervised annotation-free segmentation pipeline, (2) empirical evidence that automated metrics alone are insufficient for architecture selection, and (3) a dual-evaluation framework validating human-in-the-loop assessment as essential for industrial deployment.

Keywords: conditional generative adversarial networks; texture synthesis; deep learning; image processing; computer vision; perceptual quality assessment; industrial quality control; human evaluation

1. Introduction

In the stone processing and construction industries, digital transformation has created demand for high-fidelity virtual material representations to support virtual prototyping, digital twin applications, and mass customization workflows [1,2]. Natural stone textures, particularly marble

with its stochastic vein patterns, pose unique synthesis challenges: each slab exhibits non-repeating structures requiring both photorealistic rendering and precise designer control [3]. However, obtaining large-scale annotated datasets for training conditional generative models is extremely challenging in industrial contexts due to the prohibitive cost of manual pixel-level annotation, the proprietary nature of production data, and limited batch sizes typical of specialty materials [4]. This data scarcity problem represents a critical barrier to deploying deep learning solutions for texture synthesis in manufacturing environments.

While conditional Generative Adversarial Networks (cGANs)[5] have demonstrated impressive capabilities for image-to-image translation tasks [6,7], their application to industrial texture synthesis faces two unresolved challenges. First, existing approaches assume the availability of ground-truth semantic masks, an assumption that does not hold for proprietary manufacturing data, where manual annotation would require weeks of skilled labor [4,8]. Second, the standard evaluation paradigm relies exclusively on automated metrics, such as Fréchet Inception Distance (FID) [9], Inception Score (IS) [10], and Multi-Scale Structural Similarity Index Measure (MS-SSIM), which were originally designed for object recognition rather than texture quality. Recent studies have shown that these metrics exhibit weak or negative correlations with human perceptual judgments in texture synthesis tasks. The Inception-v3 network underlying FID was trained for object classification on ImageNet, making it architecturally designed to be texture-invariant and thus insensitive to texture artifacts (e.g., checkerboard patterns, repetitive microstructures) that are perceptually obvious to human observers [11]. This metric-perception discrepancy has been rigorously documented, Zhou et al. (2019) [12] demonstrated that across multiple datasets FID scores do not fully correlate with the scores of evaluations made by humans, while the largest human evaluation study to date, comprising over 207,000 perceptual judgments across 41 generative models, concluded that existing metrics don't show strong correlations with human evaluations [13,14]. Borji (2022) [14] further documented that FID has a blind spot regarding image quality and is particularly unsuitable for specialized domains where visual features differ substantially from those of ImageNet. Yet systematic analyses have demonstrated that FID exhibits high sensitivity to low-level image statistics, such as noise and blur, that fail to correlate with human perception, while remaining largely insensitive to high-level artifacts that human observers immediately detect [13]. This fundamental misalignment has reinforced the position that human preference constitutes the ultimate ground truth for evaluating AI-generated content [15]. The present study addresses two critical gaps: first, it provides direct empirical evidence of metric-perception divergence across conditional GAN architectures in an industrial texture synthesis context; second, it introduces an annotation-free segmentation pipeline integrating SLIC, GMM, and Graph Cut optimization, thereby eliminating the manual labeling bottleneck that has historically constrained the application of conditional generative models to proprietary manufacturing datasets.

The evolution of texture synthesis spans three major paradigms. Traditional procedural generation, exemplified by Perlin noise [16,17], can produce marble-like patterns through mathematical algorithms but often with limited realism and requiring expert tuning[18,19]. Example-based non-parametric methods leverage real source images to generate textures but struggle with large-scale structures like continuous marble veins [20–24]. Modern deep generative models have revolutionized image synthesis [25,26], with GANs emerging as efficient alternatives [27]. Neural style transfer [11,28] established the principle of separating structure and appearance, while Spatial GANs enabled scalable texture generation [29,30].

GANs enable controllable synthesis by conditioning generation on structural inputs like semantic masks. The Pix2Pix framework [31] established paired image-to-image translation using U-Net generators and PatchGAN discriminators, combining adversarial loss with L1 reconstruction to enforce structural fidelity. BicycleGAN [32] addresses mode collapse by enforcing bijective mappings between latent codes and outputs, enabling diverse generations. The most significant innovation for mask-conditioned synthesis is Spatially-Adaptive Normalization (SPADE) in GauGAN [6,31], which modulates normalization parameters as learned functions of the input mask at every layer,

preventing semantic information from being washed away by standard normalization, critical for preserving vein boundaries while synthesizing organic appearance [33].

While diffusion models have achieved state-of-the-art results, their deployment faces practical barriers in manufacturing: inference is significantly slower than GANs, and conditional variants like ControlNet [34] rely on foundation models pre-trained on billions of images. For manufacturing environments with limited proprietary data (~200–500 samples) and computational constraints, cGANs represent a practical choice. Comprehensive comparison with fine-tuned foundation models would require different experimental protocols and constitutes valuable future work [35].

The annotation bottleneck has motivated unsupervised segmentation research in medical imaging [36–38], yet adoption in manufacturing remains limited [4,8]. Unsupervised methods combine classical computer vision techniques in a principled pipeline: Simple Linear Iterative Clustering (SLIC) superpixels [39,40] reduce computational complexity by grouping pixels into perceptually coherent regions while preserving boundaries [41]; Gaussian Mixture Models (GMM) then probabilistically classify these regions into semantic classes based on color features [42,43]; finally, Graph Cut algorithms [43] or Normalized Cuts [44] enforce spatial coherence through energy minimization, removing segmentation noise while respecting natural material boundaries. Interactive methods like GrabCut [45] offer additional refinement capabilities. This work resorts to Borovec et al.'s pipeline [37] to natural stone, demonstrating that the same approach can automatically extract marble vein structures suitable for cGAN training without manual annotation.

Recent industrial applications demonstrate the transformative potential of GANs. In manufacturing quality control, GANs address data scarcity for defect detection [46] and anomaly detection [47,48]. Beyond inspection, GANs contribute to design optimization [49], product lifecycle prediction [50], and digital twin systems [51]. Strategic applications include technology road mapping [52] and custom material design [53,54]. However, prior work on natural materials like marble remains scarce [55,56], with most studies focusing on regular repeating patterns rather than stochastic geological textures.

To overcome these limitations, we propose a dual-evaluation framework for benchmarking conditional GANs in industrial texture synthesis that addresses both the annotation bottleneck and evaluation uncertainty. Our framework integrates: (1) an adapted unsupervised segmentation pipeline [37] that automatically extracts structural masks from raw production scans, eliminating manual annotation costs; and (2) a rigorous human-centered validation protocol combining Visual Turing Tests [10] and Mean Opinion Scores adapted from telecommunications standards [[57,58] to complement standard automated metrics [59]. To the best of our knowledge, this represents the first systematic application of dual-protocol evaluation (automated + human) to industrial material texture synthesis, and the first demonstration that unsupervised mask generation enables conditional GAN training for natural stone without manual labeling.

This study focuses on a marble type from a single quarry, marketed under the commercial name Exotic Ambar, providing a controlled testbed for architectural comparison without confounding variables and reflecting authentic industrial deployment scenarios in which manufacturers specialize in specific materials. We systematically compare four conditional GAN architectures, baseline cGAN, Pix2Pix, BicycleGAN, and GauGAN, selected for their shared architectural lineage (U-Net generators, PatchGAN discriminators) and trainability on modest datasets (~300 samples) using accessible GPU resources, unlike StyleGAN[60] or foundation models requiring extensive tuning or billions of parameters. The main contributions of this study are:

1. We validate an unsupervised segmentation pipeline (SLIC + GMM + Graph Cut) for automatically generating semantic masks from marble imagery, demonstrating a practical solution to the data scarcity challenge where obtaining pixel-perfect annotations is economically prohibitive.
2. We conduct a systematic benchmark comparing four cGAN architectures trained and evaluated under identical conditions on real industrial data (289 high-resolution marble scans acquired from production lines), providing evidence-based architecture selection guidance for practitioners.

3. We implement a dual-evaluation framework contrasting automated metrics (FID, IS, MS-SSIM) with human-centered assessment (Visual Turing Test [10], Mean Opinion Scores from domain experts), revealing significant metric-perception discrepancies with direct implications for deployment decisions.

4. We demonstrate that GauGAN achieves human-indistinguishable synthesis quality despite inferior FID scores, while Pix2Pix exhibits the opposite pattern, establishing empirically that automated metrics alone are insufficient for architecture selection in quality-critical manufacturing applications [61,62].

5. We provide comprehensive methodology documentation to enable replication and extension to other natural material synthesis tasks (wood, fabric, geological samples).

The remainder of this paper is organized as follows. Section 2 details our methodology: data collection, the unsupervised segmentation pipeline, cGAN implementations, and the dual-evaluation protocol. Section 3 presents the results: visual comparisons, automated metrics, human-evaluation outcomes, metric-perception discrepancy analysis, and computational performance. Section 4 discusses practical implications for industrial deployment and limitations. Section 5 concludes with actionable recommendations and directions for future research.

2. Materials and Methods

This study introduces a comprehensive pipeline for controllable marble texture synthesis that addresses two critical deployment barriers: the prohibitive cost of manual annotation for training conditional generative models and the inadequacy of automated metrics for validating perceptual quality in texture-synthesis applications. The methodology consists of three integrated components validated on real industrial scan data: an unsupervised segmentation pipeline that automatically generates conditioning masks, a systematic benchmarking of four conditional GAN architectures trained on these masks, and a dual-evaluation framework combining automated metrics with structured human assessment protocols adapted from telecommunications.

The complete methodological workflow is illustrated in Figure 1. The process initiates with data curation and pre-processing, where raw industrial scans are filtered and standardized. Next, the semantic mask generation stage employs a multi-step unsupervised algorithm to extract the binary vein structure from each image. These image-mask pairs are then used in the conditional GAN implementation phase, which involves both the construction and adversarial training of the generative models. The final stage is a comprehensive performance evaluation, where the models are systematically compared using a dual framework of human-centered qualitative assessments and objective quantitative metrics.

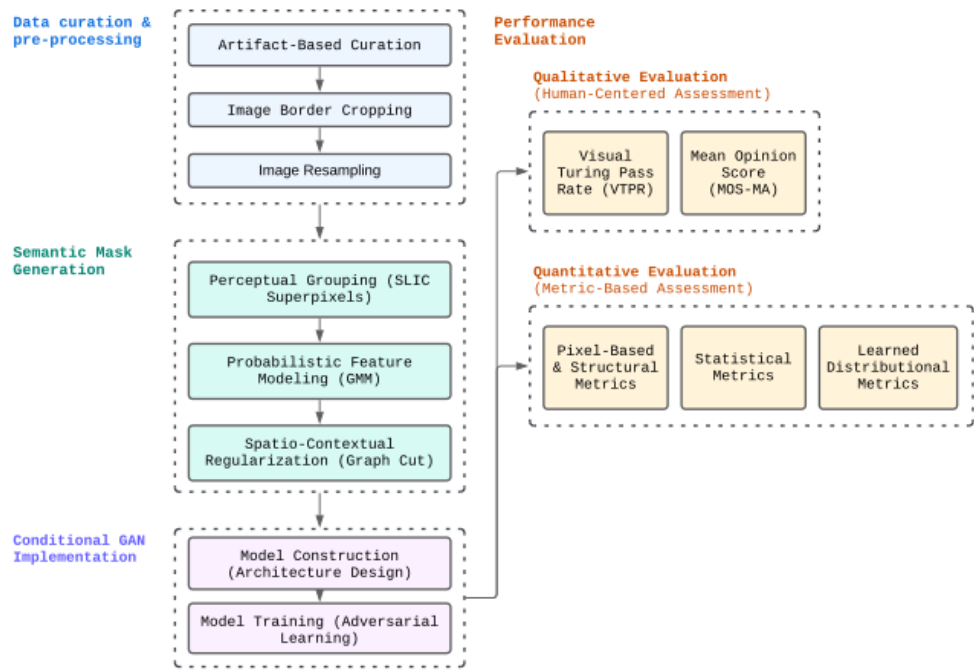


Figure 1. Flowchart of the proposed pipeline for controllable marble texture synthesis. The workflow proceeds through four stages: (1) Data curation and pre-processing, including artifact-based filtering and image standardization; (2) Semantic mask generation via SLIC superpixel segmentation, GMM-based probabilistic modeling, and Graph Cut spatial regularization; (3) Conditional GAN implementation, encompassing model construction and adversarial training; (4) Performance evaluation combining qualitative human-centered assessment (VTPR, MOS-MA) with quantitative automated metrics (pixel-based, statistical, and learned distributional).

