

Simple Localized Counterfactuals for Visual Explanation

Supplementary Material

A. Extended Related Works

Saliency & Prototype Methods. Saliency-based methods give importance maps that highlight which pixels in the image are the most influential in the classifier’s decision. Some approaches calculate these through perturbations [16, 17, 42], others through gradients [59], or class activation maps (CAM) [10, 29, 49, 63, 66]. Locating these pixels is a vital step to understanding the classifier’s decision, but it does not tell what visual concepts are being represented. Prototype-based methods [6, 11, 14, 36] use case-based reasoning by comparing with examples of similar training image patches. Although this improves in discovering the visual concepts utilized by the model, showing similar patches may still be too coarse to discover the specific visual concepts surrounding the decision boundary. Additionally, these methods are ante-hoc, requiring a pre-defined architecture with specified patch sizes and prototype count, further hindering their flexibility. In contrast, our method seeks to be a general approach for post-hoc insights into a model’s existing case-base reasoning.

Feature Disentanglement. One of the main reasons for models being difficult to interpret is the existence of polysemantic neurons [39, 47]. There are many works that investigate feature disentanglement [12, 13, 57], but we utilize the recent work of sparse autoencoders (SAEs) [23, 37, 40, 56, 58] since they are differential and lightweight to train. We refer the reader to the literature for the latest disentanglement approaches.

B. Extended Mask Details and Ablations

Extended Mask Details. We further improve the computational efficiency of the above mask generation process by computing the gradient-based mask on the update trajectory during the first stage of counterfactual generation by aggregating the gradients of the loss function $\nabla_{\Delta z} \mathcal{L}_{cf}$ with respect to the delta variable Δz , which we call latent aggregated gradients (LAG).

Our choice of using LAG for mask generation is validated in an ablation study on the Zebra-Sorrel task (details in Sec. 4), incorporating different masking approaches. Quantitatively, Table 3 shows that the use of LIG or LAG gives superior performance in flipping the prediction. We also compare our method with non-region constrained methods (DVCE, LCDE, GCD) in Table 1. The superiority suggests that the masking alone can improve counterfactual example construction without generative priors. Qualitatively, Fig. 6 shows that editing in the pixel space leads to more artifacting and is void of semantic meaning,

while editing in the latent space produces more meaningful changes.

Since we create a mask by optimizing Equation 3, our approach is a two stage approach where the second stage utilizes the mask resulting in the final counterfactual.

Mask Ablations. We conduct an ablation study regarding different counterfactual example generation strategies, including applying the modification at the pixel/latent space and different masking techniques. The results are shown in Fig. 6. When the original sorrel image is modified to create the counterfactual example zebra, the stripes concept appears for almost all strategies. However, the location and quality of the change vary. For pixel-based modifications, we see significant artifacts. Even with masks applied to the pixel space, there still exist a large number of modified pixels in the background. In contrast, latent-based methods generate fewer global modifications, especially when masks are applied.

For latent-based methods, the choice of masking is critical. Using our mask processing, of keeping a soft mask, thresholding below a certain value, and applying a Gaussian filter, the IG method fails to localize in our approach, while LAG succeeds in obtaining localized change. We note that in [54] binning gradients to image-grid blocks and threshold based on an overall area may produce successful results for IG, but this puts a hard constraint on the masking. We allow a soft mask, and thresholding is based on a normalized value, allowing for a varied mask-size-to-image ratio.

We conduct another ablation study to highlight the localization ability of our mask compared to other saliency-based approaches. Results can be viewed in Table 4. Our latent aggregated gradient approach is superior in the Zebra-Sorrel task, of which we randomly sample 100 images, and it competitive on the CUB dataset where we randomly sample 5 images per class.

C. Dataset Descriptions

- **CelebA** [34] contains over 200K celebrity face images annotated with 40 binary attributes. These attributes can be effectively leveraged for different categories in counterfactual experiments.
- **CelebA-HQ** [33] is a higher-resolution subset of CelebA with improved image quality and more realistic details.
- **ImageNet** [46] provides 1,000 object categories with diverse natural images, enabling us to apply the counterfactual generation on real-world scenarios.
- **CUB** [62] consists of 200 bird species, which acts as a challenging benchmark of fine-grained tasks.

