Faithful Counterfactual Visual Explanations (FCVE)

Bismillah Khan[a], Syed Ali Tariq[a], Tehseen Zia[a], Muhammad Ahsan[b], David Windridge[c]

[a]*Department of Computer Science, COMSATS University Islamabad, Pakistan*
[b]*The City School, Ravi Campus, Pakistan*
[c]*Middlesex University London, United Kingdom*

Abstract

Deep learning models in computer vision have made remarkable progress, but their lack of transparency and interpretability remains a challenge. The development of explainable AI can enhance the understanding and performance of these models. However, existing techniques often struggle to provide convincing explanations that non-experts easily understand, and they cannot accurately identify models' intrinsic decision-making processes. To address these challenges, we propose to develop a counterfactual explanation (CE) model that balances plausibility and faithfulness. This model generates easy-to-understand visual explanations by making minimum changes necessary in images without altering the pixel data. Instead, the proposed method identifies internal concepts and filters learned by models and leverages them to produce plausible counterfactual explanations. The provided explanations reflect the internal decision-making process of the model, thus ensuring faithfulness to the model.

Keywords: Explainable AI, visual explanation, counterfactual

1. Introduction

Deep convolution neural networks (DCNNs) are at the leading edge of technology in many advanced areas of computer vision applications such as healthcare [1], criminal justice [2], banking finance decisions [3], transportation [4], agriculture [5], fraud detection [6] and scene segmentation [7], etc. The extensive use of deep convolutional neural networks over a conventional neural network is due to the fact that they are computationally competitive,

automatically learn a hierarchy of representations from the input data [8], and are agile compared to neural networks [9].

However, DCNNs are opaque in nature as their innards are not properly understood and visible, making them a black-box [10]. The DCNN models need to be transparent for safety-critical applications such as healthcare and criminal justice that involve dealing with human life [11, 12] and driverless vehicles [13], etc., in which the effect of inaccurate or undesired decisions have significant consequences [14, 15, 16]. Several studies show that DCNNs often regard dataset bias [17] and rely on undesired or inappropriate features to take decisions. The DCNNs also produce incorrect results when subtle changes are made to the input [18]. Adversarial attacks cause risk to several security-critical applications, for instance, in driver-less vehicles where slight obstructions on traffic signs can result in undesired conclusions [19] or in surveillance systems where malevolent individuals may cause harm [20]. Therefore, DCNNs are unreliable and need explainable AI (XAI) approaches to determine their deficiencies and train trustworthy, robust, and transparent models [16, 21, 22].

Different types of (XAI) methods exist in the literature and can be categorized into two dominant groups: ante-hoc [23] and post-hoc [24]. Although the ante-hoc models have intrinsically explainable model structure, the explainability comes at the cost of lower performance. The post-hoc models tend to explain other pre-build black-box models; hence they do not compromise performance at the expense of explainability. Among many post hoc techniques, counterfactual and contrastive explanations have emerged as powerful visual explanation types.

Contrastive explanations usually identify the actual features of the input data that play an important role in model decision-making for the inferred class [25]. Such explanations are meaningful as they imitate the process of human thoughts and are easily understood. Counterfactual explanations describe what features need to be modified and to what extent to flip the decision of the model (i.e., to reverse an undesired outcome). Counterfactual explanations offer recourse by trying to find the minimum change in the input data to obtain a positive result [26, 27, 28, 29]. On the basis of such explanations, we come across the reasons behind the model predictions; hence we can either accept or reject the given prediction accordingly. Several contrastive and counterfactual explanation methods have been proposed recently [30, 31, 25, 32, 33], in which certain input data pixels are perturbed to alter the model's prediction.

However, a critical shortcoming of these existing approaches is that they are not faithful (or aligned) to the model and do not make the model transparent (i.e. glass-box, rather than black-box) in terms of its reasoning process. Further, these methods aim to find the optimal combination of pixels for perturbing or in-filling, and they are computationally expensive [28]. Addressing the issue, a recent study [34] deals with super-pixels rather than pixels to find crucial decisive concepts that, when deleted from or added to the query image, affect the model's decision. Despite generating useful explanations, this method is not faithful and glass-box transparent as it generates explanations by operating on pixel data. Another line of research aims to identify whether a particular concept has some significance to a given model [35]. This approach, however, neither investigates the internal reasoning of DCNN models nor provides counterfactual and contrastive explanations.

This research is based on a recent study [36], which deals with identifying counterfactuals and contrastive filters of DCNN models rather than pixels of an image. Despite generating counterfactual and contrastive filters, the approach does not provide visual explanations of the generated counterfactual and contrastive ones. Hence, this study aims to suggest a post-hoc explainability method that visually explains the predictive identification of counterfactuals and contrastive filters in a DCNN model.

In this regard, the proposed solution is a Faithful Counterfactual Visual Explanation (FCVE) model.

2. Related Work

The authors in [30] propose a method for generating counterfactual visual explanations to provide insights into the decision-making process of deep learning models. The authors employ GANs to generate alternative images that would have led to different model predictions. They use a conditional GAN framework where the generator is conditioned on the input image and a desired output class. While the generated counterfactuals are visually plausible, the evaluation of faithfulness, i.e., the degree to which the generated explanations accurately reflect the model's internal reasoning, is not extensively discussed. In the paper [37] authors introduce an intriguing approach to generate counterfactual explanations using latent space transformations. The authors propose a method that leverages the power of generative models, specifically CycleGAN, to produce counterfactual instances by mapping an original instance to a counterfactual representation in the latent space

and then back to the input space. This cycle-consistency constraint ensures that the generated counterfactuals retain important features of the original instance while introducing meaningful modifications. The paper not only focuses on generating counterfactuals but also discusses their utility in enhancing interpretability and fairness in machine learning models. This broader perspective strengthens the paper's significance and relevance in addressing the need for explainable AI systems. While the paper presents an innovative and promising approach, it also has some limitations worth considering. One limitation is the reliance on CycleGAN, which may not capture all the complexities of the original input space. Exploring alternative generative models or incorporating additional constraints could further improve the fidelity and relevance of the generated counterfactual explanations.

