Are Wavelets More Reliable? Comparing The Topology of Image Generator Metrics

Manas Patil*, Amit Tiwari*, Abhijat Bharadwaj*, and Vikram Gadre
Indian Institute of Technology Bombay
Email: manaspatil@ee.iitb.ac.in, {amit.t, bharadwaj.abhijat, vmgadre}@iitb.ac.in

***Abstract*—Existing evaluation metrics of image generative models typically rely on perceptual features extracted from pretrained deep neural networks, yet such metrics often exhibit unreliability and limited interpretability. We propose a Generator Distance Metric (GDM) based on generation fidelity and diversity, and introduce a topological data analysis framework to assess the stability of these metrics. We evaluate Fréchet Wavelet Distance (FWD), Fréchet Inception Distance (FID), FD-DINOv2, and CLIP Maximum Mean Discrepancy (CMMD) across image generators trained on CIFAR-10 and FFHQ-64. Our results indicate that FWD-based GDM induces topological and geometric structures that exhibit greater stability and provide qualitative interpretations than those based on DNN backbone-based metrics.**

***Index Terms*—Image generation, metric reliability, topological data analysis, wavelet packets.**

I. Introduction

Modern image generation models such as Generative Adversarial Networks (GANs) [1], [2], [3], Variational Autoencoders (VAEs) [4], and diffusion models [5], [6] have undergone significant advances over the last decade. These image generators are evaluated on perceptual properties such as realism and diversity of their generated images, which traditional metrics fail to accurately capture. Current evaluation metrics, such as Inception Score (IS) [7], Fréchet Inception Distance (FID) [8], FD-DINOv2 [9], etc., compare empirical distributions of images in the feature space of a deep neural network (DNN) pretrained on large image datasets.

Studies like [10], [11], [12] have shown that DNN-based metrics are uninterpretable, sample-size-sensitive, biased toward specific domains and image generators, and prone to incorrect assumptions, raising concerns about their reliability and fairness. Recent developments such as the Fréchet Wavelet Distance (FWD) [11] and CLIP Maximum Mean Discrepancy (CMMD) [12] aim to address some of the above problems. FWD uses a domain-agnostic wavelet packet transform (WPT) to create an unbiased, interpretable metric. CMMD uses CLIP [13] embeddings and relaxes the normality assumption. However, the reliability of such metrics remains uncertain.

In this work, we propose to solve this problem from a topological data analysis (TDA) perspective. TDA-based approaches have been shown to be effective for interpreting loss landscapes [14]. Using an evaluation metric, we measure the separation between image generators on two key parameters: *ground-truth fidelity* and *cross-generator dissimilarity*. We further use this measure to form a topological space, namely a Vietoris-Rips complex, of image generators. Through persistent homology, we evaluate the stability of the topological spaces generated by four prominent metrics, viz. FID, FD-DINOv2, CMMD, and FWD using various image generators. The main contributions of the work are summarized as follows:

1) We introduce Generator Distance Metric (GDM) and a novel TDA-based framework for assessing the reliability of image generation metrics.
2) Using this framework, we evaluate the metrics derived from three prominent pretrained backbones (Inception-v3 [15], DINO-v2 [16], and CLIP [13]) and a multi-resolution backbone (WPT) across image generators with diverse characteristics.
3) Our analysis shows that FWD yields a stable topological representation of image generators and demonstrates substantially higher reliability than metrics based on pretrained backbones.
4) We further provide novel insights into the geometric and topological structures induced by FWD across different image generators.

II. Related Work

Image Generator Metrics. Inception Score (IS), introduced by Salimans *et al.* [7], measures sample quality and diversity via KL-divergence between the conditional and marginal label distributions produced by an Imagenet pretrained Inception classifier. Fréchet Inception Distance (FID) [8] measures the Fréchet distance between the distributions of Inception embeddings of generated and real images, approximated to a Gaussian. Chong and Forsyth [10] identified sample size dependence in the metrics and proposed IS_∞ and FID_∞. Studies also analyze Fréchet distance metrics using various backbones [9], [17]. Veeramacheneni *et al.* [11] proposed FWD, correcting for domain bias in pretrained networks. CMMD [12] utilizes the CLIP [13] model and relaxes normality assumptions, demonstrating better human perceptual alignment than FID.

