Évaluation des générateurs d'images à partir de peu d'exemples : calculer le FID avec 10 fois moins d'images, c'est possible

Nicolas AUDEBERT[1, 2] Arnaud BRELOY[2]

[1]Univ. Gustave Eiffel, ENSG, IGN, LASTIG, F-94160 Saint-Mandé, France

[2]Conservatoire national des arts et métiers, CEDRIC, EA4629, F-75141 Paris, France

Résumé – La distance Inception de Fréchet (*Fréchet Inception Distance* ou FID) est une métrique standard pour l'évaluation des modèles génératifs images. Construite sur la distance de Wasserstein, le FID mesure l'écart entre la distribution des images générées et celle des images réelles. Cependant, le FID nécessite une grande quantité d'images réelles pour être calculé de façon fiable. Cela rend son utilisation peu adaptée à l'évaluation de modèles génératifs entraînés sur peu de données. Dans cet article, nous proposons un remplacement au FID, strictement compatible, mais qui produit une estimation fiable de la distance de Wasserstein avec quelques milliers d'images seulement. Plus précisément, nous remplaçons l'estimateur classique de la distance de Wasserstein par une variante issue de la théorie des matrices aléatoires (RMT). Nous montrons que le RMT FID est plus robuste que le FID classique au travers de l'évaluation des performances de StyleGAN2 sur deux jeux de données : CIFAR-10 et AFHQ-Cats.

Abstract – Fréchet Inception Distance (FID) is a standard evaluation metric for image generation models. Built upon the Wasserstein distance, FID measures the gap between the distribution of generated images and the distribution of real images. However, computing the FID requires a large number of real images, otherwise its fiability drops sharply. This makes evaluating generative models with FID unsuitable for low-data configurations. In this work, we introduce a replacement to FID that is strictly compatible, but that can be used to produce an estimation of the Wasserstein distance even with only a few thousands real images. We replace the classical Wasserstein distance estimator by a variant from random matrix theory (RMT). We show that RMT FID is more robust than classical FID by evaluating the performances of StyleGAN2 on two datasets: CIFAR-10 and AFHQ-Cats.

1 Introduction

La distance *Inception* de Fréchet (FID, de l'anglais *Fréchet Inception Distance*) [4] est une métrique communément employée pour l'évaluation des générateurs d'image. Elle représente la distance entre la distribution des images générées et une distribution d'images de référence. Le FID se construit à partir de la distance de Wasserstein entre deux distributions μ et ν de $\mathbb{R}^n$, définie par :

$$d_W(\mu,\nu)^2 = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^n\times\mathbb{R}^n} \|x_1 - x_2\|^2 \,\mathrm{d}\gamma(x_1,x_2)\ . \quad (1)$$

Concrètement, soit $g : z \in \mathbb{R}^d \to x \in \mathbb{R}^{W\times H\times 3}$ un générateur d'image, prenant en entrée un vecteur de bruit aléatoire réel de dimension d et produisant une image couleur de dimensions $W \times H$. Le FID évalue la qualité visuelle des images générées en les comparant aux images réelles dans l'espace latent d'un réseau de neurones convolutif. Soit f un réseau profond pré-entraîné, typiquement le modèle Inception v3 [16] entraîné sur ImageNet. f agit comme extracteur de caractéristiques, c'est-à-dire une fonction $f : x \in \mathbb{R}^{W\times H\times 3} \to w \in \mathbb{R}^p$ qui prend une image et renvoie un vecteur de caractéristiques w de dimension p. Le FID compare alors les distributions des caractéristiques obtenues sur les images réelles $\mu = \{w = f(x) | x \in \mathcal{D}_{\text{réelles}}\}$ issues du jeu de données qui a servi à l''entraînement du générateur g, et des caractéristiques obtenues sur les images générées $\nu = \{w' = f(g(z)) | z \sim \mathcal{N}(0,1)\}$. On suppose que les caractéristiques w suivent une distribution gaussienne, ce qui permet de remplacer Équation (1) par la formule analytique de la distance de Wasserstein entre deux gaussiennes[1] :

$$d_W(\mu,\nu)^2 = \|m_1 - m_2\|_2^2 + \mathrm{tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right) \quad (2)$$

avec $\mu = \mathcal{N}(m_1,\Sigma_2)$ et $\nu = \mathcal{N}(m_2,\Sigma_2)$, et ou $\Sigma^{\frac{1}{2}}$ est l'unique racine dans $\mathcal{S}_p^+$ (ensemble des matrices symétriques semi-définies positives) de la matrice $\Sigma \in \mathcal{S}_p^+$. En pratique, m et Σ sont remplacées par leurs estimations empiriques sur les jeux de données d'images réelles et générées.