The authors in [38] present a method that generates counterfactual explanations for image classifiers. The approach utilizes GANs to generate alternative images by perturbing the input image in a semantically meaningful way. By incorporating contrastive loss and regularization terms, the authors aim to ensure the plausibility and faithfulness of the generated counterfactuals. However, the evaluation of faithfulness is not explicitly addressed, and more rigorous analysis is necessary to determine the extent to which the explanations align with the internal reasoning of the classifier. In [39], the main focus is on generating plausible counterfactual and semi-factual explanations for deep learning models. The authors propose a method that combines an encoder-decoder architecture with variational autoencoders (VAEs) to generate counterfactual explanations. The generated explanations are evaluated based on their plausibility and faithfulness. The authors provide qualitative analyses and comparisons to demonstrate the faithfulness of their approach, but a more comprehensive quantitative evaluation would further strengthen their claims. Cocox, a framework for generating conceptual and counterfactual explanations, is introduced in [34]. The authors propose a two-step process: first, they learn concept prototypes using GANs, and then generate counterfactual explanations by manipulating latent variables within the GAN framework. While the paper primarily focuses on conceptual explanations, the faithfulness of the generated counterfactual explanations is not explicitly discussed or evaluated. The article [40] addresses the challenge of generating semantically consistent visual counterfactual explanations. The authors aim to generate plausible counterfactual images that maintain semantic coherence, ensuring that changes to the image do not introduce unrealistic or incoherent elements. The study presents a novel framework that lever-

ages GANs to generate semantically consistent visual counterfactuals. The authors propose a two-step approach consisting of modification and regularization phases. In the modification phase, they use a conditional GAN to generate counterfactual images by introducing changes to the original image. The GAN is trained to preserve the image semantics while incorporating user-specified changes. The regularization phase involves a semantic consistency loss term that encourages the generated images to maintain semantic coherence throughout the modification process. The authors evaluate their framework using qualitative and quantitative assessments. They compare their method with existing approaches and demonstrate that it produces visually realistic and semantically consistent counterfactual images. They perform user studies to measure the generated counterfactuals' perceived plausibility and semantic coherence, obtaining favorable results. The paper addresses an important aspect of counterfactual explanation generation, emphasizing the need for explanations that align with human perception and understanding. The authors in [41] present an approach that revolutionizes the field of interpretable machine learning. By combining Generative Adversarial Networks (GANs) and StyleSpace analysis, they introduce a method that generates visually captivating explanations for classifier decisions. The authors demonstrate the efficacy of their framework by manipulating the latent space of a GAN to create images that clarify the underlying rationale behind a classifier's output. The disentangled properties of StyleGAN enable the generation of interpretable visual attributes, showcasing the ability of the proposed method to capture essential features driving classifier decisions.

The novelty of this paper lies in its fusion of GANs and StyleSpace analysis to produce explanations that surpass conventional textual justifications. By exploiting the unique characteristics of StyleGAN, the authors unlock the potential to manipulate specific visual attributes within the generator's latent space. This approach allows for the creation of visually intuitive explanations that go beyond traditional methods, ensuring that the generated images are both interpretable and relevant. While the paper's reliance on labeled data and its focus on image-based explanations present limitations. When the classifier exhibits biases or errors, the StylEx may inadvertently captures and amplify these inaccuracies, due to its dependence on the quality of the underlying classifier. Additionally, the performance of StylEx could be impacted while dealing with complex datasets, where attributing changes in classifier decisions to specific visual attributes may be challenging. Moreover, the effectiveness of this method in handling multi-attribute counterfactual ex-

planations decreases.
StylEx's limitations could be overcome through the consideration of multiple enhancements. Firstly, The underlying classifier training process can be made more robust and fair attribute extraction by incorporating techniques for bias detection and mitigation. The classifier model could be less susceptible to inaccuracies if it is regularly audited and updated, particularly when there is biased data. Additionally, researchers could concentrate on improving StylEx to better handle counterfactual explanations that involve multiple attributes. The authors in [42] introduce a compelling method for generating counterfactual explanations. By leveraging generative models and enforcing cycle-consistency, the authors provide a valuable contribution to the field of interpretable machine learning. The paper's comprehensive evaluation, along with its focus on interpretability and fairness, highlights its potential impact in improving transparency and trust in AI systems. However, there are some limitations to consider. When managing very complex generative model latent spaces may be a challenge for the method, to find clear paths for attribute changes. Additionally, accurately training the shift predictor, an important part of the process, can be tricky. These limitations could impact the overall effectiveness. To address limitations, enhance the method's attribute disentanglement by incorporating advanced techniques such as disentangled representation learning. Validate and generalize the proposed approach across diverse datasets and image classification tasks to ensure broader applicability. Conduct thorough experiments to assess the effectiveness of the pipeline in detecting and mitigating bias in image classification systems under various real-world scenarios. An innovative approach is presented in [43] to generate visual counterfactual explanations using diffusion models. The authors propose a method that leverages the power of diffusion models to transform an input image into a counterfactual representation by iteratively updating the pixel values. By incorporating a contrastive loss function, the generated counterfactual explanations highlight the minimal changes required to alter the classification decision of a deep neural network. The paper provides a thorough evaluation of the proposed method, demonstrating its effectiveness in generating interpretable visual explanations and its potential impact on enhancing transparency and interpretability in deep learning systems. Although DVCEs can provide promising insights into image classifier decisions, there are certain limitations that must be taken into consideration. Even with adaptive parameterization, it's still a challenge to produce semantically meaningful changes. DVCEs' effectiveness can be

demonstrated by choosing the right hyperparameters, and the optimal selection may vary across different datasets and classifiers. Furthermore, it might require careful parameter tuning. Further investigation is necessary to determine if DVCEs can be applied to a wider range of classifiers and datasets, and the method's sensitivity to variations of input and model architectures should be thoroughly investigated. To reduce the computational cost of multiple iterations, enhancing computational efficiency can be achieved through optimization and parallelization. Examining alternative denoising methods for insecure models and enhancing approximation techniques to strengthen the theoretical foundations of DVCEs. The authors in [44] present an effective baseline method for generating reverse counterfactual explanations. The simplicity of the proposed method, combined with its competitive performance, makes it a valuable addition to the field of interpretable machine learning. Addressing potential biases and exploring the generalizability of the approach would be valuable directions for future research. This paper provides a solid foundation for generating reverse counterfactual explanations and opens avenues for further advancements in this area. The method discussed in the paper, called Latent-CF, can be utilized effectively for particular types of data, such as images and loan details. The method might not work as smoothly if the datasets have many different features. In simpler terms, it's like a tool that works well for specific datasets, but we are not entirely sure how it handles different tasks. Future research should utilize diverse datasets to address limitations, consisting of diverse table and high-dimensional datasets, and conduct a comprehensive investigation of alternative optimization strategies within the latent space, such as genetic algorithms or Bayesian-driven approaches, to enhance the generalizability and robustness of the proposed Latent-CF method. The paper [45] presents a comprehensive exploration of the application of diffusion models for generating counterfactual explanations. The authors propose an approach that utilizes the power of diffusion models to generate plausible and interpretable counterfactual instances by iteratively updating the input data. The paper provides a thorough analysis of the benefits and limitations of diffusion models in the context of counterfactual explanation generation, highlighting their ability to capture complex data distributions and generate meaningful modifications. The extensive evaluation on various datasets and comparison with existing methods demonstrate the effectiveness and superiority of diffusion models for generating high-quality counterfactual explanations. Even though DiME has been successful, it is necessary to admit some limitations.

