Pattern Recognition Letters 156 (2022) 112–118

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# VANT-GAN: Adversarial Learning for Discrepancy-Based Visual Attribution in Medical Imaging

Tehseen Zia[a,b,c,*], Shakeeb Murtaza[d,c], Nauman Bashir[b,c], David Windridge[e], Zeeshan Nisar[f]

[a] National Center for Artificial Intelligence, Prince Muhammad Bin Fahad University, Saudi Arabia
[b] Medical Imaging and Diagnostics Lab, National Center of Artificial Intelligence, Pakistan
[c] COMSATS University Islamabad, Pakistan
[d] Laboratoire d'imagerie, de vision et d'intelligence artificielle, École de technologie supérieure, Montreal, Canada
[e] Middlesex University, UK
[f] ICube, University of Strasbourg, CNRS (UMR 7357), France

ARTICLE INFO

*Article history:*
Received 30 March 2021
Revised 31 January 2022
Accepted 5 February 2022
Available online 8 February 2022

Edited by: Jiwen Lu

*Keywords:*
Visual Attribution
Domain Translation
Alzheimer
Generative Adversarial Networks

ABSTRACT

Visual attribution (VA) in relation to medical images is an essential aspect of modern automation-assisted diagnosis. Since it is generally not straightforward to obtain pixel-level ground-truth labelling of medical images, classification-based interpretation approaches have become the de facto standard for automated diagnosis, in which the ability of classifiers to make categorical predictions based on class-salient regions is harnessed within the learning algorithm. Such regions, however, typically constitute only a small subset of the full range of features of potential medical interest. They may hence not be useful for VA of medical images where capturing all of the disease evidence is a critical requirement. This hence motivates the proposal of a novel strategy for visual attribution that is not reliant on image classification. We instead obtain normal counterparts of abnormal images and find discrepancy maps between the two. To perform the abnormal-to-normal mapping in unsupervised way, we employ a Cycle-Consistency Generative Adversarial Network, thereby formulating visual attribution in terms of a discrepancy map that, when subtracted from the abnormal image, makes it indistinguishable from the counterpart normal image. Experiments are performed on three datasets including a synthetic, Alzheimer's disease Neuro imaging Initiative and, BraTS dataset. We outperform baseline and related methods in both experiments.



## 1. Introduction

Medical image classification is becoming a vital aspect of patient stratification, disease progression assessment, treatment response and disease severity grading within a modern medical setting. Consequently, it is increasingly important for practitioners to understand the salient information underlying these automated classifications, which in turn motivates the study of *visual attribution* (VA) [1–5]. The need for VA arises because machine diagnosis typically differs from that of human experts in key respects; for instance, a radiologist is trained via observation of many abnormal/normal images such that they are able to transfer their internally-learned representation of the disease to novel image settings. Their training hence enables them to analyze a image by finding abnormalities that differ from a conjectured *counterpart normal* representation of the equivalent healthy patient. A supervised classification system, by contrast, will typically seek to identify key features indicative of the distinction between normal and abnormal tissue [1].

Inspired by this conjectured expert *modus operandi*, we seek a methodology capable of the production of a counterpart normal image in relation to an input image such that we may use this 'normal' image to analyze the input image. By accompanying input images with their counterpart normal images it hence becomes possible to provide a visual analogy-based counterfactual explanation for the automated diagnostic decision.

Visual attribution is currently addressed by initially training a deep neural network (DNN) based image classifier and then using one of the following two approaches: 1) application of forward propagation (or activation) to find regions of the input image responsible for making predictions (e.g. [1], or else 2) using back-propagation to analyze the gradient of the prediction with respect to the input image [2]. Neural network classification based visual attribution approaches consequently tend to exhibit common limitations, potentially leading to undesirable outcomes in certain settings. In particular, since neural network classifiers are trained to minimize mutual information between inputs and outputs, they

* Corresponding author.
*E-mail address:* tehseen.zia@comsats.edu.pk (T. Zia).

https://doi.org/10.1016/j.patrec.2022.02.005
0167-8655/

are implicitly conditioned to utilise the fewest possible input features. Consequently, DNN classifiers typically make predictions based on certain salient regions rather than entire objects of interest. In other words, a classifier may disregard low-discrimination features when dominant features with a sufficiency of information about the target are available (an early study [6] demonstrated that if evidence of a particular class is present in multiple regions of an image, e.g. multiple indicators of disease within a medical image, a DNN classifier will likely disregard a significant fraction of this evidence).

Within the domain of medical image diagnosis, by contrast, it is highly desirable to visually attribute evidence of a disease in such a way as to capture *all* of the disease effects present. This hence motivates us to propose a novel strategy for visual attribution that is not reliant on image classification, as distinct the majority of extant techniques. We instead aim to obtain normal counterparts[1] of abnormal images and find discrepancy maps between the two.

To do this, we leverage, as a stepping-off point, a recently-proposed generative adversarial network architecture (ANT-GAN) [7] capable of generating normal-looking correlates (if not strict counterparts) of the abnormal images. Since it is generally unrealistic to obtain contemporaneous normal and abnormal pairs practically, ANT-GAN learns to perform the abnormal-to-normal mapping in unsupervised way via the application of the Cycle-Consistecy GAN principle, in which an inverse mapping and cycle consistency (i.e. forwards-backwards) loss is introduced to the GAN in order to tackle tasks for which paired training data does not exist.

