Concepts Guide and Explain Diffusion Visual Counterfactuals

Franz Motzkus[1,2,*], Ute Schmid[2]

[1]Continental Automotive GmbH, Max-Urich-Straße 3, 13355 Berlin, Germany
[2]Universität Bamberg, An der Weberei 5, 96047 Bamberg, Germany

Abstract

Diffusion models enable the generation of diverse yet realistic image features, which is crucial in counterfactual generation answering "what if" questions of what needs to change to make an image classifier change its prediction. Current methods generate authentic counterfactuals, but lack transparency in the feature changes. To address this limitation, we introduce Concept-guided Latent Diffusion Counterfactual Explanations (CoLa-DCE), a concept-guided approach for any classifier that provides a high degree of control over concept selection and spatial conditioning. The counterfactuals comprise an increased granularity through minimal feature changes, improved comprehensibility through feature visualization, and increased transparency by localizing feature changes. We show the advantages of our approach in minimality and interpretability extensively across multiple datasets, classification, and diffusion models and demonstrate how our CoLa-DCE explanations make model errors like misclassification cases comprehensible.

Keywords

explainable AI (xAI), Counterfactuals, Image-to-Image Diffusion, Concept Encodings

1. Introduction

Recent advancements in generative models have sparked new interest in counterfactual explanations for computer vision tasks [1, 2, 3]. By answering what would need to change to induce a different outcome, counterfactual explanations are well-aligned with human reasoning [4, 5] and are deemed plausible, if they are consistent with the user's beliefs - realistic, and with minimal effort to change towards the counterfactual [5]. Diffusion models generate realistic high-resolution images with diverse features within the data distribution [6, 7, 8], designating them for generating counterfactual images [1, 3, 2]. Previous diffusion-based counterfactuals optimize all image features but lack clarity on specific feature changes and their relation to the model prediction, making detection and tracking feature changes challenging. As humans perceive minimal counterfactual differences semantically rather than pixel-wise, defining minimality in the number of feature changes is better suitable.

CoLa-DCE solves both problems: We guide the counterfactual generation with a restricted number of semantic concepts, further enabling a high level of control by concept selection. We additionally include feature visualization capabilities, allowing for direct comprehensibility of features that represent the difference between the original and the counterfactual class. Hereby, CoLa-DCE provides semantic as well as spatial guidance and visualization, simultaneously enabling control and better transparency.

1. We introduce CoLa-DCE for the diffusion-based generation of counterfactuals using semantic concept-guidance. We show how local counterfactual targets and concept-guided feature changes derived from the classifier's perception increase the quality of counterfactuals.
2. We extend our concept guidance with spatial conditioning, guiding and revealing localized feature changes, that are made comprehensible via concept visualization and localization maps.
3. We show how CoLa-DCE assists in model debugging by making cases of misclassification more understandable by exposing feature-level information.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

[*] Corresponding author.

franz.walter.motzkus@continental-corporation.com (F. Motzkus); ute.schmid@uni-bamberg.de (U. Schmid)
0009-0009-4362-7907 (F. Motzkus); 0000-0002-1301-0326 (U. Schmid)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

