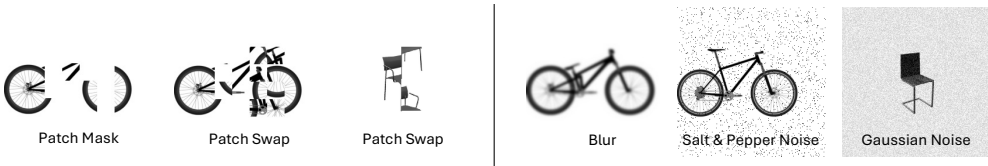
Deep generative models for engineering design

Fan, J.

Chapter 3

Evaluating the Plausibility with Deep Learning

We have pointed out the evaluation issue in the previous chapter, i.e., the evaluation of FID does not align well with human judge in terms of the design plausibility. In this chapter, we are addressing this issue as well as the research question 2: *How to automatically evaluate the plausibility of designs generated by DGMs?*. The content of this chapter has been published in the paper [164].



(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

Figure 3.1: From which side (*left* or *right*) are the design images more plausible? (a) Structural implausibility; (b) visual artifacts. Recent works [10, 46, 57] discover 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.

3.1 Introduction

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 works rely on a subjective evaluation [167, 99], 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 [99]. 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 [11, 112, 14]. DGM developers [70, 58, 72] are heavily relying on the Fréchet inception distance (FID) [58] metric, which extracts latent features from real and generated images with an Inception-V3 [163] model pre-trained on ImageNet [38] 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 [18, 161, 83] 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 [134], VQ-VAE [173] and DINOv2 [119], etc. According to the most recent work by Stein et al. [161], where they compared 17 metrics using encoders from 9 various networks, $\mathrm{FD}_{\text{DINO-V2}}$ 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 [10, 46, 57, 167]. 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 Figure 3.1. 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 [86, 82]. 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.

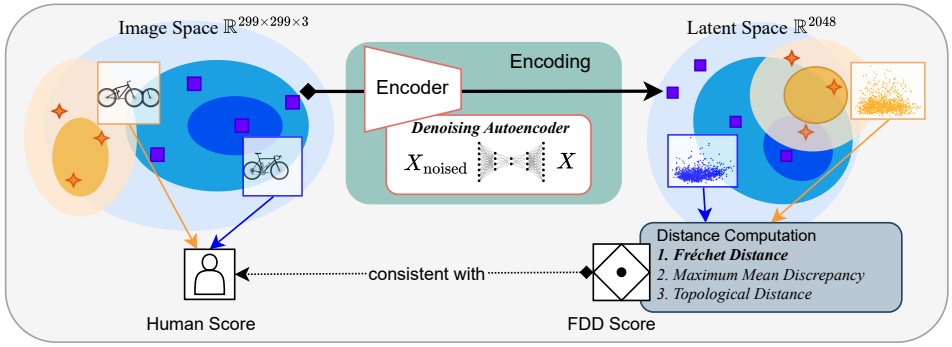


Figure 3.2: Plausibility evaluation using Fréchet denoised distance. *Blue area and squares* visualizing the distribution and samples of real data in the image space and in the DAE-encoded latent space; *Orange area and stars* illustrating the distribution and samples of generated data in the image space and in the DAE-encoded latent space.

Finally, we propose the Fréchet denoised distance (FDD) by replacing the Inception-V3 model within the FID framework with a denoising autoencoder (DAE) [176] 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) [62] (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 [161, 148, 83] 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 [16], TD [62] and training the network on structural images [46], to design new DAE-based metrics, i.e., kernel denoised distance (KDD), topology denoised distance (TDD), and FDD $(\cdot)$, respectively. We test these metrics on BIKED and the results show that these DAE-based metrics are highly correlated and FDD performs relatively best.