2.1. Dataset and Unsupervised Mask Generation

The dataset comprises 289 high-resolution images of Exotic Ambar marble slabs captured on an industrial production line using a factory-calibrated line-scan camera. Each slab measures 0.5–2.5 m in longest dimension, scanned at mean resolution 7185×4166 pixels with controlled illumination. From an initial set of 327 scans, 38 samples (12%) exhibiting protective film artifacts or scanner malfunctions were excluded through visual inspection, ensuring the dataset reflects genuine marble appearance variation rather than imaging defects. Examples of excluded samples are documented in Appendix A.4. All images underwent standardized pre-processing: 200-pixel border cropping to remove frame artifacts, bicubic resampling to 1280×720 pixels, and normalization to [-1, 1] range. The dataset was deterministically split into 232 training (80%) and 57 validation (20%) samples, with no data augmentation applied to avoid interpolation artifacts at vein boundaries. The entire pre-processing pipeline was implemented using TensorFlow’s API to ensure bit-wise identical handling of data during both training and inference.

The central challenge addressed here is economic feasibility. Manual pixel-level annotation of marble veins requires specialized expertise, and the time-consuming process of obtaining precise annotations restricts the scalability and practicality of supervised approaches in industrial contexts [4]. Supervised deep learning approaches, such as U-Net, are therefore impractical for specialty materials with limited production volumes. To circumvent this annotation bottleneck, we implemented an unsupervised three-stage segmentation pipeline combining established computer vision techniques: SLIC superpixel over-segmentation, Gaussian Mixture Model color clustering, and graph cut spatial regularization.

The SLIC algorithm first reduces each image to approximately 3,000 perceptually uniform superpixels by clustering in 5D CIELAB-spatial space (nominal size 20 px, compactness 0.3),

preserving vein boundaries while reducing computational complexity. Superpixels provide a mid-level representation that captures local texture and color homogeneity, providing a robust set of primitive regions for further analysis [42]. Each superpixel is characterized by a 9-dimensional feature vector encoding CIELAB mean, standard deviation, and median. A two-component Gaussian Mixture Model trained via Expectation-Maximization provides initial probabilistic class assignments (vein vs. matrix) based solely on color distribution. However, simple color-based clustering often fails in practice for natural materials due to substantial variation in hue, saturation, and brightness across regions of a single slab. These raw probabilities are then refined using Maximum A Posteriori estimation in a Markov Random Field [22], where the optimal binary labeling minimizes an energy function balancing GMM data fidelity against spatial smoothness (regularization weight $\lambda = 5.0$). This energy minimization is solved globally via graph cut optimization, yielding spatially coherent masks that preserve delicate vein bifurcations while suppressing isolated noise. Graph cut efficiently finds the minimum-cut solution that best respects both the clustering cues and spatial continuity, producing a clean segmentation of veins versus matrix. Alternative graph-based methods include Normalized Cuts for balanced partitions and GrabCut for interactive foreground-background separation. All 289 generated masks were visually inspected and accepted without manual correction, demonstrating the pipeline's robustness across diverse vein densities, orientations, and matrix colorations characteristic of the marble with commercial name Exotic Ambar. Figure 2 shows four representative examples from the dataset: the top row shows marble slabs exhibiting natural variation in vein patterns, while the bottom row displays the corresponding binary masks produced by the unsupervised segmentation pipeline. The high-quality segmentation of fine vein structures without manual intervention validates the pipeline's suitability for large-scale industrial deployment. A detailed visualization of all pipeline stages is provided in Appendix A.1.

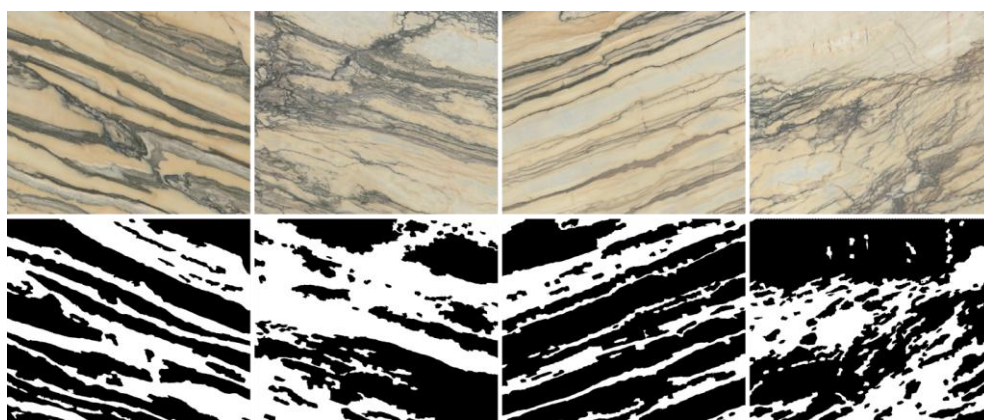


Figure 2. Representative marble slabs and unsupervised mask generation. Top row: Four samples showing natural variation in vein density and orientation from dataset of the marble with commercial name Exotic Ambar. Bottom row: Corresponding binary masks generated automatically via SLIC + GMM + Graph Cut pipeline.

2.2. Conditional GAN Architectures and Training

To establish a reproducible benchmark for mask-conditioned texture synthesis, we trained four seminal conditional GAN architectures representing evolutionary advances in conditioning strategies: the original conditional GAN [5], Pix2Pix [7], BicycleGAN [32], and GauGAN with Spatially-Adaptive Normalization [6]. All models employ identical PatchGAN discriminators (70×70 receptive field) to isolate architectural differences in the generators. Complete architecture diagrams are provided in Appendix A.2.

The baseline conditional GAN uses a U-Net generator (8 encoder blocks with skip connections to 7 decoder blocks) conditioned on both the binary vein mask and a 100-dimensional latent vector sampled from a standard normal distribution, with dropout (rate 0.5) in the first three decoder layers to promote output diversity. The U-Net encoder-decoder framework is renowned for its efficacy in

image-to-image translation tasks due to skip connections that preserve high-frequency spatial details during reconstruction.

Pix2Pix represents a deterministic variant that removes explicit latent sampling, instead relying solely on the input mask with dropout during inference providing mild stochasticity, while adding an L1 pixel-wise reconstruction loss ($\lambda = 100$) to the adversarial objective to enforce fidelity to a paired ground-truth image. Pix2Pix demonstrated that a single cGAN framework can synthesize plausible photos from diverse input representations, effectively separating content from style [7].

BicycleGAN extends this framework by introducing a dedicated encoder network that learns to invert generated images back to latent codes, enforcing a bijective latent-to-image mapping through cycle consistency ($\lambda_{\text{latent}} = 10$) to combat mode collapse and enable diverse texture generation from identical mask inputs. This addresses the one-to-many mapping inherent in image translation, where a given mask could correspond to multiple realistic appearances.

GauGAN represents a fundamental architectural departure: rather than using a standard U-Net encoder, synthesis begins from a learned constant tensor progressively upsampled through six residual blocks, each incorporating Spatially-Adaptive Normalization (SPADE) layers that modulate normalization parameters (γ, β) as learned spatial functions of the input mask, thereby preserving semantic structure at every layer without the washing away effect of standard batch normalization that plagues earlier architectures. SPADE ensures that the output strictly conforms to the specified structure while the network fills in photorealistic textures and details, achieving state-of-the-art fidelity in synthesizing images from segmentation maps.

We implemented all models in TensorFlow and trained them on a single NVIDIA H100 GPU. Training employed Adam optimization with $\beta_2 = 0.999$, batch size 4, and architecture-specific learning rates ranging from 5×10^{-5} to 2×10^{-4} determined through preliminary hyperparameter sweeps. All network weights were initialized from a Gaussian distribution with mean 0 and standard deviation 0.02, a standard GAN initialization. Regularization included one-sided label smoothing, gradient clipping, and early stopping with patience of 2,000 epochs, monitoring validation FID [63]. Models were trained for up to 10,000 epochs with convergence-based termination when validation metrics plateaued. Table 1 reports the optimized hyperparameter configurations. These values were determined through preliminary sweeps optimizing validation FID.

Table 1. Final hyperparameter settings per architecture, determined through validation FID optimization.

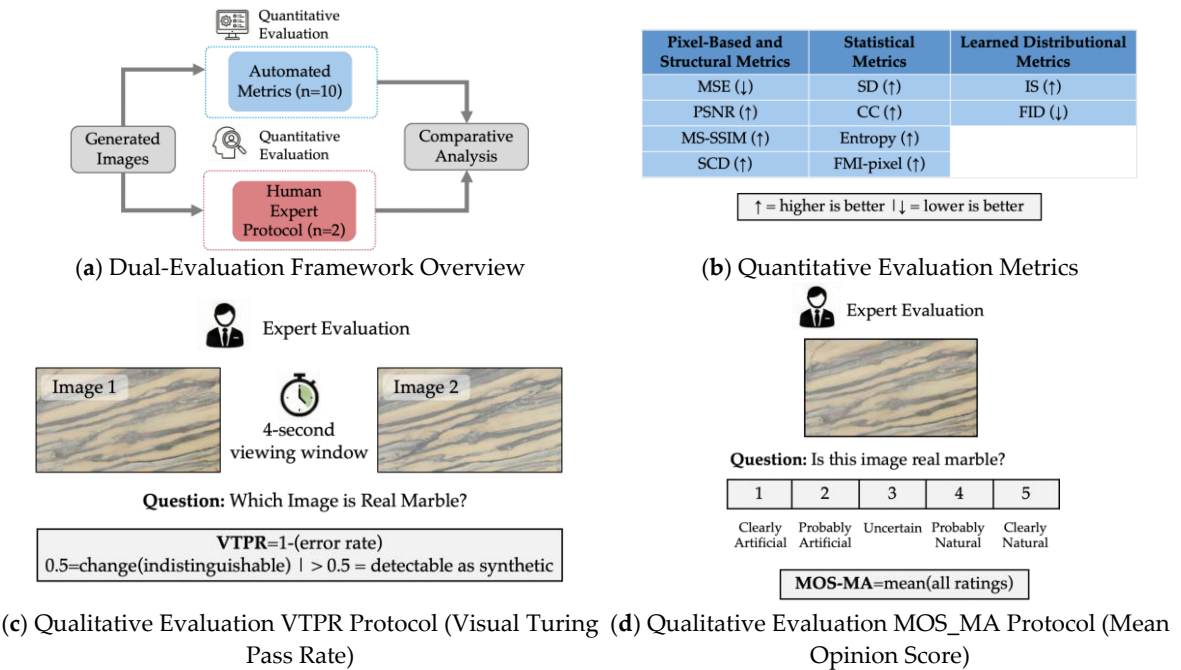
Architecture	Generator LR	Discriminator LR	Encoder LR	λ_{L1}	λ_{latent}	Latent Dim	Total Training Epochs
cGAN	1×10^{-4}	5×10^{-5}	—	—	—	100	6,900
Pix2Pix	2×10^{-4}	2×10^{-4}	—	100	—	—	5,100
BicycleGAN	2×10^{-4}	2×10^{-4}	2×10^{-4}	100	10	8	3,400
GauGAN	5×10^{-5}	5×10^{-5}	—	350	—	—	3,100

2.3. Dual-Evaluation Framework

Standard GAN evaluation protocols rely almost exclusively on automated metrics, particularly FID and IS, computed from Inception-v3 features originally trained for object classification on ImageNet. However, systematic studies have demonstrated fundamental limitations of this approach. Zhou et al. (2019) [12] established the HYPE benchmark, a rigorous human evaluation framework grounded in psychophysics research, showing that automated metrics serve as noisy indirect proxies for perceptual quality. Stein et al. (2023) [13] conducted the largest human evaluation study to date with over 1,000 participants and 207,000 judgments, finding that FID computed with Inception-v3 fails to generalize beyond ImageNet-like distributions and that models producing more perceptually realistic images paradoxically score worse on FID. Borji (2022) [14], in a comprehensive survey of over 24 evaluation measures, documented that FID has a blind spot for image quality and that reliable evaluation is particularly relevant in domains where humans can be less attuned to evaluate quality samples, directly relevant to industrial texture synthesis, where domain expertise

shapes quality perception. Despite this evidence, most recent industrial GAN papers defer human evaluation to future work, relying solely on FID and SSIM [59], a practice inadequate for quality-critical applications in which end-user perception determines deployment success.

Our methodological contribution is a dual-evaluation framework (Figure 3 (a)) that systematically compares automated metrics against structured human assessment, establishing whether metric-based optimization aligns with perceptual authenticity in the industrial texture synthesis context. The quantitative component (Figure 3 (b)) employs a battery of 10 automated metrics spanning three complementary families. Pixel-based and structural metrics include Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR), which quantify pixel-level reconstruction fidelity; Multi-Scale Structural Similarity Index (MS-SSIM), which assesses perceptual structural similarity across multiple scales; and Structural Content Dissimilarity (SCD), which measures structural differences in content representation. Statistical metrics include Standard Deviation (SD), Correlation Coefficient (CC), Entropy (EN), and Feature Mutual Information (FMI-Pixel) [64], which quantify texture characteristics such as intensity distribution variability, linear dependence between generated and real images, information content, and feature co-occurrence patterns. The learned distributional metrics employed, namely IS and FID, measure high-level feature similarity using Inception-v3 embeddings. These metrics provide complementary perspectives on reconstruction accuracy, structural preservation, texture statistics, and learned feature similarity, with FID serving as the primary convergence criterion during training.



