Computers in Biology and Medicine 166 (2023) 107521

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Visual attribution using Adversarial Latent Transformations

Tehseen Zia [a,b,*], Abdul Wahab [a,b], David Windridge [c], Santosh Tirunagari [c], Nauman Bashir Bhatti [d]

[a] *COMSATS University Islamabad, Pakistan*
[b] *Medical Imaging and Diagnostics Lab, National Center of Artificial Intelligence, Pakistan*
[c] *Middlesex University, UK*
[d] *York University, Canada*

ARTICLE INFO

Keywords:
Visual attribution
Domain translation
Generative adversarial networks
Medical imaging
Explainable artificial intelligence

ABSTRACT

The ability to accurately locate all indicators of disease within medical images is vital for comprehending the effects of the disease, as well as for weakly-supervised segmentation and localization of the diagnostic correlators of disease. Existing methods either use classifiers to make predictions based on class-salient regions or else use adversarial learning based image-to-image translation to capture such disease effects. However, the former does not capture all relevant features for visual attribution (VA) and are prone to data biases; the latter can generate adversarial (misleading) and inefficient solutions when dealing in pixel values. To address this issue, we propose a novel approach Visual Attribution using Adversarial Latent Transformations (VA2LT). Our method uses adversarial learning to generate counterfactual (CF) normal images from abnormal images by finding and modifying discrepancies in the latent space. We use cycle consistency between the query and CF latent representations to guide our training. We evaluate our method on three datasets including a synthetic dataset, the Alzheimer's Disease Neuroimaging Initiative dataset, and the BraTS dataset. Our method outperforms baseline and related methods on all datasets.

1. Introduction

The capability to identify disease effects at an individual level, referred to as Visual Attribution (VA) [1], is essential for various medical applications. These include utilizing VA for weakly supervised localization or segmentation of diseases [2–8], gaining a deeper understanding of disease effects [9,10], monitoring the progression and severity of diseases [11], and identifying different subtypes of complex diseases like Alzheimer's and schizophrenia [1].

Currently, the most common approaches for VA of medical images use deep neural network (DNN) classifiers for either applying forward propagation (or activation) to identify the regions of the input image responsible for the predictions, or using backpropagation to analyze the gradient of the prediction with respect to the input image [6,8,12–16]. However, these approaches have two limitations that can lead to uninformative and undesirable results in certain situations: 1) DNN classifiers are trained to minimize mutual information between inputs and outputs, which means they tend to rely on the fewest possible input features to make predictions. As a result, DNN classifiers may make decisions based on only a few salient regions of the input image, rather than considering the entire object of interest. This can be problematic in the context of medical image diagnosis where it is important to capture all of the disease effects present in an image [1,17,18]. 2) According to user studies, these methods are not as informative to humans as the simple nearest neighbors from the training set. Humans may prefer to see examples that are similar to the natural images rather than mere heatmaps or other visualizations. Counterfactual (CF) explanations, which produce an example that is similar to the explanation subject but predicted as a different category by the model, have been shown to be more useful to humans in understanding the diagnosis [19–24]. CF explanations have also been advocated by social scientists as a preferred mode of explanation [25].

To address the limitations, adversarial learning-based visual attribution (VA) methods have been proposed [1,17,18]. These methods use techniques from domain translation to translate abnormal images into their normal counterparts, and then identify the discrepancies (as VA maps) between the two. However, it is often impractical to obtain contemporaneous normal and abnormal image pairs, so these VA methods perform the abnormal-to-normal mapping in an unsupervised way using cyclic-consistency GANs. These approaches learn a discrepancy map that, when added to the abnormal image, makes it indistinguishable from the normal image. However, there are several limitations to these approaches as well. Since they directly optimize

[*] Correspondence to: Department of Computer Science, COMSATS University Islamabad, Park Road, Tarlai Kalan, Islamabad 45550, Pakistan.
E-mail address: tehseen.zia@comsats.edu.pk (T. Zia).

https://doi.org/10.1016/j.compbiomed.2023.107521
Received 20 February 2023; Received in revised form 2 September 2023; Accepted 19 September 2023
Available online 23 September 2023

for perturbations in the input space, they may lead to adversarial solutions (i.e., discrepancy maps could be adversarial perturbations) that manipulate the predictions of the discriminator with imperceptible changes leading to the generation of a noisy VA map [26]. Adversarial examples are typically off the data manifold, where DNN-based discriminators can be fooled because they do not generalize to data that has never been seen in training. We also show in the results that these methods optimize for minimal changes in the abnormal images in order to generate the normal CF and do not fully attribute disease-affected regions. Consequently, there is a need for research that addresses these limitations and develops more effective and efficient VA methods for medical image diagnosis.

