Visual DNA: Representing and Comparing Images using Distributions of Neuron Activations

Benjamin Ramtoula Matthew Gadd Paul Newman Daniele De Martini
Mobile Robotics Group, University of Oxford
{benjamin, mattgadd, pnewman, daniele}@robots.ox.ac.uk

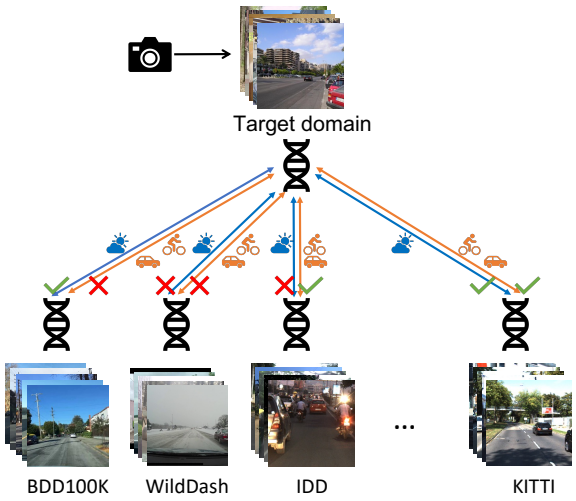
Abstract

Selecting appropriate datasets is critical in modern computer vision. However, no general-purpose tools exist to evaluate the extent to which two datasets differ. For this, we propose representing images – and by extension datasets – using Distributions of Neuron Activations (DNAs). DNAs fit distributions, such as histograms or Gaussians, to activations of neurons in a pre-trained feature extractor through which we pass the image(s) to represent. This extractor is frozen for all datasets, and we rely on its generally expressive power in feature space. By comparing two DNAs, we can evaluate the extent to which two datasets differ with granular control over the comparison attributes of interest, providing the ability to customise the way distances are measured to suit the requirements of the task at hand. Furthermore, DNAs are compact, representing datasets of any size with less than 15 megabytes. We demonstrate the value of DNAs by evaluating their applicability on several tasks, including conditional dataset comparison, synthetic image evaluation, and transfer learning, and across diverse datasets, ranging from synthetic cat images to celebrity faces and urban driving scenes.

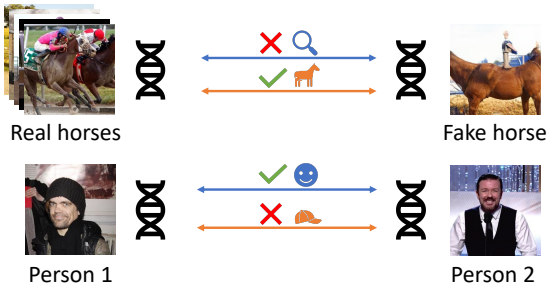
1. Introduction

Being able to compare datasets and understanding how they differ is critical for many applications, including deciding which labelled dataset is best to train a model for deployment in an unlabelled application domain, sequencing curricula with gradually increasing domain gap, evaluating the quality of synthesised images, and curating images to mitigate dataset biases.

However, we currently lack such capabilities. For example, driving datasets available covering many domains [2, 4, 9, 16, 17, 22, 35, 54, 62, 64] were collected under diverse conditions typically affecting image appearance (*e.g.* location, sensor configuration, weather conditions, and post-processing). Yet, users are limited to coarse or insufficient meta-information to understand these differences. Moreover, depending on the application, it might be desirable to



(a) When deploying a vision model in a new target domain, selecting models pre-trained on the most relevant datasets can help. However, there are no methods to measure dataset similarities automatically. A general distance between datasets would be sensitive to many variation types, but DNAs provide sufficient granularity to customise the comparison to focus on features of interest. For example, DNA comparisons can be customised to ignore weather conditions or focus on semantic content



(b) The DNA can also be used to compare individual images to datasets – for example, to measure the realism and semantic consistency of a synthetic image – or pairs of images – for example, to verify the presence of similar attributes such as smiling or wearing a hat.

Figure 1. Example use-cases of the DNA representation.

Project page and code: bramtoula.github.io/vdna

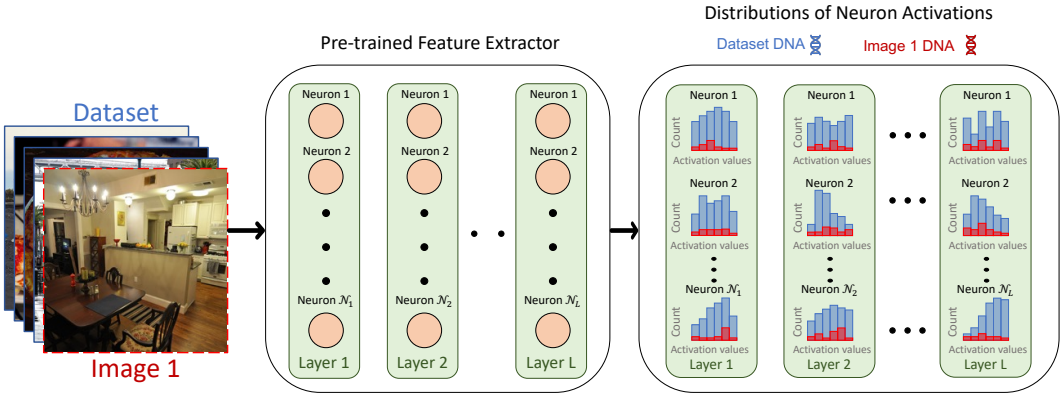


Figure 2. We propose representing images by passing them through a pre-trained frozen feature extractor network and collecting neuron activations. We then create a descriptor called the Distribution of Neuron Activations (DNA) by fitting a distribution (the histogram in the illustration) to the activations at each neuron. We can quantitatively measure the similarity of different datasets or images by comparing their DNAs. Neuron combination strategies that are sensitive to specific attributes can also allow for customised comparisons of DNAs.

compare datasets *only* on controlled sets of attributes while ignoring others. For self-driving, these may be weather, road layout, driving patterns, or other agents' positions.

