Enhancing plausibility evaluation for generated designs with denoising autoencoder

Fan, J.; Trigui, A.; Bäck T.H.W.; Wang, H.; Leonardis, A.; Ricci, E.; ... ; Varol, G.

Enhancing Plausibility Evaluation for Generated Designs with Denoising Autoencoder

Jiajie Fan[1,2(✉)], Amal Trigui[1], Thomas Bäck[2], and Hao Wang[2]

[1] BMW Group, Bremer Str. 6, 80788 Munich, Germany
{jiajie.fan,amal.trigui}@bmw.de
[2] LIACS, Leiden University, Niels Bohrweg 1, 2333 Leiden, CA, The Netherlands
{t.h.w.baeck,h.wang}@liacs.leidenuniv.nl

Abstract. A great interest has arisen in using Deep Generative Models (DGM) for generative design. When assessing the quality of the generated designs, human designers focus more on structural plausibility, *e.g.*, no missing component, rather than visual artifacts, *e.g.*, noises or blurriness. Meanwhile, commonly used metrics such as Fréchet Inception Distance (FID) may not evaluate accurately because they are sensitive to visual artifacts and tolerant to semantic errors. As such, FID might not be suitable to assess the performance of DGMs for a generative design task. In this work, we propose to encode the to-be-evaluated images with a Denoising Autoencoder (DAE) and measure the distribution distance in the resulting latent space. Hereby, we design a novel metric Fréchet Denoised Distance (FDD). We experimentally test our FDD, FID and other state-of-the-art metrics on multiple datasets, *e.g.*, BIKED, Seeing3DChairs, FFHQ and ImageNet. Our FDD can effectively detect implausible structures and is more consistent with structural inspections by human experts. Our source code is publicly available at https://github.com/jiajie96/FDD_pytorch.

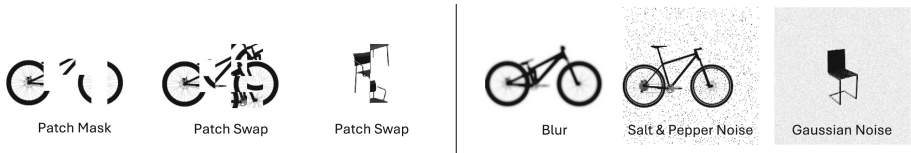
Keywords: Evaluation metric · Generative design · Structural plausibility

1 Introduction

Following the swift development of Deep Generative Models (DGMs) in general image generation tasks [17,22,26,37], a great interest has arisen in using DGMs to enable generative design [14,15,34,39], where DGMs are able to create innovative designs based on specific input requirements provided by users. In this particular domain, design data is responsible for representing the design object with structural and geometric patterns, which are required to be recognizable and plausible. In order to rank models during the development of generative models, recent

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-73229-4_6.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15136, pp. 88–105, 2025.
https://doi.org/10.1007/978-3-031-73229-4_6



(a) Better scored by the SOTA metrics (*e.g.*, FID, KID and $FD_{DINO\text{-}V2}$)

(b) Better scored by human designers and our FDD

Fig. 1. From which side (*left* or *right*) are the design images more plausible? (a) Structural implausibility; (b) visual artifacts. Recent work [2,16,20] discovers that the SOTA metrics (FID, KID and $FD_{DINO\text{-}V2}$) tend to penalize visual artifacts more than structural implausibility, which matters more to the human designers. In contrast, our FDD consists better with human designers and is able to focus on shapes.

works rely on a subjective evaluation [14,32], where human experts apply an established set of criteria to manually assess a significant quantity of generated data. This evaluation method yields reliable results, serving as "ground truth" for model ranking, but it is time-consuming and hard to reproduce [32]. Hence, for developing DGMs for design generation, it is necessary to have an automated metric, which is able to reliably quantify the goodness of the target DGM.

Meanwhile, the evaluation of generated images is still an unsolved challenge among other general tasks in the DGM domain [3,4,33]. DGM developers [7, 24,26] are heavily relying on the Fréchet Inception Distance (FID) [21] metric, which extracts latent features from real and generated images with an Inception-V3 [44] model pre-trained on ImageNet [11] respectively and then quantifies their difference using Fréchet Distance as the final FID score. As the primary metric in the DGM field, FID is able to measure the fidelity and diversity and present them in a single value. However, a lot of studies [6,28,43] disclose that FID does not always align with human evaluation and claim that this limitation is due to the reliance on the pre-trained Inception-V3 model. Hence, novel metrics are delivered by replacing the Inception-V3 model by other backbone networks, *e.g.*, Clip, [36] VQ-VAE [45] and DINOv2 [35], *etc.*. According to the most recent work of Stein *et al.* [43], where they compared 17 metrics using encoders from 9 various networks, $FD_{DINO\text{-}V2}$ has the most reliable performance in terms of consistency with human judgment in their experiments.

On the other hand, recent works have pointed out that the Inception-V3 model and Inception-powered metrics perform poorly on shapes [2,14,16,20]. Our work investigates this finding and observes that the state-of-the-art (SOTA) metrics generally suffer from this issue: they are sensitive to visual artifacts like noises, yet they have a high tolerance towards semantic failures, *e.g.*, part missing in a bicycle, as illustrated in Fig. 1. Besides, human experts are able to recognize the same structural representation of the observed design image regardless of minor noise and they tend to penalize the evaluation based on the implausibility of the design more, rather than the presence of visual artifacts [27,29]. Motivated by this, our work aims to create a novel metric for generative design that is robust to visual corruption of the observed images and biased towards the design plausibility.

