**DOI: 10.37943/24MMIK3887**

**Marya Ryspayeva**
PhD candidate, Department of Software
marya.rys1@mail.ru, orcid.org/0000-0001-5055-4149
Akhmet Baitursynuly Kostanay Regional University, Kazakhstan

**Olga Salykova**
Candidate of Technical Sciences, Associate professor,
Department of Software
solga0603@mail.ru, orcid.org/0000-0002-8681-4552
Akhmet Baitursynuly Kostanay Regional University, Kazakhstan

# IMPACT OF LOSS FUNCTION ON SYNTHETIC BREAST ULTRASOUND IMAGE GENERATION

**Abstract:** The BUSI (Breast Ultrasound Images) dataset is small and imbalanced, which limits the effective training of deep learning diagnostic models. Generative Adversarial Networks (GANs) offer a promising and increasingly popular solution for synthesizing realistic medical images to augment scarce training data and improve overall model generalization. This study investigates the impact of loss function selection in our previously published Deep Generative Adversarial Network with Wasserstein Gradient Penalty and Transfer Learning (DGAN-WP-TL). Two configurations were evaluated: one trained using Wasserstein GAN with Gradient Penalty (WGAN-GP) and another trained using Binary Cross-Entropy (BCE) loss. The experiments were conducted on the BUSI dataset with perceptual loss weights $\lambda = 0.5$, $3.0$, $5.0$, $7.0$, and 10.0. Model performance was comprehensively assessed using Fréchet Inception Distance (FID), Kernel Inception Distance (KID), Learned Perceptual Image Patch Similarity (LPIPS), and Multi-Scale Structural Similarity Index (MS-SSIM). Results demonstrate that WGAN-GP consistently outperformed BCE across all $\lambda$ values, generating images with higher fidelity, improved realism, and greater visual diversity. The superiority was most pronounced for $\lambda = 3.0$ and $\lambda = 5.0$, where WGAN-GP achieved the lowest KID and FID and the most balanced diversity–fidelity trade-off. The best-performing DGAN-WP-TL configuration (WGAN-GP, $\lambda = 5.0$) achieved KID = 0.14, FID = 179.42, LPIPS (fake–fake) = 0.49, and MS-SSIM (fake–fake) = 0.18. These results highlight the crucial role of loss function design in medical image synthesis. Overall, the study confirms that WGAN-GP provides superior image realism and variability, making it the preferred choice for high-quality, clinically relevant synthetic data generation, while BCE remains a lightweight and practical alternative for constrained computational environments.

**Keywords:** BUSI dataset; DGAN-WP-TL; WGAN-GP; BCE loss; synthetic medical images; loss function analysis

## Introduction

Breast cancer is one of the world's most prevalent cancers and the leading cause of death in females [1], [2]. Screening for breast cancer with ultrasound imaging is established because the modality is safe, inexpensive, and universally available, particularly in healthcare-impoverished communities [2]. Despite the clinically useful breast ultrasound data being available, the data are often sparse and unevenly distributed, with the malignant cases underrepresented. The imbalance in this type of data negatively affects the training of deep networks, which require large and varied datasets to be effective in terms of diagnostics [3].


Received: 10.09.2025 Accepted: 20.10.2025 Published: 30.10.2025

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

Generative Adversarial Networks (GANs), first introduced by Goodfellow et al. [4], have established remarkable promise in the transcending of data scarcity via the creation of realistic synthetic images. GANs in medical imaging applications have been implemented in diverse modalities, including in magnetic resonance imaging (MRI) [5], computed tomography (CT) [6], histopathology [7], and ultrasound [8]. The synthetic data produced via GANs could be employed in mitigating the problem of class imbalance, augmenting the amount of training sets, and improving the generalizability of machine learning systems in medical applications [9].

The working of GANs is quite dependent on the choice of the loss function. The Binary Cross-Entropy (BCE) loss is employed in standard GANS, which evaluates the data distribution divergence of the synthesized and real data. Although widely used, the BCE loss is found to suffer from vanishing gradients and instability during training, often resulting in mode collapse as well as low diversity. The Wasserstein GAN (WGAN) with the application of the gradient penalty (WGAN-GP) was therefore introduced, claiming more stable optimization along with better-quality samples [10]. Existing medical images research holds the view that techniques based on WGAN are capable of yielding more anatomically consistent images, along with better capture of minority pathological classes.

Under breast ultrasound imaging, applications of the generative models were classification [11], [12], segmentation [13], [14], and image-to-image translation tasks [15]. Although the methodologies report progress, the experiments in most cases relied on architectural design or supervised scenarios, with limited efforts devoted to the role of loss functions in the quality of synthesized images. Especially, no systematic comparison in the breast ultrasound synthesis scenario is presented among the widely used BCE and WGAN-GP, despite their popularity.

The gap motivates the present study. The objective is the study of the influence of the selection of the loss function on the quality and the diversity of the synthesized breast ultrasound images. Of particular interest is the comparison of the behavior of the DGAN-WP-TL [16] when initialized from the training data with the BCE loss and the WGAN-GP. The experiments are conducted on the BUSI dataset across a range of $\lambda$ values (0.5, 3.0, 5.0, 7.0, 10.0), with evaluation using widely accepted metrics, including Fréchet Inception Distance (FID), Kernel Inception Distance (KID), Learned Perceptual Image Patch Similarity (LPIPS), Multi-Scale Structural Similarity Index (MS-SSIM) and Sliced Wasserstein Distance (SWD) with PCA and UMAP.

Overall, the principal contributions of this paper are threefold:

- We provide the first systematic comparison of BCE and WGAN-GP loss functions during the synthesis of breast ultrasound.
- We analyze the $\lambda$ scaling sensitivity of the generative ability under multiple evaluation metrics.
- We find that WGAN-GP is always more diverse and faithful, while the training of the BCE is unstable but comparable in some settings.

By detailing the function of loss function design, this work is beneficial in optimizing breast ultrasound imaging synthetic data generation pipelines as well as more broadly optimizing medical image analysis data augmentation techniques.