The qualitative component implements two human-centered protocols adapted from telecommunications quality assessment standards [57,58]. Protocol 1 (Visual Turing Pass Rate, VTPR; Figure 3 (c) operationalizes the adversarial objective [10,27] following psychophysical best practices established by Zhou et al. (2019) [12], who demonstrated that time-constrained evaluation leveraging the ~150ms threshold for human image processing provides reliable discrimination. We present three domain experts (stone-processing engineers with 10+ years of experience) with 20 each two-alternative forced-choice (2AFC) trials: each trial displays one real and one synthetic marble image in random order for 4 seconds, and experts identify which is real. This time constraint captures glance-level authenticity corresponding to feedforward visual processing [65] while preventing extended artifact-focused scrutiny, consistent with the HYPE benchmark methodology that achieved strong statistical reliability at approximately \$60 per assessment [12]. VTPR is computed as 1 minus the mean identification error rate, where 0.5 indicates perfect indistinguishability (chance performance) and values exceeding 0.5 indicate detectably synthetic outputs. Protocol 2 (Mean Opinion Score on Marble Authenticity, MOS-MA; Figure 3 (d) provides graded quality assessment [58]: the same three experts rate 20 images (15 generated plus 5 real quality-control samples) on a 5-point Likert scale (1 = “Clearly Artificial”, 5 = “Clearly Natural”) without time constraints. MOS-MA is computed as the mean score across all ratings, with standard error and 95% confidence intervals assuming independent expert judgments. Together, these protocols provide both forced discrimination (VTPR) and absolute quality scaling (MOS-MA), enabling detection of architectures that achieve metric-based optimization at the expense of perceptual authenticity, the core hypothesis motivating this evaluation strategy.

3. Results

Results are organized to demonstrate the three core claims of this work: (1) unsupervised mask generation enables training without annotation cost, (2) human-centered evaluation reveals architecture-specific perceptual quality that automated metrics fail to capture, and (3) a critical divergence exists between metric-based optimization and human perception in industrial texture synthesis. We begin with qualitative visual assessment of generated textures, present comprehensive quantitative metrics, introduce structured human evaluation findings, and conclude with computational efficiency analysis to inform deployment decisions.

3.1. Qualitative Assessment: Visual Comparison Across Architectures

All four conditional GAN architectures successfully learned to synthesize photorealistic marble textures from binary vein masks after training on the 232-sample dataset with unsupervised annotations. Figure 4 presents a systematic comparison across four representative samples with varying vein patterns: given identical mask inputs (column a), each architecture generates plausible marble appearances exhibiting correct vein placement, naturalistic matrix coloration, and appropriate texture granularity relative to ground-truth real marble images (column b).

Visual inspection reveals distinct architectural characteristics. The baseline conditional GAN (column c) produces diverse outputs due to explicit latent sampling, successfully generating variation in matrix tone and vein texture while maintaining structural fidelity to the input mask. Pix2Pix (column d) generates sharp, high-contrast outputs with excellent mask adherence and strong vein definition, producing textures that appear highly realistic at standard viewing distances. BicycleGAN (column e) successfully produces outputs with controlled diversity through its latent embedding mechanism, though this diversity sometimes manifests as variation in global lighting and color temperature rather than localized material texture properties. GauGAN (column f) exhibits smooth texture quality with particularly organic vein-to-matrix transitions, producing outputs where the boundary between vein and matrix regions appears naturally graduated rather than sharply delineated.

Across all four samples spanning different vein geometries, from parallel diagonal structures (Samples 1-2) to curved organic patterns (Sample 3) and complex scattered networks (Sample 4), the

architectures demonstrate consistent synthesis capability. Each architecture maintains its characteristic visual signature across diverse structural inputs, indicating that observed quality differences stem from fundamental architectural design choices rather than mask-specific overfitting. All generated outputs are visually plausible and structurally faithful to their conditioning masks, validating the efficacy of the unsupervised segmentation pipeline for providing geometric guidance to the generative process.

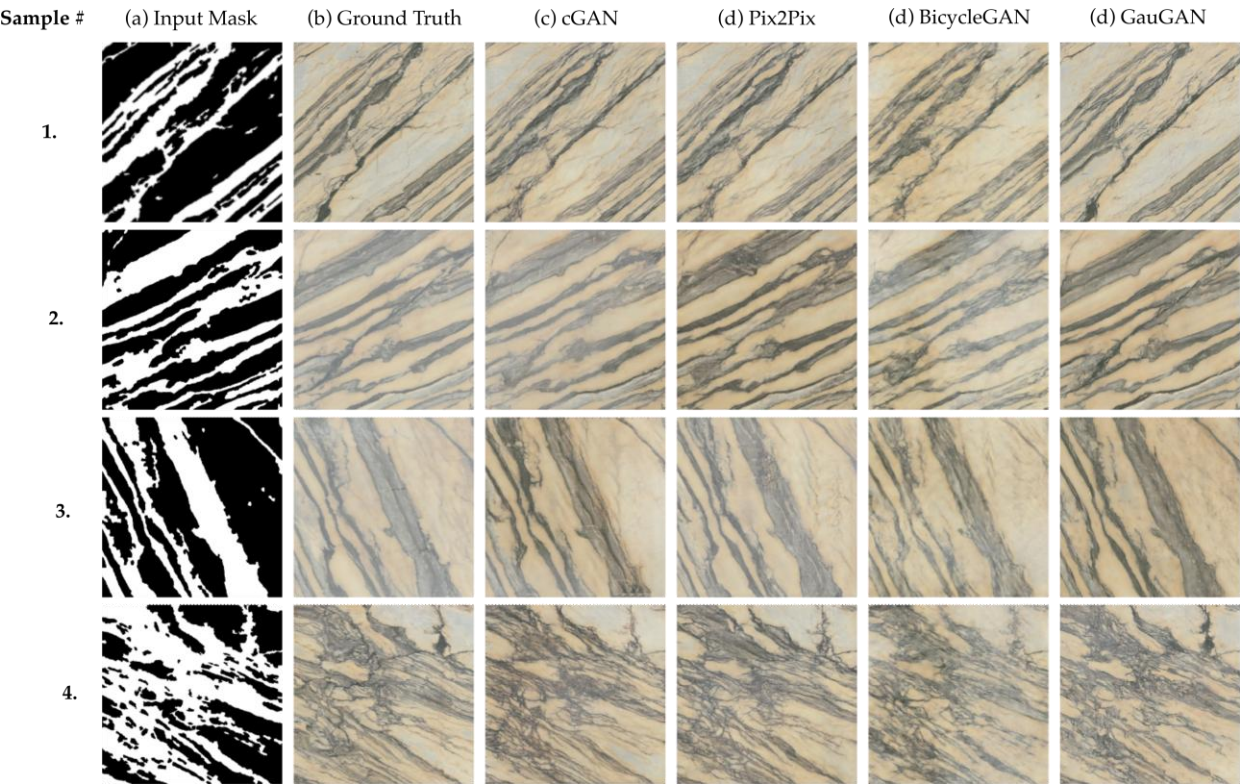


Figure 4. Qualitative comparison of marble texture synthesis across four cGAN architectures. (a) Input binary mask. (b) Ground-truth real marble. (c) cGAN output. (d) Pix2Pix output. (e) BicycleGAN output. (f) GauGAN output. Rows show samples with varying vein density and orientation.

3.2. Quantitative Metrics: Automated Performance Evaluation

Table 1 presents comprehensive quantitative evaluation across 10 automated metrics computed on all 57 validation samples. Pix2Pix achieved the best performance on distributional metrics, including the widely-used Fréchet Inception Distance (FID = 85.286, 2.3% lower than GauGAN) and Inception Score (IS = 1.940). This consistent metric superiority reflects Pix2Pix’s L1 reconstruction loss enforcing pixel-level fidelity to ground-truth training data. GauGAN ranked second on distributional metrics but demonstrated the strongest performance on reconstruction fidelity measures: structural similarity (MS-SSIM = 0.713), pixel accuracy (PSNR = 22.626 dB, MSE = 0.006), correlation (CC = 0.847), and mask adherence (FMI-pixel = 0.886). BicycleGAN and baseline conditional GAN showed weaker metric performance, particularly on distributional measures (FID > 94), suggesting their latent diversity mechanisms produce outputs that deviate further from training distribution statistics in Inception-v3 feature space. Full metric distributions across all validation samples are shown in Appendix A.3.

Table 2. Automated performance metrics on validation set (57 samples, mean ± std). Bold indicates best performance per metric.

Pixel-Based and Structural Metrics		Statistical Metrics	Learned Distributional Metrics
------------------------------------	--	---------------------	--------------------------------

Architecture	MSE (↓)	PSNR (↑)	MS-SSIM (↑)	SCD (↑)	SD (↑)	CC (↑)	Entropy (↑)	FMI-pixel (↑)	IS (↑)	FID (↓)
cGAN	0.009 ± 0.004	20.931 ± 1.898	0.636 ± 0.065	0.980 ± 0.010	0.123 ± 0.015	0.780 ± 0.047	6.433 ± 0.226	0.643 ± 0.093	1.790 ± 0.154	94.623
Pix2pix	0.007 ± 0.003	21.710 ± 1.883	0.680 ± 0.076	0.981 ± 0.010	0.125 ± 0.011	0.829 ± 0.056	6.439 ± 0.222	0.794 ± 0.154	1.940 ± 0.275	85.286
BicycleGAN	0.007 ± 0.004	22.165 ± 2.207	0.693 ± 0.080	0.982 ± 0.009	0.119 ± 0.011	0.839 ± 0.057	6.385 ± 0.214	0.826 ± 0.171	1.766 ± 0.197	100.071
GauGAN	0.006 ± 0.003	22.626 ± 2.514	0.713 ± 0.098	0.983 ± 0.010	0.120 ± 0.014	0.847 ± 0.065	6.381 ± 0.231	0.886 ± 0.225	1.903 ± 0.264	87.308

3.3. Human-Centered Evaluation: Perceptual Quality Assessment

The structured human evaluation protocols reveal a striking divergence from automated metric rankings. Figure 5(a) presents Visual Turing Pass Rates (VTPR): GauGAN achieved the lowest (best) pass rate (0.533, 95% CI: 0.400–0.667), indicating expert evaluators could not reliably distinguish GauGAN outputs from real marble at better-than-chance levels. In contrast, Pix2Pix, the highest-performing architecture in metric-based evaluation, achieved the highest (worst) VTPR (0.650, 95% CI: 0.583–0.717), meaning experts correctly identified Pix2Pix outputs as synthetic in 65% of trials despite its 2.3% FID advantage over GauGAN. BicycleGAN (0.633, 95% CI: 0.567–0.700) and baseline cGAN (0.583, 95% CI: 0.517–0.650) achieved intermediate performance. Mean Opinion Scores (Figure 5(b)) corroborate this ranking: GauGAN received the highest naturalness ratings (MOS-MA = 2.889, 95% CI: 2.578–3.200), while Pix2Pix scored lowest among GAN architectures (2.333, 95% CI: 2.022–2.644). The baseline cGAN achieved a MOS-MA of 2.644 (95% CI: 2.333–2.955), and BicycleGAN scored 2.667 (95% CI: 2.356–2.978).

These human-centered evaluations showed a somewhat inverted ranking relative to automated metrics: the architecture optimized for FID (Pix2Pix) performs worst in perceptual authenticity, while GauGAN, which achieves the second-best FID, produces textures that expert evaluators perceive as indistinguishable from real marble. This metric-perception divergence challenges the foundational assumption that minimizing Inception-based distributional distance yields perceptually superior outputs, particularly for stochastic texture synthesis tasks where fine-scale irregularity defines naturalness.

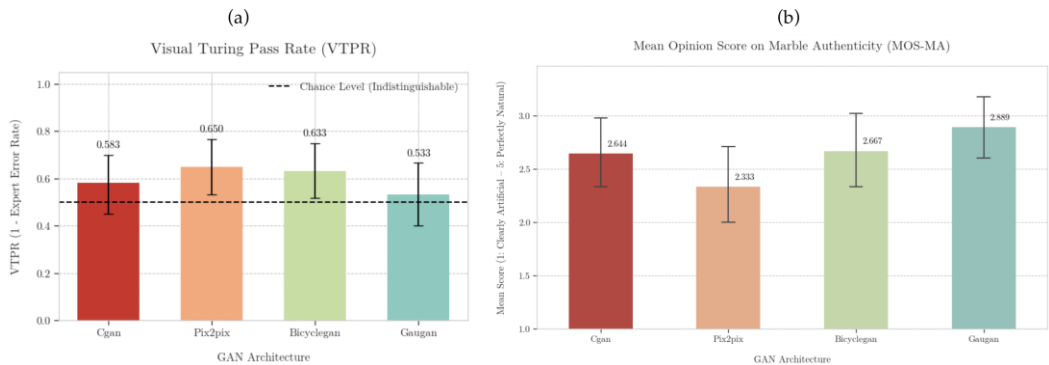


Figure 5. Human-centered evaluation reveals ranking inversion relative to automated metrics. (a) Visual Turing Pass Rate (VTPR): fraction of trials where expert evaluators correctly identified synthetic images as real (lower values indicate more realistic outputs that fool experts). Dashed line at 0.5 indicates chance level performance (perfect indistinguishability). GauGAN’s 95% confidence interval (error bars) spans 0.5, demonstrating expert-level photorealism. Pix2Pix, despite achieving the best FID score, is most easily detected by human evaluators. (b) Mean Opinion Score on Marble Authenticity (MOS-MA) on 5-point Likert scale (1 = clearly artificial, 5 = perfectly natural). GauGAN achieves highest authenticity ratings; Pix2Pix rated significantly worse despite metric superiority. Error bars represent 95% confidence intervals for all architectures.

