Direct Adversarial Latent Estimation to Evaluate Decision Boundary Complexity in Black Box Models

Ashley S. Dale[1] and Lauren Christopher[1]

[1]Affiliation not available

April 24, 2024

Abstract

A trustworthy AI model should be robust to perturbed data, where robustness correlates with the dimensionality and linearity of feature representations in the model latent space. Existing methods for evaluating feature representations in the latent space are restricted to white-box models. In this work, we introduce *Direct Adversarial Latent Estimation* (DALE) for evaluating the robustness of feature representations and decision boundaries for target black-box models. A surrogate latent space is created using a Variational Autoencoder (VAE) trained on a disjoint dataset from an object classification backbone, then the VAE latent space is traversed to create sets of adversarial images. An object classification model is trained using transfer learning on the VAE image reconstructions, then classifies instances in the adversarial image set. We propose that the number of times the classification changes in an image set indicates the complexity of the decision boundaries in the classifier latent space; more complex decision boundaries are found to be more robust. This is confirmed by comparing the DALE distributions to the degradation of the classifier F1 scores in the presence of adversarial attacks. This work enables the first comparisons of latent-space complexity between black box models by relating model robustness to complex decision boundaries.

Introduction

Trustworthy Artificial Intelligence (A.I.) for computer vision continues to be an important research area. In particular, robustness–one of several proposed dimensions of AI Trustworthiness (Floridi, 2019; Wing, 2021; Kaur et al., 2022)–is a property of a model where a perturbation of an input does not cause the model to return an inaccurate prediction (Drenkow et al., 2021; Han et al., 2023). Adversarial attacks on a model leverage a lack of robustness by causing features to be misassigned either during the feature extraction (i.e., mapped to an incorrect place in the latent space (Gat et al., 2022)), or during a secondary task (e.g., features assigned to the incorrect class (Wei et al., 2022)). Our contribution is a metric that relates a black-box model's robustness with the complexity of decision boundaries in the latent space.

Quantifying the robustness of a model is a challenging task (Li et al., 2023; Weber et al., 2023; Celdran et al., 2023). Model robustness may be tested through untargeted attacks that perturb inputs through the addition of noise or simple transformations (Akhtar et al., 2021) or more sophisticated targeted adversarial attacks (Lee & Kim, 2023). In each test, data are mapped to available features in the model's latent space, and then a secondary task (e.g., classification, tracking, and/or segmentation) is performed on the extracted features by another portion of the model architecture. Robustness in the secondary task relies partially on the robustness of the features produced by the feature extractor backbone; the decision boundaries represent the organization of these features in the latent space.

Latent space adversarial vulnerability exists in a transfer-learning context, where pretrained feature extraction backbones are implemented with frozen weights. This makes the model features and their relationships

inaccessible. Black-box model attacks are possible with an adversarial example from a surrogate model that successfully transfers and attacks the target black-box model (Gu et al., 2023). This transferability has been explained using white-box models, where an adversarial latent space is shared by the surrogate and target models. The shared latent space is characterized by similar decision boundaries (Tramèr et al., 2017) that exist in a region which is locally linear (Cubuk et al., 2017; Godfrey et al., 2023). However, the transferability of an adversarial example does not depend on the degree of perturbation as measured by the distance from a classification boundary (Tramèr et al., 2017; Waseda et al., 2023). Transferability gives us the connection between two models, and indicates that the latent space of a source model may represent the latent space of a target model.

We connect cross-model robustness and latent space adversarial vulnerabilities in the following ways: **First** , we use a surrogate model to estimate the complexity of the target model latent space, where the surrogate feature space is trained on disjoint data from the target model and there is no shared information during the training process.

Second , we introduce *Direct Adversarial Latent Estimation* (DALE), a new method for quantifying the complexity of decision boundaries in the latent space. State-of-the-art metrics for latent space complexity directly leverage latent feature representations by analyzing the similarities of model layer weights; examples include Canonical Correlation Analysis (Hardoon et al., 2004), Centered Kernel Alignment (Kornblith et al., 2019), Singular Vector Canonical Correlation Analysis (Raghu et al., 2017), and Graph-Based Similarity (Chen et al., 2021). Because these metrics require access to the model weights, they are unsuitable for black-box models. To our knowledge, a method for comparing the complexity of different latent spaces without the use of model weights has not been proposed until this paper.

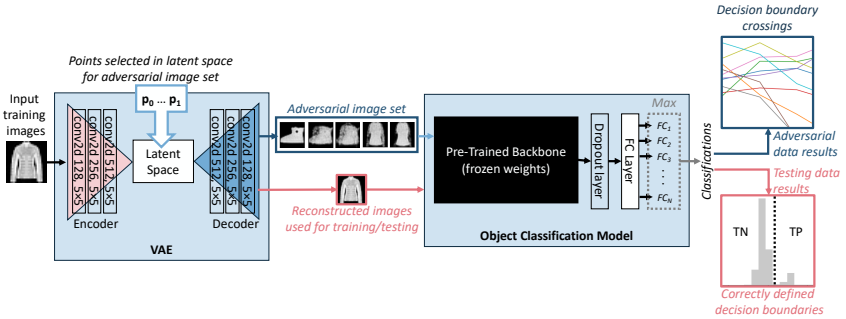


Figure 1: Experimental procedure discussed in §. A Variational Autoencoder (VAE) is trained to reconstruct images from a source dataset such as Fashion-MNIST (Xiao et al., 2017). An object classification model is initialized with a black-box feature extraction backbone, where only the fully connected (FC) layer and drop out layer of the model are trained with the VAE reconstructions. The classification results for each FC perceptron show that true positives (TP) for the perceptron class are separable from the true negatives (TN). A set of adversarial images generated by sampling the latent space of the VAE can then be used to map decision boundary crossings in the latent space of the target model by evaluating shifts in the distribution of values produced by the FC layer.

Third , we show that the number of decision boundaries crossed in the classification model latent space by an adversarial image set does not depend on the amount of surrogate latent space perturbation, but is correlated with the robustness of the target model to adversarial attacks. We contextualize the lack of dependence on the amount of perturbation by considering the expression of perturbation as a vector in a high-dimensional non-Euclidean space; any distance calculation assuming a Euclidean space such as the standard L2-norm will fail to accurately capture the degree of perturbation.

