Fréchet Inception Distance is Failing to Preserve Rank Consistency for Synthetic Out-of-Distribution Samples

Linghui Liu[1] Henrike Stephani[1] Janis Keuper[2,3]
[1]Fraunhofer ITWM, [2]Offenburg University, [3]Mannheim University
linghui.liu@itwm.fraunhofer.de henrike.stephani@itwm.fraunhofer.de janis.keuper@hs-offenburg.de

Abstract

Fréchet Inception Distance (FID) is widely used to evaluate how well synthetic images match a real target distribution. However, FID is known to depend strongly on the ImageNet-pretrained Inception representation, and it remains unclear whether its similarity rankings remain reliable when data move away from the ImageNet distribution. We study this question through controlled natural-image shifts, non-natural image datasets, and synthetic-to-real comparisons. Across these settings, we find that FID-based rankings become less reliable outside the ImageNet domain, and that this instability is driven mainly by the evaluation feature space. Our controlled shift experiments show that ranking consistency is largely preserved under some augmentations like photometric perturbations, but breaks down under others including Gaussian blur and noise. This instability is mainly caused by the mismatch between the evaluated data and the ImageNet-pretrained feature space. With domain-adapted representations, rankings become much more consistent across backbones and discrepancy metrics, and align better with the underlying data structure. Preliminary synthetic-to-real results further suggest that synthetic data evaluation can benefit from target-domain representations and complementary metrics.

1. Introduction

Synthetic data is increasingly used as a scalable substitute or complement to human-labeled training data in computer vision [30]. It enables controllable generation of rare events, long-tail scenarios, and privacy-sensitive content, but it also introduces a domain gap between synthetic and real data. As a result, reliable evaluation is central: it allows to quantify the distribution mismatch between real and synthetic data and to quantitatively compare different synthetic datasets or generation pipelines.

Fréchet Inception Distance (FID) [18] is among the most widely used metrics for this purpose. It provides a single scalar that is typically interpreted as "lower is better", and it is often reported as the primary quantitative indicator of synthetic–real similarity. However, FID is defined in the feature space of an ImageNet-pretrained Inception network. This choice is convenient, but it is not guaranteed to be appropriate when the data depart from ImageNet-like semantics and statistics, as in medical or scientific imaging. Prior work [5, 22, 36] has shown that FID can be sensitive to perturbations and may disagree with perceptual judgments, but these analyses mostly examine changes in the absolute score. In contrast, the stability of the relative ordering induced by FID, especially under distribution shift and in non-natural image domains, remains underexplored.

In this work, we study FID through the lens of ranking stability under controlled covariate shifts and representation choices. Concretely, we investigate the following three research questions:

Question 1: When investigated data distributions depart from the ImageNet pretraining distribution, does FID preserve a reliable ranking of dataset similarity?
Question 2: How strongly does this behavior depend on the choice of feature backbone that defines the embedding space?
Question 3: Can fine-tuning the feature extractor on shifted or target-domain data mitigate ranking instability and recover more domain-consistent orderings?

To systematically answer these questions, we make three contributions. First, we provide the first in-depth empirical study of FID ranking stability under controlled covariate shifts and in non-natural image domains, rather than only on absolute score changes. Second, we isolate the role of representation choice by comparing pretrained and domain-adapted embeddings across multiple backbone architectures, and we validate that the observed trends are consistent across complementary distribution metrics computed on the same embeddings. Third, we show that domain-aware fine-tuning can recover stable and domain-consistent rankings,

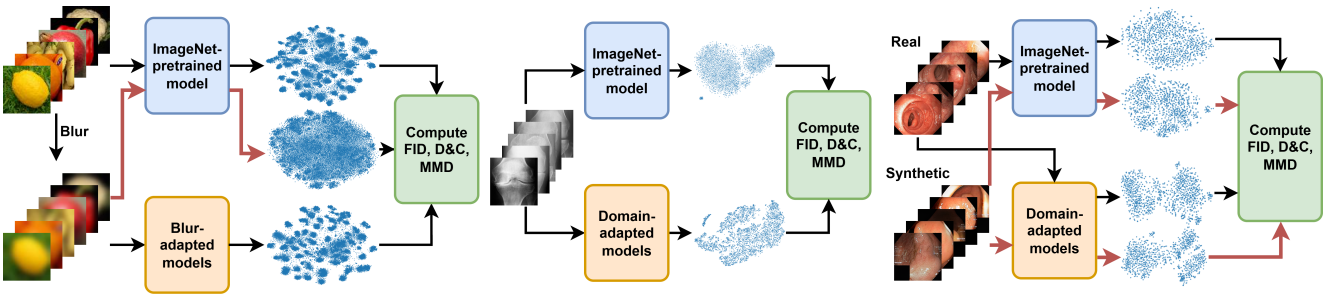


Figure 1. **Overview of the three experimental settings used to study representation effects on FID-style similarity rankings. Left**: For natural images, we first obtain a clean reference ranking in the ImageNet-pretrained feature space. After induced distribution shifts (like Gaussian blur), we compare rankings obtained in the same feature space with those from blur-adapted models to examine whether the clean reference structure is preserved. **Middle**: For non-natural images, we compare ImageNet-pretrained and domain-adapted feature spaces on medical data using FID, D&C, and MMD. **Right**: For synthetic images, we use the same metrics to compare synthetic samples against real data in pretrained and domain-adapted feature spaces.

and we derive practical guidance for evaluating synthetic data beyond natural-image settings.

Fig. 1 summarizes the experimental design of this work. In the natural-image setting, we analyze how controlled distribution shifts influence class-similarity rankings and whether feature adaptation can restore the clean reference rankings. In the non-natural-image setting, we compare ImageNet-pretrained and domain-adapted representations on medical datasets using distribution-based distances to assess whether the resulting inter-class relationships better align with domain semantics. In the synthethic-image setting, we compare distribution discrepancies between synthetic and real data in both ImageNet-pretrained and domain-adapted feature spaces.

2. Related Work

Synthetic Training Data. Synthetic data has become a practical source of training data in non-natural imaging domains. Recent surveys [26, 35] summarize its value for mitigating data scarcity, class imbalance, high annotation cost, and privacy constraints. It has been used for training in industrial surface defect inspection [40], in X-ray image analysis [14, 25], and in pathology tasks such as nuclei segmentation [12]. However, a key challenge in these settings is the synthetic-to-real domain gap, and many works therefore quantify distributional similarity between synthetic and real image sets using metrics such as FID [12, 25, 26].

Measuring Distribution Similarity with FID. Heusel et al.[18] introduced the FID together with the Two Time-Scale Update Rule training scheme. FID evaluates distribution mismatch in the Inception-V3[46] feature space by comparing the mean and covariance of real and generated embeddings, which corresponds to the Fréchet distance [13] under a multivariate Gaussian assumption. Because it provides a computationally efficient means for comparisons across models, FID rapidly became one of the most widely used metrics for evaluating image generative models. It is routinely reported in influential GAN and diffusion-model papers, including BigGAN[7], StyleGAN2[24], and DDPM[19].

Known Limitations of FID. Despite its popularity, a growing body of work has identified limitations of FID. These limitations can be grouped into issues in how FID is estimated in practice, and issues stemming from the particular feature space used to compute it.