3.4. The Metric-Perception Divergence

The contradiction between automated and human assessments is evident in the comparative results: architectures with better (lower) FID scores do not consistently achieve better human perceptual scores. This misalignment contradicts the foundational assumption of metric-based GAN development, that minimizing FID produces perceptually superior outputs. Pix2Pix represents the most extreme example of this divergence, achieving the best FID (85.286) yet worst human scores (VTPR = 0.650, MOS-MA = 2.333). Conversely, GauGAN achieved the best human evaluation scores (VTPR = 0.533, MOS-MA = 2.889) despite having a higher (worse) FID of 87.308 compared to Pix2Pix. This inverted pattern, where the architecture that best optimizes the standard training metric performs worst in human assessment, reveals a fundamental limitation in using Inception-v3-based metrics as the sole validation criterion for texture synthesis tasks.

Visual artifact analysis (Figure 6) reveals the mechanism underlying this divergence through comparative magnification of real marble, GauGAN output, and Pix2Pix output. While all three appear photorealistic at standard viewing distances, magnified inspection reveals critical qualitative differences. Real marble exhibits stochastic, aperiodic texture where fine-scale features, such as grain patterns, vein edge irregularities, and matrix crystallization details, show continuous local variation with no repeating motifs. GauGAN successfully replicates this natural stochasticity: magnified regions reveal organic texture variation indistinguishable from real marble, where adjacent areas maintain natural uniqueness without systematic pattern repetition.

In contrast, Pix2Pix outputs exhibit a subtle but systematic failure mode at magnification: identical or near-identical texture motifs recur multiple times within local neighborhoods (indicated by arrows in Figure 6(c)). These repetitive microstructures, such as specific vein branching geometries, matrix grain arrangements, or edge detail patterns appearing 3-5 times in the same orientation, violate the aperiodic character of natural mineral crystallization. While real marble and GauGAN outputs show rich local variation where no two adjacent regions share identical fine-scale structure, Pix2Pix occasionally replicates learned texture patterns during generation, producing spatially repetitive structures. This artifact manifests with sufficient frequency that expert evaluators, trained to assess natural stone quality, immediately perceive it as a synthetic regularity inconsistent with geological formation processes.

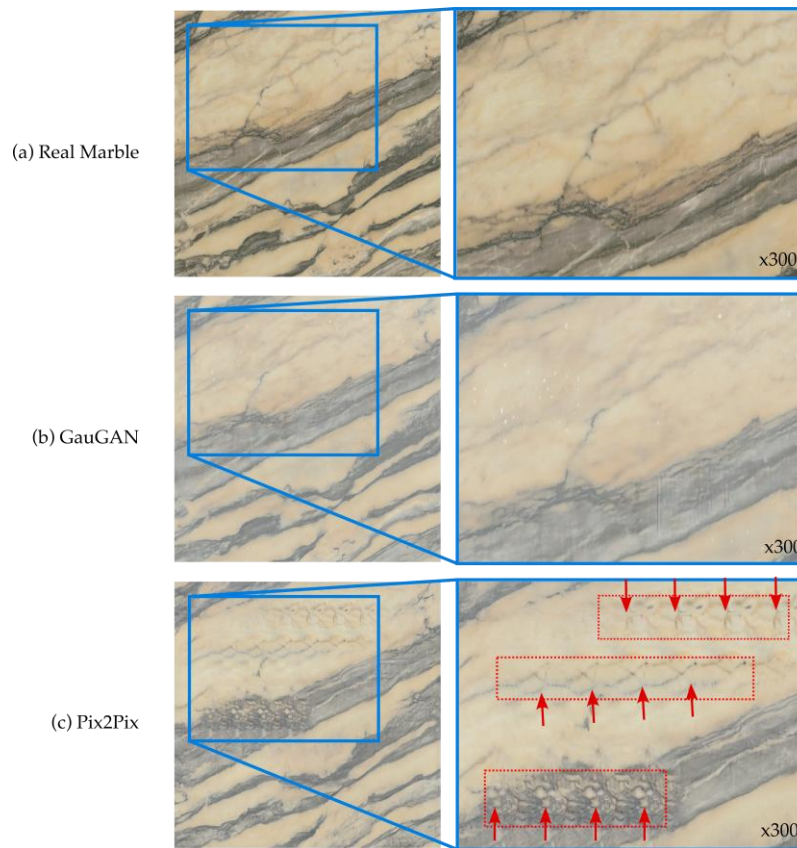


Figure 6. This finding validates recent literature documenting Inception-v3's inadequacy for texture quality assessment in stochastic material synthesis, but extends it to an industrial context with economic stakes: optimizing for FID in a product design application would select Pix2Pix, delivering textures that professional users immediately recognize as artificial due to perceptible local regularity, a costly error with direct consequences for architectural visualization, virtual prototyping, and design workflows where material authenticity determines client acceptance.

3.5. Computational Efficiency and Deployment Feasibility

Figure 7 presents the training evolution across all four architectures, revealing distinct convergence patterns that explain their final performance characteristics. The FID evolution (Figure 7(c)) demonstrates that GauGAN achieved the fastest and smoothest convergence, reaching its final FID of 87.308 within 1,000 epochs and maintaining stability thereafter. Pix2Pix exhibited slower but consistent FID improvement, achieving its best score of 85.286 at epoch 3,000. The baseline cGAN and BicycleGAN showed more volatile training dynamics with higher final FID values (94.623 and 100.071, respectively), suggesting that explicit latent diversity mechanisms complicate distributional alignment with Inception-v3 feature statistics.

Generator loss trajectories (Figure 7(a)) reveal the architectural differences in learning dynamics. GauGAN exhibited the most dramatic initial loss decrease, dropping from ~45 to ~21 within the first 1,000 epochs before stabilizing, a pattern indicating rapid feature learning enabled by SPADE's multi-scale semantic injection. BicycleGAN showed similar convergence behavior but with lower initial loss (~22), while Pix2Pix maintained relatively stable generator loss throughout training (~13-14), consistent with its deterministic mapping and L1 regularization. The baseline cGAN exhibited the lowest generator loss (~2-4), but this did not translate to superior FID performance, highlighting the disconnect between generator loss magnitude and perceptual quality.

Discriminator loss evolution (Figure 7(b)) provides insight into adversarial training stability. GauGAN's discriminator initially struggled (loss ~0.4, indicating overly confident predictions) before stabilizing around 0.6-0.7, suggesting the generator learned to produce challenging outputs that

maintained discriminator uncertainty. Pix2Pix's discriminator loss decreased steadily from ~ 1.2 to ~ 0.2 , indicating the discriminator became increasingly confident in detecting synthetic images, potentially explaining why human evaluators also found Pix2Pix outputs more detectable despite superior FID. The baseline cGAN and BicycleGAN maintained more balanced discriminator losses (~ 0.5 - 0.6), consistent with the Nash equilibrium in adversarial training.

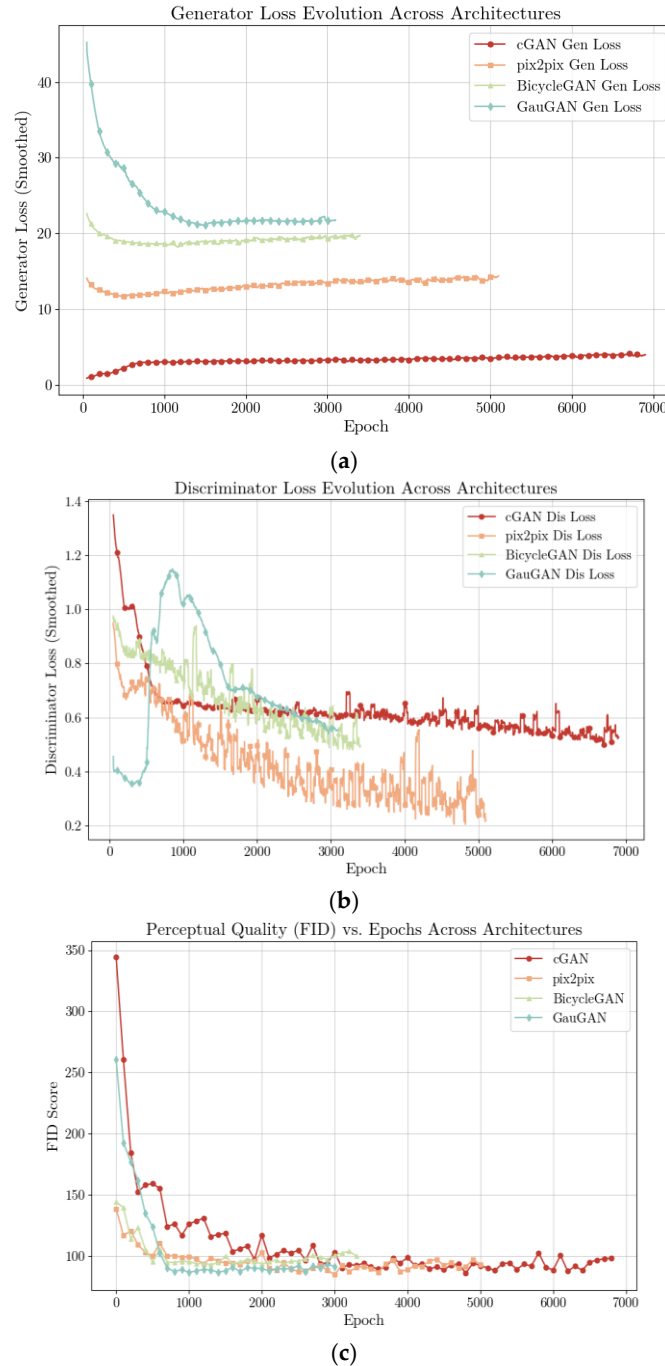


Figure 7. Training dynamics across the four architectures. (a) Generator loss convergence. (b) Discriminator loss evolution. (c) FID evolution over epochs.

Table 3 reports computational characteristics relevant for industrial deployment. GauGAN requires the fewest trainable parameters (53.1M) despite architectural complexity, as SPADE normalization layers are parameter-efficient relative to U-Net encoder stacks. However, GauGAN demands the highest training cost (0.82 min/epoch, 1761.3 GFLOPS) due to the computational expense of per-pixel normalization parameter prediction. This translates to a total training time of 42.4 hours to convergence (3,100 epochs) compared to 10.2 hours for Pix2Pix (5,100 epochs $\times$ 0.12

min/epoch) and 12.7 hours for the baseline cGAN (6,900 epochs $\times$ 0.11 min/epoch). Pix2Pix offers competitive training efficiency (0.12 min/epoch) with parameter count (61.4M). At inference time, all models achieve performance on consumer GPUs for 1280 $\times$ 720 outputs consistent with their usage in interactive design tools. Given GauGAN’s superior perceptual quality validated through human evaluation, the 4.2 $\times$ total training time penalty relative to Pix2Pix (42.4 vs. 10.2 hours) is justified for quality-critical applications like architectural visualization, where human perception is the ultimate arbiter.

Table 3. Model complexity and training efficiency metrics.

Architecture	# of Parameters (Millions)	Computational cost [GFLOPS]	Avg. Time/Epoch (min)	Total Epochs
cGAN	63.9	170.2	0.11	6900
pix2pix	61.4	170.2	0.12	5100
BicycleGAN	85.9	105.4	0.17	3400
GauGAN	53.1	1761.3	0.82	3100

4. Discussion

This section synthesizes the experimental findings to interpret their broader implications for industrial texture synthesis, analyzes the training dynamics and architectural insights, and discusses the study’s limitations while proposing concrete directions for future research.

4.1. Synthesis of Findings and Implications

The comprehensive evaluation confirms that conditional GANs are highly effective at synthesizing photorealistic, structurally controlled marble textures. However, the results reveal a significant misalignment between automated quantitative metrics and expert human judgment, a finding with profound implications for how the research community validates generative models. The quantitative-perceptual divergence between Pix2Pix (best FID: 85.286, worst MOS-MA: 2.333) and GauGAN (FID: 87.308, best MOS-MA: 2.889) demonstrates that Inception-based metrics fail to capture human-relevant texture quality. This finding aligns with large-scale empirical evidence that models producing more perceptually realistic images paradoxically score worse on FID, suggesting that replacing Inception-v3 with alternative encoders could improve human-metric alignment [13]. Our results extend these findings to another industrial context, confirming Borji’s (2022) [14] observation that FID is particularly unsuitable for specialized domains where visual features differ from those in natural images. The mechanism underlying this divergence is architectural: Pix2Pix’s transposed convolutional layers are prone to introducing subtle patterns that, although they marginally affect structural similarity metrics, are immediately perceptible to human observers as unnatural regularity. In contrast, GauGAN’s SPADE-based generator avoids transposed convolutions entirely, producing smooth organic textures that experts could not distinguish from real marble at statistically significant levels.