* Authors contributed equally to the work.

Geometric and Topological Analysis of DNNs. Seminal works in DNN interpretability by Garipov *et al.* [18], Draxler *et al.* [19], and Kuditipudi *et al.* [20] study loss landscapes and the existence of distinct minima on low-loss manifolds in stochastic gradient descent. Complementarily, TDA-based approaches study neural architectures, decision boundaries, activation-space topology via persistent homology [14]. Horoi *et al.* [21] apply TDA, specifically, persistent homology, to neural network loss landscapes and demonstrate that these topological features correlate with training dynamics and generalization, providing structural insights beyond traditional gradient-based methods. Recent work shows that Jacobian-based local geometry (scaling, rank, complexity) of generative manifolds predicts sample quality, diversity, and memorization, with poorer samples exhibiting lower values [22].

III. Preliminaries

Vietoris–Rips Complex. [23] Given a finite point set P in a metric space and a scale parameter $\epsilon \geq 0$, the Vietoris–Rips complex $VR_\epsilon(P)$ is an abstract simplicial complex defined as follows: a k-simplex $\{p_0, \ldots, p_k\}$ is included in $VR_\epsilon(P)$ if and only if $d(p_i, p_j) \leq \epsilon$ for all pairs i, j. Intuitively, this construction encodes proximity relationships among points at a given scale.

Homology Groups. [23] Let K be a simplicial complex. A k-chain is a formal sum of k-simplices with coefficients in $\mathbb{Z}_2$. The boundary operator ∂_k maps k-chains to $(k-1)$-chains. The set of k-cycles is defined as $Z_k = \ker(\partial_k)$, and the set of k-boundaries as $B_k = \operatorname{im}(\partial_{k+1})$. Since $\partial^2 = 0$, we have $B_k \subseteq Z_k$. The k-th homology group is the quotient space $H_k(K) = Z_k/B_k$, capturing k-dimensional topological features.

Filtration. [23] A filtration is a nested sequence of simplicial complexes $\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$, typically obtained by varying a scale parameter.

Persistent Homology. [23] Persistent homology studies the evolution of homological features across a filtration. A feature is said to be born at scale r if it appears in $H_k(K_r)$ but not earlier, and dies at scale $s > r$ if it becomes trivial or merges with an older feature in $H_k(K_s)$. The persistence of a feature is given by its lifetime $s - r$, reflecting its significance across scales.

The 0-th homology group H_0 characterizes connected components of the space, while the 1-st homology group H_1 captures independent 1-cycles, commonly interpreted as loops or holes. In our context, persistent H_1 cycles can be interpreted as regions where generator families pass through characteristic regimes.

IV. Proposed Framework

Let $\mathcal{G} = \{G_1, \ldots, G_M\}$ denote M image generators, each trained on a common image dataset, with distribution $\mathcal{D}_0$ in the feature space of some encoder $\mathcal{E}$. The output images of these generators follow distributions $\mathcal{D}_1, \ldots, \mathcal{D}_M$, respectively, in the feature space of $\mathcal{E}$. For a distance function $f_\mathcal{E}(\mathcal{D}_i, \mathcal{D}_j)$ measuring the distance between the image distributions $\mathcal{D}_i$ and $\mathcal{D}_j$, define a Generator Distance Metric (GDM) $r_{i,j}$ as

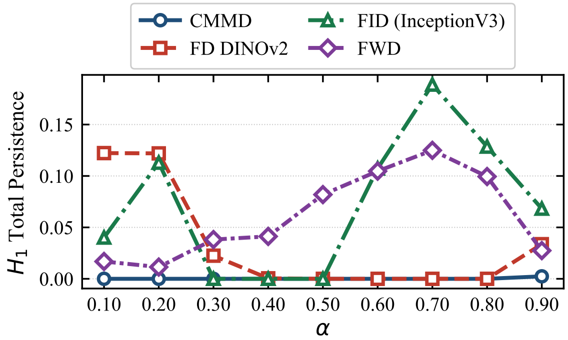
$$r_{i,j} = \alpha \underbrace{\left| f_\mathcal{E}(\mathcal{D}_i, \mathcal{D}_0) - f_\mathcal{E}(\mathcal{D}_j, \mathcal{D}_0) \right|}_{r_{\text{fidelity}}} + \beta \underbrace{f_\mathcal{E}(\mathcal{D}_i, \mathcal{D}_j)}_{r_{\text{generator}}} \quad (1)$$