3.2 Related Work on DGM Evaluation

Accurately ranking generative models remains an unresolved challenge [11, 112, 14]. 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 [161]. Overcoming the flaws of previous metrics, e.g., SSIM [202], LPIPS [198] and IS [143], currently most reported evaluation methods, e.g., Fréchet inception distance (FID) [58] and kernel inception distance (KID) [16], have largely addressed the challenge of automated evaluation and are employed as the primary metric for model ranking in the field of DGM. 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 [163] 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 [16] uses the squared maximum mean discrepancy (MMD) [50] with a polynomial kernel [14].

Concerns about the over-reliance on the Inception-V3 model have been raised and researchers claim that an ImageNet [38] classifier like the Inception-V3 model brings a significant bias to the evaluation with FID [112, 119]. Furthermore, FID is proved to be vulnerable to manipulation [83], especially when there exists a significant domain discrepancy between the data set of interest such as BIKED [137] and ImageNet [38]. Consequently, the results measured by FID often show a poor correlation with human judgments. Similarly, KID [16] 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. [161] studied 17 metrics with 9 different encoders (e.g., CLIP [134], SwAV [24] and DINOv2 [119]). Their finding concludes that $\text{FD}_{\text{DINO-V2}}$ [161] demonstrates the most reliable performance over various perspectives, e.g., fidelity, diversity, rarity, and memorization of generative models. Previous works [10, 46, 57] 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. They revealed the limitation of ImageNet-trained CNNs in recognizing shapes. This flaw may explain the inconsistency of CNN-based metrics with human judgments when evaluating design images, where human experts prefer to use shape information for assessment [86, 82]. In other studies, new metrics have been

proposed to evaluate fidelity and diversity, including density and coverage [112], as well as precision and recall [141, 84].

Recent studies have introduced autoencoder-based metrics for evaluation purposes: for instance, Buzuti et al. [21] leveraged the VQ-VAE [173] 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 [95], FFHQ [72], and AFHQ [31]. 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 [86, 82]. Meanwhile, Horak et al. [62] 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.

3.3 Preliminaries

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

$$\mathrm{FD} = \|\mu_{\boldsymbol{w}} - \mu_{\boldsymbol{w'}}\|_2^2 + \mathrm{Tr}(\Sigma_{\boldsymbol{w}} + \Sigma_{\boldsymbol{w'}} - 2(\Sigma_{\boldsymbol{w}}\Sigma_{\boldsymbol{w'}})^{\frac{1}{2}}), \tag{3.1}$$

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

Denoising Autoencoder (DAE) The DAE [176] 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 $(E_\theta \colon \boldsymbol{x} \to \boldsymbol{w})$ and a decoder $(D_\theta \colon \boldsymbol{w} \to \boldsymbol{x})$. In the training phase, source images $\boldsymbol{x} \in \mathbb{R}^{w \times h \times c}$ are corrupted with Gaussian noises $\boldsymbol{x}_\eta = \boldsymbol{x} + \eta$,

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 $\boldsymbol{x}_\eta$ into its lower-dimension latent representation $\boldsymbol{w} = E_\theta(\boldsymbol{x}_\eta)$, then the decoder restores the latent representation back into pixel-based image space $\hat{\boldsymbol{x}} = D_\theta(\boldsymbol{w}) = D_\theta \circ E_\theta(\boldsymbol{x}_\eta)$. The network is trained minimizing the following loss function:

$$\min_{E_\theta, D_\theta} \Delta(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - D_\theta \circ E_\theta(\boldsymbol{x}_i + \eta))^2 \,, \tag{3.2}$$

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 $\boldsymbol{x}$ and its latent representation $\boldsymbol{w}$ [176]. More specifically, DAE bypasses the noisy corruption between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$. This allows the latent representation $\boldsymbol{w}$ to contain meaningful information about the source $\boldsymbol{x}$, even thought DAE only sees the corrupted input $\hat{\boldsymbol{x}}$.

3.4 Method

We implement the encoder $E_\theta(\boldsymbol{x}_\eta)$ of the denoising autoencoder (DAE) as the feature extractor. First, we design a DAE architecture (refer to Section 3.5.2 for more information on this architecture) and train it on the ImageNet [38] 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 $\boldsymbol{x}$ and generated images $\boldsymbol{x}'$ into the latent features $\boldsymbol{w} \in \mathbb{R}^{2048}$ and $\boldsymbol{w}' \in \mathbb{R}^{2048}$, 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 Section 3.3, to quantify the difference between the two manifolds $\boldsymbol{w}$ and $\boldsymbol{w}'$. Hereby, we design the Fréchet denoised distance (FDD), illustrated with an explanatory diagram in Figure 3.2.

