Looking in the Mirror: A Faithful Counterfactual Explanation Method for Interpreting Deep Image Classification Models

Supplementary Material

7. Mirror CFE for multi-class classification network

$\mathbf{q}_{(cf,t)} > \mathbf{q}_{(cf,s)}$ for $k = 0.5 + \epsilon$ is only true for a binary classification setting. It may not hold in a multi-class classification setting due to interference from the other classes in the softmax probability calculation. While the exact $\mathbf{z}_p$ and $\mathbf{z}_r$ positions are not directly computable in the multi-class classification setting, we use Eq. (1) obtained $\mathbf{z}_r$ to initialize the L-BFGS [11] algorithm with the optimization goal that a new $\mathbf{z}_{r'}$ location will satisfy $\min_{\mathbf{z}_r} \|\mathbf{l}_r - \mathbf{l}_{r'}\|$,. This process is depicted in Fig. 5. The reflection point's multi-class logits are calculated as $\mathbf{l}_r = \mathbf{W}^\top \mathbf{z}_r + \mathbf{b}$ and $\mathbf{l}_{r'}$ a hypothetical reflection point's logit that flips the confidence between the source and target classes: $\mathbf{l}_{(r',s)} = \mathbf{l}_{(s,t)}$ and $\mathbf{l}_{(r',t)} = \mathbf{l}_{(s,s)}$, the rest class logits remain unchanged. For the KFE points on the line from $\mathbf{z}_s$ to $\mathbf{z}_{r'}$ the computation of their location is the fraction of the travel indicated by k. For example, we calculate the new projection point $\mathbf{z}_{p'} = \frac{\mathbf{z}_s + \mathbf{z}_{r'}}{2}$. Note that $\mathbf{z}_{p'}$ can achieve $\mathbf{p}_{(p',s)} \approx \mathbf{p}_{(p',t)}$, but $\mathbf{p}_{(p',s)}, \mathbf{p}_{(p',t)} \neq 0.5$ due to the presence of other classes. Finally, we found the L-BFGS can find reasonably accurate logits that reflect the intended probabilities. In Fig. 6, we show our generated multi-class classification reflection points in $\mathcal{Z}$ by using T-SNE visualization. It can be seen that the generated reflection points are correctly located in the T-SNE computed clusters.

Our Mirror-CFE supports multi-class training, but we followed the binary settings for fair comparison with baselines. Multi-class results in Table 2 are similar to the pairwise counterparts. CelebA-HQ has binary attributes, hence there is no difference using multi-class.

		L1↓	LPIPS↓	FID ↓	D.Val.↑	Val.↑
MNIST	Pairwise	**0.26**	**0.33**	**3.20**	**0.99**	**1.0**
	Multiclass	0.27	0.34	6.94	**0.99**	**1.0**
FMNIST	Pairwise	0.25	0.21	**4.40**	0.96	**0.99**
	Multiclass	**0.20**	**0.19**	5.83	**0.97**	0.98

Table 2. Comparison between Multi-class and pairwise settings of our Mirorr-CFE on MNIST and Fashion-MNIST.

8. Computing KFE encoder feature from latent space encoding

Given a KFE latent code $\mathbf{z}_k$ can be computed by using Eq. (1) for any k from $\mathbf{z}_s$, we show how $\mathbf{f}_k^l$ is computed from $\mathbf{z}_k$:

