A. Technical details

A.1. Feature extractors

A.1.1 Features considered

We present details of the locations of neurons considered in Tab. 4. For the Stable diffusion v1.4 encoder, we expect the latent space to contain sufficient information to reconstruct the original image. As such, we only consider the latent space, treating each element as a neuron.

Feature extractor architecture	Layers considered	Number of neurons per layer
Inception-v3	Outputs of the first and second max pooling layers, input features to the auxiliary classifier, and output of the final average pooling layer	64 / 192 / 768 / 2048
Stable diffusion v1.4 encoder	Latent space produced by the VAE encoder	4096
ResNet-50	Output of each block	$3 \times [256]$ / $4 \times [512]$ / $6 \times [1024]$ / $3 \times [2048]$
Vision Transformer (ViT-B/16)	Output of each of the 12 self-attention layers, and output of the final normalisation layer	$13 \times [768]$
Vision Transformer (ViT-L/16)	Output of each of the 24 self-attention layers, and output of the final normalisation layer	$25 \times [1024]$
HRNet-W48	Outputs of each stage, where we treat each branch of the stage as providing a different layer. We also include the output of the classifier	256 / (48/96) / (48/96/192) / (48/96/192/384) / 194

Table 4. Details of layers and neurons considered for all architectures.

A.1.2 Spatial accumulation for Vision Transformers

In Sec. 3.2, we describe feature maps as having spatial dimensions $S_h^l \times S_w^l$. In the case of vision transformers, we treat the patch/token dimensions as the only spatial dimension. Activations over the patch/token dimension are accumulated in the histogram of the corresponding neuron.

A.2. Data processing

Activation normalisation As discussed in Sec. 4, activations at each neuron are normalised to ensure comparable scales over different neurons. The normalisation relies on tracking minimum and maximum activation values over the following datasets: StyleGANv2 generated images of cars, cats, churches, faces, and horses, LSUN cars, LSUN cats, LSUN churches, LSUN horses, FFHQ, Metfaces, CelebA, Cityscapes, KITTI, IDD, ADE20K, BDD100K, Mapillary, Wilddash, COCO, and SUN-RGBD. Using the minimum and maximum values observed, $\mathtt{a}_{\text{min}}$ and $\mathtt{a}_{\text{max}}$, for each neuron over all of these datasets, we recompute normalised activations $\mathtt{a}_{\text{norm}}$ from the activations observed during inference $\mathtt{a}$ to produce histograms as follows:

$$\mu = \frac{1}{2}(\mathtt{a}_{\text{max}} + \mathtt{a}_{\text{min}}) \tag{9}$$

$$\sigma = \frac{1}{2}(\mathtt{a}_{\text{max}} - \mathtt{a}_{\text{min}}) \tag{10}$$

$$\mathtt{a}_{\text{norm}} = \frac{\mathtt{a} - \mu}{\sigma} \tag{11}$$

Cropping Images are centre-cropped to squares and resized depending on the feature extractor, as described in Tab. 5.

Feature extractor	Image size used
Inception-v3	299×299
CLIP image encoder (ViT-B/16)	224×224
Stable diffusion v1.4 encoder	256×256
Random weights (ResNet-50)	224×224
DINO (ResNet-50)	224×224
DINO (ViT-B/16)	224×224
Mugs (ViT-B/16)	224×224
Mugs (ViT-L/16)	224×224
HRNet-W48	360×360

Table 5. Image sizes used for different feature extractors.

A.3. Random image transformations for Cityscapes retrieval task

In Sec. 5.1, we attempt to retrieve randomly augmented Cityscapes images. The augmentations applied to each image are 2 randomly selected operations applied sequentially. The operations considered are `Identity`, `Shear X`, `Shear Y`, `Translate X`, `Translate Y`, `Rotate`, `Adjust brightness`, `Adjust saturation`, `Adjust contrast`, `Adjust sharpness`, `Posterize`, `Auto contrast`, `Equalize`, `Salt-and-pepper noise`, `Gaussian noise`, and `Blur`.

A.4. Weights optimisation for attribute removal

For the optimisation of weights used in Sec. 5.3, we use the Adam optimiser with a learning rate of $1\mathrm{e}{-5}$, and first- and second-moment decays rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Weights are initialised to a constant value of $\frac{1}{\text{number of neurons}}$.

We perform optimisation of the loss in Eq. (7) on DNAs from the training set, which were generated using up to 50000 images for each DNA. We use early stopping by monitoring the loss on the validation set, stopping after 50 iterations without improvements. The results in Tab. 2 are then reported on the testing set.

A.5. Neuron selection for StyleGANv2 synthetic images

In Sec. 5.4, we select neurons to rank images from one class by selecting the most sensitive neurons to differences between `real` and `fake` images of all other classes. More precisely, this is done by creating a $\text{DNA}^{\text{hist.}}$ representing all `real` images from the other classes, and a DNA representing all `fake` images from the other classes. Combining the $\text{DNA}^{\text{hist.}}$ of multiple datasets is done by summing the counts for each histogram from the $\text{DNAs}^{\text{hist.}}$ to combine. This is equivalent to creating a new single $\text{DNA}^{\text{hist.}}$ using all the images from the different datasets. The sensitivity of neurons is then computed using the neuron-wise EMD between the combined `real` $\text{DNA}^{\text{hist.}}$ and the combined `fake` $\text{DNA}^{\text{hist.}}$.

A.6. Example histograms

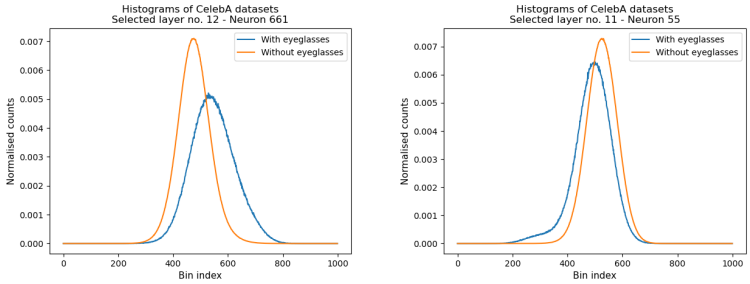
In Figs. 8 and 9, we present examples of histograms from specific neurons of the $\text{DNA}^{\text{hist.}}$ of CelebA images with or without some attributes. We observe that their shapes would not always be well approximated by a Gaussian distribution, as is done with $\text{DNAs}^{\text{Gauss.}}$.

B. Additional results details

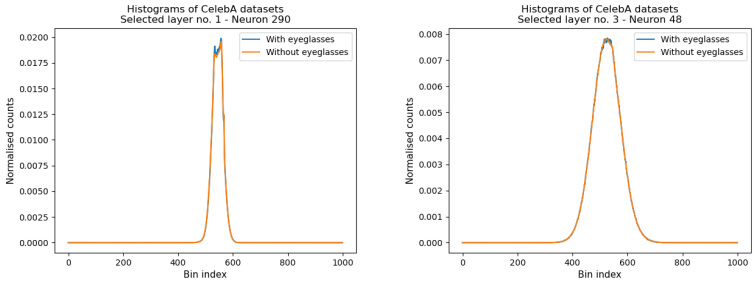
The following sections present more detailed results. Unless mentioned otherwise, the results use `dna-emd` with a Mugs (ViT-B/16) feature extractor.