where the constants $0 < \alpha, \beta < 1$, and $\beta = 1 - \alpha$. Note that $r_{i,j}$ is a distance $\forall$ distance function $f_\mathcal{E}$. $f_\mathcal{E}$ The first term (r_{fidelity}) captures the absolute difference in the *ground-truth fidelity* of the two generators, and the second term ($r_{\text{generator}}$) captures their mutual dissimilarity or *cross-generator diversity*. Parameter α (and thus, β) controls the trade-off between the two terms in the generator distance. Note that $\forall i, j$: 1) $r_{i,j} = r_{j,i}$, 2) $r_{i,0} = f_\mathcal{E}(\mathcal{D}_i, \mathcal{D}_0)$, and 3) at $\alpha \to 0$, $r_{i,j} \to f_\mathcal{E}(\mathcal{D}_i, \mathcal{D}_j)$. Thus, GDM is a generalization of the original metric $f_\mathcal{E}$, with an added constraint measuring data fidelity.

Given the $M \times M$ distance matrix R with entries $R_{ij} = r_{i,j}$ for all $1 \leq i, j \leq M$, compute the Vietoris–Rips persistent homology. The resulting persistence diagrams capture topological features characterized by their birth and death filtration scales. Interpreting the generators $G_i \in \mathcal{G}$ as points (vertices) in the underlying space, the persistent H_0 features identify generators that merge into connected components at various filtration levels. In contrast, the persistent H_1 features capture non-trivial 1-dimensional cycles reflecting higher-order geometric relationships between the generators. We observe that the H_0 features obtained through GDM for $f_\mathcal{E} \in$ {FID, FWD, FD-DINOv2, CMMD} almost always preserve their rank order for all variations in α, indicating that GDM creates a reliable topological space with varied α (this is discussed further in the supplementary material). Our framework analyzes the reliability of a metric ($f_\mathcal{E}$) on the stability of persistence of homology groups with variations in α.

V. Experiments and Observations

Experimental Setup. In this work, we analyze the following metrics ($f_\mathcal{E}$): FID (Inception-v3), FD-DINOv2, FWD (Haar wavelet), and CMMD. Since FID, FWD, and FD-DINOv2 use squared distances as scores, we use their square roots as distance metrics. The Ripser [24] backend was used to compute persistent homology. We provide the analysis on generators trained on CIFAR-10 [25] and FFHQ-64×64 [26] datasets. The image generators are taken from the following pretrained families: DCGAN [1], DDGAN [2], DDPM [5] (with both DDPM and DDIM [27] inference samplers), EDM [6], VAE [4], and StyleGAN2-ADA [3], with varying hyper-parameters and configurations (see Table I). For each configuration, the feature distribution was empirically estimated over 2000 generated images. Note that at higher α, the GDM collapses to r_{fidelity}, leading to lower total persistence across all metrics.



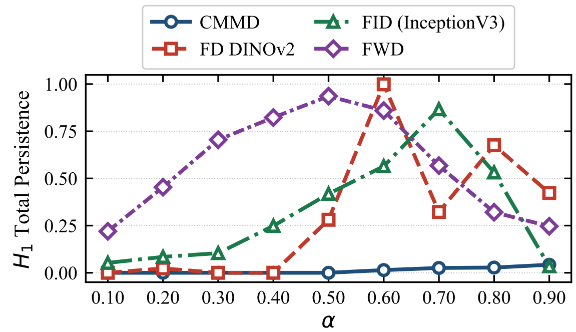


Fig. 1. Stability of the H_1 features reported in the Rips filtration of distance matrices made up of each metric by calculating total persistence of all reported H_1 cycles for the dataset CIFAR-10 (left) and FFHQ-64 $\times$ 64 (right).