In this work, DALE was leveraged to analyze target object classification models initialized with a pretrained

feature extraction backbone and a classification head trained using image reconstructions from a Variational Autoencoder (VAE). Each target model was tested using adversarial image sets generated by traversing the surrogate latent space of the VAE model. Classification results show transitions in object classifier predictions as shown in Fig. 1 and generates the DALE distributions shown in Fig. 5 that maps the decision boundaries of the target model's latent space relative to the surrogate latent feature space. We then test each object classifier model with untargeted adversarial attacks, and show that the DALE distributions correlate with the robustness of the object classifier to attacks.

Background

In this section, we discuss previous analyses of robust latent spaces, the use of the VAE as a surrogate model of the object classifier latent space, and the role of augmented and synthetic data in testing latent space feature representations.

Latent space robustness

Neural networks have been successfully used for computer vision tasks. The feature vectors generated by the model backbone are samples taken from an N-dimensional latent space manifold encoding features in the training dataset. Sampling along a direction on the manifold results in a set of feature vectors, where features may be drawn from multiple classes to form a perturbed vector.

A robust latent space is composed of robust features, and the development of robust latent spaces is an active area of research (Fang et al., 2017; Li & Han, 2024). A robust latent space serves two purposes: First, it improves model performance by hardening a model against adversarial examples (Khaire & Dhanalakshmi, 2022; Huang et al., 2021). However, it is known that specifically training a model with adversarial examples does not significantly impact the model's decision boundaries (Tramèr et al., 2017). Second, robust features facilitate explainable model predictions (Din et al., 2020). Recent research has produced explainable latent spaces through feature disentangling methods applied during training (Lin et al., 2019), visualization methods (Liu et al., 2019; Sainburg et al., 2019), and loss functions designed to maximize the information of the latent space (Higgins et al., 2017; Dubois et al., 2019; Hou et al., 2017).

Latent space complexity therefore has a significant role in the model robustness. A high-dimensional and locally-linear latent space is the source of adversarial examples (Godfrey et al., 2023; Gu et al., 2023), and this region may be large due to the approximately piece-wise linear structure of common activation functions such as ReLU and sigmoid (Warde-Farley & Goodfellow, 2016).

VAE as a latent space surrogate

When using a black box model with a pre-formed latent space–such as in the context of deploying pretrained weights during transfer learning–the robustness of the latent space is fixed and can only be evaluated. The dataset used to form the black-box latent space may be inaccessible, and the model may be deployed in a context other than that for which it was trained. In these situations, the use of surrogate models for developing adversarial attacks on a target model's latent space is already well established (Waseda et al., 2023; Akhtar et al., 2021; Ilyas et al., 2019).

Intuitively, one object classification model may approximate another object classification model. Trade-offs in model size and performance are frequent, where a model with worse performance and lower resource consumption is selected over a larger model with better performance; the smaller model approximates the larger model. This motivates knowledge distillation (KD), where a large target model is compressed by training a smaller surrogate model to perform the same task. KD is effective even when the architectures of the target and surrogate models are entirely dissimilar (Wang & Yoon, 2021); this is the case when the latent space of a Variational Autoencoder Encoder (VAE) is used to approximate the latent space of an

object classification model. KD relies on minimizing the difference between the surrogate's and target's inputs to their respective softmax layers; this is equivalent to forcing similar latent space encodings for a given input. Theoretically, Wang and Yoon, (Wang & Yoon, 2021) attributes the success of this approach to a maximization of mutual information between the two models.

The mutual information perspective is complemented by understanding the role of individual features during training. In a feature-based context, (Olah et al., 2020) posits the universality of features learned across all networks by examining individual neurons, and both (Cianfarani et al., 2022) and (Jones et al., 2022) present results suggesting that the robustness of a model is correlated with the universality of the features implemented by the model. Robust features have been found in the features learned by more-brittle models (Jones et al., 2022), and feature robustness was improved by forcing intermediate latent space geometries to be smooth (Lassance et al., 2021). These results, along with the mutual information perspective, suggest similar latent spaces across models regardless of training regime or architecture, and perhaps even independent of training data within a knowledge domain.

Role of different types of data in testing model robustness

In this work, we use real, synthetic, and adversarial data distributions to test the robustness of object classification models.

Synthetic, or Virtual World (VW) data is commonly used to make large datasets, and the efficacy of synthetic training data is well documented (Chen et al., 2021). In open source datasets containing synthetic data (Shermeyer et al., 2021; Naphade et al., 2021; Geiger et al., 2013), synthetic instances are crafted to contain desirable features, e.g., weather conditions, which may be hard to empirically capture. In this way, features are added to the dataset to better match the expected deployment environment. Synthetic data is helpful as long as there is a minimal domain gap between the real and synthetic data instances (Dale et al., 2023).

Latent space generated data is another source of synthetic images (Guibas et al., 2017; Jetchev et al., 2016; Sampath et al., 2021). Here, "realistic" images are created from features in the original distribution that are combined to create new instances. This generative data interpolates between features already present in the distribution rather than introducing new features.

Adversarial instances may be generated using a perturbation δ added to a dataset image x to create an adversarial instance $x' = x + \delta$, where for a given perturbation budget ϵ, $\delta \leq \epsilon$. The perturbation δ has various sources, depending on whether the goal is to degrade performance as in an untargeted attack or force a misclassification as in a targeted attack.

In this work, we create adversarial image sets to stress the model decision boundaries. For our adversarial image set, the perturbation budget ϵ is not a fixed constant, and instead depends on points p_0, p_1 in the VAE surrogate latent space. A trajectory in the latent space between p_0 and p_1 is described by the line $f(\beta) = (p_1 - p_0)\beta + C$ where C is an arbitrary constant, β is a fractional step size between p_0 and p_1, and $f(\beta)$ is a new point on the latent space manifold. If we take the perturbation budget to be $\epsilon = (p_1 - p_0)$ and $C = 0$, then a point $f(\beta) = p_n$ sampled along a specific trajectory in the VAE latent space can be generated using

$$p_n = p_0 + \epsilon n \tag{1}$$

where $n \in [0, 1]$, p_0 is the initial latent space point so that $p_1 = p_0 + \epsilon$. Each point p_n is then transformed into an adversarial image x' by the VAE decoder sub-architecture. Sampling n sequentially between $[0, 1]$

creates an ordered set of latent space points along a trajectory between p_0 and p_1. This set of ordered points can then generate ordered sets of adversarial images. We intentionally stress the latent space of the classifier backbone with adversarial image sets of inter-class features. Observing the effects of the mixed-class features on the predictions generated by the model classification head allows us to quantify the complexity of the decision boundary entanglement in the object classifier latent space.