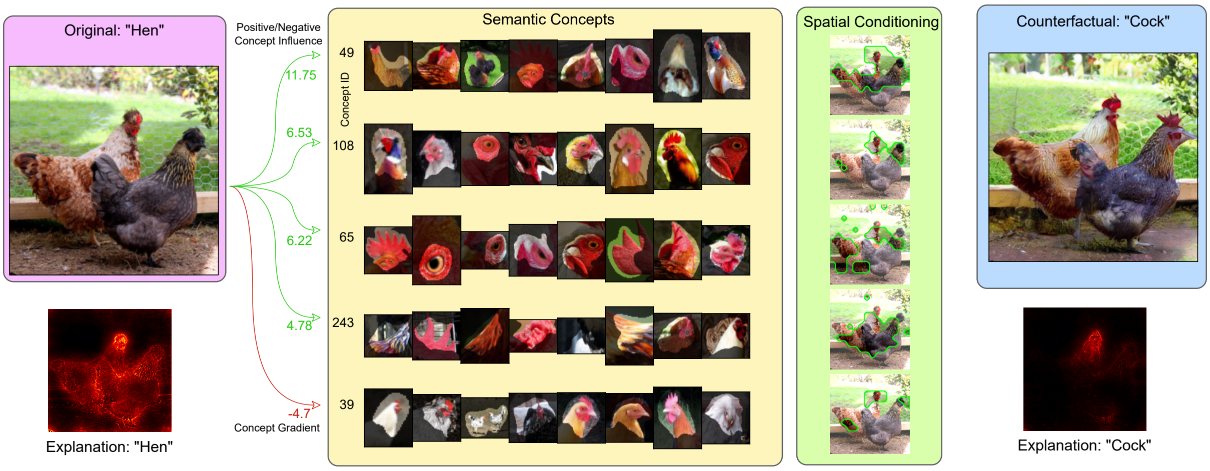


Figure 1: Example image of a CoLa-DCE counterfactual consisting of the top-k concepts with reference samples, a localization map per concept indicating the concept regions, and the generated counterfactual.

2. Related Work

2.1. Counterfactual Image Generation

Diverse approaches for computing image counterfactuals exist. CVE [9] replaces feature regions in an image with matching image patches from a distractor image of the counterfactual class. Other methods directly optimize the image using specific loss functions [10, 11] or use an autoencoder to control the optimization or modification in a disentangled latent space [12] or simplified interpretable space [13].

DiME [2] introduces diffusion models for generating counterfactuals, using a classifier to guide the diffusion process. However, it is limited to small perturbations as required in the CelebA dataset. ACE [14] is a two-step process consisting of computing pre-explanations and refining them. A localization mask for the most probable feature change is computed before repainting the image by combining the generated counterfactual within the mask with the original image outside.

DVCE [1] includes an additional robust classifier to relax the robustness constraint for the tested classifier and aligns the gradients of both models. However, generated features might be induced by the robust classifier, decreasing the faithfulness towards the original classifier. LDCE [3] overcomes the robustness requirement for the classifier by constructing a consensus mechanism, aligning the gradients of the external and the diffusion model's implicit classifier directly. However, feature changes are hard to track due to the optimization on all features and thus lack transparency. A concept-based approach can improve the transparency and comprehensibility to the user by modifying features on a semantic concept level while enforcing semantic minimality in the number of feature changes.

2.2. Local Concept Attribution

Layer-wise Relevance Propagation (LRP) [15] describes a local attribution method that backpropagates a modified gradient to assign pixel-wise importance scores for a target class. Concept-wise Relevance Propagation (CRP) [16] extends LRP to concept space by defining every neuron or channel in the latent space as a concept. A concept mask filters attributions during the LRP backward pass, retaining the concept attribution in input space. Relevance Maximization [16] assigns semantic meaning to channels by analyzing the constrained explanations across samples. Our approach generalizes the concept masking for a gradient manipulation and applies Relevance Maximization to visualize key concepts.

3. Concept-guided Latent Diffusion Counterfactual Explanations

As depicted in Figure 2 CoLa-DCE introduces three major improvements: Selecting sample-specific targets (yellow), conditioning on a set of concepts (orange), and further spatial conditioning (purple).

