Applications of GANs in Explainable AI (XAI): A Survey

Dheeraj Varma Chittari
Technical University of Kaiserslautern
Email: macharav@rhrk.uni-kl.de

Abstract*—Nowadays, Machine Learning and Artificial Intelligence models are becoming more sophisticated and complex to solve real-world challenging problems. As the complexity of the model increases, the interpretability of such systems decreases greatly and therefore they are known as 'black-box' models. The need to understand and interpret how the model behaves in different situations is highly crucial to employing complex AI models in safety-critical systems such as autonomous driving, security surveillance, credit, and banking industries. In the healthcare domain, complex AI models are employed to perform robotic surgeries and the need to understand how complex AI model works is very important in explaining the functionality of the robotic device. Therefore the need to effectively and efficiently understand a black-box model is highly necessary. Explainable AI (XAI) is a branch of Artificial Intelligence that deals with explaining the decisions of a complex black-box model. In the last ten years, there has been tremendous progress in the field of Explainable AI (XAI). Many techniques and methods have been proposed to improve the interpretability of a complex AI system. One such technique is employing Generative Adversarial Networks in the XAI domain. In this paper, we will present how the advancements in Generative Adversarial Networks (GANs) can be used to further improve the explainability of black-box models. Generative Adversarial Networks are a class of deep neural networks that can generate realistic samples from input distribution. We identify the state-of-the-art methods that use GAN or apply properties of GANs to the existing XAI techniques to further improve the interpretability of a complex AI system. We analyze each method's use case and working principle. Finally, we present an overview of each method and its advantages, limitations and propose important characteristics and future work.

1. Introduction

Generative Adversarial Networks (GANs) belong to the family of generative models in the deep learning domain that were introduced by Ian J. Goodfellow and his team [1] in 2014. The primary purpose of GANs is to learn the input data distribution and generate new samples. Because of their ability to generate realistic data, generative adversarial networks are employed in various domains such as image generation [23], [50], video generation [58], [59], text generation [2], [3], voice generation [60]. Consider the AOT-GAN [50] which is an image generation GAN method. It is used for the photo inpainting technique of generating a part of the image that is missing from the original image. GANs can generate visually distinguishable counterfactual examples to a given input. GANs can produce multiple images for a particular scene that differ from each other by a single or a subset of feature values that enable the images to form a certain hierarchical order among themselves such as low memorable image to high memorable image or low aesthetic to high aesthetic image.

Explainable AI also known as interpretable AI aims at providing human interpretable reasoning for complex AI model predictions. We have seen in recent years, Deep Neural Network (DNN) models have been achieving state-of-the-art performance on various Machine Learning and Artificial Intelligence tasks due to their greater model complexities employing billions of parameters. As the model grows larger with millions of parameters, the interpretability of the model reduces significantly. Thus, it is difficult to understand and interpret how the model works for a given input. Therefore the need for explaining the decisions made by the complex model on the given input is critical in earning trust and employing these DNN models. A myriad of XAI methods and techniques exists to explain a black-box model. Attribution methods are very popular post-hoc XAI methods. Most of the attribution methods generate feature perturbation for the input to explain the classifier's decision. They can be categorised as model-agnostic vs model-dependent methods or global-explanation vs local-explanation methods. Generative models are highly sophisticated networks that can produce high-quality realistic images with slightly altered features. Therefore employing generative models in XAI (attribution) methods will enhance the interpretability of the black-box classifiers. Currently, there is no literature survey in the field of applications of Generative Adversarial Networks (GANs) to Explainable AI (XAI) attribution methods.

In this paper, we present various applications of Generative Adversarial Networks to Explainable AI (XAI) methods. We analyze each application's use case, their working principle, qualitative and quantitative evaluations, and finally discuss their advantages and limitations. We identify and present common characteristics among all the XAI methods that use GANs. The remainder of the paper is structured with the following sections: Background, Methods and Applications, Discussion, and Conclusion.

2. Background

Here, we briefly discuss the early developments, progress, and present state of XAI systems as well Generative models separately.

2.1. XAI Methods

According to the comprehensive study presented by Mueller et al. [11] explanation systems can be categorized into three generations: In the first generation the explanations were used to incorporate the domain knowledge from an expert into low-level machines languages, in the second generation they were mainly used for human-computer interactions to enhance human cognition, in the third generation the explanation produced are used to solve black-box models for undesired biases, unethical uses, lack of transparencies. In this paper, we refer to the third-generation systems when we say XAI methods. There have been many methods proposed to interpret the decisions of a complex AI model.

Attribution Methods They are the algorithms that explain the importance of features in an input that is responsible for the black-box classifier's decision. In general, each input feature gets a normalized score of [0-1] where a score of 1 means that the feature is of high relevance to the classifier's decision and a score of 0 suggests that the feature is not relevant. Most of the attribution methods are used to interpret the post-hoc classifiers meaning they can only explain the classifier after its decision has been made on the input.
Attribution methods can be broadly classified into:

- Global Agnostic Model: The algorithms describe the average behavior of an entire AI system. Examples are Partial Dependence Plot (PDP), accumulated local effects and global surrogate models.
- Local Agnostic Model: These methods explain the individual input predictions. Examples are Local Interpretable Model-Agnostic Explanations (LIME) and SHAP.

Based on pixel value manipulations [12], attribution methods can also be classified as:

- Gradient-based: This method computes the gradient of the prediction with respect to the input features. Grad-CAM and DeconvNet are the examples of gradient-based approach.
- Perturbation-based: In this method, a certain amount of pixels of the input are modified to generate perturbed examples similar to the original input. The generated examples along with the input are passed to the attribution methods to explain the decisions based on a black-box prediction. Local Interpretable Model-Agnostic Explanations (LIME) and SHAP are some examples.

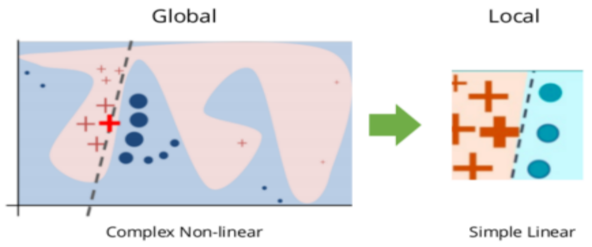

Figure 1. Perturbed samples are created around the input using the original complex model. The input along with the perturbed samples are trained on a simple linear machine learning model. [42]

2.1.1. SHAP.

SHAP attribution method is based on *shapley* values from the game theory domain. In many cases, a feature present in the input might be dependent on other features for its contribution towards the classifier's prediction. In this approach, each individual feature's contribution to the final black-box classifier's output is computed with and without taking into account the presence of other features [16]. This has a very high time complexity in the order of exponential time. However certain strategies are proposed to choose the Shapley values wisely by iterating only over local feature areas, approximating importance using samples from the training dataset [17].

2.1.2. LIME.

Local Interpretable Model-Agnostic Explanations [15] (LIME) is used to interpret the individual input samples prediction by the complex black-box model through an interpretable simple model as seen in figure 1. Select the instance of interest for which the interpretation from the black-box model is required. Then perturb those samples that are relatively close to the instance of interest and get their black-box predictions. Assign weights to the perturbed samples according to their proximity with respect to the instance of interest sample. Train a simple model on the newly obtained perturbed dataset. Explain the prediction of the instance of interest with respect to the simple model. Thus the newly learned model can have a good estimate of the local behavior with respect to the original black-box model but it cannot give a reliable global explanation.