L'objectif des générateurs d'image est ainsi d'atteindre le FID le plus faible possible, indiquant que les images générées suivent la même distribution que les images réelles. Cette métrique est devenue un indicateur incontournable du progrès de la qualité en synthèse d'images par modèles profonds [8, 9, 5, 12]. Bien que la version initiale repose sur un modèle pré-entraîné pour la classification d'*images*, le FID a de multiples déclinaisons : distance audio de Fréchet (FAD) [10], distance vidéo de Fréchet (FVD) [18] et même distance ChemNet de Fréchet (FCD) [14] pour la génération de molécules médicales. Il « suffit » de remplacer f par un extracteur de caractéristiques approprié afin de construire une nouvelle métrique.

Néanmoins, le FID n'est pas exempt de limitations. Sa reproductibilité est parfois difficile car son calcul dépend des poids du modèle Inception, des différences d'implémentations matérielles et du type d'interpolation appliqué pour redimensionner les images [13]. Toutefois, le choix du classifieur Inception v3 a des conséquences plus lourdes que les aspects techniques.

[1]C'est cette définition qui vaut au FID le nom de distance de Fréchet.

Le pré-entraînement sur ImageNet induit des invariances dans l'espace latent qui rend le FID insensible à certaines distorsions incluses comme augmentation de données, ce qui représente un biais intrinsèque au modèle [7]. KYNKÄÄNNIEMI et al. [11] recommande ainsi de remplacer Inception par un modèle auto-supervisé, comme CLIP [15], dont les caractéristiques seraient plus génériques. Cependant, le choix de la distance de Wasserstein comme métrique sous-jacente pose elle-même différents problèmes. À l'origine, HEUSEL et al. [4] proposent de calculer le FID avec au moins 50 000 images réelles et 50 000 images générées. KARRAS et al. [9] (section 4.3) notent par la suite que le FID présente un biais élevé dans des régimes à « faible quantité de données », de l'ordre de quelques milliers d'images. Puisqu'il est possible de générer une infinité d'images synthétiques, le nombre d'images réelles est donc le facteur limitant. Pour certains jeux de données, il peut être de l'ordre de 1000 images, et donc même inférieur à la dimension des caractéristiques extraites par f (typiquement, $p = 2048$ pour Inception v3). En pratique, cela conduit le FID a être estimé entre μ une faible quantité d'images réelles (quelques milliers) et ν une très grande quantité d'images générées pour compenser (quelques dizaines de milliers). Le FID_∞ a ainsi été proposé comme solution pour débiaiser le FID [2] lorsqu'il est estimé sur peu d'images, au prix d'un coût plus élevé introduit par le calcul de plusieurs estimations sur le même jeu de données. Ces difficultés ont mené certains auteurs à proposer des métriques alternatives ne dépendant plus de la distance de Wasserstein, comme le KID [1] ou CMMD [6].

Dans cet article, nous montrons que ces inquiétudes ne sont pas fondées et qu'il est possible d'estimer le FID sur des jeux de données de taille modeste sans sacrifier la précision. En particulier, nous nous appuyons sur des travaux en estimation de distances entre distributions qui utilisent la théorie des matrices aléatoires, qui a engendré de nombreuses façons d'approximer des distances entre matrices de covariances [3]. Beaucoup de distances entre distributions font partie de cette catégorie, comme la distance de Fisher ou la divergence de Kullback-Leibler. Comme le montre l'Équation (2), la distance de Wasserstein en fait également partie. TIOMOKO et COUILLET [17] ont ainsi introduit un estimateur spécifique afin d'estimer la distance de Wasserstein entre gaussiennes. Cet estimateur converge nettement plus rapidement que l'estimateur classique de l'Équation (2), en particulier dans le régime où la quantité de données n est proche de leur dimension p. En pratique, pour Inception, les caractéristiques extraites sont de dimension $p = 2048$. Nous allons ainsi montrer qu'il est possible d'estimer le FID de façon robuste avec quelques milliers d'exemples réels et autant d'exemples générés.