Methods

Two Variational Autoencoders (VAEs) and six object classification models were each trained on a single NVIDIA RTX A6000 GPU with 48 GB RAM. The latent space of each VAE was sampled to create sets of adversarial images according to Eq. 1. These image sets were then classified using the trained object classifiers by considering the index of the max value of the output vector of the last fully connected (FC) layer, equivalent to a one-hot vector label. The number of times the class label changes during an adversarial image set is related to the number of decision boundaries crossed in the object classifier latent space. The distribution of object classifier decision boundary crossings is then visualized as a function of VAE perturbation ϵ, and results are presented in §. The remainder of this section discusses details of the datasets used during transfer learning, the data generated by traversing the latent space of the VAE, and the object classifier backbones.

Real and synthetic datasets

This study used two datasets for training: the Fashion-MNIST test split (Xiao et al., 2017) and RGBZFO (Dale, 2021). All images were rescaled to $128 \times 128 \times 3$ for the VAE and as per object classification backbone input requirements.

The Fashion-MNIST test split consists of 10k black and white images in ten classes(Xiao et al., 2017). Of the original 10k images, the training set consisted of 8k images, with 1k reserved for validation and testing each. The black and white images were chosen since the backbones were pre-trained using RGB data.

RGBZFO consists of 9160 real and synthetic RGB images belonging to 10 classes. A total of 8736 synthetic images were created and used in the dataset following the methods previously discussed in (Dale, 2021). An additional 424 photographs of real object instances were combined with the synthetic images to create the RGBZFO dataset. This data was then randomly divided into training (5863 images), validation (1465 images), and testing (1832 images) sets with the ratio of synthetic to real images held constant across all splits. This dataset was chosen over open-source datasets to ensure that the pretrained object classifier backbone had never seen any of the images before, and only universal features were shared between the in-house dataset and the object classifier backbone latent space.

Creation of adversarial images

Three different kinds of adversarial images were generated for this study: image sets generated from the latent space of the VAE to calculate DALE metric statistics, adversarial images corrupted with normally distributed noise to conduct a black-box attack on the model, and adversarial images generated using the Fast Gradient Signed Method (FGSM) (Goodfellow et al., 2014) for a white-box attack on the model. The details of these datasets are discussed below.

Adversarial Image Sets

The RGBZFO training and validation splits were combined to create a dataset of 7328 images; these images were used to train a Variational Autoencoder (VAE) after the method presented by (Chollet & others, 2015). A second, identical VAE was trained using the entire 10,000 images from the Fashion-MNIST testing split.

The VAE encoder architectures consisted of 3 convolutional layers of dimensions {64, 64, 128|32, 32, 256|16, 16, 512} (reversed for the decoder) followed by a fully connected layer of 512 neurons, and coupled by a 32-dimensional latent space as shown in Fig. 1. The VAE loss function is

$$\mathcal{L} = \left(\frac{\alpha_{GM}}{N}\sum_{i}^{N} GM_i\right) + (\alpha_{BCE} \cdot BCE) + (\alpha_{KL} \cdot KL) \tag{2}$$

where the binary cross entropy (BCE) is used as the reconstruction error of the image, KL is the Kullback-Leibler divergence quantifying the information stored in the latent space (Asperti & Trentin, 2020), GM_i is the Gram Matrix comparing the outputs of convolutional layers as shown in supplementary Fig. S1 to quantify the style loss of the input and reconstructed image (Asperti & Trentin, 2020), and α is the corresponding weighting of each term in the loss function. Here, $\alpha_{GM} = 10^5$, $\alpha_{KL} = -0.5$, and $\alpha_{BCE} = 128^2$. The VAE was trained for 1000 epochs with a batch size of 32. The VAE image reconstructions of the original Fashion-MNIST and RGBZFO datasets were reserved for training the object classifier models.

The VAE latent space was traversed by randomly selecting two images from the VAE training dataset, generating their 32-dimensional vectors p_0, p_1 in the VAE latent space, linearly interpolating between these two vectors according to Eq. 1, and using these interpolated vectors to create a set of images from the VAE-decoder. There is no ground truth for the adversarial image sets. A total of 10,000 latent space pairs were sampled for the RGBZFO dataset, with 100 interpolated steps between them, for a total of $10,000 \times 100 =$ 1E6 images divided across 1E4 image sets.

For each pair of instances in the 10,000 image set, the distance in the latent space between the initial point p_0 and the final point p_1 was approximated by the L2 norm. However, since latent space manifolds are curved, their distances are not well represented by a Euclidean function (Arvanitidis et al., 2017). In the case of the VAE, the manifold has a Gaussian prior, and a random sample of distances traveled on the manifold also has a Gaussian distribution. A uniform representation of sampled distances was recovered by applying the cumulative distribution function of the normal distribution to transform the distances. The transformed distance data (normalized by the maximum distance traveled ϵ_{MAX}) is shown in Fig. 5 and discussed further in §. The transformed distance traveled is taken to be the perturbation budget ϵ for that set of 100 adversarial image instances. Accordingly, when points p_0 and p_1 are close, the set of adversarial instances generated between p_0 and p_1 will have a small step size and a perturbation budget $\epsilon/\epsilon_{MAX} < 1$.

Adversarial images for black-box attack

The simplest attack for a black box model is the addition of noise to an input image. Following the notation introduced in §, the maximum perturbation budget ϵ_{WGN} is taken to be the standard deviation σ of the normal distribution $\mathcal{N}(\mu = 0, \sigma)$. Accordingly, adversarial examples x' were created for each instance x in the training dataset partitions by adding normally distributed noise δ, where the strength of the attack was increased by allowing σ to take increasing values $[0.001, 0.01, 0.1, 0.2, 0.3, ..., 1.0]$. The strength of the attack was quantified using peak signal-to-noise ratio (PSNR) in the image domain, to allow for comparison across attacks.

Adversarial images for white-box attack

The Fast Gradient Signed Method (FGSM) leverages knowledge of the model gradients to increase the success of the attack (Goodfellow et al., 2014). For each instance x with ground truth label y in the training

datasets partitions, an adversarial instance x' was created so that

$$x' = x + \epsilon_{FGSM} \cdot \text{sign}\left(\nabla_x J(\theta, x, y)\right) \tag{3}$$

where θ represents the model weights, J is the loss function used to train the model, and ϵ_{FGSM} is the perturbation budget quantifying the strength of the attack. The FGSM attack is uniquely suited to this study as it maximizes the loss by perturbing the image in the direction of increasing gradient, that is, moving the instance towards a decision boundary. A model with simpler decision boundaries in the latent space is then expected to be more vulnerable to FGSM attacks (Kim et al., 2022). As for the black-box attack, the final strength of the attack was quantified using PSNR taken in the image domain.

