# Augmentation by Counterfactual Explanation - Fixing an Overconfident Classifier

Sumedha Singla*
University of Pittsburgh
sumedha.singla@pitt.edu

Nihal Murali*
University of Pittsburgh
nihal.murali@pitt.edu

Forough Arabshahi
Meta AI
forough@meta.com

Sofia Triantafyllou
University of Crete
sof.triantafillou@gmail.com

Kayhan Batmanghelich
University of Pittsburgh
kayhan@pitt.edu

## 1. Implementation Details

### 1.1. Dataset

We focus on improving classification models based on deep convolution neural networks (CNN) as most state-of-the-art performance models fall in this regime. In our experiments, we consider classification models trained on following datasets:

1. AFHQ [2]: Animal face high quality (AFHQ) dataset is a high resolution dataset of animal faces with 16K images from cat, dog and wild labels. In our experiments, we consider a multi-class classifier over cat and dog labels. We consider images with "wild" label as near-OOD. The classifier is trained at an image resolution of $256 \times 256$.

2. Dirty MNIST [8]: The dataset is a combination of original MNIST [5] and simulated Ambiguous-MNIST dataset. Each sample in Ambiguous-MNIST is constructed by decoding a linear combination of latent representations of two different MNIST digits from a pre-trained VAE [4]. The training dataset of the classifier comprises of 60K clean-MNIST and 60K Ambiguous-MNIST samples, with one-hot labels. In our experiments, we consider classifier trained on seven classes over digits '0' - '6'. We consider images from digits '7' - '9' as near-OOD samples. The original dataset consists of grayscale images of size $28 \times 28$ pixels. We consider a classification model trained on $64 \times 64$ resolution.

3. Skin lesion (HAM10K) [9]: The HAM10000 is a dataset of 100K dermatoscopic images of pigmented skin lesions. It contains seven different lesion types – Melanocytic Nevi (nv), Melanoma (mel), Benign Keratosis (bkl), Actinic Keratoses and Intraepithelial Carcinoma (akiec), Basal Cell Carcinoma (bcc), Dermatofibroma (df), Vascular skin lesions (vasc). In our experiments, we consider classifier trained to distinguish the majority class nv from mel and bkl. We consider images from rest of the lesions as near-OOD. The classifier is trained at an image resolution of $256 \times 256$.

4. CelebA [7] : Celeb Faces Attributes Dataset (CelebA) is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 binary attributes annotations per image. In our experiments, we consider a two-class classifier over attributes "Young" and "Smiling" trained on CelebA dataset. Our AiD samples comprises of middle-aged people who are arguably neither young nor old. To obtain such data, we use aleatoric uncertainty estimates from MC-Dropout averaged across 50 runs on test-set of CelebA. The classifier is trained at an image resolution of $256 \times 256$. We center-crop the images as a pre-processing step.

### 1.2. Classification Model

We used DenseNet architecture as the classification model. In DenseNet, each layer implements a non-linear transformation based on composite functions such as Batch Normalization (BN), rectified linear unit (ReLU), pooling, or convolution. The resulting feature map at each layer is used as input for all the subsequent layers, leading to a highly convoluted multi-level multi-layer non-linear convolutional neural network. We aim to improve such a model in a post-hoc manner without accessing the parameters learned by any layer or knowing the architectural details. Our proposed approach can be used for any DNN architecture.