Chong and Forsyth [10] show that the standard finite-sample estimator FID_N is biased: with limited samples, the expected score can systematically deviate from the population value, which can make comparisons across models unreliable. Parmar *et al.* [36] demonstrate that low-level pipeline differences, such as resizing, quantization, and compression, can noticeably change FID, even though these operations are not meant to alter the semantic content of images. Jayasumana *et al.* [20] argue that the commonly used closed-form FID formula is only strictly justified under a multivariate Gaussian assumption for both feature distributions, yet typical Inception embeddings can violate this assumption, making the standard computation theoretically mismatched to the intended distance. They also propose *CMMD*, which compares real and generated distributions by computing *Maximum Mean Discrepancy* (MMD) [15] distance on CLIP embeddings [39] using a Gaussian RBF kernel. Betzalel *et al.*[5] further show that, although FID captures coarse performance trends but fluctuates more near convergence, making fine-grained comparisons unreliable.

A recurring theme is that FID is strongly shaped by the chosen embedding network (usually ImageNet-pretrained Inception-V3), so it may reflect representation-specific biases rather than perceptual realism. Kynkäänniemi *et al.*

[27] show that standard FID is closely tied to ImageNet [11] classification structure: simply aligning Top- $\mathcal{N}$ predicted-class statistics between real and generated sets can substantially reduce FID without a corresponding visual improvement. They further show this effect weakens when the feature space is changed, highlighting backbone dependence. Jung and Keuper [22] report that FID inherits robustness biases from Inception-V3, being relatively tolerant to photometric changes but more sensitive to edge and texture corruptions. Therefore, FID-based rankings can favor models that reproduce high-frequency structure over color statistics. They also report that directly optimizing for FID can lower the metric without guaranteeing better-looking samples, indicating misalignment when FID is treated as a training objective. Stein *et al.* [43] provide large-scale human-subject evidence that diffusion models can be judged most realistic by people while being downranked by FID, and they argue that replacing Inception features with self-supervised features (e.g., DINOv2 [33]) yields evaluations that better track perceptual realism.

Jiralerspong *et al.* [21] show that FID deteriorate sharply under nearly imperceptible image transformations (e.g., light blur, mild posterization, high-quality JPEG), even when FID is computed in a DINOv2 feature space. This suggests that absolute FID values can be unstable to subtle nuisance factors, so conclusions based purely on score differences should be treated with care.

Building on these observations, our experiments study ranking stability rather than treating absolute FID values as the main target. Under controlled covariate shift induced by augmentation, we test whether the class-level FID ordering that matches human similarity judgments remains stable, and whether fine-tuning the feature extractor on the shifted data can recover the original ordering. We further compare multiple backbones to quantify how representation choice affects ranking stability. Beyond FID, we include complementary metrics computed in the same representation space. We report *density and coverage* (D&C) [32] and use their harmonic mean as a single summary score to make the resulting rankings directly comparable to FID. We also compute MMD [20] in the same embeddings as our FID variants to keep comparisons across augmentations and backbones controlled.

Domain Specific Investigations of FID. FID is also widely used outside natural images, for instance in histopathology generation [37], material microscopy modality transfer [3], and remote-sensing image synthesis [49]. However, in these domains the gap to ImageNet-style semantics can make "low FID" less informative. In echocardiography, FID may fail to capture subtle but clinically meaningful structural differences [1]. For 3D volumetric data, the standard Inception-V3 backbone is trained on 2D ImageNet and cannot directly evaluate volumes, so evaluation is often performed on 2D slices instead [47].

To improve domain relevance, many works in medical imaging replace the ImageNet pretrained Inception feature space with domain-specific encoders, such as RadImageNet[31] for 2D data [34], MedicalNet[9] for 3D data [44], or M3D-CLIP [4] for a pathology-aware medical FID [45]. However, medical imaging does not have a single universally accepted feature extractor for "clinical similarity", and different choices can change rankings. Woodland *et al.* [48] show that rankings from medical-trained extractors can be volatile and may disagree with expert judgment. Motivated by these domain-specific uncertainties, we use non-natural datasets to test whether adapting the feature extractor yields inter-class distance rankings that better reflect the target domain.

3. Methods

This section introduces the three distribution-comparison metrics used in our experiments: FID, D&C, and MMD. For fair comparison, all three metrics are computed in the same embedding space defined by the corresponding feature extractor.

3.1. FID

FID measures the discrepancy between real and generated image distributions in the feature space. Concretely, images are embedded by a pretrained Inception-V3 network, and the resulting penultimate layer features for real and generated sets are approximated as multivariate Gaussians $\mathcal{N}(\mu_r, \Sigma_r)$ and $\mathcal{N}(\mu_g, \Sigma_g)$. The FID is then computed as the Fréchet distance which is equivalent to the closed-form 2-Wasserstein distance for Gaussians[13]:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2\,(\Sigma_r\Sigma_g)^{\frac{1}{2}}\Big)\,. \quad (1)$$

3.2. D&C

D&C [32] measures how well one distribution matches the density and support coverage of another. Let $\{x_i\}_{i=1}^N$ and $\{y_j\}_{j=1}^M$ denote the embedded features of real and generated images, respectively. Define $B(x, r) = \{z \mid \|z - x\|_2 \le r\}$ and let $\mathrm{NND}_k(x_i)$ be the distance from x_i to its k-th nearest neighbour among the real features. Density counts how many real neighbourhoods contain each generated sample, reflecting how much generated samples fall into dense regions of the real data manifold. While coverage measures the fraction of real samples whose neighbourhood is hit by at least one generated sample, reflecting how well the gen-

erator covers the real support:

$$\text{Density} := \frac{1}{kM} \sum_{j=1}^{M} \sum_{i=1}^{N} \mathbb{I}[y_j \in B(x_i, \text{NND}_k(x_i))] \,, \tag{2}$$

$$\text{Coverage} := \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\exists\, j \text{ s.t. } y_j \in B(x_i, \text{NND}_k(x_i))] \,. \tag{3}$$

Note that density is not necessarily upper-bounded by 1. For direct comparability with single-score metrics, we summarize D&C by the harmonic mean

$$H(D, C) := \frac{2\,\tilde{D}\,C}{\tilde{D} + C}, \tag{4}$$

where we clip density as $\tilde{D} = \min(1, D)$ to keep the score bounded.

3.3. MMD

MMD is a kernel-based measure of the distribution discrepancy that compares distribution without assuming Gaussian statistics. We adopt the unbiased estimator with a Gaussian RBF kernel $k(x, y) = \exp\big(-\|x - y\|_2^2/(2\sigma^2)\big)$, where $\sigma = 10$ [20]:

$$\begin{aligned} \hat{\text{dist}}^2_{\text{MMD}}(X, Y) = & \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} k(x_i, x_j) \\ & + \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} k(y_i, y_j) \\ & - \frac{2}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} k(x_i, y_j), \end{aligned} \tag{5}$$

We report the MMD score scaled by a factor of 1000, consistent with [20].

Lower FID and MMD values indicate a closer match between the real and generated distributions in the chosen feature space, whereas a higher harmonic mean of density and coverage indicates better agreement in terms of both fidelity and support coverage.

4. Experiments

This section presents three sets of experiments designed to examine how distribution-similarity metrics behave under changes in the feature representation. First, on natural images, we introduce controlled distribution shifts to test whether similarity rankings remain stable and whether feature adaptation can recover them. Second, on non-natural image datasets, we evaluate whether adapting the feature extractor leads to inter-class relationships that better reflect domain semantics. Third, in a synthetic-to-real setting, we compare rankings of synthetic image sets across metrics and feature spaces to assess how reliably these rankings transfer after adaptation.