B.1. Comparing images to a reference dataset with different neurons

In Fig. 10, we present ranked images from different datasets with specific neurons from their $\text{DNAs}^{\text{hist.}}$ compared to the $\text{DNA}^{\text{hist.}}$ of the Cityscapes dataset. Compared to Fig. 4 in which all neurons are considered, here we show results with neurons from the first and last selected layers of the feature extractor. When using neurons from the first layer, colours and textures appear much more important in producing the score. The best matches sometimes do not correspond to similar types of scenes, such as in COCO, but display similar colour profiles. Worst matches tend to contain high-frequency patterns

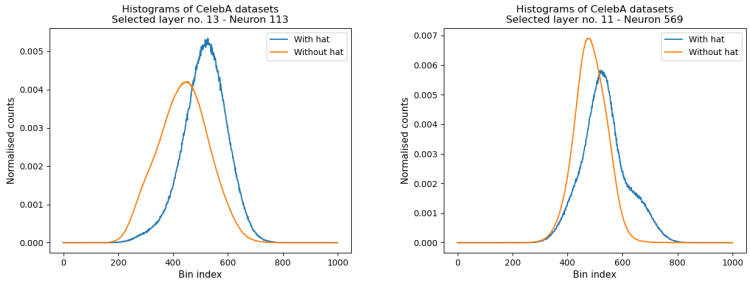


(a) Examples of histograms resulting in the largest Earth Mover's Distances

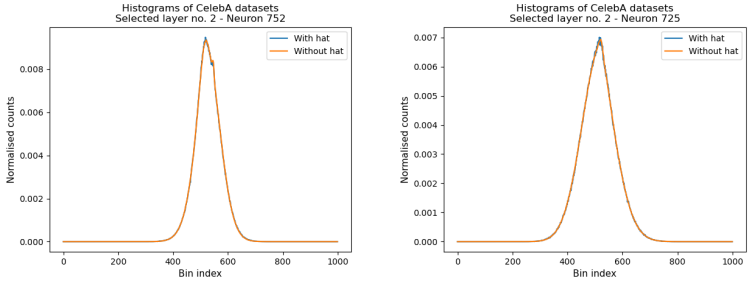


(b) Examples of histograms resulting in the lowest Earth Mover's Distances

Figure 8. Visualisation of normalised histograms for specific neurons from DNAs$^{\text{hist.}}$ (Mugs ViT-B/16) of images from the CelebA dataset with and without **`eyeglasses`**.



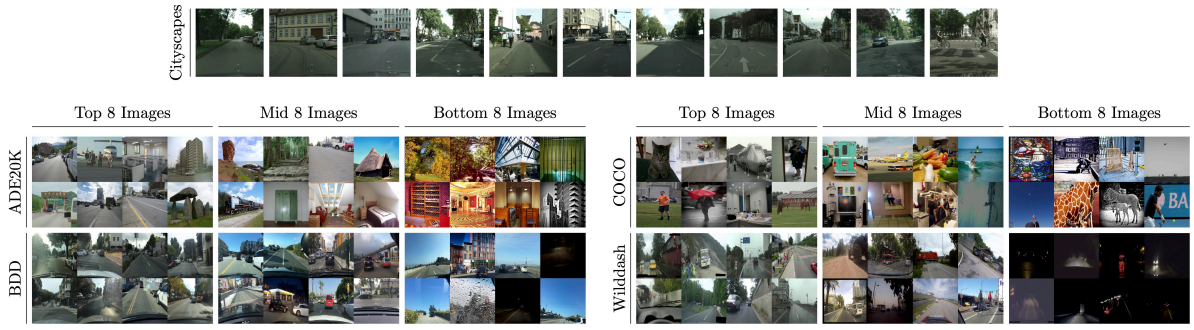
(a) Examples of histograms resulting in the largest Earth Mover's Distances



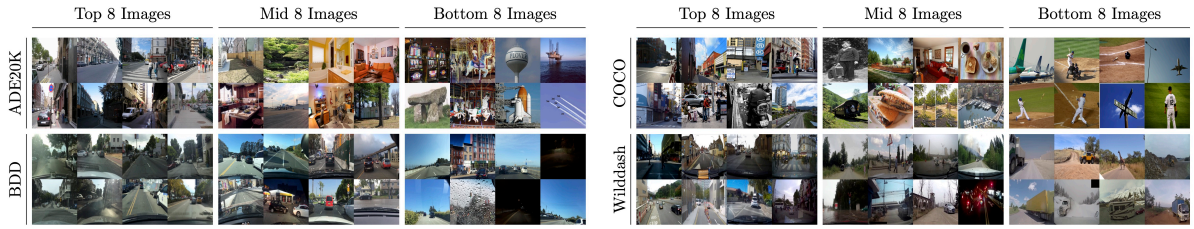
(b) Examples of histograms resulting in the lowest Earth Mover's Distances

Figure 9. Visualisation of normalised histograms for specific neurons from DNAs$^{\text{hist.}}$ (Mugs ViT-B/16) of images from the CelebA dataset with and without **`wearing hat`**.

or few features. On the other hand, when using neurons from the last layer, semantic content seems to be the main factor. Top-ranking images always show the same type of environment but do not always have the same colour profiles.



(a) Using all neurons from the **first layer** of the Mugs (ViT-B/16) feature extractor.



(b) Using all neurons from the **last layer** of the Mugs (ViT-B/16) feature extractor.

Figure 10. Images from different datasets organised by `dna-emd` when compared to Cityscapes [9] using different neurons from the feature extractor.

B.2. CelebA attribute sensitivity removal

B.2.1 Visualisation of `dna-emd` neuron-wise differences and weights learned

In Fig. 11, we compare the neuron-wise EMD between $\text{DNAs}^{\text{hist.}}$ of datasets with and without specific attributes to the weights learned in Sec. 5.3. Distances appear larger at later layers, but the attention from the weights allows us to focus on specific neurons spread over all layers and thus ignore the attribute, highlighting the need for granularity and a multi-layered approach.

B.2.2 Standard deviations of scores

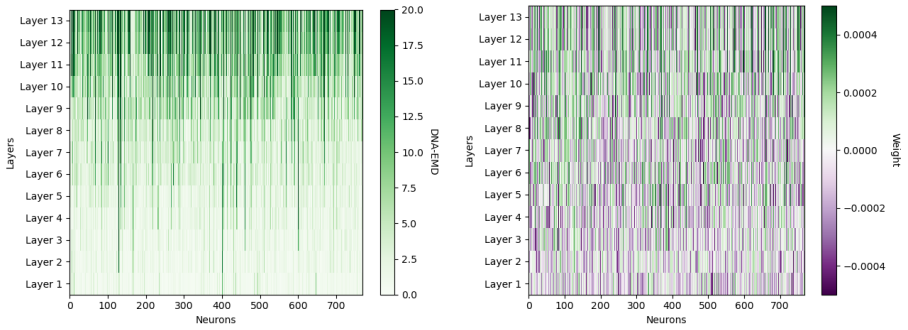
Tab. 6 details the standard deviations computed over all forty attributes for the results in Tab. 2 in Sec. 5.3.