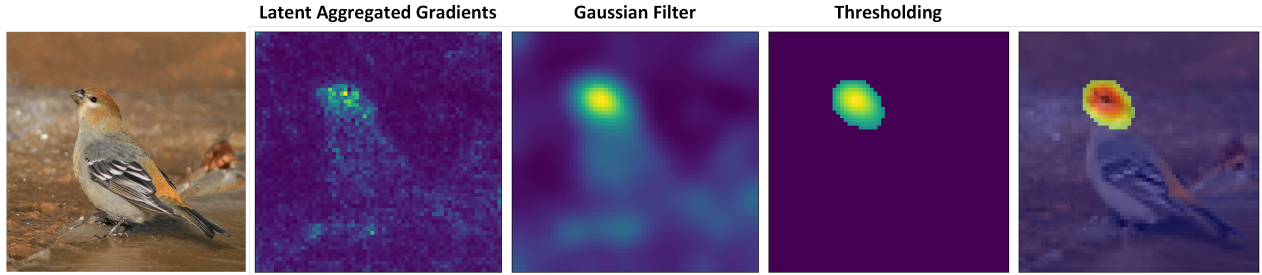


Figure 5. **Masking.** Masks are post-processed with a Gaussian filter and thresholding before being passed as a region constraint in the second stage of editing.

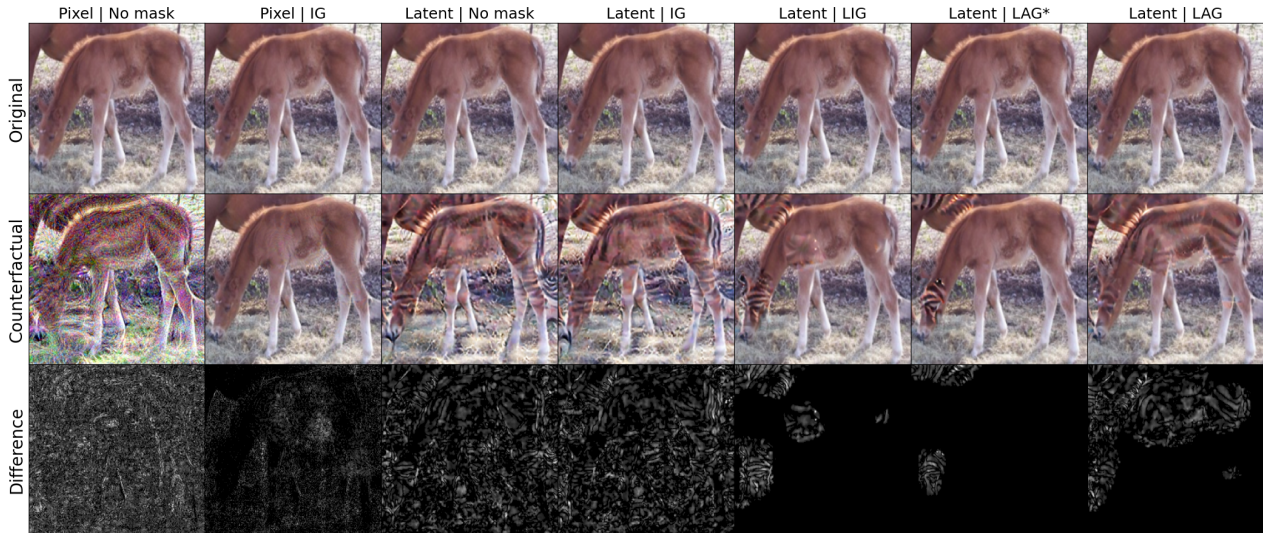


Figure 6. **Mask Ablation.** Corresponding to Table 3, this figure shows the different visual counterfactuals resulting from different editing and masking setups. The first row is the original image (repeated for viewing convenience), the second is the resulting counterfactual, and the third is a difference map between the two.

D. Training and Optimization Hyperparameters

All experiments, unless otherwise stated, have a learning rate of 0.01 using the Adam [31] optimizer, compute a maximum of 25 steps for the first stage or until the prediction is flipped to the target to obtain gradients for computing the mask. σ is set to 3, and t is set to 0.5 as defined in Equation 5. During the second stage of optimization, a small amount of perturbation is added to the masked latent area. Optimization halts either when the target class is predicted or when a maximum of 500 steps is reached. The λ_s is set to 0.1 and λ_{train} is set to 0.1. For CelebA and CelebA-HQ, we set *sigma* to 2 and for CUB we set it to 4 and set t to 0.7.