## Methods and Materials

### *Data Materials*

In this study, we focused exclusively on the Breast Ultrasound Images (BUSI) dataset [17]. The dataset consists of grayscale ultrasound scans collected from female patients in a clinical setting, with expert annotations provided by radiologists. Each image is categorized into one of two classes: benign or malignant.

The dataset contains 437 benign and 210 malignant tumor images. Lesions vary in size, shape, and background complexity, reflecting realistic diagnostic challenges. Importantly, the

malignant class is significantly underrepresented, which is a common limitation in breast cancer imaging datasets. Such underrepresentation usually makes supervised learning-based algorithms biased against majority classes and insensitive in the detection of malignant tumors.

***Methodology***

The paper compares the impact of loss function selection in the context of GAN-based breast ultrasound image synthesis. We employ an architecture of DGAN-WP-TL proposed in [16], which consists of an ImageNet-pretrained, frozen VGG-19 backbone as the perceptual feature extractor [16, 18, 19]. Two architectures are compared with one another: one is trained using Binary Cross-Entropy (BCE) loss, while the other is trained using Wasserstein GAN with Gradient Penalty (WGAN-GP). Both of the setups are compared using multiple perceptual loss weights ($\lambda = 0.5, 3.0, 5.0, 7.0, 10.0$) in order to analyze their impact on the perceived image's fidelity, diversity, and structure consistency.

***Generator Architecture***

The generator is designed so that the 100-dimensional latent noise vector, sampled from the standard Gaussian distribution, is converted into a synthetic breast ultrasound image of size 512×512 pixels. The network is founded on the progressive upsampling framework in which the spatial resolution is expanded step-wise through the sequence of transposed convolution layers.

The architecture begins with a single dense projection layer that restructures the input into an 8×8×512 feature map, providing the initial low-resolution representation. Next are five upsampling blocks that consist of transposed convolution (4×4 kernel size, stride 2, same padding), batch normalization, and LeakyReLU activation. The feature maps across the layers are upscaled sequentially into 16×16, 32×32, 64×64, 128×128, 256×256, and then finally back to 512×512.

The final layer outputs the transposed convolution, which changes the high-level feature representation into a one-channel grayscale image. The hyperbolic tangent (tanh) activation is used in order to scale the output values so the generated images end up in the correct domain for ultrasound data.

***Discriminator and Perceptual Backbone***

The discriminator is trained to recognize the realism of synthesized breast ultrasound images with the added advantage of perceptual supervision from a pretrained network. The input to the discriminator is the grayscale image of size 512×512, repeated across three channels so that it aligns with the ImageNet-pretrained VGG19 backbone. The VGG19 network is shortened at the *block2_pool* step so that mid-level semantic and textural characteristics, which are informative enough for the analysis of ultrasound images, are retained.

Simultaneously, the original input is processed in a custom convolutional pathway. The pathway is composed of multiple convolutional layers with increasingly deeper filters (128–512), separated with LeakyReLU activation in order to maintain non-linear representations of features. Two residual blocks are included to refine the learning structure and mitigate vanishing gradients. The two 3×3 convolutions in each residual block with skip connection are followed by LeakyReLU activation, allowing the network to learn both low-level texture information and high-order dependencies.

The features extracted from the frozen VGG19 branch are flattened and combined with the output of the custom pathway. Such a combination allows the discriminator to combine domain-agnostic perceptual embeddings with domain-related texture information. The com-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

bined features pass into the dense output layer, which outputs one scalar score that measures the input image's probability as real or fake.

***Training Protocol***

The models were trained for 500 epochs using the Adam optimizer with a learning rate of $1 \times 10^{-5}$, $\beta_1 = 0.5$, and a decay factor of 0.995 per epoch (minimum learning rate set to $1 \times 10^{-7}$). The generator was updated once for every five discriminator updates (n_critic = 5), following standard practice in adversarial training. The batch size was set to two due to GPU RAM constraints.

Two training configurations were compared that differed only in the adversarial loss function. The discriminator was optimized with the Wasserstein loss with gradient penalty (WGAN-GP) in the first condition. For the purpose of the study of the effect of the regularization strength, the individual models were trained with the values of the coefficients of the gradient penalties $\lambda$ as 0.5, 3.0, 5.0, 7.0, and 10.0. The discriminator was trained with the Binary Cross-Entropy (BCE) loss in the second configuration. The values of $\lambda$ in this scenario were employed only as the weights within the perceptual loss term because no gradient penalty was used in the training with BCE.

All experiments were conducted on a single NVIDIA GeForce RTX 3060 GPU with 12 GB of VRAM. The experiments were implemented in TensorFlow 2.11 with Python 3.8, using ImageNet-pretrained VGG19 weights as a fixed perceptual feature extractor.

***Evaluation Metrics***

To comprehensively evaluate the quality and diversity of the generated breast ultrasound images, we employed a set of complementary metrics. These measures capture both fidelity to real data and intra-class variability within synthetic samples, which are critical when generating underrepresented malignant cases in the BUSI dataset.

*Fréchet Inception Distance (FID).* FID quantifies the similarity between the distributions of real and generated images in a deep feature space. It is computed from the mean ($\mu$) and covariance ($\Sigma$) of feature embeddings [20], as shown in Eq. (1):

$$FID = \left\|\mu_r - \mu_g\right\|_2^2 + Tr\left(\sum_r + \sum_g - 2\left(\sum_r \sum_g\right)^{1/2}\right) \quad (1)$$

where $\mu_r$, $\mu_g$ are the means and $\sum_r$ ,$\sum_g$ are the covariances of real and generated features. $Tr$ denotes the trace of a matrix, i.e., the sum of its diagonal elements. In the FID formula, the trace operation reduces the covariance difference term to a scalar, making the overall distance computable as a single number.

Lower FID values indicate closer alignment to real data and improved fidelity.

