applied sciences

MDPI

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, 1049-001 Lisbon, Portugal; carlosmadiogo@tecnico.ulisboa.pt (C.M.A.D.); gustavo.paneiro@tecnico.ulisboa.pt (G.P.)
[2] Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal; margarida.figueiredo@tecnico.ulisboa.pt
[3] LAETA, IDMEC, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal; pedro.amaral@tecnico.ulisboa.pt
* 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

Deploying conditional Generative Adversarial Networks (cGANs) for industrial texture synthesis faces two barriers: the prohibitive cost of manual data annotation and the uncertain alignment between automated evaluation metrics and human perception. This study addresses both challenges for marble texture synthesis using 289 high-resolution industrial scans. 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 without manual annotation. Four cGAN architectures—baseline cGAN, Pix2Pix, BicycleGAN, and GauGAN—are benchmarked using a dual-evaluation protocol contrasting ten automated metrics with structured human-centered assessment. The results reveal a significant metric–perception discrepancy. Pix2Pix achieved the best Fréchet Inception Distance (FID = 85.3) yet received the lowest human ratings due to periodic texture artifacts. GauGAN produced textures statistically indistinguishable from real marble, achieving a Visual Turing Pass Rate (VTPR) of 0.533 and a Mean Opinion Score on Marble Authenticity (MOS-MA) of 2.89, despite an inferior FID (87.3). These findings make three contributions: an annotation-free segmentation pipeline, empirical evidence that automated metrics alone are insufficient for architecture selection, and a dual-evaluation framework that establishes human-in-the-loop assessment as essential for quality-critical industrial deployment.

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

Academic Editors: Sandra Pereira, António Manuel Trigueiros Da Silva Cunha and Paulo Jorge Coelho

Received: 9 February 2026
Revised: 5 April 2026
Accepted: 15 April 2026
Published: 21 April 2026

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. This challenge stems from 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. Firstly, 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]. Secondly, the standard evaluation paradigm relies exclusively on automated metrics such as the Fréchet Inception Distance (FID) [9], Inception Score (IS) [10], and MS-SSIM, which were originally designed for object recognition rather than texture quality and have been shown to exhibit weak or negative correlations with human perceptual judgments in texture synthesis tasks [11–14]. Specifically, Zhou et al. [12] demonstrated across multiple datasets that FID scores fail to correlate with human judgments. Stein et al. [13] confirmed this across 207,000 perceptual judgments, and Borji [14] documented the FID's blind spot for domain-specific image quality. This fundamental misalignment reinforces the position that human-centered evaluation is essential for quality-critical industrial applications [15].

The evolution of texture synthesis spans three major paradigms. Traditional procedural generation, exemplified by Perlin noise [16,17], produces marble-like patterns through mathematical algorithms but achieves limited realism and requires expert tuning [18,19]. Example-based non-parametric methods leverage real source images but struggle with large-scale structures such as continuous marble veins [20–24]. Modern deep generative models, particularly GANs [25–27], have overcome these limitations through learned synthesis: neural style transfer [11,28] established the principle of separating structure and appearance; the Pix2Pix framework [29] established paired image-to-image translation using U-Net generators and PatchGAN discriminators; BicycleGAN [30] addressed mode collapse through bijective latent-to-image mappings; and GauGAN [6] introduced Spatially Adaptive Normalization (SPADE), modulating normalization parameters as learned functions of the input mask at every generator layer, critical for preserving vein boundaries while synthesizing organic appearance [30]. While diffusion models have achieved state-of-the-art results, their deployment faces practical barriers in manufacturing environments with limited proprietary data (~200–500 samples), making cGANs a practical choice. Conditional variants like ControlNet [31] rely on foundation models pre-trained on billions of images, and comprehensive comparison constitutes valuable future work [32].

The annotation bottleneck has motivated unsupervised segmentation research in medical imaging [33–35], yet adoption in manufacturing remains limited [4,8]. Unsupervised pipelines combining SLIC superpixels [36,37] for perceptually coherent region grouping [38], Gaussian Mixture Models [39,40], and graph cut [40] or Normalized Cuts [41] for spatial regularization can automatically extract structural features without manual labeling. This work adapts Borovec et al.'s pipeline [34] to natural stone, demonstrating

that marble vein structures suitable for cGAN training can be extracted from raw industrial scans without annotation.

Recent industrial applications demonstrate the transformative potential of GANs. In manufacturing quality control, GANs address data scarcity for defect detection [42] and anomaly detection [43,44]. Beyond inspection, GANs contribute to design optimization [45], product lifecycle prediction [46], and digital twin systems [47]. Strategic applications include technology road mapping [48] and custom material design [49,50]. However, prior work on natural materials like marble remains scarce [51,52], with most studies focusing on regular repeating patterns rather than stochastic geological textures. Beyond visual synthesis, generative and adversarial training paradigms have demonstrated broad utility across diverse industrial scenarios. These include GAN-based data augmentation for rotating machinery fault diagnosis under data scarcity [53], domain adaptation for unsupervised fault detection in manufacturing equipment [54], and multimodal deep learning for anomaly detection in railway infrastructure [55,56]. This breadth underscores the versatility of adversarial frameworks for applied engineering problems.

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 [34] 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 the marble type commercially known as Exotic Ambar from a single quarry, providing a controlled testbed for architectural comparison without confounding variables. 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. This provides a practical solution to the annotation bottleneck where pixel-perfect labeling is economically prohibitive.
2. We conduct a systematic benchmark comparing four cGAN architectures trained and evaluated under identical conditions on 289 high-resolution industrial marble scans. This provides 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. This establishes 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 begins 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.

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 the longest dimension, scanned at a mean resolution of 7185 $\times$ 4166 pixels under 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.1. All images underwent standardized pre-processing: 200-pixel border cropping to remove frame artifacts, bicubic resampling to 1280 $\times$ 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.

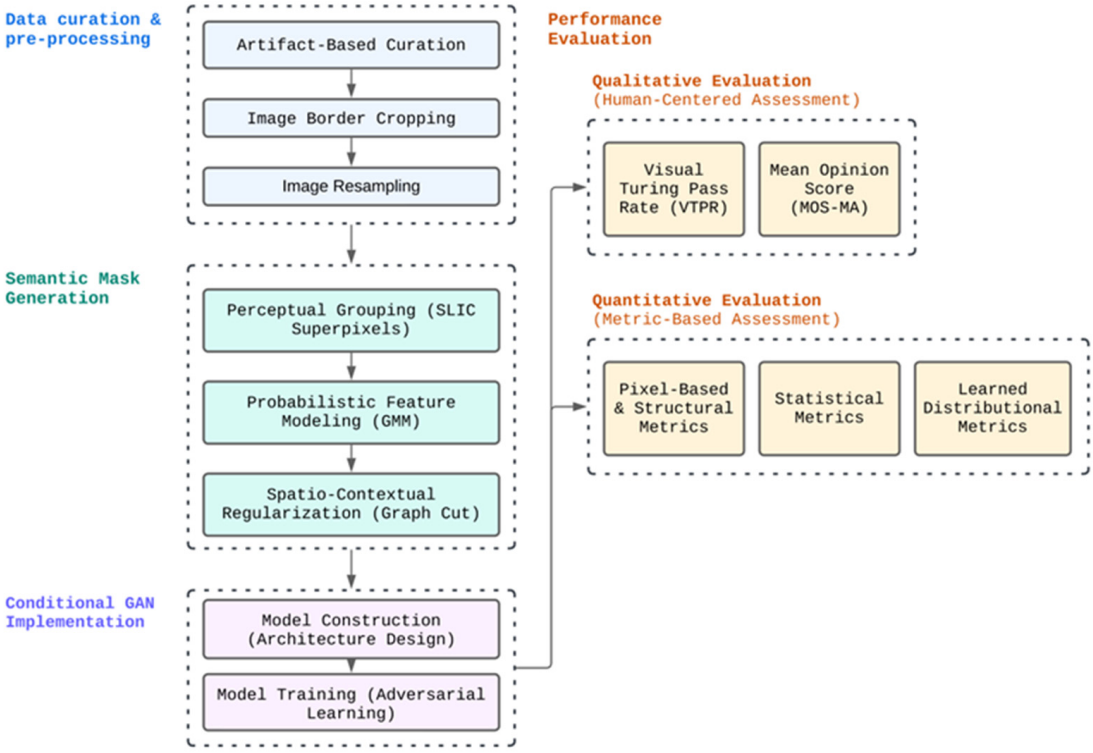


Figure 1. Flowchart of the proposed pipeline for controllable marble texture synthesis. The workflow proceeds through four stages: (1) data curation and pre-processing, including artifact-based filtering and image standardization; (2) semantic mask generation via SLIC superpixel segmentation, GMM-based 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).

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.

Step 1—SLIC Superpixel Over-Segmentation. The SLIC algorithm reduces each image to approximately 3000 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, yielding a robust set of primitive regions for further analysis [39]. Each superpixel is characterized by a 9-dimensional feature vector encoding the CIELAB mean, standard deviation, and median.

Step 2—GMM Probabilistic Color Classification. 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.

Step 3—Graph Cut Spatial Regularization. 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 the 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.2.

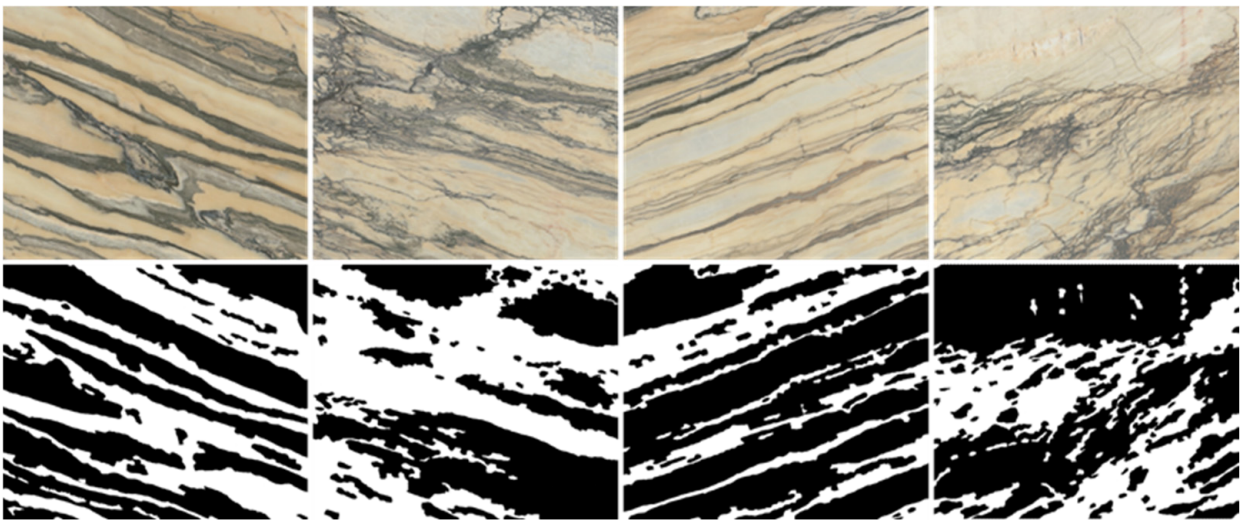


Figure 2. Representative marble slabs and unsupervised mask generation. Top row: Four samples showing natural variation in vein density and orientation from dataset of the marble with commercial name Exotic Ambar. Bottom row: Corresponding binary masks generated automatically via SLIC + GMM + graph cut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

Modern foundation models such as SAM [63] and self-supervised clustering approaches based on DINOv2 [64] or DeepCluster [65] represent powerful alternatives for unsupervised segmentation. However, these methods introduce domain-dependency constraints that conflict with our annotation-free objective: SAM has been shown to struggle with out-of-distribution industrial materials data and requires domain-specific fine-tuning to achieve reliable segmentation, thereby reintroducing annotation requirements [66]. DINOv2-based approaches similarly depend on large-scale pre-training on natural image corpora, creating a domain shift risk for proprietary manufacturing textures. The SLIC + GMM + graph cut pipeline was selected precisely to avoid these dependencies: it requires no pre-training, no external model, and is fully governed by interpretable, tunable parameters—making it directly deployable on new material types without any labeled data.