This finding has significant implications for industrial applications. Optimizing for FID alone would select Pix2Pix, delivering textures that professional users immediately recognize as artificial, a costly error in architectural visualization, virtual prototyping, and digital arts, where material authenticity determines client acceptance. For applications in which the end user is human, structured, human-centered evaluations should be considered an essential component of the validation pipeline. While automated metrics remain invaluable for guiding training dynamics, they are insufficient as the sole arbiters of perceptual quality.

Furthermore, the success of GauGAN in generating controllable, high-fidelity textures signals a potential paradigm shift in industrial material design. Our work demonstrates that conditional GANs successfully unify the control of traditional procedural methods with the realism of data-driven techniques, enabling explicit structural control through binary masks while synthesizing

photorealistic local appearance. This capability could transition design processes from selecting materials from predefined catalogs to actively creating bespoke, digitally native materials on demand, thereby supporting virtual prototyping workflows and digital twin applications.

4.2. Training Dynamics and Architectural Insights

The training dynamics reveal fundamental architectural trade-offs between convergence speed, computational cost, and perceptual quality. GauGAN's rapid FID convergence demonstrates SPADE's efficiency in learning texture distributions through multi-scale semantic injection. However, this advantage comes with a substantial per-epoch computational cost: GauGAN requires 1,761.3 GFLOPS per forward pass (10.3× higher than Pix2Pix) due to per-pixel convolutions in each SPADE layer. The discriminator loss patterns provide additional insight into the metric-perception divergence. Pix2Pix's steadily decreasing discriminator loss indicates that the discriminator has learned to reliably detect synthetic outputs, aligning with human evaluation results. In contrast, GauGAN maintained discriminator uncertainty throughout training, suggesting that its outputs remained difficult to classify even for networks explicitly trained to detect them. This adversarial balance correlates with human indistinguishability, validating the original GAN objective. For industrial deployment, these findings suggest clear decision criteria: applications requiring rapid iteration should prioritize Pix2Pix despite detectable artifacts, whereas quality-critical applications justify GauGAN's 4.2× training-time investment to achieve human-level perceptual authenticity.

4.3. Practical Validation of Unsupervised Mask Generation

The successful training of all architectures to photorealistic quality levels validates the unsupervised segmentation pipeline as a practical solution to the annotation bottleneck. Masks generated from SLIC superpixels, GMM clustering, and Graph Cut optimization were sufficient to condition high-fidelity synthesis across all 289 marble slabs without manual correction, which is critical for industrial deployment, where annotation costs can be high for datasets of this scale.

The robustness to mask imperfections is noteworthy: the L1 reconstruction loss during GAN training implicitly corrects minor mask inaccuracies by learning to fill vein regions with textures that match the training data statistics. This aligns with recent work demonstrating that two-stage generative pipelines can be effective even with imperfect label maps [38]. However, systematic segmentation failures would propagate through the synthesis pipeline, suggesting that future work exploring foundation models like Segment Anything could further improve robustness.

4.4. Limitations and Directions for Future Research

While this study establishes a foundational benchmark for mask-conditioned synthesis of stochastic natural materials, several limitations warrant acknowledgment.

First, our methodology was validated on a single marble type. While this material exhibits rich vein patterns providing a challenging test case, geological materials display enormous visual diversity. Future work should extend this framework to taxonomically diverse natural stones to assess whether the findings generalize across material classes with different mechanisms of texture generation.

Second, the statistical power of human evaluation is constrained by the sample size. While sufficient for exploratory validation and aligned with practices in perceptual quality assessment literature, future work should employ larger panels to establish robust effect sizes and enable subgroup analyses across evaluator expertise levels. Consumer preferences may differ systematically from expert judgments, and expanding evaluation through crowdsourced platforms following ITU-standardized protocols would strengthen generalizability.

Third, we deliberately excluded diffusion models from this benchmark for methodological rigor: comparing models with fundamentally different training paradigms (adversarial vs. denoising), data requirements, and computational profiles would introduce confounding variables that obscure

architecture-specific insights. State-of-the-art diffusion models like Stable Diffusion and ControlNet leverage massive pre-trained foundation models, making direct comparison methodologically complex. A dedicated diffusion model comparison using the same dataset and evaluation protocols is underway as a follow-up study, with careful attention to disentangling architectural effects from pre-training data scale.

Fourth, extending this 2D framework to 3D volumetric texture generation would enable more immersive visualization in architectural applications where marble veins penetrate through material depth. Recent work on 3D-aware generative models and neural radiance fields provides promising foundations for this extension.

Finally, the observed metric-perception misalignment suggests a critical need for developing new evaluation metrics that better align with human judgment for texture synthesis tasks. Research directions include learning-based perceptual metrics trained on human judgments, multi-scale texture descriptors that capture stochastic properties, and hybrid frameworks that combine efficient automated screening with targeted human validation. The dataset and protocols established here could serve as benchmarks for developing such metrics.

5. Conclusions

This study demonstrated that conditional GANs can synthesize photorealistic, structurally controlled marble textures from automatically generated masks, eliminating manual annotation costs. Through systematic evaluation of four architectures on 289 industrial scans, we revealed a critical metric-perception divergence: Pix2Pix achieved the best FID (85.286) but worst human ratings, while GauGAN produced textures statistically indistinguishable from real marble (VTPR: 0.533, MOS-MA: 2.889) despite inferior FID. This finding establishes that human-in-the-loop evaluation is essential for deployment decisions in quality-critical applications. For industrial texture synthesis, we recommend GauGAN for applications requiring perceptual authenticity and Pix2Pix only when computational efficiency outweighs quality concerns. Future work will extend this framework to diverse geological materials and compare with conditional diffusion models.

Author Contributions: Conceptualization, A.A.C., M.F. and G.P.; methodology, A.A.C. and M.F.; software, A.A.C. and C.D.; validation, A.A.C.; formal analysis, A.A.C. and M.F.; investigation, A.A.C. and M.F.; resources, G.P. and P.A.; data curation, A.A.C. and C.D.; writing—original draft preparation, A.A.C.; writing—review and editing, A.A.C., M.F., C.D., G.P. and P.A.; visualization, A.A.C. and C.D.; supervision, P.A.; project administration, P.A.; funding acquisition, G.P. and P.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work received support from National Funds from FCT – Fundação para a Ciência e a Tecnologia, I.P., through the project UIDB/04028/2025 and FCT-UIDP/04028/2025 of CERENA – Centro de Recursos Naturais e Ambiente. The authors acknowledge to the Sustainable Stone by Portugal project, proposal number C644943391 - 00000051 co-financed by the PRR – Recovery and Resilience Plan of the European Union (Next Generation EU).

Informed Consent Statement: All participants provided informed consent prior to evaluation. This study was conducted in accordance with institutional ethical guidelines for non-invasive behavioral research, and formal ethical review was waived as the study involved only expert-quality assessment tasks without the collection of personal or sensitive data.

Data Availability Statement: While the raw industrial scans remain proprietary, the extracted binary masks and trained model weights will be made available upon publication to support reproducibility.

Acknowledgments: This work received support from National Funds from FCT – Fundação para a Ciência e a Tecnologia, I.P., through the project UIDB/04028/2025 and FCT-UIDP/04028/2025 of CERENA – Centro de Recursos Naturais e Ambiente. The authors acknowledge to the Sustainable Stone by Portugal project, proposal

number C644943391 - 00000051 co-financed by the PRR – Recovery and Resilience Plan of the European Union (Next Generation EU).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

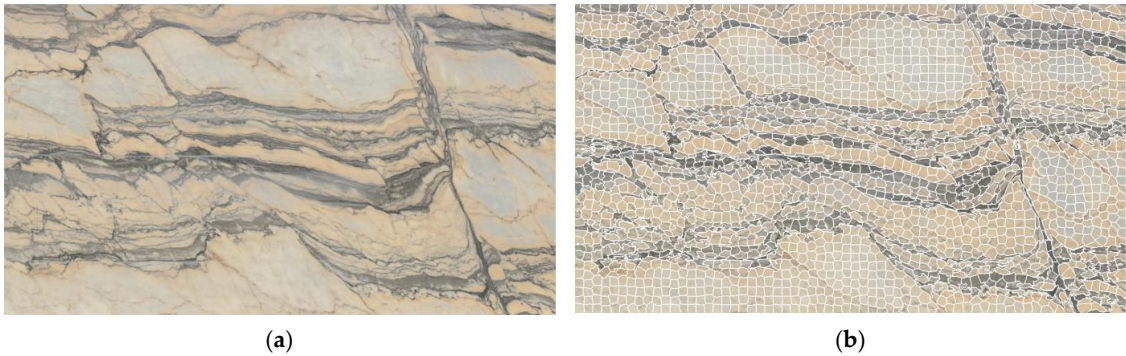
The following abbreviations are used in this manuscript:

2AFC	Two-Alternative Forced Choice
CC	Correlation Coefficient
cGAN	Conditional Generative Adversarial Network
FID	Fréchet Inception Distance
FMI	Feature Mutual Information
GAN	Generative Adversarial Network
GFLOPS	Giga Floating Point Operations Per Second
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
IS	Inception Score
MOS-MA	Mean Opinion Score on Marble Authenticity
MS-SSIM	Multi-Scale Structural Similarity Index
MSE	Mean Squared Error
PSNR	Peak Signal-to-Noise Ratio
SCD	Structural Content Dissimilarity
SLIC	Simple Linear Iterative Clustering
SPADE	Spatially-Adaptive Denormalization
VTPR	Visual Turing Pass Rate

Appendix A

Appendix A.1. Unsupervised Segmentation Pipeline Details

Figure A1 provides a detailed visualization of the unsupervised segmentation pipeline, showing all intermediate processing stages from raw input to final binary mask.



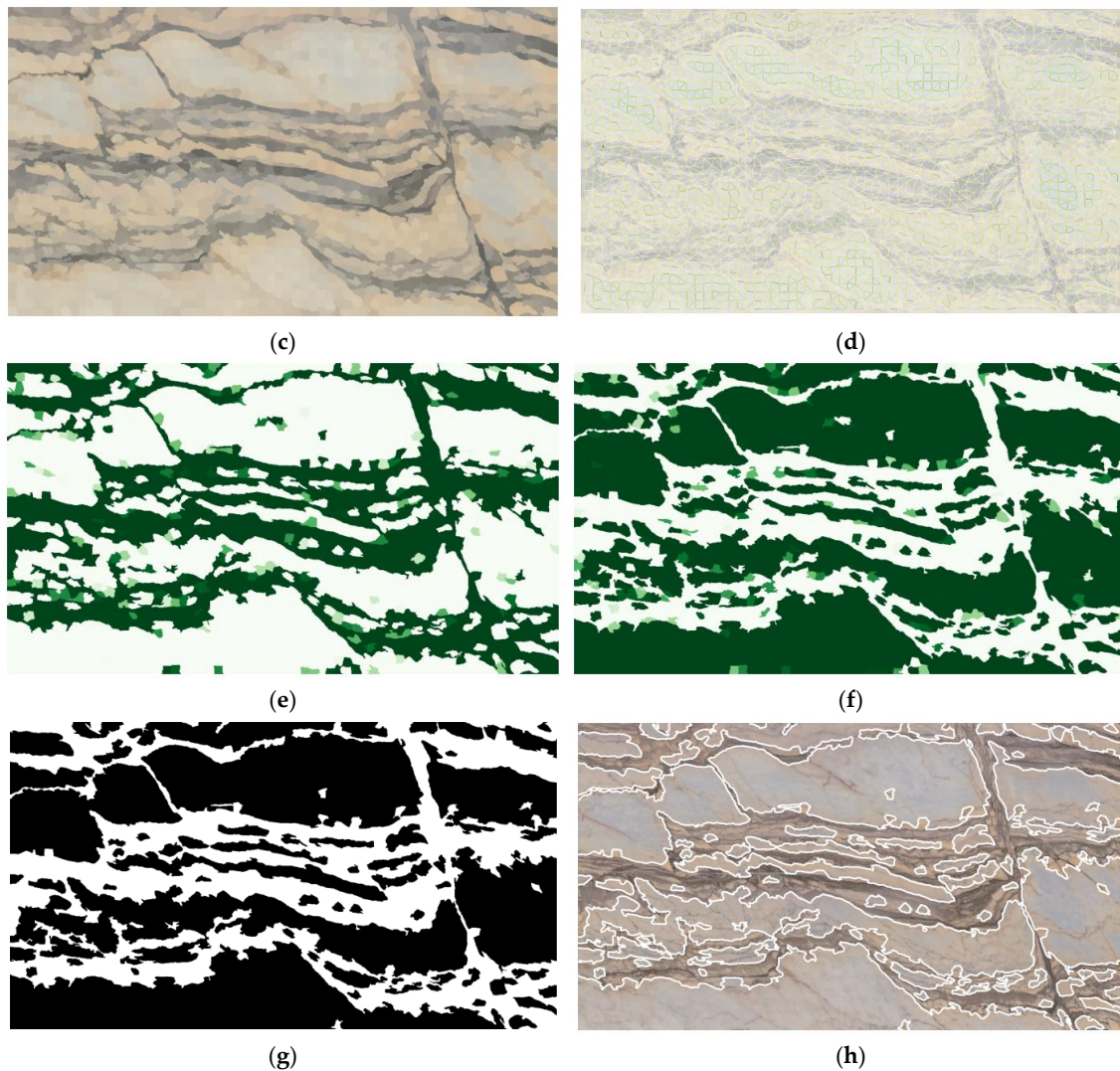
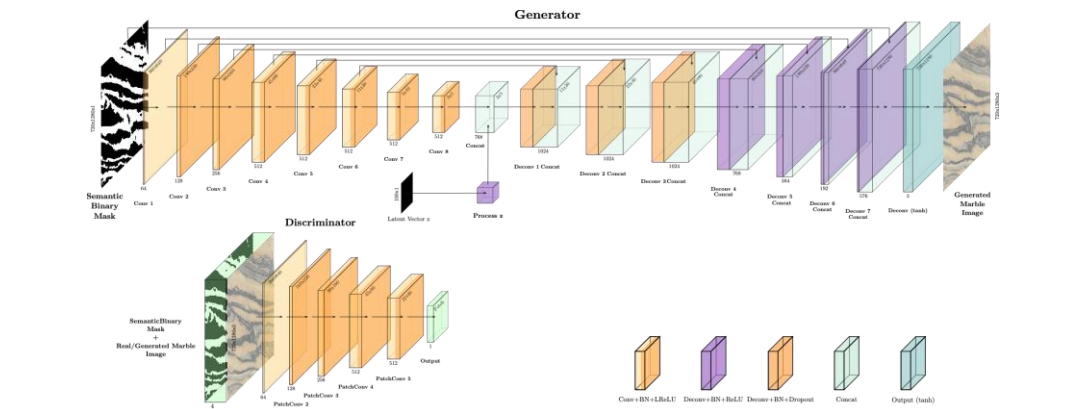