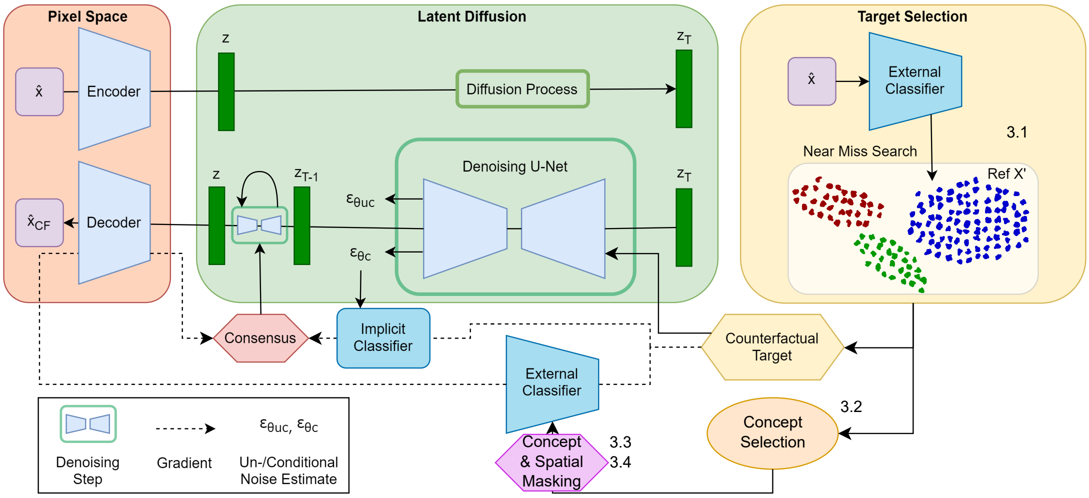


Figure 2: A simplified overview of the model architecture for our CoLa-DCE approach, including the target selection (right) and the concept-conditioning for guiding the diffusion denoising (middle).

3.1. Local Counterfactual Targets

To select the counterfactual target class, we use the model's perception of the respective data sample and compare it to the perception of a reference dataset X'.The model perception can hereby be derived by either computing the activation of the model for each sample in a selected layer or by computing the intermediate attribution using a local eXplainable Artificial Intelligence (xAI) method like LRP [15]. As the model perception of the data shall be represented, the class predictions are used to determine class affiliation. For a new sample $\hat{x} \in \hat{X}$ that we want to generate a counterfactual for, we derive the model prediction and feature space encoding $\kappa(\hat{x})$ and compare it to the encodings of the reference dataset. Hereby, based on the feature space encodings, the closest reference point with a differing class prediction is extracted, resembling the near miss approach [17]. The counterfactual target y_c is then defined as the predicted class of the reference point x'.

$$y_c = f(argmin_{x' \in X'} d(\kappa(x'), \kappa(\hat{x})) \quad \text{and} \quad f(x') \neq f(\hat{x})) \tag{1}$$

3.2. Concept Selection

For the counterfactual target y, the gradient $\nabla_x p(x|y)$ of a sample x can be extracted at each network layer. For a selected layer l, the intermediate gradient is summed over the spatial dimensions to obtain a one-dimensional representation over the channels, encoding a particular concept each [16]. Taking the absolute value of the summed gradients, the top-k concepts with $k \in \mathcal{N}(1, K)$, and K denoting the overall number of channels, are selected, which are most likely to induce a change towards the counterfactual class. The concepts are visualized with feature visualization methods like CRP [16].

3.3. Concept Conditioning

Classifier-free diffusion guidance [8] separates the conditioning into an unconditional part and a conditional part, where the difference between both parts can be used as an implicit classifier score:

$$\nabla_x \log p_\eta(x|c) = \nabla_x \log p(x) + \eta \nabla_x \log p(c|x). \tag{2}$$

This gradient-based score can be further modified by gradient manipulation to control the counterfactual generation. Motivated by the LDCE [3] algorithm, we condition the diffusion process solely on the selected concepts. The conditions require precomputation and remain fixed during the counterfactual generation. Instead the original gradient of the external classifier $\nabla_x p(x|y)$ for target y, the conditioned gradient with regards to the selected concepts λ_1, ...,λ_k with binary constraints θ_1, ..., θ_k is used. With layer l splitting the model into two parts $p(x|y) = h(g(x|y)|y)$, the conditioned gradient is computed as:

$$\begin{aligned} \nabla_x p(x|y, \theta_1 ... \theta_k) &= \nabla_x (h(g(x))|y, \theta_1 ... \theta_k) \\ &= \delta(\nabla_{g(x)} h, \theta_1 ... \theta_k) \cdot \nabla_x g \\ \text{with} \quad \delta(\nabla_{g(x)} h, \theta_1 ... \theta_k)_j &= \begin{cases} \nabla_{g(x)} h_j, & \text{if } j \in \{\theta_1, ..., \theta_k\} \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

with δ indicating binary masking the latent space gradient in the selected layer. The masked latent gradient can be backpropagated to the input without further constraints.