Feature extractor	Target attribute sensitivity removal Δ_{rem} std. dev. (%)			Other attributes sensitivity deviation $\lvert\Delta_{\text{rem}}\rvert$ std. dev. (%)		
	Fréchet Distance	DNA-Fréchet Distance	DNA-EMD	Fréchet Distance	DNA-Fréchet Distance	DNA-EMD
Inception-v3 [55]	7.55	4.30	5.65	7.40	6.99	6.37
CLIP image encoder (ViT-B/16) [44]	13.84	5.78	4.90	12.80	4.24	3.59
Stable Diffusion v1.4 encoder [47]	-	9.62	11.05	-	6.83	6.32
Random weights (ResNet-50) [45]	12.13	28.05	19.58	11.91	17.12	9.97
DINO (ResNet-50) [7]	13.03	13.28	5.19	6.91	9.72	4.34
DINO (ViT-B/16) [7]	12.26	4.50	4.19	12.47	5.62	4.83
Mugs (ViT-B/16) [67]	12.18	5.47	4.76	11.62	6.33	5.39
Mugs (ViT-L/16) [67]	17.89	5.18	4.28	16.91	5.71	4.62

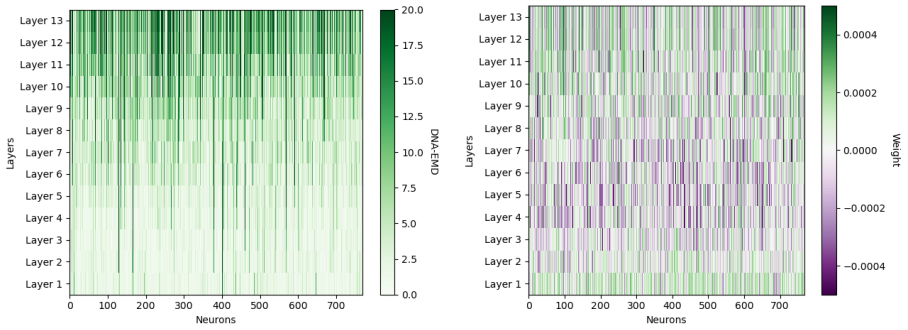
Table 6. Standard deviations over all forty attributes for scores presented in Tab. 2.

B.2.3 Detailed deviations

In Sec. 5.3, we weighted distances over different neurons to remove sensitivity to one attribute while preserving others. In Fig. 12, we visualise which attributes deviate most when ignoring another. We see that some attributes are particularly challenging to disentangle, often when we expect them to be correlated. For example, when ignoring the `no beard` attribute,



(a) Distances (**left**) and optimised weights (**right**) for the *eyeglasses* attribute.



(b) Distances (**left**) and optimised weights (**right**) for the *wearing hat* attribute.

Figure 11. Comparison of `dna-emd` distances between $\text{DNAs}^{\text{hist.}}$ with and without an attribute, and optimised weights for attribute removal.

we cause a large deviation in the `goatee` attribute. We might expect these to react to similar neurons. Still, we believe improvements in the optimised loss might help reduce this entanglement.

We also visualise overlaps between attributes in the CelebA dataset in Fig. 13.

B.3. FFHQ image pair comparisons

In this section, we present additional details for the results shown in Fig. 6. In addition to showing middle and bottom-ranked matches, we also consider different neurons for comparisons. Fig. 14 presents the matches ranked using all neurons from different layers of the feature extractor. Top matches from the first layer do not always focus on having similar semantic attributes. However, they all contain similar backgrounds and colours. Top matches from the last layer have much more diverse colour profiles and better match other images of the person in the reference image.

In Figs. 15 and 16, we present ranked matches using different numbers of selected neurons to focus on specific attributes. For the `wearing hat` attribute, we see that using too few or too many neurons can lead to not focusing on the desired attribute anymore. For the `eyeglasses` attribute, we are able to focus on the correct matches with all numbers of neurons. Even when using all neurons and no selection strategy, images with the attribute seem sufficiently favoured to be better ranked.

B.4. StyleGANv2 ranked images

We present more results of synthetic StyleGANv2 image rankings for different numbers of selected neurons for realism in Figs. 17 to 21.

B.5. Cross-dataset generalisation

Finally, in Tabs. 7 to 11 we present detailed cross-dataset generalisation results that are used to produce Tab. 3. Details are provided for the HRNet-W48 and Mugs models using `dna-emd`, and for the Mugs model with `dna-fd` and `fd`.

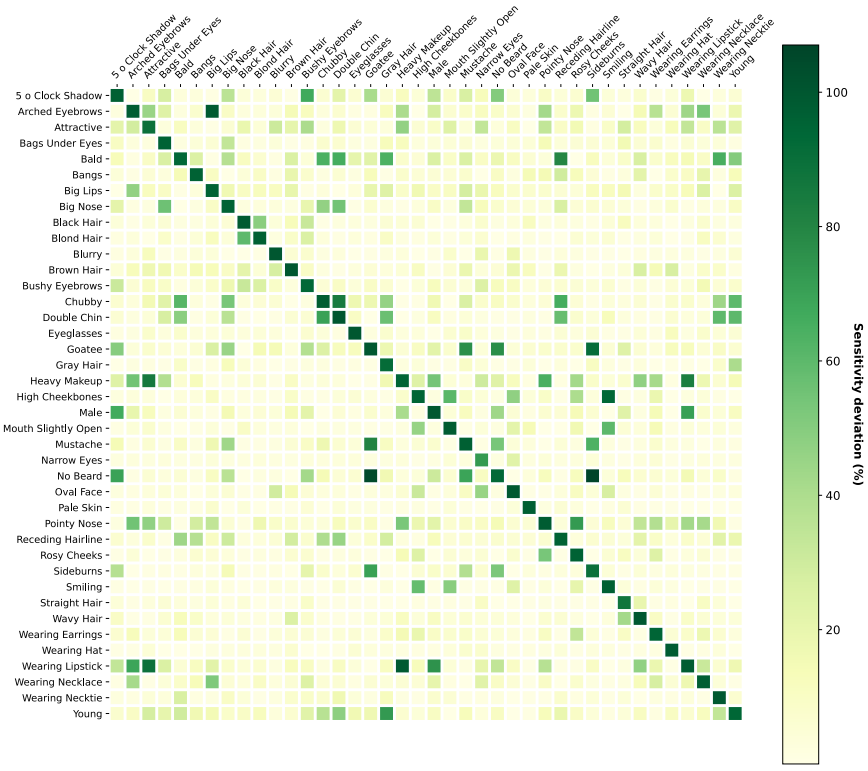


Figure 12. For each ignored attribute in the rows, we show the sensitivity deviation of all other attributes from the columns. The diagonals describe the relative drop in distances on the ignored attribute.

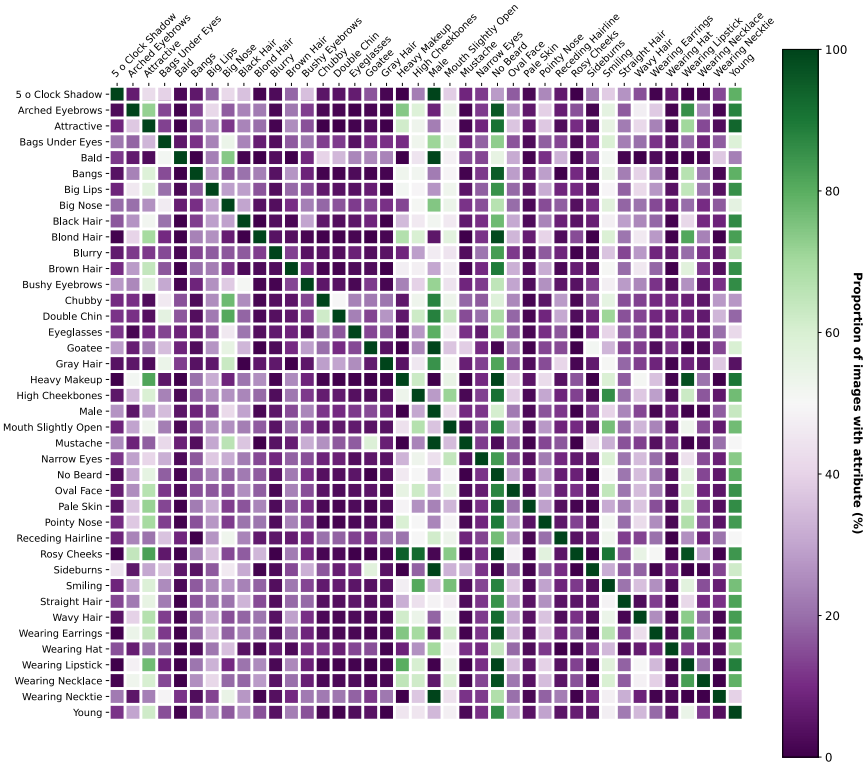


Figure 13. Illustration of correlations between attributes in the CelebA dataset. For each attribute in the rows, we show the percentage of images containing that attribute that also possess the attribute in the columns. Attributes can be negatively correlated when we observe values close to 0%, or positively correlated with values close to 100%.