*Kernel Inception Distance (KID).* KID measures the squared Maximum Mean Discrepancy (MMD) between real and generated features using a polynomial kernel [21], as defined in Eq. (2):

$$KID(x, y) = MMD^2(\varphi(x), \varphi(y)) \quad (2)$$

The expanded $MMD^2$ definition is provided in Eq. (3):

$$MMD^2(X, Y) = \frac{1}{m(m-1)} \sum_{i \neq j} k\left(x_i, x_j\right) + \frac{1}{n(n-1)} \sum_{i \neq j} k\left(y_i, y_j\right) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, x_y) \quad (3)$$

where $k(x_i, x_j)$ is a polynomial kernel and $\varphi(x)$, $\varphi(y)$ are feature embeddings of real and generated images.

Lower KID values indicate closer alignment with real data. Unlike FID, KID provides an unbiased estimator and is particularly reliable in small datasets such as BUSI.

DOI: 10.37943/24MMIK3887



*Learned Perceptual Image Patch Similarity (LPIPS).* LPIPS captures perceptual similarity by comparing distances between image patches in a pretrained feature space. Two modes of evaluation were applied [22]:

- Real–fake LPIPS: measures fidelity using perceptual similarity metric from real and synthesized samples. Lower values indicate better perceptual similarity.
- Fake–fake LPIPS: measures diversity by assessing variability among generated samples. Higher scores indicate greater intra-class variability and reduced risk of mode collapse.

The metric is formally defined in Eq. (4):

$$LPIPS(x,y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w_l \odot (f^l(x) - f^l(y)) \|^2 \quad (4)$$

where $f^l(x))$ are deep features at layer $l$, $w_l$ are learned weights, and $H_l$ $W_l$ are spatial dimensions.

*Multi-Scale Structural Similarity Index (MS-SSIM).* MS-SSIM evaluates structural similarity across multiple spatial resolutions by jointly considering luminance, contrast, and structure terms. Similar to LPIPS, two evaluation modes were applied [23]:

- Real–fake MS-SSIM: assesses structural fidelity between real and generated samples. Lower values suggest reduced structural consistency, while higher values indicate stronger alignment.
- Fake–fake MS-SSIM: evaluates structural diversity among generated images. Lower values reflect higher diversity, whereas higher values suggest mode collapse.

The metric is mathematically defined in Eq. (5):

$$MS-SSIM(x,y) = \prod_{j=1}^{M} \left[ l_j(x,y) \right]^{\alpha_j} \cdot \left[ c_j(x,y) \right]^{\beta_j} \cdot \left[ s_j(x,y) \right]^{\gamma_j} \quad (5)$$

where $l_j\,(x,y)$, $c_j\,(x,y)$, $s_j\,(x,y)$ are the luminance, contrast, and structure comparisons at scale $j$, while $\alpha_j$, $\beta_j$, $\gamma_j$ are their respective weights.

*Sliced Wasserstein Distance (SWD) with PCA and UMAP.* As a supplement to the distributional and perceptual measures, we applied the Sliced Wasserstein Distance (SWD) in the spaces that were extracted using Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) [24].

- PCA-SWD evaluates alignment in the context of a linear feature space, where distributions are aligned along the principal components. Though efficient, the PCA can under-estimate structured, non-linear artifacts [25].
- UMAP-SWD provides a non-linear mapping that preserves both local neighborhoods and globally topological structures, with the end product being a more realistic representation of the underlying manifolds of the data. This makes UMAP-SWD a better generative aligner in medical imaging [26].

The SWD is computed as defined in Eq. (6):

$$SWD(P,Q) = \frac{1}{K} \sum_{k=1}^{K} W_1(P \cdot \theta_k, Q \cdot \theta_k) \quad (6)$$

where $P$, $Q$ are real and generated distributions, $\theta_k$ are random projection directions, and $W_1$ is the 1-Wasserstein (Earth Mover's) distance between the one-dimensional projected distributions.

Lower SWD values indicate closer alignment between real and synthetic data distributions.

All the metrics were computed only on the malignant component of the BUSI dataset in order to achieve consistency in the class distribution. Testing was performed separately on the WGAN-GP loss and the BCE loss learned models across different perceptual loss weights ($\lambda = 0.5, 3.0, 5.0, 7.0, 10.0$). This setup enabled a detailed analysis of how the choice of adver-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

sarial loss function influences both the fidelity and diversity of synthetic breast ultrasound images.

**Results**

The role of the loss function during the synthesis of synthetic breast ultrasound images from the BUSI dataset is discussed in this paper. The discriminator and the generator in all experiments both employ the same structure with VGG19 as the invariant perceptual backbone. Two loss settings were compared:

- WGAN-GP loss with the gradient penalty, well-known for stabilizing adversarial training.
- BCE loss, the initial GAN formulation, that directly optimizes the discriminator classification objective.

The two networks shared the same hyperparameters, with $\lambda \in \{0.5, 3.0, 5.0, 7.0, 10.0\}$ controlling the perceptual loss weight. The relatively small size of the dataset meant that KID is the primary measure of generation quality, with FID, LPIPS (real–fake and fake–fake), and MS-SSIM (real–fake and fake–fake) as supplementary measures.

***Quantitative Results***

Given the relatively small dataset size, we emphasize KID as the primary metric, since it is an unbiased estimator and more reliable than FID in low-sample regimes.

Table 1 and Table 2 present the comparative results between WGAN-GP loss and BCE loss across different $\lambda$ values. The WGAN-GP yielded smaller KID and FID scores, which meant better convergence with the actual data distribution. Compared to this, the application of BCE loss produced more sharpened images with smaller LPIPS (real–fake), but also showed reduced diversity (higher MS-SSIM fake–fake), with tendencies towards mode collapse.

The top-performing WGAN-GP model at $\lambda = 5.0$ had the best trade-off of fidelity and diversity with the smallest KID and similar perceptual scores. The top-scoring BCE-based model was at $\lambda = 3.0$, although with fake–fake scores reflecting lower variability in the synthesized samples.