TABLE I
IMAGE GENERATOR CONFIGURATIONS

Dataset ($\mathcal{D}_0$)	Family	Hyper-Parameter and Configuration Variations ($\mathcal{G}$)	Total Configurations
CIFAR-10	DCGAN	Latent noise scale $\sigma \in \{0.5, 0.7, 0.85, 1.0, 1.15, 1.3, 1.5\}$	27
	DDGAN	Sampling temperature $\tau \in \{0.7, 0.85, 1.0, 1.15, 1.3\}$	
	DDPM	DDIM sampler: $\{50, 100, 200\}$ steps; DDPM sampler: 1000 steps	
	EDM	Steps, Churn $\in \{(18, 0), (35, 0), (35, 20), (35, 40), (79, 0)\}$	
	VAE	Decoder temperature $\tau \in \{0.5, 0.7, 0.85, 1.0, 1.2, 1.5\}$	
FFHQ	DCGAN	Latent noise scale $\sigma \in \{0.5, 0.65, 0.7, 0.85, 1.0, 1.15, 1.3, 1.5\}$	29
	EDM	Steps, Churn $\in \{(18, 0), (35, 0), (40, 0), (40, 20), (40, 40), (40, 60), (79, 0)\}$	
	VAE	Decoder temperature $\tau \in \{0.5, 0.7, 0.85, 1.0, 1.15, 1.2, 1.5\}$	
	Style-GAN2	Truncation $\Psi \in \{0.3, 0.5, 0.7, 0.85, 1.0, 1.15, 1.3\}$	

A. Topological Stability of the Metrics

In TDA, homology features with high persistence are typically regarded as meaningful, whereas low-persistence features are often attributed to noise or sampling artifacts [23]. Furthermore, by the Stability Theorem [23], for a well-behaved distance function $f_{\mathcal{E}}$, the persistence diagrams of the GDM at α and $\alpha + \Delta\alpha$ ($\Delta\alpha \ll \alpha$) differ only slightly. Consequently, smoothly varying α should induce only gradual changes in the persistence of homological features.

To assess the stability of the topological space induced by the GDM, we vary α from 0.1 to 0.9 and compute the total persistence (i.e., the sum of lifetimes) of all H_1 features observed over the filtration. Figure 1 reveals that out of the four metrics, FWD shows the smoothest variation in the total persistence. The total persistence of H_1 features from CMMD is extremely small (close to zero $\forall\alpha$), suggesting that within our setting under the GDM construction, CMMD does not encode sufficient structure to yield persistent H_1 features. FID and FD-DINOv2 show erratic total persistence, as against the stability of the same in FWD-based filtration. This indicates that FWD is more topologically stable than pretrained backbone metrics such as FID, FD-DINOv2, and CMMD across our set of image generators (see Table I).

B. Geometric Interpretations

We observed that, in addition to topological stability analysis, GDM provides a geometric perspective on the interpretations of image generator behaviour. We demonstrate the results for $\alpha = 0.6$, as it balances fidelity and generator dissimilarity terms. We use non-metric multidimensional scaling (MDS) [28] to visualize the higher-dimensional space created by GDM using various metrics by mapping it to a lower, two-dimensional space. It preserves the rank order of the original pairwise dissimilarities between points. The stress factor evaluates the goodness-of-fit of a Multidimensional Scaling, based on the monotonic transformation of the original data [28].

Table II shows the stress factor for the metrics under study. Consistently, the MDS reliability is in the order: CMMD $>$ FWD $>$ FID $>$ FD-DINOv2. Due to the inconclusive nature of CMMD for our set of generators (see Section V-A), we present the results on MDS obtained for FWD, which has significantly smaller stress than FID and FD-DINOv2. Over a scalar FWD metric, Figure 2 provides directional interpretation of the relative distances of various image generators and the real dataset. It may be interesting to note that different configurations of the same generator family tend to fall into a cluster or onto a smooth curve, indicating intra-family similarities.