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 [67], and GauGAN with Spatially Adaptive Normalization [6]. All four architectures share a common training infrastructure held fixed across experiments to isolate architectural differences in the generators. Each model employs an identical PatchGAN discriminator (70×70 receptive field), processing images at 1280×720 pixels. Training used the Adam optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$) with batch size 4 on a single NVIDIA H100 GPU, with all network weights initialized from a Gaussian distribution ($\mu = 0$, $\sigma = 0.02$). Regularization included one-sided label smoothing and gradient clipping across all architectures. Training ran for a maximum of 10,000 epochs with an early stopping patience of 2000 epochs, monitoring the validation FID [68], and convergence-based termination when validation metrics plateaued. Architecture-specific hyperparameters (learning rates, loss weights, latent dimensions) are reported in Table 1, determined through preliminary sweeps that optimized the validation FID. Complete architecture diagrams are provided in Appendix A.3.

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

Architecture	Generator LR	Discriminator LR	Encoder LR	λL1	λ_{latent}	Latent Dim	Total Training Epochs
cGAN	1×10^{-4}	5×10^{-5}	—	—	—	100	6900
Pix2Pix	2×10^{-4}	2×10^{-4}	—	100	—	—	5100
BicycleGAN	2×10^{-4}	2×10^{-4}	2×10^{-4}	100	10	8	3400
GauGAN	5×10^{-5} †	5×10^{-5}	—	350	—	—	3100

† GauGAN learning rate follows an exponential decay schedule (decay rate 0.995 every 2500 steps) applied identically to both generator and discriminator optimizers; the value reported is the initial learning rate.

Baseline cGAN. 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. Separate learning rates were applied to the generator (1×10^{-4}) and discriminator (5×10^{-5}). 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 [5].

Pix2Pix. 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 to the adversarial objective to enforce fidelity to a paired ground-truth image. The combined objective is as follows:

$$\mathcal{L}_{Pix2Pix} = \mathcal{L}_{\text{cGAN}}(G, D) + \lambda \mathcal{L}_{\mathcal{L}1}(G) \quad (1)$$

where $\mathcal{L}_{\mathcal{L}1}(G) = \mathbb{E}[|y - G(x, z)|_1]$ enforces pixel-level fidelity to the paired ground-truth image y, and $\lambda = 100$ controls the relative weight of reconstruction versus adversarial training. Both the generator and discriminator were trained with a learning rate of 2×10^{-4}. Pix2Pix demonstrated that a single cGAN framework can synthesize plausible photos from diverse input representations, effectively separating content from style [7].

BicycleGAN. BicycleGAN extends this framework by introducing a dedicated encoder network that learns to invert generated images back to latent codes (latent dimension

$z = 8$), enforcing a bijective latent-to-image mapping through both an L1 reconstruction loss ($\lambda_{L1} = 100$) and a latent regression loss ($\lambda_{latent} = 10$) to combat mode collapse and enable diverse texture generation from identical mask inputs. The generator, discriminator, and encoder were each trained with a learning rate of 2×10^{-4}. This addresses the one-to-many mapping inherent in image translation, where a given mask could correspond to multiple realistic appearances [67].

GauGAN. 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 block incorporates Spatially Adaptive Normalization (SPADE) layers that modulate the normalization parameters (γ, β) as learned spatial functions of the input mask [6], thereby preserving the semantic structure at every layer. This avoids the washing-away effect of standard batch normalization that plagues earlier architectures, critical for preserving vein boundaries while synthesizing organic appearance [30]. GauGAN was trained with an exponentially decaying learning rate (initial rate 5×10^{-5}, decay rate 0.995 every 2500 steps) and an L1 reconstruction weight of $\lambda = 350$. Formally, for activation h^i at layer i, SPADE computes.

$$\hat{h}^i_{n,c} = \gamma^i_c(\boldsymbol{m}) \frac{h^i_{n,c} - \mu^i_c}{\sigma^i_c} + \beta^i_c(\boldsymbol{m}) \quad (2)$$

where μ^i_c and σ^i_c are the per-channel mean and standard deviation of h^i, and $\gamma^i_c(\boldsymbol{m})$ and $\beta^i_c(\boldsymbol{m})$ are spatially varying scale and bias parameters learned as convolutional functions of the input segmentation mask $\boldsymbol{m}$, ensuring the semantic structure is preserved at every layer of the generator [6]. This architectural choice has a direct consequence for artifact generation. Pix2Pix's U-Net decoder relies on transposed convolutions for upsampling. When the kernel size is not a multiple of the stride, the transposed convolution produces uneven overlap across output pixels, causing certain positions to receive disproportionate contributions and generating a periodic, grid-like intensity pattern known as the checkerboard artifact [56]. GauGAN avoids this entirely by upsampling via nearest-neighbor interpolation followed by standard convolution, so every output pixel receives identical contributions from its neighborhood. The SPADE layers then re-inject spatial structure from the input mask at every resolution level, ensuring semantic fidelity without any dependence on transposed convolution upsampling.

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, despite evidence that these metrics correlate poorly with human perceptual judgments [12–14]. Most recent industrial GAN papers nonetheless defer human evaluation to future work, relying solely on the 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 3a) 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 3b) 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) [69], 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. Together, these metrics provide complementary perspectives across reconstruction accuracy, structural preservation, texture statistics, and learned feature similarity, with the FID serving as the primary convergence criterion during training.

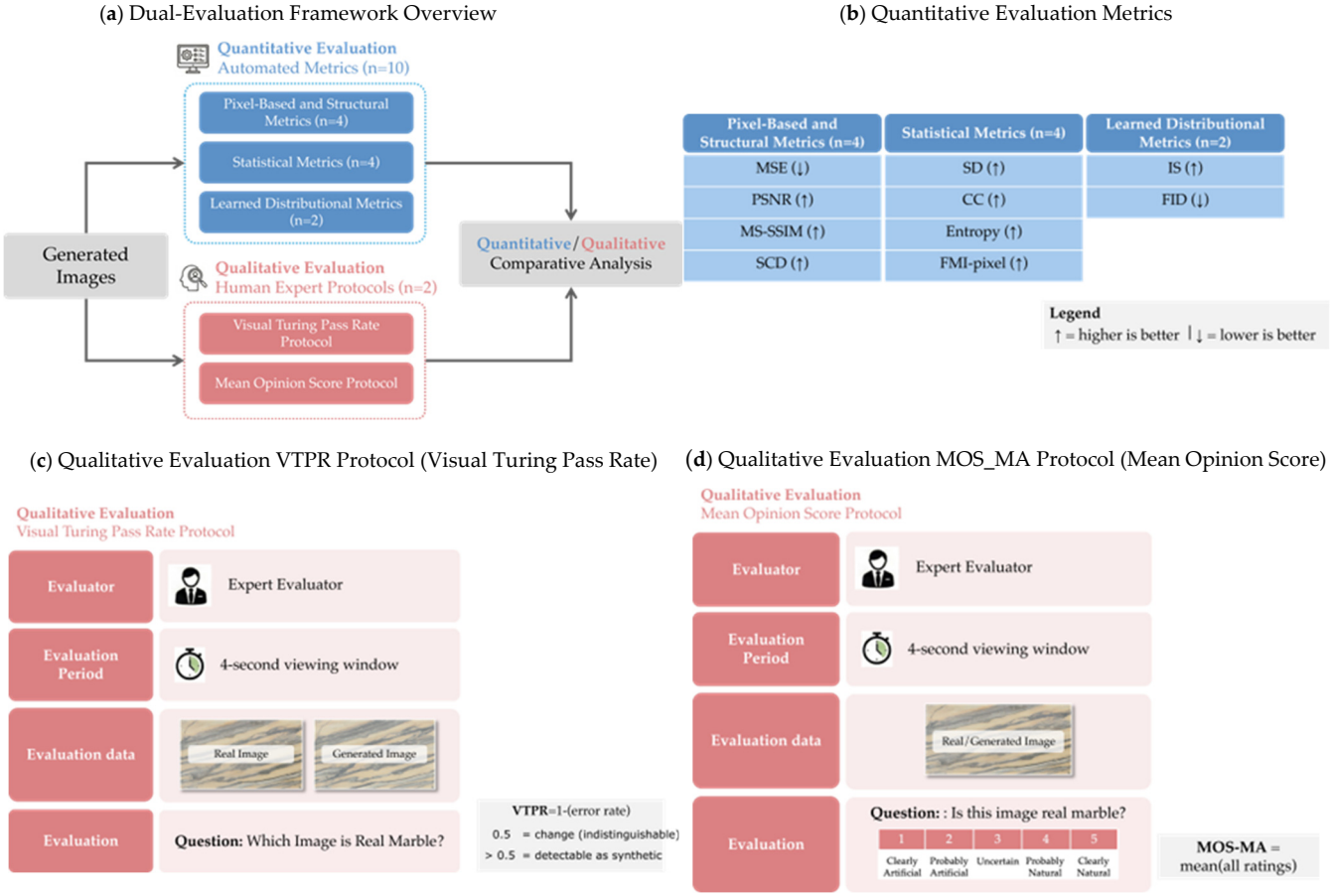


Figure 3. Dual-evaluation framework for comparative assessment of automated metrics versus human perceptual judgment. (**a**) Framework overview: generated images undergo parallel quantitative (10 automated metrics) and qualitative evaluation (2 protocols with human expert evaluators), with results compared to identify metric–perception alignment or divergence. (**b**) Quantitative metrics battery: ten metrics spanning pixel-based and structural (MSE, PSNR, MS-SSIM, SCD), statistical (SD, CC, Entropy, FMI-pixel), and learned distributional families (IS, FID). Arrows indicate optimization direction (↑ = higher is better; ↓ = lower is better). (**c**) VTPR protocol: three domain experts perform 60 two-alternative forced-choice trials, identifying real versus synthetic marble within 4 s viewing windows. VTPR = 1-(error rate); 0.5 indicates perfect indistinguishability (chance level), values >0.5 indicate detectably synthetic outputs. (**d**) MOS-MA protocol: experts rate images on 5-point Likert scale (1 = "Clearly Artificial", 5 = "Clearly Natural") without time constraints. MOS-MA = mean rating across all trials.