Object Classifier Backbones and Training Details

Three object classifier backbones were selected from the TensorFlow Hub library: efficientnetv2-xl-21k (Tan et al., 2020), resnet_v2_50 (He et al., 2016), and inception_v3 (Szegedy et al., 2017) due to their popularity, differing architectures and sizes. While ResNet-50 and Inception Net are similar in size (26E6 and 24E6 parameters respectively), they have distinct architectures that could affect the decision boundary complexity. Inception Net uses the Inception Block, where features at multiple scales are calculated simultaneously then concatenated before being passed to the next layer (Szegedy et al., 2017), while ResNet features progress in size and the layers have residual connections (He et al., 2016). In contrast to both, Efficient Net has 10× the number of parameters at 208E6 (Tan et al., 2020). Its structure implements inverted residual blocks, which are related to the residual units in ResNet (Tan et al., 2020).

A dropout layer was added to each backbone to prevent over-fitting, followed by a fully-connected layer with an L2 regularizer; this is visualized in Fig. 1. The dropout layer and the fully-connected layer of each model were trained from scratch. The backbone was initialized with weights pretrained with the ILSVRC2012 "ImageNet" dataset (Deng et al., 2009) and frozen during training. The Keras Sparse Categorical Cross Entropy loss from logits was used during training. Each model trained for 5 epochs until the training and validation losses converged.

The softmax layer typically found in classification sub-architectures was omitted. The softmax layer assigns a probability that an instance belongs to a certain class, but there is no guarantee these probabilities resemble the class distribution (Corbière et al., 2019). Omitting the softmax layer avoids false intuition about the classification result of adversarial image sets and prevents the fully connected layer from updating with information from the softmax layer during backpropagation. The output of the FC layer therefore represents only the features extracted by the backbone.

Results and Discussion

Three experiments are presented in this section. First, the VAE latent space is shown to be a valid approximation of the object classifier's latent space by showing the high separability between the true positives (TP) and true negatives (TN) for each perceptron in the fully connected layer. Next, results from classifying adversarial image sets are presented along with DALE analysis. Finally, the robustness of the object classifiers is analyzed using black-box and white-box attacks, and these results are explained in the context of the DALE analysis.

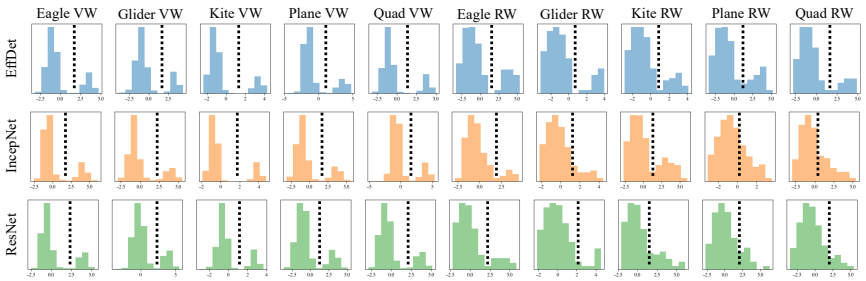


Figure 2: Histogram of the values produced by the perceptrons in the fully connected (FC) layer of the model for each image in the test dataset. (TOP) Efficient Det. (MIDDLE) Inception Net. (BOTTOM) ResNet-50. The distributions on the right of each dashed line represent the correct detections scored by the perceptron (true positives). The larger distributions to the left represent the FC values for other classes as interpreted by that particular node (true negatives).

Object classifier performance on reconstructed images

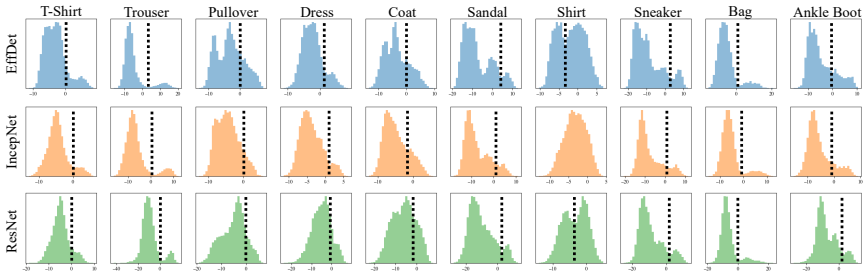


Figure 3: The output of the FC-layer perceptrons for models trained using Fashion-MNIST data. The separability between the true-positives to the right of the dashed line and true-negatives is less obvious in these distributions. This agrees with a slight decrease in F1 score performance for Fashion-MNIST compared to the RGBZFO dataset.

A correct classification occurred when the number of the perceptron containing the max value of the FC layer corresponded with the ground truth class number (one-hot label). The F1 scores are above 0.75 for all models and all classes, as shown in supplementary Fig. S2, and the networks performed comparably regarding misclassifications. There are some minor differences between network performance for individual classes; these are illustrated by the confusion matrices shown in supplementary Fig. S3. Because only the fully connected layer of the object classifier was trained using VAE reconstructions and no features in the model backbones were updated, these results show that the first goal of demonstrating that the VAE latent space successfully learned features present in each of the feature extraction backbones is accomplished.

In Fig. 2, the histogram of values from the object classifier's FC perceptron for each class in RGBZFO have been plotted for every item in the testing dataset for each backbone. These histograms allow us to interpret results from the adversarial data FC layers; a confidence threshold can be empirically identified as the FC perceptron value which separates the two distributions. The true-positive data points belonging to the named RGBZFO class titling each histogram are in a small distribution to the right of the threshold (corresponding to the maximum of the values for the FC perceptron), and the remaining true-negative data points not belonging to the named class are in a large distribution to the left of The threshold. There is a clear separation between the distributions, which corresponds to the high prediction accuracy of each network.

The analysis is repeated for models trained using Fashion-MNIST, and these results are presented in Fig. 3. A struggle to correctly classify some images is evidenced by a lack of separability between the true-positive and true-negative distributions for some perceptrons. Some FC perceptron outputs, such as the ones for "Bag' and "Trouser", are easily separable into high values for the true-positives, and low values for the true-negatives. Other classes in Fashion-MNIST are harder to distinguish, such as "Dress" and "Coat".