In this paper, we propose a novel approach for visual attribution (VA) of medical images using generative adversarial networks (GANs) that optimizes for a nonlinear transformation in the latent space rather than directly in the image space. This potentially enables the model to learn a more general function, as it is not tied to specific pixel values but rather deals with features such as C3LT (Cycle-consistent counterfactuals by latent transformations) [27]. Our proposed transformation morphs the latent code of an abnormal image into a residual latent vector that, when added to the latent code of the abnormal image, can decode a counterpart normal image that looks similar to the abnormal image but has semantically meaningful, perceptible differences that allow the discriminator to classify it as a normal image. We adopt a cycle-consistency principle [28], in which an inverse mapping and a cycle consistency (i.e., forwards–backwards) loss is introduced to the GAN to tackle tasks for which paired training data does not exist. We demonstrate that it is possible to generate VA maps using abnormal-to-normal translation in the latent space. Although our approach is based on the formulation for generating VA maps as proposed in [1,17] (i.e., discrepancy maps that, when subtracted from the abnormal image, make it indistinguishable from the counterpart normal image), we learn the abnormal-to-normal translation in the latent space rather than the image space. This differs from [27], in which an input image is explicitly transformed into a counterpart (i.e., counterfactual) image. Instead, we learn an implicit transformation through the VA map. Our approach offers a promising solution for addressing the limitations of current VA methods for medical image diagnosis and has the potential to improve the accuracy and efficiency of diagnostic tools utilizing AI.

The contribution of the paper is to propose a novel approach, called VA2LT, for the visual attribution of medical images using generative adversarial networks (GANs). Unlike previous methods such as VA-GAN, VANT-GAN and C3LT, which performs abnormal-to-normal transformations in pixel space, VA2LT optimizes this transformation in latent space. By adding the latent code of an abnormal image into a latent map code, VA2LT generates a counterpart normal image that exhibits semantically meaningful and perceptible differences. This approach restricts the generator from making adversarial changes at the pixel level and instead focuses on capturing semantic changes based on feature vectors. The use of the latent space allows for computationally efficient generation of counterfactual images with reduced overfitting compared to the pixel space.

2. Related work

2.1. Visual attribution in medical images

Class Activation Maps (CAM) are commonly used for visual attribution (VA) in medical images. Originally, CAM used global pooling to identify important image parts for CNN's decision [14]. Grad-CAM improved this by using gradient-based feature attribution [15], and guided grad-CAM further enhanced the maps using guided backpropagation [29]. Despite their widespread adoption, CAM-based methods [30–36] have limitations. They often produce low-resolution visualizations, necessitating post-processing [17]. The classifiers in these methods prioritize highly discriminative features, neglecting low-discrimination ones, resulting in imperfect VA [1]. Misalignment issues can arise due to VA upsampling [37].

Overcoming these issues, a GAN-based method with Wasserstein loss was proposed [1] but produced artifacts due to the lack of alignment between normal and abnormal images [17,18]. VANT-GAN improved alignment using cyclic-consistency loss, but direct optimization of input perturbations could lead to adversarial solutions [27]. The adversarial examples might deceive CNN-based discriminators unaccustomed to off-manifold examples. Additionally, it focuses on minimal changes for abnormal-to-normal translation, not fully attributing disease-affected areas.

In contrast, VA2LT optimizes transformations in the latent space, yielding semantically meaningful differences and avoiding pixel-level manipulation. VA2LT generates maps from feature vectors, enabling better semantic understanding. Latent space's lower dimensionality enhances computational efficiency and reduces overfitting compared to pixel space.

