# DiG-IN: Diffusion Guidance for Investigating Networks - Uncovering Classifier Differences, Neuron Visualisations, and Visual Counterfactual Explanations

Maximilian Augustin  Yannic Neuhaus  Matthias Hein
Tübingen AI Center – University of Tübingen

## Abstract

*While deep learning has led to huge progress in complex image classification tasks like ImageNet, the black-box nature of their decisions is problematic. We address these problems by generating images that optimize a classifier-derived objective using a framework for guided image generation. We analyze the decisions of image classifiers by visual counterfactual explanations, maximally disagreeing images for two classifiers and visualization of neurons and spurious features. In this way, we validate existing observations, e.g. the shape bias of adversarially robust models, and new failure modes, e.g. systematic errors of zero-shot CLIP classifiers.*

## 1. Method

We introduce a Stable Diffusion (SD) [10] based framework where the image generation is directly guided by one or multiple classifiers (classifier disagreement and VCEs) or their properties (maximizing and minimizing neuron activations). We formulate our explanation tasks as an optimization problem using a loss function $L$ on the generated image. In particular, we optimize the starting latent, the conditioning vector, and the null-text that determine the output of the diffusion process and search for ones that generate an image that optimizes our loss $L$ without the need for manual prompt tuning or other forms of human supervision. We call this plug-and-play diffusion guidance framework DiG-IN.

### 1.1. Maximizing Classifier Disagreement

We generate maximally disagreeing images for a pair of two classifiers to highlight differences caused by different training types, architectures or pre-training. Forcing disagreement shifts the focus from prototypical examples of a class and makes this approach especially suitable for discovering unexpected failure modes on out-of-distribution images. Results are shown in Fig. 1.

### 1.2. Visual Counterfactual Explanations

Counterfactual reasoning has become a valuable tool for understanding the behavior of models. For image classifiers, a Visual Counterfactual Explanation (VCE) [1, 2] for input $\hat{x}$, target class $y$ and classifier $f$ is a new image $x$, that **i)** is classified as $y$ by $f$ (actionable), **ii)** looks realistic (on the natural image manifold), **iii)** contains minimal changes to the input $\hat{x}$. To generate VCEs, we invert the diffusion process [8] and use our optimization framework to maximize the confidence into the target class while simultaneously minimizing the distance to the original image outside of an automatically generated foreground mask. Our method is training-free and produces VCEs for *any* classifier trained on *any* dataset containing natural images. We thus refer to our generated counterfactuals as *Universal VCEs (UVCE)*. Results can be found in Fig. 2.

### 1.3. Neuron Visualisations

Now we visualize the semantic meaning of specific neurons in the last layer of a classification model.
**Synthetic Neuron Visualizations** Our goal is to generate prototypical examples that visualize a target neuron $n$. We use CogAgent [4] to list the objects in the most activating train images for that neuron. For each object, we use SD to generate images for that object and use the one with the highest mean activation as initialization and maximize the activation of the target neuron using DiG-IN (See Fig. 3).
**Neuron Counterfactuals** It has been shown that the neurons that are the most impactful for a classifier's decision are often activated by the image background instead of the class object [9, 12]. To visualize this, we max- or minimize the activation of a potentially spurious neuron starting from the same inversion of a *real* image we used in Sec. 1.3. Unlike for UVCEs, we now want to allow background changes to insert or remove the spurious feature while preserving the class object. To achieve this we now enforce similarity in the automatically generated foreground mask [6] while simultaneously optimizing the activation of the neuron in the image background. See Fig. 4 for results.