The qualitative component implements two human-centered protocols adapted from telecommunications quality assessment standards [57,58]. Protocol 1 (Visual Turing Pass Rate, VTPR; Figure 3c) 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 ~150 ms 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 s, and experts identify which is real. This time constraint captures glance-level authenticity corresponding to feedforward visual processing [70] while preventing extended artifact-focused scrutiny, consistent with the HYPE benchmark methodology that achieved strong statistical reliability at approximately $60 per assessment [12]. The 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 3d) 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 the detection of architectures that achieve metric-based optimization at the expense of perceptual authenticity, the core hypothesis motivating this evaluation strategy.

2.4. Frequency-Domain Artifact Analysis

To quantitatively characterize the periodic texture artifacts identified in Section 2.2, we performed power spectral density (PSD) analysis on the outputs generated by all four architectures. For each architecture, all generated images were decomposed into 256 × 256-pixel patches with a 50% stride overlap. Each patch was converted to grayscale and zero-mean normalized before application of a 2D Fast Fourier Transform (FFT). The resulting power spectrum was radially averaged to yield a 1D profile of mean log power as a function of spatial frequency (cycles/pixel), collapsing directional information while preserving periodic structure at any orientation. Profiles were averaged across all patches and all images per architecture. To isolate artifact band behavior, difference curves were computed by subtracting the real marble mean profile from each architecture's profile; positive values indicate spectral excess above natural marble texture at that frequency.

3. Results

This section presents the results across four evaluation dimensions: (1) visual quality: qualitative comparison of generated textures across architectures; (2) automated metric performance: quantitative evaluation across 10 metrics on the 57-sample validation set; (3) human-centered assessment: structured expert evaluation using VTPR and MOS-MA protocols; and (4) metric–perception divergence analysis: direct comparison of automated and human rankings. All quantitative results are reported on the 57-sample held-out validation set, independent of training data. The central finding is a critical ranking inversion: Pix2Pix achieves the best automated metric performance (FID = 85.286) yet receives the worst human ratings (MOS-MA = 2.333), while GauGAN produces textures statistically indistinguishable from real marble (VTPR = 0.533) despite an inferior FID (87.308). The computational efficiency results are reported in Section 3.5 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 automatically generated masks' 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).

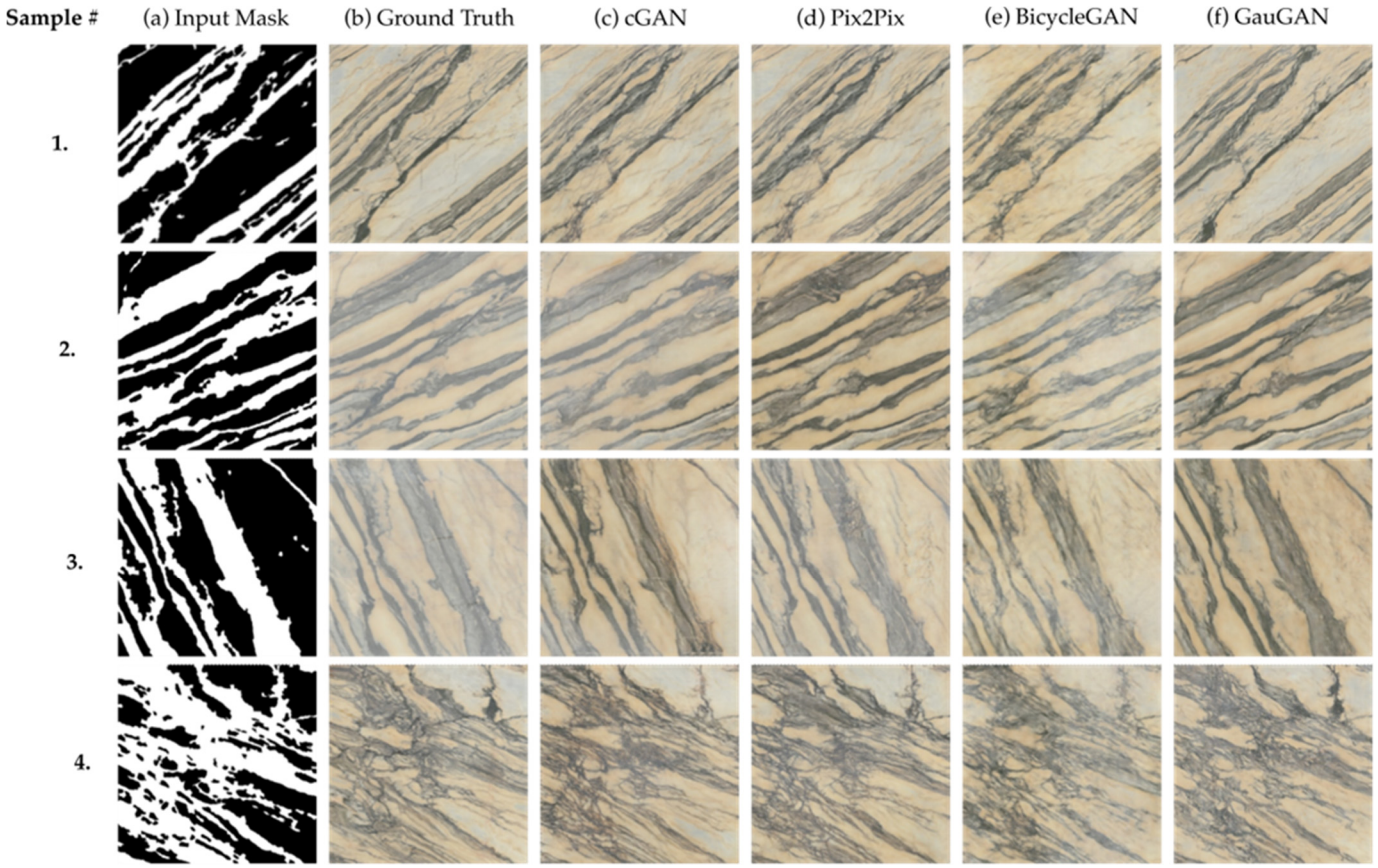


Figure 4. Qualitative comparison of marble texture synthesis across four cGAN architectures. (**a**) Input binary mask. (**b**) Ground-truth real marble. (**c**) cGAN output. (**d**) Pix2Pix output. (**e**) BicycleGAN output. (**f**) GauGAN output. Rows show samples with varying vein density and orientation.

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. However, 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

unsupervised segmentation pipeline as an effective source of geometric guidance for the generative process.

3.2. Quantitative Metrics: Automated Performance Evaluation

Table 2 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, which enforces 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 for all architectures across the 57-sample validation set are shown as sorted-value plots in Appendix A.4 (Figure A4a–h). These distributions confirm that GauGAN's perceptual and reconstruction advantages are consistent across the entire validation set and not driven by outliers: GauGAN's PSNR, MS-SSIM, and CC curves remain above those of competing architectures at nearly every sample rank position, while Pix2Pix's texture contrast advantage (SD, Entropy) is similarly consistent rather than sample-specific.

Table 2. Automated performance metrics on validation set (57 samples, mean ± std). Bold indicates best performance per metric. (↑ Higher is better; ↓ Lower is better.)

	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 5a presents the Visual Turing Pass Rates (VTPRs): 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. This constitutes a full ranking inversion between automated and human evaluation criteria, as summarized in Table 3.

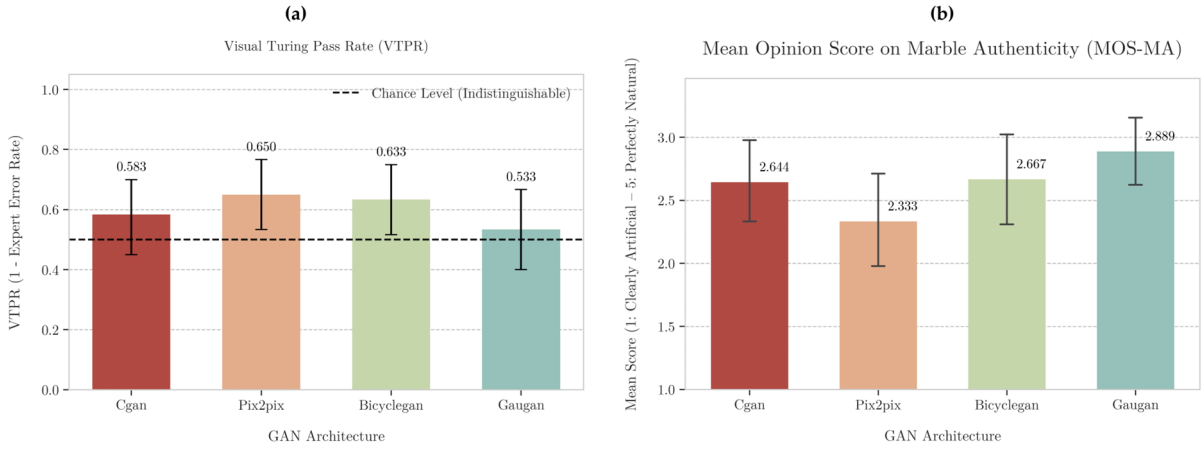


Figure 5. Human-centered evaluation reveals ranking inversion relative to automated metrics. (**a**) Visual Turing Pass Rate (VTPR): fraction of trials where expert evaluators correctly identified synthetic images as real (lower values indicate more realistic outputs that fool experts). Dashed line at 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.

Table 3. Architecture ranking under automated (FID) versus human (MOS-MA) evaluation criteria. Lower FID rank = better automated performance; higher MOS-MA rank = better perceptual performance. *(↑ Higher is better; ↓ Lower is better.)*

Architecture	FID Rank (↓ Better)	FID Score	MOS-MA Rank (↑ Better)	MOS-MA Score
Pix2Pix	1st	85.286	4th	2.333
GauGAN	2nd	87.308	1st	2.889
cGAN	3rd	94.623	3rd	2.644
BicycleGAN	4th	100.071	2nd	2.667

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 5b) 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 an MOS-MA of 2.644 (95% CI: 2.333–2.955), and BicycleGAN scored 2.667 (95% CI: 2.356–2.978), scoring comparably in the intermediate range.

These human-centered evaluations showed a somewhat inverted ranking relative to automated metrics: the architecture optimized for the 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 the Inception-based distributional distance yields perceptually superior outputs, particularly for stochastic texture synthesis tasks where fine-scale irregularity defines naturalness.

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 bet-

ter human perceptual scores. This misalignment contradicts the foundational assumption of metric-based GAN development: that minimizing the 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.

This finding validates the 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.

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 6c). 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 an 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.

To move beyond qualitative observation, we applied power spectral density (PSD) analysis to provide quantitative frequency-domain evidence for this artifact structure (Figure 7). The Pix2Pix's U-Net decoder employs four stride-2 transposed convolution layers, generating a theoretically predicted artifact period of 2^4 = 16 pixels (0.0625 cycles/pixel). Radially averaged PSD profiles were computed across all generated images for each architecture and subtracted from the real marble baseline to isolate spectral excess. Pix2Pix is the only architecture exhibiting positive spectral elevation above real marble at the predicted frequency (Δ log power = +0.0068 at 0.0625 c/px). GauGAN and BicycleGAN, both architecturally free of stride-2 transposed convolutions, show negative elevation (-0.0454 and -0.0337, respectively), indicating spectrally cleaner outputs than real marble in this band. The alignment between the prior architectural prediction and the measured spectral elevation confirms that the Pix2Pix artifact is structural and deterministic, not a consequence of training instability or dataset characteristics.