2.2. Visual attribution for segmentation of medical images

Recent research in medical image segmentation has utilized transformer-based networks [38–42]. Some studies have focused on brain tumor segmentation using datasets like BraTS [38,39], while [40] evaluated approaches on various medical datasets such as DSB18, TNBC, and Kvasir-SEG. TransUnet integrates CNNs with transformers [41], achieving state-of-the-art performance. HCT-Net [42] combines U-shaped CNNs with transformers, optimizing via neural architecture search. However, these methods rely on ground truth labels for segmentation, which can be unavailable for certain datasets like ADNI. In segmentation evaluation, methods have been proposed to assess quality without ground truths. REC-Net [43] reconstructs images from masked versions for quality assessment. Another study [44] uses an ensemble of segmentations to estimate probabilistic ground truth for nuclei segmentation in MTI. However, these methods often rely on pre-trained models and subjective evaluation.

The proposed method employs visual attribution for generating disease maps, crucial in medical images with limited pixel-level labeling. Visual attribution can enhance segmentation accuracy with reduced expert input compared to traditional methods relying on ground truths.

2.3. Counterfactual visual explanation

Most prior work on counterfactual visual explanation has focused on natural images. One early approach, [20], generates counterfactual explanations (CFs) by exhaustively searching for feature replacements between the latent features of the query image and the CF image. However, this method is slow and the generated CF images may not be representative of the data manifold. Another approach, [24], uses attribution maps to identify informative regions for the query or CF classes, but this method does not generate CF images, and the explanations are limited to highlighting regions on images. It also requires the use of CF images to render the explanations. In contrast, our work does not rely on pre-selected CF images and produces explanations in the form of counterfactual images that exist on the data manifold. [19] introduces a contrastive explanation framework that finds minimal and sufficient input features or perturbations to justify a prediction or change the classifier's prediction from the query class to a CF one. However, this approach does not provide explanations in terms of counterfactual instances and the generated explanations may be adversarial and off the data manifold. [45] generates CF images by filling a masked area on the input with a generator, but this method requires masks of diseases, which may not be available in all cases. [46] builds a graph of candidates from the training set and selects CFs from it that respect the underlying data distribution, but this assumes that a counterfactual example for the query image can be found in the training set, which may not be true for medical images. [47], on the other hand, uses cycle-consistent adversarial training to learn an unpaired abnormal-to-normal translation, but this method requires post-processing to generate visual explanations (VAs) using CF normal images. [17], similar to [47], uses cycle-consistent adversarial learning to generate VA maps.

These approaches, which directly optimize for perturbations in the input space, may produce adversarial solutions that are off the data manifold and may fail to render semantically meaningful CFs or VAs. [48] uses a cycle-consistency GAN to generate CFs for explaining the decisions of a medical image classifier, but this method has a different goal than ours and does not produce VAs of the query image in terms of its CF. [27], on the other hand, learns a latent transformation that generates visual CFs by steering in the latent space of generative models, but it translates the query image directly into a CF (rather than translating using the generated VA map) and cannot produce VAs of the query image with respect to the CF. Our work is distinct in that we aim to generate VAs of the query (abnormal) image in terms of its CF (normal) image, rather than directly explaining the decision of a classifier.

2.4. Impact of proposed work on related domains

The existing research on counterfactual visual explanation presents various contributions and limitations. Our method, Visual Attribution using Adversarial Latent Transformations (VA2LT), introduces a novel approach focusing on medical imaging data. VA2LT enhances disease detection and understanding by identifying and transforming salient regions of abnormal medical images. Unlike segmentation methods such as Multilevel Thresholding Image Segmentation (MTIS) or Improved Ant Colony Optimization Algorithm (XMACO) [49], VA2LT goes beyond delineation, actively converting abnormal regions into counterfactual normal states. This transformation process can provide clinicians with a deeper comprehension of disease attributes, potentially improving accuracy in diagnosing conditions like COVID-19 [49,50].

Moreover, VA2LT's benefits extend to other healthcare domains, such as colonoscopy for colorectal cancer detection. VA2LT's ability to discern and modify salient regions could aid in distinguishing between polyps and healthy tissues, in contrast to methods like NeutSS-PLP [51] that focus on specific polyp region extraction. VA2LT's adversarial learning in a latent space offers a holistic perspective on abnormality transformations, potentially enhancing accuracy for earlier intervention and better patient outcomes.

Our proposed VA2LT method, with its unique adversarial learning approach, differentiates itself from existing methods primarily centered around image segmentation optimization or specific region extraction. While VA2LT demonstrates promising results in detecting disease indicators in neuroimaging data, its potential extends to improving the quality and efficacy of various medical imaging modalities and disease domains when integrated with current techniques.