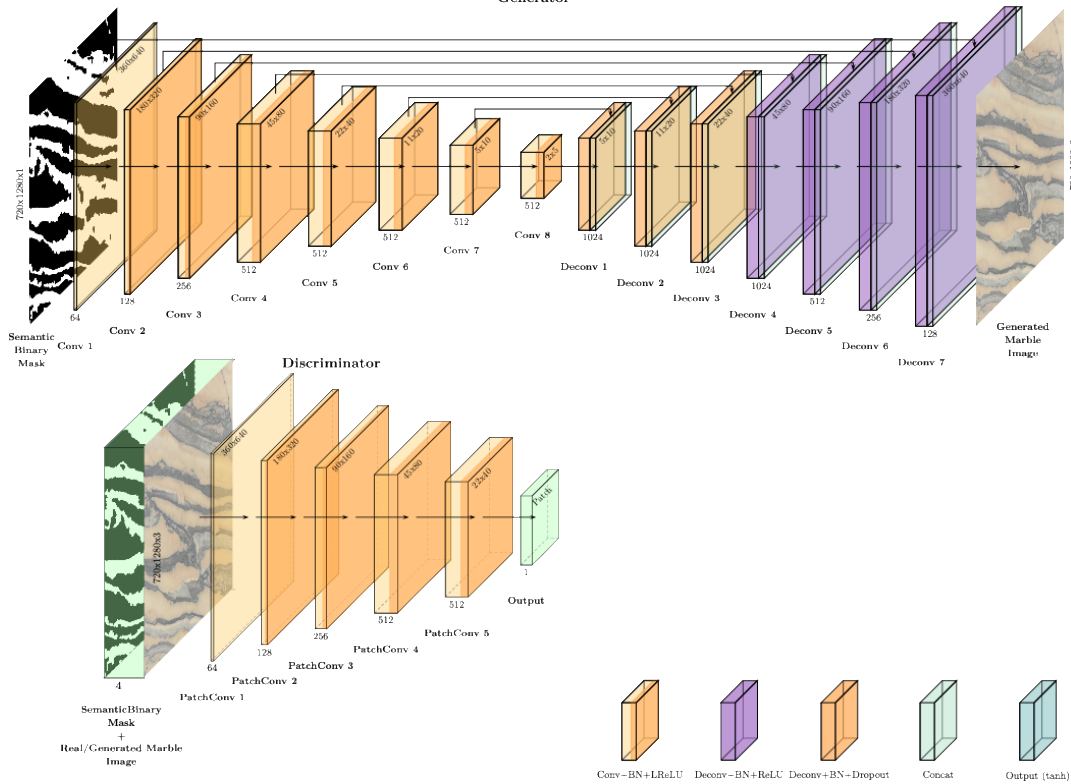
Figure A1. Detailed visualization of the unsupervised segmentation pipeline for a representative Exotic Ambar marble slab. (a) Original RGB input image at 1280×720 resolution; (b) SLIC superpixel tessellation showing approximately 3,000 perceptually uniform regions with nominal size 20 px and compactness 0.3; (c) Graph structure visualization for spatial regularization, where nodes represent superpixels and edges encode adjacency relationships; (d) Heatmap of GMM-based unary costs for the “vein” class, with brighter regions indicating higher probability of vein membership; (e) Heatmap of GMM-based unary costs for the “matrix” class; (f) Final binary mask after Graph Cut optimization with regularization weight $\lambda = 5.0$, overlaid on the original scan to demonstrate boundary preservation. The pipeline automatically extracts vein structures without manual annotation, achieving 100% acceptance rate across all 289 samples upon visual inspection.

Appendix A.2. Conditional GAN Architecture Diagrams

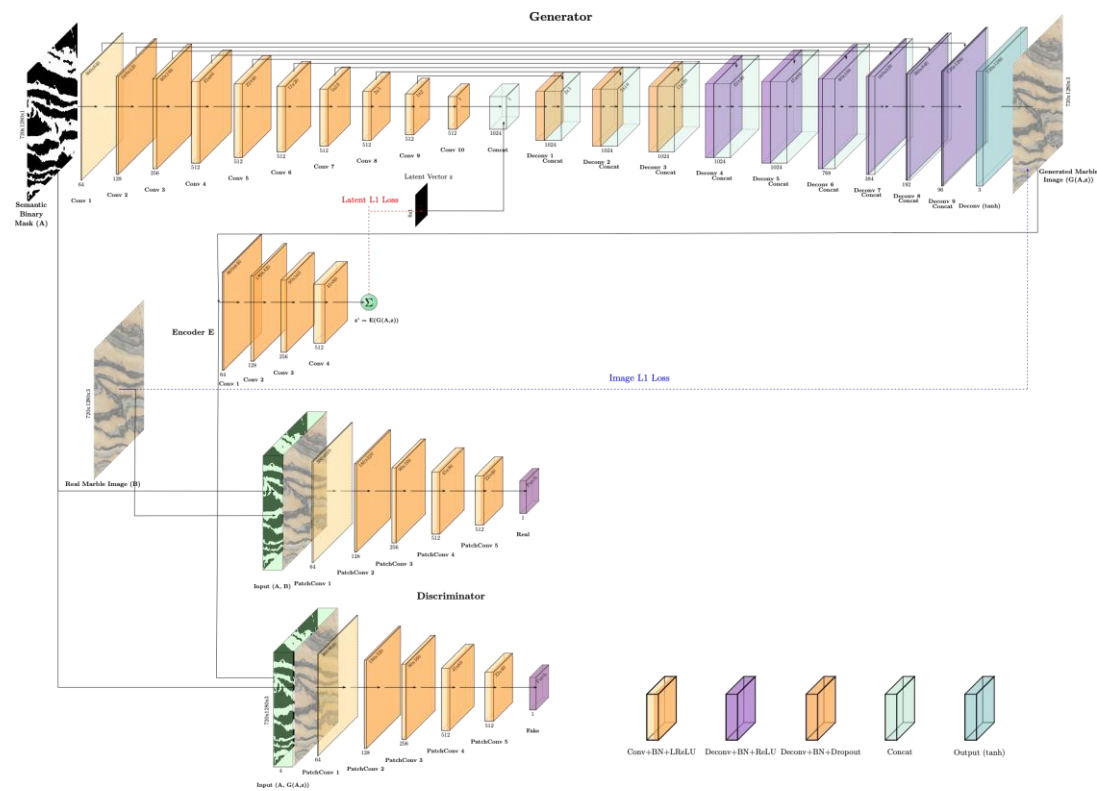
Figure A2 presents the detailed architecture diagrams for all four conditional GAN models evaluated in this study, illustrating the structural differences in their generator designs.