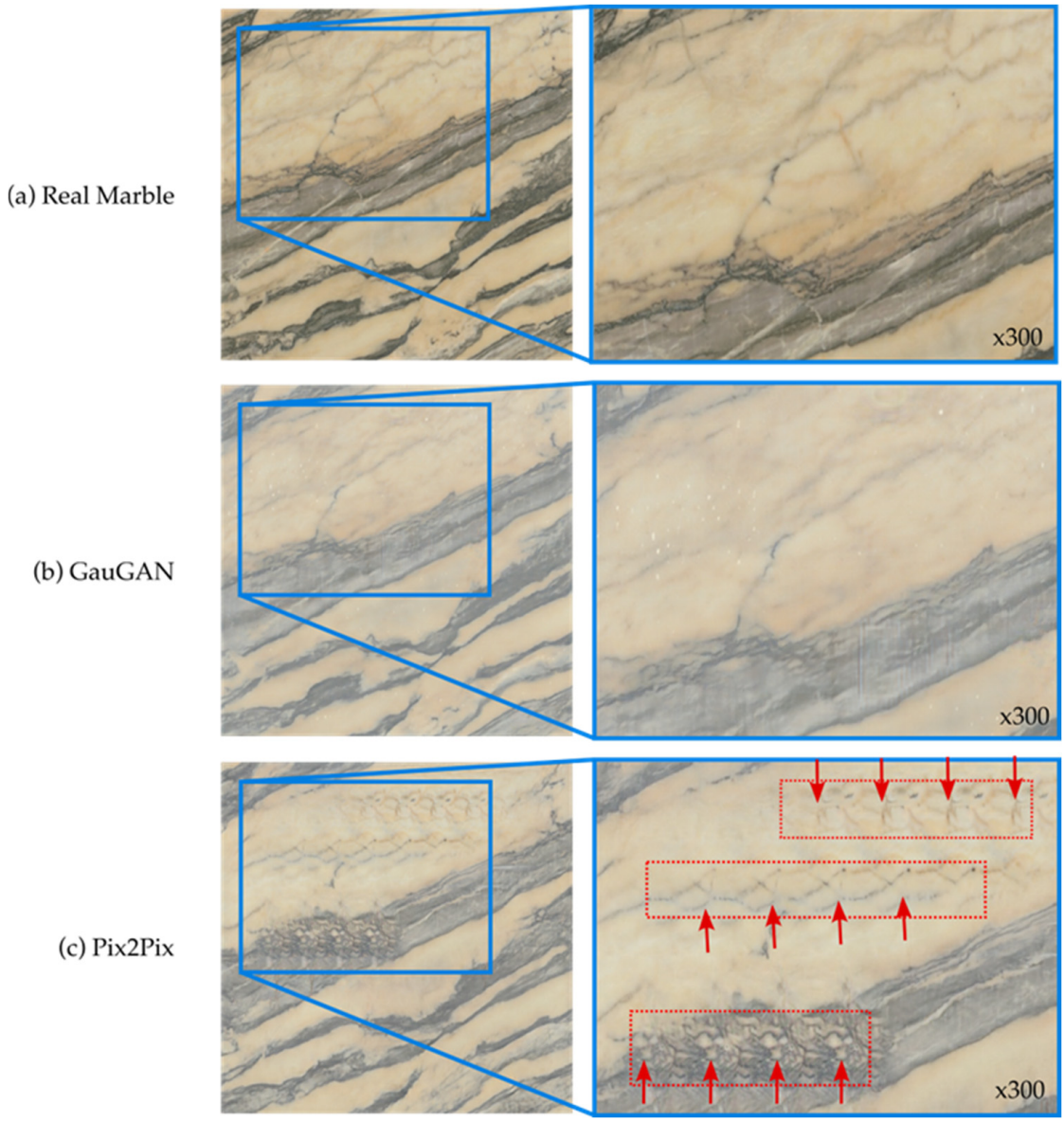


Figure 6. Visual artifact analysis through magnified comparison of synthesized marble textures. (**a**) Real marble: aperiodic, stochastic microstructure exhibiting continuous local variation in vein morphology and matrix crystallization, with no repeating spatial motifs. (**b**) GauGAN output: organic texture variation consistent with natural marble under magnification, with smooth vein-to-matrix transitions and no detectable periodicity. (**c**) Pix2Pix output: arrows indicate recurring texture motifs—specific vein branching geometries and grain arrangements—appearing repeatedly in identical orientations within local neighborhoods, violating the aperiodic character of natural mineral crystallization.

3.5. Computational Efficiency and Deployment Feasibility

Figure 8 presents the training evolution across all four architectures, revealing distinct convergence patterns that explain their final performance characteristics. The FID evolution (Figure 8c) demonstrates that GauGAN achieved the fastest convergence with most stable trajectory, reaching its final FID of 87.308 within 1000 epochs and maintaining stability thereafter. Pix2Pix exhibited a slower but consistent FID improvement, achieving its best score of 85.286 at epoch 3000. 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.

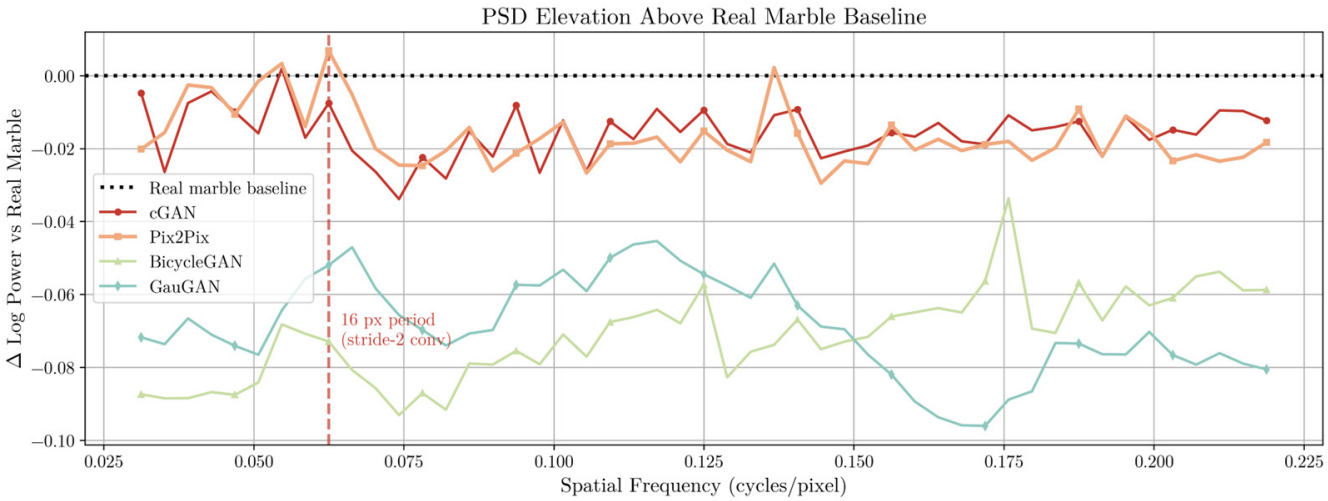


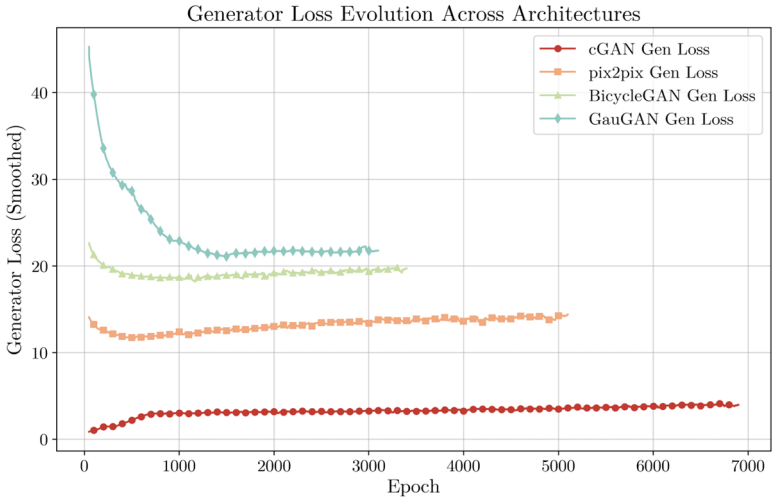
Figure 7. Power spectral density elevation above real marble baseline (Δ log power) as a function of spatial frequency for all four cGAN architectures. The vertical dashed line marks 0.0625 cycles/pixel (16-pixel period), the theoretically predicted artifact frequency for a U-Net decoder with four stride-2 transposed convolution layers. Pix2Pix is the only architecture with positive elevation at this frequency, confirming the spectral signature of the checkerboard artifact. GauGAN and BicycleGAN show negative elevation, indicating spectrally cleaner outputs than real marble in the artifact band.

The generator loss trajectories (Figure 8a) reveal the architectural differences in learning dynamics. GauGAN exhibited the most dramatic initial loss decrease, dropping from ~45 to ~21 within the first 1000 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.

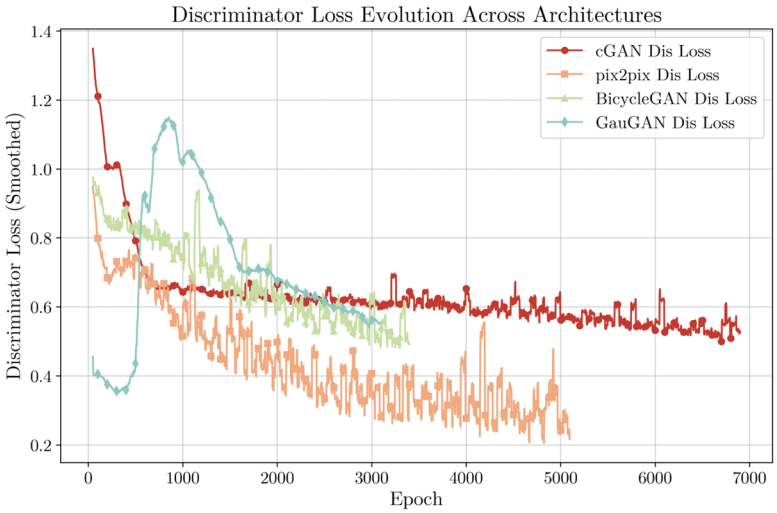
Discriminator loss evolution (Figure 8b) provides insight into adversarial training stability. GauGAN's discriminator initially struggled (loss ~0.4, indicating overly confident predictions) before stabilizing around 0.6–0.7, suggesting the generator learned to produce challenging outputs that maintained discriminator uncertainty. Pix2Pix's discriminator loss decreased steadily from ~1.2 to ~0.2, indicating the discriminator became increasingly confident in detecting synthetic images. This progressive detectability may explain why human evaluators also found Pix2Pix outputs more detectable despite a superior FID. The baseline cGAN and BicycleGAN maintained more balanced discriminator losses (~0.5–0.6), consistent with the Nash equilibrium in adversarial training.

