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 CA Leiden, The Netherlands
{t.h.w.baeck,h.wang}@liacs.leidenuniv.nl

In this supplementary material, we offer additional details on the model structure and datasets, qualitative examples, and expanded experiments that couldn't be included in the main paper due to space constraints.

1 Model Architecture

1.1 DAE trained on ImageNet

In Fig. 1, we illustrate the architecture of the Denoising Autoencoder (DAE) [9] trained on the ImageNet [2] dataset. The features extracted from the encoder are utilized to compute the Fréchet Denoised Distance (FDD) score.

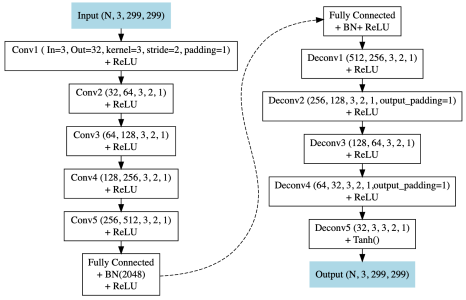


Fig. 1: Architecture of the Denoising Autoencoder Model trained on ImageNet, *left*: encoder, *right*: decoder.

1.2 DAE trained on BIKED

We also present the architecture of a DAE model trained on the BIKED [5] dataset in Fig. 2. This model shares a similar backbone with the one described in Section 1, with an input shape of $256 \times 256 \times 1$ and a smaller latent space dimension of $D_{\boldsymbol{w}} = 64$. The features extracted from the encoder in this model are employed to calculate the FDD (BIKED) metric.

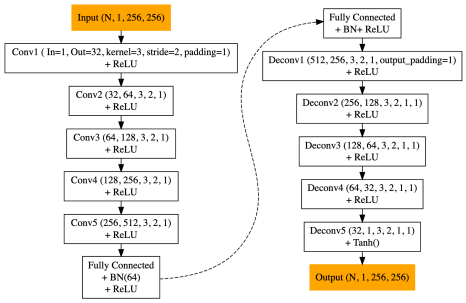


Fig. 2: Architecture of the Denoising Autoencoder Model DAE trained on BIKED, *left*: encoder, *right*: decoder.

2 Datasets

Below we list the details of the implemented datasets:

ImageNet We employ a subset of the ImageNet [2] dataset of 50 000 samples with dimension $299 \times 299 \times 3$, properly chosen to cover a wide range of 1k classes. The dataset is divided into 45 000 training samples and 5 000 test samples and is implemented for training the DAE model.

BIKED The BIKED [5] dataset is a compilation of 4 512 unique bicycle designs, contributed by various designers. The images are preprocessed into gray-scaled images with a resolution of 256×256. We have allocated 1 000 images for testing, 100 images for validation, and the remaining 3 412 images for training purposes.

Seeing3DChairs For our study, we also employ the Seeing3DChairs [1] dataset of 1 477 chair designs. For each chair design, there exists a set of images sampled from 62 consecutive viewpoints. We focus on the chair images with viewpoint numbers between 017-021. Hereby, we collect 6 970 samples, from which 100 images are utilized for validation, 1 000 are used for test and rest serve as training data.

Flickr-Faces-HQ (FFHQ) Our study incorporates a subset of the Flickr-Faces-HQ (FFHQ) [4] dataset, which contains over 70 000 high-resolution color images of human faces. Specifically, we select 1 000 samples from the FFHQ subset with a resolution of $256 \times 256 \times 3$.

3 Reconstruction with Fréchet Denoised Distance

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 [2]

Fig. 3: DAE reconstruction of images from ImageNet. *Top*: original images, *Middle*: noised images, *Bottom*: reconstructed images.