For exploration purpose, we simulate the design processes of KID [16] and TD [62] 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 [46] that trains a ResNet-50 [55] 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 [137] 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 $D_{\boldsymbol{w}} = 64$. 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 Section 3.5.

3.5 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 3.1 for more detailed information). Note that in our work, the TD metric refers to TD-Inception [62] unless otherwise explained.

3.5.1 Datasets

We select a variety of datasets covering different aspects. For a fair comparison with the FID, we train the DAE on the ImageNet [38] dataset, ensuring that the learned feature manifold is similar to the one of the Inception-V3 model. We employ a subset of the ImageNet [38] 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 [137] and Seeing3DChairs [7], to address the interests of human designers. Additionally, we incorporate the color-channeled FFHQ [72] dataset and the test samples of ImageNet [38] into our metric testing to confirm the metric's adaptability to general image generation tasks.

3.5.2 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 3.1.

Table 3.1: 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 [58]	Inception-V3 [163]	$299 \times 299 \times 3$	2048	ImageNet [38]	Fréchet distance
KID [16]					MMD
$FD_{DINO\text{-}V2}$ [161]	DINOv2 [119], ViT [78]	$224 \times 224 \times 3$	1024	LVD-142M [161]	Fréchet distance
TD-Inception [62]	Inception-V3 [163]	$299 \times 299 \times 3$	2048	ImageNet [38]	Topology distance
TD-ResNet [62]	ResNet18 [55]	$224 \times 224 \times 3$	512	Fashion-MNIST [183]	
FDD	DAE [176]	$299 \times 299 \times 3$	2048	ImageNet [38]	Fréchet distance
KDD	DAE [176]	$299 \times 299 \times 3$	2048	ImageNet [38]	MMD
TDD					Topology distance
FDD (·)		$256 \times 256 \times 1$	64	Target Dataset	Fréchet distance

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.

Training settings The training process uses a subset of 45 000 images from ImageNet [38], 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.

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 [58, 62]. The details of the disturbances are outlined below:

- **Pepper Noise.** Salt & Pepper Noise is characterized by the random conversion of image pixels to black or white. In our experiments, we specifically target pixels to turn black (i.e., pepper noise), considering the prevalent white backgrounds in most design images. The proportion of image pixels altered to black, effectively setting their value to 0, is determined by a factor α within the set $[0, 0.01, 0.02, 0.03]$.

- **Gaussian Noise.** We generate a random Gaussian noise in matrix form, $\boldsymbol{\eta} = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then we create noisy images $\boldsymbol{x}'$ by adding the defined Gaussian noise to the source image $\boldsymbol{x}$: $\boldsymbol{x}' = (1-\alpha)x + \alpha\mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\alpha \in [0, 0.1, 0.2, 0.3]$ refers to the intensity of the noise. The larger α is, the more intensive the disturbance of the source data is.

- **Gaussian Blur.** We apply a Gaussian blur to the images using a convolution operation with a Gaussian kernel. The standard deviation of the kernel, determined by α, varies from $[0, 1, 2, 3]$, resulting in progressively more blurred images.

- **Patch Mask.** For design images (BIKED and Seeing3DChairs), we evenly divide the focus area of each image (where the design object is usually located) into 16 patches. For the FFHQ-256 dataset, the entire image is segmented into 64 patches. Afterward, we randomly select a portion of patches denoted by $\alpha \in [0, 0.25, 0.5, 0.75]$ and apply a white mask to them.

- **Pepper Swap.** Using the same patch division approach as the Patch Mask, we randomly select a subset of patches, indicated by $\alpha \in [0, 0.25, 0.5, 0.75]$, and swap their positions pair-wisely.