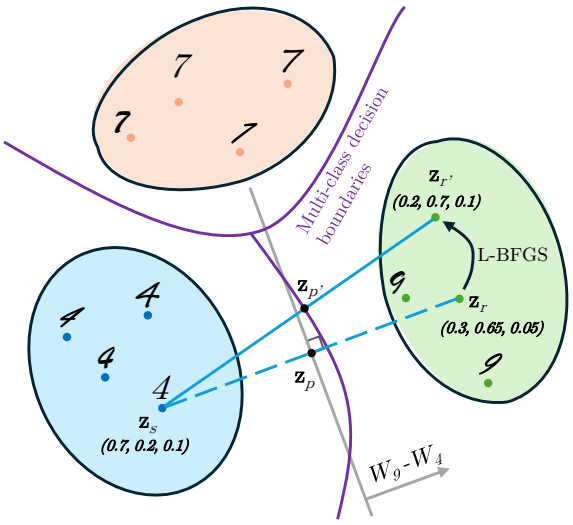

Figure 5. Illustration of the multi-class reflection point calculation that depicts a 3-class toy problem. The introduction of class 7 led to non-linear classification decision boundaries such as shown in the purple color. The source (class 4) and target (class 9) computed binary classification boundary will unlikely to provide a reflection point that flips the class confidence between class 4 and 9. Therefore, we use L-BFGS to estimate a location $\mathbf{z}_{r'}$ which flips the confidence. The confidence are noted as, *e.g.*, (0.2, 0.7, 0.1) for classes 4, 9, and 7 respectively.

$$\mathbf{z}_s = \text{GAP}(\mathbf{f}_k^i) = \sum_{h,w \in H_l, W_l} \frac{\mathbf{f}_s^l}{H_l \times W_l}. \tag{15}$$

Assuming $\mathbf{z}_k = \sum_{h,w \in H_l, W_l} \frac{\mathbf{f}_k^l}{H_l \times W_l}$ was also computed from the GAP, and using Eq. 1 to calculate $\mathbf{z}_k$:

$$\mathbf{z}_k = \mathbf{z}_s - 2k(\mathbf{W}_m^\top \mathbf{z}_s + \mathbf{b}_m)\hat{\mathbf{W}}_m. \tag{16}$$

Let $\mathbf{z}_\Delta = -2k(\mathbf{W}_m^\top \mathbf{z}_s + \mathbf{b}_m)\hat{\mathbf{W}}_m$, we show the equality:

$$\begin{aligned} \sum_{h,w \in H_l, W_l} \frac{\mathbf{f}_k^l}{H_l \times W_l} &= \sum_{h,w \in H_l, W_l} \frac{\mathbf{f}_s^l}{H_l \times W_l} + \mathbf{z}_\Delta, \\ &= \sum_{h,w \in H_l, W_l} \frac{\mathbf{f}_s^l + \mathbf{z}_\Delta}{H_l \times W_l} \end{aligned} \tag{17}$$

hence, $\mathbf{f}_k^l = \mathbf{f}_s^l + \mathbf{z}_\Delta$ is a trivial solution that assumes all elements in $\mathbf{f}_k^l$ take the same rate of travel from $\mathbf{f}_s^l$.

For the multi-class classification scenario, the calculation is slightly different. We first follow Sec. 7 to find the updated position of the reflection point $\mathbf{z}_{r'}$. Then we compute $\mathbf{z}_\Delta$ differently as $\mathbf{z}_\Delta = k(\mathbf{z}_r - \mathbf{z}_s)$ and finally $\mathbf{f}_k^l = \mathbf{f}_s^l + k(\mathbf{z}_r - \mathbf{z}_s)$.

9. Quantitative Analysis of CelebA-HQ

The quantitative comparison between C3LT and Mirror-CFE is in Table 3 for CelebA. In general, both our 1st CFE and mirror CFE $(k-1)$ show significant improvement in L1, LPIPS, FID, and denoised validity in comparison to C3LT. For validity, Mirror-CFE obtained 0.94, and C3LT achieved 1. We note that the Mirror-CFE model was trained with 30 epochs, while 200 epochs were used for C3LT.

Method	L1↓	LPIPS↓	FID ↓	D.Val.↑	Val.↑
C3LT	0.21	46.41	35.28	0.81	**1.0**
Our (1st CFE)	**0.08**	**0.051**	**8.78**	0.87	0.94
Our ($k=1$)	**0.08**	0.053	8.97	**0.89**	0.94

Table 3. Quantitative analysis of CelebA-HQ.