We propose representing datasets using their Distributions of Neuron Activations (DNAs), allowing efficient and controllable dataset and image comparisons (Fig. 1). The DNA creation exploits the recent progress in self-supervised representation learning [13, 18] and extracts image descriptors directly from *patterns* of neuron activations in neural networks (NNs). As illustrated in Fig. 2, DNAs are created by passing images through an off-the-shelf pre-trained frozen feature extraction model and fitting a distribution (*e.g.* histogram or Gaussian) to the activations observed at each neuron. This DNA representation contains multi-granular feature information and can be compared while controlling attributes of interest, including low-level and high-level information. Our technique was designed to make comparisons easy, avoiding high-dimensional feature spaces, data-specific tuning of processing algorithms, model training, or any labelling. Moreover, saving DNAs requires less than 15 megabytes, allowing users to easily inspect the DNA of large corpora and compare it to their data before committing resources to a dataset. We demonstrate the results of using DNAs on real and synthetic data in multiple tasks, including comparing images to images, images to datasets, and datasets to datasets. We also demonstrate its value in attribute-based comparisons, synthetic image quality assessment, and cross-dataset generalisation prediction.

2. Related Works

2.1. Studying Image Datasets

Early dataset studies focused on the limitations of the datasets available at the time. Ponce *et al.* [43] highlighted the need for more data, realism, and diversity, focusing on object recognition and qualitative analysis (*e.g.* "average" images for each class). Torralba and Efros [56] found evidence of significant biases in datasets by assessing the ability of a classifier to recognise images from different datasets and measuring cross-dataset generalisation of object classification and detection models. Nowadays, datasets abound, and the approaches used to compare them in those early works would be prohibitive to scale or generalise, often requiring training models for each dataset of interest and access to labels.

Compressed datasets representations allow learning models with comparable properties with reduced dataset sizes. Dataset distillation approaches [59] *synthesise* a small sample set to approximate the original data when used to train a model. Core-set selection approaches [15], instead, *select* existing samples, with image-based applications including visual-experience summarisation [42] and active learning [50]. While achieving compression of important data properties, these approaches do not produce representations that allow easy dataset comparisons, as our DNA does. Modelverse [34] performs a content-based search of generative models. Similarly to DNAs, they represent multiple datasets – generated by different generative models – using distribution statistics of extracted features from the images. However, their work does not focus on granular and controllable comparisons but on matching a query to the closest distribution.

Synthetic data evaluation for generative models such as Generative Adversarial Networks (GANs) [3] is usually framed as a dataset comparison problem, measuring a distance between datasets of real and fake images. One of the most widely used metrics is the Fréchet Inception Distance (FID) [21], which embeds all images into the feature space of a specific layer of the Inception-v3 network [55]. A multivariate Gaussian is fit to each real and fake embed-

ding, and the Fréchet distance (FD) between these distributions is computed. The Kernel Inception Distance (KID) [1] is another popular approach, which computes the squared maximum mean discrepancy between Inception-v3 embeddings. There are many other variations, such as using precision and recall metrics for distributions [10, 30, 49, 51], density and coverage [37], or rarity score [19]. These approaches rely on high-dimensional features from one layer, while our approach considers neuron activations across layers. Furthermore, while these measure dataset differences, they have mainly been employed to compare real and synthetic datasets within the same domain, not real ones with significant domain shifts. Moreover, recent evidence suggests the embeddings typically used can cause a heavy bias towards ImageNet class probabilities [29], motivating more perceptually-uniform distribution metrics. Additionally, these high-dimensional embeddings make gathering information about specific attributes of interest challenging and lead to computational issues (*e.g.* when clustering).

2.2. Representation Learning

Feature extractors can provide useful multi-granular features (*e.g.* containing information about low-level lighting conditions but also the high-level semantics), motivating our design of DNAs. Work on the interpretability of NNs supports this assumption. Indeed, Olah *et al.* [5, 39] explored the idea that NNs learn features as fundamental units and that analogous features form across different models and tasks. Neurons can react more to specific inputs, such as edges or objects [40]. Combining neuron activations from several images can be a good way to investigate what a network has learned through an activation atlas [8].

Existing uses of pre-trained feature extractors include evaluating computer vision tasks such as Inception-v3 features for FID and KID, as above. Pre-trained networks on large datasets also provide generally useful representations [24], which are often fine-tuned for specific applications. Notably, Evci *et al.* [14] found that selecting features from subsets of neurons from all layers of a pre-trained network allows better fine-tuning of a classifier head for transfer learning than using only the last layer, suggesting that relevant features are accessible by selecting appropriate neurons. Moreover, a pre-trained VGG network [52] has been used to improve the perceptual quality of synthetic images [46, 60] or to judge photorealism [46, 65].

Self-supervised training relies on pretext tasks, foregoing labelled data and learning over larger corpora, yielding better representations [24]. Morozov *et al.* [36] showed that using embeddings from self-supervised networks such as a ResNet50 [20] trained with SwAV [6] leads to FID scores better aligned to human quality judgements when evaluating generative models. We explore different feature extractors but exploit a ViT-B/16 [11] trained with Mugs [67] by default, a recent multi-granular feature technique.

2.3. The need for a more general tool

FID [21] or KID [1] use representation learning to tackle similar tasks to ours; yet, our formulation extends their applicability. Quantitative and holistic comparisons between different real datasets have been overlooked, despite being critical to tasks such as transfer or curriculum learning. We argue that a general data-comparison tool must allow selecting attributes of interest after having extracted a reasonably compact representation of the image(s) and permit the user to *customise* the distance between representations.