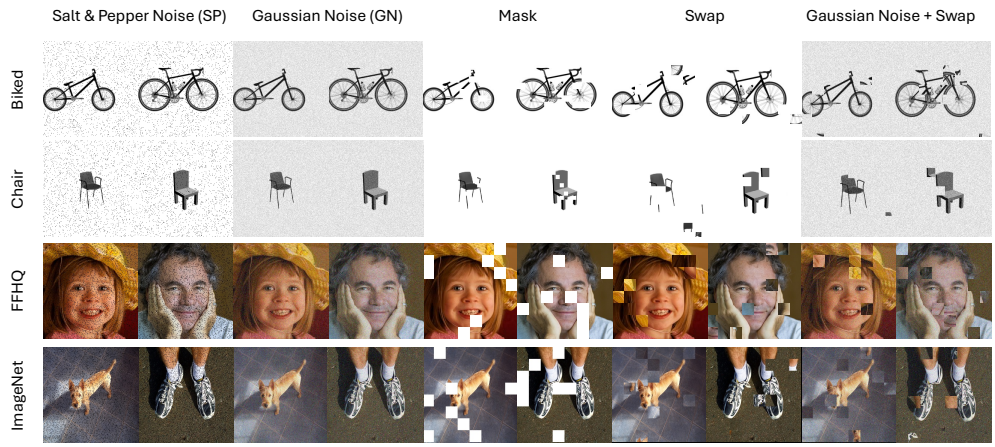


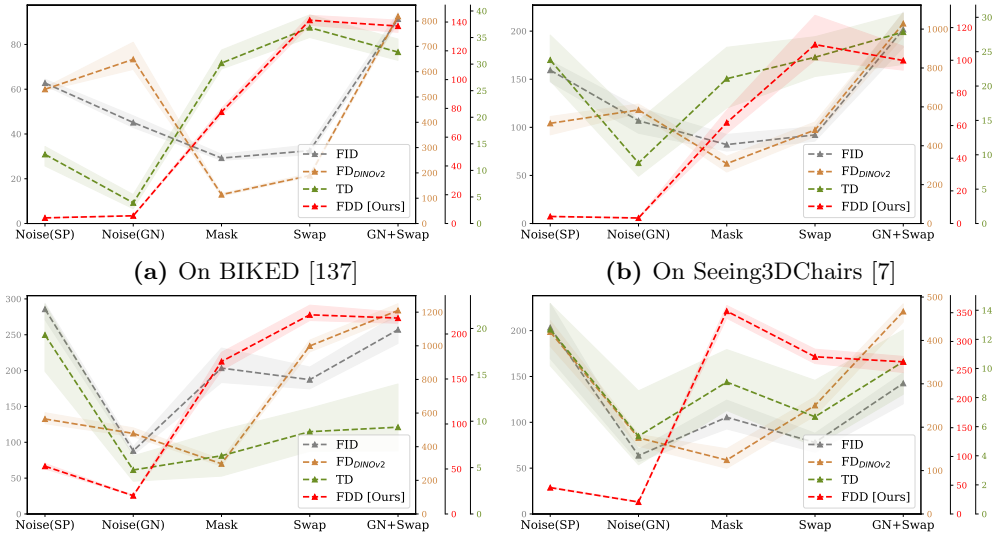
Figure 3.3: Examples of manipulated images for sensitivity test. We choose the intensity of the disturbances so that images with structural errors (i.e., mask and swap) are notably less plausible than ones with visual artifacts (i.e., salt & pepper noise and Gaussian noise).

- **Elastic Transformation.** The image is deformed by displacing a grid of control points. Each point is shifted randomly in both the horizontal and vertical directions, typically following a Gaussian distribution to determine the displacement magnitude. The degrees of the distortion are regulated by adjusting the standard deviation of the Gaussian filter $\alpha \in [0, 4, 5, 6]$.