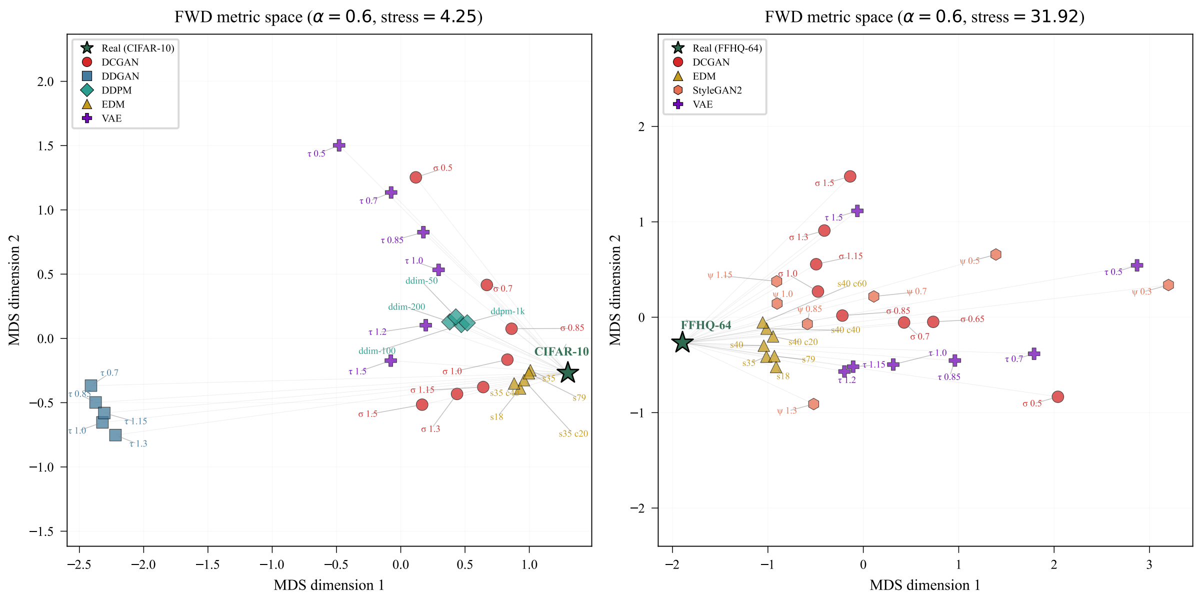


Fig. 2. 2D MDS embedding of the FWD metric space at $\alpha = 0.6$, with the generators trained on CIFAR-10 (left) and FFHQ-64 $\times$ 64 (right).

TABLE II
SUMMARIZED RESULTS: MDS STRESS, TOPOLOGICAL PERSISTENCE, AND MANN-WHITNEY U (MWU) TEST RESULTS FOR CYCLE VS. NON-CYCLE PRDC PROFILES (CIFAR-10 AND FFHQ) FOR GDM ($\alpha = 0.6$). { †p < 0.05}

Metric	**Comments on Total Persistence of H_1**	**Stress ↓**		**PRDC**	**CIFAR-10 (MWU)**			**FFHQ (MWU)**		
		CIFAR-10	FFHQ		U_{crit}	U-Stat	p-value	U_{crit}	U-Stat	p-value
FWD	**Smooth with α Significant features**	<u>4.2545</u>	<u>31.9157</u>	Precision Recall Density Coverage	38	67 **24** **37** **26**	0.65169 **0.0062**† **0.039**† **0.0085**†	59	104 **59** 88 **59**	1.0000 **0.0467**† 0.4715 **0.0467**†
FID	Sharp changes, ≈ 0 for some α	59.2685	200.5973	Precision Recall Density Coverage	47	69 **47** 73 52	0.361 **0.045**† 0.474 0.080	59	132 73 80 73	0.1019 0.1804 0.3028 0.1804
FD-DINOv2	Sharp changes, ≈ 0 for some α	2664.0697	2813.6817	Precision Recall Density Coverage	38	77 114 82 113	0.97882 **0.046**† 0.7491 0.0525	59	124 132 142 132	0.2821 0.2467 0.1112 0.2467
CMMD	Lacks significant H_1 features at any α	**0.6247**	**0.5434**	–	–	–	–	–	–	–

C. Downstream Performance Metrics

The topological features detected by FWD are not merely geometric curiosities, but they predict downstream generation quality as measured by Precision, Recall, Density, and Coverage (PRDC) [29], [30], computed in InceptionV3's embedding space. We analyze the rankings of PRDC metrics for H_1 cycle participants against non-cycle participants through the Mann-Whitney U Test (MWU) [31]. We report the U_{stat}, p-value, and U_{crit} at significance value 0.05 (standard) in Table II. The decision rules is that both distributions are indistinguishable if $U_{stat} > U_{crit}$ or p-value > 0.05. Table II shows that the H_1 features obtained by FWD consistently encode the Recall and Density metrics on CIFAR-10 and FFHQ-dataset. Neither FID nor FD-DINOv2 show correlation with the PRDC metrics. Due to inconclusive H_1 features of CMMD, it was omitted from the test.