The method uses diffusion models that may require substantial processing resources makes the computation costly during inference time, is a significant drawback. The model may be insufficient for applications that need instant interpretation due to the challenges associated with real-time explanations. These limitations need for further research to address computational efficiency and applicability in time-sensitive scenarios. In order to minimize the limitations of DiME, exploring techniques like model parallelism can help improve computational efficiency or reducing significant inference time through algorithmic optimizations. Simultaneously, by investigating transfer learning, it is possible to reduce the dependence on training data or cross-domain adaptation approaches, enabling efficient generation without requiring extensive data access. The goal of these precision enhancements is to improve DiME's efficiency, scalability, and applicability in real-time situations and environments that value privacy.

3. Proposed methodology

The proposed study aims to develop a post-hoc visual explainability method that provides plausible and faithful counterfactual visual explanations (FCVE) that are easy to understand and offer reasoning behind model decisions, reflecting the internal working process of the model. To accomplish this, we build upon a previously developed counterfactual explanation (CFE) model in [36] that identifies counterfactual filters to explain model decisions. It does this by predicting a set of minimum correct (MC) and minimum incorrect (MI) filters. The MC filters are necessary to maintain the prediction of the image to the original inferred class by the classifier. Mathematically, the MC filters can be denoted as

$$F_{MC_i} \in [0,1]^{1\times n}, \tag{1}$$

where n is the number of filters in the top convolution layer of the classifier model. Values of '1' and '0' indicate whether the corresponding filter is to be active or disabled, respectively, to maintain the prediction to the inferred class.

In contrast, the MI filters are needed to alter the classifier's decision to a chosen target class. Mathematically, the MI filters can be denoted as

$$F_{MI_i} \in [\mathbb{R}^+]^{1\times n}. \tag{2}$$

Non-zero indexes in F_{MI_i} correspond to the MI filters, and the values at these indexes indicate the magnitude by which the original filter activations are altered to modify the classifier's decision.

The CFE model operates on the last convolution layer of the classifier because these filters have the most impact and represent more abstract, high-level features, concepts, and even whole objects [46, 47, 48]. In paper [36], it is demonstrated that by enabling, disabling, or modifying these high-level filters in certain ways, it is possible to change the decisions of a pre-trained classifier to either the original inferred class or a chosen alternative class. Importantly, the CFE model probes the internal structure of a deep learning model without altering the input, allowing users to provide faithful explanations aligned with the model's internal decision-making process. Thus, in the study, we rely on these filters as changes to them may produce plausible visual explanations.

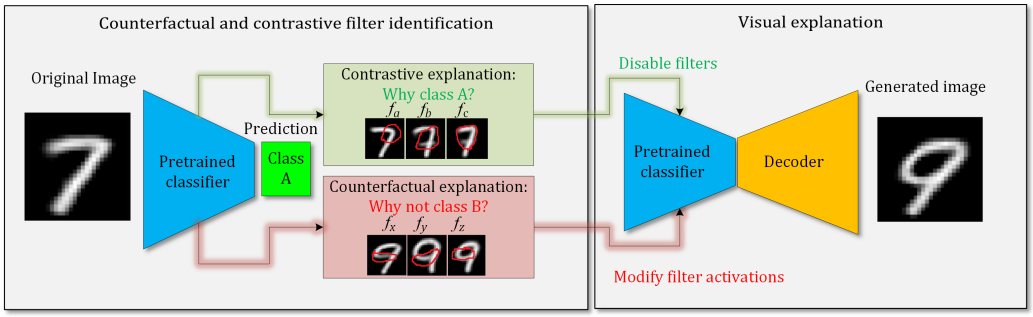

Figure 1: Block diagram of the proposed visual counterfactual explanation model. The proposed method consists of two steps: first is the identification of contrastive and counterfactual filters to explain classifier's decisions, followed by the visualization of these filters by generating images with the modified activations. The decoder is initially trained with all filters intact to recreate the input, so that when the encoder's output is altered using the identified filters, their effect is visualized in the recreated image.

3.1. ***Classifier***

To provide visual counterfactual explanations, we propose a joint CFE and classifier-decoder model. In this model, the pre-trained classifier is the model being analyzed, and the decoder model generates visualizations of the classifier's decisions by modifying the filter activations obtained from the CFE model's counterfactual and contrastive explanations. Fig. 1 presents

a block diagram illustrating the different phases involved in our proposed model.

In the counterfactual and contrastive filter identification phase, we extract the MI and MC filters to provide contrastive and counterfactual explanations of the classifier's decisions. These filters capture the necessary changes to alter the classifier's decision to a target class or maintain the original inferred class. In the visual explanation phase, we use a decoder that takes as input the encoded feature vector generated at the last convolution layer of the classifier and it tries to recreate the input image that was given to the classifier.

The idea behind the decoder model is that it is initially trained to translate the encoded classification features into the respective input while all filters are intact. Once the decoder is trained, we modify the filters using the counterfactual and contrastive explanations produced by the CFE model. This allows us to observe the filter-level changes on the regenerated images, thus generating visual explanations that reflect the alterations made by the counterfactual and contrastive filters.

In the case of Figure 1, to generate explanation for why a classifier classified an input to class A (i.e., 7 in this case) and not to class B, we can select any target counterfactual class. In this case, we selected class 9 as the counterfactual class due to similarity between them. The proposed method identified counterfactual filters that if they were active in the classification of the input 7, then the model would likely classify the input to the target class. In the case of 7, the proposed method identified key filters responsible for turning the decision from 7 to 9. And when the decoder was presented with 7 as input with the modified filters responsible for classifying 7 as 9, the decoder regenerated the 7 as a 9, demonstrating features that needed to be present in the input image to be classified into class 9.