Table 4 reports the computational characteristics relevant for industrial deployment. GauGAN requires the fewest trainable parameters (53.1 M) 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 h to convergence (3100 epochs) compared to 10.2 h for Pix2Pix (5100 epochs × 0.12 min/epoch) and 12.7 h for the baseline cGAN (6900 epochs × 0.11 min/epoch). Pix2Pix offers competitive training efficiency (0.12 min/epoch) at a parameter count of 61.4 M. At inference time, all models achieve performance on consumer GPUs for 1280 × 720 outputs appropriate for integration in interactive design tools. Given GauGAN's superior perceptual quality validated through human evaluation, the 4.2× total training time penalty relative to Pix2Pix (42.4 vs. 10.2 h)

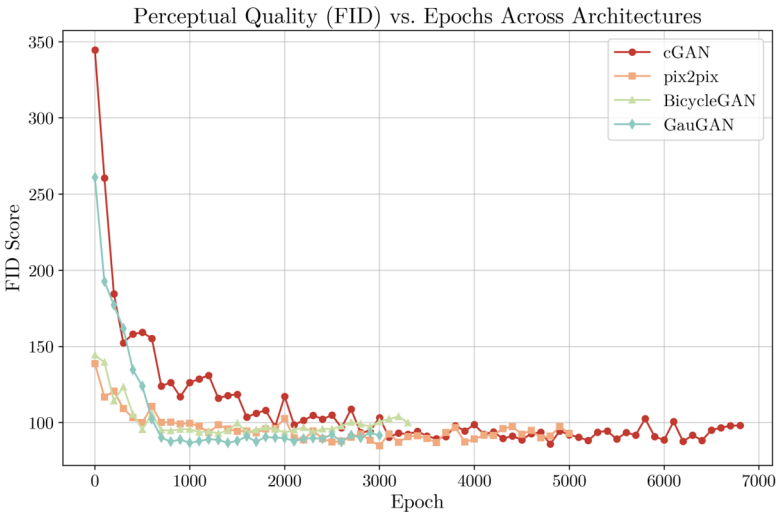
is justified for quality-critical applications like architectural visualization, where human quality perception is the ultimate arbiter.



(**a**)



(**b**)



(**c**)

Figure 8. Training dynamics across the four architectures. (**a**) Generator loss convergence. (**b**) Discriminator loss evolution. (**c**) FID evolution over epochs.

Table 4. 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. As shown in Table 2 and Figure 5, the quantitative perceptual divergence between Pix2Pix and GauGAN demonstrates that Inception-based metrics fail to capture human-relevant texture quality in stochastic material synthesis. This finding aligns with large-scale empirical evidence that models producing more perceptually realistic images paradoxically score worse on the 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 the FID is particularly unsuitable for specialized domains where visual features differ from those in natural images.

The architectural origin of this failure mode can be traced to Pix2Pix's use of transposed convolutions for spatial upsampling in its U-Net decoder. As formally characterized by Odena et al. [56], transposed convolutions with a kernel size that is not a multiple of the stride produce uneven overlap in the output: certain output pixels receive contributions from a disproportionate number of input activations, creating a periodic, grid-like intensity pattern that manifests as systematic texture repetition at characteristic spatial frequencies. This artifact is not a consequence of training instability or dataset limitations, it is an inherent structural property of the upsampling operator.

The PSD analysis in Section 3.4 provides direct spectral confirmation of this mechanism: the measured elevation at the architecturally predicted frequency (0.0625 c/px, 16 px period) is unique to Pix2Pix and absent in GauGAN, corroborating Odena et al.'s [56] theoretical framework with empirical frequency-domain evidence on industrial texture data.

GauGAN eliminates this mechanism entirely. Its generator progressively upsamples a learned constant tensor through six residual blocks using nearest-neighbor interpolation followed by standard convolution, ensuring that every output pixel receives identical contributions from its local neighborhood with no uneven overlap. The SPADE normalization layers then re-inject spatial structure derived from the input mask at each resolution level, preserving vein geometry and semantic boundaries without relying on transposed convolution upsampling. This combination of artifact-free upsampling and per-layer mask conditioning explains why GauGAN replicates the aperiodic, stochastic character of natural marble crystallization that Pix2Pix's decoder architecture structurally cannot produce.

This finding has significant implications for industrial applications: as established empirically in Section 3.4, sole reliance on the FID for architecture selection in quality-critical manufacturing contexts carries direct economic consequences for workflows in which 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.

Contextualizing these results against related work is challenging due to the scarcity of studies combining automated and human evaluation for industrial material texture synthesis. The closest prior work on marble GAN synthesis, Bernardi (2023) [51], demonstrated the feasibility of GAN-based marble texture generation but relied exclusively on automated metrics without human validation. Additionally, results were reported on a single architecture without a comparative benchmark, limiting direct performance comparison. Our dual-evaluation protocol represents, to our knowledge, the first systematic human-validated benchmark for this material class. Regarding the VTPR, GauGAN's result of 0.533 (95% CI: 0.400–0.667) is directly comparable to top-tier generator performance in large-scale human evaluation benchmarks. Specifically, the HYPE benchmark [12] reported that leading models across diverse image categories approach the chance-level threshold of 0.5, with the best generators achieving VTPR values in the 0.52–0.56 range under analogous time-constrained 2AFC conditions. Our GauGAN result falls squarely within this range, confirming that SPADE-based synthesis achieves competitive human indistinguishability for a specialized industrial domain with only 289 training samples. The FID range observed in our study (85–100) is substantially higher than the values reported for face synthesis (typically FID < 10 for state-of-the-art models) or natural scene generation (FID < 30), consistent with the known sensitivity of Inception-v3 to domain shift from its ImageNet training distribution [14]. Marble textures share little statistical structure with ImageNet object categories, inflating the absolute FID values while preserving their relative discriminative utility for cross-architecture comparison within this domain.

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 1761.3 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. Under

the experimental conditions of this study, these findings suggest tentative decision criteria: applications requiring rapid iteration may prioritize Pix2Pix despite detectable artifacts, whereas quality-critical applications justify GauGAN's 4.2× training time investment. These guidelines should, however, be validated across additional marble types before deployment-scale adoption.

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. This 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 [35]. 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

Beyond the experimental scope of this benchmark, several deployment-specific risks warrant consideration before adopting GauGAN in production environments. Firstly, domain shift poses a practical concern: all models were trained on a single marble variety (Exotic Ambar) from one quarry, and architectural rankings may not generalize to marble types with substantially different vein morphologies, colorimetric profiles, or crystallization textures. Secondly, GAN training involves inherent stochasticity; without strict random seed control, the results may exhibit run-to-run variance that could affect reproducibility in industrial quality pipelines. Thirdly, GauGAN's architectural complexity, specifically the per-pixel SPADE normalization blocks, results in a 4.2× higher training cost relative to Pix2Pix (42.4 vs. 10.2 h) and greater fine-tuning overhead as production data evolve, with implications for long-term maintenance in dynamic manufacturing environments. These considerations reinforce a conservative reading of our recommendations: GauGAN performs best under the specific conditions evaluated here, and deployment-scale adoption should be preceded by validation across a broader range of material classes.

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

Firstly, 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.

Secondly, the statistical power of human evaluation is constrained by the sample size. While sufficient for exploratory validation and aligned with practices in the 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.

Thirdly, we deliberately excluded diffusion models from this benchmark for methodological rigor: comparing models with fundamentally different training paradigms (adversarial vs. denoising diffusion), 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.

Fourthly, 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 divergence between metric and human perception: 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 an inferior FID. This finding establishes that human-in-the-loop evaluation is essential for deployment decisions in quality-critical applications. Under the experimental conditions of this study—a single marble type (Exotic Ambar), 289 industrial scans, and expert evaluation by three domain specialists—GauGAN demonstrated the strongest perceptual performance and represents the preferred architecture for quality-critical applications. Pix2Pix remains a viable option when computational efficiency is the primary constraint. However, these recommendations should be interpreted as condition-specific: validation across additional marble varieties and larger expert panels is required before broader deployment-scale conclusions can be drawn. Future work will extend this framework to diverse geological materials and compare against conditional diffusion models.

Author Contributions: Conceptualization, A.A.d.C., M.F. and G.P.; methodology, A.A.d.C. and M.F.; software, A.A.d.C. and C.M.A.D.; validation, A.A.d.C.; formal analysis, A.A.d.C. and M.F.; investigation, A.A.d.C. and M.F.; resources, G.P. and P.A.; data curation, A.A.d.C. and C.M.A.D.; writing—original draft preparation, A.A.d.C.; writing—review and editing, A.A.d.C., M.F., C.M.A.D., G.P. and P.A.; visualization, A.A.d.C. and C.M.A.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 was funded by 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).

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the non-interventional nature of the human evaluation protocol. The study involved only the perceptual assessment of industrial material images (marble textures) by three domain expert evaluators. No personal data, medical information, biometric data, or sensitive information of any kind was collected. Participants were not subjected to any physical, psychological, or social intervention. The evaluation protocol — comprising two-alternative forced-choice visual trials and Likert-scale quality ratings of marble texture images — does not meet the threshold for ethical review under the applicable national legislation (Portuguese Law No. 21/2014, of 16 April, on Clinical Research, and the guidelines of the Commission Nacional de Proteção de Dados), which exempts studies that do not involve medical procedures, sensitive personal data, or identifiable participant information.

Informed Consent Statement: Informed consent for participation was obtained verbally from all expert evaluators involved in the study prior to the evaluation sessions. Verbal consent was obtained rather than written because the study involved only non-invasive visual assessment of industrial material images, with no collection of personal, medical, or identifying information, and posed no risk to participants. All evaluators were fully informed of the study's purpose, the voluntary nature of their participation, the anonymized use of their ratings for research publication, and their right to withdraw at any time without consequence. No identifying participant information is reported in this manuscript.

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: The authors, António Alves de Campos, Carlos M. A. Diogo and Gustavo Paneiro gratefully acknowledge the support of the CERENA through FCT Project UID/04028/2025 (https://doi.org/10.54499/UID/PRR2/04028/2025, accessed on 1 February 2026). The author P.M. acknowledges Fundação para a Ciência e a Tecnologia (FCT) for its financial support via LAETA (project https://doi.org/10.54499/UID/50022/2025, accessed on 1 February 2026).

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. Excluded Samples Documentation

Figure A1 documents examples of marble slab images excluded during data curation due to imaging artifacts.