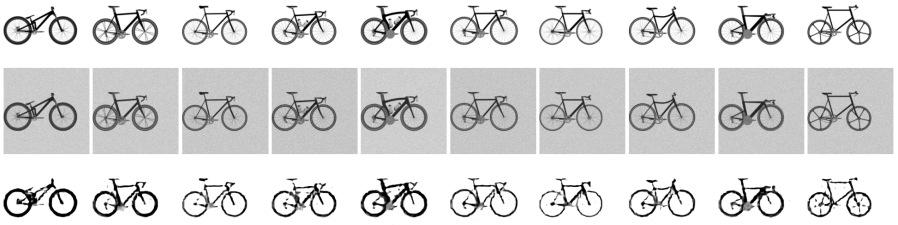


Fig. 4: DAE reconstruction of images from BIKED. *Top*: original images, *Middle*: noised images, *Bottom*: reconstructed images.

in Fig. 3, BIKED [5] in Fig. 4, Seeing3DChairs [1] in Fig. 5, and FFHQ [4] in Fig. 6. 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.

4 Topology Distance Computation

The methodology introduced by Horak et al. [3] presents a novel approach for assessing the similarity between feature spaces of real and synthetic datasets. It involves analyzing the topological properties of finite subsets of data points, denoted as F_r (features from real images) and F_g (features from generated images), sampled from their respective manifolds in $\mathbb{R}^m$. Through comparative analysis of these manifolds, the study explores the topological structure of the sampled data points.

Considering sets F_r and F_g with n data points each, distances among vectors in F_r and F_g are sorted in ascending order, leading to filtrations denoted as $VR(F_r)$ and $VR(F_g)$. These filtrations capture the connectivity evolution of the corresponding Vietoris-Rips complexes.

The persistent diagram obtained comprises n pairs (b_i, d_i), where b_i marks the point of initial appearance of observed homology groups, and d_i indicates

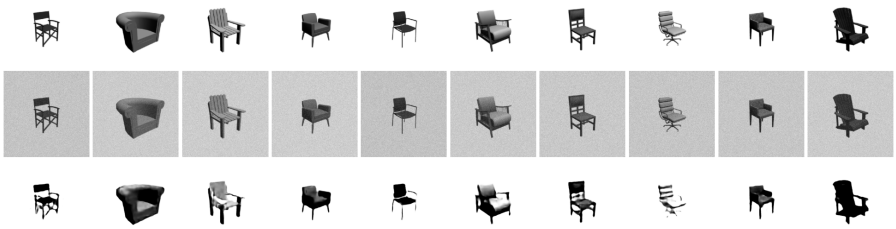


Fig. 5: DAE reconstruction of images from Seeing3DChairs. *Top*: original images, *Middle*: noised images, *Bottom*: reconstructed images.

Fig. 6: DAE reconstruction of images from FFHQ. *Top*: original images, *Middle*: noised images, *Bottom*: reconstructed images.

merging points or ∞ otherwise, for both F_r and F_g. Subsequently, longevity vectors $l(F_r)$ and $l(F_g)$ are defined, representing the sorted lifetimes of homology groups for F_r and F_g respectively. The Topology Distance (TD) between persistent diagrams, and hence between image collections, is computed as the l_2 distance between their longevity vectors:

$$TD(F_r, F_g) = \|l(F_r) - l(F_g)\|_2. \tag{1}$$

To compute the persistence homology and the persistence diagram, we opt for the `giotto-tda` library [8] .

5 Levels of disturbances

To provide a visualization of the varying degrees of disturbances applied to the datasets in assessing the sensitivity of the evaluation metrics, we plot sample images for each disturbance in Fig. 7 and Fig. 8. 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.