3. DNA - Distributions of Neuron Activations

Our system is designed around the principle of decomposing images into simple conceptual building blocks that, in combination, constitute uniqueness. Yet, to cover all possible axes of variations, it is infeasible to specify those building blocks manually. While we cannot usually link each neuron of a NN to a human concept [40], we show that they provide a useful granular decomposition of images.

Keeping track of neuron activations independently allows us to combine their statistics and study conceptually-meaningful attributes of interest. As the activations at each neuron are scalar, they can easily be gathered in 1D histograms or univariate Gaussians. While we would ideally track dependencies between neurons, this is too costly to include in our representation. Nevertheless, we show experimentally that many applications still benefit from DNAs.

3.1. Distribution choice

As in Fig. 2, in this section, we formulate DNAs using histograms to fit each neuron's activations distribution. Histograms are a good choice because they do not make assumptions about the underlying distribution; however, we can also consider other distribution approximations. We also experiment with univariate Gaussians to approximate the activations of each neuron and produce a DNA, allowing us to describe distributions with only two parameters. We denote versions using histograms and Gaussians as $\text{DNA}^{\text{hist.}}$ and $\text{DNA}^{\text{Gauss.}}$, respectively.

3.2. Generating the DNA from images

We consider a dataset of images $\mathcal{I}$ where $|\mathcal{I}| \geq 1$. We also have a pre-trained feature extractor $\mathcal{F}$ with L layers $\mathcal{L}$ manually defined as being of interest, and a set $\mathcal{N}$ of all neurons in those layers. Each layer $l \in \mathcal{L}$ is composed of N_l neurons, each producing a feature map of spatial dimensions $S_h^l \times S_w^l$. We can perform a forward pass $\mathcal{F}(i)$ of an image $i \in \mathcal{I}$ and observe the feature map f_i^l obtained at each layer l which has dimensions $N_l \times S_h^l \times S_w^l$. For each neuron $n \in \mathcal{N}$ fed with image i, we define a histogram

$h_i^n \in \mathbb{N}^B$ with B pre-defined uniform bins where bin edges are denoted $b_0, b_1, \ldots, b_B$. The count $h_i^n[k]$ in the bin of index k for neuron n of layer l is found by accumulating over all spatial dimensions that fall within the bin's edges:

$$h_i^n[k] = \sum_{s_h}^{S_h^l} \sum_{s_w}^{S_w^l} \begin{cases} 1, & \text{if } f_i^l[n, s_h, s_w] \in [b_k, b_{k+1}) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The resulting image's DNA$^{\text{hist.}}$ can then be accumulated to represent the dataset $\mathcal{I}$ as $H = \{\mathcal{H}^n\}$ for each neuron $n \in \mathcal{N}$, where the element k of $\mathcal{H}^n$ can be calculated as:

$$\mathcal{H}^n[k] = \sum_{i \in \mathcal{I}} h_i^n[k] \quad (2)$$

3.3. Comparing DNAs

Now, comparing DNAs$^{\text{hist.}}$ reduces to comparing 1D histograms for neurons of interest. Depending on the use case, different distances can be considered. Some tasks might need a distance to be asymmetrical and keep track of original histogram counts, while normalised counts and symmetric distances might be more appropriate for others.

To demonstrate straightforward uses of the representation, we experiment with a widely accepted histogram comparison metric: the Earth Mover's Distance (EMD). Earlier works have argued that the EMD is a good metric for image retrieval using histograms [48], with examples of retrieval based on colour and texture. This work uses its normalised version, equivalent to the Mallows or first Wasserstein distance [32]. The EMD can be interpreted as the minimum cost of turning one distribution into another, combining the amount of distribution mass to move and the distance. Specifically, given the normalised, cumulative histogram

$$\mathcal{F}^n[k] = \sum_{j=0}^{k} \frac{\mathcal{H}^n[j]}{\|\mathcal{H}^n\|_1} \text{ where } k \in 0, \ldots, B-1 \quad (3)$$

we can easily compute the EMD between two histograms:

$$\text{EMD}(\mathcal{H}_1^n, \mathcal{H}_2^n) = \sum_{k=0}^{B-1} |\mathcal{F}_1^n[k] - \mathcal{F}_2^n[k]| \quad (4)$$

As every neuron n is independent in the EMD formulation, we can vary the contribution of each to the calculation of the total distance. This allows us to treat neurons differently, *e.g.* when wanting to customise the distance to ignore specific attributes, as presented in Sec. 5.3. For this, we introduce the use of a simple linear combination of the histograms through scalar weights W^n for each neuron n:

$$\text{EMD}_W(H_1, H_2) = \sum_{n \in \mathcal{N}} W^n \text{EMD}(\mathcal{H}_1^n, \mathcal{H}_2^n) \quad (5)$$

In the special case of $W^n = 1/|\mathcal{N}|$, we obtain EMD_{avg} as the average EMD over individual neuron comparisons.

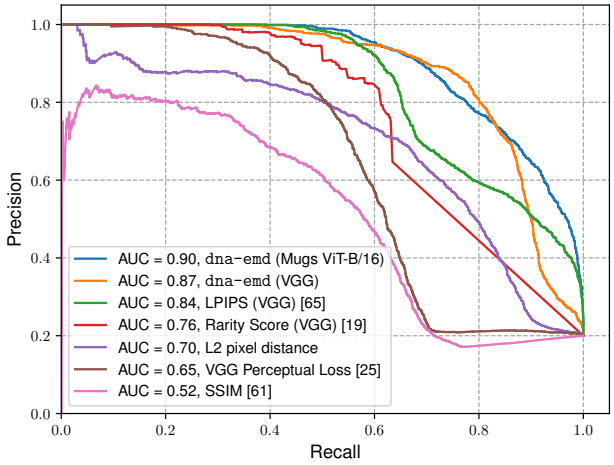


Figure 3. Precision-Recall curves and Area Under the Curve (AUC) for retrieving 2000 individual augmented Cityscapes images mixed with 8000 images from four other datasets. The retrieval compares these images to a reference set of 500 distinct non-augmented Cityscapes images.