4.1. Experiment Setup

We evaluate distribution similarity in feature spaces using four different CNN backbones and their penultimate-layer embeddings. The backbones are a pretrained Inception V3, a fine-tuned Inception V3, a fine-tuned ResNet50 [17], and a fine-tuned ConvNeXt Tiny [29]. The pretrained Inception V3 corresponds to the standard FID definition. All fine-tuned backbones are trained on the same dataset under the same protocol, so the only difference is the architecture.

For each backbone, we extract the last-layer embedding before the classifier head and compute three metrics in that shared embedding space. Besides the FID variants, we also compute D&C and MMD using the same embeddings produced by the corresponding backbone. This design isolates the impact of the feature extractor on the metric behavior.

4.2. Natural Image Distribution Shift Experiments

This experiment examines whether FID preserves a reliable similarity ranking when data distributions progressively depart from the ImageNet pretraining distribution. By introducing perturbations with increasing severity, it evaluates how different types of shift affect ranking stability. It also compares multiple feature backbones and tests whether fine-tuning on shifted data can recover the original ordering under strong corruptions.

We select 120 classes from Tiny ImageNet 200 and organize them into 20 groups of six semantically related classes. Each group contains one anchor class. The remaining five classes are chosen to have different degrees of semantic proximity to the anchor. For each group, we compute the FID between the anchor class distribution and each of the other class distributions, then sort the five classes by increasing distance to obtain a reference anchor-centric similarity ranking. Across all groups, the resulting orderings match human intuition. The full list of the 20 groups is provided in Supplementary Material Table 3.

To test the robustness of these FID-based rankings under distribution shift, we apply data augmentation to create shifted class distributions and recompute the anchor-centric rankings at increasing severity levels. The perturbations include brightness, contrast, saturation, hue, Gaussian blur, and Gaussian noise. As shown in Fig. 2, our natural-image setting uses a severity-controlled perturbation scheme, enabling a systematic analysis of how class-similarity rank-

Figure 2. **Illustration of the perturbations used in the natural-image experiments.** Starting from the original image, we apply six common image transformations and gradually increase their severity to produce controlled distribution shifts.

ings change as the shift from the clean distribution increases. We quantify ranking stability by comparing the shifted ranking to the clean ranking using Spearman's rank correlation coefficient (Spearman ρ) and Kendall's τ_b rank correlation (Kendall τ_b).

As shown in Fig. 4, color perturbations have limited impact on ranking stability, while blur and noise progressively destroy the original ordering as severity increases. This may reflect the relative invariance of Inception features to global color statistics induced by standard color-jitter augmentation, aligning with prior work [22] on FID sensitivity. In contrast, blur removes mid- and high-frequency details that many CNN features rely on for fine-grained discrimination, and additive noise injects high-frequency artifacts that can dominate early-layer responses. Both effects reshape the feature geometry in a way that is less tied to semantic content, which makes the anchor-to-class distances less reliable and destabilizes the ranking.

Fig. 3 provides an intuitive example of how blur changes FID-based rankings. For the anchor class lemon, the five classes under FID differ before and after Gaussian blur is applied. The shift suggests that blur perturbs the similarity structure captured by the pretrained feature space.

We further test whether rank stability can be improved by adapting the feature extractor to the shifted distribution. We fine-tune Inception V3, ResNet50, and ConvNeXt Tiny on the same augmented training data and recompute FID, the D&C harmonic mean, and MMD. We report Gaussian blur in Fig. 5, and Gaussian noise follows the same pattern

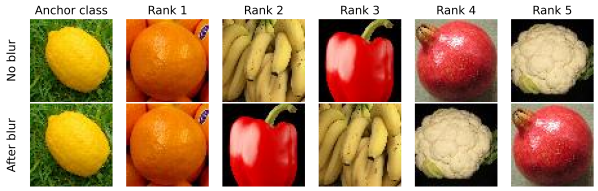


Figure 3. **Example of FID-based class ranking before and after blur.** For the anchor class lemon, the five classes are shown in ascending order of FID before and after applying Gaussian blur.

in Supplementary Material Fig. 7.

Across all three metrics, each backbone shows consistent behavior, which suggests that the observed effects are driven primarily by the embedding geometry rather than by a specific discrepancy measure. With the pretrained Inception V3, rank correlations decrease steadily as blur increases, indicating that the original similarity ordering is not preserved under strong low-pass corruption. Fine-tuned Inception V3, ResNet50 and ConvNeXt Tiny largely restore the original ordering across blur levels, indicating that their adapted representations preserve semantic neighborhoods more reliably under this shift. By comparison, D&C is systematically less stable than FID and MMD for the same backbone. Since D&C relies on nearest-neighbor neighborhoods and hard inclusion tests, small shifts in feature distances can change D&C more abruptly.

4.3. Non-Natural Image Experiments

This experiment extends the analysis to non-natural image data, where the target domain differs substantially from ImageNet. It examines how strongly inter-class similarity rankings depend on the feature backbone and whether fine-tuning on target-domain data leads to rankings that better reflect domain semantics. It also evaluates the consistency of these rankings across different backbones and discrepancy measures. To this end, FID, the D&C harmonic mean, and MMD are computed from penultimate-layer embeddings of a pretrained Inception V3 and three backbones fine-tuned on the target dataset, namely Inception V3, ResNet50, and ConvNeXt Tiny.

We consider two medical imaging benchmarks. The Breast Ultrasound Images dataset (BUSI) [2] contains three diagnostic categories corresponding to normal tissue, benign findings, and malignant tumors. The Knee Osteoarthritis severity grading dataset (KOA) [8] provides five ordered grades that represent increasing disease severity from grade 0 to grade 4.

Kendall τ_b is used to quantify ranking agreement between inter-class rankings, and the results are summarized in Table 1. Rankings induced by pretrained Inception V3 differ from those obtained after fine-tuning. In contrast, once the feature extractor is adapted, the resulting rank-