To train the proposed visual counterfactual explanation model, we follow a two-phase approach. In the first phase, we train the decoder model with mean absolute error (MAE) loss while the classifier weights are frozen to reproduce the input image given to the classifier. The loss function is defined as follows:

$$\min_{D} \frac{1}{n} \sum_{i=1}^{n} |x_i - D(C_{conv}(x_i))|_1, \tag{3}$$

where x_i represents the i^{th} input image, $C_{conv}(x_i)$ denotes the encoded feature vector produced by the last convolutional layer of the classifier model C for

the i^{th} input image, and $D(C_{conv}(x_i))$ represents the reconstructed image generated by the decoder model D using the encoded feature vector $C_{conv}(x_i)$. The mean absolute error loss is calculated as the average of the absolute differences between the input image x_i and its corresponding reconstructed image $D(C_{conv}(x_i))$ for all n input images. The training process aims to minimize this MAE, thereby ensuring the decoder effectively reproduces the input images.

In the second phase, we utilize pre-trained CFE model to generate MC and MI filters for the given classifier. The CFE model can be represented as a function that takes an input image x and produces a set of F_{MI_i} and F_{MC_i} filters as output. This can be expressed as:

$$F_{MI_i}, F_{MC_i} = \text{CFE}(x, C, \hat{c}), \tag{4}$$

where x denotes the input image, C represents the pre-trained classifier, $\hat{c}$ is the target class, and CFE is the counterfactual explanation model. The CFE model is responsible for generating the sets of MC and MI filters required to maintain the original classification decision and change it to the target class, respectively, using the following equations

$$F_{MC_i} = \text{ReLU}_t(\text{Sigmoid}(d^n(g_i))), \tag{5}$$

$$F_{MI_i} = \text{ReLU}(d^n(g_i)), \tag{6}$$

where g_i denotes the feature maps obtained after the global average pooling layer of classifier C, d^n represents a dense layer with n units, Sigmoid denotes the sigmoid activation function, and ReLU_t is a thresholded-ReLU layer with a threshold value of $t = 0.5$. The ReLUt layer produces the approximately binarized MC filter map F_{MC_i} by setting all values below the threshold to zero and leaving the other values unchanged, and ReLU denotes the rectified linear unit activation function.

These MC and MI filters are utilized to modify the filter activations of the classifier to observe their impact on the reconstructed images. The process can be described by the following equation:

$$C\big(x, F_{MC_i}, F_{MI_i}\big). \tag{7}$$

In this equation, the input image x, along with the MC and MI filters, are provided as input to the pre-trained classifier C. The classifier then generates an altered feature vector by incorporating the effects of these filters.

3.2. ***Decoder***

We designed an asymmetric encoder-decoder architecture to synthesize counterfactuals visually. The decoder is asynchronous as the encoder and decoder have variable depths (the number of deconvolution and up-sampling layers of the decoder are not equal to convolution and max pooling layers of the encoder model). The decoder has lower depth than the encoder consequently, the decoder can be trained efficiently. We train the decoder model once the encoder model is trained for the prediction of MC and MI filters. The decoder model reconstructs the latent representation provided by the CFE model. The CFE model working as encoder uses pretrained VGG16 which down sizes the input image x to the last layer as a feature vector in the latent space. The CFE model learns the extent of changes to filters in this lower-dimensional space. The decoder model is designed to up-sample these modified lower-dimensional features to higher-dimensional data equal to the dimensions of original input image x. The decoder model takes the modified feature vector as an input and produces an altered output image x' which is the counterfactual of the original input image x.

$$x' = D\Big(C\big(x, F_{MC_i}, F_{MI_i}\big)\Big). \tag{8}$$

The decoder generates the image which reflects the filter-level changes made in the latent space vector as shown in Figure 1. The reconstructed image by the decoder provides visual explanations of the features-modification made by the counterfactual and contrastive filters.This reconstructed image represents a plausible visual explanation that aligns with the internal decision-making process of the model. The procedure describing the overall approach is presented in Algorithm 1.

The proposed approach allows us to gain insights into the influence of specific filters on the model's decision-making process and the generated visual explanations provide a better understanding of how the model arrives at its decisions.

4. Results

This section presents the results and discussion of the proposed FCVE method. For the evaluation of the proposed FCVE method, we used MNIST [49] and Fashion-MNIST (FMNIST) [50] datasets and compared with related

Algorithm 1 Steps to generate counterfactual visual explanations.

Input: Image I, Classifier model C, Counterfactual explanation model CFE, target class $\hat{c}$, train dataset T

Step 1. Train a decoder

procedure TRAINDECODER(C)

Train the decoder on train dataset T using features from classifier C

$\min_D \frac{1}{n} \sum_{i=1}^{n} |x_i - D(C_{conv}(x_i))|_1$

end procedure

Step 2. Generate contrastive and counterfactual explanation using CFE model for input image I

procedure GENERATEEXPLANATION(I, CFE)

$F_{MI_i}, F_{MC_i} = \text{CFE}(I, C, \hat{c})$

end procedure

Step 3. Alter filters in classifier

procedure ALTERFILTERS(C, F_{MI_i}, F_{MC_i})

Generate feature vector g and predicted class c using classifier C:

$g, c = C(I)$

$\hat{c} = h(g \circ F_{MC_i})$ ▷ alter prediction with just MC filters enabled

$\hat{c} = h(g + F_{MI})$ ▷ alter prediction with just updated MI filters

where h represents the classification (fully-connected and softmax) layers of C

end procedure

Step 4. Use trained decoder to generate visual explanation by reconstructing input I with modified classifier

procedure VISUALEXPLANATION(D, C, F_{MI_i}, F_{MC_i})

$$I' = D\Big(C\big(I, F_{MC_i}, F_{MI_i}\big)\Big)$$

end procedure

Output: I' ▷ Reconstructed input image with counterfactual and contrastive features using the modified classifier

counterfactual explanation methods including ExpGAN [51], CEM [25], CVE [30], and C3LT [37]. In section 4.1 and 4.2, we present a visual comparison of these methods on the two datasets, followed by quantitative analysis presented in Section 4.3.