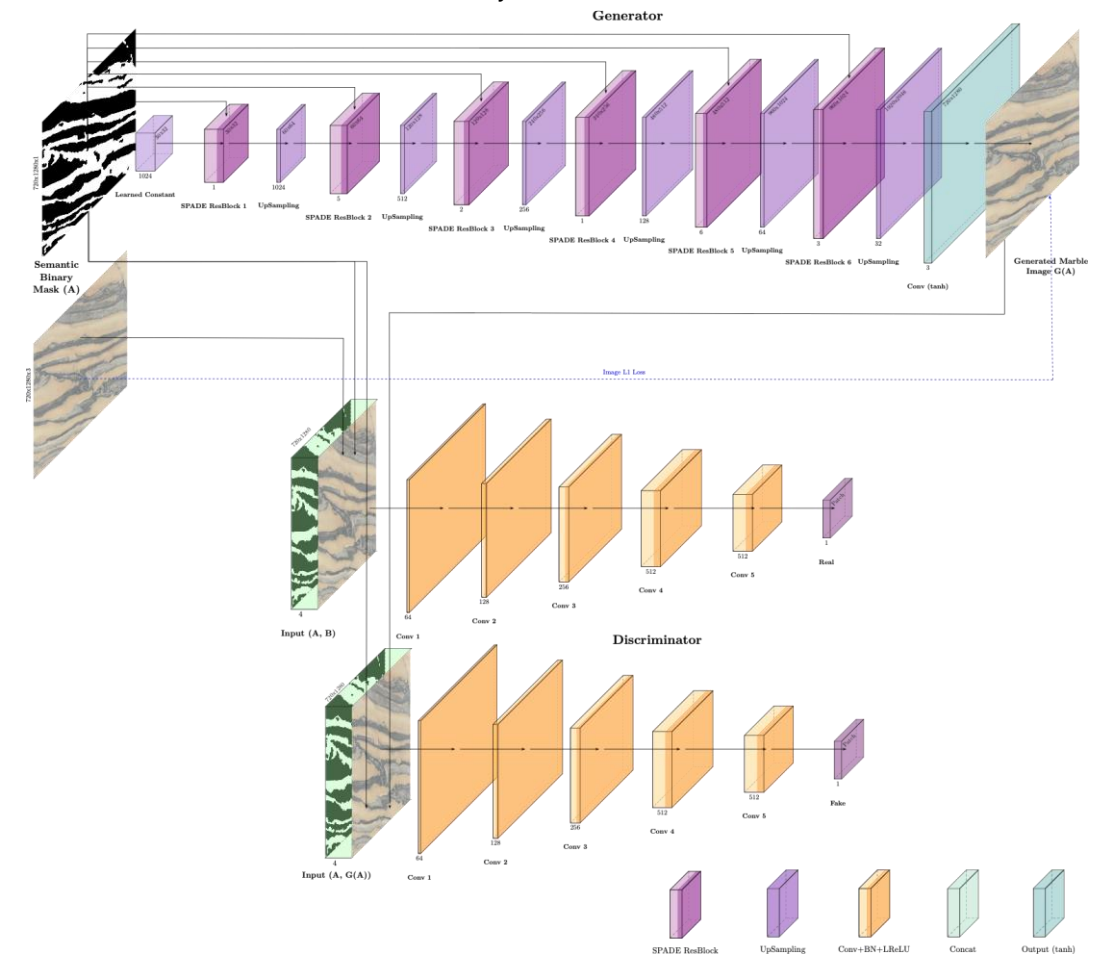
(a) cGAN architecture



(b) Pix2Pix architecture



(c) BicycleGAN architecture



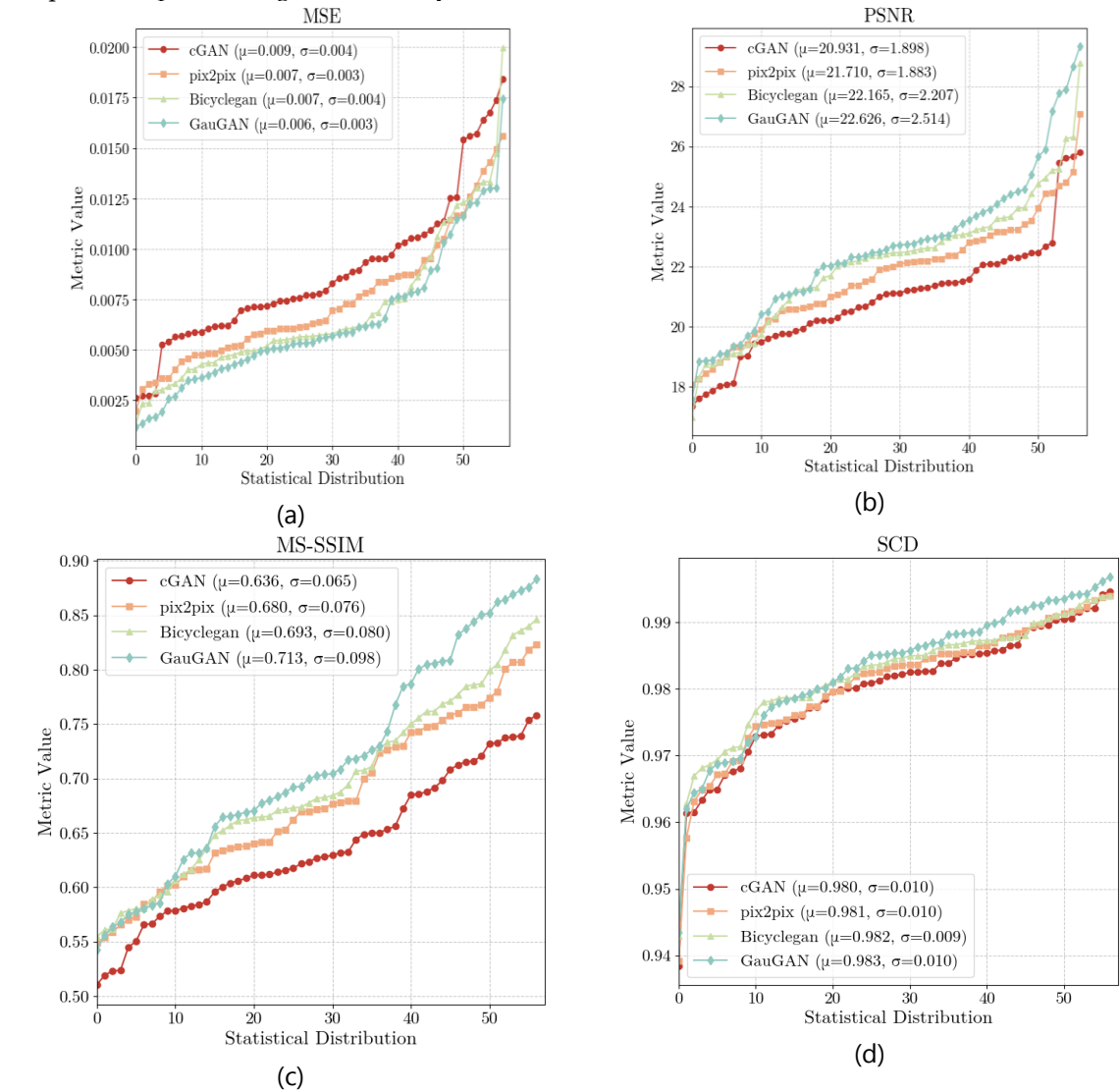
(d) GauGAN architecture

Figure A2. Architecture diagrams for the four conditional GAN models evaluated in this study. (a) Baseline conditional GAN (cGAN): U-Net generator with 8 encoder and 7 decoder blocks, conditioned on both the binary

mask and a 100-dimensional latent vector injected at the bottleneck; (b) Pix2Pix: Deterministic U-Net generator conditioned solely on the input mask, with L1 reconstruction loss ($\lambda = 100$) enforcing pixel-level fidelity; (c) BicycleGAN: Extended architecture including a dedicated encoder network for bijective latent-to-image mapping with cycle consistency ($\lambda_{\text{latent}} = 10$); (d) GauGAN: SPADE-based generator starting from a learned constant tensor, with six residual blocks incorporating Spatially-Adaptive Normalization layers that modulate activations based on the input mask at multiple scales. All architectures employ identical PatchGAN discriminators with 70×70 receptive fields to ensure fair comparison. Conv = Convolutional layer; BN = Batch Normalization; LReLU = Leaky ReLU; Deconv = Transposed Convolution.

Appendix A.3. Quantitative Metrics Distributions

Figure A3 shows the distribution of quantitative performance metrics across all validation samples, complementing the summary statistics in Table 2.



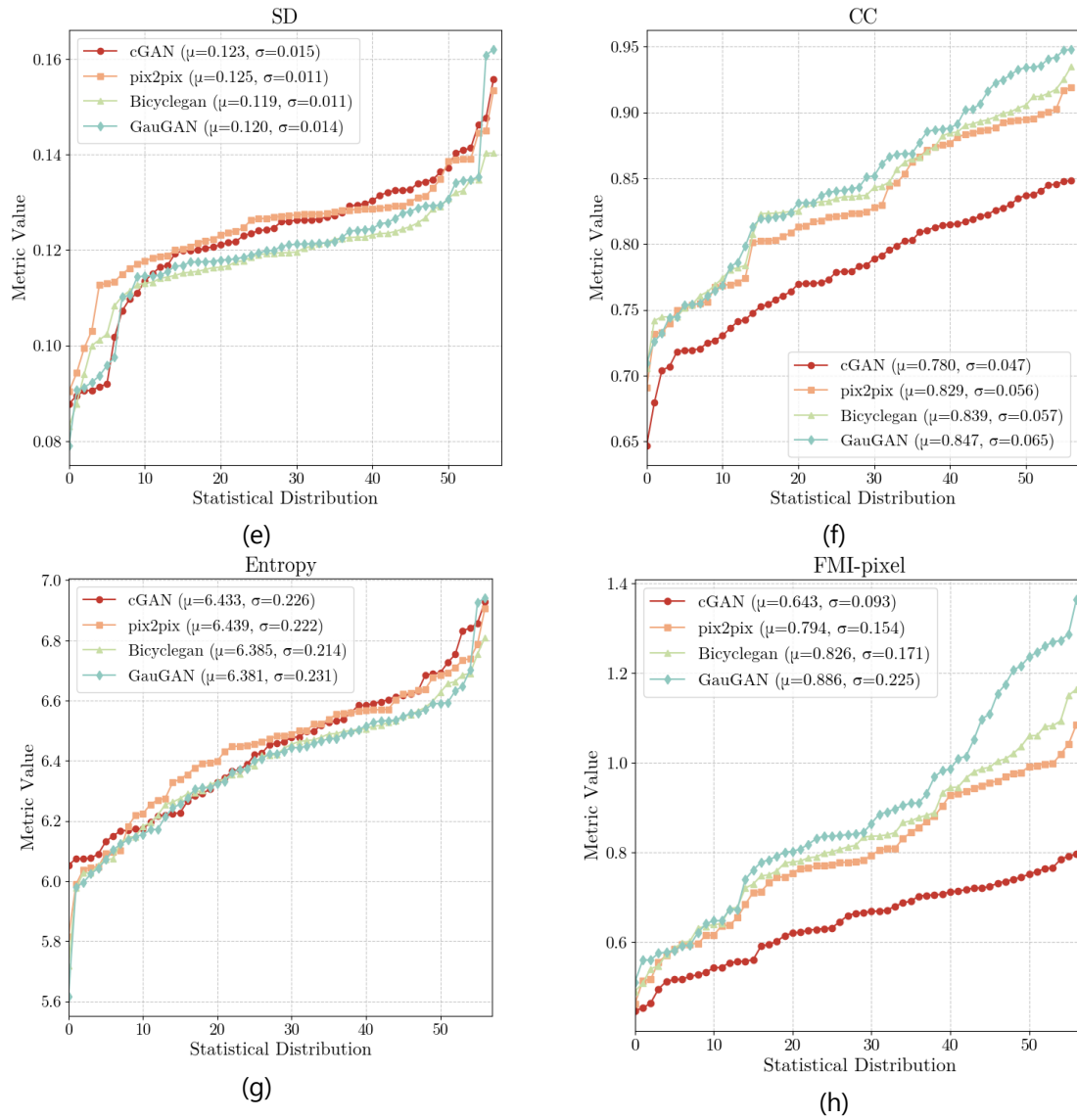


Figure A3. Distribution of quantitative performance metrics across the validation set (n = 57 samples) for all four GAN architectures. Each subplot shows sorted metric values, with the x-axis representing samples ordered by metric value and the y-axis showing the metric magnitude. Metrics shown: (a) Mean Squared Error (MSE, lower is better); (b) Peak Signal-to-Noise Ratio (PSNR, higher is better); (c) Multi-Scale Structural Similarity Index (MS-SSIM, higher is better); (d) Structural Content Dissimilarity (SCD, higher is better); (e) Standard Deviation (SD, higher indicates more contrast); (f) Correlation Coefficient (CC, higher is better); (g) Entropy (higher indicates more information content); (h) Feature Mutual Information at pixel level (FMI-pixel, higher is better). Legend indicates architecture with mean (μ) and standard deviation (σ) for each metric. GauGAN consistently demonstrates superior reconstruction fidelity (MSE, PSNR, MS-SSIM, CC), while Pix2Pix shows advantages in texture contrast (SD) and information richness (Entropy).

Appendix A.4. Excluded Samples Documentation

Figure A4 documents examples of marble slab images excluded during data curation due to imaging artifacts.

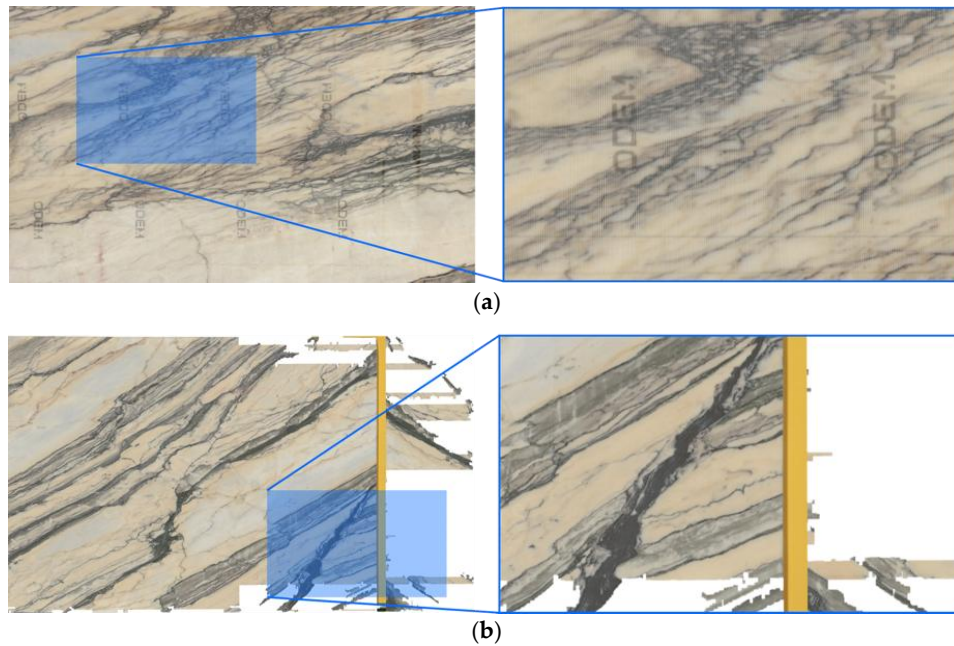


Figure A4. Examples of discarded marble slab images excluded during data curation. From an initial set of 327 scans, 38 samples (11.6%) were removed due to imaging artifacts that would compromise model training. (a) Slab with fiber-reinforced protective film: the magnified inset reveals a regular grid pattern from the reinforcement mesh that overlays the natural marble texture, introducing artificial high-frequency structure incompatible with learning genuine material appearance. (b) Slab with scanner-induced artifacts: the magnified inset shows color banding and geometric distortions resulting from line-scan camera malfunction, which would introduce spurious correlations during training. These exclusions ensure the curated dataset of 289 samples reflects genuine marble appearance variation rather than imaging defects, supporting robust generalization of the trained models.