2.1.3. Integrated Gradient.

Integrated Gradient (IG) computes an integral of the gradient from a baseline image to the input image [31]. IG can be defined as follows,

$$IG_i = (x_i - \overline{x}_i) \int_{\alpha=0}^{1} \partial S_x(\overline{x} + \alpha(x - \overline{x}))\partial\alpha \tag{1}$$

Zero input is usually considered as the baseline input. The selected baseline input is crucial for explaining one-vs-one attribution for multi-class classifiers. The zero baseline, uniform distribution baselines, focus highly on the regions where the input sample feature values are different from the baseline image values and neglect those features whose

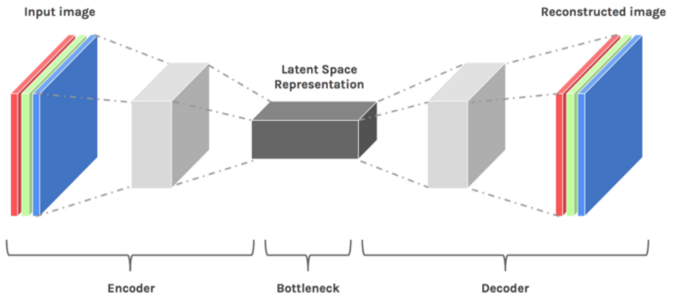

Figure 2. Auto-encoder schema with latent space representation. [18]

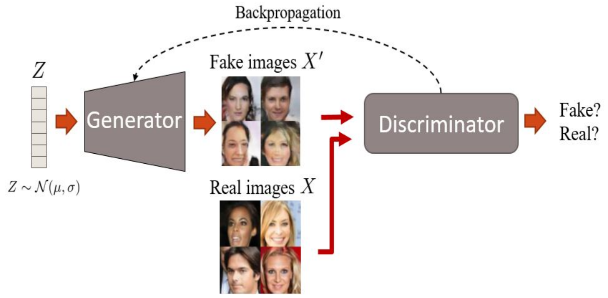

Figure 3. High level overview of GAN training. The generator takes an input from random latent space and generate images, that are fed into the discriminator network to identify real/fake. The discriminator is also trained with the actual input data. [19]

values are close to that of baseline image values. Attribution methods like Integrated Gradient [48], DeepSHAP [16] and Expected Gradient [49] makes use of baseline image to explain the classifier.

2.1.4. Partial Dependence Plot. Partial Dependence Plot (PD plot) is a global method that shows the average effect of a particular feature over all the input instances on the classification but not all the features effect for a specific input. PD plot generates the marginal effect of two features on the predicted outcome of a model [14]. This gives a clear interpretation by giving the changes in prediction due to changes in particular features. But, assumes features under the plot are not correlated with the features that are remaining. In real-world scenarios, this assumption is not true.

2.2. Generative Models

Generative models belong to the family of deep neural networks where the model tries to learn patterns and representations from the given input data distribution. Consider a generative model that has been trained on the cat's dataset. Now the model has learned the representations of various types of cats, thus it can produce new images that resemble the original data using the learned representations (z). The representations are also known as latent vectors or latent space.

The latent space is the hidden representation of all the input data in a compressed manner [18] which captures all the important input features and get rid of noise and unimportant information. From the figure 2, the middle block is considered as the latent representation. The data with similar features clusters close together in latent representations. Using this representation, the model can now generate new data that is close to the original input space. Because of their ability to generate an entirely new set of realistic data, these models have been employed in many real-world applications in a wide variety of domains.

2.2.1. Generative Adversarial Networks. GANs consists of two Deep Neural Networks (DNN), where one is a generator and another is a discriminator that are trained together. The generator's job is to produce new images from input space distribution. The discriminator is trained with the original image set and its job is to identify whether the images generated by the generator are real or fake. Over time both the generator's and discriminator's performance improves and there comes a point when the generator is able to trick the discriminator continuously by generating an image that is hard to differentiate for a discriminator between a real and fake image and eventually classifies it as real. At this point, the generator has perfectly learned to represent the original input space. A common analogy for GANs can be thought of as a counterfeiter (generator) and an examiner (discriminator) playing a game where the counterfeiter makes a collection of fake items and the examiner aims to identify if they are real or not. This process continues and after a certain period of time, the examiner and counterfeiter both get well at their processes, with the final goal being that the counterfeiter is so efficient in cheating that the fake items can pass all the checks by the examiner. In game theory, GAN model convergence is also known as Nash equilibrium.

$$min_G max_D V(D,G) = E_{x \in pdata(x)} |\log D(x)| \\ + E_{z \in pz(z)} |\log(1 - D(G(z)))|$$

GANs try to solve the objective function as defined in the above equation. The generator has one loss function whereas the discriminator has two loss functions one is for training on the original dataset another is for classifying an image produced by the generator as a real or fake one. However, GANs do not always produce desired results because they require a large amount of training data which may not be available in all the domains. Also, GANs training may suffer from *mode collapse*. A situation where the learning of discriminator gets trapped in local minima and the generator tries to produce new samples from a small distribution space that always passes the discriminator as a real one. In this case, the generator has not learned to represent the entire input dataset distribution. Another issue could be *non-convergence* where the model parameters oscillate continuously and may not converge.

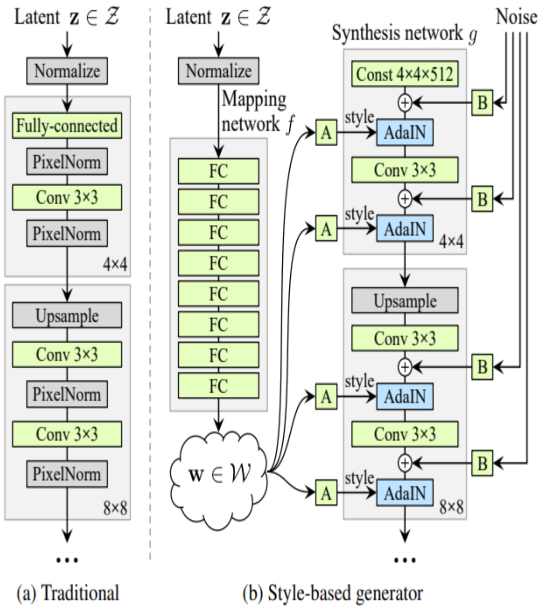

Figure 4. Comparison of Traditional GAN generator architecture with Style-based GAN generator. [24]

2.2.2. StyleGAN. The authors of StyleGAN [24] added progressive growth of GANs to the generator part. In figure 4, the latent vectors are immediately passed into the blocks after normalization, but in StyleGAN architecture the latent vectors are passed through the fully connected network then the outputs are transformed. They are passed to the individual blocks by adaptive instance normalization (AdaIN) along with random noise (B) as in figure 4. Moreover, the addition of up-sampling or down-sampling blocks, replacing a constant tensor in place of traditional inputs and mixing up of regularization in the network improved the FID score of generated images. Style mixing or regularization gave higher control over generated image styles and high-level structures such as hairstyles and sunglasses.
StyleGAN2 is introduced to mainly overcome the quality issues that occurred in StyleGAN. In the improved version of StyleGAN, the Adaptive Instance Normalization (AdaIn) is replaced with the modulations and normalizations. Weights are modified with the style and weight demodulation operation is introduced to have much control over generated images. Also, StyleGAN2 has disentangled latent space called *StyleSpace* that can capture the input features in a discrete and disjoint manner. Thus StyleSpace can be used to extract individual features.

2.2.3. Image to Image translations. It is used for creating mappings among different domains of data. Although GANs have been used in various domains, Image to Image translation in computer vision has tremendously advanced by employing GANs. Some of the variants of GANs are, CycleGAN [21] which is used in image style transfer. SRGAN [23] was developed for getting a super-resolution image. Initially, StyleGAN was introduced for face generations but later it is modified to improve the quality of images and also incorporated disentangled latent space also known as StyleSpace which can be used to get individual input features. StarGAN [20] has a standalone class discriminator and deterministic mapping that preserves the styles of the translated image. There are lot of Generative Adversarial Network variations available today. They can be classified based on architecture [26], based on latent space [27], [28], based on loss types [25], [30], based on regularization methods [29].