[*] Equal contribution

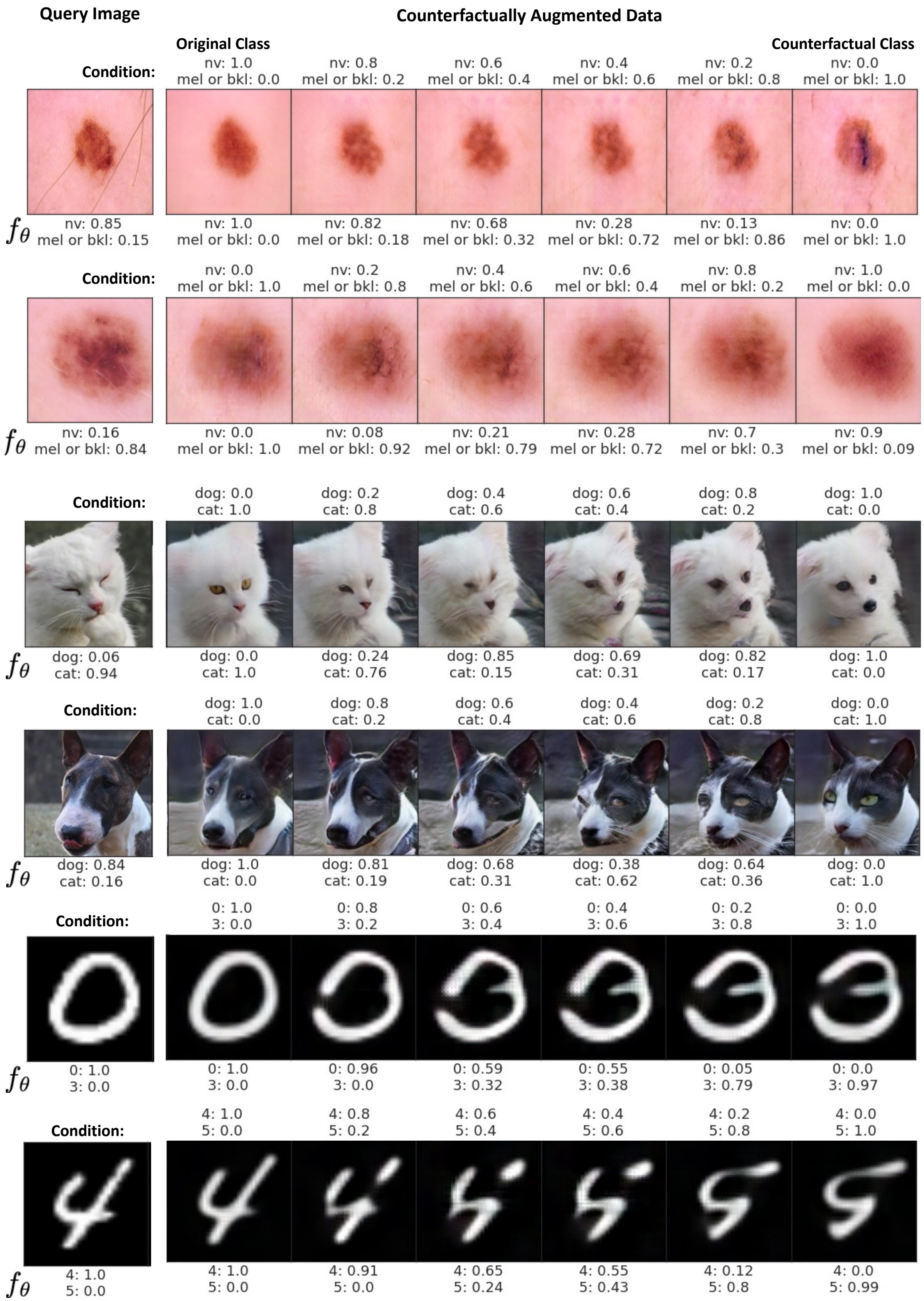


Figure 1. Examples of data augmentation using counterfactual explanations for different datasets.