10. Description of the animation files

10.1. Folder structure

We provide a collection of CFE transition animation videos, organized into dataset folders containing subfolders for specific class pairs. In the MNIST dataset, sub-folders 38, 49, and 56 represent the class pairs 3 vs. 8, 4 vs. 9, and 5 vs. 6, respectively. For the F-MNIST dataset, sub-folders 02, 46, and 79 correspond to the class pairs T-shirt vs. Pullover, Coat vs. Shirt, and Sneakers vs. Boots. For the B-MNIST dataset, subfolder 24 represents the class pair Erythroblast vs. Lymphocyte. For CelebA-HQ, subfolder 01 represents the attribute pair mouth closed (0) vs. slightly mouth open (1). Each subfolder includes an MP4 file demonstrating the CFE transition for the examples shown in Fig. 3 or Fig. 4.

10.2. File naming

The file naming convention for these animations is *fig3_rowN_A_(GT_B)_to_C_{correct/error}.mp4*, and the naming convention is:

- rowN is the row number a sample was placed in Fig. 3, or any combination of top/bottom and left/right for samples in Fig. 4.
- A is the model-predicted label, we use it as the source class for the CFE generation. A is not necessarily equal to B because a model may give error prediction.
- B is the ground truth (GT) label for reference.
- C is the target class label for the CFE generation.
- 'correct' means this case was predicted correctly, *i.e.*, A = B, or 'error' otherwise.

	MNIST	F-MNIST	B-MNIST
$\mathcal{L}_{\text{fea}}$	0.012	0.013	0.085
$\mathcal{L}_{\text{conf\_L1}}$	0.014	0.024	0.015

Table 4. The feature reconstruction (L1 distance) and KLD measurements of faithfulness to the anticipated CFE feature and target confidence for Mirror-CFE. In brief, Mirror-CFE achieves on average $< 3\%$ confidence difference to intended target confidence.

	L1↓	LPIPS↓	FID ↓	D.Val.↑	Val.↑
Skip-Connection	**0.07**	**0.02**	**4.20**	0.03	0.03
Ours (no-CSP)	0.14	0.19	40.12	0.78	0.78
Ours (full-SSC)	0.08	0.05	8.97	**0.89**	**0.94**

Table 5. Ablation study of SSC and CSP on CelebA-HQ. The 'Skip-connection' setting denotes the 'no-SPE' variant, where $\mathbf{u}_k^i = \mathbf{f}_s^i$ as defined in Eq. 14.

$\mathcal{L}_{\text{tri}}$	k	L1↓	LPIPS↓	FID ↓	D.Val.↑	Val.↑
$\alpha = 0$	$\neq 1$	0.126	0.104	2.302	0.989	0.991
$\alpha = 0.2$	$\neq 1$	0.127	0.109	2.801	0.990	0.989
$\alpha = 0.4$	$\neq 1$	0.128	0.107	3.001	**0.992**	0.990
$\alpha = 0.6$	$\neq 1$	0.116	0.102	**2.141**	0.989	0.987
✗	$\neq 1$	**0.115**	**0.099**	2.469	0.996	**0.994**
$\alpha = 0$	1	**0.185**	**0.190**	**3.364**	0.955	0.905
$\alpha = 0.2$	1	0.251	0.217	4.402	**0.960**	0.949
$\alpha = 0.4$	1	0.260	0.224	4.510	0.915	0.965
$\alpha = 0.6$	1	0.256	0.221	4.262	0.906	0.956
✗	1	0.248	0.215	4.227	0.921	**0.971**

Table 6. Ablation of $\mathcal{L}_{\text{tri}}$ loss on F-MNIST dataset, separately computed for the CFE computed from the reflection point ($k = 1$) and the rests ($k \neq 1$) (1st CFE to any CFEs with $k < 1$).

We also show additional examples in a folder called "additional samples". The naming convention is similar but changes the 'fig3_rowN' prefix to 'noN' indicating a unique file identifier (otherwise the file names could be duplicated).

10.3. Video content

Each video is structured into three sections per time frame.

The **top section** displays the Source image ($\mathbf{x}_s$) alongside the KFE image ($\mathbf{x}_k$). The displayed label of source image is 'Source($\arg\max \mathbf{p}_s$)' and the KFE image label is one of 'SFE/CFE/Reflection($\arg\max \mathbf{p}_k$)'.