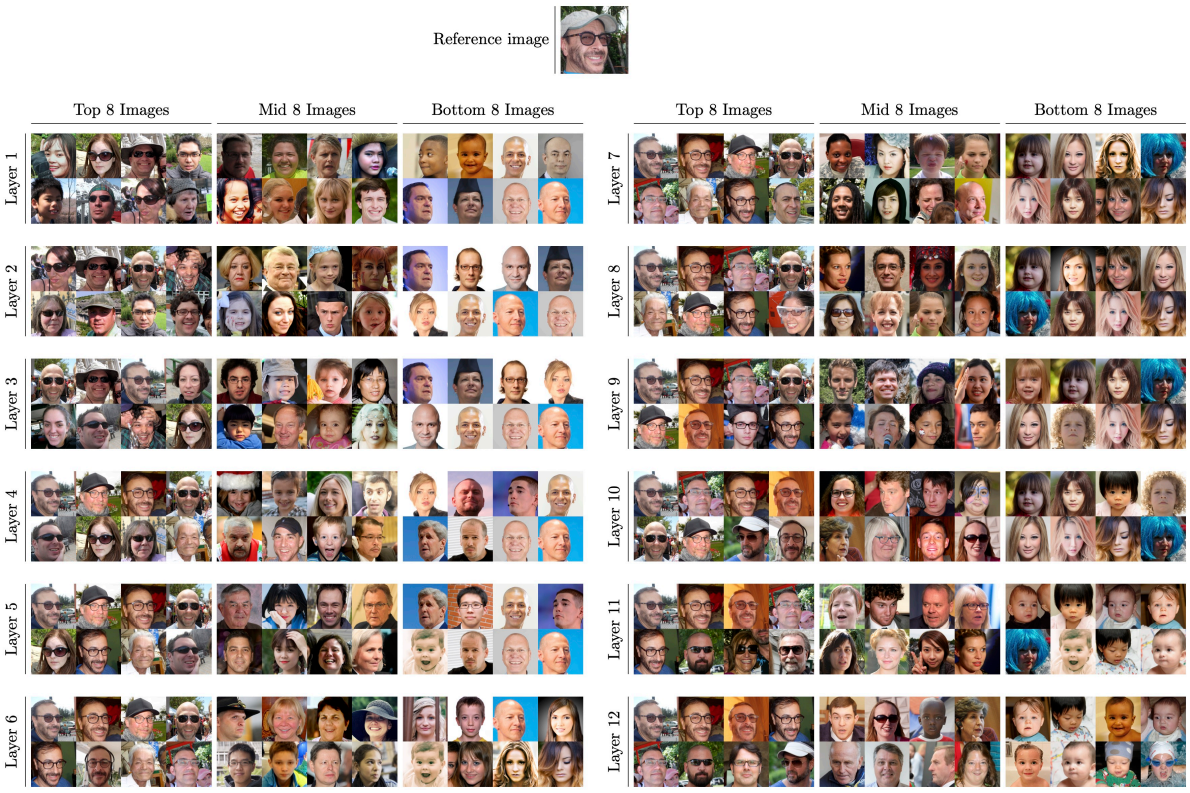


Figure 14. Ranked matches to the reference image using `dna-emd` with neurons from **different layers** of the Mugs (ViT-B/16) feature extractor.

Figure 15. Ranked matches to the reference image using `dna-emd` with neurons of the Mugs (ViT-B/16) feature extractor sensitive to the **`eyeglasses`** attribute.

Figure 16. Ranked matches to the reference image using `dna-emd` with neurons of the Mugs (ViT-B/16) feature extractor sensitive to the **`wearing hat`** attribute.

Figure 17. Generated StyleGANv2 [27] car images ranked by `dna-emd` when compared to the LSUN car images. We show the rankings for different numbers of selected neurons. The neuron selection strategy selects the most sensitive neurons when comparing real and synthetic images from other datasets.

Figure 18. Generated StyleGANv2 [27] cat images ranked by `dna-emd` when compared to the LSUN cat images. We show the rankings for different numbers of selected neurons. The neuron selection strategy selects the most sensitive neurons when comparing real and synthetic images from other datasets.

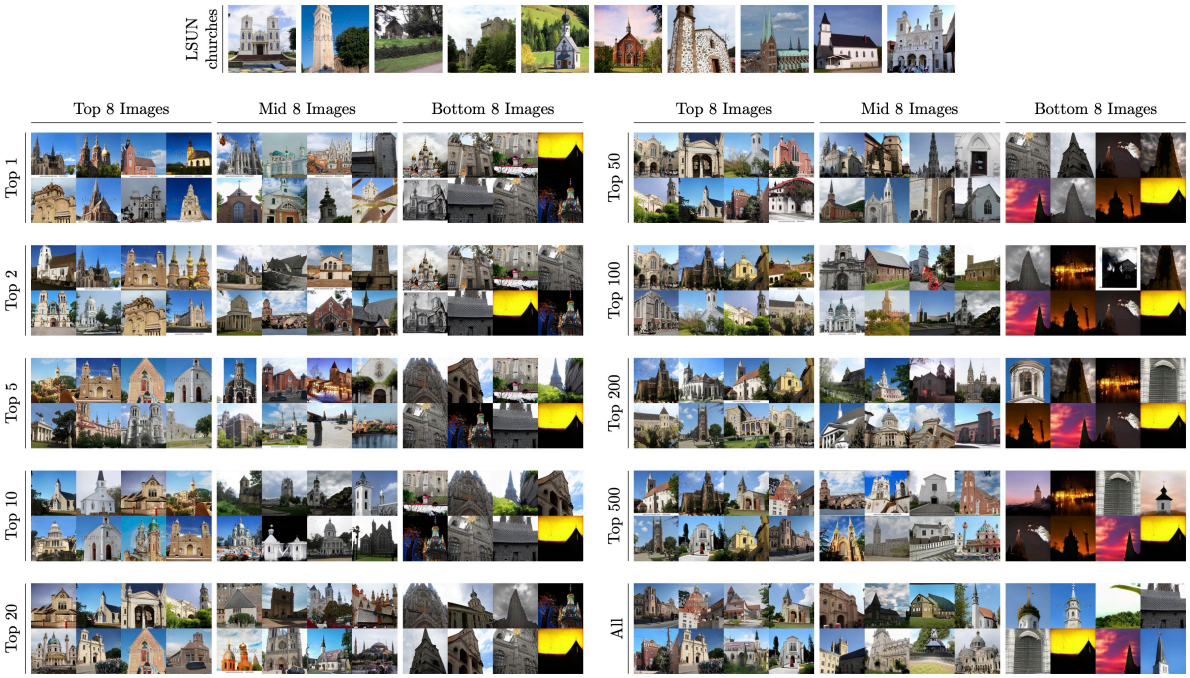


Figure 19. Generated StyleGANv2 [27] church images ranked by `dna-emd` when compared to the LSUN church images. We show the rankings for different numbers of selected neurons. The neuron selection strategy selects the most sensitive neurons when comparing real and synthetic images from other datasets.

Figure 20. Generated StyleGANv2 [27] face images ranked by `dna-emd` when compared to the FFHQ images. We show the rankings for different numbers of selected neurons. The neuron selection strategy selects the most sensitive neurons when comparing real and synthetic images from other datasets.