3.5.3 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 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 [137], ImageNet [38], FFHQ [72] and Seeing3DChairs [7]. For each dataset, we shuffle and split the samples into $n = 10$

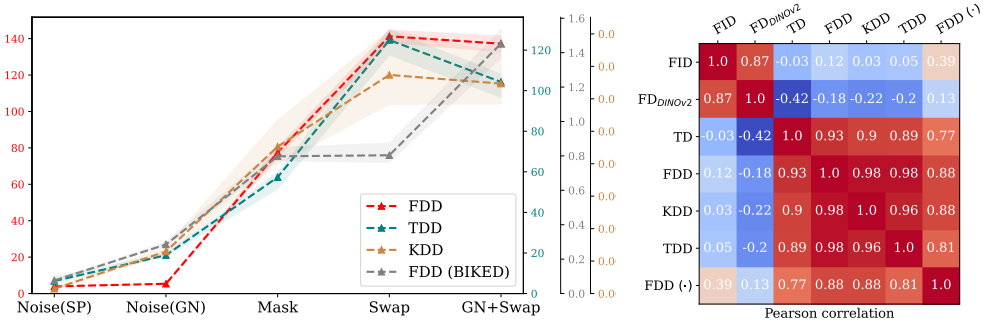


(a) On BIKED [137] **(b)** On Seeing3DChairs [7]

(c) On FFHQ [72] **(d)** On ImageNet [38]

Figure 3.4: 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.

groups, number of samples in each group varies from the dataset: $K = 300$ (for BIKED) and $K = 100$ (for Seeing3DChairs, FFHQ and ImageNet). 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 each disturbance described in Section 3.5.2: $\alpha = 0.01$ (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.



(a) Sensitivity test on BIKED [137] **(b)** Correlation test

Figure 3.5: 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.

We plot several examples of disturbed images in Figure 3.3 and record the measured results in Figure 3.4. As expected, FID [58] and $\text{FD}_{\text{DINO-V2}}$ [161] 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 Figure 3.4b and Figure 3.4c. Furthermore, as the sample size decreases within each group from 300 (BIKED [137]) to 100 (for Seeing3DChairs [7] and FFHQ [72]), 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 [16], TD [62] and training the network on structural images [46]. 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 Figure 3.5a. 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.

It is important to note that the different plausibility metrics presented in Figure 3.5a operate on inherently different scales due to the distinct formulations and normalization schemes of each method. As a result, the absolute numbers are not di-

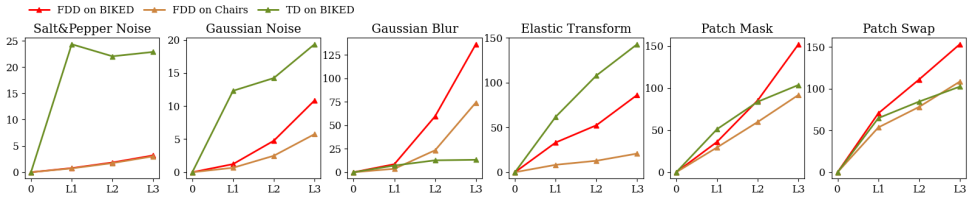


Figure 3.6: Metric comparison with increasing disturbances. The y-axis label represents the measured distance of each metric. The TD, as the most competitive metric to our FDD, performs however unstably with increasing disturbances.

rectly comparable across metrics. Despite this, the evaluation and comparison remain meaningful because the analysis focuses on relative trends and sensitivity patterns within each metric. In other words, the ability of a metric to distinguish between different types of disturbances, detect structural faults, and maintain consistency across sample groups is what informs its reliability, rather than the raw numerical range. Therefore, observing how each metric responds to noise, masking, or structural perturbations provides actionable insight, and differences in scale do not undermine the validity of the comparison.

Finally, we calculate the Pearson correlation coefficients pair-wisely among all candidate metrics, by taking the measured values from Figure 3.4a, and record the outcome in the table Figure 3.5b. 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.