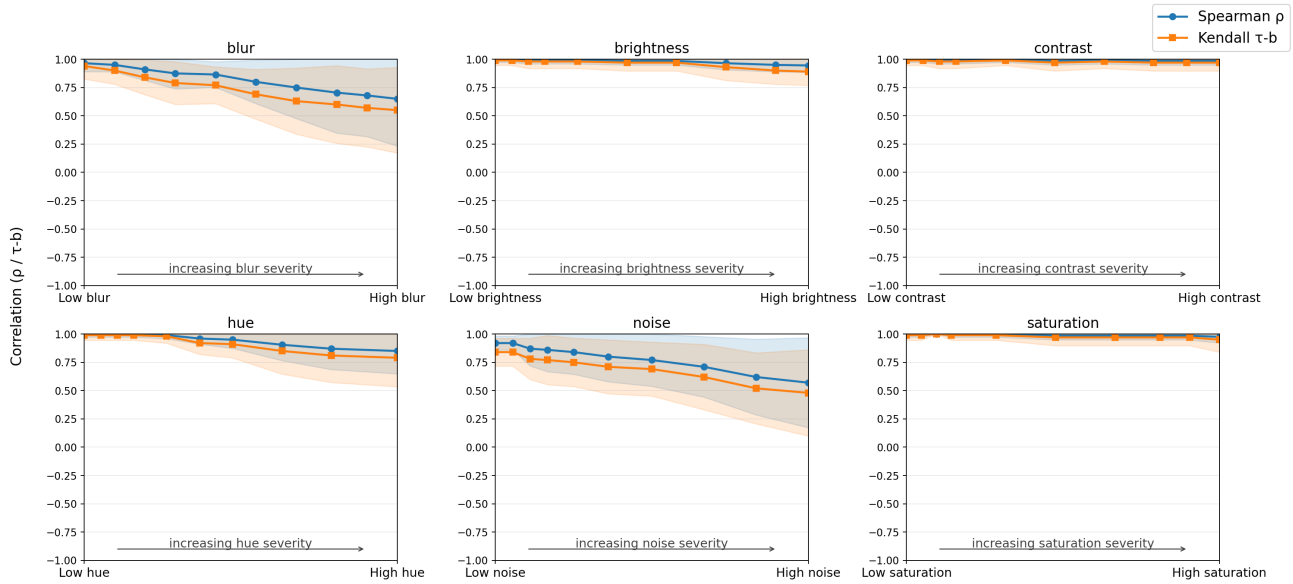


Figure 4. **FID ranking stability under natural image distribution shifts.** Spearman ρ and Kendall τ_b measure how well the anchor-based FID similarity ranking is preserved after applying each augmentation. Curves report the mean over 20 class groups, with shading showing variation across groups. Values closer to 1 indicate more consistent rankings.

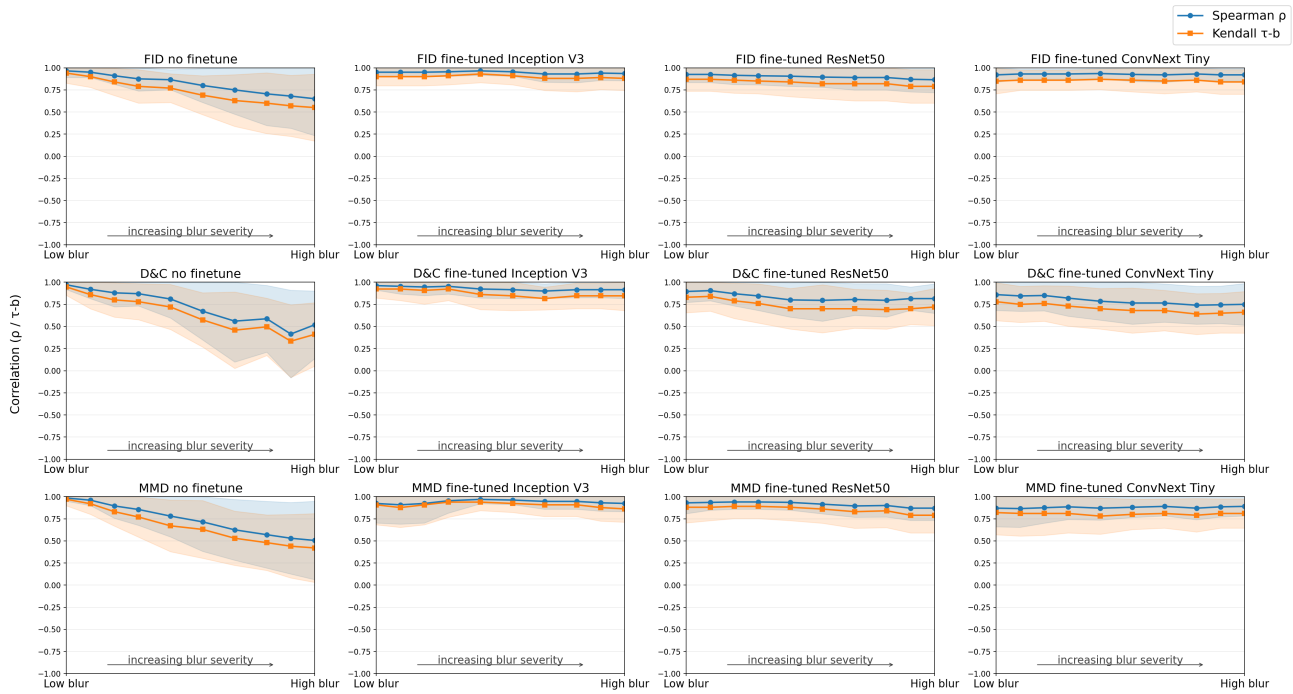


Figure 5. **Effect of feature extractor adaptation on ranking stability under Gaussian blur.** Rank correlations are compared for FID, the D&C harmonic mean, and MMD, using embeddings from a pretrained Inception V3 and three backbones fine-tuned on the shifted data.

ings are highly stable: agreement is at or near 1.0 across the three fine-tuned backbones and remains similarly high across FID, D&C, and MMD. This suggests that domain adaptation primarily determines the inter-class similarity

Table 1. **Kendall τ_b summary of ranking agreement on non-natural image datasets.** Rankings are derived from pairwise inter-class comparisons. **Pre→FT gap** averages Kendall τ_b between the pretrained Inception-V3 ranking and each fine-tuned-backbone ranking, computed separately for FID, D&C, and MMD. **FT backbone** averages Kendall τ_b over all fine-tuned backbones for each metric. **FT metric** averages Kendall τ_b over all metrics for each fine-tuned backbone. Values closer to 1 indicate more similar rankings.

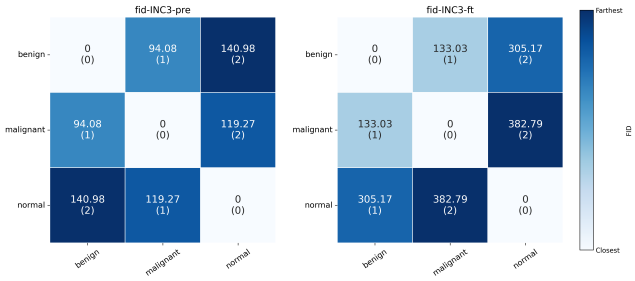
(a) Metric-wise summaries

Dataset	Summary	FID	D&C	MMD
BUSI	Pre→FT gap	0.33	0.33	0.33
	FT backbone	1.00	1.00	1.00
KOA	Pre→FT gap	0.93	0.85	0.93
	FT backbone	1.00	0.96	1.00

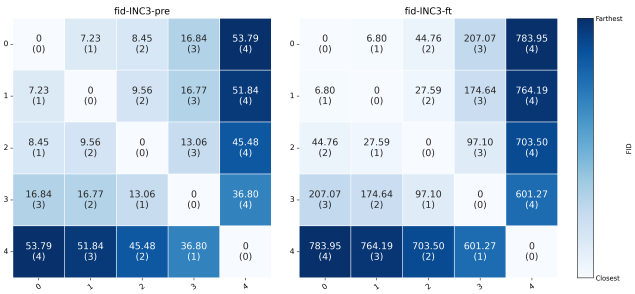
(b) Backbone-wise summary

Dataset	Summary	INC3-ft	RN50-ft	CNX-T-ft
BUSI	FT metric	1.00	1.00	1.00
KOA	FT metric	1.00	1.00	0.96

structure, while the specific choice of metric has limited impact after adaptation.



(a) **BUSI**. Pairwise FID between the three diagnostic categories.



(b) **KOA**. Pairwise FID between the five severity grades.

Figure 6. **Class-wise FID heatmaps on non-natural image datasets.** Standard FID is computed with pretrained Inception V3 features, and FID-INC3-ft uses fine-tuned Inception V3 features.