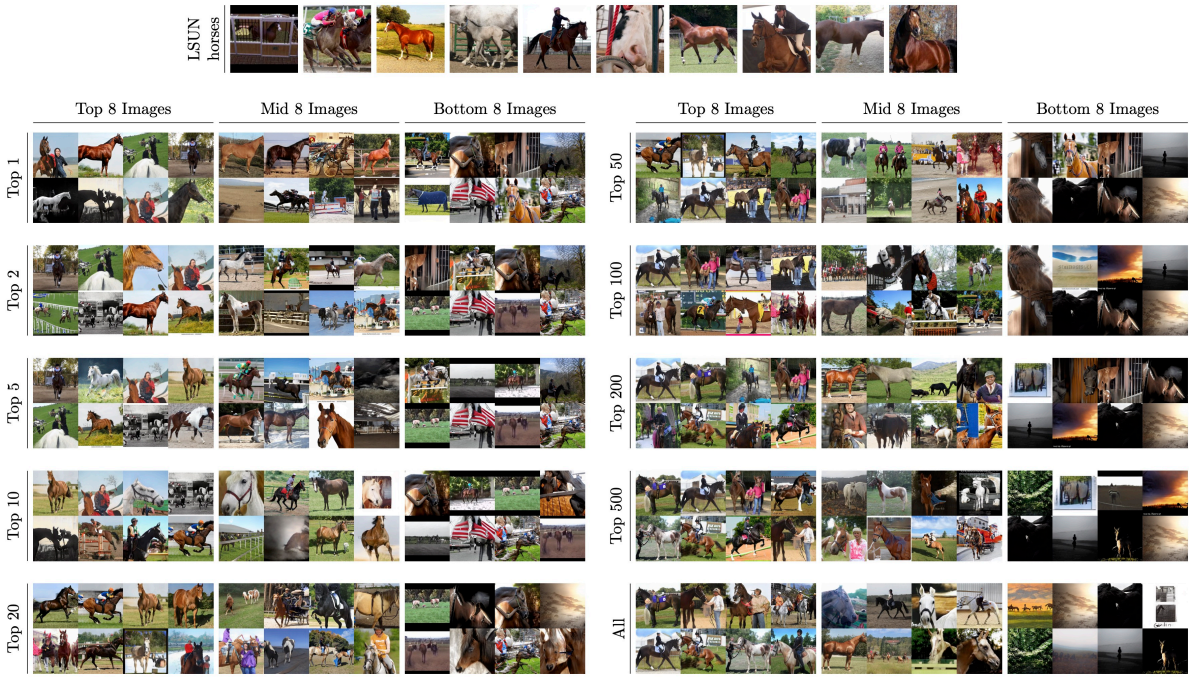


Figure 21. Generated StyleGANv2 [27] horse images ranked by `dna-emd` when compared to the LSUN horse images. We show the rankings for different numbers of selected neurons. The neuron selection strategy selects the most sensitive neurons when comparing real and synthetic images from other datasets.

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	45.3 - ADE20K	63.2 - BDD100K	77.6 - Cityscapes	52.6 - COCO	64.8 - IDD	56.2 - Mapillary	43.9 - SUN-RGBD
	19.6 - COCO	60.2 - Mapillary	69.7 - Mapillary	14.5 - ADE20K	48.2 - Mapillary	26.7 - COCO	35.3 - ADE20K
	7.1 - SUN-RGBD	45.0 - Cityscapes	60.9 - BDD100K	6.7 - Mapillary	33.9 - BDD100K	24.3 - ADE20K	29.4 - COCO
	6.2 - Mapillary	44.1 - COCO	50.2 - IDD	3.7 - BDD100K	31.3 - Cityscapes	24.3 - IDD	0.6 - IDD
	4.1 - BDD100K	43.7 - IDD	46.2 - COCO	3.3 - SUN-RGBD	31.0 - COCO	24.0 - BDD100K	0.2 - BDD100K
	3.1 - Cityscapes	41.5 - ADE20K	44.3 - ADE20K	3.1 - Cityscapes	27.0 - ADE20K	22.4 - Cityscapes	0.2 - Cityscapes
	3.1 - IDD	2.2 - SUN-RGBD	2.6 - SUN-RGBD	3.1 - IDD	1.0 - SUN-RGBD	1.1 - SUN-RGBD	0.2 - Mapillary

(a) Observed cross-dataset generalisation on semantic segmentation from Lambert *et al*. [31] (mIoU). Each column corresponds to the evaluation of one dataset (validation set). Rows are ordered by a cross-generalisation performance from training on each dataset (training set).

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	0.84 - ADE20K	3.61 - BDD100K	3.18 - Cityscapes	0.64 - COCO	4.48 - IDD	0.85 - Mapillary	2.22 - SUN-RGBD
	9.17 - COCO	8.91 - Mapillary	15.9 - Mapillary	9.2 - ADE20K	12.88 - Mapillary	9.72 - BDD100K	12.71 - ADE20K
	12.66 - SUN-RGBD	13.03 - IDD	17.13 - BDD100K	15.4 - SUN-RGBD	13.54 - BDD100K	11.52 - IDD	15.57 - COCO
	21.18 - Mapillary	17.18 - Cityscapes	17.15 - IDD	21.79 - Mapillary	18.95 - Cityscapes	16.19 - Cityscapes	26.92 - IDD
	21.57 - IDD	22.56 - ADE20K	25.61 - ADE20K	22.36 - IDD	22.53 - ADE20K	21.26 - ADE20K	27.01 - Mapillary
	22.58 - BDD100K	23.88 - COCO	25.73 - COCO	23.86 - BDD100K	23.33 - COCO	21.89 - COCO	27.96 - BDD100K
	26.17 - Cityscapes	27.96 - SUN-RGBD	30.47 - SUN-RGBD	26.41 - Cityscapes	27.51 - SUN-RGBD	26.95 - SUN-RGBD	31.08 - Cityscapes

(b) Ordered datasets ranked by `dna-emd` with corresponding EMD values. The EMD here is computed using the last layer of the Mugs (ViT-B/16) feature extractor.

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	45.3 - ADE20K	63.2 - BDD100K	77.6 - Cityscapes	52.6 - COCO	64.8 - IDD	56.2 - Mapillary	43.9 - SUN-RGBD
	19.6 - COCO	60.2 - Mapillary	69.7 - Mapillary	14.5 - ADE20K	48.2 - Mapillary	24.0 - BDD100K	35.3 - ADE20K
	7.1 - SUN-RGBD	43.7 - IDD	60.9 - BDD100K	3.3 - SUN-RGBD	33.9 - BDD100K	24.3 - IDD	29.4 - COCO
	6.2 - Mapillary	45.0 - Cityscapes	50.2 - IDD	6.7 - Mapillary	31.3 - Cityscapes	22.4 - Cityscapes	0.6 - IDD
	3.1 - IDD	41.5 - ADE20K	44.3 - ADE20K	3.1 - IDD	27.0 - ADE20K	24.3 - ADE20K	0.2 - Mapillary
	4.1 - BDD100K	44.1 - COCO	46.2 - COCO	3.7 - BDD100K	31.0 - COCO	26.7 - COCO	0.2 - BDD100K
	3.1 - Cityscapes	2.2 - SUN-RGBD	2.6 - SUN-RGBD	3.1 - Cityscapes	1.0 - SUN-RGBD	1.1 - SUN-RGBD	0.2 - Cityscapes

(c) Ordered datasets ranked by `dna-emd` with corresponding mIoU values. The ranking is taken from Tab. 7b, but we show the mIoUs for the corresponding datasets from Tab. 7a instead.