E. CelebA and CelebA-HQ Results

Quantitative Results. We use pretrained classifiers as given in [26] for CelebA and [25] for CelebA-HQ. We are

not as competitive in realism in CelebA and CelebA-HQ as shown in Table 2, but we still see superior or competitive results in sparsity as indicated by COUT. Additionally, adding our trait sparsity loss $\mathcal{L}_{\text{train}}$ (*i.e.*, Our + SAE) leads to significant increases in sparsity with the exception of CelebA-HQ on the age task.

Qualitative Results. We show that our approach is able to successfully transition between young and old by adding or removing wrinkles and move between smiling and not by exaggerating and diminishing the cheekbones in Fig. 8

F. Additional Results

Fig. 9 shows additional results on ImageNet [46]. We can see brown color added to the zebra, spots are removed from behind the mouth of the cheetah, and the Egyptian cat has spots removed and a fluffy texture is added. Additional CUB [62] results are shown in Fig. 7

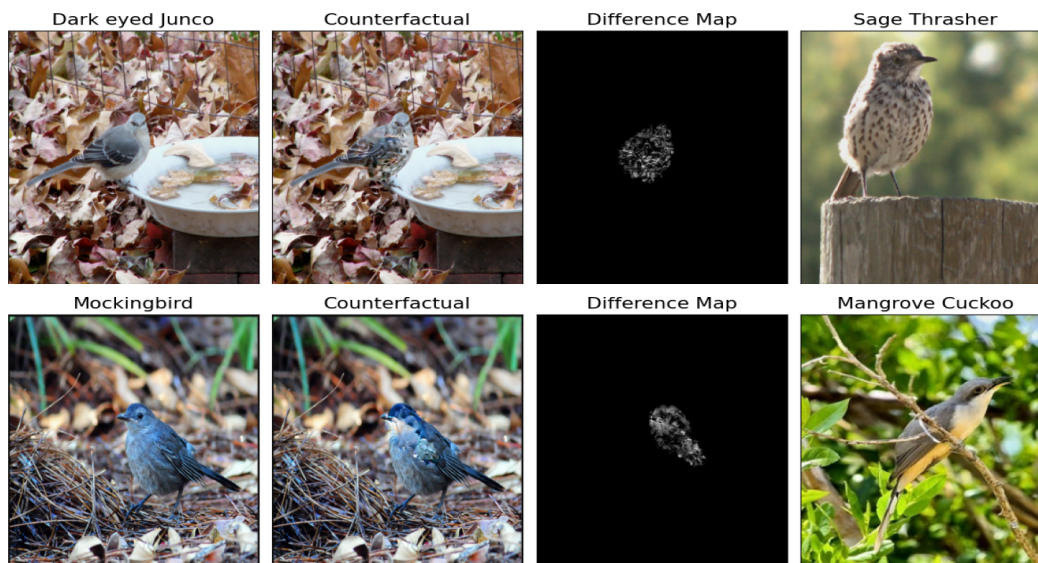


Figure 7. **CUB Qualitative Results.** Two examples are given. Each example contains four images from left to right: the original reconstruction, the visual counterfactual, the difference map between the two, and a reference example of the target class the counterfactual flipped the classifier to.



Figure 8. **CelebA and CelebA-HQ Qualitative Results.** Rows indicate the counterfactual generation task, and columns indicate the source dataset of the images shown. Each item is a triplet of images with the first being the original image reconstructed, the second being the generated counterfactual, and the third being the difference map between them.