The **middle section** visualizes the transition of the image in the latent feature space, moving from the source latent point ($\mathbf{z}_s$) to the reflection point ($\mathbf{z}_r$). The first significant transition point to obtain CFE image $\mathbf{x}_{cf}$ is highlighted in this section as the "1st CFE point". The 1st CFE means the transition first finds an image where the model predicts the image as a target sample. Additionally, the intended and predicted confidence levels for both the source and target classes are displayed as text '(S:, T:)' in orange and blue, respectively.

The **bottom section** presents a confidence plot for

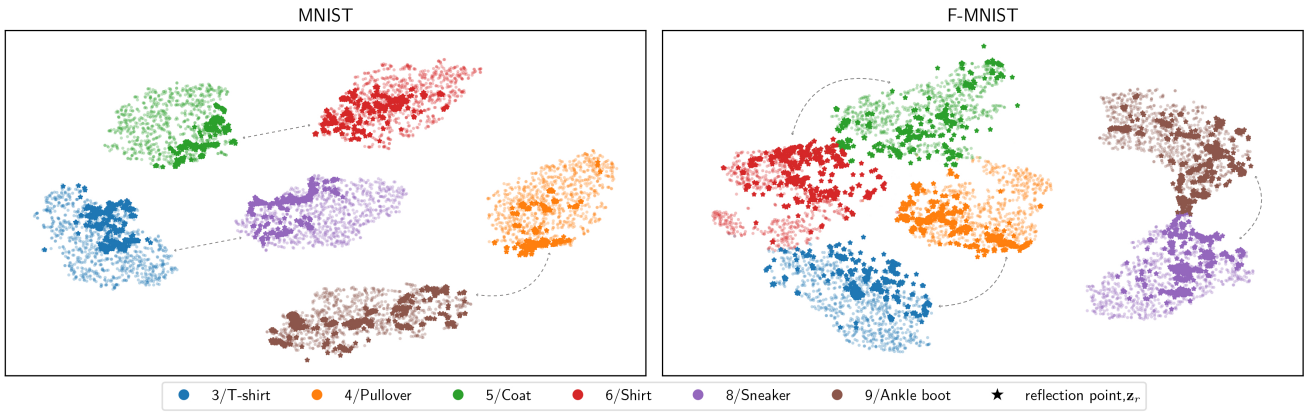

Figure 6. T-SNE visualization of MNIST (left) and F-MNIST (right). Each cluster represents a specific class, with stars marking our computed reflection points ($\mathbf{z}_{r'}$) in the latent feature space. The dashed arrow represents the class pair of source and target class in each dataset. The reflection points are well located with each class cluster, demonstrating their realism within the latent feature space.

Figure 7. Ablation of the SSC module on CelebA-HQ.

the target class. This plot also includes the difference between the source image and the KFE image (*i.e.*, $\frac{1}{C\times H\times W}\sum|\mathbf{x_k}-\mathbf{x_s}|$) at each step k (green line). This plot provides a comprehensive visualization of the transformation process across different datasets and class pairs, showing the k where the most image content change happens and how it affects the classifier's prediction (in confidence).

10.4. Additional samples

We provide additional examples to showcase the efficacy of our method. Across the MNIST dataset, all examples exhibit a smooth and meaningful transition from the source label to the target label, demonstrating the effectiveness of Mirror CFE. In F-MNIST, the transitions vary in magnitude depending on the class pair. For instance, the changes between shirt and coat[1] are minimal, while the transformations between sneaker and ankle boot[2] are considerably more pronounced. The transitions for the T-shirt and pullover[3] class pair consistently occur in the sleeve region, aligning with human expectations. In B-MNIST[4], we observe that an erythroblast nucleus is generally larger than a lymphocyte. On the other hand, lymphocyte has a smaller amount of cytoplasm than erythroblast. This distinction is also observed in the transition process for this class pair in our examples, further validating Mirror CFE's capability to visualize meaningful class-wise differences.