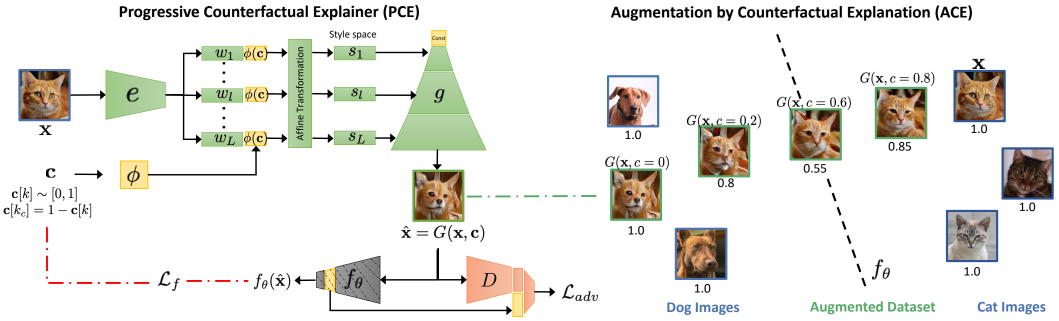


Figure 2. PCE: The encoder-decoder architecture to create counterfactual augmentation for a given query image. ACE: Given a query image, the trained PCE generates a series of perturbations that gradually traverse the decision boundary of $f_\theta$ from the original class to a counter-factual class, while still remaining plausible and realistic-looking.

### 1.3. Progressive Counterfactual Explainer

We formulate the progressive counterfactual explainer (PCE) as a composite of two functions, an image encoder $e(\cdot)$ and a conditional decoder $(g(\cdot))$ [1]. Our architecture for the conditional decoder is adapted from StyleGANv2 [1]. The image encoder converts the input image $\mathbf{x}$ into $l$ different latent codes ($w_l \in \mathbb{R}^{512}$), for each of the $L$ layers of the decoder. The decoder further transforms the layer-specific latent representation into a layer-specific style-vector as $s_l = A_l([w_l, \phi(\mathbf{c})])$ where, $A_l$ is an affine transformation and $\phi(\mathbf{c})$ is an embedding for $\mathbf{c}$. For training the StyleGANv2 decoder, we consider the default training parameters from [1]. For training the PCE, we use a randomly sampled subset ($\sim 50\%$) of the baseline training data. Given an input image, the predicted class $k$ and a counterfactual class $k_c$, we initialize the condition $\mathbf{c}$ with all zeros and then set $\mathbf{c}[k] \sim \text{Uniform}(0,1)$ and $\mathbf{c}[k_c] = 1 - \mathbf{c}[k]$. In all our experiments, we used $\lambda_{adv} = 10$, $\lambda_{rec} = 100$ and $\lambda_f = 10$. Fig. 2 summarizes our architecture.

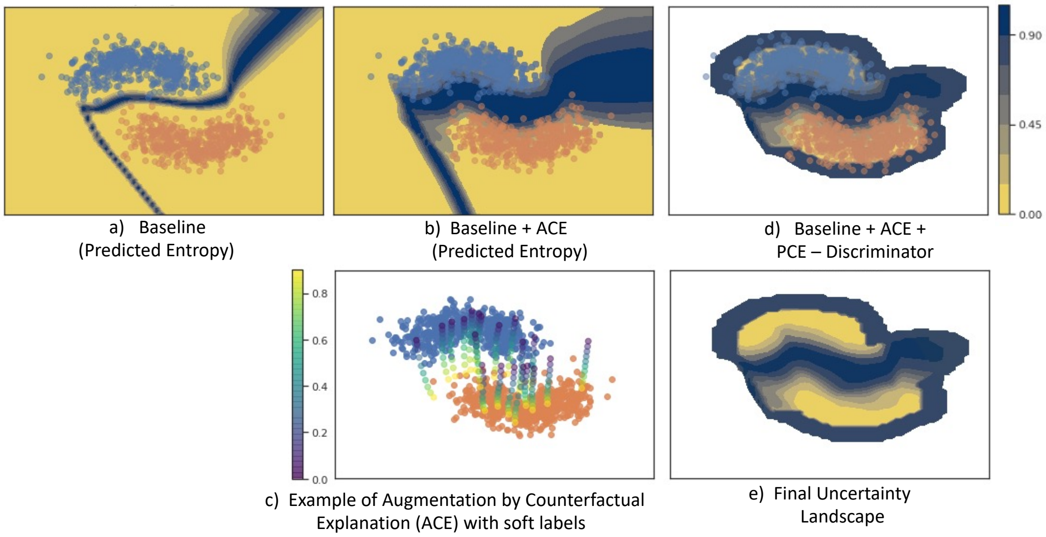
For generating counterfactually augmented data, we first consider a randomly selected subset of real training data as $\mathcal{X}_r \in \mathcal{X}$. For each image in $\mathcal{X}_r$, we generate four augmented images by randomly selecting the $\mathbf{c}[k]$. For each augmented image, we used the condition used to generate the image as the soft label while fine-tuning. Fig. 2 shows an example of our data augmentation. We denote the pool of the augmented images as $\mathcal{X}_c$. In Fig. 1, we show examples of counterfactual augmentation from different datasets.