VI. CONCLUSION

We introduced a TDA-based framework for evaluating the reliability of image generation metrics. By constructing a Vietoris–Rips filtration over a Generator Distance Metric that combines ground-truth fidelity and cross-generator dissimilarity, we showed that the wavelet-based FWD yields stable topological features that backbone-based metrics (FID, FD-DINOv2, CMMD) fail to produce. These topological features predict backbone-robust downstream metrics (recall, coverage) on both CIFAR-10 and FFHQ-64×64 using the Mann–Whitney U Test. These findings suggest that FWD-based TDA serves as a useful diagnostic tool for qualitative analysis of generative model families, warranting investigation on larger-scale datasets and alternative generators.

Acknowledgment

We acknowledge and are grateful for the support of the Anusandhan National Research Foundation (ANRF) - Partnerships for Accelerated Innovation and Research (PAIR) Project at IIT Bombay (RD/0125-ANRFPVG-001) towards this research project.

References

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

[2] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion gans," in *International Conference on Learning Representations*, 2022.

[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 8107–8116, Computer Vision Foundation / IEEE, 2020.

[4] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

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

[6] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. NeurIPS*, 2022.

[7] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

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

[9] G. Stein, J. C. Cresswell, R. Hosseinzadeh, Y. Sui, B. L. Ross, V. Villecroze, Z. Liu, A. L. Caterini, J. E. T. Taylor, and G. Loaiza-Ganem, "Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[10] M. J. Chong and D. Forsyth, "Effectively unbiased fid and inception score and where to find them," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6069–6078, 2020.

[11] L. Veeramacheneni, M. Wolter, H. Kuehne, and J. Gall, "Fréchet wavelet distance: A domain-agnostic metric for image generation," in *The Thirteenth International Conference on Learning Representations*, 2025. Also arXiv:2312.15289.

[12] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, "Rethinking fid: Towards a better evaluation metric for image generation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, 2024.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 18–24 Jul 2021.

[14] R. Ballester, C. Casacuberta, and S. Escalera, "A survey on topological data analysis for neural networks," *arXiv preprint arXiv:2312.05840*, 2024.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

[16] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024. Featured Certification.

[17] Y. Ding, L. Wang, and B. Gong, "Analyzing deep neural network's transferability via fréchet distance," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3931–3940, 2021.

[18] T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, and A. G. Wilson, "Loss surfaces, mode connectivity, and fast ensembling of dnns," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, (Red Hook, NY, USA), p. 8803–8812, Curran Associates Inc., 2018.

[19] F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht, "Essentially no barriers in neural network energy landscape," in *International Conference on Machine Learning (ICML)*, pp. 1309–1318, PMLR, 2018.

[20] R. Kuditipudi, X. Wang, H. Lee, Y. Zhang, Z. Li, W. Hu, R. Ge, and S. Arora, "Explaining landscape connectivity of low-cost solutions for multilayer nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[21] S. Horoi, J. Huang, B. Rieck, G. Lajoie, G. Wolf, and S. Krishnaswamy, "Exploring the geometry and topology of neural network loss landscapes," 2022.

[22] A. I. Humayun, I. Amara, C. Vasconcelos, D. Ramachandran, C. Schumann, J. He, K. Heller, G. Farnadi, N. Rostamzadeh, and M. Havaei, "What secrets do your manifolds hold? understanding the local geometry of generative models," 2025.

[23] H. Edelsbrunner and J. Harer, *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[24] C. Tralie, N. Saul, and R. Bar-On, "Ripser.py: A lean persistent homology library for python," *Journal of Open Source Software*, vol. 3, no. 29, p. 925, 2018.

[25] A. Krizhevsky, "Learning multiple layers of features from tiny images," tech. rep., University of Toronto, 2009.

[26] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, p. 4217–4228, Dec. 2021.

[27] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[28] M. A. A. Cox and T. F. Cox, *Multidimensional Scaling*, pp. 315–347. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[29] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," *CoRR*, vol. abs/1904.06991, 2019.

[30] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, "Reliable fidelity and diversity metrics for generative models," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185, PMLR, 13–18 Jul 2020.

[31] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.