3.4. Spatial Conditioning

While the concept conditioning focuses on semantic feature changes, the spatial dimensions of the intermediate activations provide local information. We assume that each feature should be only changed at a single location or that the feature gradient is approximately identical in equivalent locations. We add binary masking to the spatial dimensions similar to Equation 3 based on the gradient for the selected features, zeroing gradients below a threshold η. For visualization, the binary mask can additionally be upscaled to the input scale like in Net2Vec [18], yielding additional information about where a specific concept is expected to change towards the counterfactual. The spatial conditioning minimizes the feature change by restricting it locally, while the feature localization improves the comprehensibility.

4. Results

We test our approach on the ImageNet [19] validation dataset using pre-trained Torchvision models Torchvision: a VGG16 (with/without batch normalization), a ResNet18, and a ViT model. To derive appropriate targets, 90% of the validation data is used as reference data. Counterfactuals are generated and evaluated on the remaining 1000 samples, all ImageNet classes included. We inherit the parametrization parameters from LDCE [3]. Similar to [3], we evaluate the minimality via the FID score [20] as well as the L1 norm between the original and counterfactual image to measure their semantic and pixel-based distance. The flip ratio (FR) determines the accuracy of predicting the target class.

4.1. Selecting a local target results in improved counterfactuals

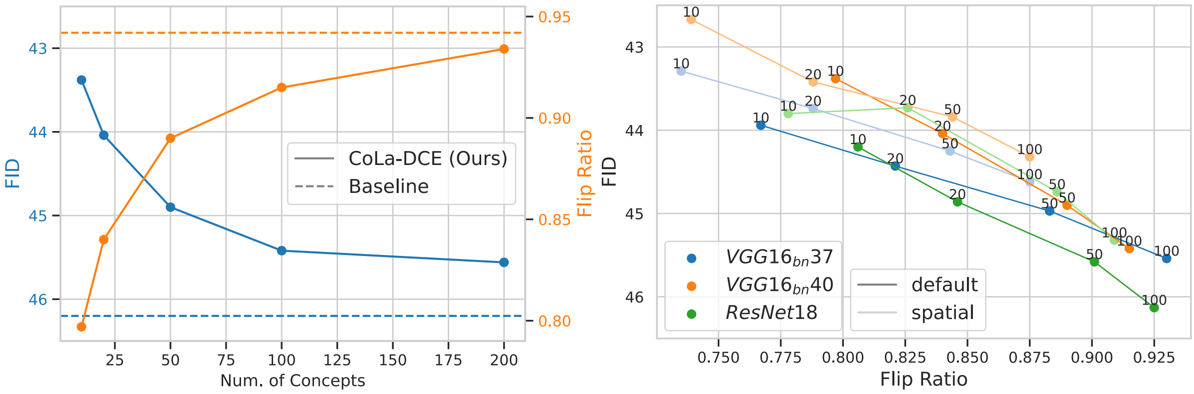
While LDCE [3] uses WordNet [21] to derive counterfactual targets based on the semantic similarity between labels, we suggest using the classifier's perception of the local input. Table 1 shows the influence of the target selection on the generated samples' quantitative performance metrics. Choosing a local (sample-based) counterfactual target on a near-miss basis leads to an improved flip ratio and confidence in all settings, demonstrating a closer decision boundary and more superficial change between the original and target class. However, retrieving the target via the intermediate activation may lead to a slightly increased FID compared to the baseline, and some counterfactual targets are not semantically connected to the original class, requiring a more substantial semantic change.

Using the intermediate LRP [15] attribution yields substantial improvements in the minimal change needed while simultaneously achieving high flip ratios. This indicates semantically similar counterfactuals close to the original images. Including the model's classification in the intermediate attribution rather than only considering the activation up to the selected layer may better represent how the features in the layer are connected toward the output, comprising top-level semantics between classes. Thus, fewer feature changes are necessary. Including the results of our CoLa-DCE method, even closer

Table 1
Quantitative comparison showing the effect of the target selection on the generated counterfactuals using the LDCE method in comparison to our CoLa-DCE method (k=20, mean score over 1000 samples). The sFID results are omitted, as they reflect the FID. The confidence score denotes the output probability.