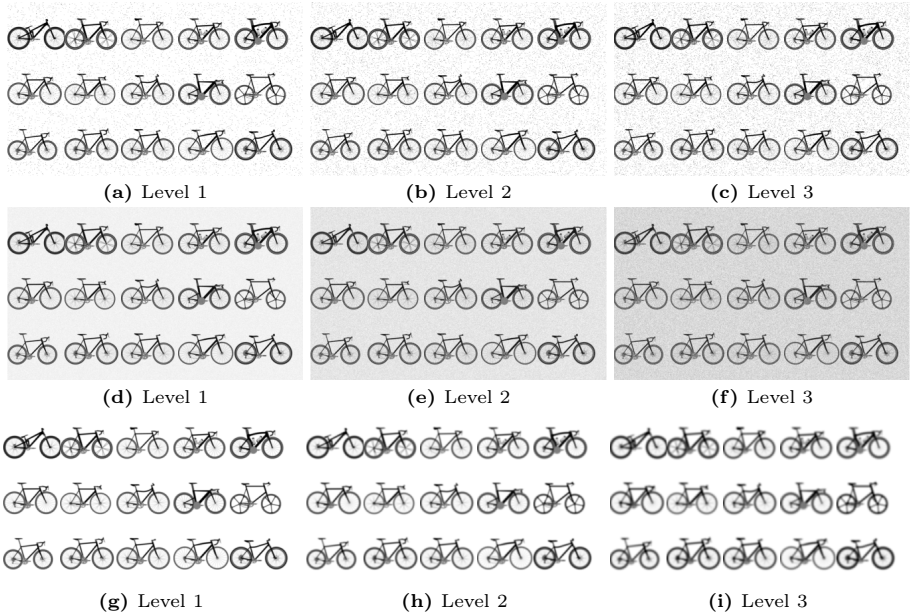
Patch Swap Using the same patch division approach as the Patch Mask, we randomly select a subset of patches, indicated by $\alpha \in [0, 0.25, 0.5, 0.75]$, and swap their positions pair-wisely.

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 $\alpha \in [0, 4, 5, 6]$.

6 Sensitivity Test on 1k images: Evaluation with FID, $\text{FD}_{\text{DINO-V2}}$, TD, and FDD

In this section, we conduct a sensitivity test on 1 000 images simultaneously, instead of dividing the dataset into groups. We record the observed scores for the FID, $\text{FD}_{\text{DINO-V2}}$, TD, and FDD metrics under the following disturbances: Salt & Pepper noise (SP), Gaussian noise (GN), patch masks, patch swap, and a combination of masks and Gaussian noise. The same test is applied to the BIKED [5] in Fig. 9a, Seeing3DChairs [1] in Fig. 9b, and FFHQ [4] in Fig. 9c.

In Fig. 9a, FID seems to perform with larger sample sizes better than it does with smaller sizes. However, it remains highly sensitive to noises, shown



(a) Level 1 **(b)** Level 2 **(c)** Level 3

(d) Level 1 **(e)** Level 2 **(f)** Level 3

(g) Level 1 **(h)** Level 2 **(i)** Level 3

Fig. 7: *Top*: Salt & Pepper Noise, *Middle*: Gaussian Noise, and *Bottom*: Gaussian Blur.

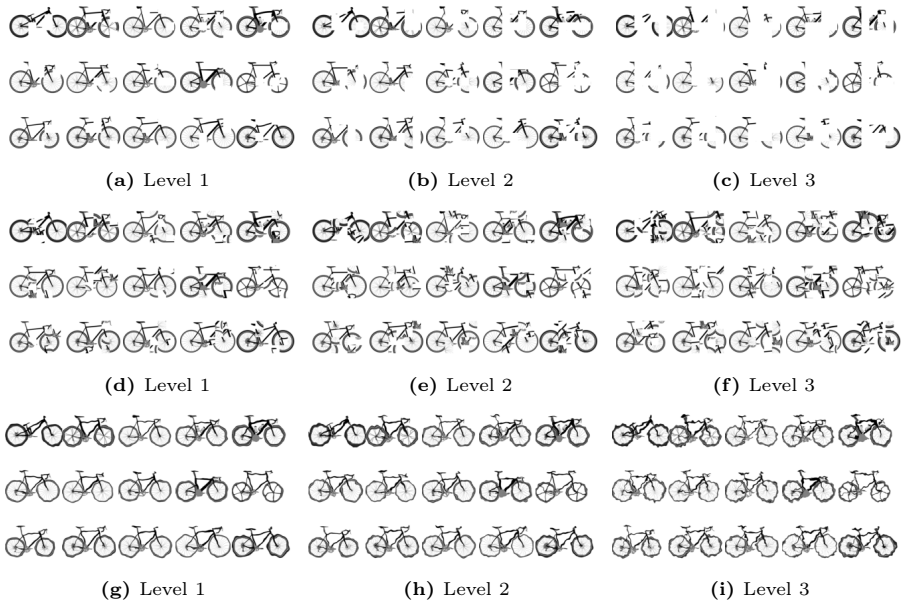
in GN+Swap, and still encounters failure when it is applied on other datasets, shown in Fig. 9b and Fig. 9c.