Validation datasets						
ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	2.7	0.0
0.0	1.3	0.0	3.4	0.0	0.0	0.0
0.0	0.9	0.0	3.0	0.0	1.9	0.0
1.0	2.2	1.9	0.2	4.0	0.3	0.0
1.0	2.6	1.9	0.6	4.0	4.3	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
Average absolute mIoU difference: 0.76						

(d) Differences between the mIoUs of the reference ranking (Tab. 7a) and the mIoUs for the predicted `dna-emd` ranking (Tab. 7c).

Table 7. Detailed results comparing the observed cross-dataset generalisation of the HRNet-W48 semantic segmentation models to predictions relying only on datasets using **`dna-emd`** with the **Mugs (ViT-B/16) feature extractor**. The final value reported in Tab. 3 corresponds to the average of the values in Tab. 7d.

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	45.3 - ADE20K	63.2 - BDD100K	77.6 - Cityscapes	52.6 - COCO	64.8 - IDD	56.2 - Mapillary	43.9 - SUN-RGBD
	19.6 - COCO	60.2 - Mapillary	69.7 - Mapillary	14.5 - ADE20K	48.2 - Mapillary	26.7 - COCO	35.3 - ADE20K
	7.1 - SUN-RGBD	45.0 - Cityscapes	60.9 - BDD100K	6.7 - Mapillary	33.9 - BDD100K	24.3 - ADE20K	29.4 - COCO
	6.2 - Mapillary	44.1 - COCO	50.2 - IDD	3.7 - BDD100K	31.3 - Cityscapes	24.3 - IDD	0.6 - IDD
	4.1 - BDD100K	43.7 - IDD	46.2 - COCO	3.3 - SUN-RGBD	31.0 - COCO	24.0 - BDD100K	0.2 - BDD100K
	3.1 - Cityscapes	41.5 - ADE20K	44.3 - ADE20K	3.1 - Cityscapes	27.0 - ADE20K	22.4 - Cityscapes	0.2 - Cityscapes
	3.1 - IDD	2.2 - SUN-RGBD	2.6 - SUN-RGBD	3.1 - IDD	1.0 - SUN-RGBD	1.1 - SUN-RGBD	0.2 - Mapillary

(a) Observed cross-dataset generalisation on semantic segmentation from Lambert *et al*. [31] (mIoU). Each column corresponds to the evaluation of one dataset (validation set). Rows are ordered by a cross-generalisation performance from training on each dataset (training set).

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	0.02 - ADE20K	0.08 - BDD100K	0.04 - Cityscapes	0.01 - COCO	0.07 - IDD	0.02 - Mapillary	0.03 - SUN-RGBD
	0.3 - SUN-RGBD	0.31 - IDD	0.35 - Mapillary	0.27 - Mapillary	0.32 - Mapillary	0.26 - COCO	0.3 - ADE20K
	0.32 - BDD100K	0.34 - SUN-RGBD	0.37 - COCO	0.31 - IDD	0.32 - COCO	0.33 - IDD	0.31 - BDD100K
	0.51 - IDD	0.34 - ADE20K	0.46 - IDD	0.37 - Cityscapes	0.35 - BDD100K	0.34 - Cityscapes	0.47 - IDD
	0.6 - COCO	0.51 - Mapillary	0.7 - BDD100K	0.54 - SUN-RGBD	0.45 - Cityscapes	0.56 - BDD100K	0.54 - COCO
	0.68 - Mapillary	0.52 - COCO	0.75 - SUN-RGBD	0.56 - BDD100K	0.47 - SUN-RGBD	0.63 - SUN-RGBD	0.63 - Mapillary
	0.81 - Cityscapes	0.66 - Cityscapes	0.82 - ADE20K	0.61 - ADE20K	0.52 - ADE20K	0.7 - ADE20K	0.75 - Cityscapes

(b) Ordered datasets ranked by `dna-emd` with corresponding EMD values. The EMD here is computed using the last layer of the HRNet-W48 feature extractor trained on all domains.

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	45.3 - ADE20K	63.2 - BDD100K	77.6 - Cityscapes	52.6 - COCO	64.8 - IDD	56.2 - Mapillary	43.9 - SUN-RGBD
	7.1 - SUN-RGBD	43.7 - IDD	69.7 - Mapillary	6.7 - Mapillary	48.2 - Mapillary	26.7 - COCO	35.3 - ADE20K
	4.1 - BDD100K	2.2 - SUN-RGBD	46.2 - COCO	3.1 - IDD	31.0 - COCO	24.3 - IDD	0.2 - BDD100K
	3.1 - IDD	41.5 - ADE20K	50.2 - IDD	3.1 - Cityscapes	33.9 - BDD100K	22.4 - Cityscapes	0.6 - IDD
	19.6 - COCO	60.2 - Mapillary	60.9 - BDD100K	3.3 - SUN-RGBD	31.3 - Cityscapes	24.0 - BDD100K	29.4 - COCO
	6.2 - Mapillary	44.1 - COCO	2.6 - SUN-RGBD	3.7 - BDD100K	1.0 - SUN-RGBD	1.1 - SUN-RGBD	0.2 - Mapillary
	3.1 - Cityscapes	45.0 - Cityscapes	44.3 - ADE20K	14.5 - ADE20K	27.0 - ADE20K	24.3 - ADE20K	0.2 - Cityscape

(c) Ordered datasets ranked by `dna-emd` with corresponding mIoU values. The ranking is taken from Tab. 8b, but we show the mIoUs for the corresponding datasets from Tab. 8a instead.

Validation datasets						
ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
0.0	0.0	0.0	0.0	0.0	0.0	0.0
12.5	16.5	0.0	7.8	0.0	0.0	0.0
3.0	42.8	14.7	3.6	2.9	0.0	29.2
3.1	2.6	0.0	0.6	2.6	1.9	0.0
15.5	16.5	14.7	0.0	0.3	0.0	29.2
3.1	2.6	41.7	0.6	26.0	21.3	0.0
0.0	42.8	41.7	11.4	26.0	23.2	0.0
Average absolute mIoU difference: 9.40						

(d) Differences between the mIoUs of the reference ranking (Tab. 8a) and the mIoUs for the predicted `dna-emd` ranking (Tab. 8c).

Table 8. Detailed results comparing the observed cross-dataset generalisation of the HRNet-W48 semantic segmentation models to predictions relying only on datasets using **`dna-emd`** with the **HRNet-W48 feature extractor trained on all domains**. The final value reported in Tab. 3 corresponds to the average of the values in Tab. 8d.

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	45.3 - ADE20K	63.2 - BDD100K	77.6 - Cityscapes	52.6 - COCO	64.8 - IDD	56.2 - Mapillary	43.9 - SUN-RGBD
	19.6 - COCO	60.2 - Mapillary	69.7 - Mapillary	14.5 - ADE20K	48.2 - Mapillary	26.7 - COCO	35.3 - ADE20K
	7.1 - SUN-RGBD	45.0 - Cityscapes	60.9 - BDD100K	6.7 - Mapillary	33.9 - BDD100K	24.3 - ADE20K	29.4 - COCO
	6.2 - Mapillary	44.1 - COCO	50.2 - IDD	3.7 - BDD100K	31.3 - Cityscapes	24.3 - IDD	0.6 - IDD
	4.1 - BDD100K	43.7 - IDD	46.2 - COCO	3.3 - SUN-RGBD	31.0 - COCO	24.0 - BDD100K	0.2 - BDD100K
	3.1 - Cityscapes	41.5 - ADE20K	44.3 - ADE20K	3.1 - Cityscapes	27.0 - ADE20K	22.4 - Cityscapes	0.2 - Cityscapes
	3.1 - IDD	2.2 - SUN-RGBD	2.6 - SUN-RGBD	3.1 - IDD	1.0 - SUN-RGBD	1.1 - SUN-RGBD	0.2 - Mapillary

(a) Observed cross-dataset generalisation on semantic segmentation from Lambert *et al*. [31] (mIoU). Each column corresponds to the evaluation of one dataset (validation set). Rows are ordered by a cross-generalisation performance from training on each dataset (training set).