Model/Method	Target	Layer	FID ↓	L1 ↓	Flip Ratio ↑	Confidence ↑
VGG16 - LDCE	Base	-	55.46	12458	0.851	0.81
VGG16 - LDCE	Act	feat.37	59.12	12456	0.936	0.89
VGG16 - LDCE	Attr	feat.37	45.56	**12443**	**0.956**	**0.92**
VGG16 - CoLa-DCE	Attr	feat.37	**44.43**	13915	0.821	0.81
Resnet18 - LDCE	Base	-	55.86	12518	0.846	0.79
Resnet18 - LDCE	Act	4.1.c1	57.46	12502	**0.96**	**0.91**
Resnet18 - LDCE	Attr	4.1.c1	46.28	**12465**	0.957	**0.91**
Resnet18 - CoLa-DCE	Attr	4.1.c1	**44.86**	13933	0.846	0.84
ViT - LDCE	Base	-	59.48	**12533**	0.833	0.81
ViT - LDCE	Act	encoder	53.75	14024	0.913	0.88
ViT - LDCE	Attr	encoder	53.24	14028	**0.917**	**0.89**
ViT - CoLa-DCE	Attr	encoder	**53.21**	14003	0.847	0.83

counterfactuals are generated with flip ratios on par with the LDCE baseline. Reconsidering the hard constraint on the number of concepts, damping the gradient signal, CoLa-DCE yields much more transparent counterfactuals while still being competitive to the baseline.



(a) Tradeoff between FID and Flip Ratio

(b) CoLa-DCE Model Comparison

Figure 3: Quantitative evaluation specifying the tradeoff between the number of concepts and the flip ratio/FID. The results in 3a are derived for the VGG16bn with target layer `feat.40`. The baseline is the LDCE method.

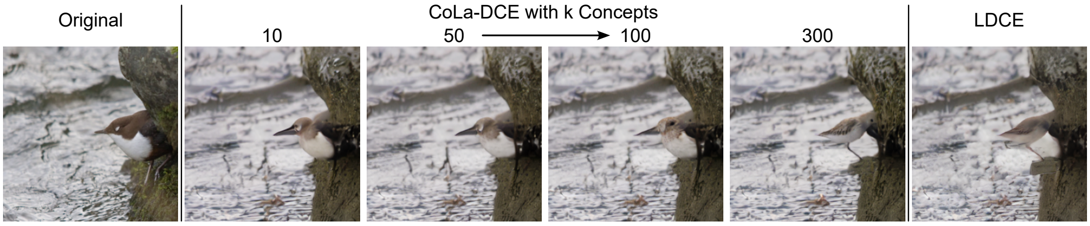


Figure 4: CoLa-DCE explanations (“water ouzel” to “red-backed sandpiper”) with a differing number of concepts k and and the VGG16bn with concept layer `40`. Limiting the concept number induces more fine-grained feature perturbations than the baseline LDCE, flipping the shown bird completely.

4.2. The number of concepts is a tradeoff between accuracy and comprehensibility

For the minimality constraint, we express the minimal semantic change in the number of feature/concept changes. While mostly a handful of concepts is used [16, 22], restricting the latent space gradient from hundreds to few channels significantly reduces the gradient for diffusion guidance. Our quantitative study (Figure 3) assesses how the concept number influences the quality of counterfactuals regarding accuracy (flip ratio) and minimality (FID). Restricting the concept number improves the FID (minor change) while the flip ratio decreases. Masking the gradient causes fewer feature changes, but also attenuates the shift towards the counterfactual class. Only ten concepts can already achieve a good performance $> 75\%$ regarding the flip ratio, while the FID score outperforms the baseline. Thus, CoLa-DCE offers concept-based transparency and control without losing much detail or accuracy. Figure 3b depicts the tradeoff between minimality and accuracy for multiple model architectures and settings. Adding spatial constraints per concept slightly lowers flip ratios, but improves the FID. Figure 4 illustrates how the number of concepts affects the counterfactual generation. Restricting the concepts causes minor changes that alter the target object semantically, while too many concepts (like LDCE) induce an alteration of the image composition.