References

1. Jimeno-Morenilla, A.; Azariadis, P.; Molina-Carmona, R.; Kyratzi, S.; Moulianitis, V. Technology Enablers for the Implementation of Industry 4.0 to Traditional Manufacturing Sectors: A Review. *Comput. Ind.* **2021**, *125*, doi:10.1016/j.compind.2020.103390.
2. Loy, J.; Canning, S.; Little, C. Industrial Design Digital Technology. *Procedia Technology* **2015**, *20*, 32–38, doi:10.1016/j.protcy.2015.07.007.
3. Xian, W.; Sangkloy, P.; Agrawal, V.; Raj, A.; Lu, J.; Fang, C.; Yu, F.; Hays, J. TextureGAN: Controlling Deep Image Synthesis with Texture Patches. **2018**.
4. Weinberger, P.; Gall, A.; Heim, A.; Yosifov, M.; Kastner, J.; Schwarz, L.; Fröhler, B.; Bodenhofer, U.; Sascha, S. Unsupervised Segmentation of Industrial X-Ray Computed Tomography Data with the Segment Anything Model. **2024**, doi:10.21203/rs.3.rs-3900291/v1.
5. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. **2014**.
6. Park, T.; Liu, M.-Y.; Wang, T.-C.; Zhu, J.-Y. Semantic Image Synthesis with Spatially-Adaptive Normalization. **2019**.
7. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017; Institute of Electrical and Electronics Engineers Inc., November 6 2017; Vol. 2017-January, pp. 5967–5976.
8. Era, I.Z.; Ahmed, I.; Liu, Z.; Das, S. An Unsupervised Approach towards Promptable Defect Segmentation in Laser-Based Additive Manufacturing by Segment Anything. *arXiv:2312.04063v3 [cs.CV]* **2024**.
9. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Proceedings of the Proceedings of the 31st

- International Conference on Neural Information Processing Systems NIPS'17; December 4 2017; pp. 6629–6640.
10. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved Techniques for Training GANs. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016); D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett, Eds.; 2016.
 11. Gatys, L.A.; Ecker, A.S.; Bethge, M. A Neural Algorithm of Artistic Style. *Computing Research Repository (CoRR)* **2015**.
 12. Zhou, S.; Gordon, M.L.; Krishna, R.; Narcomey, A.; Fei-Fei, L.; Bernstein, M.S. HYPE: A Benchmark for Human EYE Perceptual Evaluation of Generative Models. *Proceedings of the 33rd International Conference on Neural Information Processing Systems* **2019**, 3449–3461.
 13. Stein, G.; Cresswell, J.C.; Hosseinzadeh, R.; Sui, Y.; Leigh Ross, B.; Vilecroze, V.; Liu, Z.; Caterini, A.L.; Eric Taylor, J.T.; Loaiza-Ganem, G. *Exposing Flaws of Generative Model Evaluation Metrics and Their Unfair Treatment of Diffusion Models*; 2023;
 14. Borji, A. Pros and Cons of GAN Evaluation Measures: New Developments. *Computer Vision and Image Understanding* **2022**, 215, doi:10.1016/j.cviu.2021.103329.
 15. Liu, L.; Duan, H.; Hu, Q.; Yang, L.; Cai, C.; Ye, T.; Liu, H.; Zhang, X.; Zhai, G. F-Bench: Rethinking Human Preference Evaluation Metrics for Benchmarking Face Generation, Customization, and Restoration. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* **2025**, 10982–10994.
 16. Perlin, K. Improving Noise. In Proceedings of the Proceedings of the 29th annual conference on Computer graphics and interactive techniques; San Antonio, TX, 2002; pp. 681–682.
 17. Perlin, K. An Image Synthesizer. In Proceedings of the Proceedings of the 12th annual conference on Computer graphics and interactive techniques; San Francisco, CA, 1985; Vol. 19, pp. 287–296.
 18. Turk, G. *Generating Textures on Arbitrary Surfaces Using Reaction-Diffusion*; 1991; Vol. 25;.
 19. Worley, S. A Cellular Texture Basis Function. *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniq* **1996**, 291–294.
 20. Efros, A.A.; Leung, T.K. Texture Synthesis by Non-Parametric Sampling. In Proceedings of the IEEE International Conference on Computer Vision; Corfu, Greece, September 1999.
 21. Efros, A.A.; Freeman, W.T. Image Quilting for Texture Synthesis and Transfer. In Proceedings of the SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques; ACM Digital Library, 2001; pp. 341–346.
 22. Kwatra, V.; Schödl, A.; Essa, I.; Turk, G.; Bobick, A. Graphcut Textures: Image and Video Synthesis Using Graph Cuts. *ACM Transactions on Graphics (TOG)* **2003**, 22, 277–286, doi:10.1145/882262.88226.
 23. Levina, E.; Bickel, P.J. Texture Synthesis and Nonparametric Resampling of Random Fields. *Ann. Stat.* **2006**, 34, 1751–1773, doi:10.1214/009053606000000588.
 24. Aguerreberre, C.; Gousseau, Y.; Tartavel, G. Exemplar-Based Texture Synthesis: The Efros-Leung Algorithm. *Image Processing On Line* **2013**, 3, 223–241, doi:10.5201/ipol.2013.59.
 25. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, 43, 4217–4228, doi:10.1109/TPAMI.2020.2970919.
 26. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* **2022**, 10674–10685, doi:10.1109/CVPR52688.2022.01042.
 27. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv: Machine Learning (stat.ML)* **2014**.
 28. Gatys, L.A.; Ecker, A.S.; Bethge, M. Texture Synthesis Using Convolutional Neural Networks. *Computing Research Repository (CoRR)* **2015**.
 29. Nikolay Jetchev; Urs Bergmann; Roland Vollgraf Texture Synthesis with Spatial Generative Adversarial Networks. *Computing Research Repository (CoRR)* **2016**, doi:10.48550/arXiv.1611.08207.
 30. Bergmann, U.; Jetchev, N.; Vollgraf, R. Learning Texture Manifolds with the Periodic Spatial GAN. *Computing Research Repository (CoRR)* **2017**.
 31. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the Proceedings of the IEEE International Conference on

- Computer Vision; Institute of Electrical and Electronics Engineers Inc., December 22 2017; Vol. 2017-October, pp. 2242–2251.
32. Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward Multimodal Image-to-Image Translation. **2018**.
 33. Tan, Z.; Chen, D.; Chu, Q.; Chai, M.; Liao, J.; He, M.; Yuan, L.; Hua, G.; Yu, N. Efficient Semantic Image Synthesis via Class-Adaptive Normalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 4852–4866, doi:10.1109/TPAMI.2021.3076487.
 34. Zhang, L.; Rao, A.; Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision; Institute of Electrical and Electronics Engineers Inc., 2023; pp. 3813–3824.
 35. Cao, P.; Zhou, F.; Yang, L.; Huang, T.; Song, Q. Image Is All You Need to Empower Large-Scale Diffusion Models for In-Domain Generation. *CVPR2025* **2025**.
 36. Borovec, J. Fully Automatic Segmentation of Stained Histological Cuts. In Proceedings of the Poster 2013 : 17th International Student Conference on Electrical Engineering; Prague, May 16 2013.
 37. Borovec, J.; Svihlík, J.; Kybic, J.; Habart, D. Supervised and Unsupervised Segmentation Using Superpixels, Model Estimation, and Graph Cut. *J. Electron. Imaging* **2017**, *26*.
 38. Andreini, P.; Ciano, G.; Bonechi, S.; Graziani, C.; Lachi, V.; Mecocci, A.; Sodi, A.; Scarselli, F.; Bianchini, M. A Two-Stage GAN for High-Resolution Retinal Image Generation and Segmentation. *Electronics (Switzerland)* **2022**, *11*, doi:10.3390/electronics11010060.
 39. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2281, doi:10.1109/TPAMI.2012.120.
 40. Giraud, R.; Clément, M. Superpixel Segmentation: A Long-Lasting Ill-Posed Problem. *arXiv: Computer Vision and Pattern Recognition (cs.CV)* **2024**.
 41. Jampani, V.; Sun, D.; Liu, M.-Y.; Yang, M.-H.; Kautz, J. Superpixel Sampling Networks. In Proceedings of the European Conference on Computer Vision – ECCV 2018; Vittorio Ferrari, Martial Hebert, Eds.; Munich, Germany, November 8 2018.
 42. Fouad, S.; Randell, D.; Galton, A.; Mehanna, H.; Landini, G. Unsupervised Superpixel-Based Segmentation of Histopathological Images with Consensus Clustering. In Proceedings of the Communications in Computer and Information Science; Springer Verlag, 2017; Vol. 723, pp. 767–779.
 43. Boykov, Y.Y.; Jolly, M.-P. Interactive Graph Cuts for Optimal Boundary & Region Segmentation Of Objects in N-D Images. In Proceedings of the Proc. 8th IEEE Int. Conf. on Computer Vision (ICCV); July 2001; Vol. I, pp. 105–112.
 44. Shi, J.; Malik, J. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*.
 45. Rother, C.; Kolmogorov, V.; Blake, A. “GrabCut”-Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Transactions on Graphics (TOG)* **2004**, *23*, 309–314, doi:10.1145/1015706.1015720.
 46. Liu, B.; Zhang, T.; Yu, Y.; Miao, L. A Data Generation Method with Dual Discriminators and Regularization for Surface Defect Detection under Limited Data. *Comput. Ind.* **2023**, *151*, doi:10.1016/j.compind.2023.103963.
 47. Jha, S.B.; Babiceanu, R.F. Deep CNN-Based Visual Defect Detection: Survey of Current Literature. *Comput. Ind.* **2023**, *148*, doi:10.1016/j.compind.2023.103911.
 48. He, X.; Chang, Z.; Zhang, L.; Xu, H.; Chen, H.; Luo, Z. A Survey of Defect Detection Applications Based on Generative Adversarial Networks. *IEEE Access* **2022**, *10*, 113493–113512, doi:10.1109/ACCESS.2022.3217227.
 49. Gan, Y.; Ji, Y.; Jiang, S.; Liu, X.; Feng, Z.; Li, Y.; Liu, Y. Integrating Aesthetic and Emotional Preferences in Social Robot Design: An Affective Design Approach with Kansei Engineering and Deep Convolutional Generative Adversarial Network. *Int. J. Ind. Ergon.* **2021**, *83*, doi:10.1016/j.ergon.2021.103128.
 50. Kumar, V.; Hernández, N.; Jensen, M.; Pal, R. Deep Learning Based System for Garment Visual Degradation Prediction for Longevity. *Comput. Ind.* **2023**, *144*, doi:10.1016/j.compind.2022.103779.
 51. Hu, W.; Wang, T.; Chu, F. A Wasserstein Generative Digital Twin Model in Health Monitoring of Rotating Machines. *Comput. Ind.* **2023**, *145*, doi:10.1016/j.compind.2022.103807.
 52. Kim, S.; Jang, H.; Yoon, B. Developing a Data-Driven Technology Roadmapping Method Using Generative Adversarial Network (GAN). *Comput. Ind.* **2023**, *145*, doi:10.1016/j.compind.2022.103835.

53. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. **2016**.
54. Wang, R. Research on Image Generation and Style Transfer Algorithm Based on Deep Learning. *Open Journal of Applied Sciences* **2019**, *09*, 661–672, doi:10.4236/ojapps.2019.98053.
55. Bernardi, M. Generating Realistic Marble Textures Using Generative Adversarial Networks, Università degli Studi di Padova: Padova, 2023.
56. Guo, Y.; Smith, C.; Hašan, M.; Sunkavalli, K.; Zhao, S. MaterialGAN: Reflectance Capture Using a Generative SVBRDF Model. *ACM Trans. Graph.* **2020**, *39*, doi:10.1145/3414685.3417779.
57. INTERNATIONAL TELECOMMUNICATION UNION *Subjective Video Quality Assessment Methods for Multimedia Applications*; 2008;
58. Radiocommunication Bureau, I. *Recommendation ITU-R BT.500-14 Methodologies for the Subjective Assessment of the Quality of Television Images*; 2020;
59. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multi-Scale Structural Similarity for Image Quality Assessment. In Proceedings of the Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers; Pacific Grove, CA, January 9 2003.
60. Bermano, A.H.; Gal, R.; Alaluf, Y.; Mokady, R.; Nitzan, Y.; Tov, O.; Patashnik, O.; Cohen-Or, D. State-of-the-Art in the Architecture, Methods and Applications of StyleGAN. *Computer Graphics Forum* **2022**, *41*, 591–611, doi:10.1111/cgf.14503.
61. Salih, M.E.; Zhang, X.; Ding, M. Two Modifications of Weight Calculation of the Non-Local Means Denoising Method. *Engineering* **2013**, *05*, 522–526, doi:10.4236/eng.2013.510b107.
62. Chen, Y. *3D Texture Mapping for Rapid Manufacturing*; 2007; Vol. 4.
63. Saad, M.M.; Rehmani, M.H.; O'Reilly, R. Early Stopping Criteria for Training Generative Adversarial Networks in Biomedical Imaging. In Proceedings of the IEEE Irish Signals and Systems Conference (ISSC 2024); May 31 2024; pp. 1–7.
64. Haghighat, M.; Razian, M.A. Fast-FMI: Non-Reference Image Fusion Metric. In Proceedings of the 8th IEEE International Conference on Application of Information and Communication Technologies, AICT 2014 - Conference Proceedings; Institute of Electrical and Electronics Engineers Inc., 2014.
65. Fabre-Thorpe, M. The Characteristics and Limits of Rapid Visual Categorization. *Front. Psychol.* **2011**, *2*.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.