By utilizing and extending this capacity of cycle-consistent GANs to produce abnormal-to-normal translation medical image pairs, we shall demonstrate that it is possible to re-formulate visual attribution in terms of a *discrepancy map* that, when subtracted from the abnormal image, will make it indistinguishable from the counterpart normal image. To this end, we propose a class of generative models for learning discrepancy maps as a function of abnormal images. In particular, we propose an VA-extended ANT architecture, dubbed Visually-Attributed Abnormal-to-Normal Translation GAN (VANT-GAN), that learns to generate discrepancy maps *simultaneously* to learning to perform abnormal-to-normal translation.

Our approach thus aims to improve on the Visual Attribution GAN (VA-GAN) method proposed in [8], in which a map is learned that, when added into an abnormal image, renders it indistinguishable from images of the normal class. Since the map-generating function in the VA-GAN case does not aim to produce the normal counterpart of an abnormal image but rather *any* normal-looking image, the learned image translation may depict discrepancies irrelevant to medical diagnosis. We shall, in contrast, set out to constrain the unconstrained abnormal-to-normal image translation function of [8] by generating normal *counterparts* of abnormal images in order to reduce false-positive visual attributions.

## 2. Current State-of-the-Art in Medical Visual Attribution

Visual attribution (VA) in relation to medical images is currently performed predominantly via the Class Activation Map (CAM) [1] paradigm of classifier explanation. CAM in its original form used global pooling to highlight the discriminative regions of the input image most important to the CNN in reaching a decision; however, the method was later improved by replacing global average pooling with gradient-based feature attribution (referred to as grad-CAM [2]). Since grad-CAM tends to produce a coarse-grained visualization, the authors in [3] proposes *guided grad-CAM* that utilises a guided-backpropagative approach. A similar attempt to enhance the resultant maps, smooth-Grad [4], adds visual noise of differing magnitudes to an image, taking the average of the produced sensitivity maps for the final enhanced mapping. CAM-based VA techniques have found wide use across the medical domain e.g. in digital pathological images for bladder cancer prediction [5], prostate cancer detection [9], benign and malignant cutaneous tumors classification [10], Covid-19 detection from CT scans [11], Alzheimer diagnosis in MRI images [12], malarial parasite detection in thin blood-smear images [13], bone age assessment [14], interpretable CNN based cervical cancer [15], brain gender detection [16], tuberculosis visualization in Chest X-rays [17,18], diabetic retinopathy classification and visualization [6,19,20].

Despite their widespread adoption, CAM-based methods are limited in their resolution by the final layer of the model. Consequently, post-processing is often required to enhance the output resolution. Though few methods are recently devised to provide concept-level attribution [21], they rely on humans to provide concept-level details. A further issue with CAM-based methods is that the DNN classifiers employed in these methods tend to preferentially select highly discriminative features, while ignoring low-discrimination features leading to imperfect VA [8]. Also, it is reported that there is a misalignment in CAM-based VA due to upsampling of VA in CAM [22].

To mitigate the disadvantages of CAM-based techniques, a generative VA method for medical images was proposed [8]. The method uses a generative adversarial network (GAN) with Wasserstein loss function to transform abnormal medical images so as to make them indistinguishable from normal medical images. Although the method outperforms CAM-based techniques on medical images, the generated VA often contains undesirable artifacts as a consequence of the unconstrained abnormal-to-normal translation; that is, since normal and abnormal images are not aligned, the generator also learns to attribute irrelevant discrepancies between unpaired images. In order to constrain the abnormal-to-normal translation, a GAN architecture employing a cycle-consistent loss function was proposed in [7]. Here, VA is expressed via the difference between an abnormal image with corresponding normal image. The main disadvantage of the method, however, is its requirement of post-processing in order to deal with the resolution mismatch of abnormal and synthesised normal images.

In order to addressed these limitations in the current SotA relating to GAN-based medical image VA, we seek in this paper to develop an architecture that, rather than conducting explicit abnormal-to-normal translation, instead exhibits the capacity to learn a VA map similar to that of the Residual-GAN [7], that when added to an abnormal image, will directly translate it into the corresponding normal image. However, in common with [7], we employ cyclic-consistency loss function to constrain the abnormal-to-normal image translation. The Cycle-Consistency GAN has recently shown to be promising to augment Chest x-rays, particularly for Covid detection [23].

## 3. Proposed VANT-GAN Methodology

We indicate normal medical images by $x^n$ and abnormal images by $x^a$. We further assume that the $x^n$ and $x^a$ observations are sampled from distributions $p_n(x)$ and $p_a(x)$, respectively, and that an abnormal image differs from its corresponding normal image (i.e. from same patient) only by the characteristic disease markers. Within this setting, when given an abnormal image as input, we seek to produce a *disease effect map/visual attribution map* that contains all of the features that distinguish an abnormal image $x_i^a$ from its counterpart normal image $x_i^n$. In other words, we wish to generate a map that, when subtracted from the abnormal image $x_i^a$, produces an image indistinguishable from its counterpart nor-

[1] That is, visual counterparts that would be indistinguishable to the abnormal image, were it not to exhibit the effects of disease.

mal image $x_i^n$. Mathematically,

$$x_i^n = x_i^a - M(x_i^a) \tag{1}$$

where $x_i^n$, $x_i^a$ and $M(x_i^a)$ are of the same dimensions.