For fine-tuning the given baseline with consider a combination of the original training dataset $\mathcal{X}$ and the augmented data $\mathcal{X}_c$. We randomly selected a subset of samples from the two distributions and fine-tune the baseline for 5 to 10 epochs. We used the expected calibration error and the test-set accuracy to choose the final checkpoint. Our model does not require access to OOD or AiD dataset during fine-tuning. During evaluation we compute predicted entropy (PE) for original test-set and OOD samples and measure for a range of thresholds how well the two are separated. We report the AUC-ROC and the true negative rate (TNR) at 95% true positive rate (TPR) (TNR@TPR95) in our results (*see Table 1 and 2*). We will release the GitHub for the project after the review process.

## 2. Toy-Setup: Two-Moons

In this section, we demonstrate our method on a toy setup: the Two Moons dataset. We used the experimental set-up from DDU [8] for this experiment. We use scikit-learn's datasets package to generate 2000 samples with a noise rate of 0.1. For baseline classification model, we use a 2-layer dense-layer architecture, with ReLU activation and batch normalization. The 2-D input data is projected to a 64-D latent space and then to 1D space to make final binary prediction. In Fig. 3.a, we show the uncertainty estimates (predicted entropy PE) from the baseline classifier. The baseline classifier is uncertain only along the decision boundary, and certain elsewhere (low PE).

Given the baseline classifier, we train a PCE to generate augmented data. We use an encoder with two fully-connected layers that map 2-D input data to a 64-D latent space. The condition is also projected to a 64-D space and is concatenated to the output of the encoder. The decoder also have two fully-connected layers that maps the concatenated 128-D latent vector back to a 2-D input space. In Fig. 3.c, we show example of augmentation by counterfactual explanation (ACE). Given a query point, we generate series of augmented data by gradually changing the condition such that the decision of the baseline is flipped. The color of the dot represents the conditioned used to create the augmented sample. Next, we fine-tune the classifier using the original and the counterfactually augmented data. In Fig. 3.b, we show the PE estimates from the fine-tuned classifier (baseline + ACE). Fine-tuning with our augmented data widen the decision boundary. Finally, we used the discriminator of the PCE as a density estimator, to identify and reject OOD data. The discriminator is trained on real/fake samples near



a) Baseline (Predicted Entropy)

b) Baseline + ACE (Predicted Entropy)

d) Baseline + ACE + PCE – Discriminator

c) Example of Augmentation by Counterfactual Explanation (ACE) with soft labels

e) Final Uncertainty Landscape

Figure 3. Uncertainty results on Two Moons dataset. Yellow indicates low uncertainty, while blue indicates uncertainty. a) The baseline classifier is uncertain only along the decision boundary, and certain elsewhere. b) Fine-tuning baseline model on ACE data improves uncertainty estimates near the decision boundary. c) An example of augmented data and corresponding soft labels. d) The discriminator from PCE rejects OOD samples, hence the rejected space have no uncertainty values (white color). e) The final uncertainty landscape, the improved classifier is certain on in in-distribution regions and rejects OOD data.

the training distribution. Hence, we used a threshold of 0.5 on the discriminator to reject everything that is far from the training distribution. In Fig. 3.d, the white color show the input space that is rejected by the discriminator. In Fig. 3.e, we show the final uncertainty landscape without overlaying the training data. We improved the baseline model, to have high certainty only in in-distribution regions. The uncertainty increases as we go near the decision boundary. Thus in addition to image classifiers, our strategy improves the uncertainty estimates even for a classifier trained on a small 2D setup like Two Moons.