According to the results of sensitivity test, TD is the most competitive candidate against our FDD, whereas FID and $\text{FD}_{\text{DINO-V2}}$ do not exhibit a competitive performance.

7 Grad-CAM

In this section, we conduct the additional Grad-CAM [6] experiment on images selected from the Bike class of ImageNet [2] dataset. Note that in this experiment, both Inception-V3 model and the DAE model are supposed to look at all objects in the whole image, instead of focusing at one object as the observation of GRAD-CAM experiment on BIKED in the main paper. We first simply test the Grad-CAM on the Inception-V3 [7] model.As the heatmaps in Figure 10 illustrate, the Inception-V3 model effectively detects and localizes the bicycle within the image. However, it overlooks other regions of the image that may also contribute to the overall evaluation. On the other hand, as shown in Fig. 11, the DAE model trained on ImageNet is able to capture the structural information present in the input image, thereby enriching the evaluation process with a more comprehensive understanding of the scene.

As an extension of the Grad-CAM visualizations presented in the main paper, we provide additional heatmaps for the BIKED dataset in Fig. 12 and Fig. 13.



(a) Level 1 **(b)** Level 2 **(c)** Level 3

(d) Level 1 **(e)** Level 2 **(f)** Level 3

(g) Level 1 **(h)** Level 2 **(i)** Level 3

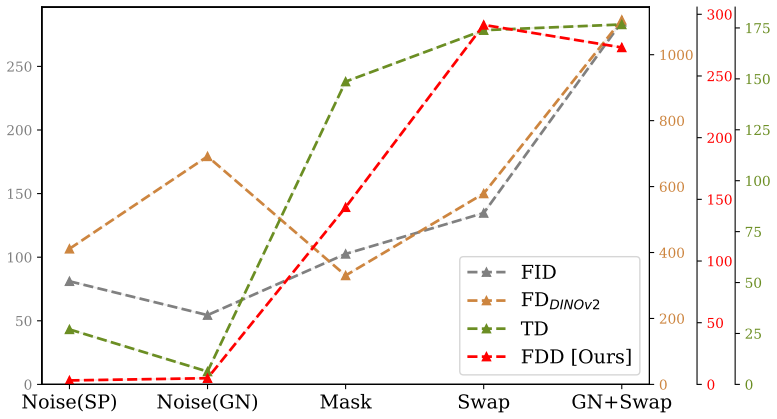
Fig. 8: *Top*: Mask, *Middle*: Patch Swap, and *Bottom*: Elastic Transformation.

These supplementary visualizations offer further insight into the model's attention.