Strong separability between fully-connected (FC) layer perceptron values for true-positive and true-negative distributions indicates well-defined decision boundaries for the true-positive class, with true-negative instances of the class commonly producing a negative perceptron output value. The results in Figures 2 and 3 indicate a mix of well-defined decision boundaries and ambiguous boundaries. Accordingly, the overall complexity of the model decision boundaries is expected to vary with the features used to test the boundary.

Decision boundary complexity analysis via direct adversarial latent estimation (DALE)

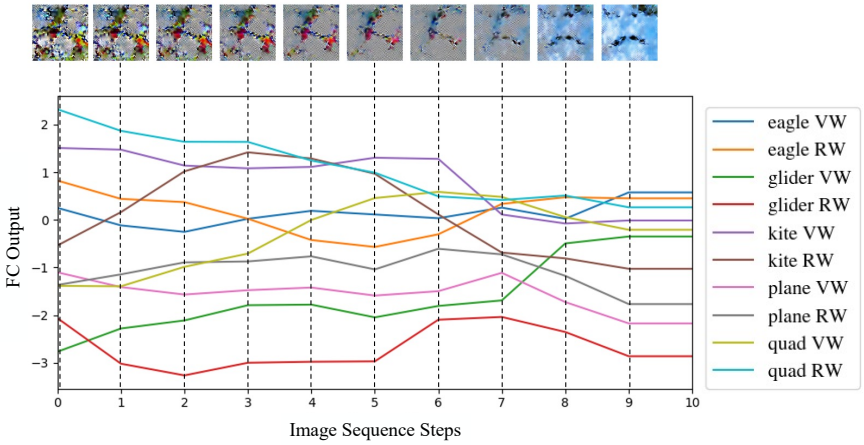


Figure 4: Object classifier fully connected (FC) layer output as a function of VAE latent space traversal adversarial images, shown at the top. Each image was classified using the Efficient Net FC layer. The FC perceptron returning the highest value is closest to the top of the plot. From image sequence step 0 to 3, the quadRW perceptron has the highest value, representing a classification of Real World Quadcopter for the images shown above the plot. At step 4, the top-most line changes from quadRW to kiteRW, representing a crossing of the decision boundary between the two classes in the latent space. The top-most line changes classes an additional 4 times, representing a total of five decision boundaries crossed for this image set.

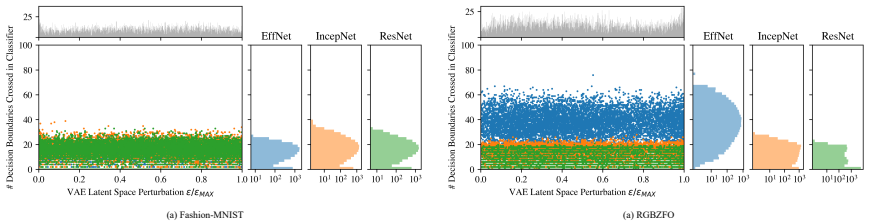


Figure 5: Direct adversarial latent estimation (DALE) distributions showing the number of times the object classifier class decision boundaries are crossed in an adversarial image set, plotted as a function of the normalized CDF of the perturbation ϵ/ϵ_{MAX} in the VAE latent space. A correlation between large perturbations and the number of decision boundaries crossed is not observed. Although 100 samples were taken for each perturbation budget ϵ/ϵ_{MAX}, the number of decision boundaries crossed is consistently lower than 80 for RGBZFO and 50 for Fashion-MNIST.

Decision boundary complexity can now be explored by evaluating how the output of the object classification fully connected (FC) layer changes during a set of ordered adversarial images sampled from the VAE latent space trained on the same data source using direct adversarial latent estimation (DALE) distributions. As shown in Fig. 4, an ordered set of VAE-generated latent space images are classified by the object classification model. The object classifier FC layer produces a value for each perceptron, with one perceptron per class. The example image set shown is intentionally presented using data that a human might struggle to classify to demonstrate the strong signal generated by the combined-class latent features; only ten instances are shown out of the one hundred instances in every adversarial set. The FC layer consistently identifies features that produced a positive value for a specific class node. In the figure shown, as the latent space is traversed, classification using the maximum value of the FC layer would return "RW Quadcopter" for the first three Image Sequence Steps, followed by "RW Kite", "VW Kite", "VW Quadcopter", "RW Quadcopter", "RW Eagle", and finally "VW Eagle". When adding in the empirically determined confidence threshold determined from the histograms in Fig. 2, FC Output > 1 would be considered a high-confidence prediction. This shows that subtle changes in the latent space potentially result in significant changes to classification results.

This can be confirmed by considering the distribution of FC output values for every image in the testing dataset. In Fig. 2, the distinct distributions in the abscissa range of $[1, 5]$ of each class plot represent the correct detections (true positives, TP) produced by the dataset. The larger distributions centered below zero represent the FC values for other classes (true negatives, TN). With this knowledge, the values shown in Fig. 2 can be better interpreted.

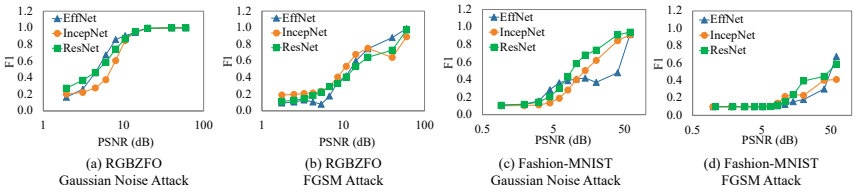
The complexity of the FC-layer decision boundaries is now quantifiable by plotting the number of times a decision boundary changes as a function of the maximum perturbation ϵ/ϵ_{MAX} to form a *direct adversarial latent estimated* (DALE) distribution. The number of times the classification changes in the object classifier's FC layer indicates the complexity of the object classifier's decision boundaries in the model's latent space. This is shown in Fig. 5 for the set of adversarial image sets generated from the RGBZFO-trained VAE and the Fashion-MNIST-trained VAE. The lack of dependence on ϵ/ϵ_{MAX} indicates that the distance defined by ϵ/ϵ_{MAX} is not correlated to the number of decision boundaries crossed.