Table 1. Evaluation Metrics for λ = 0.5, 3.0, 5.0, 7.0, 10.0 at 500 epochs of WGAN-GP loss

| λ | KID↓ | FID↓ | LPIPS real-fake↓ | LPIPS fake-fake↑ | MS-SSIM real-fake↓ | MS-SSIM fake-fake↑ |
|---|---|---|---|---|---|---|
| 0.5 | 0.24±0.02 | 251.57 | 0.54±0.06 | 0.44±0.06 | 0.19±0.11 | 0.17±0.13 |
| 3.0 | 0.17±0.01 | 193.91 | 0.51±0.06 | 0.47±0.06 | 0.17±0.11 | 0.18±0.12 |
| **5.0** | **0.14±0.01** | **179.42** | **0.51±0.05** | **0.49±0.06** | **0.18±0.12** | **0.18±0.13** |
| 7.0 | 0.18±0.02 | 202.07 | 0.51±0.06 | 0.48±0.06 | 0.19±0.12 | 0.19±0.13 |
| 10.0 | 0.17±0.01 | 198.64 | 0.53±0.06 | 0.50±0.07 | 0.19±0.12 | 0.19±0.15 |

Table 2. Evaluation Metrics for λ = 0.5, 3.0, 5.0, 7.0, 10.0 at 500 epochs of BCE loss

| λ | KID↓ | FID↓ | LPIPS real-fake↓ | LPIPS fake-fake↑ | MS-SSIM real-fake↓ | MS-SSIM fake-fake↑ |
|---|---|---|---|---|---|---|
| 0.5 | 0.25±0.01 | 243.68 | 0.51±0.05 | 0.39±0.05 | 0.18±0.11 | 0.25±0.20 |
| **3.0** | **0.24±0.01** | **236.55** | **0.50±0.06** | **0.35±0.05** | **0.15±0.10** | **0.20±0.19** |
| 5.0 | 0.27±0.02 | 255.94 | 0.51±0.05 | 0.40±0.06 | 0.15±0.09 | 0.22±0.20 |
| 7.0 | 0.31±0.02 | 277.78 | 0.52±0.05 | 0.40±0.05 | 0.16±0.11 | 0.23±0.19 |
| 10.0 | 0.32±0.02 | 277.05 | 0.51±0.06 | 0.38±0.05 | 0.15±0.10 | 0.20±0.19 |

DOI: 10.37943/24MMIK3887



Fig. 1 offers an in-depth insight into the progression of KID and FID over the training iterations for both WGAN-GP and BCE loss setups.

Fig. 1(a) displays the KID trends, in which WGAN-GP consistently achieves smaller values within $\lambda$ configurations, indicating more stable convergence towards the genuine information distribution. The BCE styles get better as time passes, but always remain bigger, particularly at $\lambda = 0.5$ as well as $\lambda = 10.0$.

Fig. 1(b) compares FID dynamics, with WGAN-GP again outperforming BCE in terms of quicker convergence and significantly lower values at 500 epochs. The gap between WGAN-GP and BCE is most significant in mid-to-later training (300–500 epochs), reflecting the regularization effect of the gradient penalty.

***Comparison with StyleGAN2 and StyleGAN3 Baseline***

To further validate the effectiveness of our proposed method, we compared DGAN-WP-TL at $\lambda = 5.0$ and DGAN-BCE-TL at $\lambda = 3.0$ against state-of-the-art baselines StyleGAN2 and StyleGAN3 (Table 3) [16, 27, 28].

StyleGAN3 achieved the lowest KID (0.09) and FID (177.99), confirming its strength in distributional alignment. However, this performance came at the cost of reduced diversity. In particular, StyleGAN3 exhibited lower LPIPS fake–fake (0.46) and higher MS-SSIM fake–fake (0.23) compared to DGAN-WP-TL, indicating less variability among generated samples. By contrast, DGAN-WP-TL at $\lambda = 5.0$ maintained strong generative quality (KID = 0.14, FID = 179.42) while outperforming StyleGAN3 in perceptual similarity and structural diversity (LPIPS fake–fake = 0.49, MS-SSIM fake–fake = 0.18).

StyleGAN2, in turn, lagged behind all models across metrics, showing higher KID and FID along with extremely poor diversity (MS-SSIM fake–fake = 0.87). The DGAN-BCE-TL variant produced competitive LPIPS real–fake values (0.50), but its higher KID (0.24) and FID (236.55) confirmed weaker distributional alignment.

Overall, while StyleGAN3 remains a powerful baseline in terms of distributional fidelity, DGAN-WP-TL at $\lambda = 5.0$ provides a superior balance between realism and diversity, yielding more structurally varied and perceptually convincing synthetic ultrasound images.

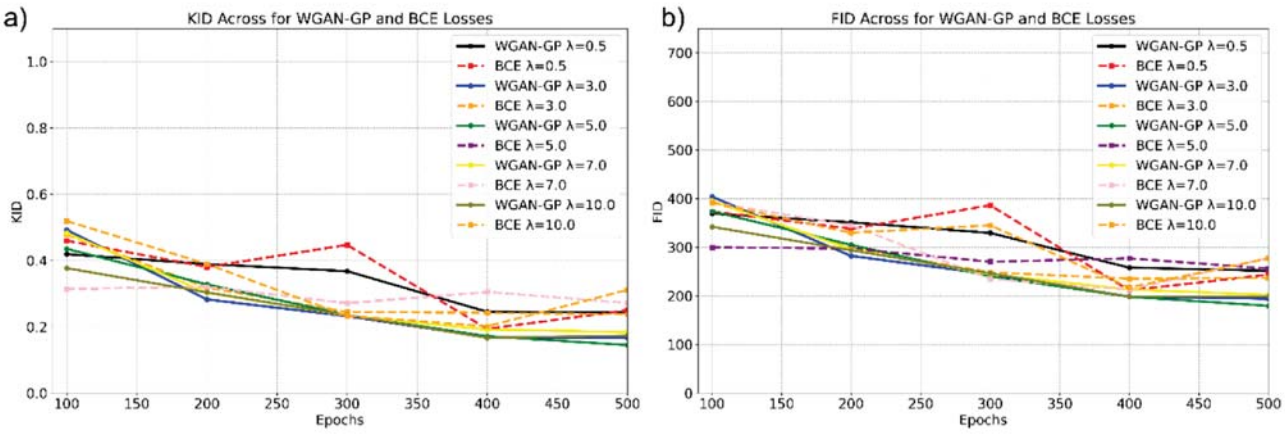


Figure 1. Evolution of KID and FID across training epochs for WGAN-GP and BCE losses at different $\lambda$ values on the BUSI dataset