2 Estimateur RMT FID

TIOMOKO et COUILLET [17] ont introduit un estimateur RMT Wasserstein de la distance de Wasserstein entre deux distributions μ et ν de $\mathbb{R}^p$. On suppose disposer de n observations pour chaque distribution, c'est-à-dire n images réelles et n images générées. Soient $\hat{m}_1$ et $\hat{m}_2$ les moyennes estimées de μ et ν et $\hat{\Sigma}_1$ et $\hat{\Sigma}_2$ leurs matrices de covariances estimées sur les n exemples. On note λ le vecteur colonne des valeurs propres $\lambda_1 \leq \ldots \leq \lambda_p$ – par ordre croissant – du produit $\hat{\Sigma}_1\hat{\Sigma}_2$. Soit $\Lambda = \text{diag}_{1\leq i\leq p}(\lambda_i)$ la matrice diagonale des valeurs propres λ_i. Considérons $\xi_1 \leq \ldots \leq \xi_p$ les valeurs propres croissantes de la matrice $\tilde{\Lambda} = \Lambda - \frac{1}{n}\sqrt{\lambda}\sqrt{\lambda}^t$. Alors, par [17],

$$\hat{d}_W(\mu,\nu)^2 = \|m_1 - m_2\|_2^2 + \frac{2n}{p}\sum_{j=1}^{p}\left(\sqrt{\lambda_j} - \sqrt{\xi_j}\right) \quad (3)$$

est un estimateur de la distance de Wasserstein entre μ et ν.

Cet estimateur suppose que $D_{\text{réel}}$ et $D_{\text{généré}}$ sont de même cardinal, c'est-à-dire qu'il y a autant d'images générées que d'images réelles. En réalité, un estimateur existe pour le cas $n_1 \neq n_2$, cependant celui-ci n'a pas d'expression close [17] : son calcul nécessite une intégration numérique dont la stabilité n'est pas garantie. Nous nous concentrons donc sur le cas $n_1 = n_2$ et nous allons voir qu'il est largement suffisant.

3 Résultats expérimentaux

Le code nécessaire à la reproduction de ces expériences est disponible sur `https://github.com/nshaud/rmtfid`.

3.1 Données synthétiques

Dans un premier temps, nous pouvons comparer le comportement asymptotique des deux estimateurs. TIOMOKO et COUILLET [17] montrent empiriquement que le RMT Wasserstein converge nettement plus rapidement, cependant les expériences se contentent de dimensions relativement faibles, jusqu'à $p = 512$. Or, nous l'avons vu, le FID utilise les caractéristiques de Inception v3, et donc $p = 2048$. Combien faut-il de données en pratique pour que les estimateurs convergent ?

La Figure 1 illustre les estimations de la distance de Wasserstein entre deux gaussiennes de dimension 100, centrées et de covariances identiques $\Sigma_1 = \Sigma_2 = \Sigma$ une matrice de Toeplitz avec $\Sigma[i,j] = 0.2^{|i-j|}$. La distance de Wasserstein théorique entre ces deux distributions est nulle. Nous générons deux jeux d'observations de taille $n_{\text{max}} = 200\,000$ et comparons le comportement selon n de trois estimateurs :

- $d_W(n)$, l'estimateur classique en estimant m et Σ à partir d'autant d'exemples dans chaque distribution,
- $d_W(n_{\text{max}}, n)$, l'estimateur classique asymétrique, c'est-à-dire considérant 200 000 observations pour l'une des distributions et d'un nombre variable pour l'autre,
- $d_{\text{RMT}}(n)$ l'estimateur RMT avec autant d'exemples dans chaque distribution.

L'estimateur classique exhibe une convergence lente vers la distance réelle. Même avec 200 000 exemples (soit $2000\times$ plus que de dimensions), sa valeur est significativement non nulle ($\approx 0{,}05$), tandis que l'estimateur RMT est $< 0{,}05$ dès quelques centaines d'observations et d'erreur négligeable à partir de quelques milliers d'exemples.

Toutefois, le FID est calculé dans un espace de dimension $p = 2048$. Qu'en est-il dans cette configuration « haute dimension » ? La Figure 2 montre que la situation est pire. La valeur communément acceptée de 50 000 commet une erreur d'approximation du FID significative. En comparaison, l'estimateur RMT est extrêmement précis, même avec un nombre d'exemples variant entre 1 à 10 fois la dimension des données.