10.5. 'Error' cases

For the error cases, the source image was not predicted as its GT source label ($\arg\max p_s \neq s$). This is an instinct learning problem of the classifier, not because of Mirror CFE. In such instances, Mirror CFE presents a unique feature to analyze how to rectify the error cases, which are unavailable from the compared methods.

When the source image is misclassified we take the predicted class as the source label and set t as the ground truth class. For example, we present an error case[5] in the MNIST dataset involving the class pair 4 and 9.Here, a source image of the digit 4 is misclassified as digit 9 and we construct the CFE from 9 to 4, highlighting the modifications necessary for it to be classified as 4. Specifically, we observe the top portion of the digit 4 bends upwards when k is nearly 0.8, which dramatically increases its confidence as being correctly recognized as a 4 by the classifier.

Another example[6] can be observed in the F-MNIST dataset, involving the class pair 'Shirt' and 'Coat'. In this case, a coat is misclassified as a shirt. Notably, many shirts in F-MNIST feature checkered patterns, which are also present in this specific misclassified sample. Hence, the classifier was making a reasonable assumption that it is a shirt. Our method 'cleaned' the checkered patterns in the image, showing users what it takes to make the image

[1] fmnist/46/additional samples
[2] fmnist/79/additional samples
[3] fmnist/02/additional samples
[4] bmnist/24/additional samples
[5] mnist/49/additional samples/no31_9(GT_4)_to_4_error.mp4
[6] fmnist/46/additional samples/no5_Shirt(GT_Coat)_to_Coat_error.mp4

correctly classified as a coat, giving users a more intuitive understanding of the class difference, specifically illustrated in this sample.

11. Limitations

Mirror-CFE requires the existence of a Global Average Pooling (GAP) layer in the classifier's architecture to facilitate the functionality of the CAM-guided spatial prior (CSP) module. We choose Class Activation Maps (CAM) because CAM is a feed-forward computation that can be computed in the same feed-forward iteration–much faster than gradient-based and sampling-based attention methods. We acknowledge that GAP-enabled models are rare in Transformers, nor do they provide meaningful CAMs off-the-shelf; hence, the usage of Mirror-CFE would require heuristic implementation to estimate the network attention, such as modeling the Transformer attention via [1, 14].