To evaluate faithfulness of the explanations provided by the proposed method, we refer the reader to [36] that used the class recall metric to demonstrate that the identified counterfactual and contrastive filters are faithful to their respective classes. In [36], it was shown that disabling around 31–44 most imported filters of a class (out of 512 total filters) resulted in a significant decrease in the class recall, whereas the overall model accuracy was reduced by just 2%–3%. On the contrary, it was shown that randomly disabling the same number of filters had a negligible effect on class recall. This shows that the counterfactual and contrastive filters predicted by the CFE model represent features exclusive to a particular class and disabling them slightly affects the overall model accuracy while significantly reducing the particular class's recall score, thus demonstrating the faithfulness of the detected filters used in the decision-making process of the classifier. In the proposed work, we mainly focus on the visual aspect of faithful explainable approach.

4.1. Visual results comparison with related methods

Figure 2 represents a comparison of the counterfactual explanation results. The first column shows the query images from MNIST and FMNIST, while the other five columns display the counterfactuals generated by ExpGAN [51], CEM [25], CVE [30], C3LT [37], and our proposed model (FCVE), respectively. Our method generates counterfactuals by manipulating the internal activations of the model, resulting in counterfactuals that are more meaningful and realistic compared to other methods. We ensured that the source and target classes were selected to maintain the counterfactual proximity property.

Among the baseline models, the results of C3LT are somewhat interpretable but mainly unrealistic. The counterfactuals obtained from C3LT are adversarial to the target classes (e.g., generating 8 and 9 from 3 and 4, respectively, and a shirt from a coat). The counterfactuals obtained from ExpGAN are not smooth (e.g., 9, 6, and pullover). The counterfactuals from CEM and CVE are unrecognizable (e.g., 9 and 8) or mostly unchanged (e.g., 6, short, boot, and pullover). In contrast, the counterfactuals generated by our method are easily recognizable and more realistic.

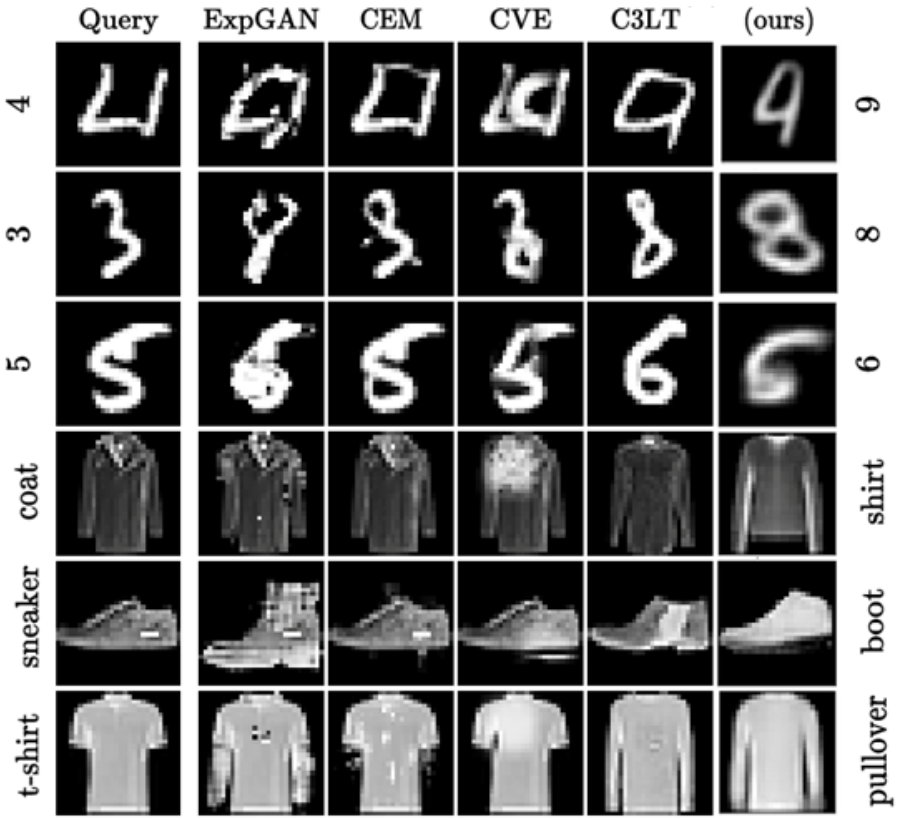

Figure 2: Visual comparison of counterfactual explanation methods. The first column shows the query images from MNIST and FMNIST, while the other five columns display the counterfactuals generated by ExpGAN [51], CEM [25], CVE [30], C3LT [37], and our proposed model (FCVE), respectively. The proposed method generates counterfactuals by manipulating the internal activations of the model, resulting in counterfactuals that are more meaningful and realistic compared to other methods.

4.2. Qualitative analysis of proposed method

This section presents an additional qualitative analysis of the proposed FCVE methods in terms of generating plausible visual counterfactuals for MNIST and FMNIST datasets.

4.2.1. MNIST counterfactuals

Figure 3 displays the counterfactuals generated for the digit seven as the source class and the digit nine as the target class. Despite the non-identical writing styles of the input images for the same digit, our method successfully generates plausible counterfactuals. This ability to generate counterfactuals indicates that the model has learned the underlying data patterns and can generalize well.

The first three input images (1st row) of the digit seven vary in writing style compared to the last three images, which include an extra intersection line. While humans can easily differentiate between these variations of the digit seven, it can be a challenging task for an algorithm to identify such subtle changes.

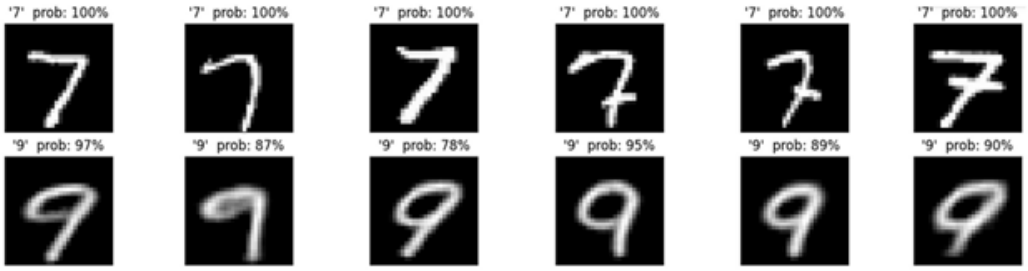

Figure 3: Plausible counterfactuals generated for digit seven as a source class and digit nine as target class. The proposed method finds the minimal changes to neuron activations such that the input of one class is transformed into another.