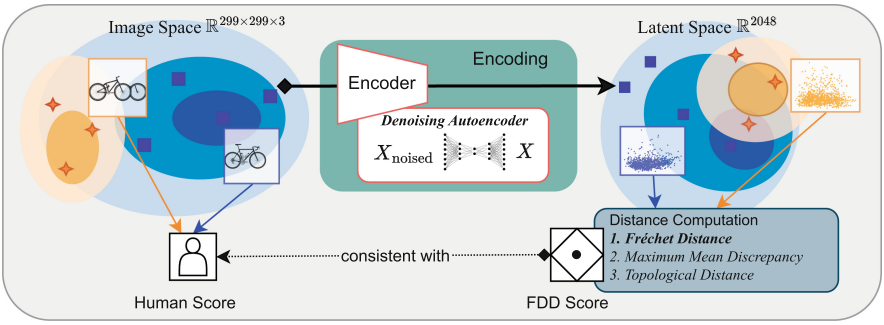


Fig. 2. Plausibility evaluation using Fréchet Denoised Distance. *Blue area and stars* visualizing the distribution and samples of real data in the image space and in the DAE-encoded latent space; *Orange area and stars* illustrating the distribution and samples of generated data in the image space and in the DAE-encoded latent space. (Color figure online)

Finally, we propose the Fréchet Denoised Distance (FDD) by replacing the Inception-V3 model within the FID framework with a Denoising Autoencoder (DAE) [46] that has been also pre-trained on ImageNet dataset and capable of encoding images into latent features with an Inception-comparable dimension of $\mathbb{R}^{2048}$. The DAE is able to observe the same structural representation in the image regardless of the noisy disturbances, which can be utilized as a strong method to extract the structural feature from the noisy input. Our work compares our FDD with other SOTA metrics, *e.g.*, FID, $\text{FD}_{\text{DINO-V2}}$ and Topology Distance (TD) [23] (since their results show a similar bias to our intention), based on the following experiments: (1) sensitivity test over visual artifacts and structural failures; (2) consistency test with increasing disturbances; (3) consistency test with human judgment in model ranking. As a result, our FDD has the most stable performance among all the experiments. In order to explain the performance of FDD, we visualize the "focus" of our FDD metric compared to the FID using a GradCAM [28,41,43] test, hereby showcasing that the DAE model has a better assessment regarding the requirements of human designers. In addition, our work explores the potential for further improvement of DAE-based metrics, where we build-upon concepts from existing works, *i.e.*, KID [5], TD [23] and training the network on structural images [16], to design new DAE-based metrics, *i.e.*, Kernel Denoised Distance (KDD), Topology Denoised Distance (TDD), and FDD (·), respectively. We test these metrics on BIKED and the results show that these DAE-based metrics are highly correlated and FDD performs relatively best.

2 Related Work

Unlike the swift development in the field of Deep Generative Models (DGMs) [12, 17,26], accurately ranking generative models remains an unresolved challenge [3,4,33]. Humans are able to give the ground-truth evaluation in assessing a

limited number of generated images, but quantifying the performance of a DGM requires an automated evaluation method [43]. Overcoming the flaws of previous metrics, *e.g.*, SSIM [49], LPIPS [48] and IS [40], currently most reported evaluation methods, *e.g.*, Fréchet Inception Distance (FID) [21] and Kernel Inception Distance (KID) [5], have largely addressed the challenge of automated evaluation and are employed as the primary metric for model ranking in the field of DGM [7,24,26]. They leverage a two-step procedure: encode real and generated images into latent features in a lower-dimensional space with a representation extractor and then use a distance critic to quantify the difference between their features. Both FID and KID utilize the Inception-V3 [44] model pre-trained on ImageNet, which has a 2048-dimensional latent space. Regarding the measurement of latent distance: FID fits the Inception features from real and generated images into a multivariate Gaussian before computing the Fréchet Distance (also known as the Wasserstein-2 distance) between them; whereas KID [5] uses the squared Maximum Mean Discrepancy (MMD) [18] with a polynomial kernel [4].

Concerns about the over-reliance on the Inception-V3 model have been raised and researchers claim that an ImageNet [11] classifier like the Inception-V3 model brings a significant bias to the evaluation with FID [33,35]. Furthermore, FID is proved to be vulnerable to manipulation [28], especially when there exists a significant domain discrepancy between the data set of interest such as BIKED [38] and ImageNet [11]. Consequently, the results measured by FID often show a poor correlation with human judgments. Similarly, KID [5] encounters the same issue as it also leverages the pre-trained Inception-V3 model. Most recently, in order to find a perceptual representation space superior to the inception manifold, Stein *et al.* [43] studied 17 metrics with 9 different encoders (*e.g.*, CLIP [36], SwAV [9] and DINOv2 [35]). Their finding concludes that $\text{FD}_{\text{DINO-V2}}$ [43] demonstrates the most reliable performance over various perspectives, *e.g.*, fidelity, diversity, rarity, and memorization of generative models. Previous works [2,16,20] shed light on the role played by the image attributes, *e.g.*, edges, shapes, textures, and colors in various computer vision tasks, *e.g.*, classification and segmentation.

Recent studies have introduced autoencoder-based metrics for evaluation purposes: for instance, Buzuti *et al.* [8] leveraged the VQ-VAE [45] and showed that their unsupervised model-based metric Fréchet AutoEncoder Distance (FAED) outperforms FID in terms of consistency with increasing disturbance when evaluating on human and animal faces, *i.e.*, CelebA HQ [31], FFHQ [26], and AFHQ [10]. By cross-comparing their measured values among various types of disturbance, their FAED noticeably penalizes visual artifacts more severely than structural implausibility with comparable intensity. This may still lead to unfair comparisons of DGMs for design synthesis, where human experts prefer to use shape information for assessment [27,29]. Meanwhile, Horak *et al.* [23] proposed a more competitive shape-based evaluation metric by investigating the topological characteristics of potential flow shapes and proposed topological distance (TD) as a complementary metric for FID. Thus, we choose TD as for the later comparison.

3 Method

Considering the preference of human designers, the metric required by design generation should "deprioritize" the visual artifacts and instead focus on evaluating the underlying shape. To achieve this, we come to the idea of replacing the Inception-V3 [44] model with the encoder of a trained Denoising Autoencoder (DAE) [46]. Following this, our work introduces the Fréchet Denoised Distance (FDD).