4. Experimental settings

Our experiments evaluate the DNA's efficacy on various tasks, datasets, and with diverse feature extractors.

Datasets and weights We use real and synthetic images from very varied domains – comparing pairs of images, pairs of datasets, and individual images to datasets. Images are processed using tools provided by Parmar *et al.* [41]. Notably, no additional tuning is performed for any experiments, i.e. *the feature extractors' weights are frozen*.

Settings Activation ranges vary for different neurons; it is therefore important to adjust the histogram settings for each neuron to get balanced distances between neurons. We thus monitor each neuron's activation values over a large set of datasets and track the minimum and maximum values observed, adding a margin of 20% and using these extremes to normalise activations between -1 and 1. Notably, the *only* hyperparameter for the DNA$^{\text{hist.}}$ is the number of bins, B, which we set to 1000 for our experiments.

Benchmarking Our primary baseline, `fd`, is the Fréchet distance [21], which measures the distance of two multivariate Gaussians that fit the samples in the embedding space of entire layers of the extractor. We use the acronym "Fréchet distance (FD)" rather than "Fréchet Inception Distance (FID)" as we explore different feature extractors than Inception-v3. Traditionally, `fd` has been used on a single layer, but we show its performance on different combinations of the extractor layers. Here, `dna-emd` denotes EMD comparisons of our DNAs$^{\text{hist.}}$, and `dna-fd` denotes FD comparisons of our DNAs$^{\text{Gauss.}}$. These three settings allow us to verify our approach, showing the effectiveness of

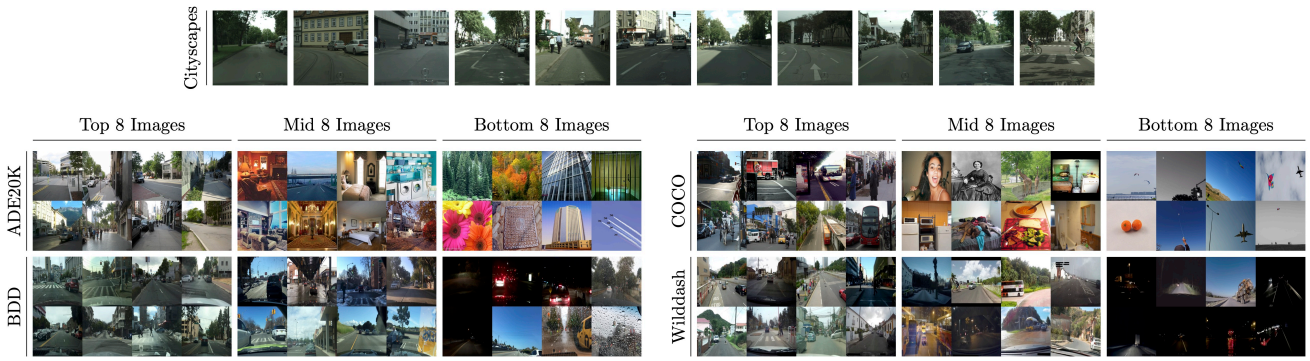


Figure 4. Images from different datasets organised by `dna-emd` over all neurons against Cityscapes. We successfully discriminate based on the scene type and its visual aspect. COCO and ADE20K images get poorly ranked when they do not contain city street scenes. The visual aspect and image quality are also considered, as seen with poorly ranked images when Wilddash images have different lighting or BDD images have challenging conditions such as rain on the windshield. Images ranked in the middle for BDD tend to contain scenes without any obvious anomaly but with brightness and colours further from Cityscapes images than better-ranked images.

considering every neuron as independent *and* not constraining the activations to fit specific distributions.

We do not consider the Kernel Distance [1] as it is unclear how to compress the required information in a compact representation.

Memory footprint We provide details on memory complexity and DNA storage size in Tab. 1.

Method	Complexity	Theoretical size	Observed size
Features	$N \times n \times S$ floats	1.10 TB	-
Spatially averaged features	$N \times n$ floats	5.59 GB	2.91 GB
$\text{DNA}^{\text{Gauss.}}$	$2 \times n$ floats	159 kB	84.7 kB
$\text{DNA}^{\text{hist.}}$	$B \times n$ ints	79.9 MB	14.8 MB

Table 1. Memory footprint of data for N images, n feature extractor neurons with an average of S elements in their feature maps, and B bins. Example with FFHQ (89.1 GB, N=70k), Mugs ViT-B/16 (n=9984, S=197) and B=1000 bins. Observed sizes are from files saved using NumPy's `savez_compressed` function.

5. Results

5.1. Finding most similar images with domain shifts

We first show the ability of DNAs to find real images similar to a reference dataset. We have created two datasets: a *reference* D_r contains 500 random Cityscapes [9] images; a *comparison* D_c contains 2000 images from each of ADE20K [66], BDD [62], COCO [33], Wilddash [64] as well as 2000 randomly augmented images (*e.g.* noise, blur, spatial and photometric) from Cityscapes that are not present in D_r. We rank each image from D_c in terms of its distance to D_r as a whole. We expect the top-ranked D_c images to all be Cityscapes augmentations. We compare the use of `dna-emd` using the EMD_{avg} (Sec. 3.3) to other perceptual comparison baselines: a perceptual loss [25], LPIPS [65], SSIM [61], the L2 pixel distance, and rarity score [19]. All approaches are evaluated using features from VGG [52]. For approaches comparing image pairs (all except ours & rarity score), we define the distance for one D_c image to D_r as its average distance to each image in D_r.

Fig. 3 shows DNAs performing best at this task *while not requiring expensive pairwise image comparisons*. Results are further improved when using features from a self-supervised approach, Mugs [67], instead of VGG.