3. Methods and Applications

In this section, we present various methods and approaches to enhance interpretability of a black-box model with the help of GANs. We discuss the problem that each method addresses, their working principle, followed by evaluation and their advantages and limitations.

3.1. GANMEX: One-vs-One Attributions using GAN-based Model Explainability

Nowadays, feature perturbation based attribution methods are being employed frequently to identify key features that are responsible for classifier's decision. Most of the attribution methods use a reference baseline input for feature perturbation. Selection of a proper baseline image plays an important role in model explainability for attribution methods that focus on one-vs-one interpretability, which means explaining why an input belongs to a target class but not the other class. In general, naive approaches such as Minimum Distance Training Sample (MDTS), zero or average pixel value (for images) methods are considered while generating a baseline image. The attribution methods for these static baselines only highlight the areas in the input that differ the most from the baseline images and ignore those regions that have feature values close to that of baseline image pixel values. They fail to explain the model accurately.

The authors of GANMEX [10] proposed some assumptions to improve the attribution method's interpretability by generating an accurate baseline input. The assumptions are:

- The baseline has to be realistic in nature.
- The baseline must belong to the target class.
- The baseline has to be close to the original input.

The baseline generated by GANMEX produced superior results for attribution based model explainability when compared to zero, MDTS approaches.

3.1.1. Working Principle. Generative Adversarial Networks (GANs) are highly sophisticated deep neural networks that are capable of generating realistic images. GANMEX [10] uses a variant of GAN called STARGAN [20]

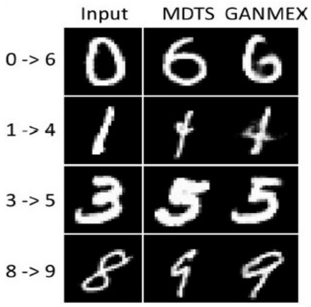

Figure 5. The image shows comparison of baselines generated by Minimum Distance Training Sample (MDTS) and GANMEX methods for MNIST dataset [41] for explaining why an input belongs to class 0 but not 6. The same explanation is for other input (1,3,8) and targets (4,5,9) respectively. [10]

to generate a baseline image that satisfies the above three assumption. Pretrained classifier can be incorporated as a discriminator in the STARGAN architecture and also it can preserve the styles of the translated images. The training of GANMEX is similar to that of STARGAN training, but GANMEX tries to solve the following objective function defined as:

$$B_{c_t}(x) = argmin_{\overline{x} \in R^N} (|x - \overline{x}| - \log R(\overline{x}) - \log S_{c_t}(\overline{x})) \tag{2}$$

where x is input image, $\overline{x}$ is generated baseline image, C_t is target class, $|x - \overline{x}|$ is similarity loss to ensure generated sample is close to the input, R is the probability of generated image being realistic and S is the probability of $\overline{x}$ being present in the target class. Therefore, $B_{c_t}(x)$ represents a generator producing a realistic baseline image belonging to a target class c_t by taking an input image (x). The generated samples are used as baseline images for attribution methods.

3.1.2. Qualitative and quantitative Evaluation. The qualitative evaluation is based on human consensus, surveys and inferences. Quantitative evaluation specifies a metric to compare different methods performance. There is no defined standard metrics available to compare various attribution methods in XAI [33]. This is indeed a very active research area where only a few papers and literature have been published so far.

The baselines generated by GANMEX for MNIST [41] dataset satisfy three properties. The baseline images are realistic, close to the input and belongs to target class, whereas the baselines from Minimum Distance Training Sample (MDTS) fails to satisfy all three properties. In figure 5, the baseline images produced by MDTS are not close to the input however GANMEX baseline images are close to the input as it can be observed for the example of why an input is classified as 0 and not 6, GANMEX generated sample has straight curve on the left side but MDTS sample has an oval shape.

The figure 6, indicates the Integrated Gradients and DeepLIFT attribution method's saliency maps produced by Zero, MDTS, GANMEX baseline images for a given input image. Consider the saliency map one-vs-one explanation generated by Integrated Gradients attribution method for why an input is 0 but not 6 from SVHN [57] dataset in figure 6 can be interpreted as follows, GANMEX produced sample image only highlights (red or blue region) the upper half of the region that are responsible for the classification of image as class 0 but not class 6. We know that, the lower region of number 0 and number 6 is almost same and modifying this region is invariant to classifier's decision. However, the attribution methods that uses Zero, MDTS baselines also highlighted the lower unnecessary region. The similar explanations can be done for other examples, where attribution methods using GANMEX generated baseline images highlight only necessary portions in the image that are important to classifier's decision.

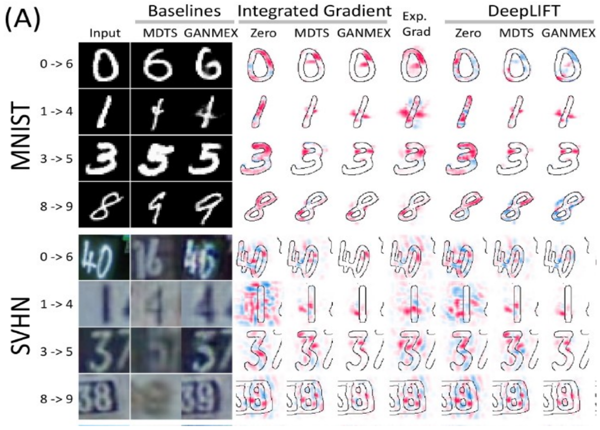

Figure 6. Comparison of Zero, MDTS, GANMEX baseline method's saliency maps generated by Integrated Gradients and DeepLIFT attribution method's for MNIST [41] and SVHN [57] datasets. [10]

	Zero	MDTS	GANMEX
Different object	0.711	0.850	0.459
Different scene	2.440	1.254	1.027
overall	1.591	0.852	0.747

Table 1. Benchmarking Attribution methods (BAM) based on projecting foreground objects onto background scenes. It shows the inverse localisation metric for BAM datasets. Here, the lower the better. [10]

Benchmarking Attribution Methods (BAM) dataset [50] is used to evaluate attribution methods efficiency using different baseline approaches. BAM was designed to have a set of background scenarios and foreground objects. This allows the ground truth information of foreground or background areas to benchmark attribution methods. Therefore, original and target class image sharing the same background but having a different object, then attribution is expected to highlight the foreground object region. If original and target images have the same object but different background then saliency map highlights the background region. From table (1), GANMEX performs better compared to other baseline approaches such as Zero, MDTS for Integrated Gradient attribution method on BAM inverse localisation metric (the

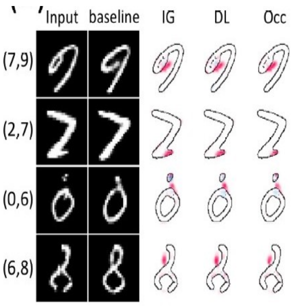

Figure 7. The figure shows mis-classification analysis of inputs from MNIST dataset [41] by a certain classifier. The first row contains why an input belongs to class 7 but not class 9. The GANMEX generated baseline image is used in Integrated Gradient (IG) [31], DeepLIFT, Occlusion attribution methods to explain the reason for wrongly classified inputs. [10]

lower the value the better the model) [10].

Both qualitative or quantitatively, it is evident that the baseline images generated by GANMEX produce more accurate attribution based one-vs-one explanations.

3.1.3. Advantages and limitations. GANMEX is the first of its kind to use a realistic baseline sample generated using GANs for one-vs-one explanations generated by feature perturbation based attribution methods for a classifier's decision. This approach can be employed in other domains such as Natural language Processing (NLP) where deleting an input feature is unclear. Another advantage of GANMEX is the method is model agnostic and therefore can used for any attribution method that uses baseline model for feature perturbations [15], [31]. GANMEX provides target class baseline images that are highly useful for one-vs-one explanations and analysis for mis-classified inputs. Figure 7, shows the mis-classification analysis of inputs from MNIST dataset [41]. The images in each row shows why an input belongs to class original but not class target by producing a realistic baseline image which is used by attribution method to produce saliency maps.