The difference in observed complexities between the RGBZFO and the Fashion-MNIST data shown in Fig. 5 can be partially attributed to the complexity of the data. The RGBZFO dataset is feature-rich compared to Fashion-MNIST. Fashion-MNIST is exclusively grayscale, while RGBZFO has RGB channels. Fashion-MNIST was originally 28×28 pixels, while RGBZFO was originally 320×320 pixels. In addition to being gray-scale, rescaling the Fashion-MNIST images to $512 \times 512 \times 3$ for EfficientNet smooths the edges of the images; since the object classification backbone can be considered an "edge library", this means the Fashion-MNIST dataset had fewer features for the model to leverage when making a prediction.

However, in both (a) and (b) of Fig. 5 an average and maximum complexity emerges in terms of the number of times the decision boundary for classifying an image set was crossed; this is visualized by the distributions on the far right of each scatter plot. The mean and standard deviations of each DALE distribution are shown in Table 1.

Table 1: The mean (μ) and standard deviation (σ) of the DALE distributions shown in Fig. 5.

Model	Dataset	DALE ($\mu \pm \sigma$)
Efficient Net	Fashion-MNIST	13.38 ± 5.67
Inception V3	Fashion-MNIST	15.45 ± 6.56
ResNet 50	Fashion-MNIST	15.10 ± 6.21
Efficient Net	RGBZFO	37.13 ± 10.25
Inception V3	RGBZFO	11.79 ± 6.39
ResNet 50	RGBZFO	3.80 ± 5.88



(a) RGBZFO Gaussian Noise Attack (b) RGBZFO FGSM Attack (c) Fashion-MNIST Gaussian Noise Attack (d) Fashion-MNIST FGSM Attack

Figure 6: Effect of attacks on models trained with different data. A stronger attack (low PSNR) predictably decreases the F1 score of the model. However, the models trained with the Fashion-MNIST data are more impacted by the addition of the noise than models trained with the RGBZFO data. Furthermore, the robustness of the Efficient Net model relative to the other models is shown to depend on the original data used to generate the attack; this agrees with the DALE distributions shown in Fig. 5.

There is no correlation between ϵ/ϵ_{MAX} the distance traveled in the VAE latent space and the number of decision boundary crossings. A significant difference between the latent space structure of a VAE and the latent space structure of a generalized object classifier is that the structure of the VAE latent space manifold is predetermined (An & Cho, 2015), while the curvature of the object classifier latent space is unknown. However, we know that for the RGBZFO dataset, the image set generated from the VAE latent space adequately tests the object classifier latent space because for the 10,000 adversarial image sets sampled from the latent space, each 100 instances long, the maximum number of boundary decision changes found is 80 regardless of the distance traveled in the VAE latent space; the maximum number of changes it is possible to observe is 100. For the models trained with Fashion-MNIST data, this maximum complexity drops to 45.

Fig. 5 further indicates that for the RGBZFO dataset, the decision boundaries of the EfficientNet latent space are significantly more complicated than those in the InceptionNet and ResNet latent spaces. However, for the Fashion-MNIST dataset, the latent feature representations of all three architectures have comparable complexity, with the EfficientNet DALE distribution shifting slightly closer to zero than the InceptionNet and ResNet distributions.

This comparison clarifies that it is not the size of the model alone which accounts for decision boundary complexity. Although EfficientNet is ten times larger than InceptionNet and ResNet, the complexity measured is also heavily dependent on the data set used to train and probe the FC layer. The RGBZFO image dataset contains more complex features due to the three-channel images and the mixture of real and synthetic data, and adversarial image sets generated with this data indicated a higher number of boundary crossings than

the Fashion-MNIST dataset. Adversarial image sets generated using the VAE trained with Fashion-MNIST data did not stress the object classifier feature extraction backbones to the same amount.

This accomplishes the second goal of this work, and contributes a new method–*Direct Adversarial Latent Estimation* (DALE)–for measuring complexity of the object classifier latent space decision boundary using a test signal generated by sampling the surrogate VAE latent space. The DALE metric directly applies to testing the robustness of a model given an adversarial perturbation of the input data. What may be perceived as a small perturbation in image space (as shown by the image set in Fig. 4) or the feature space surrogate can still cause large effects in the latent space of the object detection backbone. This can result in numerous classifications for even a small amount of perturbation.

Robustness of object classifier to adversarial attacks

To test the robustness of the features in the object classifier latent space more directly, adversarial images prepared with normally-distributed noise and adversarial images prepared with noise generated using the Fast Gradient Signed Method (FGSM) were classified, and the average F1 score of all images was calculated as a function of the strength of the attack as measured by the peak signal-to-noise ratio (PSNR) in the image space. The results are presented in Fig. 6. As expected, a low PSNR reduced the F1 score of the model. However, the classification models trained with the Fashion-MNIST data were more impacted by the addition of noise than classification models trained with the RGBZFO data. This agrees with the prediction that models with less complex decision boundaries should be more vulnerable to FGSM attacks (Kim et al., 2022).

Furthermore, the robustness of the model to each attack directly correlates with the DALE analysis, as shown in Fig. 7. Each model's robustness to an attack is quantified by calculating the area under the PSNR-F1 curve (AUC) of Fig. 6, and plotting the DALE metric against this value for each model and each attack. For models trained using the RGBZFO data, the EfficientNet model maintains an F1 score greater than 0.8 for smaller PSNR values than either InceptionNet or ResNet; this corresponds to EfficientNet/RGBZFO scoring the highest in Table 1 with a DALE metric of 37.13 ± 10.25. The situation is reversed for models trained with Fashion-MNIST, where the EfficientNet/Fashion-MNIST model now has the lower F1 score for the weakest attack (high PSNR), and also the smallest DALE metric of the Fashion-MNIST models in Table 1 of 13.38 ± 5.67. This is clearly seen in Fig. 7, where for the blue marker representing the Efficient Net model trained on Fashion-MNIST consistently has the smallest AUC value of the attack and the lowest DALE score. The reverse is true for the Efficient Net models trained on RGBZFO; the blue marker is the farthest to the right of each attack type.

The Fashion-MNIST results shown in Figs. 5 and 7 suggests that the decision boundaries probed by the Fashion-MNIST data are less defined and less robust in all models tested. In Fig. 5, the maximum number of decision boundaries crossed was < 50 indicating a smoother (and more linear) latent space, and even very weak attacks (PSNR = 60 dB) could cause the model F1 score to drop significantly as shown in Fig. 6. In Fig. 7, the decreased AUC value range of 20 to 50 (compared to the RGBZFO AUC value range of 40 to 59) and the overlapping error bars demonstrate that the lack of observed robustness and the DALE metrics are correlated.