Ideally, to model the function $M$, a data-set consisting of normal and abnormal image pairs is required, however, this is something that it is generally unrealistic to obtain in real clinical practice. A previous study proposed *VAGAN* (Visual Attribution Generative Adversarial Network) for learning the function $M$ in an under-constrained setting; i.e. by aiming to translate an abnormal image into an arbitrary normal image, rather than a strict counterpart normal image. Consequently, the disease effect map $M$ produced by this approach may contain many false positives, reflective of irrelevant discrepancies between the unpaired normal and abnormal images. We instead build on recent developments in abnormal-to-normal image translation (ANT) in order to learn an $M$ capable of translating an abnormal image into its counterpart normal image, rather than an arbitrary normal-looking image. The ANT model is described in Subsection 3.1 and the $M$ model in Subsection 3.2.

### 3.1. Abnormal-To-Normal Translation

Sun et. al proposed in [7] a generative adversarial network based ANT model (a.k.a. ANT-GAN) for generating normal counterparts to abnormal images. The main component of their model is a generator $\mathcal{G}_{A2N}$ that takes, as input, an abnormal image $x$ and produces as output the normal counterpart $\mathcal{G}_{A2N}(x)$. For learning to converge, the generator must produce a realistic normal $\hat{x}^n = \mathcal{G}_{A2N}(x^a)$ capable of fooling the normal discriminator $D^N$. The cycle consistency regularization principle is leveraged via a generator $\mathcal{G}_{N2A}$ and a discriminator $D^A$ that constrain the model to produce a *counterpart* normal. The ANT-GAN model can thus be defined as an objective function $\mathcal{L}$ consisting of three distinct parts: a GAN model $\mathcal{L}_{GAN}$, a cycle-consistent loss $\mathcal{L}_{CC}$, and an anomaly mask loss $\mathcal{L}_{AM}$. Mathematically,

$$\mathcal{L} = \mathcal{L}_{GAN} + \lambda_{CC}\mathcal{L}_{CC} + \lambda_{AM}\mathcal{L}_{AM} \tag{2}$$

where $\mathcal{L}_{GAN}$ is used to simultaneously train generators $\mathcal{G}_{A2N}$ and $\mathcal{G}_{N2A}$ and is defined as follows:

$$\begin{aligned}\mathcal{L}_{GAN} = {} & \mathbb{E}_{pa}\left[\ln D^A(x^a)\right] + \mathbb{E}_{pn}\left[\ln D^N(x^n)\right] \\ & + \mathbb{E}_{pn}[ln(1 - D^A(\mathcal{G}_{N2A}(x^n)))] + \mathbb{E}_{pa}\left[\ln(1 - D^N(\mathcal{G}_{A2N}(x^a)))\right]\end{aligned} \tag{3}$$

The cyclic-consistent loss $\mathcal{L}_C$ is used to transform normal and abnormal images into one another, and helps in the learning of $\mathcal{G}_{A2M}$ and $\mathcal{G}_{N2A}$:

$$\begin{aligned}\mathcal{L}_{CC} = {} & \mathbb{E}_{pa}[||\mathcal{G}_{N2A}(\mathcal{G}_{A2N}(x^a)) - x^a||_2] \\ & + \mathbb{E}_{pn}[||\mathcal{G}_{A2N}(\mathcal{G}_{N2A}(x^n)) - (\mathcal{G}_{A2N}(x^a))||_2]\end{aligned} \tag{4}$$

$\mathcal{L}_{CC}$ allows additional information to be transferred between abnormal and normal medical images while learning their corresponding generators. The first term aims to reconstruct a given abnormal image following its translation into a normal image, and the second term aims to reconstruct a given normal image following its translation into an abnormal image.

$\mathcal{L}_{AM}$ is used to isolate and modify the disease markers within the image while keeping the normal region within the image unchanged. $\mathcal{L}_{AM}$ is defined as:

$$\mathcal{L}_{AM} = \mathbb{E}_{pa}(x)\left[\ ||(1 - M_x) \odot \mathcal{G}_{A2N}(x^a) - x^a||_2\ \right] \tag{5}$$

where $\odot$ denotes element-wise multiplication, 1 represents an input-sized all-ones matrix and $M_x$ is an image-sized marker matrix.

### 3.2. VANT-GAN Visual Attribution Map Generation Model

In contrast to the ANT-GAN model, in which a generator explicitly synthesizes a normal counterpart from an abnormal image as $\mathcal{G}_{A2N} : x_i^a \rightarrow x_i^n$, VANT-GAN seeks to embody a generator $\mathcal{G}_{A2M}$ capable of taking an abnormal image $x^a$ as input so as to produce a map, $M_{x^a}$, via $\mathcal{G}_{A2M} : x_i^a \rightarrow M_{x^a}$, that when subtracted from the abnormal image, $x_i^a$, outputs a normal image $x_i^n$. If the generator $\mathcal{G}_{A2M}$ has converged effectively, then the discriminator $D^N$ ideally cannot distinguish between the real and fake (or synthesized) $x_i^n$. In contrast to VAGAN's under-constrained map generator, VANT-GAN further embodies a cycle consistency loss in order to constrain the generator $\mathcal{G}_{A2M}$ so as to be able to translate an abnormal image into its normal counterpart. To achieve cycle consistency, VANT-GAN hence learns an additional generator $\mathcal{G}_{N2A}$ by using discriminator $D^A$ in a similar manner to ANT-GAN.