Method	CelebA (Smile)				CelebA (Age)				CelebA-HQ (Smile)				CelebA-HQ (Age)			
	FID (↓)	sFID (↓)	COU ^T (↑)	FR (↑)	FID (↓)	sFID (↓)	COU ^T (↑)	FR (↑)	FID (↓)	sFID (↓)	COU ^T (↑)	FR (↑)	FID (↓)	sFID (↓)	COU ^T (↑)	FR (↑)
ACE l_1 [28]	1.3	4.0	0.78	97.6	1.5	4.1	0.72	96.2	<u>3.2</u>	20.2	0.55	95.0	5.3	21.7	0.40	95.0
ACE l_2 [28]	<u>1.9</u>	<u>4.6</u>	0.62	84.3	<u>2.1</u>	<u>4.6</u>	0.56	77.5	6.9	22.0	0.59	95.0	16.4	28.2	0.53	95.0
DiME [26]	3.2	4.9	0.53	97.2	4.2	5.9	0.44	99.0	18.1	27.7	0.65	<u>97.0</u>	18.7	27.8	0.56	97.0
FastDiME [65]	4.2	6.1	0.45	99.0	4.8	6.8	0.36	98.6	-	-	-	-	-	-	-	-
FastDiME-2 [65]	3.3	5.5	0.44	99.4	4.0	6.0	0.37	99.3	-	-	-	-	-	-	-	-
FastDiME-2+ [65]	3.2	5.2	0.41	98.9	3.6	5.6	0.32	98.7	-	-	-	-	-	-	-	-
LCDE-txt [48]	-	-	-	-	-	-	-	-	13.6	25.8	0.34	-	14.2	25.6	0.33	-
GCD [53]	7.2	7.7	0.40	97.2	8.7	9.1	0.32	96.0	7.3	7.9	0.51	97.5	9.5	10.1	0.30	96.0
RCSB [54]	3.0	4.8	0.87	99.8	2.9	4.9	<u>0.81</u>	99.3	3.0	20.0	<u>0.83</u>	98.9	4.9	27.3	0.80	<u>99.4</u>
Ours	9.8	17.8	0.70	100.0	9.8	12.8	0.75	100.0	4.4	<u>19.4</u>	0.79	100.0	<u>5.0</u>	<u>19.9</u>	<u>0.64</u>	99.9
Ours + SAE	9.4	12.3	0.76	<u>99.9</u>	10.3	13.2	0.82	<u>99.9</u>	5.6	20.2	0.88	<u>99.9</u>	<u>5.0</u>	<u>19.9</u>	0.62	97.7

Table 2. **CelebA & CelebA-HQ Tasks.** Results for popular counterfactual tasks smile and age are given. FastDiME [65] does report results for CelebA-HQ, and LCDE [48] does not report results for CelebA nor for flip rate (FR). Results for other methods are from [28, 48, 65]

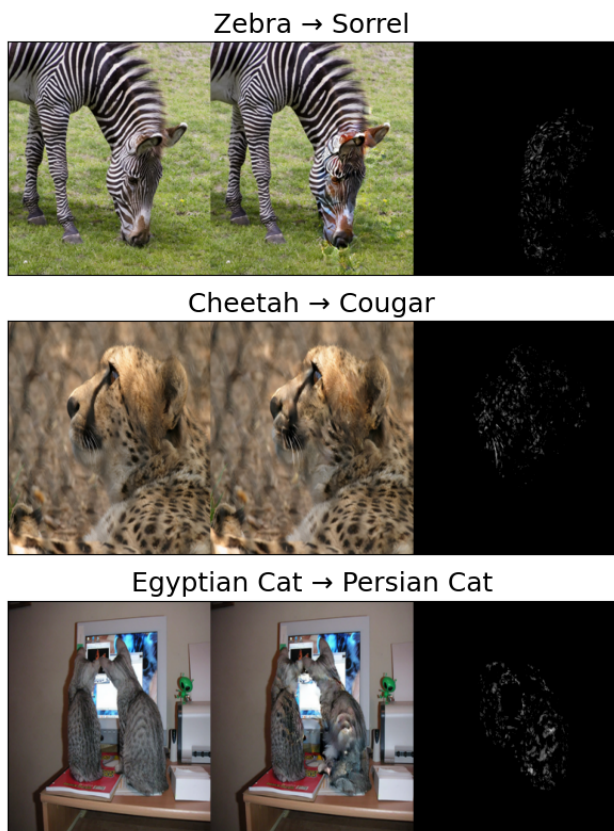


Figure 9. **ImageNet Qualitative Results.** Each image triplet is titled with the counterfactual generation direction and is comprised of the original image reconstructed (left), the visual counterfactual (middle), and the difference map between them (right).