3.5.4 Consistency with Increasing Disturbances

In this section, we test the consistency of the FDD metric in response to escalating levels of disturbances outlined in Section 3.5.2. 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 [137] and Seeing3DChairs [7] datasets, respectively. Afterwards, we report the scores in Fig-

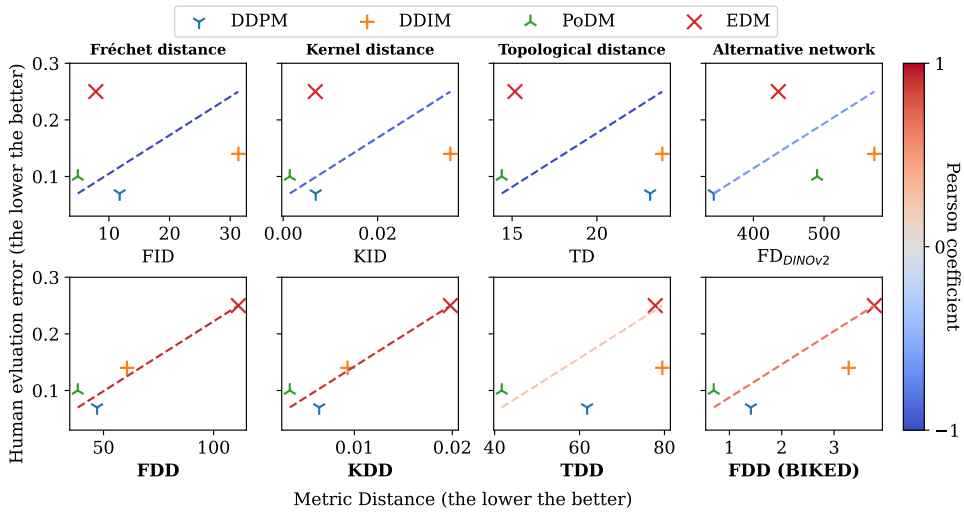


Figure 3.7: Metrics comparison in the task of model ranking. We color-code the *diagonal lines* after the measured Pearson correlation coefficient between metric results and human judgments, *dark red* refers to a strong positive correlation between metric distances and human judgments.

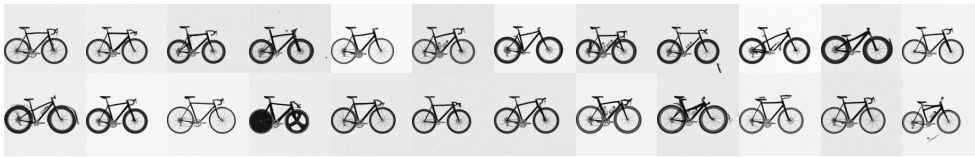
ure 3.6 and demonstrate the consistent performance of the proposed FDD metric. While FID has been noted to exhibit inconsistency in detecting the disturbance level induced by salt and pepper as documented in Heusel et al. [58], our FDD successfully measures the levels of various deformations, spanning from visual to structural distortions.

3.5.5 Model Ranking

In the model ranking, we employ five deep generative models, e.g., DDPM [61], DDIM [156], EDM [71] and PoDM [167], 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.

Meanwhile, we apply the candidate metrics, including FID, KID, $\text{FD}_{\text{DINO-V2}}$, 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

(a) DDPM [61] (FID: 11.77, $FD_{DINO\text{-}V2}$: 342.82, FDD: 48.08)



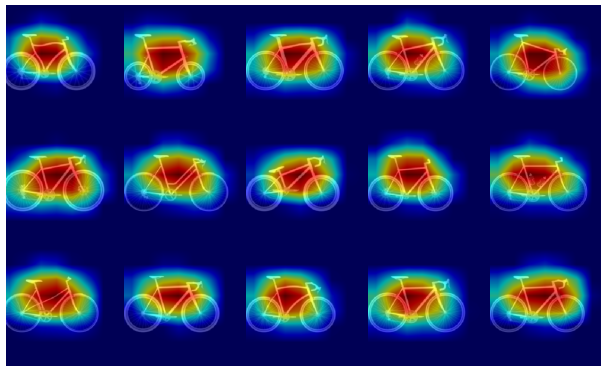
(b) DDIM [156] (FID: 31.35, $FD_{DINO\text{-}V2}$: 571.21, FDD: 60.66)