We also verify qualitatively how images from different datasets from D_c are ranked when compared to D_r using Mugs features. Fig. 4 shows the successful discrimination of scene types and visual aspects in all comparisons.

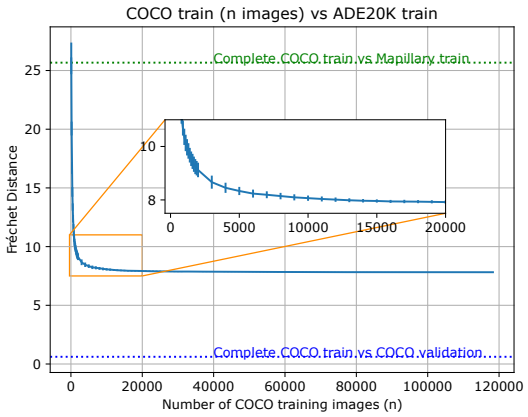
5.2. Number of images required for dataset DNAs

`fd` is known to work poorly with scarce data [1]. We assess this in Fig. 5, comparing the distance between the entire ADE20K training set [66] and increasingly larger subsets of COCO's training set [33]. Here, `fd` reaches a steady value after 10000 samples while `dna-emd` needs only 400, making it a reliable representation even for small datasets.

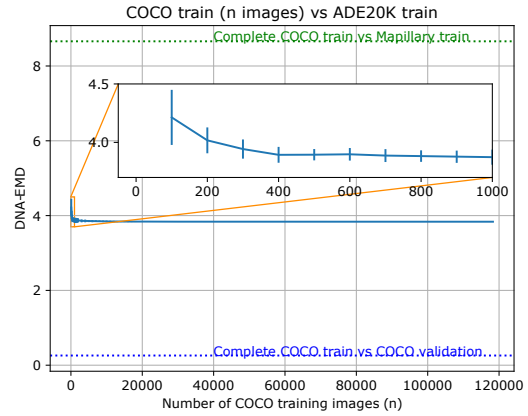
5.3. Ignoring specific attributes

Here, we demonstrate the granularity provided by individual neurons by considering: given DNAs of two datasets, can we measure the distance between them while ignoring contributions due to specifically-selected attributes?

Attribute datasets For this experiment, we split the CelebA training set according to each of the 40 labelled attributes, *e.g. smiling* or *wearing a hat*, where A is the set of all attributes. We use the "in the wild" version of images which are not cropped and aligned around faces, allowing us to assess robustness to different locations and scales of attributes. Considering one attribute $a \in A$, we compute two DNAs, $\mathcal{D}_a$, with images *with* the attribute, and $\mathcal{D}_{\bar{a}}$, with images *without* the attribute. Neurons whose distributions



(a) Fréchet Distance (`fd`)



(b) Ours (`dna-emd`)

Figure 5. Influence of the number of images on dataset distances using DINO (ViT-B/16) features. `dna-emd` requires significantly fewer samples than `fd` from the COCO training set to reach a steady value when compared with ADE20K's training set. Here, the distances using all images are non-zero, as there is a domain shift between the datasets. Dashed lines illustrate the distances obtained comparing COCO's training set to its validation set and to ADE20K to illustrate the scale of the error. Results are averaged over ten seeds, with vertical lines showing the standard deviations.

vary greatly between these DNAs – *i.e.* are *sensitive* – correlate with the attribute.

Learned sensitivity removal and deviation We input the neuron-wise (for `dna-fd` and `dna-emd`) or layer-wise (for `fd`) distances between $\mathcal{D}_a$ and $\mathcal{D}_{\bar{a}}$ into a linear layer, which produces a weighted distance with which we can ignore differences of a specific attribute while maintaining sensitivity to the other attributes. Its parameters correspond to the weights W for the linear combination in Eq. (5). Next, we define the *sensitivity deviation* of attribute a. For `dna-emd`:

$$\Delta_a = 1 - \frac{\text{EMD}_W(\mathcal{D}_a, \mathcal{D}_{\bar{a}})}{\text{EMD}_{\text{avg}}(\mathcal{D}_a, \mathcal{D}_{\bar{a}})} \tag{6}$$

This and all the following calculations can be applied to `fd` and `dna-fd` with the FD. If a is the only attribute that changes between $\mathcal{D}_a$ and $\mathcal{D}_{\bar{a}}$, and W is optimised to ignore a, then the EMD_W should not be sensitive to a and $\text{EMD}_W(\mathcal{D}_a, \mathcal{D}_{\bar{a}}) = 0$, $\Delta_a = 1$. For instance, we have datasets at night and datasets at day but want to compare only considering the types of vehicles present. For attributes to which we want the distance to remain sensitive, $b \in A \setminus a$, we can also measure deviations from the original distance caused by the weights W using $|\Delta_b|$, indicating the change in sensitivity of the distance to this attribute. We want no deviation for these attributes, *i.e.* $|\Delta_b| = 0$. Finally, we impose (and back-propagate, using Adam [28]) a loss:

$$L_a = |1 - \Delta_a| + \frac{1}{|A| - 1} \sum_{b \in A \setminus a} |\Delta_b| \tag{7}$$

meaning that we will optimise W to desensitize the EMD_W to a but remain sensitive to all other attributes in $b \in A \setminus a$.

CelebA sensitivities Tab. 2 presents results as averaged over all attributes (*i.e.* with $a =$ `hat` and b being all other attributes, then $a =$ `glasses` and b all others, etc.). The results clearly show that neuron granularity is crucial for success as `fd`, which operates layer-wise, falls short against `dna-emd` and `dna-fd`. Averaged over all attributes, our approach can discard 95.5% of the distance over the attributes on which we remove sensitivity, while only causing a 9.6% of deviation in distances over other attributes. `dna-emd` performs slightly better than `dna-fd`, but both do very well. We observe that all feature extractors considered can somewhat succeed at the task, including the ResNet-50 with random weights, which we expect to still produce valuable features [45]. However, we obtain the best results using self-supervised models which are likely to produce more informative features.