To achieve this, we import ANT-GAN's loss function in Equation 3 into $\mathcal{L}_{VANT-GAN}$ as follows:

$$\begin{aligned}\mathcal{L}_{VANT-GAN} = {} & \mathbb{E}_{pa}\left[\ln D^A(x^a)\right] + \mathbb{E}_{pn}\left[\ln D^N(x^n)\right] \\ & + \mathbb{E}_{pn}\left[\ln(1 - D^A(\mathcal{G}_{N2A}(x^n)))\right] \\ & + \mathbb{E}_{pa}\left[\ln(1 - D^N(x^a + \mathcal{G}_{A2M}(x^a)))\right]\end{aligned} \tag{6}$$

We redefine the cycle-consistency $\mathcal{L}_{VANT-CC}$ by changing $\mathcal{L}_{CC}$ as follows:

$$\begin{aligned}\mathcal{L}_{VANT-CC} = {} & \mathbb{E}_{pa}[||\mathcal{G}_{N2A}(x^a + \mathcal{G}_{A2M}(x^a)) - x^a||_1] \\ & + \mathbb{E}_{pn}[||x^a + \mathcal{G}_{A2M}(\mathcal{G}_{N2A}(x^n)) - (x^a + \mathcal{G}_{A2M}(x^a))||_1]\end{aligned} \tag{7}$$

The cycle consistency loss $\mathcal{L}_{VANT-CC}$ (Equation 7) consists of two terms: first, *forward cycle consistency* which aims to bring back an abnormal image $x^a$ after translating it to its counterpart normal image $x^n$, i.e. $x^a \rightarrow x^a - \mathcal{G}_{A2M}(x^a) \rightarrow \mathcal{G}_{N2A}(x^a - \mathcal{G}_{A2M}(x^a))$, and second, *backward cycle consistency* which aims to reproduce $x^n$ after translating it to $x^a$, i.e. $x^n \rightarrow \mathcal{G}_{N2A}(x^n) \rightarrow x^n = x^a - \mathcal{G}_{A2M}(\mathcal{G}_{N2A}(x^n))$.

Finally, we define a loss $\mathcal{L}_{VANT}$ that optimizes VANT-GAN as follows:

$$\mathcal{L}_{VANT} = \mathcal{L}_{VANT-GAN} + \lambda\mathcal{L}_{VANT-CC} \tag{8}$$

Once VANT-GAN is trained, we retain only the generator $\mathcal{G}_{A2M}$ and discard the generator $\mathcal{G}_{N2A}$ and discriminators $D^A$ and $D^N$. In the application stage, we thus input an instance of the positive class to the network $\mathcal{G}_{A2M}$ in order to obtain the visual attribution map $M(x^a)$.

The proposed model is illustrated in Fig. 1.

## 4. Implementation

### 4.1. Network Architecture

The generator network is adapted from [24], which demonstrates excellent results in unpaired image-to-image translation. This network has three components: an encoder, a set of residual blocks, and a decoder. The encoder shrinks the representation of input image while increasing the number of channels. This encoder is comprised of three Convolution-InstanceNorm-Relu layers. The first convolution layer in the encoder is of kernel_size=7, stride=1, with k=64 filters. The remaining two convolutions are of kernel_size=3, stride=2, with k=128 filters. We utilize the concept of reflection padding to reduce the artefacts produced by these convolution layers. The encoder block is then passed to a set of either 6 or 9 residual blocks with k=256 filters (9 residual blocks are used where the input image size is greater than 128 by 128 pixels). The output of this residual block is then re-expanded in the decoder section via two transpose-convolutions with $k$ filters, each followed by an InstanceNormalization and a Relu layer. An

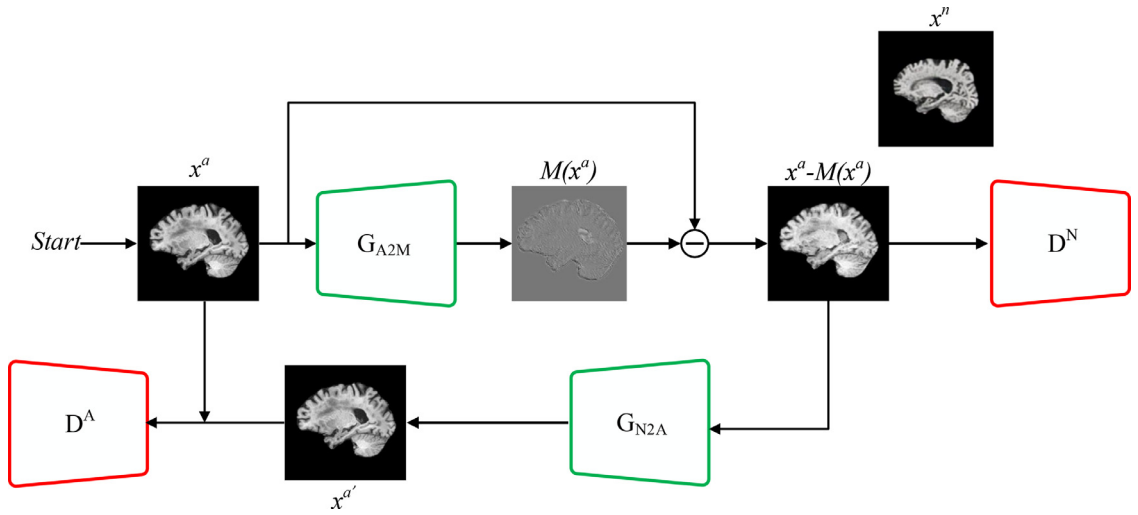


**Fig. 1.** VANT-GAN model diagram with an example image from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Four networks are employed: a generator trained to output discrepancy maps, a discriminator trained to discriminate normal images, a generator trained to synthesise abnormal images given a generated normal image, and a discriminator to discriminate abnormal images.

additional convolutional layer is applied to produce the final output in RGB format. For our Discriminator network, we utilize the concept of PatchGAN as adopted in [25]. PatchGAN was introduced to identify whether overlapping image patches are real or fake. We use four convolution layers with kernel_size=4, stride=2 and $k$ increasing filters followed by InstanceNormalization and LeakyRelu activation with a slope of 0.2. Finally, the last convolution layer is applied with kernel_size=1 to produce a 1-dimensional output.