(c) EDM [71] (FID: 7.84, $FD_{DINO\text{-}V2}$: 435.25, FDD: 111.25)

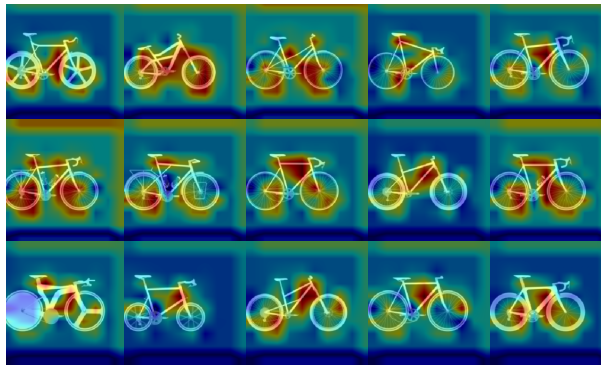
Figure 3.8: Qualitative evaluation of generated bicycle designs. DDPM and DDIM yield structurally more plausible results than EDM, but FID and $FD_{DINO\text{-}V2}$ fail to agree with human judgments, whereas our FDD ranks the models with the perspective of structural plausibility and penalizes the visual artifacts as well.

and human error rates are visualized in Figure 3.7. 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 [161], 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 Figure 3.8, 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.



(a) FID (Inception-V3)



(b) Our FDD (DAE trained on ImageNet)

Figure 3.9: 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.

3.5.6 Grad-CAM Visualization

The Grad-CAM [148, 91] is designed to visualize the focus on the input image as perceived by the classifier/segmentation model up to the last fully connected 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 $\mu_{\boldsymbol{w}}$ and the covariance $\Sigma_{\boldsymbol{w}}$ of the extracted features. Then, we obtain the attention maps of FID and FDD by back-propagating the value of ${\mu_{\boldsymbol{w}}}^2 + \Sigma_{\boldsymbol{w}}$, 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 Figure 3.9) visualize the area observed by the corresponding metric in the BIKED [137] 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 [161, 83], 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 Figure 3.14 and Figure 3.15, we provide also Grad-CAM analysis on general images, where Inception-V3 model tends to drop structural information while DAE captures it.

3.5.7 Reconstruction with Denoising Autoencoder

In this section, we demonstrate the restoration power of the DAE model trained on the ImageNet on noisy images from various datasets, e.g., the ImageNet [38] in Figure 3.10, BIKED [137] in Figure 3.11, Seeing3DChairs [7] in Figure 3.12, and FFHQ [72] in Figure 3.13. For the reconstruction, we apply Gaussian noise to the original images using the formula $\mathbf{x}_\eta = \mathbf{x} + \eta$, where $\eta \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \mathbf{I})$ and $\sigma^2 = 0.5$ and then restore the noised images with the DAE model. First, the DAE reliably filters out the additive Gaussian noise, which shows that the learned latent representation encodes the essential structural content of the image (rather than simply reproducing pixel-level noise). Second, because it is trained to ignore superficial perturbations (noise) and focus on the underlying image content, the encoder part of the DAE produces representations that are robust to visual artifacts.

Here, FID, based on the Inception-V3 network pretrained on ImageNet, lacks the explicit restoration objective that encourages latent representations to focus on structural coherence and clean image manifolds. FID therefore uses features optimized for classification (or generic image-recognition) rather than for denoising or explicitly modeling a clean-image manifold. As a result, FID's feature space may be overly sensitive to visual noise, artifacts, or superficial texture differences, but less sensitive to deeper structural implausibilities of generated designs (for instance, incorrect shape

Figure 3.10: DAE reconstruction of images from ImageNet. *Top*: original images, *Middle*: noised images, *Bottom*: reconstructed images.