Finding similar images Qualitatively, we expect neurons to react to general and consistent features, which should also apply to comparing image pairs and different datasets. To verify this, we compare image pairs from a different dataset, FFHQ [26], with and without specific attributes (`eyeglasses` and `wearing hat`) and select the neuron(s) with the highest `dna-emd` sensitivity on CelebA. Using the selected neurons, we compare the $\text{DNA}^{\text{hist.}}$ of a selected reference image to $\text{DNAs}^{\text{hist.}}$ of 2000 random FFHQ samples. We present our results in Fig. 6. We can verify that very few neurons are required to focus on high-level semantic attributes, even when selected on a different dataset. We still observe some errors, possibly due to neurons reacting to several attributes simultaneously.

5.4. Synthetic Data

Related systems have been used in the evaluation of synthetic image-creation techniques. We thus qualitatively investigate the use of `dna-emd` to evaluate the quality

Feature extractor	Mean target attribute a sensitivity removal Δ_a (%) ↗			Mean other attributes sensitivity deviation $\frac{1}{\lvert A\rvert-1}\sum_{b\in A\setminus a}\lvert\Delta_b\rvert$(%) ↘		
	Fréchet Distance	DNA-Fréchet Distance	DNA-EMD	Fréchet Distance	DNA-Fréchet Distance	DNA-EMD
Inception-v3 [55]	9.6	94.8	92.3	9.1	11.9	10.9
CLIP image encoder (ViT-B/16) [44]	20.1	93.7	94.3	17.6	7.2	7.4
Stable Diffusion v1.4 encoder [47]	-	87.7	81.4	-	19.4	19.3
Random weights (ResNet-50) [45]	11.6	72.1	83.4	10.1	33.0	20.1
DINO (ResNet-50) [7]	15.8	87.3	93.5	8.9	16.2	9.4
DINO (ViT-B/16) [7]	19.0	93.9	94.2	16.3	10.2	9.6
Mugs (ViT-B/16) [67]	20.0	93.7	95.5	16.7	10.3	9.6
Mugs (ViT-L/16) [67]	34.6	93.3	95.3	28.0	9.4	9.1
Mean	18.7	89.6	**91.2**	15.2	14.7	**11.9**

Table 2. Customising different dataset comparison techniques to be insensitive to specific attributes. For each of the 40 attributes in CelebA, we use a weighted combination of distances over different layers or neurons with weights optimised such that the resulting distance between images with and without the attribute becomes zero. This is captured in the "target attribute sensitivity removal", which measures the relative drop in distance. We must ensure that the distance remains sensitive to the other 39 attributes. The "other attributes sensitivity deviation" measures the relative deviation in the original distance caused by the customisation. We show averages over all attributes on the CelebA testing set. `fd` can only combine distances over individual layers, making it challenging to ignore some attributes while preserving others. On the other hand, neuron-wise metrics such as `dna-fd` and `dna-emd` provide sufficient granularity for customising the distance to ignore one attribute while preserving the others. We only consider the latent space of Stable Diffusion v1.4, which we treat as a single layer – hence we cannot perform a weighted combination of layers for the `fd` approach.

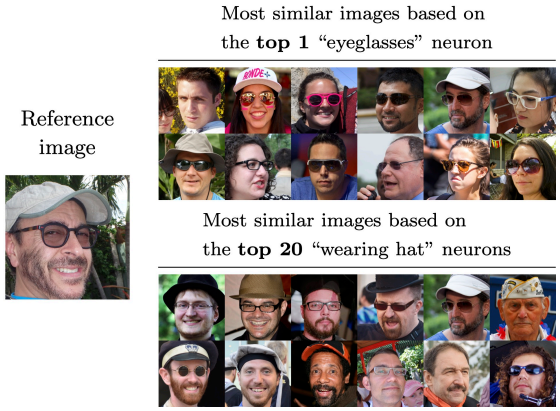


Figure 6. We seek to find the closest match to a reference image, on the left, from FFHQ, according to specific attributes – here, `wearing hat` and `eyeglasses`. To do so, we select the neurons with the highest attribute sensitivity from CelebA and use them for comparisons in FFHQ, demonstrating the generalisability of these neurons. We show that very few neurons suffice to recover images with `eyeglasses` and `wearing hat`.

– *i.e.* closeness to the distribution of real images – of StyleGANv2 [27] generated images. Here, we collect the DNAs$^{\text{hist.}}$ for the datasets of real and generated images containing various classes [26, 27, 63]. We use these to select the most sensitive neurons (as above in Fig. 6) to differences between `real` and `fake` images, which we expect to be good indicators of realism. These neurons are used to compare a separate dataset with generated images of one class not included in the datasets responsible for neuron selection – *e.g.* when evaluating realism for cars, we select neurons based on cats, horses, churches, and faces, focusing on general realism rather than car-specific features.

Our results are reported in Fig. 7. We clearly identify outliers in the generated samples using either selected or all neurons. However, when using all neurons, top matches do not always match our perceptual quality assessment. By selecting a small number of neurons reacting to realism, we favour images with fewer synthetic generation artefacts.

5.5. Generalisation prediction under domain shifts

Above, we have compared images and datasets from similar domains. However, many applications require comparing datasets from distinct domains. Here, we show the power of DNAs in cross-dataset generalisation prediction, which can serve, for instance, to select the best dataset for training when performing transfer learning.