### 4.2. Training Details

To stabilize the training we use least-square loss instead of negative log likelihood as adopted in [26]. Following the strategy of [27] we update the discriminator using a buffer of 50 previously generated outputs. The training procedure is carried out on a batch of size 1 via the Adam optimizer set to a 0.0002 initial learning rate, linearly decaying to zero over half of the total epochs. Initial weights are initialized randomly from a Gaussian distribution of $\mathcal{N}(0, 0.02)$. The loss weights ($w_{cycle}$= 10 and $w_{identity}$= 0.5$w_{cycle}$) are copied from our baseline architecture [24].

## 5. Experiments

We perform experiments on a synthetic dataset and two publicly available medical imaging datasets, the ADNI and BratS dataset. We evaluate the proposed VANT-GAN VA approach against the immediately comparable visual explanation methods indicated in the Introduction, namely; CAM [1], gradCAM [2], and VA-GAN [8]. Note that, whereas CAM and gradCAM utilize classification networks, VA-GAN, ANT-GAN and the proposed VANT-GAN employ image-translation networks. To further assist comparison, all tested networks are built using a similar discriminator architecture to that of the proposed method. However, for CAM methods, we replace the last two layers with a global average pooling layer followed by a dense prediction in order to create class-specific activation maps for visual explanation, as described in [1]. We quantitatively compare the respective methodologies using the Dice-Coefficient, Intersection over Union (IoU) and normalized cross correlation (NCC) evaluation metrics for synthetic and BraTs since ground truths are available for these datasets. We follow [8] to assess performance of the compared models on the ADNI dataset. As ground truths are not available for the ADNI dataset, NCC score is used to evaluate the models.

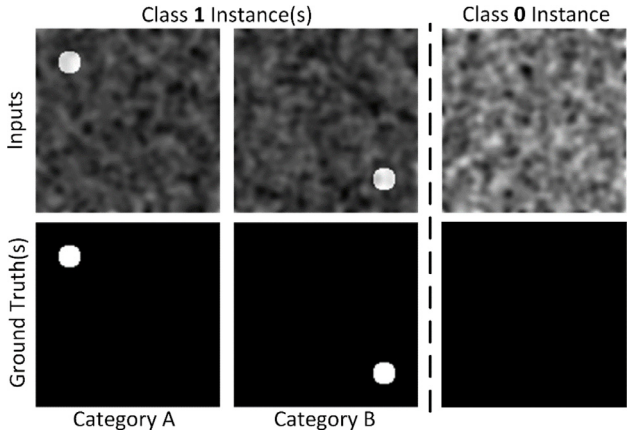


**Fig. 2.** Synthetic data examples. Left of the dotted line are samples of Class 1 (i.e. the disease class) and right of the dotted line are samples of Class 0 (i.e. the normal class). The upper row shows the input and the bottom row shows the ground truth.

### 5.1. Experiments on Synthetic Data

#### 5.1.1. Synthetic dataset

Alongside the indicated real medical imaging datasets, we evaluate the proposed and related VA approaches on a synthetically generated dataset consisting of 10000 128x128 images separated into two label classes such that one half of the dataset represents the healthy control group (label 0) and the remaining half represents the patient group (label 1). The images are generated via the data generation process set out in [8]. Healthy control group images are constructed by convolution of random iid Gaussian noise with a Gaussian blurring filter. Images of the patient control group are produced via the same noise generation process; however, they also contain effects attributable to one of two distinct disease processes. These effects are visually-manifested through insertion of a circle in the top left side of the image (disease process $A$), or a circle at the bottom right-hand side (disease process $B$) (note that both diseases processes share the same Class 1 label). The circles are placed randomly with a maximum 5-pixel offset in each direction via uniform random sampling in order to add further visual variety. Samples images are shown in Fig. 2.

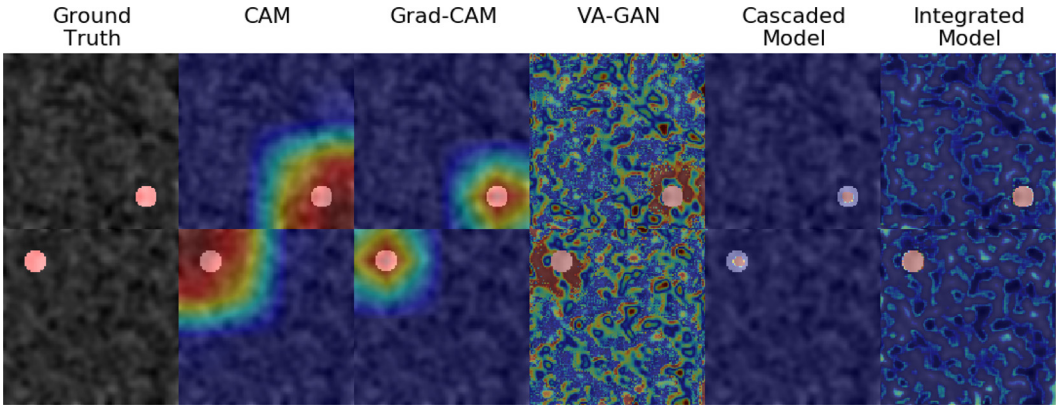


**Fig. 3.** Examples of visualization maps of the compared methods on synthetic data.

**Table 1**
IoU, Dice Scores and NCC Scores of evaluated methods on synthetic data.

| Method | IoU | Dice |
|---|---|---|
| CAM | 10.4 | 18.8 |
| gradCAM | 30.7 | 47 |
| VA-GAN | 872 | 92.8 |
| ICAM | 89.3 | 93.1 |
| VANT-GAN | 91.4 | 95.5 |