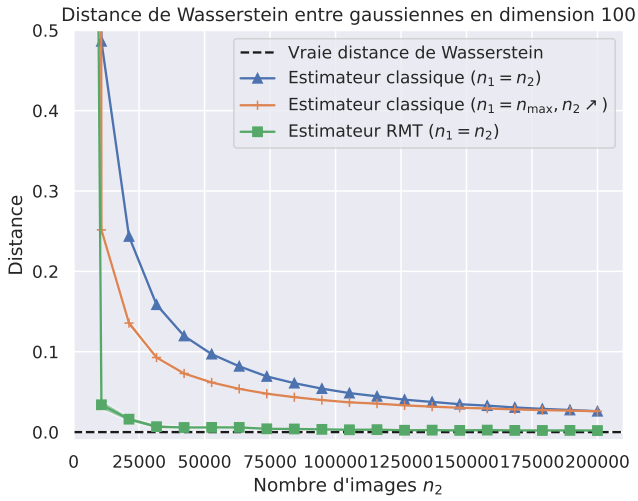

FIGURE 1 : Estimations de la distance de Wasserstein entre gaussiennes ($p = 100$) en fonction du nombre d'observations.

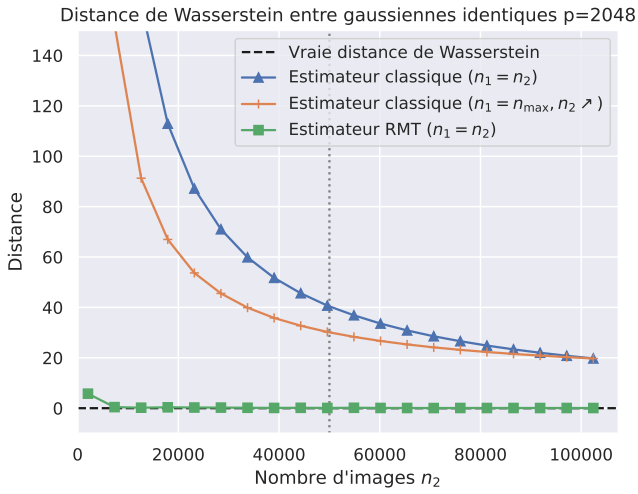

FIGURE 2 : Estimations de la distance de Wasserstein entre gaussiennes ($p = 2048$) en fonction du nombre d'observations.

3.2 Images générées

La comparaison empirique des estimateurs de la distance de Wasserstein est inquiétante sur des données gaussiennes synthétiques, mais le FID est calculé sur les caractéristiques des images réelles et générées, extraites à l'aide du modèle Inception v3. Le problème se pose-t-il encore dans cette configuration ? Pour le vérifier, nous calculons le FID du modèle StyleGAN2-ADA [9]. La distribution de référence est constituée des 50 000 images du jeu d'apprentissage de CIFAR-10, que nous comparons à 50 000 images générées par StyleGAN2-ADA. Les auteurs indiquent un FID de 2,42 avec l'estimateur classique FID@50k, ce que nous reproduisons sans problème (2,49). Comme précédemment, la Figure 3 illustre les estimations du FID en faisant varier le nombre d'images utilisé.

Sur ces données réelles, les estimations sont légèrement plus bruitées mais le constat est sans appel. Le FID à 50 000 est vraisemblablement surestimé. Par ailleurs, l'estimateur RMT permet d'obtenir une valeur stable dès ≈ 5000 exemples, soit $10\times$ moins que ce qui est habituellement préconisé ! Autrement dit, rien ne justifie donc le protocole « FID@50k » de la littérature : le calcul du FID pourrait être réalisé sur quelques milliers d'images générées, réduisant ainsi les temps de calcul et par conséquent les coûts économiques et écologiques de l'évaluation des modèles génératifs.