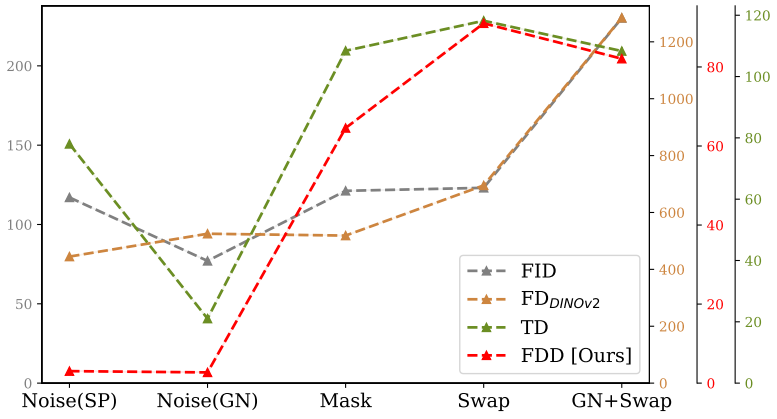
References

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. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255. IEEE (2009), https://ieeexplore.ieee.org/abstract/document/5206848/
3. 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, https://doi.org/10.1609/aaai.v35i9.16943
4. 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)
5. Regenwetter, L., Curry, B., Ahmed, F.: BIKED: A Dataset for Computational Bicycle Design With Machine Learning Benchmarks. Journal of Mechanical Design

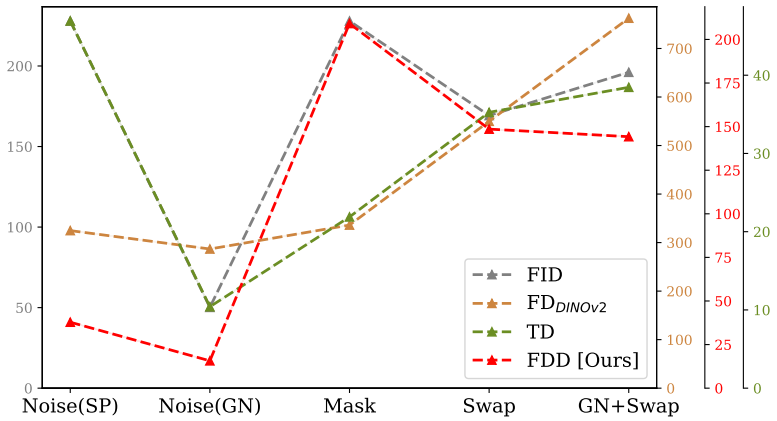
144(3) (10 2021). https://doi.org/10.1115/1.4052585, https://doi.org/10.1115/1.4052585, 031706
6. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
7. 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
8. Tauzin, G., Lupo, U., Tunstall, L., Pérez, J.B., Caorsi, M., Medina-Mardones, A.M., Dassatti, A., Hess, K.: giotto-tda:: A topological data analysis toolkit for machine learning and data exploration. Journal of Machine Learning Research **22**(39), 1–6 (2021)
9. 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



(a) BIKED



(b) Seeing3DChairs



(c) FFHQ

Fig. 9: Sensitivity Test with 1k samples

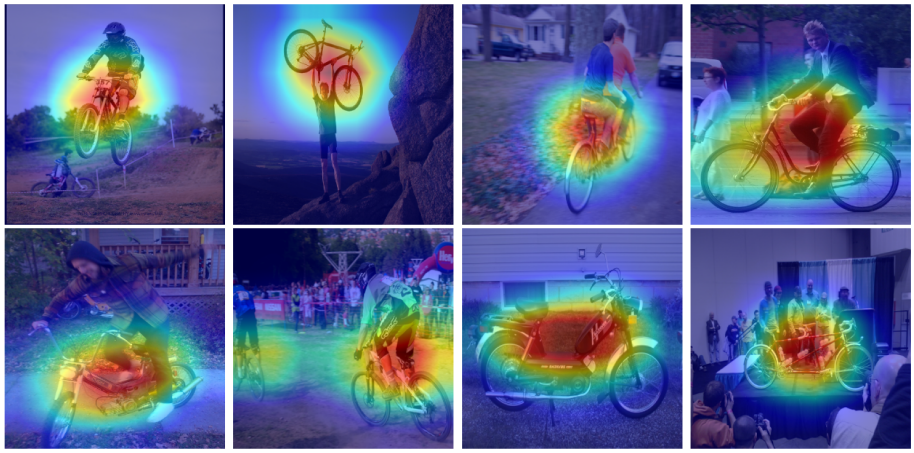


Fig. 10: Heatmaps of the Inception-V3 model on ImageNet images from the bike class

Fig. 11: 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.