Figure 4 displays the counterfactuals generated for randomly selected source and target classes of the MNIST dataset. The input images in the first row (i.e., 9, 4, 4, 5, 1, and 6) are chosen randomly, while the images in the second row (i.e., 8, 9, 9, 6, 0, and 0) represent the counterfactuals generated by our model. Our model aims to generate counterfactuals by adding or subtracting features from the original input image. For example, the first counterfactual (8) is obtained by adding a line to the input image (9). Similarly, the counterfactuals of 0, 9, and 6 are generated by the same principle from the input images 1, 4, and 5, respectively. Additionally, a counterfactual of 0 is obtained by removing a portion of 6.

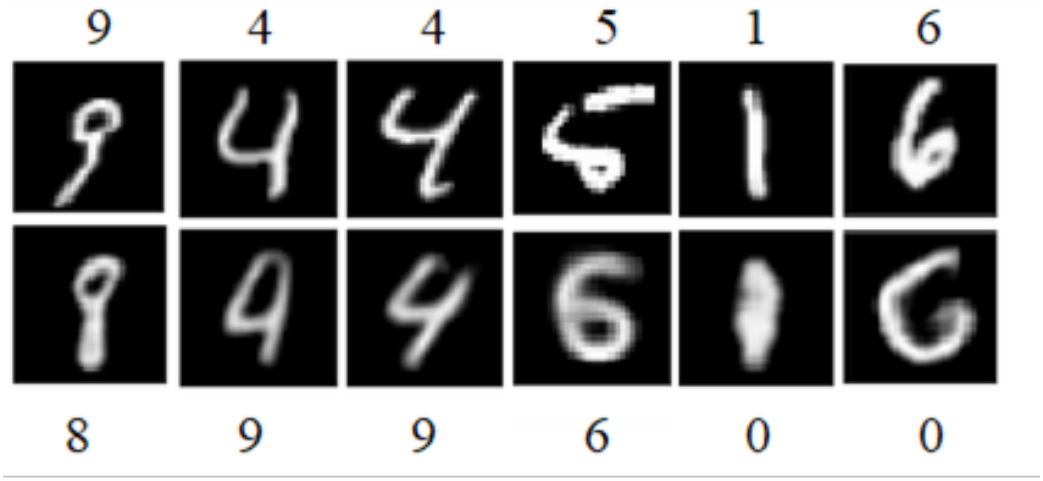

Figure 4: Counterfactuals generated for random source and target classes of MNIST dataset. Similar to Fig. 3, the proposed method finds the minimal changes to neuron activations such that the input of one class is transformed into another.

4.2.2. FMNIST counterfactuals

Figure 5 displays the results of the counterfactual visual explanations obtained using the proposed method. In this analysis, the source class is "Pullover," represented by the images in the first row. The goal is to transform these pullover images into counterfactual representations of the target classes, which are "Dress" and "Coat" displayed in the second and third rows, respectively. These counterfactuals are generated by the FCVE model, utilizing the source class image as a starting point.

It can be seen that the proposed method successfully modifies the source class images to generate plausible visual counterfactuals that accurately represent the target classes. The generated counterfactuals exhibit visual characteristics and features associated with the respective target classes, showcasing the effectiveness of the FCVE model in capturing and manipulating the underlying data patterns.

Figures 6 and 7 showcase additional examples of counterfactual image generation from visually identical and non-identical classes, respectively. In Figure 6, the first row comprises actual images of t-shirts from the FMNIST dataset, while the second row displays the counterfactuals generated by our proposed model, for the target class of "Pullover". The source class (t-

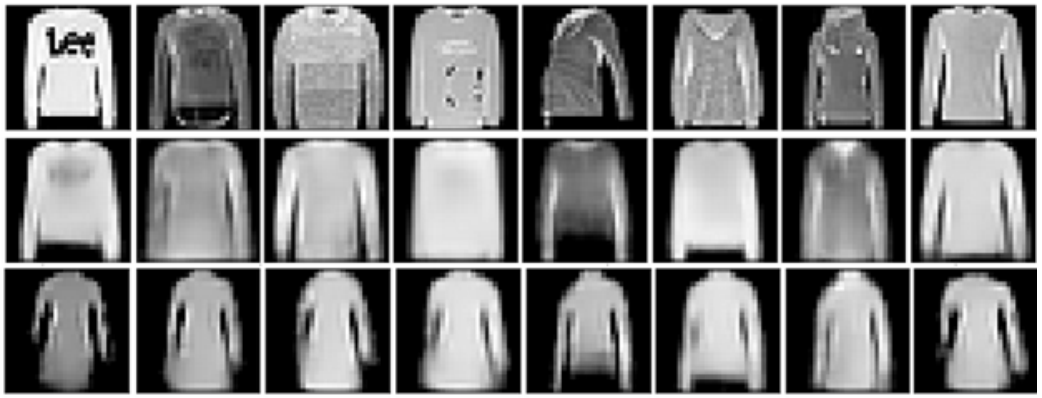

Figure 5: Plausible counterfactuals generated for the FMNIST dataset. The first row is the source class "Pullover". Second and third rows are target classes of "Dress" and "Coat" into which the source image is transformed into by altering the filter activations.

shirts) and target class (pullover) belong to visually similar categories, and the proposed method effectively transforms the t-shirts into pullovers with distinctive features, such as long sleeves. It is worth noting that the model accurately captures the shape of the target class while sacrificing some finer details, such as patterns on the t-shirts. This suggests that the shape is a more crucial feature than the specific patterns when differentiating between these classes.

Similarly, in figure 7, the counterfactuals generated by our proposed model are presented, focusing on visually diverse source and target classes. In this case, the source class is "Trouser", while the target class is "Shirt". These classes exhibit noticeable visual differences in terms of shape, texture, and overall appearance. Despite the visual disparity between the source and target classes, our proposed model consistently produces realistic target class images by transforming the source images. This demonstrates the effectiveness of our approach in generating accurate and visually coherent counterfactual representations.

4.3. Quantitative comparison

This section provides a quantitative analysis of the proposed FCVE method and compares it with existing methods in terms of the proximity measure and Fréchet Inception Distance (FID).

Figure 6: Counterfactual generation from visually identical classes of "T-shirt" (source class, row 1) and "Pullover" (target class, row 2) in FMNIST. The proposed method effectively transforms the t-shirts into pullovers with distinctive features, such as long sleeves.