**Table 2**
Scores of evaluated methods on ADNI dataset

| | NCC Score on ADNI dataset | |
|---|---|---|
| Method | Mean | Std |
| CAM | 0.09 | 0.07 |
| gradCAM | 0.11 | 0.09 |
| VAGAN | 0.27 | 0.15 |
| ICAM | 0.30 | 0.28 |
| VANT-GAN | 0.36 | 0.35 |

#### 5.1.2. Evaluation Protocol

We divide the data on the basis of a 80-20 train/test set split, following the protocol of [8]. For quantitative evaluation, we calculate IoU and Dice score between the disease maps and the visual explanation. We use the maximum pixel value as a threshold to convert the visual explanation map into a binary mask. Following [8], we also employ the normalized cross correlation (NCC) measure between ground-truth maps and the predicted visual explanation maps.

#### 5.1.3. Results

Quantitative results with respect to the synthetic data are reported in Table 1 for all of the tested methods. Results clearly indicate the relative supremacy of the proposed method; examples of visual explanation maps for all of the methods are shown in Fig. 3. It is apparent that the CAM-based methods tend to focus on areas where the circles are distributed uniformly, and are unable to provide fine-grained visualization maps (the effect can clearly be observed from the visualization map of the CAM-based methods in Fig. 3). It is further apparent that VA-GAN produces noisy visualization maps due to its under-constrained mapping from unaligned noisy images; the noisy maps contain many false positives which degrade VA-GAN performance (this effect can be seen in the visual explanation map of VA-GAN from Fig. 3). Contrarily, the proposed method produces far more plausible visual explanation maps primarily due to the constrained CycleGAN-based mapping, VANT-GAN can thus better describe the input image w.r.t. the generated CI.

### 5.2. Experiments on Medical Imaging Data

#### 5.2.1. Datasets

**Alzheimer's Disease Neuroimaging Initiative dataset:** From the ADNI cohort, we selected 5778 3D T-1 weighted MR images of 1288 subjects with two of the labels: MCI (label 0) or AD (label 1). A 1.5T magnet is used to obtain 2839 of the total images with the rest of the images obtained using a 3T magnet. A number of subjects are converted from MCI to AD over the years, scanned at regular intervals. Although these correspondences are not used for the training here, however, we exploit their advantages. Examples of normal and abnormal images from the ADNI dataset are shown in Fig. 5.

Standard operations in the FSL toolbox are used to pre-process each of the images in data; this pre-processing includes reorientation, registering images to MNI space, cropping and correcting inhomogeneous fields. The ROBEX algorithm is then applied to skull-strip the images. Finally, the images are resampled to 1.3 mm3, followed by normalization to a range between -1 to 1. The final voxel size is 128x256x256.

**BraTs dataset:** The dataset contains brain MRIs classified into normal and tumorous classes. We preprocess the data to filter out MRI slices that contain the full brain. The dataset contains 3174 images where 2711 are tumorous and 463 are non-tumorous. We split each set into 80-20 train/test sets, resulting in 2538 training images and 636 testing images. The filtered slices are resized to 256 - 256 and the data normalized to the 0-to-1 range. We further increase the data size by performing run-time augmentation on training sets through random jittering and mirroring. For augmenting, the images are scaled to 286 - 286 and then randomly cropped to 256 - 256.

#### 5.2.2. Evaluation

We use the visual explanation maps generated by the networks for semantic segmentation of disease affected regions. We split each dataset into 80-20 train/test sets. To gauge the efficacy of the respective networks, we employ mean IoU and Dice coefficient (i.e. the standard metrics to evaluate semantic segmentation methods). To calculate these metrics, we convert the visual explanation maps into binary masks. The highest value of the explanation map is used as a threshold to convert the visual explanation map into a binary mask.

#### 5.2.3. Results

Table 2 sets out quantitative results for each of these experiments. The proposed method significantly outperforms the other

**Table 3**
IoU and Dice Scores of evaluated methods on BraTS datasets

| | BraTS dataset | |
|---|---|---|
| Method | IoU Score | Dice Score |
| CAM | 30.8 | 45.1 |
| gradCAM | 54.7 | 60.3 |
| ANT-GAN | 76.3 | 80.1 |
| VA-GAN | 89.5 | 93.2 |
| VANT-GAN | 91.4 | 94 |