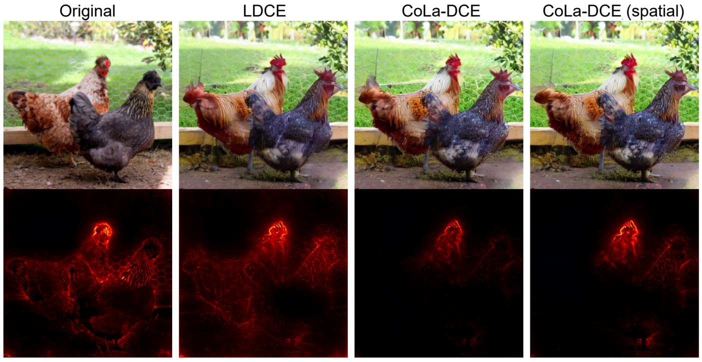


Figure 5: Comparison of the counterfactual images and their explanations for LDCE and our proposed method CoLa-DCE w/o and with spatial constraints.

4.3. Spatial constraints per concept improve the focus

Assuming each feature is locally restricted, we add spatial constraints per concept by thresholding the gradient. In Figure 1, changes towards the cockscomb are only reasonable near the hen's head, with the gradient set to zero elsewhere. In Figure 5, CoLa-DCE yields much more sparse explanations than LDCE, highlighting fewer and more concentrated feature changes. With added spatial constraints, a stronger focus in the explanation becomes apparent, either having more sparse explanations or reflecting a stronger focus on single semantic features. Performance-wise, the spatial conditioning further decreases the FID for the better, while only slight drawbacks regarding the flip ratio occur.

4.4. How can concept-based counterfactuals help in explaining model failures?

Counterfactuals are especially useful when explaining samples at the classifier's decision boundary between two classes. When misclassified samples and their correctly classified counterfactuals are inspected using our CoLa-DCE approach, the root cause of the misclassification in terms of identified or missing features becomes apparent. Figure 6 describes a misclassification case where the original image lacks specific evidence of belonging to the label “brambling”. The sample seems to represent a rare case of the class where the classifier is missing essential concepts shown in the CoLa-DCE explanation for a correct classification. Hence, a dataset or model adaptation is required, where more samples of the class showing the necessary concepts can be included in a model finetuning.

5. Limitations

As ground truth information of an optimal counterfactual image does not exist, only heuristics containing desired properties for counterfactuals can be optimized. However, the right balance between minimally deviating the image while maximizing the flip ratio depends on a rough estimate of the user's preferences. Parameter optimization is also required to balance the influence of the external gradient and the reconstruction accuracy, like in LDCE[3]. We acknowledge that the diffusion model's ability to accurately reconstruct an image and generate similar concept information as the external classifier highly influences the counterfactual quality. Poor results are expected for out-of-distribution data, as the needed features are naturally not captured by the diffusion model and cannot be generated.

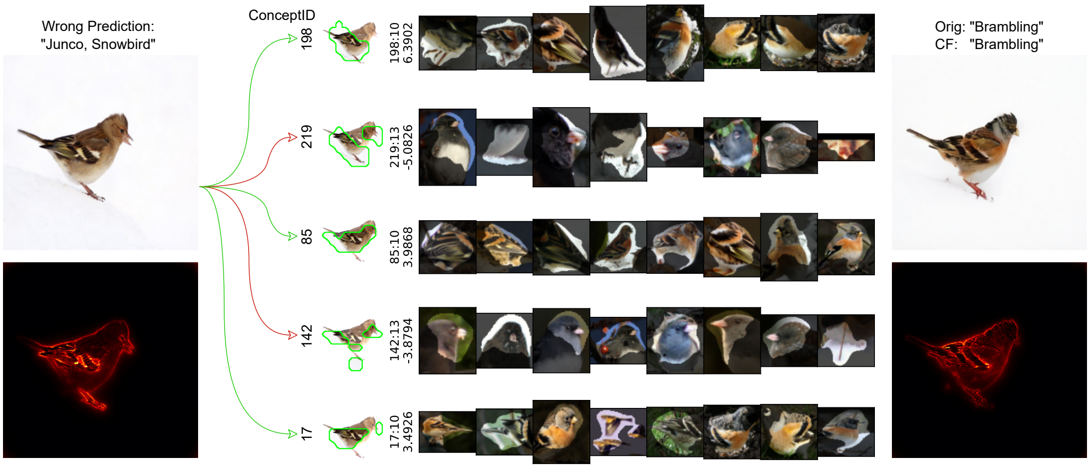


Figure 6: A CoLa-DCE explanation for a misclassified sample, that the VGG16bn classifies as "Junco, Snowbird". For a correct classification as "Brambling", the orange chest color, a different feather pattern, and a gray-blueish head are needed. Head and beacon shall look less similar to "Junco, Snowbird".