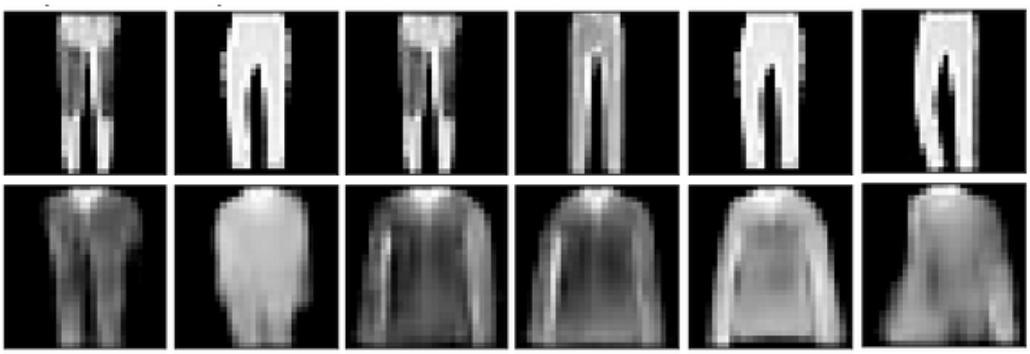

Figure 7: Counterfactual generation from visually non-identical classes of "Trouser" (source class, row 1) and "Shirt" (target class, row 2) in FMNIST. Despite the differences between the source and target classes, the proposed method produces realistic target class images by transforming the source images.

4.3.1. Proximity

Proximity property explanations the counterfactuals, meaning Faithful to the original instance. The generated counterfactual explanations are considered the best as they are closest to the original instance. Proximity is the mathematical formula to quantify the closeness of two instances (query image and counterfactual) using L1 distance. Satisfying, this minimal feature change property to generate counterfactual examples, the proximity metric

can be defined mathematically in terms of distance function as,

$$\text{proximity} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{dsict}(x_i - x_i')}{CHW} \tag{9}$$

where x_i and x_i' represent the ith query image and counterfactual example from the set being evaluated, and C, W and H are the channels count, width, and height of the query image, respectively. The Lower levels of proximity suggest methods that produce counterfactuals that are closer to the original data points.

4.3.2. Plausibility

Plausibility property depicts the counterfactual explanations are realistic, feature values are coherent to the domain set. The feature values of counterfactuals should not be an outlier in consideration with domain set. Enhancing trust in the explanation is facilitated by plausibility. The approach we are using to check plausibility is FID score calculation. It is feature-wise subtraction of the query images and their respective counterfactuals. Plausibility contributes to the robustness and stability of counterfactual explanations. The formula for calculating the FID is as follows:

$$FID = |\mu - \mu'|^2 + \text{Tr}\left(x + x' - 2\sqrt{x.x'}\right) \tag{10}$$

Where $\mu, \mu', x, x', \text{Tr}(\cdot), |\cdot|^2$ denotes mean feature vectors of the real and generated image distributions, the covariance matrices of the real and generated image distributions, the trace of a matrix and the squared Euclidean norm. the FID metric measures the similarity between the generated images and real images, focusing on the distribution of features. A lower FID score indicates better-quality images and greater realism.

Table 1 presents a comparison of the counterfactual explanation methods based on both the proximity and FID metrics. These metrics were obtained from various baseline models, including ExpGAN [51], CEM [25], CVE [30], and C3LT [37]. From the table, it is evident that the proposed FCVE method achieves a significantly lower FID score compared to the compared methods. This result indicates that the proposed method generates high-quality counterfactuals that closely resemble the real data, demonstrating its effectiveness in generating realistic and meaningful counterfactual explanations.

Table 1: Comparison of counterfactual explanation methods on MNIST and FMNIST datasets based on proximity and FID scores.

Method	ExpGAN [51]		CEM [25]		CVE [30]		C3LT [37]		FCVE (our)	
	MNIST	FMNIST	MNIST	FMNIST	MNIST	FMNIST	MNIST	FMNIST	MNIST	FMNIST
Proximity	0.074	0.135	0.016	0.013	0.055	0.054	0.072	0.116	0.098	0.198
FID	41.12	76.52	50.03	96.87	47.53	83.77	22.83	62.31	0.50	2.02

5. Conclusion

The development of explainable AI techniques plays a crucial role in addressing the transparency and interpretability challenges associated with deep learning models in computer vision. While significant progress has been made, existing methods still face limitations in providing convincing explanations that are easily understandable to non-experts and accurately capture the intrinsic decision-making processes of the models.

To overcome these challenges, we have proposed a counterfactual explanation (CE) model that aims to strike a balance between plausibility and faithfulness. Our model generates visual explanations that are not only easy to comprehend but also faithfully represent the model's internal decision-making process. Importantly, these explanations are generated by making minimal changes to the original images, without altering the pixel data.

Instead of relying solely on pixel-level manipulations, our approach identifies and leverages the internal concepts and filters learned by the model. By understanding and manipulating these internal representations, our model produces plausible counterfactual explanations that reflect the model's underlying decision-making process, making the provided explanation faithful to the model.

Through qualitative and quantitative analysis, we have demonstrated the effectiveness of our proposed FCVE method. The qualitative analysis highlights the close resemblance between the generated counterfactuals and the original data instances, indicating the high quality of the explanations. Furthermore, the quantitative analysis using Fréchet Inception Distance (FID) scores confirms that our method outperforms the baseline models in generating realistic and diverse counterfactuals.

Future research directions could focus on extending the proposed method to other domains and exploring additional evaluation metrics to further validate the effectiveness of counterfactual explanations in different contexts.

References

[1] M. Z. Uddin, M. M. Hassan, Activity recognition for cognitive assistance using body sensors data and deep convolutional neural network, IEEE Sensors Journal 19 (19) (2018) 8413–8419.

[2] F. Schiliro, A. Beheshti, N. Moustafa, A novel cognitive computing technique using convolutional networks for automating the criminal investigation process in policing, in: Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 1, Springer, 2021, pp. 528–539.

[3] Y. Abakarim, M. Lahby, A. Attioui, Towards an efficient real-time approach to loan credit approval using deep learning, in: 2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC), IEEE, 2018, pp. 306–313.

[4] O. Alfarraj, Internet of things with bio-inspired co-evolutionary deep-convolution neural-network approach for detecting road cracks in smart transportation, Neural Computing and Applications (2020) 1–16.

[5] X. Zhang, Y. Qiao, F. Meng, C. Fan, M. Zhang, Identification of maize leaf diseases using improved deep convolutional neural networks, Ieee Access 6 (2018) 30370–30377.

[6] A. Chouiekh, E. H. I. E. Haj, Convnets for fraud detection analysis, Procedia Computer Science 127 (2018) 133–138.