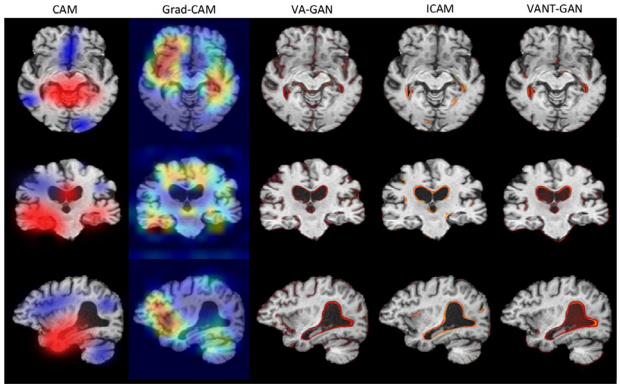


**Fig. 4.** Example visualization maps of the compared methods with respect to the ADNI dataset

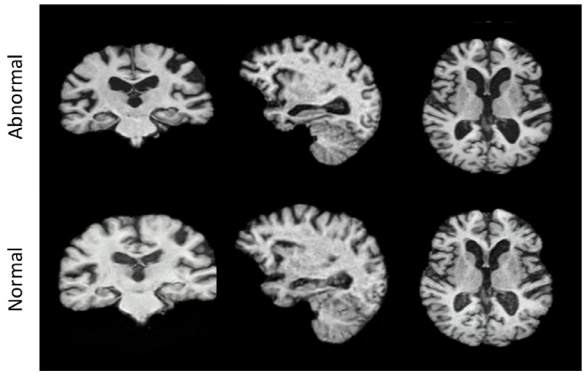


**Fig. 5.** Example of Normal and Abnormal Images from the ADNI dataset

methods. Examples of the visual explanation maps for the ADNI data are as depicted in Fig. 4.

The results and findings are consistent with the synthetic data. We believe that CAM-based methods show limited performance as a result of focusing only on a minimal set of the most discriminative features while disregarding the rest. The visual explanations of the CAM method are hence noisy, low resolution and often falsely-oriented. gradCAM improves on the explanations of the CAM approach in terms of noise and resolution. However, the gradCAM explanation region is much smaller than that of the ground truth. VA-GAN can detect edges around the infected area; however, the explanation is appreciably noisy (especially so on zooming-in the visualization map). ICAM reduces the noise in the visual attribution map; however, the explanation is not exclusive in terms of its coverage of the affected region. The proposed VANT-GAN method, by contrast, outperforms other methods in its exclusive coverage of the affected region.

We also compare our method with the baseline ANT-GAN method, as shown in the visual results in Fig. 6. For this comparison, we use ANT-GAN to translate an abnormal image into normal counterpart and then subtract the abnormal image from the normal image to obtain the visual attribution map. Finally, we apply

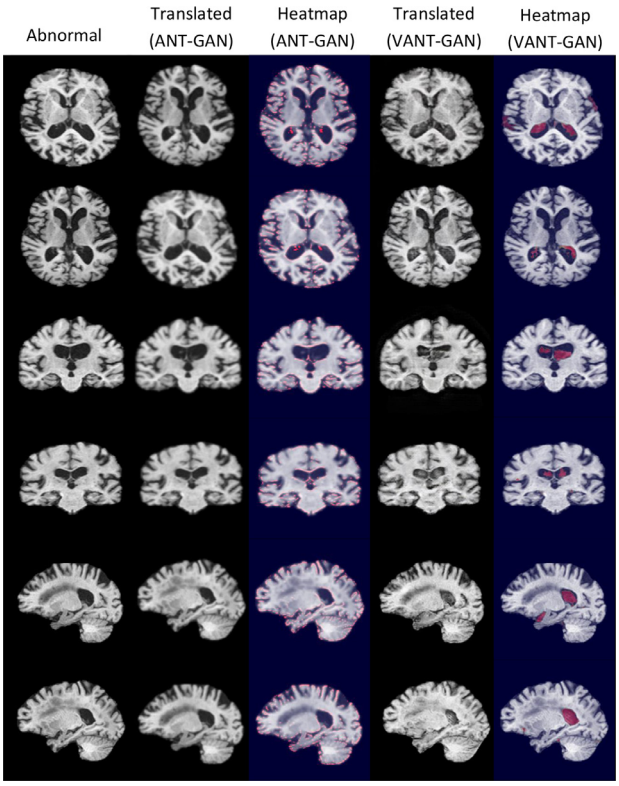


**Fig. 6.** Comparison of ANT-GAN with VANT-GAN

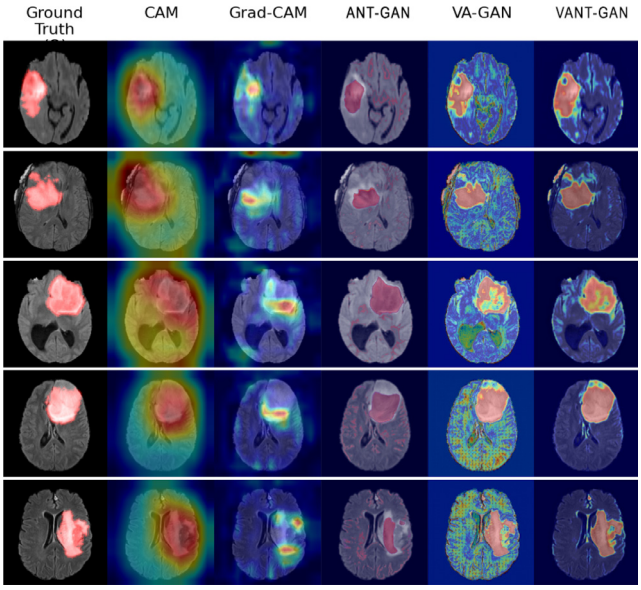


**Fig. 7.** Example visualization maps of the compared methods with the BraTS dataset

a threshold on the residual image to obtain the reported results. It may be seen that, as ANT-GAN does not explicitly learn a visual attribution map, it is unable to cope with minor changes (such as translation) that typically occur during the translation phase. Further, ANT-GAN cannot explicitly regularize the characteristics of the visual attribution map (since visual attribution is implicitly produced as by-product of ANT-GAN's abnormal-to-normal translation, we cannot explicitly regularize the shape of the visual attribution map).

The quantitative and visual results on BraTs dataset are shown respectively in Table 3 and Fig. 7. We notice that these results are consistent with our previous results on synthetic and ADNI datasets.