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

Table 3. Comparative evaluation of DGAN-WP-TL, DGAN-BCE-TL, StyleGAN2, and StyleGAN3 on the BUSI dataset

| Model | KID↓ | FID↓ | LPIPS real-fake↓ | LPIPS fake-fake↑ | MS-SSIM real-fake↓ | MS-SSIM fake-fake↓ |
|---|---|---|---|---|---|---|
| DGAN-WP-TL at $\lambda$ =5.0 | **0.14±0.01** | **179.42** | **0.51±0.05** | **0.49±0.06** | **0.18±0.12** | **0.18±0.13** |
| DGAN-BCE-TL at $\lambda$ =3.0 | 0.24±0.01 | 236.55 | 0.50±0.06 | 0.35±0.05 | 0.15±0.10 | 0.20±0.19 |
| StyleGAN2 | 0.42±0.01 | 383.4863 | 0.78±0.05 | 0.10±0.09 | 0.21±0.07 | 0.87±0.11 |
| StyleGAN3 | **0.09±0.01** | **177.99** | 0.51±0.07 | 0.46±0.08 | 0.20±0.12 | 0.23±0.14 |

***Qualitative Results***

To complement the quantitative analysis, we inspected visual realism, anatomical plausibility, and distributional alignment of the generated images.

Real BUSI images are shown alongside samples from each model in Fig. 2. DGAN-WP-TL ($\lambda = 5.0$) produces images with realistic speckle patterns, lesion heterogeneity, and continuous tissue layers that closely resemble real scans. DGAN-BCE-TL ($\lambda = 3.0$) often looks sharp but exhibits reduced variety across samples (repeated textures and lesion shapes). StyleGAN2 tends to generate over-smoothed images with weak tumor boundaries and limited anatomical detail. StyleGAN3 consistently introduces artifacts—horizontal banding/striping, haloing near borders, and locally distorted textures—visible across nearly all examples. These artifacts explain why StyleGAN3 can achieve strong distributional scores while still being perceptually less reliable for clinical use.

DOI: 10.37943/24MMIK3887



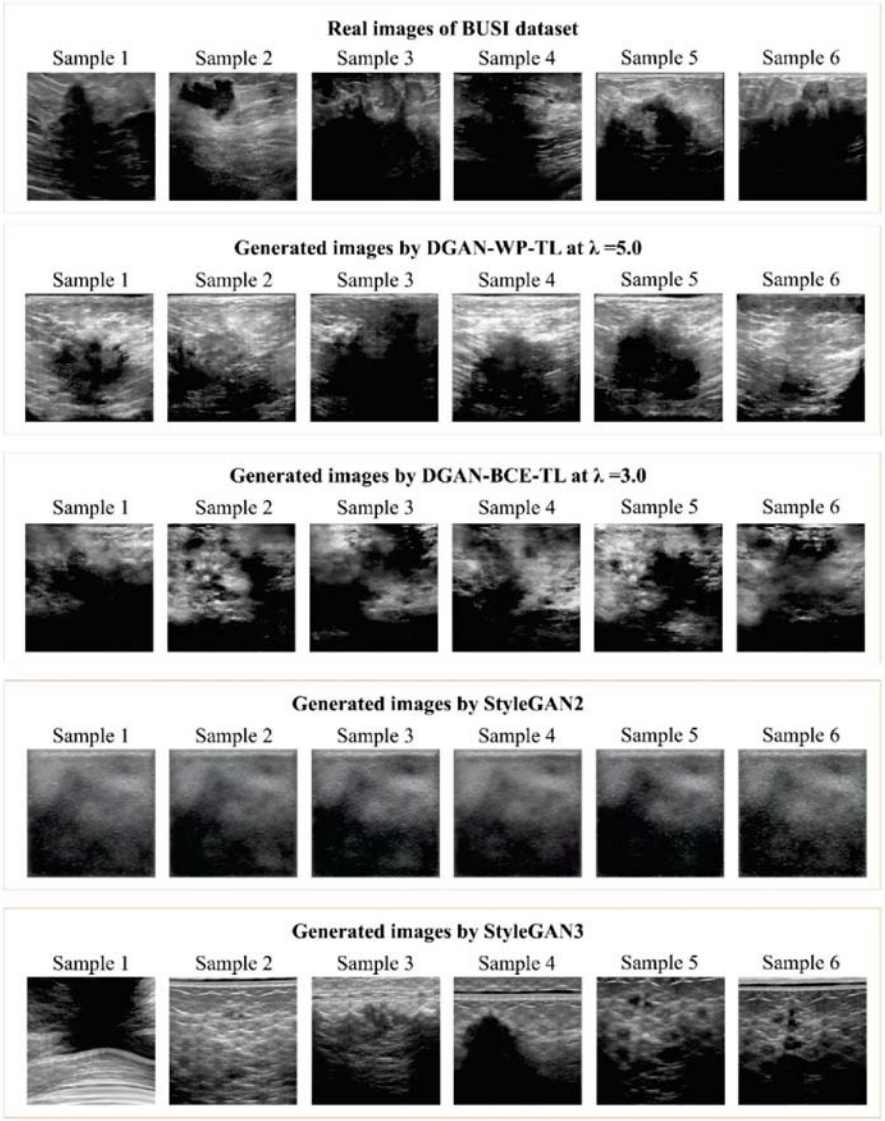


Figure 2. Real BUSI images compared to generated samples from DGAN-WP-TL ($\lambda$=5.0), DGAN-BCE-TL ($\lambda$=3.0), StyleGAN2, and StyleGAN3.