For this, we compare our ranking of distances from dataset DNAs to the measured cross-dataset generalisation of a semantic-segmentation network reported in Tab. 3 of Lambert *et al.* [31]. This reference provides mIoUs from an HRNet-W48 [58] semantic segmentation model architecture trained on seven datasets: ADE20K [66], COCO [33], BDD100K [62], Cityscapes [9], IDD [57], Mapillary [38], and SUN-RGBD [53], and evaluated on all seven corresponding validation sets. Cross-generalization varies widely for different pairs of datasets, with mIoUs ranging from 0.2 (training on Cityscapes and validating on SUN-RGBD) to 69.7 (training on Mapillary and validating on Cityscapes). Therefore, for each validation set $v \in V$, we have a ranking of which dataset's training sets transferred best in terms of mIoU. We denote by $T_v^{\text{gt}}[i]$ the training set used by the model producing the i-th highest mIoU for validation set v, and by $\text{mIoU}_v(t)$ the mIoU observed with a model trained on t and evaluated on the validation set v.

A good dataset distance metric will produce similar mIoU rankings – and importantly, without training a model. We therefore compare all pairs of datasets using `fd`, `dna-fd`, and `dna-emd`, and rank them by distance. We denote by $T_v^{\text{pred}}[i]$ the training set ranked at the i-th position when compared to validation set v. To aggregate results, we

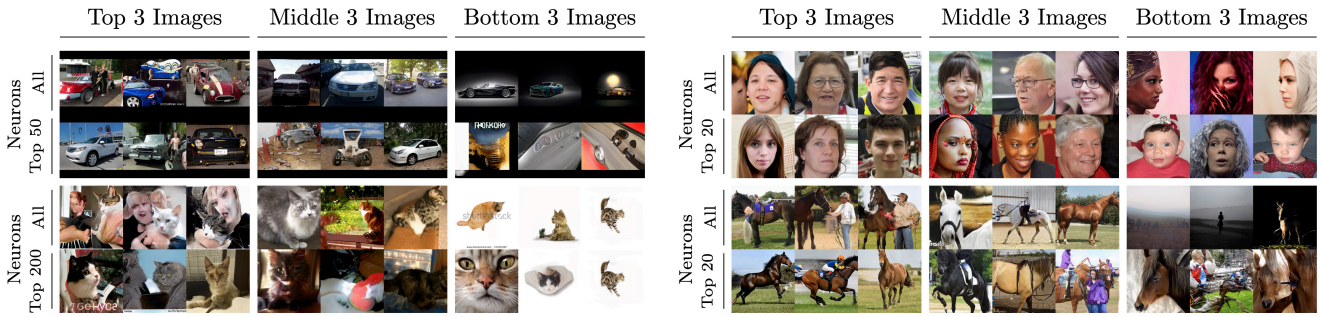


Figure 7. Generated StyleGANv2 [27] images of cars, faces, cats, and horses, ranked by `dna-emd` similarity to the corresponding real dataset's $\text{DNA}^{\text{hist.}}$. We selected neurons for realism by comparing real and synthetic images only featuring other classes (for general realism rather than focusing on class-specific differences) and compared this to using all neurons. Generally, selecting neurons results in rankings that better align with perceptual quality.

measure the discrepancy d between predicted and reference rankings using average mIoU differences:

$$d = \frac{1}{|V||T|} \sum_{v \in V} \sum_{i=1}^{|T_v^{\text{gt}}|} \left| \text{mIoU}_v(T_v^{\text{gt}}[i]) - \text{mIoU}_v(T_v^{\text{pred}}[i]) \right| \quad (8)$$

This discrepancy penalises out-of-rank predictions based on the difference of mIoU at those ranks.

In addition to the Mugs feature extractor, we also consider domain-specific feature extractors. We evaluate cross-dataset generalisation using features extracted from an HRNet-W48 semantic segmentation model trained on MSeg [31] which combines all datasets used in the experiment. We also use HRNet-W48 models trained on the validation domains. We report results relying on the features from the last layer of each model. We present the summary results for different feature extractors and metrics in Tab. 3.

Using `dna-emd` with a self-supervised network provides the best cross-dataset generalization. While being specifically adapted to the task and datasets considered, HRNet-W48 models fail to perform as well, likely due to the less general features not allowing to measure domain shifts as well. The average mIoU error in ranking datasets with `dna-emd` with Mugs features is only 0.76, indicating very good predictions of cross-generalization performance without training a model, markedly superior to `fd` and `dna-fd`.

Feature extractor	Fréchet Distance	DNA-Fréchet Distance	DNA-EMD
Mugs (ViT-B/16)	1.66	1.79	**0.76**
HRNet-W48 (all domains)	9.63	11.18	9.40
HRNet-W48 (val. domain)	13.9	14.5	6.85
Random ordering	14.93 ± 1.86 (50 samples)		

Table 3. Effectiveness of using dataset comparisons to predict semantic segmentation transfer learning performance. We compare the ranking of training datasets by a model's transfer learning performance to the ranking of datasets based on their distance to the validation set. We measure the severity of errors in predicted ranking by calculating the average difference in mIoU scores of reference models when ranked by mIoU and when ranked by the distance between their training and validation sets.

6. Limitations

Labelled data requirements for neuron selection Our neuron selection experiments in this work rely on labelled images to find neuron combination strategies. This is not always available, in which case unsupervised clustering techniques such as deepPIC [23] could be used.

Combining neurons Many neurons are likely to be polysemantic [12, 39], meaning that they are likely to react to multiple unrelated inputs. The approaches used in this paper to combine information from different neurons might be too limited to properly isolate specific attributes.

Discarded information in the DNA representation To make DNAs practical and scalable, we have discarded information about features. This includes spatial information about where activations occur and dependencies between activations of all neurons. These could help to obtain an even better representation.

7. Conclusion

We have presented a general and granular representation for images. This representation is based on keeping track of distributions of neuron activations of a pre-trained feature extractor. One DNA can be created from a single image or a complete dataset. Image DNAs are compact and granular representations which require no training, hyperparameter tuning, or labelling, regardless of the type of images considered. Our experiments have demonstrated that even with simplistic comparison strategies, DNAs can provide valuable insights into attribute-based comparisons, synthetic image quality assessment, and dataset differences.