## 6. Conclusion

In this paper, a novel visual attribution (VA) technique is developed with respect to medical images (although intrinsically applicable to general images), one that leverages the capacity of cycle-consistent GANs in conjunction with the concept of the Residual GAN to generate counterpart normal images in relation to abnormal (i.e. diseased) input images. The resulting VANT-GAN model is thus capable of providing a conjectured 'healthy' image to medical practitioners in order to highlight the process by which automated disease classification is arrived at. The model thus deploys a cycle-consistent GAN architecture for joint learning of both a normal and a visual attribution map.

Experimental results demonstrate that, by contrast with back-propagation-based and pre-existing counterfactual VA techniques, the proposed method produces significantly more refined visual attribution maps for highlighting disease markers in the input image than the current state-of-the-art.

Because the proposed approach relies on translation across two domains, it intrinsically only caters for VA of a single disease. VA for multiple diseases potentially thus requires computationally expensive training of multiple individual models per disease. Future work will thus investigate the possibility of explicitly multiclass VA techniques. It will also be of interest to investigate whether spatial regularization of the generated map would further improve results and potentially also enable direct generation of the binary mask. Finally, it will be of interested to apply the proposed approach to Covid-19 datasets, in particular in relation to issue of 'long Covid' diagnosis.

## Declaration of Competing Interest

No conflict of interest exists. We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

## References


[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[2] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad–cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[3] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806 (2014).

[4] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, arXiv preprint arXiv:1706.03825 (2017).

[5] Z. Zhang, Y. Xie, F. Xing, M. McGough, L. Yang, Mdnet: A semantically and visually interpretable medical image diagnosis network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6428–6436.

[6] T. Nguyen-Duc, H. Zhao, J. Cai, D. Phung, Med-tex: Transferring and explaining knowledge with less data from pretrained medical imaging models, arXiv preprint arXiv:2008.02593 (2020).

[7] L. Sun, J. Wang, Y. Huang, X. Ding, H. Greenspan, J. Paisley, An adversarial learning approach to medical image synthesis for lesion detection, IEEE Journal of Biomedical and Health Informatics (2020).

[8] C.F. Baumgartner, L.M. Koch, K. Can Tezcan, J. Xi Ang, E. Konukoglu, Visual feature attribution using wasserstein gans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8309–8319.

[9] J. Akatsuka, Y. Yamamoto, T. Sekine, Y. Numata, H. Morikawa, K. Tsutsumi, M. Yanagi, Y. Endo, H. Takeda, T. Hayashi, et al., Illuminating clues of cancer buried in prostate mr image: Deep learning and expert approaches, Biomolecules 9 (11) (2019) 673.

[10] S.S. Han, M.S. Kim, W. Lim, G.H. Park, I. Park, S.E. Chang, Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm, Journal of Investigative Dermatology 138 (7) (2018) 1529–1538.

[11] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, arXiv preprint arXiv:1711.05225 (2017).

[12] C. Yang, A. Rangarajan, S. Ranka, Visual explanations from deep 3d convolutional neural networks for alzheimers disease classification, in: AMIA Annual Symposium Proceedings, 2018, American Medical Informatics Association, 2018, p. 1571.

[13] S. Rajaraman, K. Silamut, M.A. Hossain, I. Ersoy, R.J. Maude, S. Jaeger, G.R. Thoma, S.K. Antani, Understanding the learned behavior of customized convolutional neural networks toward malaria parasite detection in thin blood smear images, Journal of Medical Imaging 5 (3) (2018) 034501.

[14] C. Zhao, J. Han, Y. Jia, L. Fan, F. Gou, Versatile framework for medical image processing and analysis with application to automatic bone age assessment, Journal of Electrical and Computer Engineering 2018 (2018).

[15] I. Rio-Torto, K. Fernandes, L.F. Teixeira, Understanding the decisions of cnns: An in-model approach, Pattern Recognition Letters 133 (2020) 373–380.

[16] K. Gao, H. Shen, Y. Liu, L. Zeng, D. Hu, Dense-cam: Visualize the gender of brains with mri images, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–7.

[17] C. Dasanayaka, M.B. Dissanayake, Deep learning methods for screening pulmonary tuberculosis using chest x-rays, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization (2020) 1–11.

[18] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, D. Pfeiffer, Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization, Scientific reports 9 (1) (2019) 1–9.

[19] Y. Jang, J. Son, K.H. Park, S.J. Park, K.-H. Jung, Laterality classification of fundus images using interpretable deep neural network, Journal of digital imaging 31 (6) (2018) 923–928.

[20] H. Jiang, J. Xu, R. Shi, K. Yang, D. Zhang, M. Gao, H. Ma, W. Qian, A multi-label deep learning model with interpretable grad-cam for diabetic retinopathy classification, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, pp. 1560–1563.

[21] T. Zia, N. Bashir, M.A. Ullah, S. Murtaza, Softnet: A concept-controlled deep learning architecture for interpretable image classification, Knowledge-Based Systems (2022) 108066.

[22] P. Xia, H. Niu, Z. Li, B. Li, On the receptive field misalignment in cam-based visual explanations, Pattern Recognition Letters 152 (2021) 275–282.

[23] G. Bargshady, X. Zhou, P.D. Barua, R. Gururajan, Y. Li, U.R. Acharya, Application of cyclegan and transfer learning techniques for automated detection of covid-19 using x-ray images, Pattern Recognition Letters 153 (2022) 67–74.

[24] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[25] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

[26] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2794–2802.

[27] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, Learning from simulated and unsupervised images through adversarial training, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2107–2116.