## 3. Additional Results

Much of the prior work has focused on obtaining uncertainty estimates from a pre-trained DNN output using threshold-based scoring functions. Liu et al [6] in their paper show how energy functions can be used not only as scoring functions but also as a trainable cost-function to shape the energy surface explicitly for OOD-detection. Hendrycks et al [3] propose the Outlier Exposure method which regularizes the softmax outputs to be a uniform distribution for outlier data. We compare these commonly-used methods against our technique (ACE) and show the results in Tables 1 and 2. Our method is consistent and competitive, if not outperforming, across all datasets and AiD/OOD categories.

## 4. Ablation Study

We conducted an ablation study over the three loss terms of PCE in Eq. 5. The three terms of the loss function enforces three properties of counterfactual explanation, data consistency: explanations should be realistic looking images, classifier consistency: explanations should produce a desired outcome from the classifier and self consistency: explanation image should retain the identity of the query image. For ablation study we consider the cat and dog classifier. We train three PCE, in each run we ablate one term from the final loss function. In Fig. 4, we show qualitative example of the counterfactual data augmentation generated through each PCE. Without data consistency, the images are blur and are no longer realistic. Without classifier consistency loss, though the images are realistic, but the output of the classifier is not changing with the condition, hence such PCE won't generate augmented samples near the decision boundary, which is the goal of our proposed strategy. With self consistency, the generated images are not a gradual transformation of a given query image.

Further, in Fig. 5 we present quantitatively compare the uncertainty estimates from the baseline, before and after the fine-tuning with ACE. In each row, we represent a different ablation over the three loss terms. Fig. 5.A. shows the predicted entropy (PE) of **in-distribution (iD)** samples. Ideally, fine-tuning should minimally effect the PE distribution over iD samples. Without classification consistency loss (second row), the PE distribution of iD samples changed significantly. Fig. 5.B and Fig. 5.C shows the PE distribution over **ambiguous in-distribution (AiD)** samples and **near-OOD** samples, respectively. The data augmentation derived from PCE without adversarial loss or reconstruction loss, is not able to separate AiD samples or near-OOD from rest of the test set. In Fig. 5.D, we use the discriminator of

Table 1. Additional results on identifying **ambiguous in-distribution (AiD)** samples. For all metrics, higher is better.

| **Train Dataset** | **Method/ Model** | **Test-Set** Accuracy | **Identifying AiD** AUC-ROC | TNR@TPR95 |
|---|---|---|---|---|
| AFHQ | Baseline+energy [6] | 99.44±0.02 | 0.87±0.06 | 49.00±1.64 |
| | Energy w/ fine-tune [6] | 99.45±0.11 | 0.69±1.28 | 30.36±2.52 |
| | Outlier Exposure [3] | 99.50±0.14 | 0.85±0.01 | 41.07±0.75 |
| | **Baseline+ACE** | **99.52±0.21** | **0.91±0.02** | **50.75±3.9** |
| Dirty MNIST | Baseline+energy [6] | 95.68±0.02 | 0.80±0.03 | 17.60±0.55 |
| | Energy w/ fine-tune [6] | 96.17±0.02 | 0.39±0.04 | 11.59±0.25 |
| | Outlier Exposure [3] | **96.30±0.07** | 0.63±0.07 | 17.6±2.88 |
| | **Baseline+ACE** | 95.36±0.45 | **0.86±0.01** | **34.12±2.60** |
| CelebA | Baseline+energy [6] | 89.36±0.96 | 0.57±0.28 | 4.87±0.32 |
| | Energy w/ fine-tune [6] | **90.22±0.96** | 0.53±1.25 | 5.06±0.28 |
| | Outlier Exposure [3] | 86.65±1.22 | 0.53±0.46 | 5.06±0.19 |
| | **Baseline+ACE** | 86.80±0.79 | **0.74±0.06** | **22.36±2.30** |
| Skin-Lesion (HAM10K) | Baseline+energy [6] | 85.88±0.75 | 0.77±0.12 | 18.40±0.51 |
| | Energy w/ fine-tune [6] | **86.56±0.53** | 0.64±0.06 | 17.45±1.78 |
| | Outlier Exposure [3] | 86.37±0.46 | 0.73±0.02 | 13.21±2.70 |
| | **Baseline + ACE** | 81.21±1.12 | **0.84±0.05** | **71.60±3.80** |