Acknowledgements This work was supported by EPSRC Programme Grant "From Sensing to Collaboration" (EP/V000748/1), the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems [EP/S024050/1], and Oxbotica. The authors would like to acknowledge the use of Hartree Centre resources and the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work. We thank Adam Caccavale, Pierre-Yves Lajoie, Pierre Osselin, and David Williams for helpful discussions and inputs that contributed to this work.

References

[1] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 3, 5

[2] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[3] Ali Borji. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019-02-01. 2

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628. IEEE, 2020. 1

[5] Nick Cammarata, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. Curve Detectors. *Distill*, 5(6):e00024.003, 2020. 3

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pages 9912–9924. Curran Associates Inc., 2020. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 7, 15

[8] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation Atlas. *Distill*, 4(3):e15, 2019. 3

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223. IEEE, 2016. 1, 5, 7, 15

[10] Josip Djolonga, Mario Lucic, Marco Cuturi, Olivier Bachem, Olivier Bousquet, and Sylvain Gelly. Precision-Recall Curves Using Information Divergence Frontiers. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 2550–2559. PMLR, 2020. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, page 21, 2021. 3

[12] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. 8

[13] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022. 2

[14] Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C Mozer. Head2Toe: Utilizing intermediate representations for better transfer learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6009–6033. PMLR, 17–23 Jul 2022. 3

[15] Dan Feldman. Introduction to core-sets: an updated survey. *arXiv preprint arXiv:2011.09384*, 2020. 2

[16] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[17] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi Autonomous Driving Dataset. 2020. 1

[18] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and Benchmarking Self-Supervised Visual Representation Learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6390–6399. IEEE, 2019. 2

[19] Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity score: A new metric to evaluate the uncommonness of synthesized images. *arXiv preprint arXiv:2206.08549*, 2022. 3, 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016. 3

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2, 3, 4

[22] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018. 1

[23] Nikita Jaipuria, Katherine Stevo, Xianling Zhang, Meghana L. Gaopande, Ian Calle Garcia, Jinesh Jain, and Vidya N. Murali. deepPIC: Deep Perceptual Image Clustering For Identifying Bias In Vision Datasets. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4792–4801. IEEE. 8

[24] Longlong Jing and Yingli Tian. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 2019. 3

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing. 5

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. 6, 7

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116. IEEE, 2020. 7, 8, 20, 21, 22

[28] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[29] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. In *Proc. ICLR*, 2023. 3

[30] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved Precision and Recall Metric for Assessing Generative Models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[31] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A Composite Dataset for Multi-domain Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 10, 2020. 7, 8, 23, 24, 25, 26, 27

[32] E. Levina and P. Bickel. The Earth Mover's distance is the Mallows distance: Some insights from statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 251–256. IEEE Comput. Soc, 2001. 4

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755. Springer International Publishing, 2014. 5, 7

[34] Daohan Lu, Sheng-Yu Wang, Nupur Kumari, Rohan Agarwal, David Bau, and Jun-Yan Zhu. Content-based search for deep generative models. *arXiv preprint arXiv:2210.03116*, 2022. 2

[35] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 1

[36] Stanislav Morozov, Andrey Voynov, and Artem Babenko. On self-supervised image representations for gan evaluation. In *International Conference on Learning Representations*, 2020. 3

[37] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable Fidelity and Diversity Metrics for Generative Models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020. 3

[38] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5000–5009. IEEE, 2017. 7

[39] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 5(3):e00024.001, 2020. 3, 8

[40] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature Visualization. *Distill*, 2(11):e7, 2017. 3

[41] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022. 4

[42] Rohan Paul, Dan Feldman, Daniela Rus, and Paul Newman. Visual precis generation using coresets. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1304–1311, 2014. 2

[43] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset Issues in Object Recognition. In *Toward Category-Level Object Recognition*, volume 4170 of *Lecture Notes in Computer Science*, pages 29–48. Springer Berlin Heidelberg, 2006. 2

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 7, 15

[45] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's Hidden in a Randomly Weighted Neural Network? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11890–11899. IEEE. 6, 7, 15

[46] Stephan R. Richter, Hassan Abu Al Haija, and Vladlen Koltun. Enhancing Photorealism Enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 7, 15

[48] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. 4

[49] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing Generative Models via Precision and Recall. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 3

[50] Ozan Sener and Silvio Savarese. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*, 2018. 2

[51] Loic Simon, Ryan Webster, and Julien Rabin. Revisiting precision recall definition for generative modeling. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5799–5808. PMLR, 2019. 3

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 3, 5

[53] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576. IEEE, 2015. 7

[54] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2443–2451. IEEE, 2020. 1

[55] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. IEEE, 2016. 2, 7, 15

[56] A Torralba and AA Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011. 2

[57] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and C.V. Jawahar. IDD: A Dataset for Exploring Problems of Autonomous Navigation in Unconstrained Environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1743–1751. IEEE, 2019. 7

[58] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019. 7

[59] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 2

[60] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807. IEEE, 2018. 3

[61] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[62] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1, 5, 7

[63] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 7

[64] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernández Domínguez. WildDash - Creating Hazard-Aware Benchmarks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11210 of *Lecture Notes in Computer Science*, pages 407–421. Springer International Publishing, 2018. 1, 5

[65] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595. IEEE, 2018. 3, 5

[66] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130. IEEE, 2017. 5, 7

[67] Pan Zhou, Yichen Zhou, Chenyang Si, Weihao Yu, Teck Khim Ng, and Shuicheng Yan. Mugs: A multi-granular self-supervised learning framework. *arXiv preprint arXiv:2203.14415*, 2022. 3, 5, 7, 15