6. Conclusion

Our CoLa-DCE method successfully tackles the lack of transparency and fine-grained control in diffusion-based counterfactual generation methods. Starting from an improved target selection, we show how our concept-based approach yields semantically fewer image changes, enforcing the minimality requirement. By restricting concepts and applying spatial constraints, the counterfactual generation is more focused on small, localized feature perturbations, which are additionally more comprehensible due to the concept grounding. From our CoLa-DCE explanations, it is directly deducible which feature changes at which location cause the prediction change of the classifier, strongly improving the transparency and understandability to a human user. With the high degree of control in generating images with CoLa-DCE, we are confident to induce further work using fine-grained concept guidance for image alteration tasks.

Declaration on Generative AI

During the preparation of this work, the authors used Chat-GPT-4 and Grammarly for grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Augustin, V. Boreiko, F. Croce, M. Hein, Diffusion visual counterfactual explanations, in: Advances in Neural Information Processing Systems, volume 35, 2022, pp. 364–377.

[2] G. Jeanneret, L. Simon, F. Jurie, Diffusion models for counterfactual explanations, in: Computer Vision – ACCV 2022, Springer Nature Switzerland, Cham, 2023, pp. 219–237.

[3] K. Farid, S. Schrodi, M. Argus, T. Brox, Latent diffusion counterfactual explanations, 2023. arXiv:2310.06668.

[4] D. Lewis, Counterfactuals and comparative possibility, Journal of Philosophical Logic 2 (1973) 418–446.

[5] R. M. J. Byrne, Précis of the rational imagination: How people create alternatives to reality, Behavioral and Brain Sciences 30 (2007) 439–453. doi:10.1017/S0140525X07002579.

[6] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 6840–6851.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10674–10685. doi:10.1109/CVPR52688.2022.01042.

[8] J. Ho, T. Salimans, Classifier-free diffusion guidance, 2022. arXiv:2207.12598.

[9] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, Counterfactual visual explanations, in: Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 2376–2384.

[10] M. Augustin, A. Meinke, M. Hein, Adversarial robustness on in- and out-distribution improves explainability, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 228–245.

[11] V. Boreiko, M. Augustin, F. Croce, P. Berens, M. Hein, Sparse visual counterfactual explanations in image space, in: Pattern Recognition, Springer International Publishing, Cham, 2022, pp. 133–148.

[12] P. Rodríguez, M. Caccia, A. Lacoste, L. Zamparo, I. Laradji, L. Charlin, D. Vazquez, Beyond trivial counterfactual explanations with diverse valuable explanations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1056–1065.

[13] M. Zemni, M. Chen, E. Zablocki, H. Ben-Younes, P. Pérez, M. Cord, Octet: Object-aware counterfactual explanations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15062–15071.

[14] G. Jeanneret, L. Simon, F. Jurie, Adversarial counterfactual visual explanations, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16425–16435.

[15] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLOS ONE 10 (2015) 1–46.

[16] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin, From attribution maps to human-understandable explanations through concept relevance propagation, Nat. Mac. Intell. 5 (2023) 1006–1019. doi:10.1038/S42256-023-00711-8.

[17] J. Rabold, M. Siebers, U. Schmid, Generating contrastive explanations for inductive logic programming based on a near miss approach, Machine Learning 111 (2022) 1799–1820.

[18] R. Fong, A. Vedaldi, Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2018.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009.

[20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. 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.

[21] G. A. Miller, Wordnet: a lexical database for english, Commun. ACM 38 (1995) 39–41.

[22] M. Dreyer, R. Achtibat, T. Wiegand, W. Samek, S. Lapuschkin, Revealing hidden context bias in segmentation and object detection through concept-specific explanations, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2023, pp. 3829–3839.