The main limitation for GANMEX comes directly from the drawback of using GANs that is requiring large amount of training data to converge and produce a realistic baseline image. This method is complex and time consuming compared to zero or uniform baseline approaches. GANMEX is applicable to those attribution methods that use baseline image to produce model explanations. Also, the authors did not provide a method to handle one-vs-all explainability that why an input belongs to a particular class but not all others. One of the assumptions of GANMEX is that the baseline image must belong to the target class but for one-vs-all explanations, the selection of baseline that belongs to all other classes is a open question.

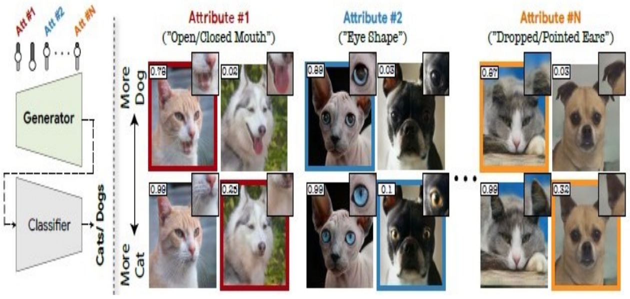

Figure 8. Attributes and their counterfactual examples generated by StylEx [33].The highlighted boundary box images represent counterfactual examples in each column for the attribute defined above. The change in probabilities caused by perturbing a particular attribute can be viewed in top left corner of the image.

3.2. Explaining in Style: Training a GAN to explain a classifier in StyleSpace (StylEx)

Saliency maps (heat maps) are the most common visual form of interpretations for a complex black-box model. They emphasize those areas in the inputs that are responsible for a classifier to make its decision. However, such maps cannot be applied if the input features are not spatially located such as size, color. Also saliency maps do not provide the direction of impact of the features on the classifier.

Counterfactual examples overcome these downsides. Counterfactual explanation can be defined as follows, if the input value x is modified to $\overline{x}$, then the classifiers decision would change from y to $\overline{y}$ where the difference between original input x and altered sample $\overline{x}$ is easy to interpret for humans. Consider a model trained on dog and cat samples. A counterfactual explanation could be if the pupils are made bigger then the probability of predicting the image as cat would decrease by some x percent. From the figure 8, the highlighted images are counterfactual images generated by StylEx [33] method where images in the same column have different classification probability based on change in the single attribute value.

The perturbations for generating adversarial examples must be visually detectable and human interpretable. Though adversarial perturbations change the classifiers decision on the input images, they cannot be considered as counterfactual examples. Because the noise added by adversarial examples are not recognizable by humans. A key benefit of using counterfactual example is it provides explanations to each input sample indicating which part of the sample is important towards classification and also how they can be modified to generate a different outcome. Recently, counterfactual explanations are becoming popular and gaining importance [45], [46], [47] in explainable AI. A smooth transformation of a set of input features from the original class to the target class will not isolate each individual feature importance and fail to explain each feature importance accurately. To explain the feature importance, we need a control over each individual feature in the input. With

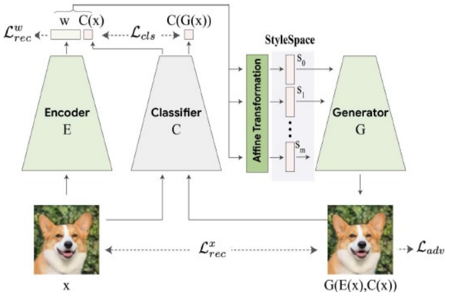

Figure 9. StylEx architecture. Here encoder, generator and discriminator are trained together. The output of the classifier is combined with the affine transformations of the encoder to ensure that the StyleSpace contains classifier related attributes. This acts as an input to the generator which then generates new images close to the input. [33]

this approach, we will be able to identify in which direction and how much an input feature has to be altered so as to change the models decision. Such counterfactual examples can be generated using StylEx approach. StylEx [33] aims to achieve this with the help of StyleGAN2 [49], a variant of Generative Adversarial Network (GAN).

3.2.1. Working Principle. StyleGAN2 [49] contains disentangled latent (hidden) representations of input data features which is also known as StyleSpace. StylEx uses the information present in the StyleSpace to derive individual attribute effects on the model's prediction. To achieve this, the authors of StylEx [33] proposed these steps: (i) Train the StyleGAN by incorporating a classifier into the GAN training to capture classifier related features into StyleSpace. (ii) Then the StyleSpace is explored for a precise set of features that affect the classifier's decision. As shown in Figure 9, the output of the encoder, that is latent vectors of an input image, is combined with the classifier's output which is then passed as an affine transformation into the generator for the StyleSpace to capture classifier specific attributes.

$$L_{cls} = D_{KL}|C(x^{'}) - C(x)| \tag{3}$$

To ensure that the generated images have same properties of the input images, a KL divergence (classifier loss) is applied to the generated and input images as shown in equation (3).

After generating the StyleSpace containing classifier related attributes, a naive approach (AttFind) is followed to identify the effect of individual features on classifier's prediction and select only top-k features among all. For each image calculate the effect of modifying each style coordinate (latent representation of feature in StyleSpace) with respect to classifiers output. Select top-k style coordinates that change the classifier's prediction upon changing the coordinates value in a certain direction with a certain magnitude. The top-3 attributes from figure 10 are 'Full Smile', 'Skin smoothing', 'Face width'. These three features are in descending order of importance to the classifier's decision of perceiving an image as younger or older. It is because altering these features within a threshold distance in the latent space (StyleSpace) has increased the classifiers decision of identifying the person as younger from 0.18 to 0.59 for the full smile attribute, from 0.18 to 0.41 for the skin smoothing attribute, and from 0.18 to 0.39 for the face width attribute. Thus, StylEx aims to explain each feature's importance per individual input image by creating counterfactual examples.

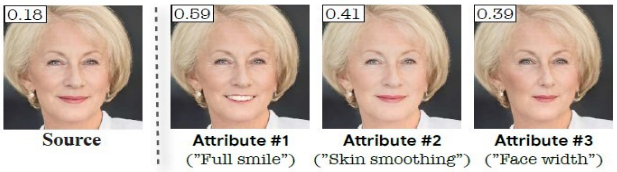

Figure 10. StylEx captures top-3 attributes when querying for making the person look younger. On top right corner of each image, we can see that the increase in probabilities compared to the source. These StylEx identified attributes also correlate with the human cognitive perception of what make a person look younger. [33]

	Wu et al. [34]	StylEx AttFind
Perceived Gender	0.783(±0.186)	0.96(±0.047)
Perceived Age	0.85(±0.095)	0.983(±0.037)
Plants	0.91(±0.081)	0.916(±0.068)

Table 2. User Study on visual coherence and distinctness for attributes of a certain classifier extract using Wu et al. [34] and StylEx AttFind methods. [33]

3.2.2. Qualitative and quantitative Evaluation. For qualitative evaluations, the authors of StylEx [33] considered visual coherence and distinctness as a quality measure for the StylEx extracted features. Wu et al. [34] proposed that features can be extracted from StyleSpace using normalised difference between the coordinate values on each labels.

From table 2, it is evident that the user studies shows the top attributes produced by StyleEx AttFind are visually distinct and coherent. In figure 8, changing the attribute values from open mouth to closed, pointed ears to dropped ears in cats vs dogs dataset, the probability have changed immensely to the opposite class and the identified attributes are visually perceptible to humans.