References

- [1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018. 1
- [2] Naveed Akhtar, Mohammad AAK Jalwana, Mohammed Bennamoun, and Ajmal Mian. Attack to fool and explain

Edit	Mask	FID (↓)	sFID (↓)	S^3 (↑)	COU ^T (↑)	FR (↑)
Zebra - Sorrel						
Pixel	-	82.7	103.4	0.74	0.51	<u>99.0</u>
Pixel	IG	26.8	68.9	0.97	0.07	95.0
Latent	-	70.1	87.9	0.74	0.56	100.0
Latent	IG	60.2	82.0	0.79	0.53	100.0
Latent	LIG	48.8	80.5	<u>0.96</u>	<u>0.80</u>	98.0
Latent	LAG*	<u>45.7</u>	<u>76.9</u>	97.0	0.78	98.0
Latent	LAG	48.6	78.5	0.95	0.81	<u>99.0</u>

Table 3. **Mask Ablation.** Different masking variants are analyzed by creating counterfactual examples on a random subset of 100 images in the Zebra-Sorrel dataset. The edit is either performed in the pixel space (removing the autoencoder) or the latent space. The mask is either removed or computed using Integrated Gradients (IG), Latent Integrated Gradients (LIG), or Latent Aggregated Gradients (LAG). LAG* is calculated by performing a single optimization step in the initial editing phase.

Saliency Method	Zebra - Sorrel	CUB
LIME [43]	0.20	252.22
IxG [51]	0.12	100.28
GC [49]	0.08	41.61
LC [29]	<u>0.07</u>	<u>45.95</u>
SC [63]	0.08	94.66
GIG [30]	0.11	87.66
GBP [55]	<u>0.07</u>	64.86
IG [59]	0.11	105.26
Ours (LAG)	0.05	48.09

Table 4. **Localization Ablation.** Pixel Flipping [5] as implemented in the Quantus [22] package measures the impact of removing pixel in descending order of their attribution value. The lower the better.

deep networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5980–5995, 2021. 1

- [3] Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022. 1, 2, 4
- [4] Maximilian Augustin, Yannic Neuhäus, and Matthias Hein. Dig-in: Diffusion guidance for investigating networks - uncovering classifier differences neuron visualisations and visual counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11093–11103, 2024. 1, 2
- [5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one*, 10(7):e0130140, 2015. 4
- [6] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12):1061–1070, 2021. 1
- [7] Valentyn Boreiko, Maximilian Augustin, Francesco Croce, Philipp Berens, and Matthias Hein. Sparse visual counterfactual explanations in image space. In *DAGM German Conference on Pattern Recognition*, pages 133–148. Springer, 2022. 1, 2
- [8] Jaret Burkett. Ostris vae - kl-f8-d16, 2024. Available at <https://huggingface.co/ostris/vae-kl-f8-d16>. 1, 4
- [9] Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025. 3
- [10] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 1
- [11] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 1
- [12] Pattarawat Chormai, Jan Herrmann, Klaus-Robert Müller, and Grégoire Montavon. Disentangled explanations of neural network predictions by finding relevant subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1
- [13] Pasquale Coscia, Angelo Genovese, Fabio Scotti, and Vincenzo Piuri. Features disentanglement for explainable convolutional neural networks. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 514–520. IEEE, 2024. 1
- [14] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10265–10275, 2022. 1
- [15] Amil Dravid, Florian Schiffrers, Boqing Gong, and Aggelos K. Katsaggelos. medxgan: Visual explanations for medical classifiers through a generative latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2936–2945, 2022. 2, 3
- [16] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1
- [17] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1
- [18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2
- [19] Yash Goyal, Ziyang Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019. 1, 2
- [20] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024. 1
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4
- [22] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. 4
- [23] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 3, 1
- [24] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. *Advances in neural information processing systems*, 32, 2019. 1
- [25] Paul Jacob, Éloi Zablocki, Hedi Ben-Younes, Mickaël Chen, Patrick Pérez, and Matthieu Cord. Steex: steering counterfactual explanations with semantics. In *European Conference on Computer Vision*, pages 387–403. Springer, 2022. 2
- [26] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian conference on computer vision*, pages 858–876, 2022. 2, 4
- [27] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16425–16435, 2023. 1, 2
- [28] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16425–16435, 2023. 1, 2, 3, 4