Symbol	Meaning/Definition
$\mathcal{X}, \mathcal{C}, \mathcal{Z}, \mathcal{I}$	These represent the image set, label set, latent feature space, and image (pixel) space, respectively.
$\|\mathcal{C}\|$ (scalar)	The size of the label set.
s, t (scalars)	Source and target classes.
$\mathbf{x} \in \mathbb{R}^{C \times H \times W}, y$ (scalar)	An image $\mathbf{x}$ associated with the class label y.
$F(\mathbf{x})$	The classifier model.
$\mathbf{z} \in \mathbb{R}^N$	Latent feature representation $\mathbf{z}$ with the size N (*e.g.*, 2048) computed as $\mathbf{z} = F(\mathbf{x})$.
C, H, W	The number of input channels and spatial dimensions of the input image.
$\mathbf{W} \in \mathbb{R}^{N \times \|\mathcal{C}\|}, \mathbf{b} \in \mathbb{R}^{\|\mathcal{C}\|}$	Weight and bias of the classifier F's classification layer respectively.
$\mathbf{p} \in \mathbb{R}^{\|\mathcal{C}\|}, \sigma(\mathbf{z})$	Probability distribution $\mathbf{p}$ over classes, computed as $\mathbf{p} = \sigma(\mathbf{z}) = \text{softmax}(\mathbf{W}^\top \mathbf{z} + \mathbf{b})$ function applied to the latent feature representation $\mathbf{z}$.
$\mathbf{l} \in \mathbb{R}^{\|\mathcal{C}\|}$	Logit before applying softmax, *i.e.*, $\mathbf{l} = \mathbf{W}^\top \mathbf{z} + \mathbf{b}$.
$\mathbf{W}_s, \mathbf{W}_t \in \mathbb{R}^N; \mathbf{b}_s, \mathbf{b}_t$ (scalars)	$\mathbf{W}_s$ and $\mathbf{W}_t$ are weight vectors which are the "slices" of $\mathbf{W}$ for the classes s or t. Similarly, $\mathbf{b}_s$ and $\mathbf{b}_t$ are the elements in the bias vector $\mathbf{b}$ for classes s and t.
$\mathbf{W}_m \in \mathbb{R}^N, \mathbf{b}_m$ (scalar)	Pairwise decision boundary between the source and target classes, *i.e.*, the mirror. Calculated as the difference of class weights ($\mathbf{W}_t - \mathbf{W}_s$) and bias ($\mathbf{b}_t - \mathbf{b}_s$).
$G(\mathbf{z})$	A mapping function G to project latent feature $\mathbf{z}$ into image space.
k (scalar)	Step factor where $k \in [0, 1]$.
$\mathbf{f}^i \in \mathbb{R}^{C_i \times H_i \times W_i}$	A spatial feature $\mathbf{f}^i$ is obtained from i-th layer of the feature encoder F. C_i, H_i and W_i denote the channel, height, and width of $\mathbf{f}^i$.
l (scalar)	The highest layer in F which produces $\mathbf{z}$: $\mathbf{z} = \text{GAP}(\mathbf{f}^l)$, hence $N = C_l$.
$\mathbf{x}_{ss}, \mathbf{z}_{ss}$	A random source class image $\mathbf{x}_{ss}$ corresponds to latent feature $\mathbf{z}_{ss}$.
$\mathbf{x}_r, \mathbf{z}_r$	A generated mirror point image $\mathbf{x}_r$ from mirror latent feature $\mathbf{z}_r$ at step factor $k = 1.0$.
$\mathbf{x}_{cf}, \mathbf{z}_{cf}$	A generated counterfactual image $\mathbf{x}_{cf}$ from latent feature $\mathbf{z}_{cf}$ at step factor $k = 0.5 + \epsilon$ where $0 < \epsilon <= 0.5$.
$\mathbf{x}_{sf}, \mathbf{z}_{sf}$	A generated semi-factual image $\mathbf{x}_{sf}$ from latent feature $\mathbf{z}_{sf}$ at step factor $k = 0.5 - \epsilon$.
$\mathbf{x}_k, \mathbf{z}_k$	A generated KFE image $\mathbf{x}_k$ from latent feature $\mathbf{z}_k$ at step factor $k \in [0, 1]$.
$\mathcal{L}_{cls}$	Classification loss ensuring the validity of counterfactual explanations.
$\mathcal{L}_{adv}$	Adversarial loss ensuring realism in generated counterfactuals.
$\mathcal{L}_{rec}$	Reconstruction loss ensuring regenerated images resemble the original inputs.
$\mathcal{L}_{fea}$	Feature reconstruction loss for consistency in latent space.
$\mathcal{L}_{tri}$	Triangulation loss maintaining proximity and realism during counterfactual generation.
α (scalar)	Scaling factor in triangulation loss $\mathcal{L}_{tri}$ where $\alpha \in [0, 1]$.
B_i, D_i	The bottleneck and decoder modules used in the i-th layer's Spatial pattern editor (SPE) module.
$\mathbf{u}_k^i \in \mathbb{R}^{C_i \times H_i \times W_i}$	Output from the SPE module in i-th layer at step k.
$\mathbf{U}_k \in \mathbb{R}^{\|\mathcal{C}\| \times H_l \times W_l}$	Unnormalized CAMs at step k computed as $\mathbf{U}_k = \mathbf{W}^\top \mathbf{f}_k^l$.
$\mathbf{N}_k \in \mathbb{R}^{\|\mathcal{C}\| \times H_l \times W_l}$	Normalized CAMs at step k where each channel is individually normalized between [0, 1].
$\mathbf{M}_k^i \in \mathbb{R}^{H_i \times W_i}$	Spatial prior mask i-th layer at step k. $\mathbf{M}_k^i$ is uniformly applied to all channels.

Table 7. Symbol lookup table.