For quantitative evaluation, since StylEx is the first of its kind to use GANs to produce counterfactual examples to analyse multi-attribute classifiers, there have not been any baseline comparisons to estimate its efficiency quantitatively [33].

3.2.3. Advantages and Limitations. StylEx [33] provides control over each key feature. The classifier's decision can be analyzed by changing the direction and magnitude of a particular feature from the given input image. It can be used to identify model biases that can arise from biased datasets or training. If the top attributes identified by this method do not correspond to reality, then it can also be viewed as

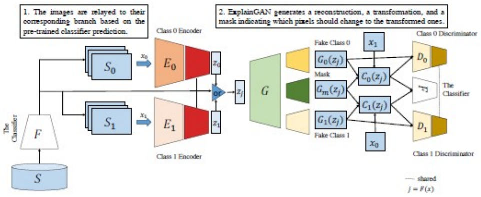

Figure 11. Architecture of ExplainGAN having two class specific encoders (E), a generator (G) and two class specific discriminators (D). [35]

the classifier being biased. Another advantage is that this method provides counterfactual explanations for multi-class classifiers and gives image specific explanations for a given classifier. The attributes identified by this model correspond to human visual cognition.

The limitations of this method are sometimes the images selected from the same class may not have the same priority of attributes. The top-k attributes identified by StylEx may be different for different images that belong to the same class, in such scenarios, the classifier's top attributes for a particular class may not be identified immediately and need to analyze each image individually which may be a time-consuming process is the major drawback of this method.

3.3. ExplainGAN

ExplainGAN [35] comes under perturbation-based XAI systems where it produces human perceptible decision boundary crossing examples. Traditional methods for a model interpretation such as saliency maps may fail to visually interpret changes required in the original input to alter the model's decision. Adversarial examples do not generate human perceptible explanations. ExplainGAN uses a generative model to produce visually identifiable boundary crossing transformations for the input images.

3.3.1. Working Principle. ExplainGAN [35] takes an input image and a binary classifier to generate an opposite class (boundary crossing) transformed image and a binary mask responsible for the change in a classifier's decision. The generated image is similar to the input image except for a visually identifiable difference, that it belongs to a different class with respect to the input. The binary mask image indicates which pixels are changed in the input to generate the transformed image.

The figure 11, describes the architecture of ExplainGAN, where we have two class specific encoders (E), a generator (G) and two class specific discriminators (D). The input is passed to the corresponding encoder which produces latent vectors (z) based on its class. The generator takes these latent vectors and produce a reconstructed image, transformed image (opposite class label image from the original image) and a binary mask indicating where the changes have happened. In order to create an explainable image transformation, ExplainGAN uses a 'prior' loss function. This comprises of *smoothness loss* that encourages localization of masks by restricting its effect to only the certain surrounding area pixels, *entropy loss* makes the mask to be binary, *count loss* which contains the prior knowledge regarding the approximate number of pixels to change.

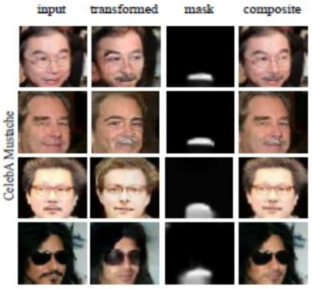

Figure 12. Qualitative analysis of ExplainGAN over celebrity dataset with a binary classification of having/not-having a mustache. [35]

3.3.2. Qualitative and quantitative Evaluation. The qualitative explanation shows the semantically consistent, human identifiable and highly localised changes occurred to the generated image.

Figure 12, shows transformed images generated by ExplainGAN. The binary mask highlights the regions that differ from input to the generated image. The model made changes to the transformed image that are highly localized (here mustache) and minimal yet human perceptible. Localized changes can be observed in the first image, the model made changes only to the mustache but not the eyeglasses.

For quantitative evaluation, the authors of ExplainGAN proposed a new metric called '**substitutability**'. It means how much of the generated images can be substituted for original images. We compare the accuracy of the original images and generated images by the classifier. The higher the accuracy, the better the quality of generated images. For example, if the accuracy of original images of class 1 is 90 percent and the generated images belonging to the same class have an accuracy of 45 percent, then the substitutability of generated images is 50 percent. Since it is the first paper to define substitutability there are no benchmark data available to quantitatively evaluate this model.

3.3.3. Advantages and Limitations. A major advantage of ExplainGAN [35] approach is that it is model-agnostic in nature. Also, the explanations produced are easy to understand even for a non-practitioner.

The limitation of this method is that it produces a binary mask, which is not suitable for creating a saliency maps which require a continuous mask. The binary mask can only be partially used as an attribution map where the pixel priority of all the pixels in a mask is unknown. Another limitation is that it cannot be applied for one-vs-all explainability. It makes use of many custom unconventional loss functions such as prior loss, smooth loss, entropy loss, count loss, thus obtaining the correct set of values for these hyper-parameters is a challenging task.

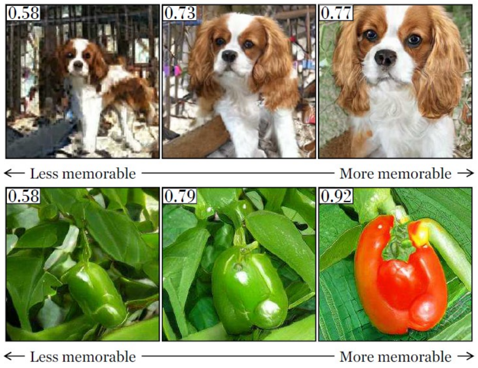

Figure 13. Visualizations generated by GANalyze. It contains series of dog and a vegetable related images sorted from less memorable to high memorable order with top right value indicating the property of interest. [43]

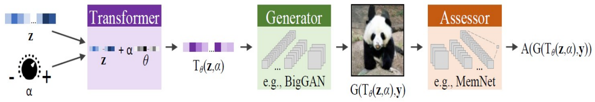
z
α
Transformer
z
+ α
θ
$T_\theta(\mathbf{z},\alpha)$
Generator
e.g., BigGAN
$G(T_\theta(\mathbf{z},\alpha),\mathbf{y})$
Assessor
e.g., MemNet
$A(G(T_\theta(\mathbf{z},\alpha),\mathbf{y}))$

Figure 14. High level architecture of GANalyze model. The transformer changes the latent vector (z) such that it produces desired set of images. The accessor predicts the interest attribute (memorability). The α is used to set required degree of change. [43]

3.4. GANalyze: Toward Visual Definitions of Cognitive Image Properties

Human cognitive properties such as memorability, aesthetics, and emotional valance do not have a concrete visual definition. There are studies [51] that explore semantic explanations such as humans are more memorable than trees present in a particular scene. However, there are no quantifiable properties that can be measured for what makes an image more memorable or more aesthetic. GANalyze aims to provide insights into interpreting cognitive properties through some quantifiable measures such as object size varying from small to large, the color varying from muted to bright, person's smile varying from normal to high rather than semantic explanations. GANalyze [43] aims to provide meaningful explanations for the cognitive properties of a scene through quantifiable properties.

Figure 13, describes a series of generated images, where we can navigate the path of increasing memorability. The visualizations responsible for obtaining images from low memorable to high memorable are analyzed. The visualizations corresponds to image properties such as object dimension (size). Though GANalyze [43] does not directly correspond to interpreting an AI system, in the above-explained way, concrete definitions can be attributed to explain the cognitive properties such as memorable, valence, aesthetics, using Generative Adversarial Networks.

3.4.1. Working Principle. A scene is best described in the form of images. Several images of the same scene differing slightly in human perceptible transitions such as object size, shape, color, and object centering form an important resource for GANalyze explanations. Generative Adversarial Networks are highly capable of generating such realistic images. To produce slight variations for the base image, a transformation is required to move the input vector along a certain direction in the generator's latent space. After generating several images for the same scene, a function is needed to access the image's property of interest such as valence and memorability, such a function is called Assessor.