For BUSI, the three categories form a clinically progression from normal appearance to benign abnormalities and to malignant lesions. Fig. 6a shows that standard FID can contradict this progression, yielding a lower value between normal and malignant than between normal and benign. After fine-tuning, the FID relationships align with the expected ordering, with benign forming an intermediate category between normal and malignant.

For KOA, the labels are ordinal by construction and reflect increasing severity. Fig. 6b shows that standard FID can produce inconsistent ordering among adjacent grades, for example, by yielding a lower FID between grades 2 and 0 than between grades 2 and 1. Fine-tuning results in inter-grade FID relationships that better align with the severity progression.

Also, we observe that FID values computed with fine-tuned Inception V3 features are substantially larger than those from the pretrained model. ImageNet-pretrained features are not optimized for medical morphology, so the resulting embeddings provide limited discrimination between clinically different classes, leading to weakly structured FID values. Fine-tuning adapts the representation to domain-specific cues, enlarges relevant class margins in feature space, and produces FID relationships that are more informative for similarity ranking.

4.4. Synthetic Image Experiments

This experiment studies how rankings of synthetic image sets depend on the feature representation used for evaluation, and whether domain-adapted feature extractors lead to more consistent ranking behavior across different embedding spaces and discrepancy measures. To this end, we use medigan [34], an open-source library that distributes pretrained generative models for medical image synthesis. We select generators trained on NODE21 [42] and HyperKvasir [6].

NODE21 consists of frontal chest radiographs with nodule annotations. We use Medigan models 17 (DCGAN[38]) and 18 (WGAN-GP[16]) to generate lung-nodule image patches. Based on the dataset metadata, we crop real nodule patches from annotated radiographs and sample matched patches from nodule-free images to construct the training set for fine-tuning Inception V3, ResNet50, and ConvNeXt Tiny. We then generate 1,000 images with each generator and compute FID, D&C, and MMD between each synthetic set and the set of real nodule patches using embeddings extracted from the pretrained model and the three fine-tuned backbone networks.

HyperKvasir is a large-scale gastrointestinal endoscopy dataset that includes polyp images. We use Medigan models 9 (PGGAN[23]), 10 (FastGAN[28]), and 11 (SinGAN[41]), which are trained on polyp segmentation data, to generate synthetic polyp images. We then fine-tune the three feature extractors on the labeled images from HyperKvasir and

compare 1,000 generated samples from each model with the real set of segmented polyp images using the same three metrics across four embedding spaces.

Table 2 summarizes ranking agreement over synthetic image sets. Within each metric–feature space configuration, we rank these synthetic sets by their synthetic-to-real discrepancy under FID, D&C, or MMD. For each medigan model, we sample three independent synthetic sets, and within each metric–backbone setting the induced ranking is identical across the three samples.

Across both datasets, the pretrained Inception ranking largely matches the ranking obtained with domain-adapted feature spaces. Moreover, D&C and MMD are fully consistent across all feature extractors in our setting, yielding the same synthetic-set ordering for each dataset. The only deviations arise for FID under specific fine-tuned backbones: on NODE21, FID with ConvNeXt-T-ft flips the ordering between the two synthetic sets; on HyperKvasir, FID with ResNet50-ft changes the relative order of the two non-best synthetic sets.

Overall, the synthetic-to-real study suggests that ranking-based conclusions depend on both the metric and the representation, and that reporting a single FID score in one feature space can be insufficient for reliable synthetic-set selection.

Table 2. **Kendall τ_b summary of ranking agreement on synthetic-to-real comparisons.** Rankings are obtained by ordering synthetic image sets according to their similarity to the corresponding real dataset under FID, D&C, and MMD. For each generator, three independently sampled synthetic sets were used. for each metric and backbone setting, the ranking was unchanged across repetitions. Values closer to 1 indicate more similar rankings.

(a) Metric-wise summaries

Dataset	Summary	FID	D&C	MMD
NODE21	Pre→FT gap	0.33	1.00	1.00
	FT backbone	-0.33	1.00	1.00
HyperKvasir	Pre→FT gap	0.78	1.00	1.00
	FT backbone	0.56	1.00	1.00

(b) Backbone-wise summary

Dataset	Summary	INC3-ft	RN50-ft	CNX-T-ft
NODE21	FT metric	1.00	1.00	-0.33
HyperKvasir	FT metric	1.00	0.56	1.00

5. Discussion

The experimental findings can be naturally interpreted through the three questions that motivate this work. The first is **whether FID preserves a reliable ranking of dataset similarity when the data distribution moves away from the ImageNet pretraining distribution**: The natural-image experiments suggest that it generally does not. FID-based rankings are mostly preserved under photometric shifts but become less stable under Gaussian blur and noise.

The second question concerns **how strongly this behavior depends on the feature backbone that defines the embedding space**: The results indicate that the main factor is the feature representation itself, particularly whether the embedding space is ImageNet-pretrained or adapted to the target domain. Across both the natural-image and non-natural image experiments, different fine-tuned backbones produce highly similar rankings. The same pattern is also observed for FID, D&C, and MMD, whose induced rankings remain largely consistent under the same representation.

The third question is **whether fine-tuning the feature extractor on shifted or target-domain data can reduce ranking instability and recover orderings that better reflect the domain**: The results suggest that it can. In the natural-image setting, once the feature extractors are adapted to the shifted data, ranking consistency is largely restored, which suggests that the instability is driven more by the representation than by the discrepancy metrics themselves. In the non-natural image setting, standard ImageNet-pretrained Inception features can produce inter-class relationships that conflict with domain semantics, whereas domain-adapted feature extractors recover clinically plausible rankings.

In the **synthetic-to-real setting**, most rankings remain stable across metrics and feature spaces, but some FID-based orderings change after adaptation. This suggest that relying on a single FID score in a single pretrained feature space may give only a partial view of synthetic-to-real similarity.

Limitations. At the same time, this study remains limited in scope, since only a small number of pretrained generators are evaluated on two medical datasets, and the rankings are derived from discrepancy to real data rather than downstream task performance. These results should therefore be viewed as initial evidence that ranking behavior can vary with the representation, and that synthetic data evaluation may benefit from reporting multiple metrics together with domain-adapted feature spaces.

References

[1] Akmalbek Bobomirzaevich Abdusalomov, Rashid Nasimov, Nigorakhon Nasimova, Bahodir Muminov, and Taeg Keun Whangbo. Evaluating synthetic medical images using artificial intelligence with the gan algorithm. *Sensors*, 23(7): 3440, 2023. 3

[2] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled,

and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 5

[3] Nicholas Amano, Bo Lei, Martin Müller, Frank Mücklich, and Elizabeth A Holm. Microscopy modality transfer of steel microstructures: Inferring scanning electron micrographs from optical microscopy using generative ai. *Materials Characterization*, 220:114600, 2025. 3

[4] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024. 3

[5] Eyal Betzalel, Coby Penso, and Ethan Fetaya. Evaluation metrics for generative models: An empirical study. *Machine Learning and Knowledge Extraction*, 6(3):1531–1544, 2024. 1, 2

[6] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020. 7

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[8] Pingjun Chen. Knee osteoarthritis severity grading dataset. *Mendeley Data*, 1(10.17632):30784984, 2018. 5

[9] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019. 3

[10] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079, 2020. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[12] Kexin Ding, Mu Zhou, He Wang, Olivier Gevaert, Dimitris Metaxas, and Shaoting Zhang. A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. *Scientific Data*, 10(1):231, 2023. 2