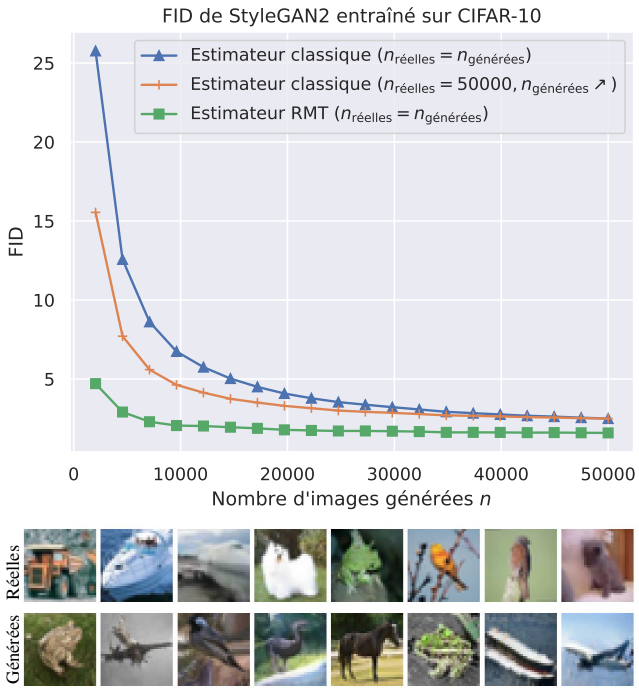

FIGURE 3 : Estimations du FID de StyleGAN2-ADA [8] entraîné sur CIFAR-10.

Cependant, l'intérêt du RMT FID ne réside pas seulement dans l'économie de la génération d'images pour l'évaluation. En effet, de nombreux jeux de données utilisés pour entraîner les modèles génératifs disposent de nettement moins que 50 000 exemples d'apprentissage[2]. Il existe de nombreux exemples de tels jeux de données mais un des plus sympathiques est le jeu de données AFHQ-Cats, constitué de 5153 images de chats. Comme pointé par KARRAS et al. [9], le FID est assez peu représentatif sur ce jeu de données à cause du nombre limité d'images. La Figure 4 illustre les estimations du FID de StyleGAN2-ADA sur AFHQ-Cats avec les différents estimateurs. À nouveau, il est vraisemblable que l'estimation classique soit nettement surestimée. Dans ce régime, le nombre de données disponibles pour le calcul du FID est limité, de $1\times$ à $2\times$ la dimension des caractéristiques. Il ne semble ainsi pas raisonnable de comparer différents générateurs d'images entraînés sur AFHQ-Cats avec l'estimateur FID classique. Cependant, l'estimateur RMT converge assez rapidement et donne un résultat stable même avec $\approx$5000 images. Dans de telles configurations, le RMT FID est donc un substitut de choix pour évaluer des modèles génératifs image.

4 Conclusion

Cet article introduit une version améliorée de la *Fréchet Inception Distance*, qui s'appuie sur un estimateur de la distance

[2]Par ailleurs, un protocole rigoureux impliquerait que les images utilisées pour l'évaluation fassent partie d'un jeu de test écarté de l'entraînement du modèle. Ce protocole est rare dans la littérature, justement car il réduit encore plus le nombre d'images disponibles pour le calcul du FID.

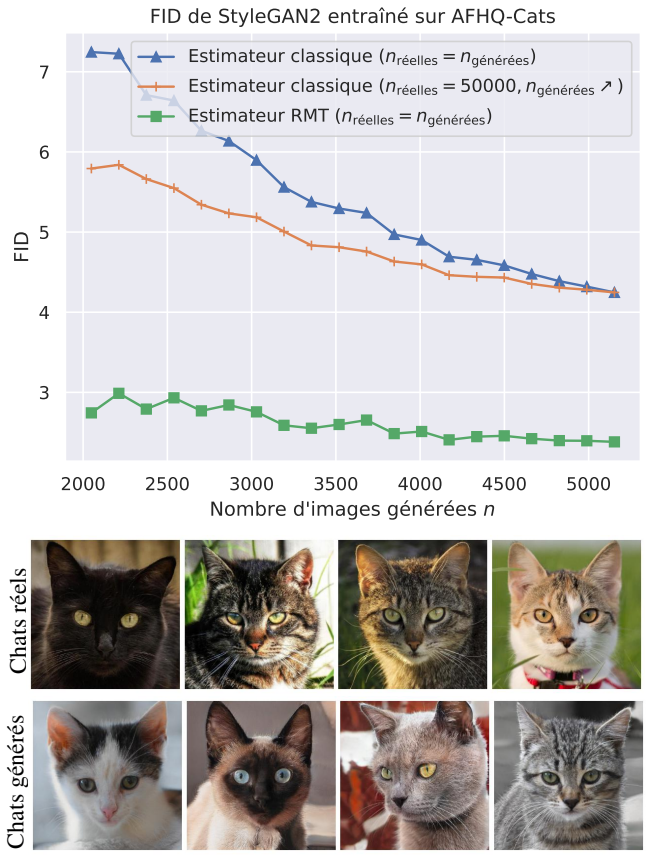

FIGURE 4 : Estimations du FID de StyleGAN2-ADA [8] entraîné sur AFHQ-Cats.