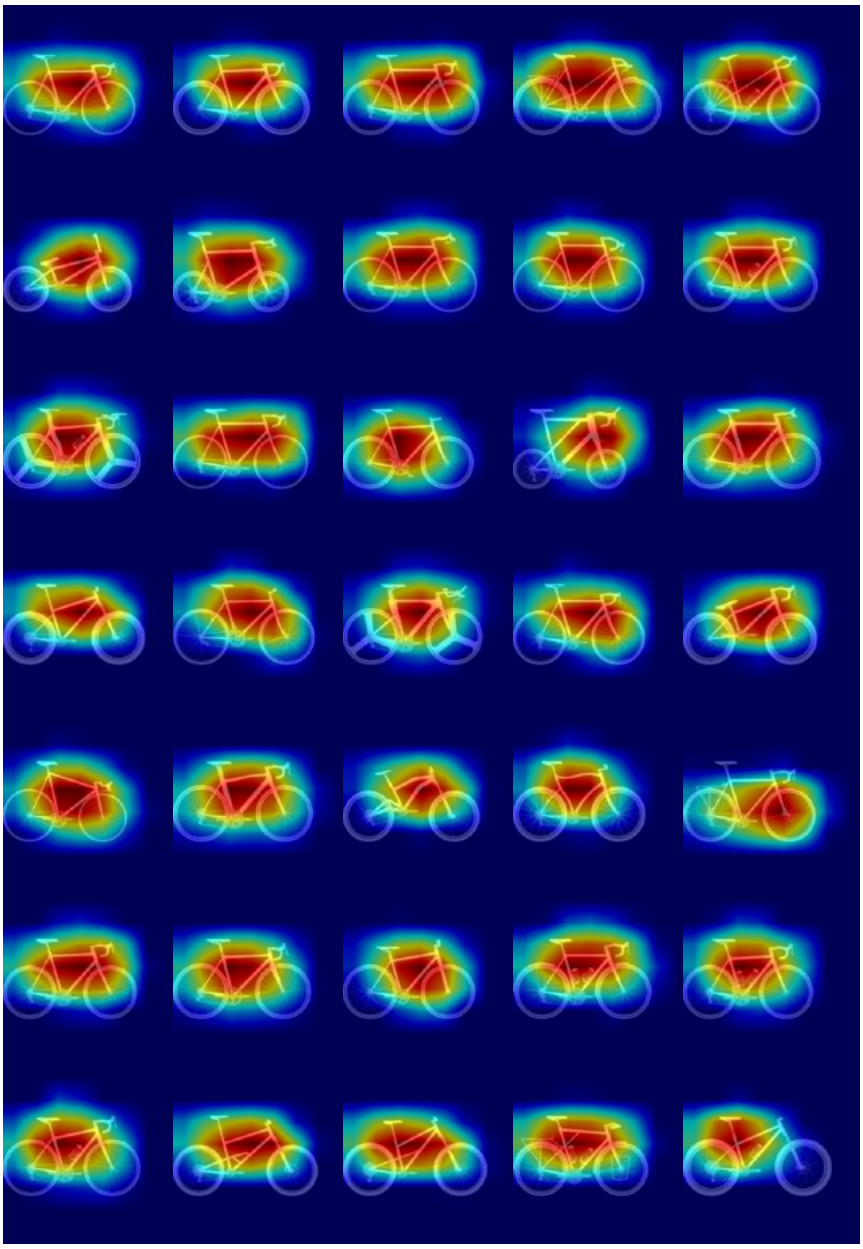


Fig. 12: Heatmaps of Inception-V3 model on BIKED

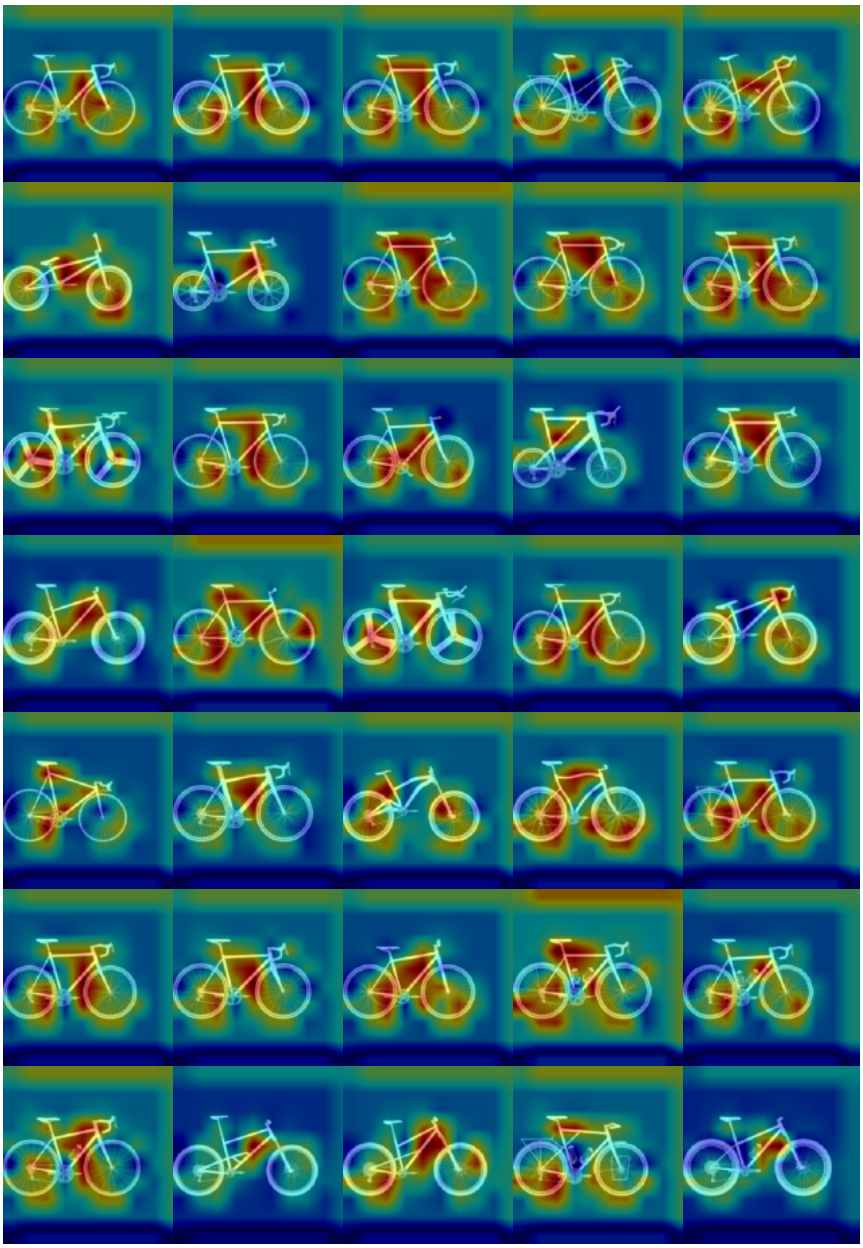


Fig. 13: Heatmaps of our DAE model on BIKED. Unlike the Inception-V3 model trained for classification, the DAE model draw out structural outlines in the heatmap, thereby providing the basis for evaluation scores