3.1 Preliminaries

Fréchet Inception Distance (FID). The FID leverages the ImageNet-trained Inception-V3 model without its last fully connected layer. Hereby, it provides a lower-dimensional latent space. Real images $\vec{x}$ and generated images $\vec{x}'$ are embedded into the Inception features $\vec{w} \in \mathbb{R}^{2048}$ and $\vec{w}' \in \mathbb{R}^{2048}$, respectively, and then separately fitted into two multivariate Gaussian distributions, with $(\mu_{\vec{w}}, \Sigma_{\vec{w}})$ and $(\mu_{\vec{w}'}, \Sigma_{\vec{w}'})$ denoting the means and covariances thereof. The difference between the two latent manifolds will be quantified with Fréchet distance with:

$$\mathrm{FD} = \|\mu_{\vec{w}} - \mu_{\vec{w}'}\|_2^2 + \mathrm{Tr}(\Sigma_{\vec{w}} + \Sigma_{\vec{w}'} - 2(\Sigma_{\vec{w}}\Sigma_{\vec{w}'})^{\frac{1}{2}}), \tag{1}$$

where $\mathrm{Tr}(\cdot)$ computes the trace of a matrix.

Denoising Autoencoder (DAE). The DAE [46] is able to observe the same structural representation in the image regardless of the noisy disturbances, which demonstrates its robustness in assessing structural plausibility. The architecture of DAE is based on an expansion of the fundamental autoencoder model, consisting of two components: an encoder $(E_\theta \colon \vec{x} \to \vec{w})$ and a decoder $(D_\theta \colon \vec{w} \to \vec{x})$. In the training phase, source images $\vec{x} \in \mathbb{R}^{w \times h \times c}$ are corrupted with Gaussian noises $\vec{x}_\eta = \vec{x} + \eta$, where $\eta \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \mathbf{I})$ and σ refers to the noise scale. The encoder (E_θ) embeds the noised image $\vec{x}_\eta$ into its lower-dimension latent representation $\vec{w} = E_\theta(\vec{x}_\eta)$, then the decoder restores the latent representation back into pixel-based image space $\hat{\vec{x}} = D_\theta(\vec{w}) = D_\theta \circ E_\theta(\vec{x}_\eta)$. The network is trained using the following loss function:

$$\min_{E_\theta, D_\theta} \Delta(\vec{x}, \hat{\vec{x}}) = \frac{1}{n}\sum_{i=1}^{n}(\vec{x}_i - \hat{\vec{x}}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\vec{x}_i - D_\theta \circ E_\theta(\vec{x}_i + \eta))^2 \,, \tag{2}$$

where n is the batch size. While minimizing the reconstruction error in the training, Denoising Autoencoder (DAE) in turn maximizes the mutual information between the original input $\vec{x}$ and its latent representation $\vec{w}$ [46]. More specifically, DAE bypasses the noisy corruption between $\vec{x}$ and $\hat{\vec{x}}$. This allows the latent representation $\vec{w}$ to contain meaningful information about the source $\vec{x}$, even thought DAE only sees the corrupted input $\hat{\vec{x}}$.

3.2 Fréchet Denoised Distance (FDD)

We implement the encoder $E_\theta(\vec{x}_\eta)$ of the Denoising Autoencoder (DAE) as the feature extractor. First, we design a DAE architecture (refer to Sect. 4.2 for more information on this architecture) and train it on the ImageNet [11] dataset with input shape of $299 \times 299 \times 3$. Second, similarly to the procedure of FID, we embed a certain number K of real images $\vec{x}$ and generated images $\vec{x}'$ into the latent features $\vec{w} \in \mathbb{R}^{2048}$ and $\vec{w}' \in \mathbb{R}^{2048}$, respectively. Note that the image is preprocessed into a shape of $299 \times 299 \times 3$ regardless of the original shape and color. Next, we follow the procedure of the Fréchet distance, introduced in Sect. 3.1, to quantify the difference between the two manifolds $\vec{w}$ and $\vec{w}'$. Hereby, we design the Fréchet Denoised Distance (FDD), illustrated with an explanatory diagram in Fig. 2.

For exploration purpose, we simulate the design processes of KID [5] and TD [23] and replace the distance measures with Maximum Mean Discrepancy (MMD) and Topology Distance (TD), hereby delivering more DAE-based metrics, *e.g.*, Kernel Denoised Distance (KDD) and Topology Denoised Distance (TDD). We also notice the work of [16] that trains a ResNet-50 [19] model on an alternative dataset of ImageNet, *i.e.*, Stylized-ImageNet, and hereby successfully develops a shape-biased classifier. Inspired by this proposal, we additionally train a DAE model from scratch on the BIKED [38] dataset. The DAE model trained on BIKED images has an input shape of $256 \times 256 \times 1$ and a smaller latent space with dimension $D_{\vec{w}} = 64$. Hereby, we design a FDD $(\cdot)$ metric, which is based on the DAE trained on the same target dataset. The evaluation of FDD $(\cdot)$ on BIKED images is demonstrated in Sect. 4.

4 Experiments

To evaluate the design plausibility of generated images, a useful metric should satisfy the following conditions: (1) bias toward design structure, (2) consistency with increasing disturbances, and (3) alignment with human judgment. Hence, we leverage correspondingly three experiments: sensitivity test, consistency test with increasing disturbances, and model ranking, over the the SOTA metrics and our metrics (see Table 1 for more detailed information). Note that in our work, the TD metric refers to TD-Inception [23] unless otherwise explained.

4.1 Datasets

We select a variety of datasets covering different aspects. For a fair comparison with the FID, we train the DAE on the ImageNet [11] dataset, ensuring that the learned feature manifold is similar to the one of the Inception-V3 model. We employ a subset of the ImageNet [11] dataset of 50 000 samples with dimension $299 \times 299 \times 3$, properly chosen to cover a wide range of 1 000 classes. The dataset is divided into 45 000 training samples and 5 000 test samples. Our comparative analysis and tests also incorporate two design datasets, BIKED [38] and Seeing3DChairs [1], to address the interests of human designers. Additionally, we

Table 1. A list of candidate performance metrics for measuring design plausibility. Below the *dashed line*, we also list other DAE-based metrics as a means of exploring further improvements. FDD (·) utilizes a DAE trained on the target dataset with the DAE architecture modified according to the dataset, *e.g.*, FDD (BIKED) utilizes a DAE trained on the BIKED dataset.