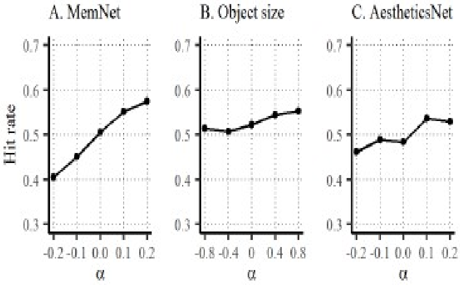

Figure 15. Result of user studies captured on various assessors. The graphs shows the relation of users hit rate and assessor score. [43]

$$T(z,\alpha) = z + \alpha.\theta \tag{4}$$

GANalyze uses a generator from BigGAN [52] which is pretrained on ImageNet [53] dataset. The Generator takes a noise input (z) and target class (y) to produce an image G(z,y). A transformer is used to modify the input vector (z) along the learned direction (θ) in the generator's latent space as in equation 4. Here α is a hyperparameter to set the desired amount of required change. To estimate the property of interest present in an image, an assessor module such as a MemNet [54] (memorability), AestheticNet [55] (aesthetic) is employed in the network. The complete GANalyze architecture is present in figure 14. The memorability score increases in general if the hyperparameter(α) value increases. Several quantifiable components that can contribute to memorability are shape, redness, color, entropy.

3.4.2. Qualitative and quantitative Evaluation. There is no benchmark evaluation available for this experiment. The evaluation is done based on user studies. Several hundreds of image sets were generated using the GANalyze framework. Each image set has multiple images of a particular scene with slight variation in the object of interest properties such as no smile to a big smile on a human face. Then each test user is shown an image from different sets and they have to select yes if they are sure that they have seen a repeat of the currently shown image.

Later, the experiment images are passed through the respective assessor to estimate the assessor score. From the figure 15, the high assessor score (high α) corresponds to the high hit rate from the user studies and vice versa. Thus, the evaluation shows that the model can successfully navigate the GAN latent space to make an image more (less) memorable to humans.

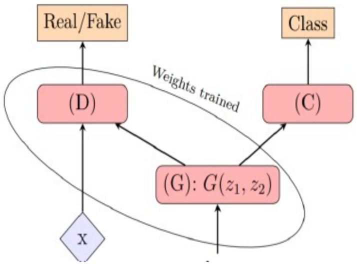

Figure 16. The figure describes the medXGAN training scheme with (C) being the medical classifier and (G),(D) are generator and discriminator respectively. [37]

3.4.3. Advantages and Limitations. GANalyze [43] is the first of its kind to demonstrate the effect of cognitive properties of a scene in a quantifiable manner rather than the semantic attributes based on mere human cognition abilities. The authors also demonstrated that the cognitive properties are interrelated to each other, for example an image modified to become more aesthetic also becomes more memorable.
The major limitation of this method is to select the appropriate quantifiable cognitive factor for a given scene. Several factors that affect memorability are color, shape, object size and to choose a particular key factor for a given scene (image) is crucial for explaining the cognitive property. The authors do not address this problem. Also, there are no benchmark results available and this experiment is highly dependent on user studies for its evaluation, cognitive properties differ very much for each individual. Thus, user results may yield differently based on different users groups as they can be from various demography's, backgrounds, stereotypes.

3.5. medxGAN: Visual Explanations for Medical Classifiers through a Generative Latent Space

Medical explanation GAN (medXGAN) [37] is a framework proposed to visually interpret a medical classifier's decisions. This method works by encoding medical domain knowledge into generators latent space through medical image dataset. In order to visualize changing class attributes, this method takes an input image (positive for certain disease) from a target class, then tries to find the image latent representation and interpolate in the latent space to get image's negative (negative for the disease) realization. It works for binary classifiers. But it can be extended multi-class classifiers by generating a negative image from any class other the input's original class.

3.5.1. Working Principle. A pretrained classifier must be incorporated into the training of the medXGAN to capture classifier related attributes in the generator's latent space. Given a positive input image, a reconstruction task enables finding of the latent vectors in the generator's latent space. Finally, interpolate the latent vectors to generate a negative realization of the positive image. Through the interpolation we can visualize the changing features that are responsible for altering the classifiers decisions. The training scheme of medXGAN is visualized in figure 16 which has a generator, discriminator and a medical classifier that gives feedback on generated images.

	MRI	X-Ray
medXGAN	0:97	0:91
Grad-CAM	0.89	0.83

Table 3. The tables shows the MRI and X-Ray classifier's accuracy after the features identified by medXGAN and Grad-CAM are perturbed [37].

3.5.2. Quantitative Evaluation. The quantitative evaluation is done based on comparing the results from Grad-CAM [56] with medXGAN. For given input image, identify the classifier specific features using medXGAN and Grad-CAM methods. Perturb the features by replacing the pixels with the average intensity. Then pass the perturbed images into the original classifier and take its predictions. The results from table 3.5.1 shows that medXGAN recorded a larger drop in classifier's prediction on perturbed images than Grad-CAM indicating that the features identified by medXGAN are more relevant to classifier's decision.

3.5.3. Advantages and Limitations. Major advantage of medXGAN is its ability to identify localised fine details in an image that are responsible for classifier's prediction. The quantitative evaluations shows that it outperforms Grad-CAM visualizations.
Limitations of medXGAN is that reconstruction task of generating negative image from positive one is highly dependent on generator learning a rich latent distribution. If the generator is too specific to training data then an input differing from input distribution may result in poor reconstruction.

3.6. xAI-GAN: Enhancing Generative Adversarial Networks via Explainable AI Systems

During the extensive research for applications of GANs in Explainable AI, it is natural to come across the approaches to improve GANs using Explainable AI methods. One such approach is XAI-GAN. Generative Adversarial Networks (GANs) excel in the process of generation of new amount of realistic data. GANs try to mimic the input data distribution. However, the training of GANs requires an enormous amount of training data which may not be available in all the domains. Hence there arise the need for data efficiency for training the GANs. Various methods have been proposed to address this issue, once such commonly employed method is Differential Augmentation [38], which increases the amount of training data by augmenting the input data using different augmentation techniques such as rotating, scaling, cropping the images.
XAI-GAN [17] is another such method that addresses the problem of insufficient data for efficient training of GANs.

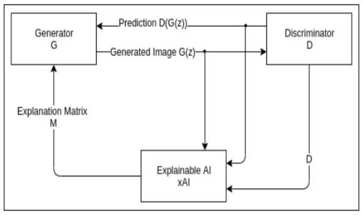

Figure 17. Architecture of XAI-GAN. This consists of XAI system in between the generator and discriminator. This provides each feature importance in form of **Explanation matrix**(E) to the generator. [17]

The corrective feedback loop from one Deep Neural Network (discriminator) to another Deep Neural Network (generator) is the key idea behind the working of GANs. Along with the standard GAN loss, XAI-GAN provides additional feedback from the discriminator to the generator making use of Explainable AI. This form of additional feedback provides more insights to the generator about how the discriminator is identifying real or fake data. Therefore, with this additional information, the generator learns quickly to produce more realistic images to fool the discriminator. Hence reducing the need to have more training data.

3.6.1. Working Principle. The XAI system is a model agnostic method and can be used with any XAI system that is capable of providing the importance of each input feature with respect to the classifier (here discriminator) in a quantifiable measure.