[7] N. Seijdel, N. Tsakmakidis, E. H. De Haan, S. M. Bohte, H. S. Scholte, Depth in convolutional neural networks solves scene segmentation, PLoS computational biology 16 (7) (2020) e1008022.

[8] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE transactions on neural networks and learning systems (2021).

[9] R. Nandhini Abirami, P. Durai Raj Vincent, K. Srinivasan, U. Tariq, C.-Y. Chang, Deep cnn and deep gan in computational visual perception-driven image analysis, Complexity 2021 (2021) 1–30.

[10] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115.

[11] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (xai): Toward medical xai, IEEE Transactions on Neural Networks and Learning Systems (2020).

[12] A. Holzinger, C. Biemann, C. S. Pattichis, D. B. Kell, What do we need to build explainable ai systems for the medical domain?, arXiv preprint arXiv:1712.09923 (2017).

[13] É. Zablocki, H. Ben-Younes, P. Pérez, M. Cord, Explainability of vision-based autonomous driving systems: Review and challenges, arXiv preprint arXiv:2101.05307 (2021).

[14] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, arXiv preprint arXiv:1708.08296 (2017).

[15] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, Explainable ai: the new 42?, in: International Cross-domain Conference for Machine Learning and Knowledge Extraction, Springer, 2018, pp. 295–303.

[16] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (5) (2019) 206–215.

[17] Q. Zhang, W. Wang, S.-C. Zhu, Examining cnn representations with respect to dataset bias, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[18] N. Akhtar, A. Mian, Threat of adversarial attacks on deep learning in computer vision: A survey, IEEE Access 6 (2018) 14410–14430.

[19] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, Robust physical-world attacks on deep

learning visual classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1625–1634.

[20] S. Thys, W. Van Ranst, T. Goedemé, Fooling automated surveillance cameras: adversarial patches to attack person detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[21] A. Ghorbani, J. Zou, Neuron shapley: Discovering the responsible neurons, arXiv preprint arXiv:2002.09815 (2020).

[22] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.

[23] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, Communications of the ACM 63 (1) (2019) 68–77.

[24] D. Vale, A. El-Sharif, M. Ali, Explainable artificial intelligence (xai) post-hoc explainability methods: Risks and limitations in non-discrimination law, AI and Ethics (2022) 1–12.

[25] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, P. Das, Explanations based on the missing: Towards contrastive explanations with pertinent negatives, in: Advances in Neural Information Processing Systems, 2018, pp. 592–603.

[26] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[27] A.-H. Karimi, G. Barthe, B. Balle, I. Valera, Model-agnostic counterfactual explanations for consequential decisions, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 895–905.

[28] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, P. Flach, Face: feasible and actionable counterfactual explanations, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 344–350.

[29] A. Van Looveren, J. Klaise, Interpretable counterfactual explanations guided by prototypes, in: Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21, Springer, 2021, pp. 650–665.

[30] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, Counterfactual visual explanations, Vol. 97 of Proceedings of Machine Learning Research, PMLR, Long Beach, California, USA, 2019, pp. 2376–2384.

[31] L. A. Hendricks, R. Hu, T. Darrell, Z. Akata, Grounding visual explanations, in: European Conference on Computer Vision, Springer, 2018, pp. 269–286.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[33] R. Luss, P.-Y. Chen, A. Dhurandhar, P. Sattigeri, Y. Zhang, K. Shanmugam, C.-C. Tu, Generating contrastive explanations with monotonic attribute functions, arXiv preprint arXiv:1905.12698 (2019).

[34] A. R. Akula, S. Wang, S.-C. Zhu, Cocox: Generating conceptual and counterfactual explanations via fault-lines., in: AAAI, 2020, pp. 2594–2601.

[35] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), Vol. 80 of Proceedings of Machine Learning Research, PMLR, Stockholmsmässan, Stockholm Sweden, 2018, pp. 2668–2677.

[36] S. A. Tariq, T. Zia, M. Ghafoor, Towards counterfactual and contrastive explainability and transparency of dcnn image classifiers, Knowledge-Based Systems (2022). doi:https://doi.org/10.1016/j.knosys.2022.109901.

[37] S. Khorram, L. Fuxin, Cycle-consistent counterfactuals by latent transformations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10203–10212.

[38] C.-H. Chang, E. Creager, A. Goldenberg, D. Duvenaud, Explaining image classifiers by counterfactual generation, arXiv preprint arXiv:1807.08024 (2018).

[39] E. M. Kenny, M. T. Keane, On generating plausible counterfactual and semi-factual explanations for deep learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 11575–11585.

[40] S. Vandenhende, D. Mahajan, F. Radenovic, D. Ghadiyaram, Making heads or tails: Towards semantically consistent visual counterfactuals, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII, Springer, 2022, pp. 261–279.

[41] O. Lang, Y. Gandelsman, M. Yarom, Y. Wald, G. Elidan, A. Hassidim, W. T. Freeman, P. Isola, A. Globerson, M. Irani, et al., Explaining in style: Training a gan to explain a classifier in stylespace, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 693–702.

[42] K. Alipour, A. Lahiri, E. Adeli, B. Salimi, M. Pazzani, Explaining image classifiers using contrastive counterfactuals in generative latent spaces, arXiv preprint arXiv:2206.05257 (2022).

[43] M. Augustin, V. Boreiko, F. Croce, M. Hein, Diffusion visual counterfactual explanations, arXiv preprint arXiv:2210.11841 (2022).

[44] R. Balasubramanian, S. Sharpe, B. Barr, J. Wittenbach, C. B. Bruss, Latent-cf: a simple baseline for reverse counterfactual explanations, arXiv preprint arXiv:2012.09301 (2020).

[45] G. Jeanneret, L. Simon, F. Jurie, Diffusion models for counterfactual explanations, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 858–876.

[46] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, A. Torralba, Understanding the role of individual units in a deep neural network, Proceedings of the National Academy of Sciences (2020).

[47] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene cnns, arXiv preprint arXiv:1412.6856 (2014).

[48] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6541–6549.

[49] Y. LeCun, The mnist database of handwritten digits, http://yann. lecun. com/exdb/mnist/ (1998).

[50] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, arXiv preprint arXiv:1708.07747 (2017).

[51] P. Samangouei, A. Saeedi, L. Nakagawa, N. Silberman, Explaingan: Model explanation via decision boundary crossing transformations, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 666–681.