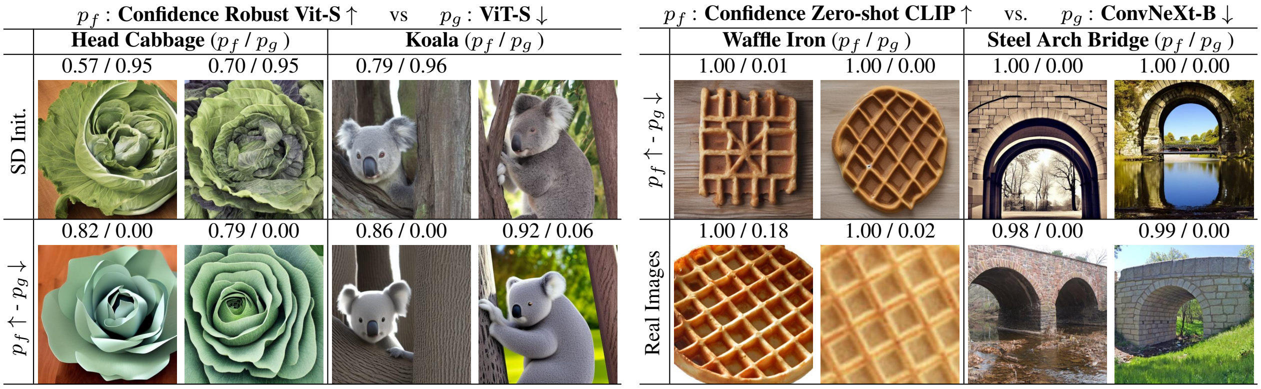
$p_f$ : Confidence Robust Vit-S ↑ vs $p_g$ : ViT-S ↓
$p_f$ : Confidence Zero-shot CLIP ↑ vs. $p_g$ : ConvNeXt-B ↓
Head Cabbage ($p_f$ / $p_g$ )
Koala ($p_f$ / $p_g$ )
Waffle Iron ($p_f$ / $p_g$ )
Steel Arch Bridge ($p_f$ / $p_g$ )
SD Init.
$p_f \uparrow$ - $p_g \downarrow$
0.57 / 0.95
0.70 / 0.95
0.79 / 0.96
0.82 / 0.00
0.79 / 0.00
0.86 / 0.00
0.92 / 0.06
$p_f \uparrow$ - $p_g \downarrow$
Real Images
1.00 / 0.01
1.00 / 0.00
1.00 / 0.00
1.00 / 0.00
1.00 / 0.18
1.00 / 0.02
0.98 / 0.00
0.99 / 0.00

Figure 1. **Classifier disagreement:** Starting from a Stable Diffusion output of "a photograph of $y$" for a given class $y$ , we maximize the confidence of a classifier $f$ while minimizing the confidence of another classifier $g$. **Left:** $f$: adversarial robust ViT-S, $g$: standard ViT-S. The resulting images retain the same shape but with smooth surfaces and little texture. **Right:** $f$: zero-shot CLIP (ImageNet), $g$: ConvNeXt-B. We find systematic misclassifications, waffles as "waffle iron" and stone bridges as "steel arch bridges", and validate them by finding similar real images in LAION-5B. The errors of CLIP are most likely an artefact of the text embeddings due to the composition of the class name.

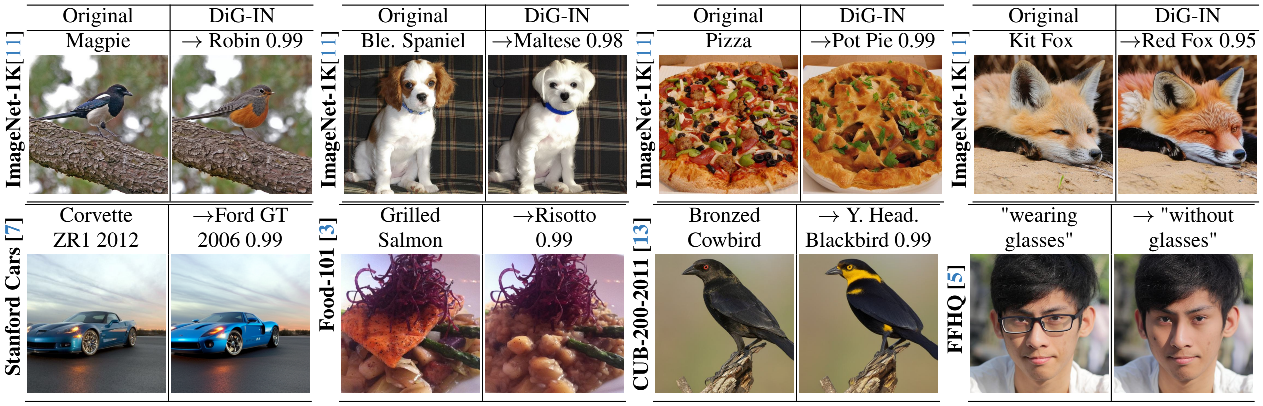

Figure 2. **UVCEs** for various datasets. DiG-IN is the first training-free method that can generate highly realistic VCEs for any dataset containing natural images without requiring a dataset-specific generative model or an adversarially robust classifier.