Metric	Backbone Model	Input Dimension	Feature Dimension	Training Dataset	Distance Measures
FID [21]	Inception-V3 [44]	$299 \times 299 \times 3$	2048	ImageNet [11]	Fréchet Distance
KID [5]					Maximum Mean Discrepancy
$FD_{DINO\text{-}V2}$ [43]	DINOv2 [35], ViT [13]	$224 \times 224 \times 3$	1024	LVD-142M [43]	Fréchet Distance
TD-Inception [23]	Inception-V3 [44]	$299 \times 299 \times 3$	2048	ImageNet [11]	Topology Distance
TD-ResNet [23]	ResNet18 [19]	$224 \times 224 \times 3$	512	Fashion-MNIST [47]	
FDD	Denoising Autoencoder [46]	$299 \times 299 \times 3$	2048	ImageNet [11]	Fréchet Distance
KDD	Denoising Autoencoder [46]	$299 \times 299 \times 3$	2048	ImageNet [11]	Maximum Mean Discrepancy
TDD					Topology Distance
FDD (·)		$256 \times 256 \times 1$	64	Target Dataset	Fréchet Distance

incorporate the color-channeled FFHQ [26] dataset and the test samples of ImageNet [11] into our metric testing to confirm the metric's adaptability to general image generation tasks. Details of the implemented datasets for comparative analysis are provided in the supplementary material.

4.2 Experimental Settings

For the reproducibility of our work, this section documents all the essential details regarding the development of our FDD metric and the experimental setups. To justify the setting choices, we aim to align our DAE model's architecture with that of the Inception-V3 model, particularly in terms of input shape and latent dimension. The model architecture and training settings describe the DAE trained on ImageNet, whereas the configurations of the DAE trained on BIKED are correspondingly adjusted as shown in Table 1.

Model Architecture. Our approach employs a DAE comprising 5 convolutional layers across both the encoder and the decoder. Here, the feature dimensions for the convolutional layers in the encoder are arranged in the following sequence $[32, 64, 128, 256, 512]$. For the decoder, these dimensions are applied in reverse order. Each layer employs a 3×3 kernel shape, a stride of 2, padding of 1, and the `Rectified Linear Unit (ReLU)` as the activation function, aligning with the Inception-V3 model. The last activation layer of the decoder uses a `Tanh` function to adjust the outputs to a pixel range of $[-1, 1]$. In alignment with the configuration parameters of the Inception-V3 model, the encoder's input shape is specified as $299 \times 299 \times 3$, and the latent vector dimension is established at 2048. See the supplementary material for visualization of the model architecture.

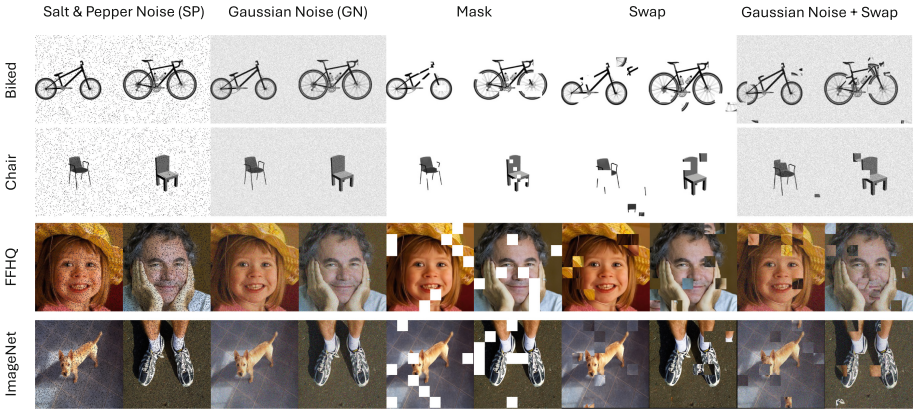


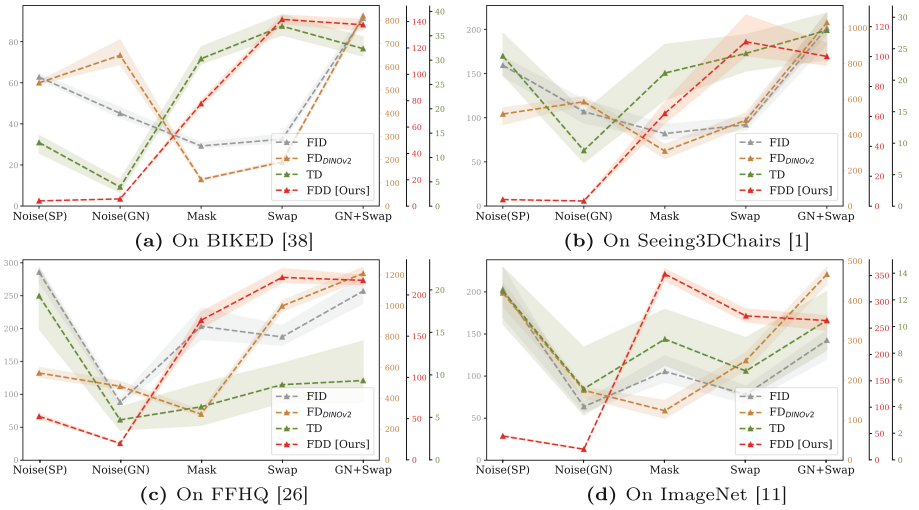
Fig. 3. Examples of manipulated images for sensitivity test. We choose the intensity of the disturbances so that images with structural errors (*i.e.*, mask and swap) are notably less plausible than ones with visual artifacts (*i.e.*, salt & pepper noise and Gaussian noise).