Table 2. OOD detection performance for different scoring-based methods.

| **Train Dataset** | **Method** | **Near-OOD (Wild)** AUC-ROC | TNR@TPR95 | **Far-OOD (CIFAR10)** AUC-ROC | TNR@TPR95 | **Far-OOD (CelebA)** AUC-ROC | TNR@TPR95 |
|---|---|---|---|---|---|---|---|
| AFHQ | Baseline+energy [6] | 0.88±0.03 | 47.77±1.10 | 0.94±0.05 | 72.68±2.69 | 0.96±0.04 | 74.75±2.89 |
| | Energy w/ fine-tune [6] | **0.93±3.06** | 45.97±2.78 | **0.99±0.00** | 0.66±0.01 | 0.94±1.86 | 68.38±3.03 |
| | Outlier Exposure [3] | 0.92±0.01 | **73.99±2.62** | 0.99±0.20 | **99.54±0.79** | 0.96±0.01 | 78.69±3.02 |
| | **Baseline+ACE** | 0.89±0.03 | 51.39±4.40 | 0.98±0.02 | 88.71±5.70 | **0.97±0.03** | **88.87±9.80** |
| Dirty MNIST | Baseline+energy [6] | 0.87±0.04 | 40.30±1.05 | 0.86±0.12 | 43.92±2.30 | 0.91±0.02 | 62.10±5.17 |
| | Energy w/ fine-tune [6] | 0.60±0.08 | 37.43±0.93 | **1.00±0.00** | **99.99±0.00** | **1.00±0.00** | 99.06±0.01 |
| | Outlier Exposure [3] | **0.94±0.01** | **65.58±1.64** | **1.00±0.00** | **99.99±0.00** | **1.00±0.00** | **99.56±0.12** |
| | **Baseline+ACE** | **0.94±0.02** | 37.23±1.90 | 0.98±0.02 | 67.88±3.10 | 0.97±0.02 | 70.71±1.10 |
| CelebA | Baseline+energy [6] | 0.76±0.51 | 9.40±0.01 | 0.94±0.08 | 32.08±1.70 | 0.85±0.76 | 17.10±0.72 |
| | Energy w/ fine-tune [6] | 0.85±1.27 | 32.81±1.92 | **0.99±0.00** | **99.99±0.00** | 0.91±0.77 | **84.35±1.29** |
| | Outlier Exposure [3] | 0.66±0.69 | 8.44±0.45 | 0.75±0.70 | 26.09±0.51 | 0.69±0.53 | 16.63±0.90 |
| | **Baseline+ACE** | **0.87±0.03** | **34.37±2.50** | 0.96±0.01 | 96.35±2.50 | **0.92±0.05** | 63.51±1.50 |
| Skin-Lesion (HAM10K) | Baseline+energy [6] | 0.70±0.04 | 10.85±0.08 | 0.70±0.14 | 7.90±0.29 | 0.65±0.20 | 2.83±1.33 |
| | Energy w/ fine-tune [6] | 0.62±0.02 | 9.80±1.81 | **1.00±0.00** | **99.77±0.33** | 0.76±0.13 | 16.04±1.08 |
| | Outlier Exposure [3] | 0.67±0.09 | 10.38±3.30 | 0.99±0.00 | 97.17±2.37 | 0.81±0.08 | 22.64±4.30 |
| | **Baseline+ACE** | **0.72±0.04** | **10.99±2.80** | 0.97±0.02 | 66.77±1.40 | **0.96±0.03** | **95.83±5.00** |