We project real and synthetic images into a feature space and visualize with PCA and UMAP (Fig. 3). DGAN-WP-TL shows the closest overlap with real data (Fig. 3a–b), indicating good coverage of the target manifold. DGAN-BCE-TL forms a tighter, more compact cluster (Fig. 3c–d), suggesting reduced diversity. StyleGAN2 clusters far from real data (Fig. 3e–f), reflecting poor alignment. StyleGAN3 partially overlaps but with an elongated, anisotropic spread (Fig. 3g–h), consistent with the artifacts seen in Fig. 2. The Sliced Wasserstein Distance (SWD) printed on the plots supports this: UMAP-SWD is lowest for DGAN-WP-TL ($\approx$1.09) vs. StyleGAN3 ($\approx$1.32), indicating better alignment on a non-linear embedding of the data manifold. PCA-SWD slightly favors StyleGAN3 ($\approx$3.30 vs. $\approx$5.58), but PCA's linearity can under-penalize structured artifacts.

The bar chart in Fig. 4 summarizes SWD across models for both PCA and UMAP. It confirms the embedding analysis: DGAN-WP-TL yields the best (lowest) UMAP-SWD among all models, with StyleGAN3 second, DGAN-BCE-TL next, and StyleGAN2 worst. Though at times favoring StyleGAN3, being distorted as a linear projection and often underestimating structured but non-linearly inseparable artifacts, PCA is defective. However, better reflecting the intrinsic manifold with preservation of local neighborhoods as well as large-scale topology, UMAP captures the manifolds more accurately. Thus, UMAP-SWD is the more reliable measure of the

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

type's generative alignment in this scenario. Considering qualitative artifacts (Fig. 2) together with the non-linear embedding advantage (Fig. 3 and 4), DGAN-WP-TL ($\lambda = 5.0$) offers the most credible trade-off—high perceptual realism and structural diversity without the systematic artifacts observed in StyleGAN3.

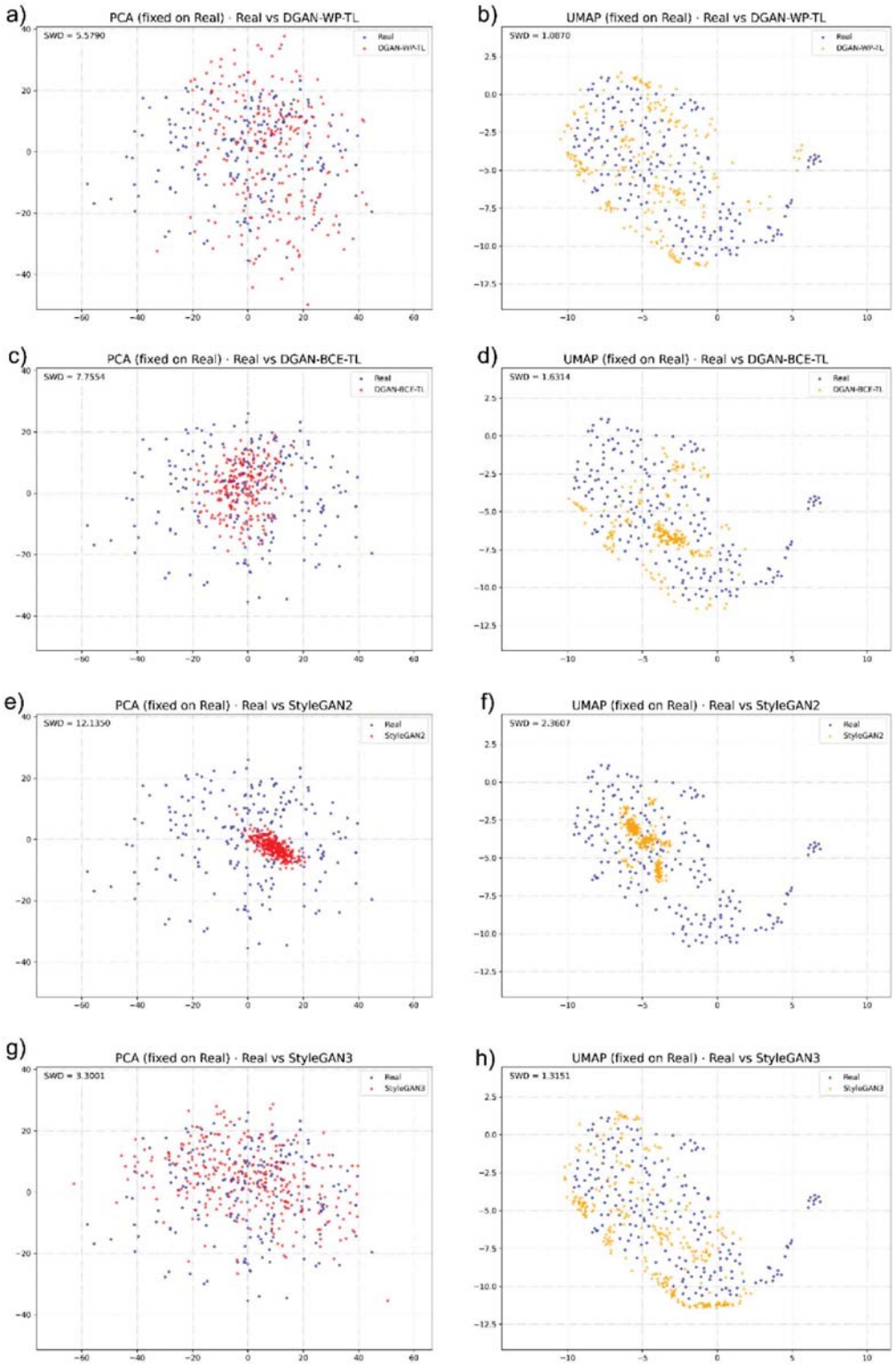


Figure 3. PCA and UMAP visualizations of real and synthetic image distributions across models. (a) PCA – Real vs. DGAN-WP-TL ($\lambda$=5.0); (b) UMAP – Real vs. DGAN-WP-TL ($\lambda$=5.0); (c) PCA – Real vs. DGAN-BCE-TL ($\lambda$=3.0); (d) UMAP – Real vs. DGAN-BCE-TL ($\lambda$=3.0); (e) PCA – Real vs. StyleGAN2; (f) UMAP – Real vs. StyleGAN2; (g) PCA – Real vs. StyleGAN3; (h) UMAP – Real vs. StyleGAN3.