Training Settings. The training process uses a subset of 45 000 images from ImageNet [11], which are rescaled to the range $[-1, 1]$. For the DAE training set-up, input images are corrupted with Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma^2 \cdot \mathbf{I})$ with $\sigma = 0.1$ before being fed into the encoder. We utilize the `Adam` Optimizer with a learning rate of 1e−3 to train the DAE with a batch size of 128 and epochs of 1 000. The reconstruction loss is assessed by calculating the mean squared error (MSE) between the original and output images. Model performance is continuously assessed during training, and the best-performing model is chosen from the saved checkpoints for further experiments. We also implement an early stop function, where the training stops if the reconstruction loss does not reduce within 20 epochs. In the supplementary material, we provide qualitative results of reconstructing images from various datasets using the DAE trained on ImageNet.

Disturbance Procedures. For conducting the sensitivity and consistency tests that exam metrics' performance in dealing with various disturbances, we design the perturbation methods, *i.e.*, salt & pepper noise, Gaussian noise, patch mask, patch swap and a mixed disturbance of Gaussian noise and patch swap, and their respective intensity levels based on previous studies [21,23]. Details of the disturbances are provided in the supplementary material.

4.3 Sensitivity Test

Despite the presence of noise, a human designer can still recognize the underlying structure in a design. However, designs with missing parts or structural errors are less usable. Thus, we design the sensitivity test with the anticipation that an appropriate metric for the design generation evaluation task should progressively



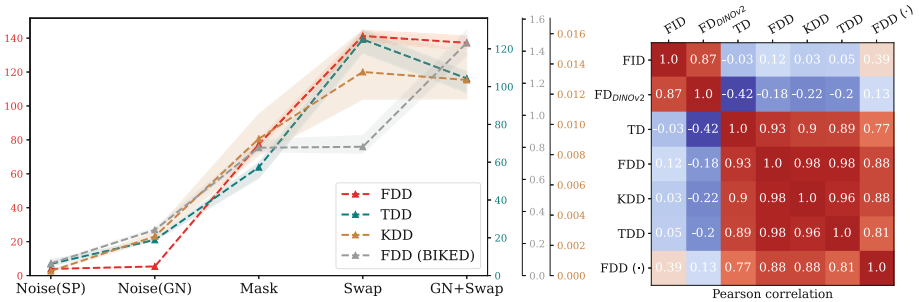
(a) On BIKED [38] (b) On Seeing3DChairs [1]

(c) On FFHQ [26] (d) On ImageNet [11]

Fig. 4. Sensitivity Comparison. The y-axis represents the score value measured by each metric, where a lower value indicates a higher similarity to source data, *i.e.*, better quality. A reliable plausibility metric should penalize more on the basis of structural errors (*e.g.*, mask and swap) than visual artifacts (*e.g.*, noise). For each metric, the *dashed line* shows the mean across the groups and the *shaded region* depicts the measured values from the groups.

demonstrate deteriorating scores from visual artifacts to structural deficiencies. Additionally, to prove the importance of structural integrity in the evaluation process, we expect that the score for a mixed disturbance of Gaussian noise and patch swap will be comparable to that of solely patch swap disturbance, thus remaining independent from the added visual artifacts. The aim of the sensitivity test is to cross-compare the metric performance in dealing with various disturbances and to see if the metric aligns with human designers.

This test involves four datasets: BIKED [38], ImageNet [11], FFHQ [26] and Seeing3DChairs [1]. For each dataset, we shuffle and split the samples into $n = 10$ groups, number of samples in each group varies from the dataset: $K = 300$ (for BIKED) and $K = 100$ (for Seeing3DChairs, FFHQ and ImageNet). Considering the effect of sample number to the metric performance, we provide results of sensitivity test with 1k images for each dataset in the supplementary material. We introduce five types of disturbances into source images and create five corrupted counterparts, *i.e.*, pepper noise, Gaussian noise, patch mask, patch swap and a mix of Gaussian noise and patch swap. The introduced disturbances adhere to a rule where visual artifacts, such as pepper noise and Gaussian noise, are intentionally kept at levels that do not significantly impact the recognition of the design. On the other hand, structural failures, such as patch masking and patch swapping, lead to designs that are implausible and consequently receive worse human evaluation scores compared to visual artifacts. We choose one level from



(a) Sensitivity test on BIKED [38] **(b)** Correlation test

Fig. 5. Experiments with FDD and other DAE-based metrics. In (a), within all DAE-based metrics, FDD shows the best performance; (b) Pearson correlation of metrics over all distances measured during the sensitivity test with BIKED.

each disturbance described in Sect. 4.2: $\alpha = 0.01$ (pepper noise, Gaussian noise), and $\alpha = 0.25$ (patch mask, patch swap). Next, we measure the distance between each one of these corrupted image sets and the original image set, using FID, $\text{FD}_{\text{DINO-V2}}$, TD, and our FDD. Since they are measures of distance quantifying the dissimilarity between observed images and source images, a smaller value indicates greater similarity to the source data.

We plot several examples of disturbed images in Fig. 3 and record the measured results in Fig. 4. As expected, FID [21] and $\text{FD}_{\text{DINO-V2}}$ [43] show a great bias towards visual artifacts, with notably higher distance assigned to pepper and Gaussian noised images compared to those with patch mask and patch swap. In contrast to FID and $\text{FD}_{\text{DINO-V2}}$, TD and our FDD provide a distinct evaluation perspective by detecting structural faults and imposing penalties accordingly. One unanticipated result was that TD exhibits a poor performance with regard to pepper noise as illustrated in Fig. 4b and Fig. 4c. Furthermore, as the sample size decreases within each group from 300 (BIKED [38]) to 100 (for Seeing3DChairs [1] and FFHQ [26]), TD shows a significant increase in standard deviation across 10-times implementations. Interestingly, our FDD shows a better stability among various noises and gives significantly worse scores to images with structural failures.