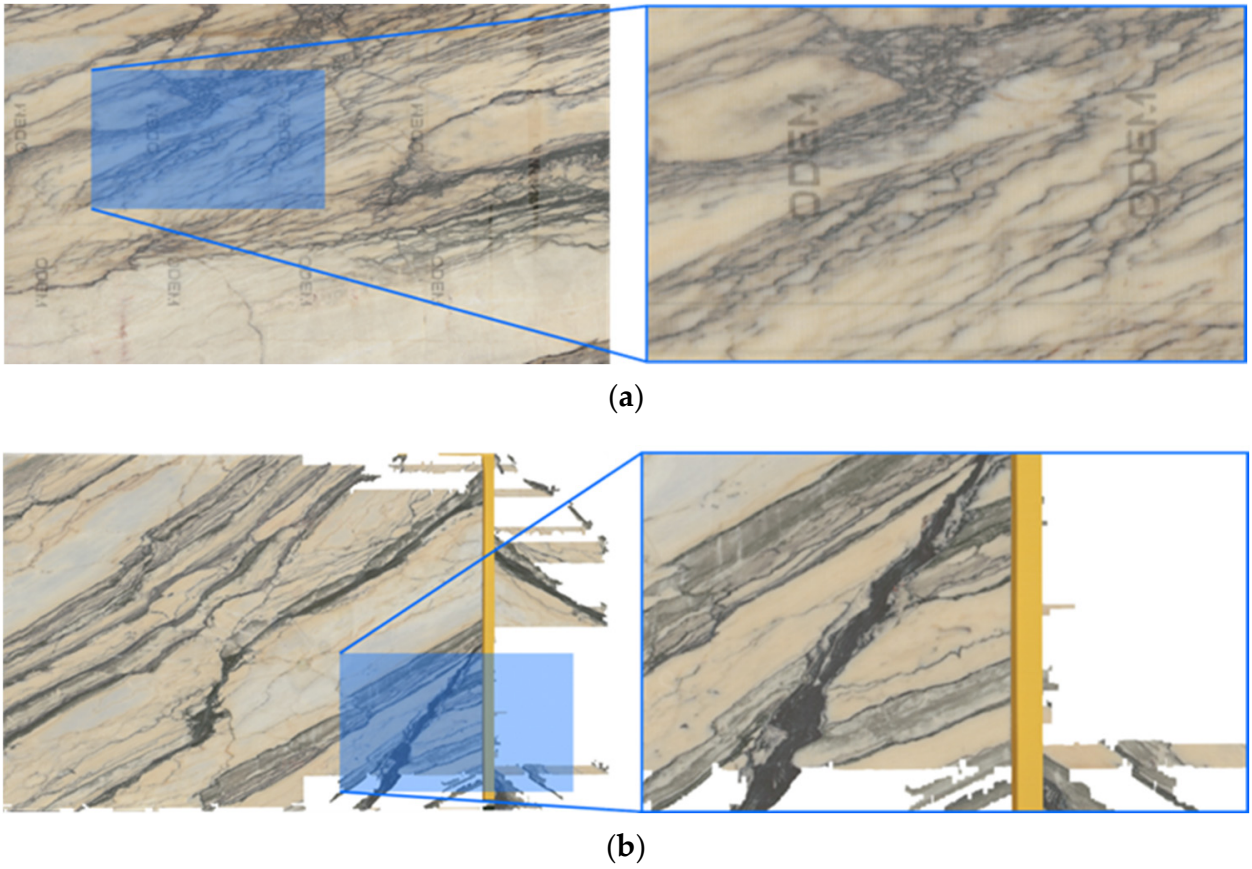


Figure A1. Examples of discarded marble slab images excluded during data curation. From an initial set of 327 scans, 38 samples (11.6%) were removed due to imaging artifacts that would compromise model training. (**a**) Slab with fiber-reinforced protective film: the magnified inset reveals a regular 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.

Appendix A.2. Unsupervised Segmentation Pipeline Details

Figure A2 provides a detailed visualization of the unsupervised segmentation pipeline, showing all intermediate processing stages from raw input to final binary mask.

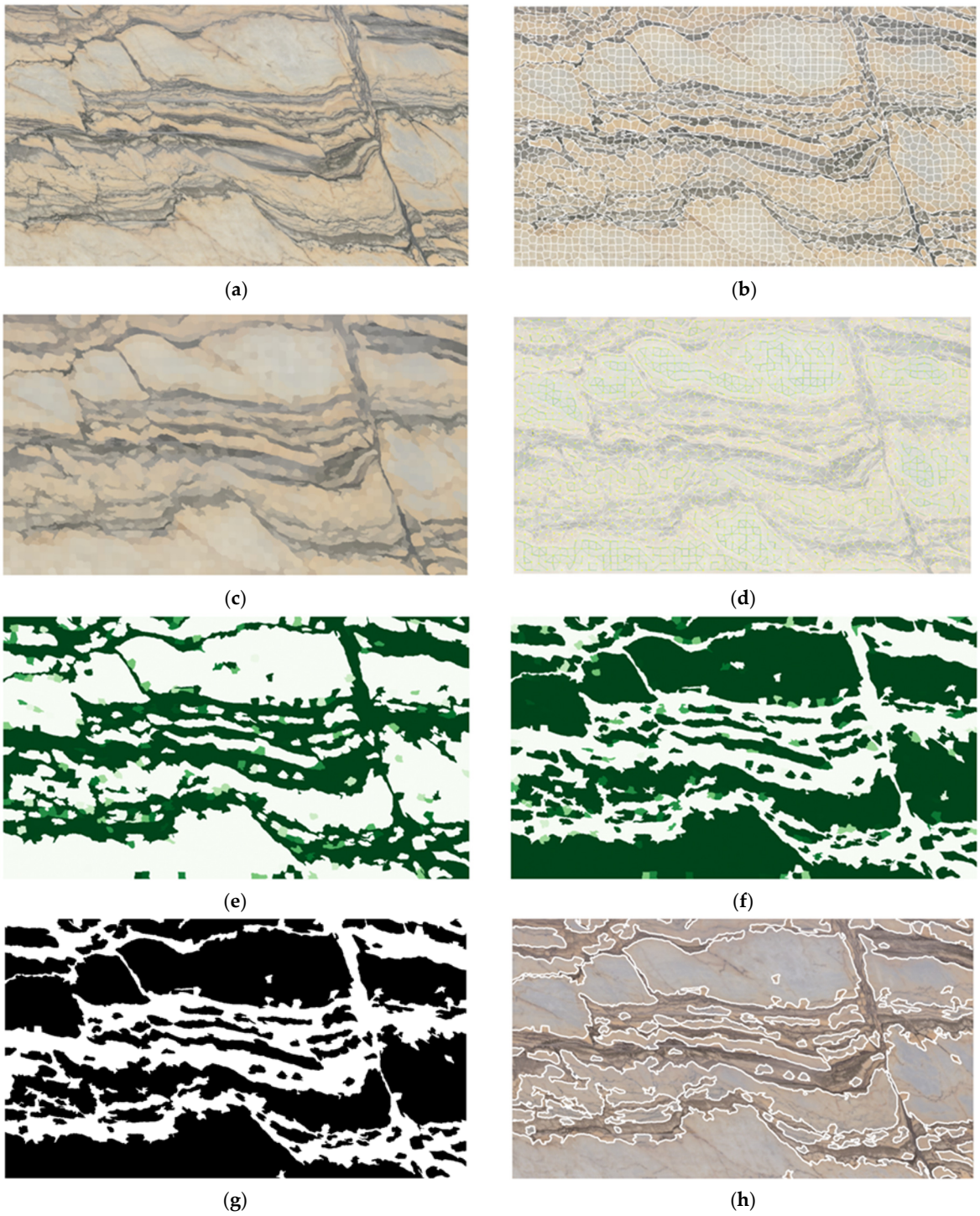
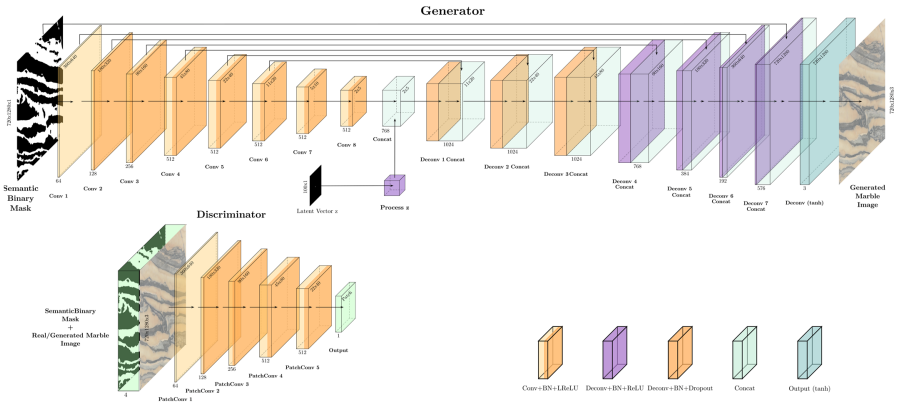


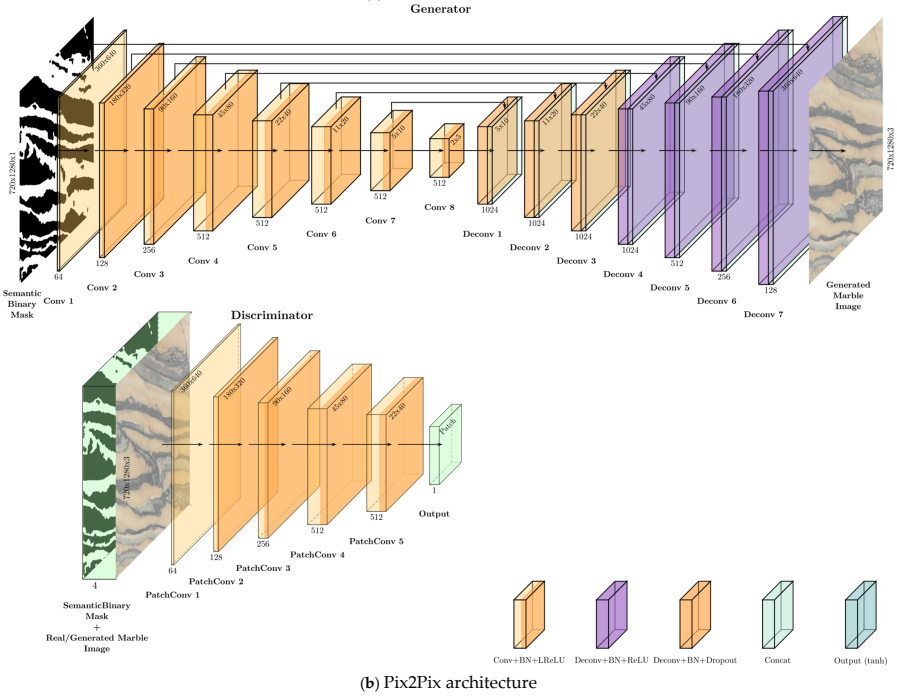
Figure A2. Detailed visualization of the unsupervised segmentation pipeline for a representative Exotic Ambar marble slab. (**a**) Original RGB input image at 1280 × 720 resolution; (**b**) SLIC superpixel tessellation showing approximately 3000 perceptually uniform regions with nominal size 20 px and 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**) heatmap of GMM-based unary costs for the "vein" class; (**g**) final binary mask after graph cut optimization with regularization weight $\lambda = 5.0$, overlaid on the original scan to demonstrate boundary preservation. (**h**) overlay of the final binary mask over the original RGB input image. The pipeline automatically extracts vein structures without manual annotation, achieving 100% acceptance rate across all 289 samples upon visual inspection.

Appendix A.3. Conditional GAN Architecture Diagrams

Figure A3 presents the detailed architecture diagrams for all four conditional GAN models evaluated in this study, illustrating the structural differences in their generator designs.