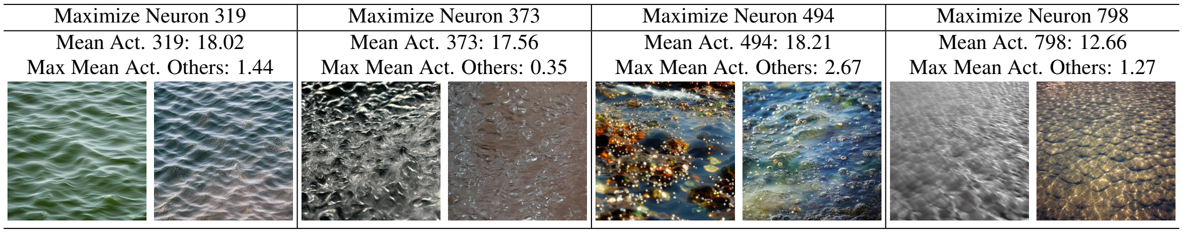

Figure 3. **Neuron visualization for a SE-ResNet-D 152 [14] trained on ImageNet:** Our neuron visualization allows to identify subtle differences between four neurons which are all activated by some kind of "water". Interestingly, the individual neurons are maximally activated only for a specific type of "water" and show no strong activations for the images generated where the other neurons are maximized.

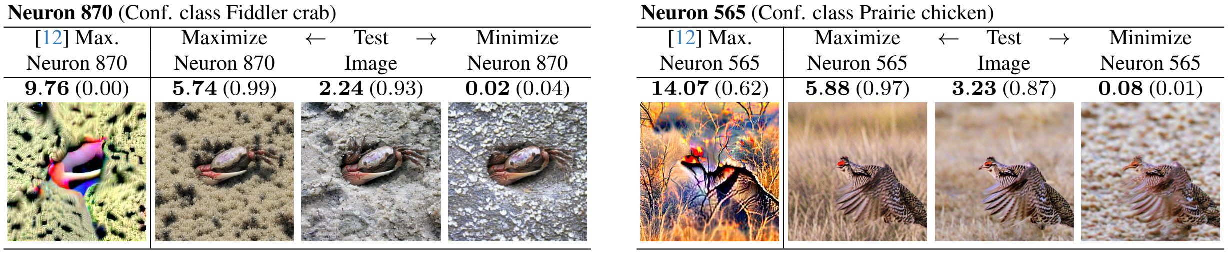

Figure 4. **Neuron Counterfactuals:** Starting from a test image, we max- and minimize the value of the corresponding spurious neuron found in [12]. As a comparison, we show the result of the maximizing feature attack [12]. Our images better convey the semantic meaning of the neuron. Maximizing the spurious neuron enhances the spurious background (left: sand, right: dry grass) while minimizing removes it.

# References

[1] Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. In *NeurIPS*, 2022. 1

[2] Valentyn Boreiko, Maximilian Augustin, Francesco Croce, Philipp Berens, and Matthias Hein. Sparse visual counterfactual explanations in image space. In *GCPR*, 2022. 1

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014. 2

[4] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023. 1

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 2

[6] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 1

[7] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 2013. 2

[8] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 1

[9] Yannic Neuhaus, Maximilian Augustin, Valentyn Boreiko, and Matthias Hein. Spurious features everywhere – large-scale detection of harmful spurious features in imagenet, 2023. 1

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 2

[12] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *ICLR*, 2022. 1, 2

[13] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 2

[14] Ross Wightman, Hugo Touvron, and Herve Jegou. Resnet strikes back: An improved training procedure in timm. In *NeurIPS Workshop on ImageNet: Past, Present, and Future*, 2021. 2