The XAI system takes the input data, discriminator, and its prediction and generates each attribute importance in form of a normalized score in a matrix. The matrix is called Explainable matrix (E) which contains each input feature's importance in normalized score (between 0 and 1), where a value of 1 means that the feature is of high importance to the classifier and modifying this feature will alter the classifier's decision. These normalized values are then passed to the gradient descent method of the generator to focus on those important features (guided gradient descent) that the discriminator uses to identify real vs fake data.

$$\delta^{'}_{G(z)} = \delta_{G(z)} + \alpha * \delta_{G(z)} * M \tag{5}$$

Equation 4 represents the XAI-GAN guided gradient descent where $\delta_{G(z)}$ is the generated image gradient, M is the Explanation matrix, and α is the hyper-parameter to control XAI information. The values in the explainable matrix are multiplied with generated image's gradient which penalizes the low-important (low score) features and benefits high-important features during the generators gradient descent.

3.6.2. Qualitative and quantitative Evaluation. XAI-GAN is quantitatively evaluated based on the quality of the generated images. Frechet Inception Distance (FID) [39] is a measure to calculate the quality of different data distributions. The higher the quality, therefore the lesser amount of data is required to successfully train a GAN.

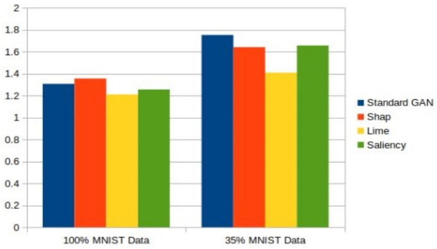

Figure 18. Frechet Inception distance (FID) calculated for standard GAN and XAI (using Shap, Lime, Saliency) over 100 percent and 35 percent of MNIST data. Lower the score the better the quality of generated images. [17]

Dataset	Standard GAN	Shap	Lime	Saliency
MNIST	1221.19	9925.1	38865.93	2215.73
FMNIST	991.11	9694.81	39237.33	2162.24

Table 4. Average time taken (seconds) for training of standard GAN vs XAI-GAN using Shap, Lime and Saliency methods on MNIST and Fashion MNIST datasets. [17]

Considering an entire MNIST dataset [41], from figure 18, XAI-GAN employing LIME XAI system's generated images showed an improvement of 7.3 percent on FID score compared to standard GANs generated images. On 35 percent of the MNIST data, XAI-GAN using LIME XAI system showed a enhancement of 19.6 percent on FID scores. Also, XAI-GAN using LIME produces an improvement of 2.2 percent when trained on only 20 percent of CIFAR data [40] whereas the standard GAN has been trained on complete 100 percent of CIFAR dataset. In this case, only 20 percent of CIFAR data is sufficient for XAI-GAN to generate images that have same quality of standard GANs generated images trained over complete 100 percent data.

3.6.3. Advantages and Limitations. The major advantage of XAI-GAN is that it requires less amount training data compared to standard GANs training in order to generate same quality of images. This method can be applied to all the domains where limited amount of training data is available to train the classifier using GANs. Since XAI-GAN is model agnostic it can be used with any interpretable AI system. XAI-GAN can even be combined with data augmentation techniques such as differential augmentation which increases the efficiency of the training process further. Since it requires us to modify the gradient descent with explainable matrix (E) values, we will have greater control over the learning process of this model.

XAI-GAN training requires a lot of time because of the overhead added by employing XAI systems in GANs training. As shown in table 4, the standard GAN training took 1221 seconds to train on MNIST dataset whereas XAI-GAN employing Shap has taken 9925 seconds. This is the major limitation of this model.

4. Discussion

In this paper, we have discussed various state-of-the-art applications of GANs in XAI domain. First, we discussed GANMEX [10], a novel approach for generating a realistic baseline input using GANs to further enhance the attribution method's one-vs-one explainability. Mis-classification of inputs can be accurately explained with the saliency maps produced by the attribution method by considering the baseline inputs generated by GANMEX. Though the authors of GANMEX did not propose a method to generate a baseline image for one-vs-all explainability, it can be solved by considering a target sample to be present in any of the classes other than the input's original class.

Then we briefed on how counterfactual examples generated using GANs can be used to interpret the decisions of a black-box model using StylEx [33]. The machine learning models can be biased based on biases in the training data. StylEx can identify the model biases if the top attributes identified by this method do not correspond to reality. But sometimes the inputs selected from the same class does not have the same priority of attributes and different input from the same class may yield different top-k attributes. There isn't any solution to handle such scenarios, we propose that the user have to evaluate all the samples present in a particular class and calculate the importance of each attribute for input, then take the mean of each attribute from all the samples to get the priority of attributes. The proposed method is a naive approach, further research is needed to come up with an ideal solution.

ExplainGAN [35] generates a counterfactual example for a given input and a binary mask that captures the necessary changes required for the classifier to predict a different class other than its original class for the given input. The explanations produced by the binary mask are easy to interpret even for a non-practitioner. However, the binary mask is not suitable to produce saliency maps as they require a mask with continuous values. The authors proposed many unconventional loss functions and obtaining a correct set of hyper-parameter values is a challenging task.

GANalyze [43] is a technique to explain the cognitive properties of an input in a quantifiable manner rather than semantic perception. Though this method does not directly correspond to interpreting a complex black-box model, concrete definitions can be attributed to explaining the cognitive properties such as memorable, valence, and aesthetics for the inputs using Generative Adversarial Networks. As different images (or scenes) may depend on different quantifiable properties such as size, shape, and color to enhance cognition, the authors of this method did not present a way to select the appropriate quantifiable components for different scenarios. Further research is required to select the appropriate measurable properties for different scenarios.

xAI-GAN [17] uses attribution methods to further improve the quality of samples generated by GANs. In general, all the GAN based XAI approaches require a lot of training data to generate accurate results, xAI-GAN can be employed in all those methods and domains that suffer from having a huge amount of training data. However, xAI-GAN takes more training time compared to normal GAN training. We believe, it depends on the individual use case where one has to consider the time vs training data trade-off to generate accurate samples using the xAI-GAN approach.

Some of the characteristics that we observed during our research on applications of GANs for explainable AI are:

- Since GANs are specialised in generating new samples, most of the presented approaches follow perturbation based model explanations.
- Most of the applications are model agnostic and can be applied to interpret any type of classifier.
- GANs rely on huge amount of training data, XAI systems employing GANs also suffer from this drawback.
- Since the GAN training does not involve any classifier specific training, all the XAI methods employing GANs have to incorporate to-be-explained model into the discriminator of the GAN architecture.
- We believe, StylEx can be used to generate baseline images satisfying all the three properties proposed by the authors of GANMEX through careful modification of input in the StyleSpace.
- As there is no definite benchmark evaluation available for XAI systems, therefore each method takes their own approach to explain (qualitatively) the usefulness of the proposed method.

5. Conclusion

In the introduction section, we have discussed the need for understanding the decisions of a black-box model. Then elaborated on Explainable AI systems and the various types of XAI systems. Later presented Generative Adversarial Networks (GANs) in detail and their working principle with their applications in various sectors. Next, we have seen various methods and their application of GANs in XAI. For each method, we have presented the problem it addresses, its working principle, qualitative and quantitative evaluation, advantages, and limitations. In the discussion section, we proposed our in-depth analysis of each method and proposed alternate solutions for some of the problems that GANMEX and StylEx methods are facing. We have also highlighted that further research is necessary to tackle the problems highlighted in the discussion section for StylEx and GANalyze methods. We presented several characteristics of applications of GANs in XAI in our discussion section.

By comparing the advantages and limitations of every above-discussed application, the advantages presented by each approach outweighs the limitations. These limitations can also be overcome in some or the other way as presented in the discussion section. Hence, we conclude that employing GANs in XAI would bring us a step closer to the ultimate goal of interpreting the decisions of any kind of complex AI model.

References

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. arXiv:1406.2661. arXiv year 2014.