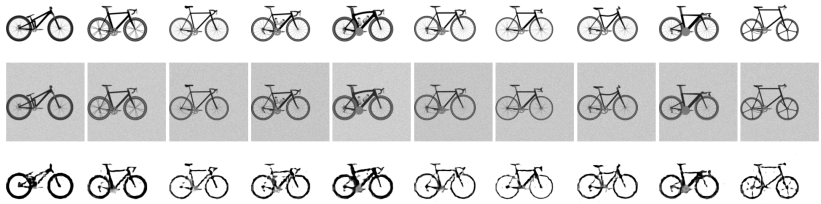


Figure 3.11: DAE reconstruction of images from BIKED. *Top*: original images, *Middle*: noised images, *Bottom*: reconstructed images.

combinations or missing functional components).

3.6 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, $\mathrm{FD}_{\text{DINO-V2}}$ 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.

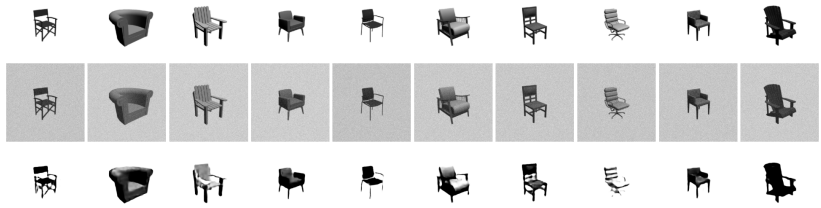


Figure 3.12: DAE reconstruction of images from Seeing3DChairs. *Top*: original images, *Middle*: noised images, *Bottom*: reconstructed images.

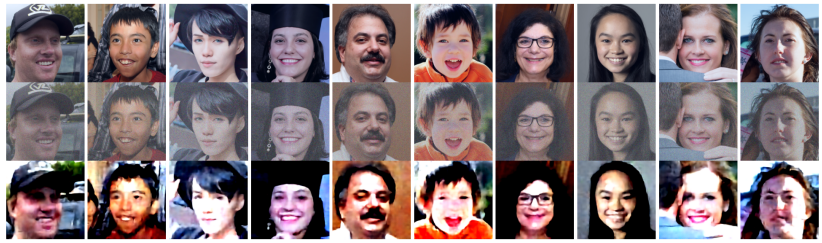


Figure 3.13: DAE reconstruction of images from FFHQ. *Top*: original images, *Middle*: noised images, *Bottom*: reconstructed images.

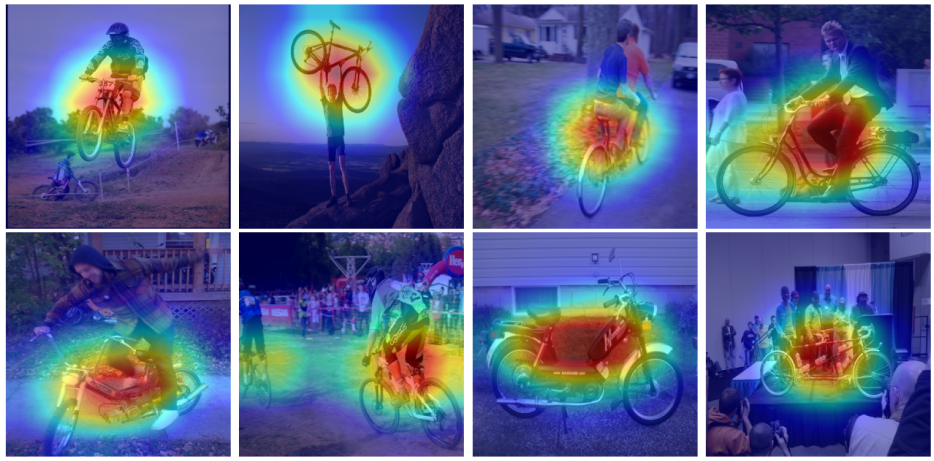


Figure 3.14: Heatmaps of the Inception-V3 model on ImageNet images from the bike class

Figure 3.15: Heatmaps of our DAE model on ImageNet images from the bike class. Inception-V3 focuses on the object from the top-classes, such as the bike, and hereby ignores the rest parts of the image, which is suboptimal for evaluating the image plausibility.