DOI: 10.37943/24MMIK3887



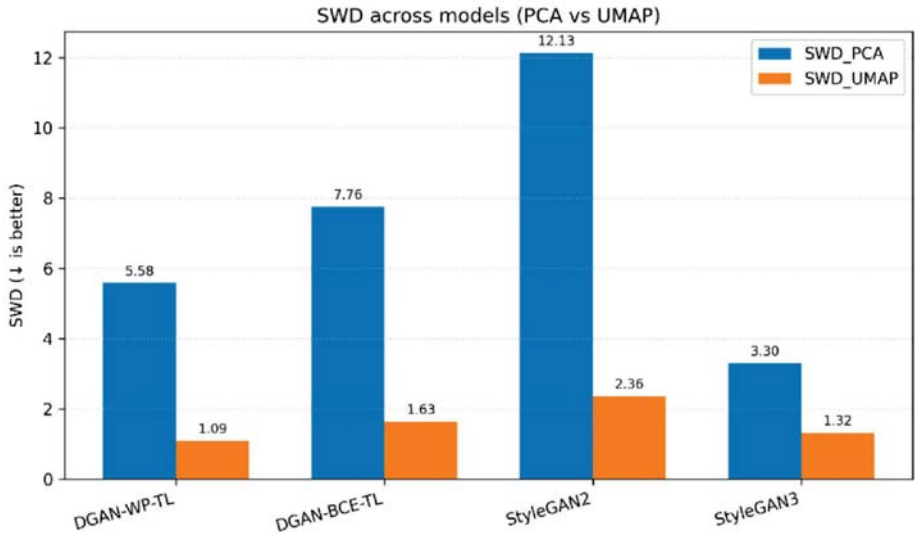


Figure 4. Bar plots of Sliced Wasserstein Distance (SWD) across models using PCA and UMAP embeddings.

**Discussion**

The impact of the choice of loss function on DGAN-WP-TL's generation of synthetic breast ultrasound images from the BUSI database were investigated, as described in the paper. Systematically comparing WGAN-GP and BCE loss functions and baselines against StyleGAN2 and StyleGAN3, we offer a distributional alignment-perceptual realism-structural diversity tradeoff study.

WGAN-GP loss, particularly for $\lambda = 5.0$, yields optimal fidelity and diversity tradeoff were observe. Whilst in certain instances, BCE loss produced images that looked sharper, higher resulting MS-SSIM fake–fake values demonstrated lower variability, in agreement with partial mode collapse. It then follows that BCE loss continues to be less suited for producing medical data, where diversity in lesion morphology is central to allowing for robust downstream diagnostic models.

Comparisons to StyleGAN baselines indicate a significant gap exists between numerical metric measurements of alignment and perceptual plausibility. StyleGAN3 had the lowest FID and KID, in agreement with its superior modelling of global distributional stats. Qualitative assessment presented, however, persistent structured artifacts, e.g., banding and local distortions, making it clinically unreliable. DGAN-WP-TL, in contrast, though with slightly higher FID/KID, had more anatomically plausible images and higher diversity-sensitive scores (MS-SSIM and LPIPS). These again prove that distributional measures may not capture the perceptual and diagnostic sufficiency of computer-generated medical images.

Embedding-based comparison, via PCA and UMAP, also supports this conclusion. While PCA-SWD tended to favour StyleGAN3 at times, linear PCA underrepresented structured artifacts. UMAP, which maintains non-linear manifolds and neighbourhood topology, on the other hand, disclosed DGAN-WP-TL as the closest to real BUSI data. Such a conclusion again highlights the importance of judging generative models not only via traditional scores but also via manifold-informed methods that are much better reflective of clinical plausibility.

More broadly, the present contribution again highlights the key role of loss function formulation in obtaining a realism-diversity balance in medical image generation. Higher-order gradient-penalized adversarial losses such as WGAN-GP provide stable optimization and higher distributional realism compared to BCE, in particular in the small and imbalanced-dataset

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

regime. Conversely, sole reliance on global distributional metrics (KID/FID) may overestimate a specific model's clinical utility, in favour of the use of perceptual and embedding-based assessments.

**Conclusion**

This work systematically examined how the choice of adversarial loss influences synthetic breast ultrasound image generation on the BUSI dataset. Using a fixed DGAN-WP-TL architecture with a VGG19 perceptual backbone, we compared WGAN-GP and BCE objectives across $\lambda \in \{0.5, 3.0, 5.0, 7.0, 10.0\}$ and evaluated realism and diversity with KID, FID, LPIPS (real–fake/fake–fake), MS-SSIM (real–fake/fake–fake), and manifold alignment (PCA/UMAP with SWD).

Across settings, WGAN-GP at $\lambda = 5.0$ delivered the most favourable trade-off: low KID/FID, high perceptual fidelity, and more substantial diversity (higher LPIPS f–f, lower MS-SSIM f–f) than BCE. Although StyleGAN3 achieved the highest KID/FID, qualitative studies and embedding analysis revealed continued structural artifacts and a comparably narrower spread, suggesting that distributional scores themselves may overestimate in-the-wild utility. BCE intermittently resulted in clear images but had lower variability and less stable training. Taken together, these findings show that WGAN-GP is the preferable objective for clinically credible BUSI synthesis, balancing fidelity with lesion-level diversity that is critical for downstream diagnostic robustness.

Our results have two practical implications. Firstly, loss design provides a primary knob for training stability and class/lesion diversity preservation in small-scale imbalanced medical data sets where architectural adjustments are not sufficient. Secondly, we suggest a distributional metric-level assessment alongside diversity-auditory and manifold-sensitive analyses, rather than relying solely on global scores, to prevent models with salient artifacts from being disproportionately favoured.

Limitations and future work. This study focused on a single dataset (BUSI), one perceptual backbone (VGG19), and offline metrics rather than reader studies. Future work will (i) validate on multi-center ultrasound cohorts and additional modalities, (ii) explore conditional and segmentation-guided synthesis to better control lesion attributes, (iii) quantify impact on downstream classifiers and detection models, and (iv) incorporate privacy-preserving training and clinical expert evaluation. These steps will further clarify how synthetic ultrasound can be safely and effectively integrated into real-world workflows.