the PCE to identify **far-OOD** samples. In all three rows, we observe sub-optimal performance of the discriminator in identifying and rejecting far-OOD samples. The legend shows the AUC-ROC for binary classification over uncertain samples and iD samples. Hence, all three loss terms are important to improve the uncertainty estimates of the baseline over all samples across the uncertainty spectrum.

## 5. Robust Generalization

In this experiment, we establish a connection between loss landscape plots and generalization of classifiers. In order to qualitatively understand the improved generalization of our method, we try to visualize the high-dimensional loss landscape via 3D weight visualization plots as shown by Li *et al.* We compute the cross-entropy loss using test set of CelebA and AFHQ and follow the method given by Li *et al.* to compare the loss landscape geometry for the baseline model and our method (ACE).

We observe that our method leads to smooth and flatter loss landscapes as compared to baseline. This shows that slight perturbation to the weight does not change the loss much, which may qualitatively explain why we obtain better generalization performance and robustness to adversarial attacks in our experiments. We do not thoroughly investigate this direction and leave it as an important direction for future research.

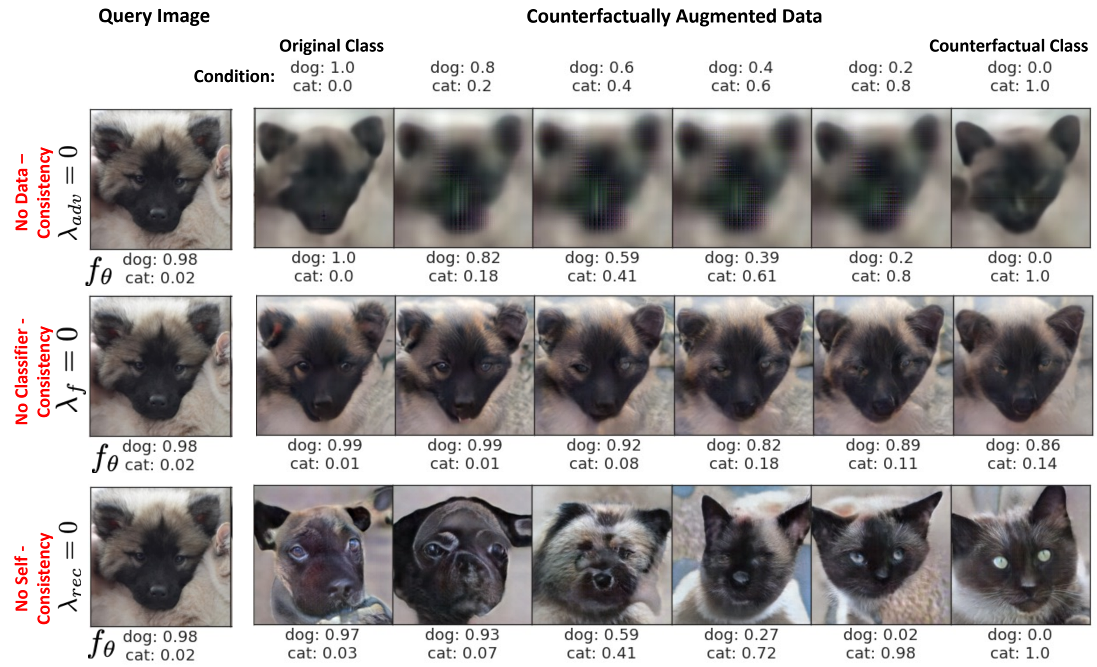


Figure 4. Examples of data augmentation while ablating different loss terms.