Furthermore, we explore other possible metrics based on DAE by incorporating concepts from existing works such as KID [5], TD [23] and training the network on structural images [16]. This adaption yields new evaluation metrics, *i.e.*, Kernel Denoised Distance (KDD), Topology Denoised Distance (TDD), and FDD (BIKED), respectively. Later on, we subject these metrics to the sensitivity test and present the results in Fig. 5a. Our analysis reveals that FDD exhibits the most consistent performance across various criteria: the most stable result across different groups and excellence in distinguishing between visual and structural disturbances.

Finally, we calculate the Pearson correlation coefficients pair-wisely among all candidate metrics, by taking the measured values from Fig. 4a, and record the

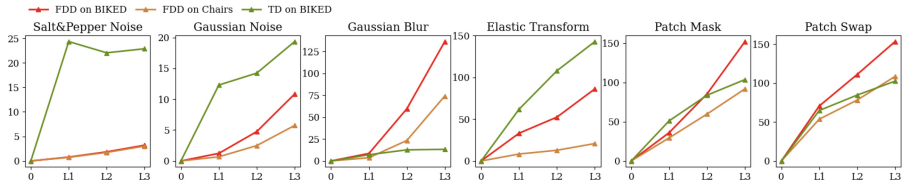


Fig. 6. Metric comparison with increasing disturbances. The y-axis label represents the measured distance of each metric. The TD, as the most competitive metric to our FDD, performs however unstably with increasing disturbances.

outcome in the table Fig. 5b. Notably, the result reveals two categories among the metrics: FID and $\text{FD}_{\text{DINO-V2}}$ are grouped together, while TD and our designed metrics demonstrate a stronger correlation with each other. This is a promising finding since TD's main perspective is the topology and geometric behavior of the latent space, hereby we argue that the latent space of our DAE maintains the topological properties of the image space well and can be captured by Fréchet Distance. Note that the KDD, TDD and FDD (BIKED) are experimental explorations. They are highly related to our FDD in the correlation test and our FDD outperforms them in the sensitivity test.

4.4 Consistency with Increasing Disturbances

In this section, we test the consistency of the FDD metric in response to escalating levels of disturbances outlined in Sect. 4.2. As a fundamental requirement, a performance metric should be able to accurately detect and respond to worsened image quality, including visual fidelity and structural plausibility. We start by adding various disturbances to a subset comprising $K = 1\,000$ images sourced from the BIKED [38] and Seeing3DChairs [1] datasets, respectively. Afterwards, we report the scores in Fig. 6 and demonstrate the consistent performance of the proposed FDD metric. While FID has been noted to exhibit inconsistency in detecting the disturbance level induced by salt and pepper as documented in Heusel *et al.* [21], our FDD successfully measures the levels of various deformations, spanning from visual to structural distortions.

4.5 Model Ranking

In the model ranking, we employ five deep generative models, *e.g.*, DDPM [22], DDIM [42], EDM [25] and PoDM [14], with the consideration that the models executed in model ranking should exhibit significant differences in visual quality and structural plausibility. These models are then trained on BIKED images with a resolution of 256×256. We generate 5k images from each model and manually evaluate them into plausible designs and implausible designs. We denote the ratio of implausible bicycle designs as human evaluation error, the lower the better, which serves as the "ground truth" in this model ranking experiment.

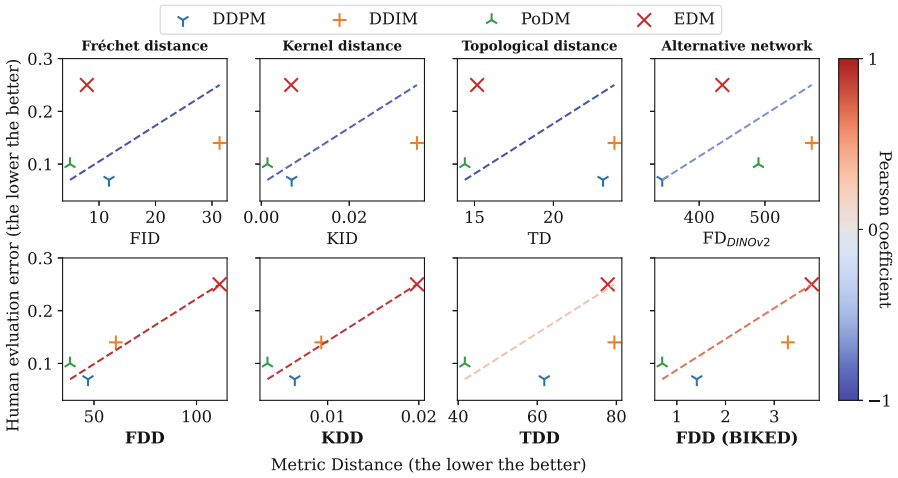


Fig. 7. Metrics comparison in the task of model ranking. We color-code the *diagonal lines* after the measured Pearson correlation coefficient between metric results and human judgments, *dark red* refers to a strong positive correlation between metric distances and human judgments.

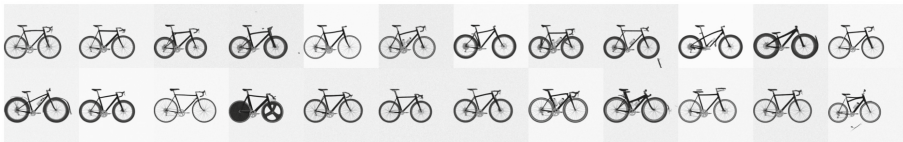
Meanwhile, we apply the candidate metrics, including FID, KID, $FD_{DINO\text{-}V2}$, TD, FDD, and other DAE-based metrics, to evaluate each generative model with their generated samples, with 1k images in each group. Subsequently, the distances measured and human error rates are visualized in Fig. 7. Note that the proximity of the plotted points (measured distances, human evaluation error) to the diagonal line signifies the consistency of the metric with human evaluation. Particularly, the expected behavior is seen in the FDD, KDD, and FDD (BIKED) measurements, which are highly associated and yield the same consistent ranking. This observation aligns with the notion proposed by [43], whereby provided a good encoder is chosen, all these metrics provide sensible ways of quantifying distances between probability distributions.