[13] DC Dowson and BV666017 Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982. 2, 3

[14] Cong Gao, Benjamin D Killeen, Yicheng Hu, Robert B Grupp, Russell H Taylor, Mehran Armand, and Mathias Unberath. Synthetic data accelerates the development of generalizable learning-based algorithms for x-ray image analysis. *Nature Machine Intelligence*, 5(3):294–308, 2023. 2

[15] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012. 2

[16] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 7

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1, 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[20] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024. 2, 3, 4

[21] Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood divergence: evaluating the generalization of generative models using samples. *Advances in Neural Information Processing Systems*, 36:33095–33119, 2023. 3

[22] Steffen Jung and Margret Keuper. Internalized biases in fréchet inception distance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*. 1, 3, 5

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 7

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[25] Bardia Khosravi, Frank Li, Theo Dapamede, Pouria Rouzrokh, Cooper U Gamble, Hari M Trivedi, Cody C Wyles, Andrew B Sellergren, Saptarshi Purkayastha, Bradley J Erickson, et al. Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. *EBioMedicine*, 104, 2024. 2

[26] Lennart R Koetzier, Jie Wu, Domenico Mastrodicasa, Aline Lutz, Matthew Chung, W Adam Koszek, Jayanth Pratap, Akshay S Chaudhari, Pranav Rajpurkar, Matthew P Lungren, et al. Generating synthetic data for medical imaging. *Radiology*, 312(3):e232471, 2024. 2

[27] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. In *ICLR*, 2023. 3

[28] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International conference on learning representations*, 2020. 7

[29] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 4

[30] Keith Man and Javaan Chahl. A review of synthetic image data and its use in computer vision. *Journal of Imaging*, 8 (11):310, 2022. 1

[31] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5): e210315, 2022. 3

[32] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pages 7176–7185. PMLR, 2020. 3

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[34] Richard Osuala, Grzegorz Skorupko, Noussair Lazrak, Lidia Garrucho, Eloy García, Smriti Joshi, Socayna Jouide, Michael Rutherford, Fred Prior, Kaisar Kushibar, et al. medigan: a python library of pretrained generative models for medical image synthesis. *Journal of Medical Imaging*, 10 (6):061403–061403, 2023. 3, 7

[35] Anthony Paproki, Olivier Salvado, and Clinton Fookes. Synthetic data for deep learning in computer vision & medical imaging: A means to reduce data bias. *ACM Computing Surveys*, 56(11):1–37, 2024. 2

[36] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11410–11420, 2022. 1, 2

[37] Pushpak Pati, Sofia Karkampouna, Francesco Bonollo, Eva Compérat, Martina Radić, Martin Spahn, Adriano Martinelli, Martin Wartenberg, Marianna Kruithof-de Julio, and Marianna Rapsomaniki. Accelerating histopathology workflows with generative ai-based virtually multiplexed tumour profiling. *Nature machine intelligence*, 6(9):1077–1093, 2024. 3

[38] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 7

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[40] Ole Schmedemann, Melvin Baaß, Daniel Schoepflin, and Thorsten Schüppstuhl. Procedural synthetic training data generation for ai-based defect detection in industrial surface inspection. *Procedia CIRP*, 107:1101–1106, 2022. 2

[41] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4570–4580, 2019. 7

[42] Ecem Sogancioglu, Bram Van Ginneken, Finn Behrendt, Marcel Bengs, Alexander Schlaefer, Miron Radu, Di Xu, Ke Sheng, Fabien Scalzo, Eric Marcus, et al. Nodule detection and generation on chest x-rays: Node21 challenge. *IEEE transactions on medical imaging*, 43(8):2839–2853, 2024. 7

[43] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023. 3

[44] Li Sun, Junxiang Chen, Yanwu Xu, Mingming Gong, Ke Yu, and Kayhan Batmanghelich. Hierarchical amortized gan for 3d high resolution medical image synthesis. *IEEE journal of biomedical and health informatics*, 26(8):3966–3975, 2022. 3

[45] Onkar Susladkar, Gayatri Deshmukh, Yalcin Tur, Gorkem Durak, and Ulas Bagci. Victr: Vital consistency transfer for pathology aware image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22772–22782, 2025. 3

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2

[47] Jie Wang, Nazia Tabassum, Tanjin T Toma, Yibo Wang, Andreas Gahlmann, and Scott T Acton. 3d gan image synthesis and dataset quality assessment for bacterial biofilm. *Bioinformatics*, 38(19):4598–4604, 2022. 3

[48] McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Suprateek Kundu, Joshua P Yung, Ankit B Patel, et al. Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 87–97. Springer, 2024. 3

[49] Matthew Yates, Glen Hart, Robert Houghton, Mercedes Torres Torres, and Michael Pound. Evaluation of synthetic aerial imagery using unconditional generative adversarial networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:231–251, 2022. 3

Fréchet Inception Distance is Failing to Preserve Rank Consistency for Synthetic Out-of-Distribution Samples

Supplementary Material

Table 3. Twenty class groups used in the natural-image distribution shift experiment. Each group contains one anchor class and five similar classes ranked by similarity in descending order (most similar → least similar).

Group	Anchor class	Most similar class → least similar class				
		1	2	3	4	5
beach gear	bikini	swimming trunks	snorkel	sunglasses	umbrella	sandal
birds fish	goose	albatross	black stork	dugong	king penguin	goldfish
building exteriors	barn	thatched roof	picket fence	birdhouse	barbershop	butcher shop
cats	egyptian cat	tabby cat	persian cat	cougar	lion	red panda
coastal scenes	seashore	lakeside	cliff	alp	coral reef	cliff house
containers	beer bottle	soda bottle	pill bottle	water jug	bucket	barrel
cooked foods	pizza	potpie	meat loaf	mashed potato	ice cream	pretzel
dogs	golden retriever	labrador retriever	standard poodle	german shepherd	Chihuahua	yorkshire terrier
fruits	lemon	orange	banana	bell pepper	pomegranate	cauliflower
furniture	rocking chair	table	chest	desk	bathtub	lampshade
home tools	refrigerator	space heater	plunger	candle	broom	lawn mower
insects	butterfly	sulphur butterfly	bee	dragonfly	ladybug	cockroach
kitchenware	teapot	water jug	wooden spoon	wok	frying pan	plate
public transit	school bus	trolleybus	police van	moving van	freight car	bullet train
reptiles amphibians	bullfrog	tailed frog	american alligator	european fire salamander	scorpion	boa constrictor
road vehicles	sports car	convertible	station wagon	limousine	go kart	tractor
sea life	snail	slug	sea slug	sea cucumber	jellyfish	brain coral
sports gear	basketball	volleyball	rugby ball	dumbbell	punching bag	stopwatch
street fixtures	parking meter	pay phone	atm	bannister	turnstile	maypole
ungulates	bison	ox	pig	bighorn sheep	camel	gazelle