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	0.01 - ADE20K	0.04 - BDD100K	0.02 - Cityscapes	0.0 - COCO	0.02 - IDD	0.0 - Mapillary	0.03 - SUN-RGBD
	0.1 - COCO	0.11 - IDD	0.15 - IDD	0.1 - ADE20K	0.08 - BDD100K	0.04 - BDD100K	0.17 - COCO
	0.14 - SUN-RGBD	0.11 - Mapillary	0.19 - BDD100K	0.12 - SUN-RGBD	0.1 - Cityscapes	0.05 - IDD	0.17 - IDD
	0.21 - BDD100K	0.15 - Cityscapes	0.24 - Mapillary	0.13 - BDD100K	0.1 - Mapillary	0.06 - ADE20K	0.19 - ADE20K
	0.23 - IDD	0.18 - ADE20K	0.31 - SUN-RGBD	0.16 - Mapillary	0.11 - ADE20K	0.07 - COCO	0.21 - Mapillary
	0.24 - Mapillary	0.24 - COCO	0.34 - ADE20K	0.18 - IDD	0.11 - COCO	0.09 - SUN-RGBD	0.22 - BDD100K
	0.37 - Cityscapes	0.28 - SUN-RGBD	0.37 - COCO	0.21 - Cityscapes	0.12 - SUN-RGBD	0.12 - Cityscapes	0.27 - Cityscapes

(b) Ordered datasets ranked by `dna-emd` with corresponding EMD values. The EMD here is computed using the last layer of the HRNet-W48 feature extractor trained on validation domains.

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	45.3 - ADE20K	63.2 - BDD100K	77.6 - Cityscapes	52.6 - COCO	64.8 - IDD	56.2 - Mapillary	43.9 - SUN-RGBD
	19.6 - COCO	43.7 - IDD	50.2 - IDD	14.5 - ADE20K	33.9 - BDD100K	24.0 - BDD100K	29.4 - COCO
	7.1 - SUN-RGBD	60.2 - Mapillary	60.9 - BDD100K	3.3 - SUN-RGBD	31.3 - Cityscapes	24.3 - IDD	0.6 - IDD
	4.1 - BDD100K	45.0 - Cityscapes	69.7 - Mapillary	3.7 - BDD100K	48.2 - Mapillary	24.3 - ADE20K	35.3 - ADE20K
	3.1 - IDD	41.5 - ADE20K	2.6 - SUN-RGBD	6.7 - Mapillary	27.0 - ADE20K	26.7 - COCO	0.2 - Mapillary
	6.2 - Mapillary	44.1 - COCO	44.3 - ADE20K	3.1 - IDD	31.0 - COCO	1.1 - SUN-RGBD	0.2 - BDD100K
	3.1 - Cityscapes	2.2 - SUN-RGBD	46.2 - COCO	3.1 - Cityscapes	1.0 - SUN-RGBD	22.4 - Cityscapes	0.2 - Cityscapes

(c) Ordered datasets ranked by `dna-emd` with corresponding mIoU values. The ranking is taken from Tab. 9b, but we show the mIoUs for the corresponding datasets from Tab. 9a instead.

Validation datasets						
ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	16.5	19.5	0.0	14.3	2.7	5.9
0.0	15.2	0.0	3.4	2.6	0.0	28.8
2.1	0.9	19.5	0.0	16.9	0.0	34.7
1.0	2.2	43.6	3.4	4.0	2.7	0.0
3.1	2.6	0.0	0.0	4.0	21.3	0.0
0.0	0.0	43.6	0.0	0.0	21.3	0.0
Average absolute mIoU difference: 6.85						

(d) Differences between the mIoUs of the reference ranking (Tab. 9a) and the mIoUs for the predicted `dna-emd` ranking (Tab. 9c).

Table 9. Detailed results comparing the observed cross-dataset generalisation of the HRNet-W48 semantic segmentation models to predictions relying only on datasets using **`dna-emd`** with the **HRNet-W48 feature extractor trained on validation domains**. The final value reported in Tab. 3 corresponds to the average of the values in Tab. 9d.

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	45.3 - ADE20K	63.2 - BDD100K	77.6 - Cityscapes	52.6 - COCO	64.8 - IDD	56.2 - Mapillary	43.9 - SUN-RGBD
	19.6 - COCO	60.2 - Mapillary	69.7 - Mapillary	14.5 - ADE20K	48.2 - Mapillary	26.7 - COCO	35.3 - ADE20K
	7.1 - SUN-RGBD	45.0 - Cityscapes	60.9 - BDD100K	6.7 - Mapillary	33.9 - BDD100K	24.3 - ADE20K	29.4 - COCO
	6.2 - Mapillary	44.1 - COCO	50.2 - IDD	3.7 - BDD100K	31.3 - Cityscapes	24.3 - IDD	0.6 - IDD
	4.1 - BDD100K	43.7 - IDD	46.2 - COCO	3.3 - SUN-RGBD	31.0 - COCO	24.0 - BDD100K	0.2 - BDD100K
	3.1 - Cityscapes	41.5 - ADE20K	44.3 - ADE20K	3.1 - Cityscapes	27.0 - ADE20K	22.4 - Cityscapes	0.2 - Cityscapes
	3.1 - IDD	2.2 - SUN-RGBD	2.6 - SUN-RGBD	3.1 - IDD	1.0 - SUN-RGBD	1.1 - SUN-RGBD	0.2 - Mapillary

(a) Observed cross-dataset generalisation on semantic segmentation from Lambert *et al*. [31] (mIoU). Each column corresponds to the evaluation of one dataset (validation set). Rows are ordered by a cross-generalisation performance from training on each dataset (training set).

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	0.0 - ADE20K	0.01 - BDD100K	0.01 - Cityscapes	0.0 - COCO	0.01 - IDD	0.0 - Mapillary	0.0 - SUN-RGBD
	0.11 - COCO	0.06 - Mapillary	0.13 - Mapillary	0.12 - ADE20K	0.09 - BDD100K	0.07 - BDD100K	0.16 - ADE20K
	0.15 - SUN-RGBD	0.08 - IDD	0.16 - IDD	0.23 - SUN-RGBD	0.09 - Mapillary	0.07 - IDD	0.24 - COCO
	0.32 - IDD	0.17 - Cityscapes	0.17 - BDD100K	0.31 - Mapillary	0.2 - Cityscapes	0.14 - Cityscapes	0.49 - IDD
	0.32 - BDD100K	0.32 - ADE20K	0.43 - COCO	0.34 - IDD	0.34 - ADE20K	0.31 - COCO	0.51 - Mapillary
	0.32 - Mapillary	0.39 - COCO	0.47 - ADE20K	0.39 - BDD100K	0.37 - COCO	0.33 - ADE20K	0.52 - BDD100K
	0.48 - Cityscapes	0.52 - SUN-RGBD	0.6 - SUN-RGBD	0.44 - Cityscapes	0.51 - SUN-RGBD	0.5 - SUN-RGBD	0.62 - Cityscapes

(b) Ordered datasets ranked by `dna-fd` with corresponding FD values. The FD here is computed using the last layer of the Mugs (ViT-B/16) feature extractor.