de Wasserstein plus efficace que la formule usuelle. Nous montrons que le RMT FID remplace avantageusement le FID classique dans un régime où peu d'images réelles sont disponibles pour évaluer le modèle génératif. En outre, nos résultats semblent indiquer que le protocole habituel impliquant d'évaluer le FID sur 50 000 images réelles et 50 000 images générées est arbitraire et que le biais est l'estimateur classique demeure élevé dans cette configuration. Le RMT FID ouvre la porte à une évaluation plus robuste des modèles génératifs tout en nécessitant moins de ressources. D'autres améliorations restent toutefois possibles. Le RMT FID ne permet pas actuellement de traiter le cas $n_1 \neq n_2$, or il serait envisageable de générer une grande quantité d'images afin d'améliorer la précision du RMT FID lorsque très peu de données réelles sont disponibles. Par ailleurs, le RMT FID, comme le FID, nécessite de passer par une estimation des matrices de covariance empiriques, ce qui est malaisé lorsque $n < p$. Dans le cas d'Inception v3, cela impose de disposer d'au moins 2048 images réelles. Enfin, nos travaux partagent le même angle mort que nos prédécesseurs : l'hypothèse de gaussianité des caractéristiques utilisées pour le calcul du FID. Si cette hypothèse peut se justifier dans le cas où beaucoup d'images sont disponibles, il est peu probable que les caractéristiques d'Inception v3 soient naturellement distribuées selon une gaussienne. Cette non-gaussianité est susceptible de remettre en question la pertinence de la distance de Wasserstein comme métrique de qualité perceptuelle. D'autres distances entre distributions pourraient s'économiser cette hypothèse et se révéler plus pertinentes pour l'évaluation des modèles génératifs.

Références

[1] Mikołaj BIŃKOWSKI et al. "Demystifying MMD GANs". In : International Conference on Learning Representations. 2018.

[2] Min Jin CHONG et 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. 2020.

[3] Romain COUILLET et al. "Random Matrix-Improved Estimation of Covariance Matrix Distances". In : *Journal of Multivariate Analysis* 174 (2019).

[4] Martin HEUSEL et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In : *Advances in Neural Information Processing Systems*. T. 30. 2017.

[5] Jonathan HO, Ajay JAIN et Pieter ABBEEL. "Denoising Diffusion Probabilistic Models". In : *Advances in Neural Information Processing Systems*. T. 33. 2020.

[6] Sadeep JAYASUMANA et al. "Rethinking FID : Towards a Better Evaluation Metric for Image Generation". In : Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[7] Steffen JUNG et Margret KEUPER. "Internalized Biases in Fréchet Inception Distance". In : NeurIPS 2021 Workshop on Distribution Shifts : Connecting Methods and Applications. 2021.

[8] Tero KARRAS et al. "Analyzing and Improving the Image Quality of StyleGAN". In : Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[9] Tero KARRAS et al. "Training Generative Adversarial Networks with Limited Data". In : *Advances in Neural Information Processing Systems*. T. 33. 2020.

[10] Kevin KILGOUR et al. "Fréchet Audio Distance : A Reference-Free Metric for Evaluating Music Enhancement Algorithms". In : Proceedings of Interspeech. 2019.

[11] Tuomas KYNKÄÄNNIEMI et al. "The Role of ImageNet Classes in Fréchet Inception Distance". In : The Eleventh International Conference on Learning Representations. 2022.

[12] Yaron LIPMAN et al. "Flow Matching for Generative Modeling". In : The Eleventh International Conference on Learning Representations. 2023.

[13] Gaurav PARMAR, Richard ZHANG et 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. 2022.

[14] Kristina PREUER et al. "Fréchet ChemNet Distance : A Metric for Generative Models for Molecules in Drug Discovery". In : *Journal of Chemical Information and Modeling* 58.9 (2018).

[15] Alec RADFORD et al. "Learning Transferable Visual Models From Natural Language Supervision". In : International Conference on Machine Learning. PMLR, 2021.

[16] Christian SZEGEDY et al. "Rethinking the Inception Architecture for Computer Vision". In : IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[17] Malik TIOMOKO et Romain COUILLET. "Random Matrix-Improved Estimation of the Wasserstein Distance between Two Centered Gaussian Distributions". In : 2019 27th European Signal Processing Conference (EUSIPCO). 2019.

[18] Thomas UNTERTHINER et al. "FVD : A New Metric for Video Generation". In : DeepGenStruct (International Conference on Learning Representations Workshop). 2019.