In comparison, the decision boundaries probed by the RGBZFO data are far more robust, with F1 scores remaining higher for stronger attacks than the Fashion-MNIST models. This is again shown by the non-overlapping error bars in Fig. 7, where Fig. 7 demonstrates the third goal of this work, to correlate the DALE metric with model robustness to adversarial attacks given an appropriate test signal and well-defined decision boundaries.

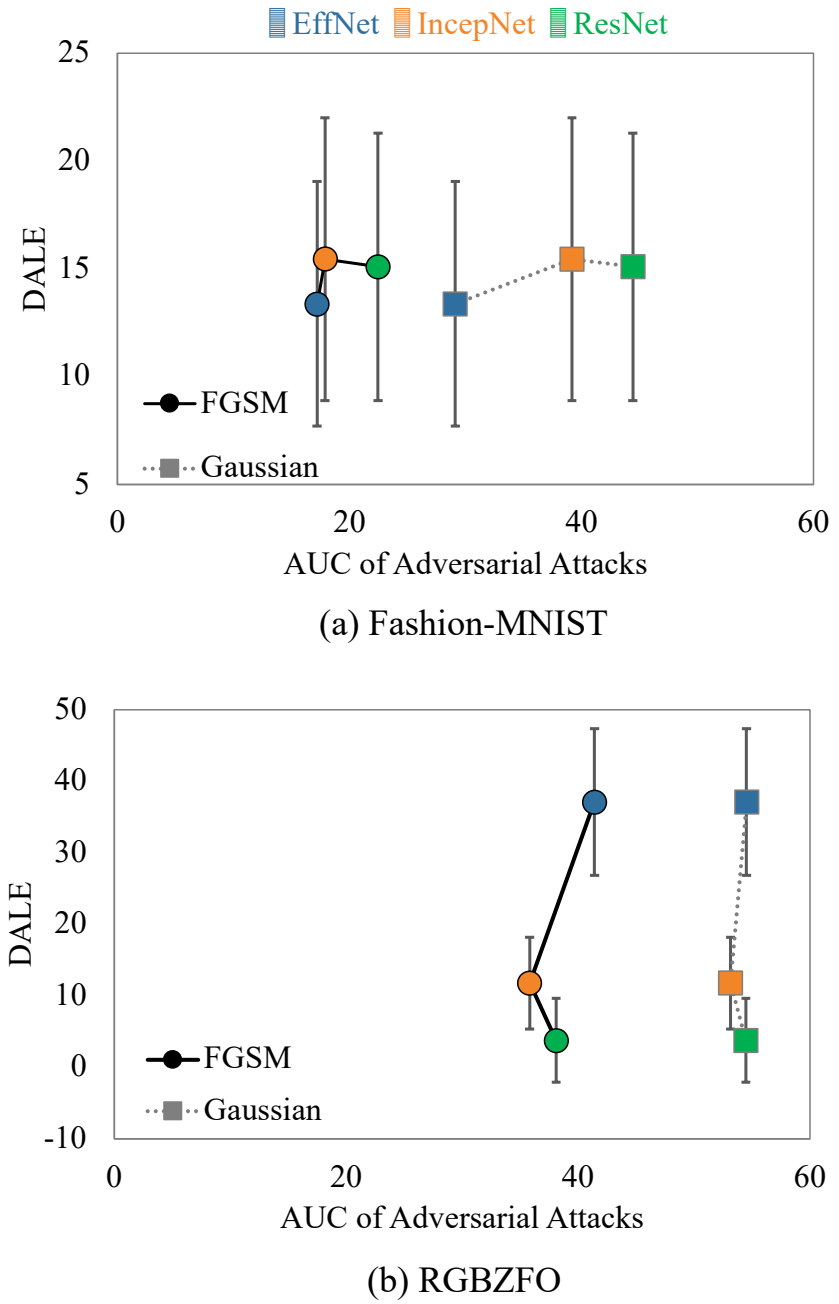


Figure 7: Correlation of the success of an adversarial attack vs. DALE metrics. (TOP) Fashion-MNIST and (BOTTOM) RGBZFO results for FGSM and Gaussian noise attacks. The Fashion-MNIST trained models demonstrate a lower robustness (smaller AUC value) that correlates with a decreased DALE score; the blue markers representing EffNet are to the lower left of each group of attacks, and the error bars plotted from the DALE-σ value are overlapping. The opposite case is present in the RGBZFO trained models, where the EffNet models are in the upper right of their attack type and the error bars are non-overlapping.

Conclusion

This work demonstrates three results. First, the use of a VAE as a surrogate model to study the latent space of black-box feature extraction backbones in the absence of any shared information during training, where the object classification head connected to the model correctly classifies image reconstructions generated by the VAE with high accuracy and precision (F1 score > 0.75 across all models and classes). This was achieved by leveraging VAE models trained on disjoint training datasets from the object classification backbones to create a surrogate latent space. Student-teacher models typically assume that information is communicated between the teacher model and the student model when the student model is trained. Here, this constraint is lifted, and the training results show that even though the object detection backbones were trained using ILSVRC2012 data and the VAE architectures were trained using either RGBZFO or Fashion-MNIST, the VAE latent space still contains features which are also found in the object detection dataset. This work then establishes the use of VAE models as surrogate latent spaces for black-box models without any shared information.

Second, a new metric *Direct Adversarial Latent Estimation* (DALE) for estimating the complexity of an object classification backbone decision boundaries is introduced, given an appropriate test signal of adversarial latent-generated data. The DALE distributions for six different models agree with and motivate the observed model performance degradation when the model is attacked using white-box and black-box noise methods. This provides the first method for probing the complexity of decision boundaries inside black-box models.

Third, we show that model robustness to perturbed input images is found to correlate with decision boundary complexity, and not with the degree of perturbation or model size. The DALE metric results suggest that model robustness could be viewed in terms of decision boundary complexity in the latent space, where more-complex decision boundaries (such as for the RGBZFO data) are more robust.

Acknowledgment

The authors would like to thank William Boler for his contributions to software design for these experiments, and Albert William and Wen Krogg for their efforts in generating synthetic data.

References

Establishing the rules for building trustworthy AI. (2019). *Nature Machine Intelligence*, *1*(6), 261–262.