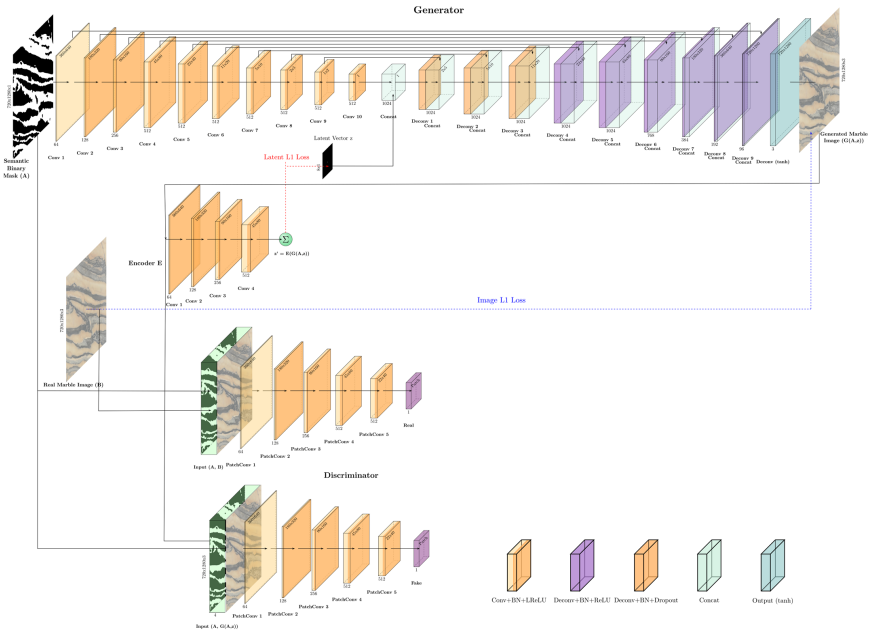


(**a**) cGAN architecture

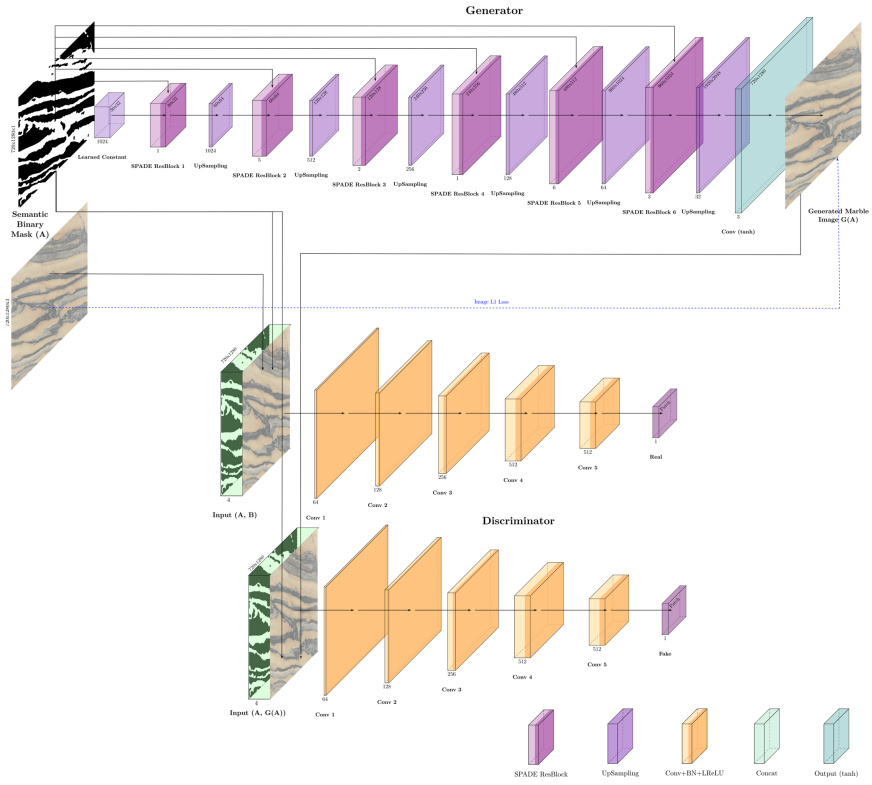


(**b**) Pix2Pix architecture

Figure A3. *Cont.*



(**c**) BicycleGAN architecture



(**d**) GauGAN architecture

Figure A3. 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 (λ = 100) enforcing pixel-level fidelity; (**c**) BicycleGAN: Extended architecture including a dedicated encoder network for bijective latent-to-image mapping with cycle consistency (λ_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 $\times$ 70 receptive fields to ensure fair comparison. Conv = Convolutional layer; BN = batch normalization; LReLU = Leaky ReLU; Deconv = transposed convolution.

Appendix A.4. Quantitative Metrics Distributions

Figure A4 shows the distribution of quantitative performance metrics across all validation samples, complementing the summary statistics in Table 2.