- [29] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 1, 4
- [30] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5050–5058, 2021. 4
- [31] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2
- [32] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 693–702, 2021. 1, 2
- [33] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5549–5558, 2020. 2, 3, 1
- [34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2, 3, 1
- [35] Tung Luu, Nam Le, Duc Le, and Bac Le. From visual explanations to counterfactual explanations with latent diffusion. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 420–429. IEEE, 2025. 1, 2, 3
- [36] Chiyu Ma, Brandon Zhao, Chaofan Chen, and Cynthia Rudin. This looks like those: Illuminating prototypical concepts using multiple visualizations. *Advances in Neural Information Processing Systems*, 36:39212–39235, 2023. 1
- [37] Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013. 2, 3, 1
- [38] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. CounterGAN: Generating realistic counterfactuals with residual generative adversarial nets. *arXiv preprint arXiv:2009.05199*, 2020. 1, 2
- [39] Laura O’Mahony, Vincent Andrearczyk, Henning Müller, and Mara Graziani. Disentangling neuron representations with concept vectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3770–3775, 2023. 3, 1
- [40] Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025. 2, 3, 1
- [41] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 4574–4594. PMLR, 2022. 1
- [42] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 1
- [43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ” why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 4
- [44] Pau Rodriguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1056–1065, 2021. 2
- [45] 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*, pages 10684–10695, 2022. 1
- [46] 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. *International journal of computer vision*, 115(3):211–252, 2015. 2, 3, 4, 1
- [47] Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*, 2022. 3, 1
- [48] Simon Schrodi, Karim Farid, Max Argus, and Thomas Brox. Latent diffusion counterfactual explanations. In *DAGM German Conference on Pattern Recognition*, pages 295–311. Springer, 2024. 1, 2, 4
- [49] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 4
- [50] Sheng-Min Shih, Pin-Ju Tien, and Zohar Karnin. Ganmex: One-vs-one attributions using gan-based model explainability. In *International Conference on Machine Learning*, pages 9592–9602. PMLR, 2021. 1, 2
- [51] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016. 4
- [52] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 4
- [53] Bartłomiej Sobieski and Przemysław Biecek. Global counterfactual directions. In *European Conference on Computer Vision*, pages 72–90. Springer, 2024. 1, 2, 4
- [54] Bartłomiej Sobieski, Jakub Grzywaczewski, Bartłomiej Sadlej, Matthew Tivnan, and Przemysław Biecek. Rethinking visual counterfactual explanations through region constraint. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 3, 4

- [55] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 4
- [56] Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models. *arXiv preprint arXiv:2502.06755*, 2025. 2, 3, 1
- [57] Jiayu Su, David A Knowles, and Raul Rabadan. Disentangling interpretable factors with supervised independent subspace principal component analysis. *Advances in Neural Information Processing Systems*, 37:37408–37438, 2024. 1
- [58] Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. Spine: Sparse interpretable neural embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2, 3, 1
- [59] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1, 4
- [60] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2
- [61] Inam Ullah, Muwei Jian, Sumaira Hussain, Jie Guo, Hui Yu, Xing Wang, and Yilong Yin. A brief survey of visual saliency detection. *Multimedia Tools and Applications*, 79:34605–34645, 2020. 1
- [62] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 3, 4, 1
- [63] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 1, 4
- [64] Yinong Oliver Wang, Eileen Li, Jinqi Luo, Zhaoning Wang, and Fernando De la Torre. Unsupervised model diagnosis. *arXiv preprint arXiv:2410.06243*, 2024. 1, 2
- [65] Nina Weng, Paraskevas Pegios, Eike Petersen, Aasa Feragen, and Siavash Bigdeli. Fast diffusion-based counterfactuals for shortcut removal and generation. In *European Conference on Computer Vision*, pages 338–357. Springer, 2024. 1, 2, 3, 4
- [66] Ziheng Zhang, Jianyang Gu, Arpita Chowdhury, Zheda Mai, David Carlyn, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. Finer-cam: Spotting the difference reveals finer details for visual explanation. *arXiv preprint arXiv:2501.11309*, 2025. 1
- [67] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*, 2017. 1, 2