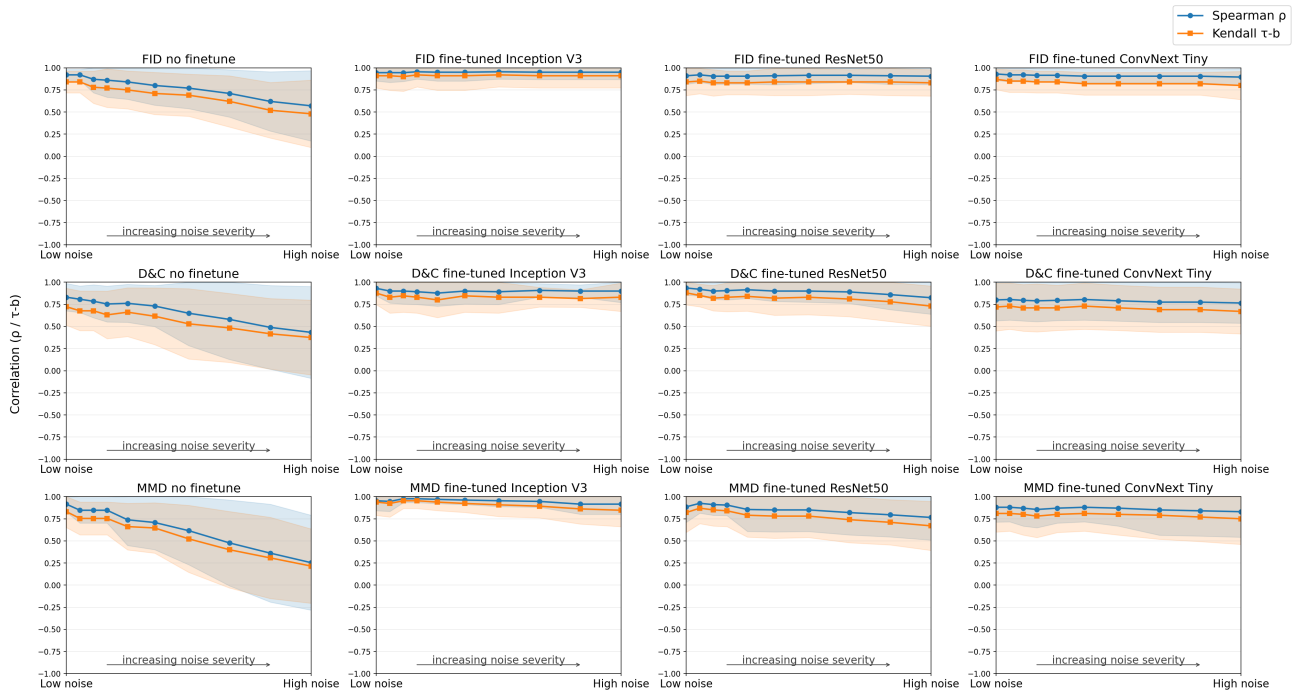


Figure 7. **Effect of feature extractor adaptation on ranking stability under Gaussian noise.** Rank correlations are compared for FID, the D&C harmonic mean, and MMD, using embeddings from a pretrained Inception V3 and three backbones fine-tuned on the shifted data.

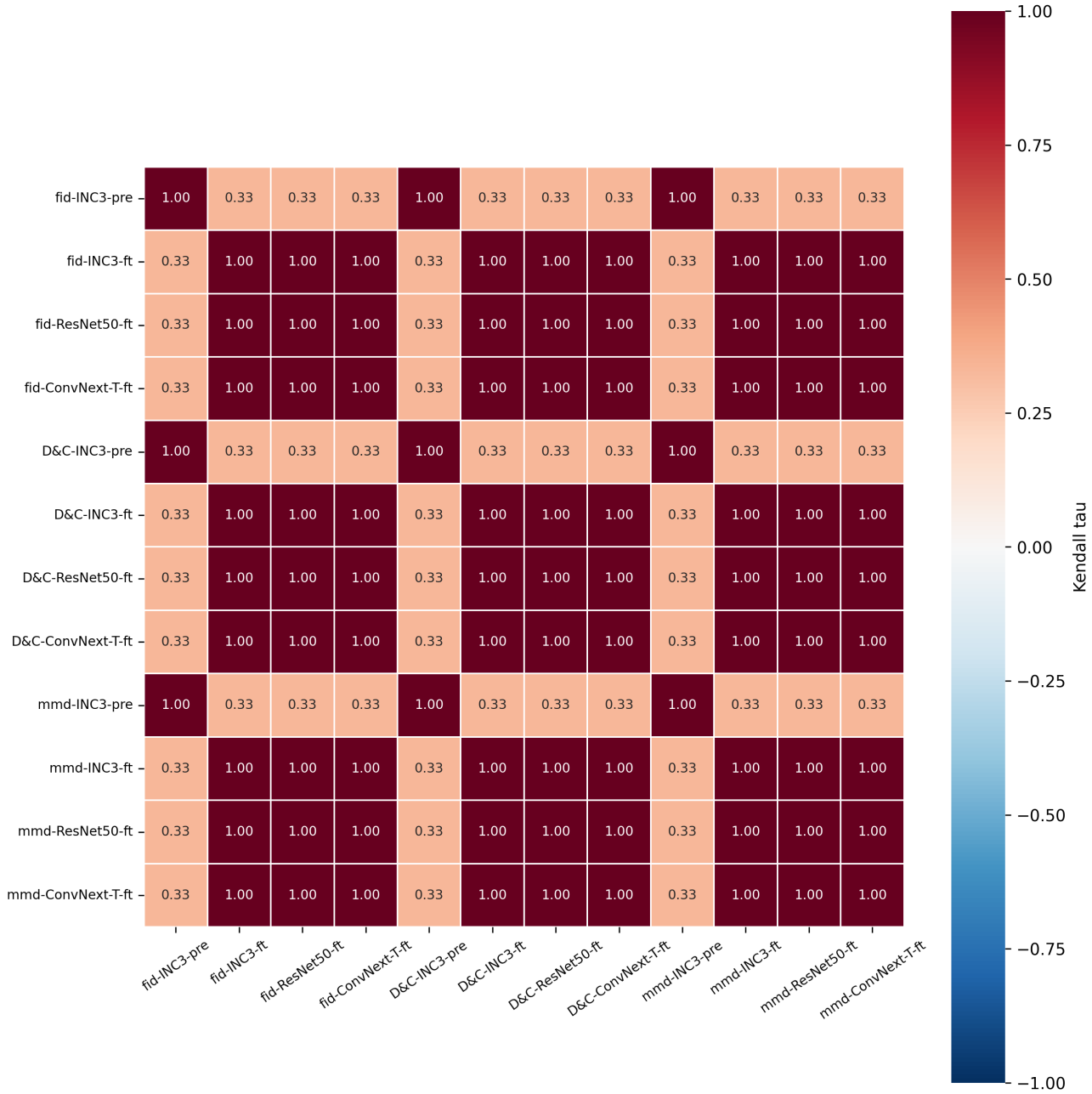


Figure 8. **Kendall τ_b matrix of ranking agreement across metrics and feature backbones on BUSI.**