Trustworthy AI. (2021). *Communications of the ACM*, *64*(10), 64–71.

Trustworthy artificial intelligence: a review. (2022). *ACM Computing Surveys (CSUR)*, *55*(2), 1–38.

A systematic review of robustness in deep learning for computer vision: Mind the gap?. (2021). *ArXiv Preprint ArXiv:2112.00639*.

Interpreting Adversarial Examples in Deep Learning: A Review. (2023). *ACM Computing Surveys*.

Latent space explanation by intervention. (2022). *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*, 679–687.

Physical Adversarial Attack meets Computer Vision: A Decade Survey. (2022). *ArXiv Preprint ArXiv:2209.15179*.

Sok: Certified robustness for deep neural networks. (2023). *2023 IEEE Symposium on Security and Privacy (SP)*, 1289–1310.

Rab: Provable robustness against backdoor attacks. (2023). *2023 IEEE Symposium on Security and Privacy (SP)*, 1311–1328.

A Framework Quantifying Trustworthiness of Supervised Machine and Deep Learning Models. (2023). *SafeAI2023: The AAAI's Workshop on Artificial Intelligence Safety*, 2938–2948.

Advances in adversarial attacks and defenses in computer vision: A survey. (2021). *IEEE Access*, *9*, 155161–155196.

Robustness of Deep Learning Models for Vision Tasks. (2023). *Applied Sciences*, *13*(7), 4422.

A survey on transferability of adversarial examples across deep neural networks. (2023). *ArXiv Preprint ArXiv:2310.17626*.

The space of transferable adversarial examples. (2017). *ArXiv Preprint ArXiv:1704.03453*.

Intriguing properties of adversarial examples. (2017). *ArXiv Preprint ArXiv:1711.02846*.

How many dimensions are required to find an adversarial example?. (2023). *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2352–2359.

Closer look at the transferability of adversarial examples: How they fool different models differently. (2023). *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1360–1368.

Canonical correlation analysis: An overview with application to learning methods. (2004). *Neural Computation*, *16*(12), 2639–2664.

Similarity of neural network representations revisited. (2019). *International Conference on Machine Learning*, 3519–3529.

Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. (2017). *Advances in Neural Information Processing Systems*, *30*.

Graph-Based Similarity of Neural Network Representations. (2021). *ArXiv Preprint ArXiv:2111.11165*.

Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. (2017).

Robust latent subspace learning for image classification. (2017). *IEEE Transactions on Neural Networks and Learning Systems*, *29*(6), 2502–2515. https://doi.org/10.1109/TNNLS.2017.2693221

Enforcing Sparsity on Latent Space for Robust and Explainable Representations. (2024). *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5282–5291.

Stability of feature selection algorithm: A review. (2022). *Journal of King Saud University-Computer and Information Sciences*, *34*(4), 1060–1073.

Exploring architectural ingredients of adversarially robust deep neural networks. (2021). *Advances in Neural Information Processing Systems*, *34*, 5545–5559.

A novel GAN-based network for unmasking of masked face. (2020). *IEEE Access*, *8*, 44276–44287.

Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation. (2019). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(4), 1254–1266.

Latent space cartography: Visual analysis of vector space embeddings. (2019). *Computer Graphics Forum*, *38*, 67–78.

Latent space visualization, characterization, and generation of diverse vocal communication signals. (2019). *BioRxiv*, 870311.

beta-vae: Learning basic visual concepts with a constrained variational framework. (2017). *International Conference on Learning Representations*.

Disentangling VAE. (2019). http://github.com/YannDubs/disentangling-vae/

Deep feature consistent variational autoencoder. (2017). *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1133–1141.

Adversarial perturbations of deep neural networks. (2016). *Perturbations, Optimization, and Statistics*, *311*(5). https://doi.org/10.7551/mitpress/10761.003.0012

Adversarial examples are not bugs, they are features. (2019). *Advances in Neural Information Processing Systems*, *32*.

Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. (2021). *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zoom in: An introduction to circuits. (2020). *Distill*, *5*(3), e00024–001.

Understanding robust learning through the lens of representation similarities. (2022). *Advances in Neural Information Processing Systems*, *35*, 34912–34925.

If you've trained one you've trained them all: inter-architecture similarity increases with robustness. (2022). *Uncertainty in Artificial Intelligence*, 928–937.

Representing deep neural networks latent space geometries with graphs. (2021). *Algorithms*, *14*(2), 39.

Synthetic data in machine learning for medicine and healthcare. (2021). *Nature Biomedical Engineering*, *5*(6), 493–497.

Rareplanes: Synthetic data takes flight. (2021). *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 207–217.

The 5th ai city challenge. (2021). *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4263–4273.

Vision meets robotics: The kitti dataset. (2013). *The International Journal of Robotics Research*, *32*(11), 1231–1237.

All patched up: effective integration of real and synthetic features into a single image for object detection. (2023). *2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–7. https://doi.org/10.1109/AIPR60534.2023.10440704

Synthetic medical images from dual generative adversarial networks. (2017). *ArXiv Preprint ArXiv:1709.01872*.

Texture synthesis with spatial generative adversarial networks. (2016). *ArXiv Preprint ArXiv:1611.08207*.

A survey on generative adversarial networks for imbalance problems in computer vision tasks. (2021). *Journal of Big Data*, *8*, 1–59.

3D object detection using virtual environment assisted deep network training. (2021). [PhD thesis]. Purdue University Graduate School.

Explaining and harnessing adversarial examples. (2014). *ArXiv Preprint ArXiv:1412.6572*.

Keras. (2015). https://keras.io

Balancing reconstruction error and kullback-leibler divergence in variational autoencoders. (2020). *IEEE Access*, *8*, 199440–199448.

Latent space oddity: on the curvature of deep generative models. (2017). *ArXiv Preprint ArXiv:1710.11379*.

Curved representation space of vision transformers. (2022). *ArXiv Preprint ArXiv:2210.05742*.

Efficientdet: Scalable and efficient object detection. (2020). *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10781–10790.

Deep residual learning for image recognition. (2016). *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Inception-v4, inception-resnet and the impact of residual connections on learning. (2017). *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*.

ImageNet: A large-scale hierarchical image database. (2009). *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Addressing failure prediction by learning model confidence. (2019). *Advances in Neural Information Processing Systems*, *32*.

Variational autoencoder based anomaly detection using reconstruction probability. (2015). *Special Lecture on IE*, *2*(1), 1–18.