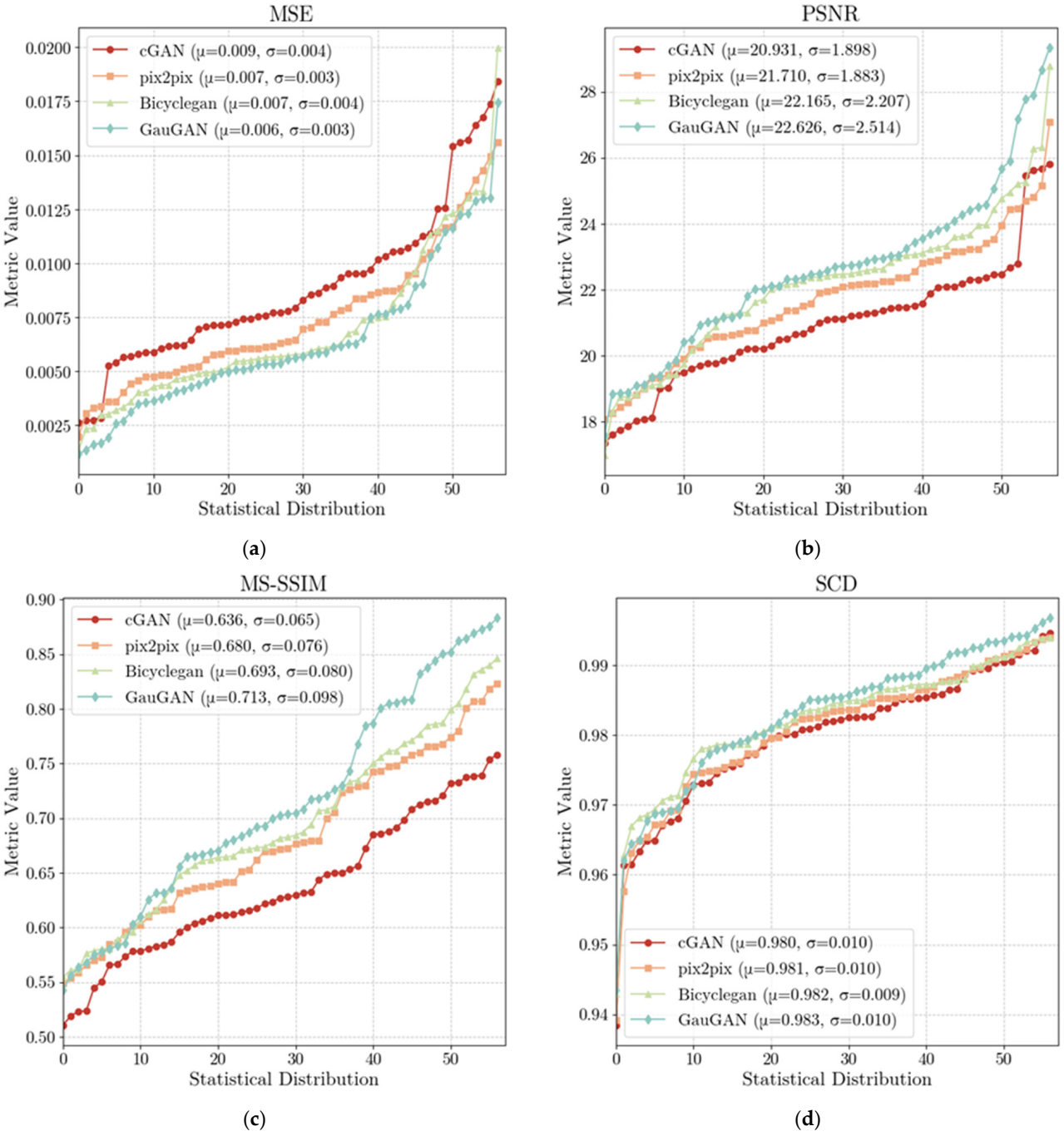


Figure A4. *Cont.*

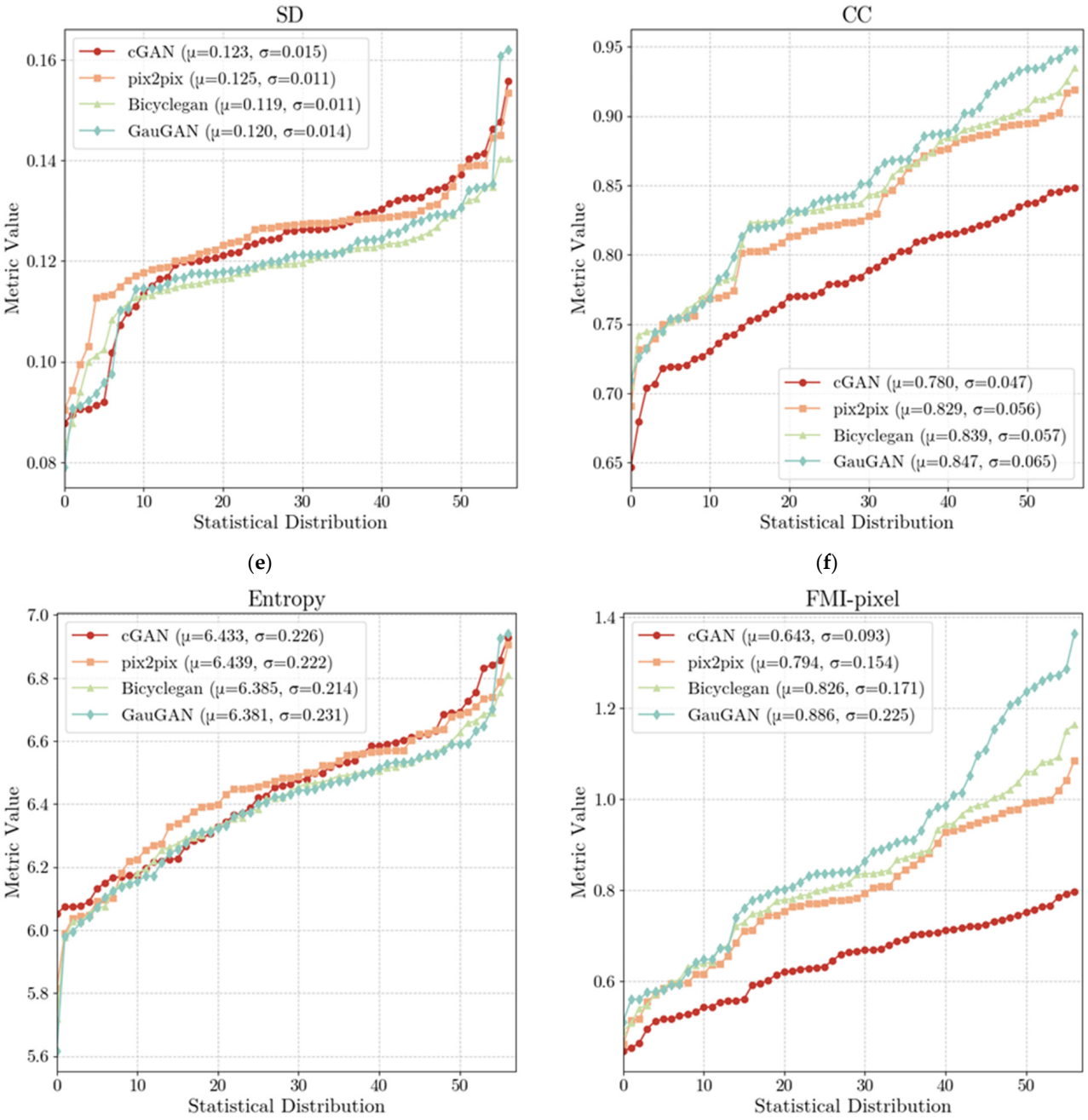


Figure A4. Distribution of quantitative performance metrics across the validation set (n = 57 samples) for all four GAN architectures. Each subplot shows sorted metric values, with the x-axis representing samples ordered by metric value and the y-axis showing the metric magnitude. Metrics shown: (**a**) 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).

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*, 103390. [CrossRef]
2. Loy, J.; Canning, S.; Little, C. Industrial Design Digital Technology. *Procedia Technol.* **2015**, *20*, 32–38. [CrossRef]
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. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 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. *Res. Sq.* 2024, *preprint*. [CrossRef]
5. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784. [CrossRef]
6. Park, T.; Liu, M.-Y.; Wang, T.-C.; Zhu, J.-Y. Semantic Image Synthesis with Spatially-Adaptive Normalization. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
7. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017*; IEEE: Piscataway, NJ, USA, 2017; 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* **2024**, arXiv:2312.04063v3.
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 31st International Conference on Neural Information Processing Systems NIPS'17, Long Beach, CA, USA, 4–9 December 2017*; ACM Digital Library: New York, NY, USA, 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)*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds.; NIPS Foundation: San Diego, CA, USA, 2016.
11. Gatys, L.A.; Ecker, A.S.; Bethge, M. A Neural Algorithm of Artistic Style. *J. Vis.* **2016**, *16*, 326. [CrossRef]
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. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2019; pp. 3449–3461.
13. Stein, G.; Cresswell, J.C.; Hosseinzadeh, R.; Sui, Y.; Leigh Ross, B.; Villecroze, 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. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 3732–3784.
14. Borji, A. Pros and Cons of GAN Evaluation Measures: New Developments. *Comput. Vis. Image Underst.* **2022**, *215*, 103329. [CrossRef]
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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 Ocotober 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 10982–10994.
16. Perlin, K. Improving Noise. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, San Antonio, TX, USA, 23–26 July 2002*; Association for Computing Machinery: New York, NY, USA, 2002; pp. 681–682.
17. Perlin, K. An Image Synthesizer. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques, San Francisco, CA, USA, 22–26 July 1985*; ACM Digital Library: New York, NY, USA, 1985; Volume 19, pp. 287–296.
18. Turk, G. Generating Textures on Arbitrary Surfaces Using Reaction-Diffusion. *Acm Siggraph Comput. Graph.* **1991**, *25*, 289–298. [CrossRef]
19. Worley, S. A Cellular Texture Basis Function. In *SIGGRAPH '96: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniq*; Association for Computing Machinery: New York, NY, USA, 1996; pp. 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, 20–27 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: New York, NY, USA, 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 Trans. Graph. (TOG)* **2003**, *22*, 277–286. [CrossRef]
23. Levina, E.; Bickel, P.J. Texture Synthesis and Nonparametric Resampling of Random Fields. *Ann. Stat.* **2006**, *34*, 1751–1773. [CrossRef]
24. Aguerrebere, C.; Gousseau, Y.; Tartavel, G. Exemplar-Based Texture Synthesis: The Efros-Leung Algorithm. *Image Process. Line* **2013**, *3*, 223–241. [CrossRef]

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. [CrossRef]
26. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 10674–10685. [CrossRef]
27. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661. [CrossRef]
28. Gatys, L.A.; Ecker, A.S.; Bethge, M. Texture Synthesis Using Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 262–270.
29. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2017; pp. 2242–2251.
30. 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. [CrossRef]
31. Zhang, L.; Rao, A.; Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE International Conference on Computer Vision*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023; pp. 3813–3824.
32. 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025*; IEEE: Piscataway, NJ, USA, 2025.
33. Borovec, J. Fully Automatic Segmentation of Stained Histological Cuts. In Proceedings of the Poster 2013: 17th International Student Conference on Electrical Engineering, Prague, Czech Republic, 16 May 2013.
34. Borovec, J.; Svihlík, J.; Kybic, J.; Habart, D. Supervised and Unsupervised Segmentation Using Superpixels, Model Estimation, and Graph Cut. *J. Electron. Imaging* **2017**, *26*, 061610. [CrossRef]
35. 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* **2022**, *11*, 60. [CrossRef]
36. 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. [CrossRef] [PubMed]
37. Giraud, R.; Clément, M. Superpixel Segmentation: A Long-Lasting Ill-Posed Problem. *arXiv* **2024**, arXiv:2411.06478.
38. 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, Munich, Germany, 8–14 November 2018*; Ferrari, V., Hebert, M., Eds.; ACM Digital Library: New York, NY, USA, 2018.
39. 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: Berlin/Heidelberg, Germany, 2017; Volme 723, pp. 767–779.
40. Boykov, Y.Y.; Jolly, M.-P. Interactive Graph Cuts for Optimal Boundary & Region Segmentation Of Objects in N-D Images. In *Proceedings of the 8th IEEE International Conference on Computer Vision ICCV, Vancouver, BC, Canada, 7–14 July 2001*; IEEE: Piscataway, NJ, USA, 2001; Volume I, pp. 105–112.
41. Shi, J.; Malik, J. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888–905. [CrossRef]
42. 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*, 103963. [CrossRef]
43. Jha, S.B.; Babiceanu, R.F. Deep CNN-Based Visual Defect Detection: Survey of Current Literature. *Comput. Ind.* **2023**, *148*, 103911. [CrossRef]
44. 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. [CrossRef]
45. 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*, 103128. [CrossRef]
46. Kumar, V.; Hernández, N.; Jensen, M.; Pal, R. Deep Learning Based System for Garment Visual Degradation Prediction for Longevity. *Comput. Ind.* **2023**, *144*, 103779. [CrossRef]
47. Hu, W.; Wang, T.; Chu, F. A Wasserstein Generative Digital Twin Model in Health Monitoring of Rotating Machines. *Comput. Ind.* **2023**, *145*, 103807. [CrossRef]
48. Kim, S.; Jang, H.; Yoon, B. Developing a Data-Driven Technology Roadmapping Method Using Generative Adversarial Network (GAN). *Comput. Ind.* **2023**, *145*, 103835. [CrossRef]

49. Johnson, J.; Alahi, A.; Li, F.-F. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
50. Wang, R. Research on Image Generation and Style Transfer Algorithm Based on Deep Learning. *Open J. Appl. Sci.* **2019**, *9*, 661–672. [CrossRef]
51. Bernardi, M. Generating Realistic Marble Textures Using Generative Adversarial Networks. Bachelor's Thesis, Università degli Studi di Padova, Padova, Italy, 2023.
52. Guo, Y.; Smith, C.; Hašan, M.; Sunkavalli, K.; Zhao, S. MaterialGAN: Reflectance Capture Using a Generative SVBRDF Model. *ACM Trans. Graph.* **2020**, *39*, 254. [CrossRef]
53. Wang, X.; Jiang, H.; Zeng, T.; Dong, Y. An Adaptive Fused Domain-Cycling Variational Generative Adversarial Network for Machine Fault Diagnosis under Data Scarcity. *Inf. Fusion* **2026**, *126*, 103616. [CrossRef]
54. Wang, X.; Jiang, H.; Mu, M.; Dong, Y. A Dynamic Collaborative Adversarial Domain Adaptation Network for Unsupervised Rotating Machinery Fault Diagnosis. *Reliab. Eng. Syst. Saf.* **2025**, *255*, 110662. [CrossRef]
55. Yan, J.; Cheng, Y.; Zhang, F.; Li, M.; Zhou, N.; Jin, B.; Wang, H.; Yang, H.; Zhang, W. Research on Multimodal Techniques for Arc Detection in Railway Systems with Limited Data. *Struct. Health Monit.* 2025, *online first*. [CrossRef]
56. Odena, A.; Dumoulin, V.; Olah, C. Deconvolution and Checkerboard Artifacts. *Distill* **2016**, *1*, e3. [CrossRef]
57. *ITU-T P.910*; Subjective Video Quality Assessment Methods for Multimedia Applications. International Telecommunication Union: Geneva, Switzerland, 2008.
58. *ITU-R BT.500-14*; Methodologies for the Subjective Assessment of the Quality of Television Images. International Telecommunication Union: Geneva, Switzerland, 2020.
59. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multi-Scale Structural Similarity for Image Quality Assessment. In Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 9–12 November 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. *Comput. Graph. Forum* **2022**, *41*, 591–611. [CrossRef]
61. Salih, M.E.; Zhang, X.; Ding, M. Two Modifications of Weight Calculation of the Non-Local Means Denoising Method. *Engineering* **2013**, *5*, 522–526. [CrossRef]
62. Chen, Y. 3D Texture Mapping for Rapid Manufacturing. *Comput.-Aided Des. Appl.* **2007**, *4*, 761–771. [CrossRef]
63. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 4015–4026.
64. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *Trans. Mach. Learn. Res.* **2024**, 1–31. Available online: https://openreview.net/forum?id=a68SUt6zFt (accessed on 5 April 2026).
65. Caron, M.; Bojanowski, P.; Joulin, A.; Douze, M. Deep Clustering for Unsupervised Learning of Visual Features. In *Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012*; ACM Digital Library: New York, NY, USA, 2019; pp. 132–149.
66. Tabassum, A.; Ziabari, A. Adapting Segment Anything Model (SAM) to Experimental Datasets via Fine-Tuning on GAN-Based Simulation: A Case Study in Additive Manufacturing. *arXiv* **2024**, arXiv:2412.11381.
67. Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward Multimodal Image-to-Image Translation. *arXiv* **2017**, arXiv:1711.11586. [CrossRef]
68. 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), Belfast, UK, 13–14 June 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–7.
69. 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, Astana, Kazakhstan, 15–17 October 2014*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2014.
70. Fabre-Thorpe, M. The Characteristics and Limits of Rapid Visual Categorization. *Front. Psychol.* **2011**, *2*, 242. [CrossRef]