On the other hand, the absence of a significant link between the SOTA metrics and human evaluation suggests a deficiency of these most reported metrics in evaluating structural design images. In Fig. 8, we plot the generated bicycles for qualitative evaluation of our FDD metric. EDM achieves the best FID of 7.84, but the generated bicycles contain a large portion of implausible designs; DDPM (FID 11.77) and DDIM (FID 31.35) are unfairly penalized, even though the results are significantly more plausible than those from EDM. Here, our FDD is able to rank the models more accurately.

4.6 Grad-CAM Visualization

The Grad-CAM [30,41] is designed to visualize the focus on the input image as perceived by the classifier/segmentation model up to the last fully connected

(a) DDPM [22] (FID: 11.77, $FD_{DINO\text{-}V2}$: 342.82, FDD: 48.08)



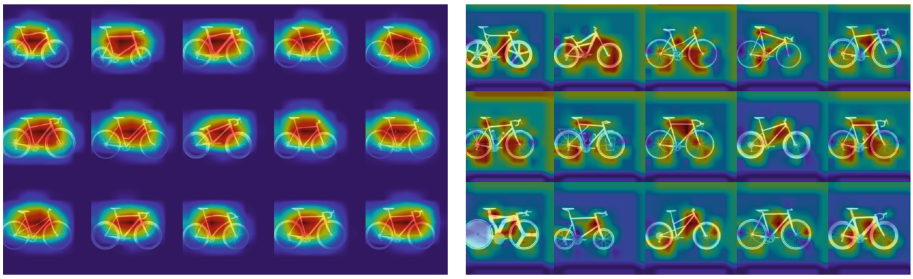
(b) DDIM [42] (FID: 31.35, $FD_{DINO\text{-}V2}$: 571.21, FDD: 60.66)

(c) EDM [25] (FID: 7.84, $FD_{DINO\text{-}V2}$: 435.25, FDD: 111.25)

Fig. 8. Qualitative evaluation of generated bicycle designs. DDPM and DDIM yield structurally more plausible results than EDM, but FID and $FD_{DINO\text{-}V2}$ fail to agree with human judgments, whereas our FDD ranks the models with the perspective of structural plausibility and penalizes the visual artifacts as well.

layer. In our work, we use the Grad-CAM visualization to compare the observation fields of the FID and FDD metrics. We first transfer the test images into inception space and latent space via the Inception-V3 model and the DAE model, respectively, which have the same dimension of $\mathbb{R}^{2048}$. We compute the mean $\mu_{\vec{w}}$ and the covariance $\Sigma_{\vec{w}}$ of the extracted features. Then, we obtain the attention maps of FID and FDD by back-propagating the value of ${\mu_{\vec{w}}}^2 + \Sigma_{\vec{w}}$, to the last convolutional layer of the Inception-V3 model (*i.e.*, *Mixed 7c.branch pool*) and the one of DAE (*i.e.*, *encoder 8*), respectively. The Grad-CAM generates a heatmap of reduced dimensions (*e.g.*, 10×10 for DAE) which is then upsampled to match the dimensions of the original image for intuitive visual comparison. The heatmaps (seen in Fig. 9) visualize the area observed by the corresponding metric in the BIKED [38] images, *i.e.*, where the metric "looks at" when it calculates the distance.

The focus of the Inception-V3 model is simply the area around the center of the main object, often mismatching the object's shape and borders. As explained in previous works [28,43], this phenomenon is caused by the model's classification training across 1 000 classes. Consequently, it prioritizes detecting the object's presence rather than its structure. On the other hand, even when also trained on ImageNet, the DAE generates an intensive attention map with positive and negative gradients surrounding the bicycle's structure, which efficiently assesses the complex details of the bicycle's shape. In supplementary material, we provide



(a) FID (Inception-V3) **(b)** Our FDD (DAE trained on ImageNet)

Fig. 9. Where does the metric look at? Heatmaps illustrating the perception of the Fréchet distance with Grad-CAM. The focus of an encoder can be demonstrated by both *bright red* and *deep blue*. The offset of the focusing area can be caused by upsampling the attention map to the image shape.

also Grad-CAM analysis on general images, where Inception-V3 model tends to drop structural information while DAE captures it.

5 Conclusion

In this work, we approached the field of evaluating generated design images and proposed a structure-biased metric Fréchet Denoised Distance (FDD) by replacing the Inception-V3 model in the FID metric with a Denoising Autoencoder, unsupervised-trained on the same dataset (*i.e.*, ImageNet) and with the same 2048-dimensional latent space. Through a series of experiments, including sensitivity test for various types of disturbance, consistency test with increasing disturbances, and alignment test with human judgment in model ranking, we found FDD to fulfill the quality requirements for serving as a metric and outperform other SOTA metrics, *e.g.*, FID, $\text{FD}_{\text{DINO-V2}}$ and TD, on design images such as BIKED and Seeing3DChairs, as well as real-world images such as human faces from FFHQ and general images from ImageNet. We explained the effectiveness of FDD with a Grad-CAM visualization, where the DAE is able to "focus" on the design structure of the observed shape.

Limitation and Future Work. FDD has a low priority towards visual artifacts, which may become problematic in contexts where the visual quality is critical. Yet, our experiments demonstrated that when there is no major structural failure, it is still capable of penalizing visual artifacts. In addition, we believe that this novel insight may also be useful in guiding DGMs to generate more reliable, plausible designs, which is of great potential to be further investigated in future work. Finally, while our research on FDD mainly focused on image space, considering the recent study of 3D Diffusion modeling (adding noise to 3D data) and the use of an autoencoders to process 3D data, we believe that FDD can also assess 3D-DGMs.

Acknowledgements. We gratefully acknowledge Laure Vuaille for her valuable insights and the support provided by BMW Group.