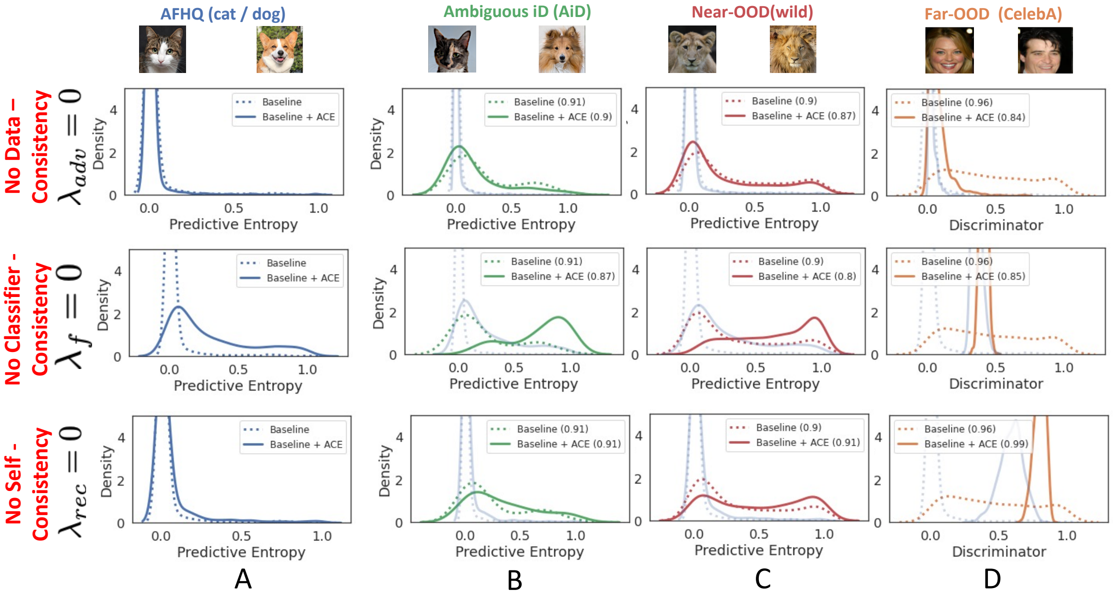


Figure 5. Comparison of the uncertainty estimates from the baseline, before and after the fine-tuning with ACE. Each row represents a different ablation over the three loss terms. A) Predicted entropy (PE) of **in-distribution (iD)** samples. Ideally, fine-tuning should minimally effect the PE distribution over iD samples. Without classification consistency loss (second row), the PE distribution of iD samples changed significantly. B) PE distribution over **ambiguous in-distribution (AiD)** samples. C) PE distribution over **near-OOD** samples. The data augmentation derived from PCE without adversarial loss or reconstruction loss, is not able to separate AiD samples or near-OOD from rest of the test set. D) We use the discriminator of the PCE to identify **far-OOD** samples. In all three rows, we observe sub-optimal performance of the discriminator in identifying and rejecting far-OOD samples. The legend shows the AUC-ROC for binary classification over uncertain samples and iD samples. Hence, all three loss terms are important to improve the uncertainty estimates of the baseline over all samples across the uncertainty spectrum.

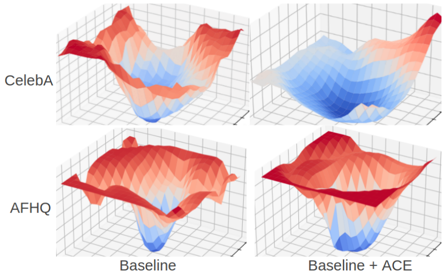


Figure 6. Weight loss landscape visualizations for baseline model and our method on CelebA and AFHQ datsets

## References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[3] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *International Conference on Learning Representations*, 2019.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

[5] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[6] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020.

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[8] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *arXiv preprint arXiv:2102.11582*, 2021.

[9] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.