[2] Zihang Dai, Zhilin Yang, Fan Yang, William W. Cohen, and Ruslan R. Salakhutdinov. 2017. Good semi-supervised learning that requires a bad GAN. In Advances in Neural Information Processing Systems. 6510–6520

[3] William Fedus, Ian Goodfellow, and Andrew M. Dai. 2018. MaskGAN: Better Text Generation via Filling in the . arXiv:1801.07736. Retrieved from https://arxiv.org/abs/1801.07736.

[4] Chris Donahue, Julian McAuley, and Miller Puckette. 2018. Synthesizing Audio with Generative Adversarial Networks. arXiv:1802.04208. Retrieved from https://arxiv.orb/abs/1802.04208.

[5] Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. 2017. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. arXiv:1706.02633. Retrieved from https://arxiv.org/abs/1706.02633.

[6] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” in Information Processing in Medical Imaging. Berlin, Germany: Springer, 2017, pp. 146–157.

[7] Amil Dravid, Florian Schiffers, Boqing Gong, Aggelos K. Katsaggelos. medXGAN: Visual Explanations for Medical Classifiers through a Generative Latent Space. arXiv:2204.05376. arXiv year 2022.

[8] P. Isola et al., “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5967–5976.

[9] C. Li and M. Wand, “Precomputed real-time texture synthesis with markovian generative adversarial networks,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 702–716.

[10] Sheng-Min Shih, Pin-Ju Tien, Zohar Karnin Proceedings of the 38th International Conference on Machine Learning, PMLR 139:9592-9602, 2021.

[11] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein, “Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai,” arXiv preprint arXiv:1902.01876, 2019.

[12] A Guide for Making Black Box Models Explainable by Christoph Molnar. Topic: Neural Network Interpretations - Pixel Attribution.

[13] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, Mohiuddin Ahmed. Explainable Artificial Intelligence Approaches: A Survey. arXiv:2101.09429, arXiv year: 2021.

[14] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” Annals of statistics, pp. 1189–1232, 2001.

[15] Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 1135–1144.

[16] Lundberg, S. M.; and Lee, S.-I. 2017a. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., Advances in Neural Information Processing Systems 30, 4765–4774. Curran Associates, Inc. URL http://papers.nips.cc/paper/7062-a-unifiedapproach- to-interpreting-model-predictions.pdf.

[17] Vineel Nagisetty, Laura Graves, Joseph Scott, Vijay Ganesh. xAI-GAN: Enhancing Generative Adversarial Networks via Explainable AI Systems. arXiv:2002.10438, arXiv year 2020.

[18] Ekin Tiu. Understanding Latent Space in Machine Learning - medium.

[19] Guim Perarnau. Fantastic GANs and where to find them. https://guimperarnau.com/blog/2017/03/Fantastic-GANs-and-where-to-find-them

[20] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 8789–8797

[21] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-image Translation Using Cycle-consistent Adversarial Networks. arXiv:1703.10593v6. Retrieved from https://arxiv.org/abs/1703.10593v6.

[22] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised dual learning for image-to-image translation. In Proceedings of the IEEE International Conference On Computer Vision. 2849–2857

[23] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 105–114.

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8110–8119, 2020.

[25] Sung Woo Park and Junseok Kwon. 2019. Sphere generative adversarial network based on geometric moment matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’19).

[26] Emily L. Denton, Soumith Chintala, Rob Fergus, et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. In Advances in Neural Information Processing Systems. 1486–1494

[27] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv:1809.11096. Retrieved from https://arxiv.org/abs/1809.11096

[28] Augustus Odena. 2016. Semi-supervised Learning with Generative Adversarial Networks. arXiv:1606.01583. Retrieved from https://arxiv.org/abs/1606.01583

[29] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. 2017. GP-GAN: Towards Realistic High-resolution Image Blending. arXiv:1703.07195. Retrieved from https://arxiv.org/abs/1703.07195

[30] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision. IEEE, 2813–2821.

[31] Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. CoRR, abs/1703.01365, 2017. URL http://arxiv.org/abs/1703.01365.

[32] Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. pp. 4765–4774,2017. URL http://papers.nips.cc/paper/ 7062-a-unified-approach-to-interpreting-model-predictions. pdf.

[33] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, Inbar Mosseri. Explaining in Style: Training a GAN to explain a classifier in StyleSpace. arXiv:2104.13369, arXiv year 2021.

[34] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for StyleGAN image generation. arXiv preprint arXiv:2011.12799, 2020.

[35] Samangouei, Pouya, Ardavan Saeedi, Liam Nakagawa and Nathan Silberman. “ExplainGAN: Model Explanation via Decision Boundary Crossing Transformations.” ECCV (2018).

[37] Amil Dravid, Florian Schiffers, Boqing Gong, Aggelos K. Katsaggelos. medXGAN: Visual Explanations for Medical Classifiers through a Generative Latent Space. arXiv:2204.05376, arXiv year :2022.

[38] Zhao, S.; Liu, Z.; Lin, J.; Zhu, J.-Y.; and Han, S. 2020. Differentiable Augmentation for Data-Efficient GAN Training. arXiv preprint arXiv:2006.10738 .

[39] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in Neural Information Processing Systems, 6626–6637.

[40] Krizhevsky, A.; Nair, V.; and Hinton, G. 2010. CIFAR-10 (Canadian Institute for Advanced Research). Unpublished Manuscript. URL: https://www.cs.toronto.edu/ kriz/cifar.html

[41] LeCun, Y.; and Cortes, C. 2010. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/.

[42] C3.ai, What is Local Interpretable Model-Agnostic Explanations (accessed on 28.07.2022) https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/

[43] Lore Goetschalckx (1 and 2), Alex Andonian (1), Aude Oliva (1), Phillip Isola (1) ((1) MIT, (2) KU Leuven). GANalyze: Toward Visual Definitions of Cognitive Image Properties. arXiv:1906.10112, arXiv year: 2019.

[44] J. Kim and H. Park, ”OA-GAN: Overfitting Avoidance Method of GAN oversampling based on xAI,” 2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN), 2021, pp. 394-398, doi: 10.1109/ICUFN49451.2021.9528594.

[45] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 607–617, 2020.

[46] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In International Conference on Machine Learning, pages 2376–2384. PMLR, 2019.

[47] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL and Tech., 31:841, 2017.

[48] Mukund Sundararajan, Ankur Taly, Qiqi Yan. Axiomatic Attribution for Deep Networks. arXiv:1703.01365.

[49] Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott Lundberg, Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. arXiv:1906.10670.

[49] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. arXiv:1912.04958.

[50] Yang, M. and Kim, B. BIM: towards quantitative evaluation of interpretability methods with ground truth. CoRR, abs/1907.09701, 2019.

[50] yan2021agg, author = Zeng, Yanhong and Fu, Jianlong and Chao, Hongyang and Guo, Baining, title = Aggregated Contextual Transformations for High-Resolution Image Inpainting, booktitle = Arxiv, pages=-, year = 2020.

[51] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7):1469–1482, jul 2014.

[52] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. CoRR, abs 1809.11096, 2018.

[53] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. Int. J. Comput. Vision, 115(3):211–252, Dec. 2015.

[54] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In The IEEE International Conference on Computer Vision (ICCV), December 2015.

[55] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, Computer Vision – ECCV 2016, pages 662–679, Cham, 2016. Springer International Publishing

[56] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra.Grad-cam: Visual explanations from deep networks viagradient-based localization. In Proceedings of the IEEE internationalconference on computer vision, pages 618–626,2017. 2, 3.

[57] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

[58] Y. Li et al., ”StoryGAN: A Sequential Conditional GAN for Story Visualization,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6329-6338.

[59] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, ”Mocogan: Decomposing motion and content for video generation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1526-1535.

[60] Rongjie Huang, Chenye Cui, Feiyang Chen, Yi Ren, Jinglin Liu, Zhou Zhao, Baoxing Huai, Zhefeng Wang.arXiv:2110.07468, arXiv:year:2021.