References

1. Aubry, M., Maturana, D., Efros, A.A., Russell, B.C., Sivic, J.: Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In: 2014 IEEE CVPR, pp. 3762–3769 (2014). https://doi.org/10.1109/CVPR.2014.487
2. Baker, N., Lu, H., Erlikhman, G., Kellman, P.J.: Deep convolutional networks do not classify based on global object shape. PLoS Comput. Biol. **14** (2018). https://api.semanticscholar.org/CorpusID:54476941
3. Barratt, S.T., Sharma, R.: A note on the inception score. ArXiv **abs/1801.01973** (2018). https://api.semanticscholar.org/CorpusID:38384342
4. Betzalel, E., Penso, C., Navon, A., Fetaya, E.: A study on the evaluation of generative models. CoRR **abs/2206.10935** (2022). https://doi.org/10.48550/ARXIV.2206.10935
5. Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings. OpenReview.net (2018). https://openreview.net/forum?id=r1lUOzWCW
6. Borji, A.: Pros and cons of GAN evaluation measures: new developments. Comput. Vis. Image Underst. **215**, 103329 (2022). https://doi.org/10.1016/j.cviu.2021.103329, https://www.sciencedirect.com/science/article/pii/S1077314221001685
7. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. ArXiv **abs/1809.11096** (2018), https://api.semanticscholar.org/CorpusID:52889459
8. Buzuti, L.F., Thomaz, C.E.: Fréchet autoencoder distance: a new approach for evaluation of generative adversarial networks. Comput. Vis. Image Underst. **235**, 103768 (2023)
9. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS 2020, Curran Associates Inc., Red Hook, NY, USA (2020)
10. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: diverse image synthesis for multiple domains. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8185–8194 (2020). https://doi.org/10.1109/CVPR42600.2020.00821
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, pp. 248–255. IEEE (2009). https://ieeexplore.ieee.org/abstract/document/5206848/
12. Dhariwal, P., Nichol, A.Q.: Diffusion models beat GANs on image synthesis. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021). https://openreview.net/forum?id=AAWuCvzaVt
13. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Fan, J., Vuaille, L., Bäck, T., Wang, H.: On the noise scheduling for generating plausible designs with diffusion models (2023)

15. Fan, J., Vuaille, L., Wang, H., Bäck, T.: Adversarial latent autoencoder with self-attention for structural image synthesis. arXiv preprint arXiv:2307.10166 (2023)
16. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F., Brendel, W.: Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. ArXiv **abs/1811.12231** (2018). https://api.semanticscholar.org/CorpusID:54101493
17. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014). http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
18. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. **13**(1), 723–773 (2012)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
20. Hermann, K.L., Chen, T., Kornblith, S.: The origins and prevalence of texture bias in convolutional neural networks. arXiv: Computer Vision and Pattern Recognition (2019). https://api.semanticscholar.org/CorpusID:220266152
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local NASH equilibrium. In: Advances in Neural Information Processing Systems, vol. 30 (NIPS 2017) (2018)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv preprint arxiv:2006.11239 (2020)
23. Horak, D., Yu, S., Khorshidi, G.S.: Topology distance: a topology-based approach for evaluating generative adversarial networks. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 7721–7728. AAAI Press (2021). https://doi.org/10.1609/AAAI.V35I9.16943
24. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. ArXiv **abs/1710.10196** (2017). https://api.semanticscholar.org/CorpusID:3568073
25. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Adv. Neural. Inf. Process. Syst. **35**, 26565–26577 (2022)
26. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
27. Kucker, S.C., et al.: Reproducibility and a unifying explanation: lessons from the shape bias. Infant Behav. Dev. **54**, 156–165 (2019). https://api.semanticscholar.org/CorpusID:53045726
28. Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., Lehtinen, J.: The role of imagenet classes in fréchet inception distance. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=4oXTQ6m_ws8
29. Landau, B., Smith, L.B., Jones, S.S.: The importance of shape in early lexical learning. Cogn. Dev. **3**, 299–321 (1988). https://api.semanticscholar.org/CorpusID:205117480
30. Liu, W., et al.: Towards visually explaining variational autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8642–8651 (2020)

31. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
32. Maiorca, A., Yoon, Y., Dutoit, T.: Evaluating the quality of a synthesized motion with the fréchet motion distance. In: ACM SIGGRAPH 2022 Posters. SIGGRAPH 2022, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3532719.3543228
33. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 7176–7185. PMLR (2020). https://proceedings.mlr.press/v119/naeem20a.html
34. Nobari, A.H., Rashad, M.F., Ahmed, F.: Creativegan: editing generative adversarial networks for creative design synthesis. CoRR **abs/2103.06242** (2021). https://arxiv.org/abs/2103.06242
35. Oquab, M., et al.: DINOv2: learning robust visual features without supervision. Trans. Mach. Learn. Res. (2024). https://openreview.net/forum?id=a68SUt6zFt
36. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021). https://api.semanticscholar.org/CorpusID:231591445
37. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016). http://arxiv.org/abs/1511.06434
38. Regenwetter, L., Curry, B., Ahmed, F.: BIKED: a dataset for computational bicycle design with machine learning benchmarks. J. Mech. Des. **144**(3) (2021). https://doi.org/10.1115/1.4052585
39. Regenwetter, L., Nobari, A.H., Ahmed, F.: Deep generative models in engineering design: a review. J. Mech. Des. **144**(7), 071704 (2022)
40. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. ArXiv **abs/1606.03498** (2016), https://api.semanticscholar.org/CorpusID:1687220
41. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
42. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv:2010.02502 (2020). https://arxiv.org/abs/2010.02502
43. Stein, G., et al.: Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. CoRR **abs/2306.04675** (2023). https://doi.org/10.48550/ARXIV.2306.04675
44. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2818–2826 (2015). https://api.semanticscholar.org/CorpusID:206593880
45. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
46. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: International Conference on Machine Learning (2008). https://api.semanticscholar.org/CorpusID:207168299

47. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. ArXiv **abs/1708.07747** (2017). https://api.semanticscholar.org/CorpusID:702279
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 586–595. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00068, http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html
49. Zhou, W.: Image quality assessment: from error measurement to structural similarity. IEEE Trans. Image Process. **13**, 600–613 (2004)