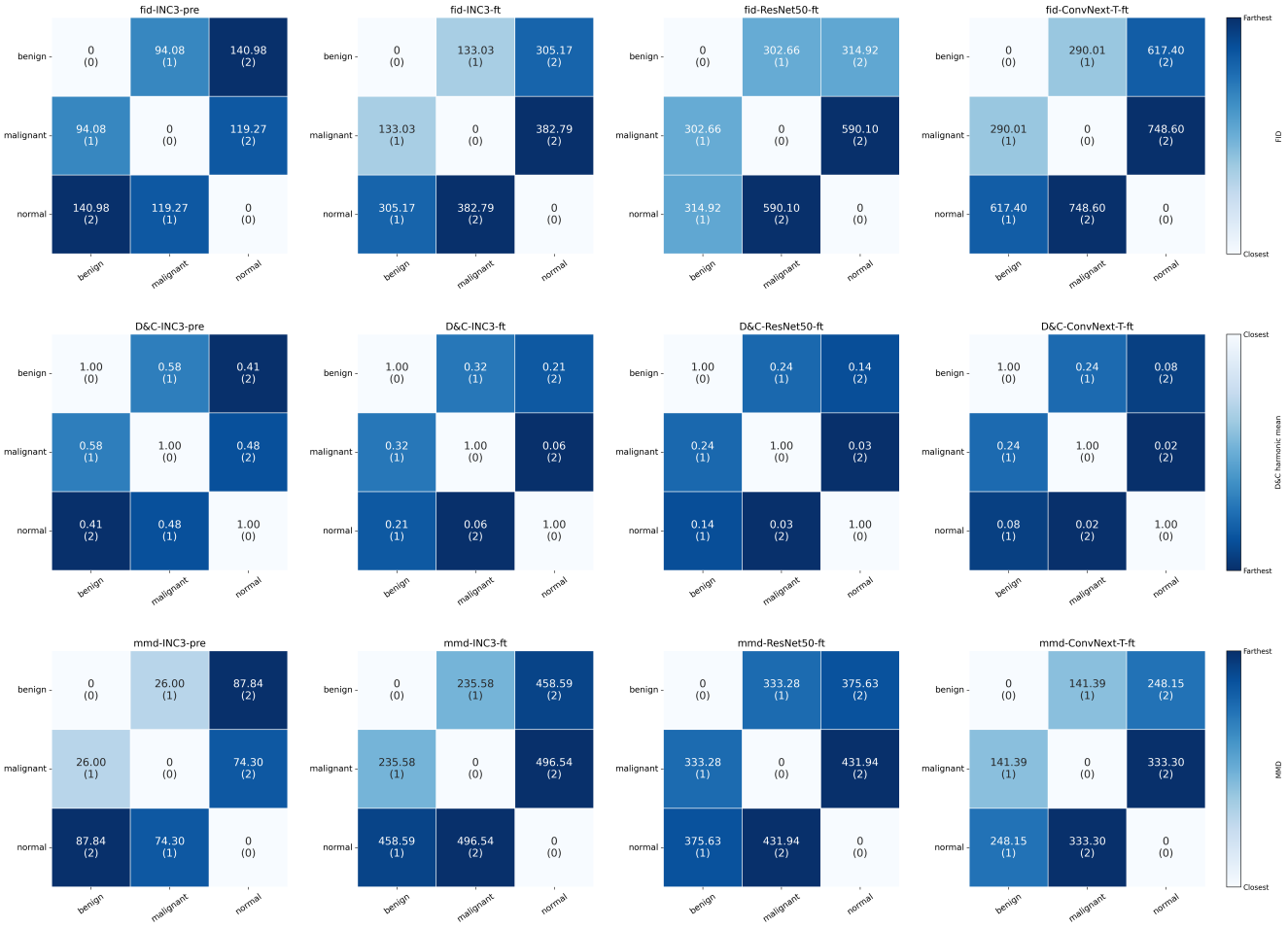


Figure 9. **Pairwise inter-class comparison heatmaps across metrics and feature backbones on BUSI.**

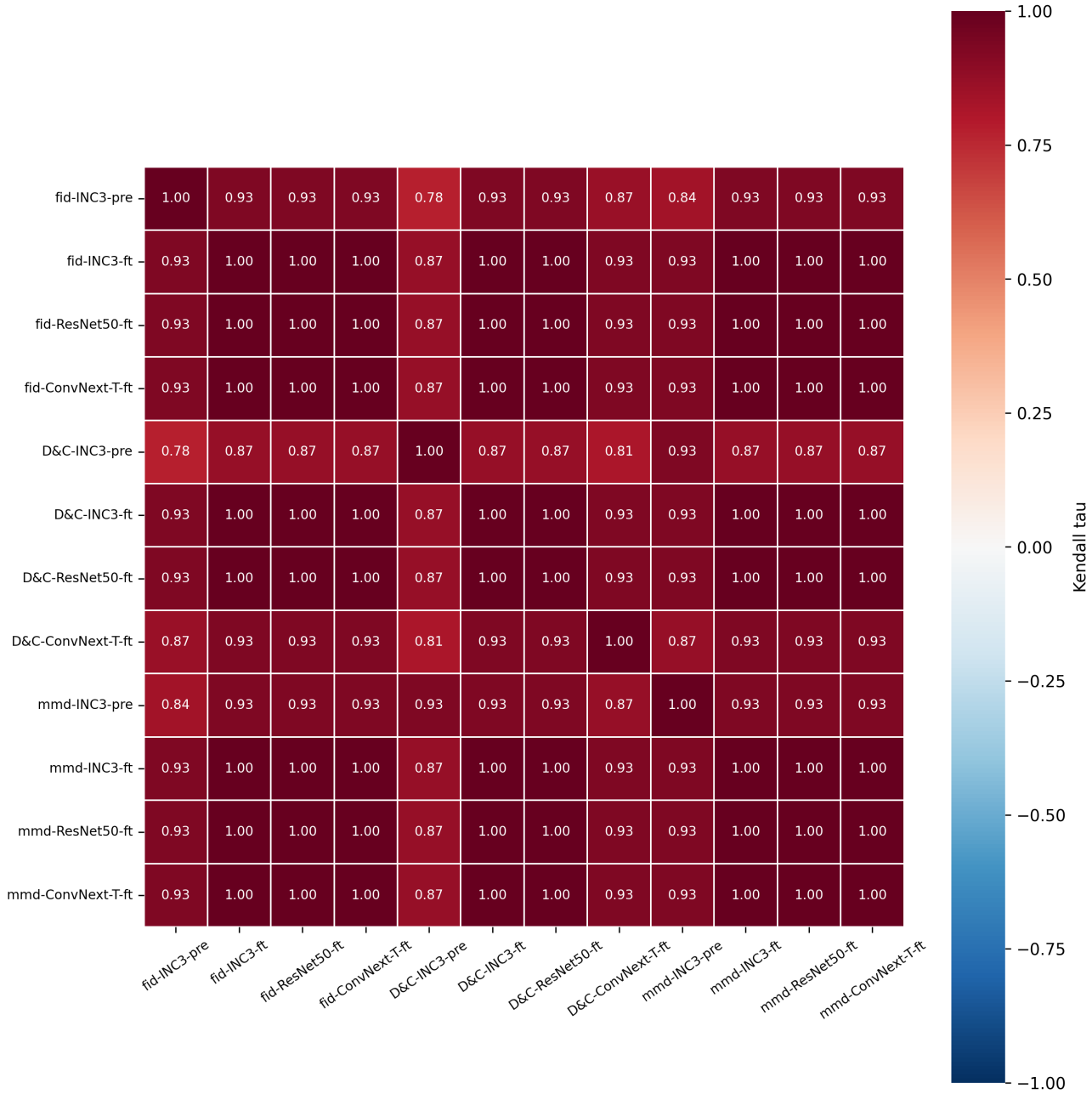


Figure 10. **Kendall τ_b matrix of ranking agreement across metrics and feature backbones on KOA.**

Figure 11. **Pairwise inter-class comparison heatmaps across metrics and feature backbones on KOA.**