## References

[1] Sechopoulos, I., Teuwen, J., & Mann, R. (2020). Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. Seminars in Cancer Biology, 72, 214–225. https://doi.org/10.1016/j.semcancer.2020.06.002

[2] Negi, A., Joseph Raj, A. N., Nersisson, R., Zhuang, Z., & Murugappan, P. (2020). RDA-UNet-WGAN: An accurate breast ultrasound lesion segmentation using Wasserstein generative adversarial networks. Arabian Journal for Science and Engineering, 45, 6909–6921. https://doi.org/10.1007/s13369-020-04480-z

[3] Ryspayeva, M. (2023). Generative adversarial network as data balance and augmentation tool in histopathology of breast cancer (pp. 99–104). Proceedings of the 2023 IEEE International Conference on Smart Information Systems and Technologies (SIST). https://doi.org/10.1109/SIST58284.2023.10223577

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. arXiv preprint arXiv:1406.2661. https://arxiv.org/abs/1406.2661

[5] Ryspayeva, M., & Salykova, O. (2025). Effect of data balancing methods on MRI Alzheimer's classification. Proceedings of the 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST) (pp. 1–7). IEEE. https://doi.org/10.1109/SIST61657.2025.11139255
[6] Motamed, S., Rogalla, P., & Khalvati, F. (2021). Data augmentation using generative adversarial networks (GANs) for GAN-based detection of pneumonia and COVID-19 in chest X-ray images. Informatics in Medicine Unlocked, 27, 100779. https://doi.org/10.1016/j.imu.2021.100779
[7] Ryspayeva, M. (2023). Generative adversarial network as data balance and augmentation tool in histopathology of breast cancer (pp. 99–104). Proceedings of the 2023 IEEE International Conference on Smart Information Systems and Technologies (SIST). IEEE. https://doi.org/10.1109/SIST58284.2023.10223577
[8] Haq, D. Z., & Fatichah, C. (2023). Ultrasound image synthetic generating using deep convolution generative adversarial network for breast cancer identification. IPTEK The Journal for Technology and Science, 34(1), 12–21. https://doi.org/10.12962/j20882033.v34i1.14968
[9] Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2019). Deep learning approaches for data augmentation and classification of breast masses using ultrasound images. International Journal of Advanced Computer Science and Applications, 10(5), 1–11. https://doi.org/10.14569/IJACSA.2019.0100579
[10] Gulrajani, I., Ahmed, F., Arjovsky, M., & Dumoulin, V. (2017). Improved training of Wasserstein GANs. arXiv preprint arXiv:1704.00028. https://doi.org/10.48550/arXiv.1704.00028
[11] Liu, Z., Lv, Q., Lee, C., & Shen, L. (2023). GSDA: Generative adversarial network-based semi-supervised data augmentation for ultrasound image classification. Heliyon, 9(8), e19585. https://doi.org/10.1016/j.heliyon.2023.e19585
[12] You, G., Qin, Y., Zhao, C., Zhao, Y., Zhu, K., Yang, X., & Li, Y. (2023). A CGAN-based tumor segmentation method for breast ultrasound images. Physics in Medicine & Biology, 68(7), 075010. https://doi.org/10.1088/1361-6560/acdbb4
[13] Han, L., Huang, Y., Dou, H., Wang, S., Ahamad, S., Luo, H., Liu, Q., Fan, J., & Zhang, J. (2020). Semi-supervised segmentation of lesion from breast ultrasound images with attentional generative adversarial network. Computer Methods and Programs in Biomedicine, 189, 105275. https://doi.org/10.1016/j.cmpb.2019.105275
[14] Xing, J., Li, Z., Wang, B., Qi, Y., Yu, B., Ghazvinian Zanjani, F., Zheng, A., Duits, R., & Tan, T. (2020). Lesion segmentation in ultrasound using semi-pixel-wise cycle generative adversarial nets. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18(3), 940–949. https://doi.org/10.1109/TCBB.2020.297847
[15] Barkat, L., Freiman, M., & Azhari, H. (2023). Image translation of breast ultrasound to pseudo anatomical display by CycleGAN. Bioengineering, 10(3), 388. https://doi.org/10.3390/bioengineering10030388
[16] Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2019). Dataset of breast ultrasound images. Data in Brief, 28, 104863. https://doi.org/10.1016/j.dib.2019.104863
[17] Ryspayeva, M., & Salykova, O. (2025). Multi-domain synthetic medical image generation and dataset balancing with DGAN-WP-TL. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 13(1). https://doi.org/10.1080/21681163.2025.2556687
[18] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y
[19] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://arxiv.org/abs/1409.1556
[20] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems, 30, 6626–6637. https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

[21] Bińkowski, M., Sutherland, D. J., Arbel, M., & Gretton, A. (2018). Demystifying MMD GANs. International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=r1lUOzWCW
[22] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 586–595. https://doi.org/10.1109/CVPR.2018.00068
[23] Wang, Z., Simoncelli, E. P., & Bovik, A.C. (2003). Multiscale structural similarity for image quality assessment. Proceedings of the 37th Asilomar Conference on Signals, Systems & Computers, 2, 1398–1402. IEEE. https://doi.org/10.1109/ACSSC.2003.1292216
[24] Deshpande, I., Zhang, Z., Schwing, A. G., & Forsyth, D. (2018). Generative modeling using the sliced Wasserstein distance. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3483–3491. https://doi.org/10.1109/CVPR.2018.00366
[25] Jolliffe, I.T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202
[26] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. https://arxiv.org/abs/1802.03426
[27] Karras, T., Laine, S., & Aila, T. (2020). A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(12), 4217–4228. https://doi.org/10.1109/TPAMI.2020.2970919
[28] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. Advances in Neural Information Processing Systems, 34, 852–863. https://proceedings.neurips.cc/paper/2021/hash/076ccd93ad68be51f23707988e934906-Abstract.html