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	45.3 - ADE20K	63.2 - BDD100K	77.6 - Cityscapes	52.6 - COCO	64.8 - IDD	56.2 - Mapillary	43.9 - SUN-RGBD
	19.6 - COCO	60.2 - Mapillary	69.7 - Mapillary	14.5 - ADE20K	33.9 - BDD100K	24.0 - BDD100K	35.3 - ADE20K
	7.1 - SUN-RGBD	43.7 - IDD	50.2 - IDD	3.3 - SUN-RGBD	48.2 - Mapillary	24.3 - IDD	29.4 - COCO
	3.1 - IDD	45.0 - Cityscapes	60.9 - BDD100K	6.7 - Mapillary	31.3 - Cityscapes	22.4 - Cityscapes	0.6 - IDD
	4.1 - BDD100K	41.5 - ADE20K	46.2 - COCO	3.1 - IDD	27.0 - ADE20K	26.7 - COCO	0.2 - Mapillary
	6.2 - Mapillary	44.1 - COCO	44.3 - ADE20K	3.7 - BDD100K	31.0 - COCO	24.3 - ADE20K	0.2 - BDD100K
	3.1 - Cityscapes	2.2 - SUN-RGBD	2.6 - SUN-RGBD	3.1 - Cityscapes	1.0 - SUN-RGBD	1.1 - SUN-RGBD	0.2 - Cityscapes

(c) Ordered datasets ranked by `dna-fd` with corresponding mIoU values. The ranking is taken from Tab. 10b, but we show the mIoUs for the corresponding datasets from Tab. 10a instead.

Validation datasets						
ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	14.3	2.7	0.0
0.0	1.3	10.7	3.4	14.3	0.0	0.0
3.1	0.9	10.7	3.0	0.0	1.9	0.0
0.0	2.2	0.0	0.2	4.0	2.7	0.0
3.1	2.6	0.0	0.6	4.0	1.9	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
Average absolute mIoU difference: 1.79						

(d) Differences between the mIoUs of the reference ranking (Tab. 10a) and the mIoUs for the predicted `dna-fd` ranking (Tab. 10c).

Table 10. Detailed results comparing the observed cross-dataset generalisation of the HRNet-W48 semantic segmentation models to predictions relying only on datasets using **`dna-fd`** with the **Mugs (ViT-B/16) feature extractor**. The final value reported in Tab. 3 corresponds to the average of the values in Tab. 10d.

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	45.3 - ADE20K	63.2 - BDD100K	77.6 - Cityscapes	52.6 - COCO	64.8 - IDD	56.2 - Mapillary	43.9 - SUN-RGBD
	19.6 - COCO	60.2 - Mapillary	69.7 - Mapillary	14.5 - ADE20K	48.2 - Mapillary	26.7 - COCO	35.3 - ADE20K
	7.1 - SUN-RGBD	45.0 - Cityscapes	60.9 - BDD100K	6.7 - Mapillary	33.9 - BDD100K	24.3 - ADE20K	29.4 - COCO
	6.2 - Mapillary	44.1 - COCO	50.2 - IDD	3.7 - BDD100K	31.3 - Cityscapes	24.3 - IDD	0.6 - IDD
	4.1 - BDD100K	43.7 - IDD	46.2 - COCO	3.3 - SUN-RGBD	31.0 - COCO	24.0 - BDD100K	0.2 - BDD100K
	3.1 - Cityscapes	41.5 - ADE20K	44.3 - ADE20K	3.1 - Cityscapes	27.0 - ADE20K	22.4 - Cityscapes	0.2 - Cityscapes
	3.1 - IDD	2.2 - SUN-RGBD	2.6 - SUN-RGBD	3.1 - IDD	1.0 - SUN-RGBD	1.1 - SUN-RGBD	0.2 - Mapillary

(a) Observed cross-dataset generalisation on semantic segmentation from Lambert *et al*. [31] (mIoU). Each column corresponds to the evaluation of one dataset (validation set). Rows are ordered by a cross-generalisation performance from training on each dataset (training set).

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	19.37 - ADE20K	31.16 - BDD100K	40.74 - Cityscapes	9.68 - COCO	54.03 - IDD	9.78 - Mapillary	71.23 - SUN-RGBD
	233.41 - COCO	101.67 - Mapillary	196.74 - Mapillary	227.36 - ADE20K	160.51 - BDD100K	96.19 - BDD100K	332.54 - ADE20K
	279.51 - SUN-RGBD	141.84 - IDD	225.68 - IDD	389.04 - SUN-RGBD	161.9 - Mapillary	123.14 - IDD	460.08 - COCO
	398.31 - Mapillary	215.55 - Cityscapes	232.53 - BDD100K	417.86 - Mapillary	255.92 - Cityscapes	171.68 - Cityscapes	638.82 - IDD
	401.76 - BDD100K	423.12 - ADE20K	543.35 - COCO	457.03 - IDD	470.54 - ADE20K	401.26 - ADE20K	652.68 - BDD100K
	416.94 - IDD	508.35 - COCO	550.05 - ADE20K	476.86 - BDD100K	519.04 - COCO	428.07 - COCO	660.09 - Mapillary
	522.29 - Cityscapes	645.3 - SUN-RGBD	701.34 - SUN-RGBD	509.45 - Cityscapes	656.16 - SUN-RGBD	630.26 - SUN-RGBD	704.86 - Cityscapes

(b) Ordered datasets ranked by `fd` with corresponding FD values. The FD here is computed using the last layer of the Mugs (ViT-B/16) feature extractor.

Training datasets	Validation datasets						
	ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
	45.3 - ADE20K	63.2 - BDD100K	77.6 - Cityscapes	52.6 - COCO	64.8 - IDD	56.2 - Mapillary	43.9 - SUN-RGBD
	19.6 - COCO	60.2 - Mapillary	69.7 - Mapillary	14.5 - ADE20K	33.9 - BDD100K	24.0 - BDD100K	35.3 - ADE20K
	7.1 - SUN-RGBD	43.7 - IDD	50.2 - IDD	3.3 - SUN-RGBD	48.2 - Mapillary	24.3 - IDD	29.4 - COCO
	6.2 - Mapillary	45.0 - Cityscapes	60.9 - BDD100K	6.7 - Mapillary	31.3 - Cityscapes	22.4 - Cityscapes	0.6 - IDD
	4.1 - BDD100K	41.5 - ADE20K	46.2 - COCO	3.1 - IDD	27.0 - ADE20K	24.3 - ADE20K	0.2 - BDD100K
	3.1 - IDD	44.1 - COCO	44.3 - ADE20K	3.7 - BDD100K	31.0 - COCO	26.7 - COCO	0.2 - Mapillary
	3.1 - Cityscapes	2.2 - SUN-RGBD	2.6 - SUN-RGBD	3.1 - Cityscapes	1.0 - SUN-RGBD	1.1 - SUN-RGBD	0.2 - Cityscapes

(c) Ordered datasets ranked by `fd` with corresponding mIoU values. The ranking is taken from Tab. 11b, but we show the mIoUs for the corresponding datasets from Tab. 11a instead.

Validation datasets						
ADE20K	BDD100K	Cityscapes	COCO	IDD	Mapillary	SUN-RGBD
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	14.3	2.7	0.0
0.0	1.3	10.7	3.4	14.3	0.0	0.0
0.0	0.9	10.7	3.0	0.0	1.9	0.0
0.0	2.2	0.0	0.2	4.0	0.3	0.0
0.0	2.6	0.0	0.6	4.0	4.3	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
Average absolute mIoU difference: 1.66						

(d) Differences between the mIoUs of the reference ranking (Tab. 11a) and the mIoUs for the predicted `fd` ranking (Tab. 11c).

Table 11. Detailed results comparing the observed cross-dataset generalisation of the HRNet-W48 semantic segmentation models to predictions relying only on datasets using **`fd`** with the **Mugs (ViT-B/16) feature extractor**. The final value reported in Tab. 3 corresponds to the average of the values in Tab. 11d.