3. Methodology

3.1. Generating VA via transformation in the latent space

In this study, we use a generative model to create a VA map to transform an abnormal image into normal counterpart examples that are close to the data manifold. To do this, we utilize the concept of steerability in the latent space of generative models and propose learning a transformation in the latent space to generate the counterparts. We adopted an autoencoder framework as a fundamental component of our study. The encoder network utilized a CNN architecture with two convolutional layers, followed by batch normalization and ReLU activation. The final output was flattened and connected to two fully connected layers to compute the latent space representation. The decoder network, also based on a CNN architecture, consisted of a fully connected layer, reshaping, batch normalization, and ReLU activation. It further employed two transposed convolutional layers with ReLU activation, culminating in the final layer using hyperbolic tangent activation. The model was optimized using stochastic gradient descent (SGD) with mean squared error (MSE) as the loss function for both the encoder and decoder.

Given an abnormal image (x) belonging to class c in the training set, a target counterfactual normal image $(\acute{x})$ from class ć, a VA generator G, a discriminator D for normal images, we aim to learn a non-linear transformation (VA) map in the generator's latent space that maps the latent code of the abnormal (z_x) to a latent code of normal $(z_{\acute{x}})$:

$$g^* = arg\ min_g \mathbb{E}_x \left[\mathcal{L}_{cls}(D(\acute{x})) + \mathcal{L}_{prx}(\acute{x}, x)\right] \tag{1}$$

$$\text{s.t.}\quad \acute{x} = De(z_{\acute{x}}),\quad z_{\acute{x}} = G(z_x) + z_x \quad z_x = En(x)$$

where $\mathcal{L}_{cls}$ is the classification loss favoring that the generated normal x́ belongs to class ć and $\mathcal{L}_{prx}$ is the proximity loss encouraging x́ to be proximal to the input x. To obtain the latent code z_x from abnormal image x, we train an encoder $En(.) : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^m$. We also decode latent code $z_{\acute{x}}$ and $G(z_x)$ into, respectively, a normal counterfactual and VA map by training a decoder $De : \mathbb{R}^m \rightarrow \mathbb{R}^{C\times H\times W}$. (Note that the encoder and decoder models used in this study are trained with combined abnormal and normal images).

One major difference between our approach and previous work on visual attribution (VA) is the way in which VA maps are generated and used. Previous VA methods generated VA maps directly from abnormal images and added these to the images to create a normal counterfactual. In contrast, our method encodes both abnormal and normal images into latent vectors and then applies the VA formulation as described [1,17]. By approaching the VA problem in latent space rather than image space, our method captures the full range of disease-affected regions. This differs from [27] in that it reformulates the process of generating counterfactual explanations as an indirect VA-based translation rather than generating a counterfactual class directly from the input image. Specifically, our method generates a VA map using the G function (unlike [27], where G generates counterfactuals directly), which translates the abnormal latent code into a normal one.

3.2. Generating the normal counterfactual using cycle-consistency

We aim to find a transformation G that translates abnormal images into counterpart normals. However, finding such a transformation is highly under-constrained, and there may be multiple solutions to the optimization problem that are equally valid. To address this issue, we introduce regularization into the optimization process by incorporating cycle consistency between the latent codes of abnormal and counterpart normal images. This is achieved by introducing an additional transformation $h : \mathbb{R}^m \rightarrow \mathbb{R}^m$ to estimate the inverse of g to transform generated normal latent vector back to the input abnormal latent vector, i.e. $z_x \approx z_x^{cyc}$ where $z_x^{cyc} = h(g(z_x))$. We define the cycled query image as $x^{cyc} = \acute{G}(z_x)$ and add the cycle loss to the objective function (1) (note, here, that x^{cyc} is cycled abnormal whereas x́ is the generated normal counterfactual). Ǵ(.) is thus distinct from G(.) and optimizes the following objective:

$$h^* = arg\ min_g \mathbb{E}_{\acute{x}} \left[\mathcal{L}_{cls}(\acute{D}(\acute{x})) + \mathcal{L}_{prx}(\acute{x}, x)\right] \tag{2}$$

$$\text{s.t.}\quad x = De(z_x),\quad z_x = \acute{G}(z_{\acute{x}})$$

Note that this objective differs from objective 1 in that the latent vectors of normals are directly transformed into the latent vectors of the corresponding abnormal image. The direct transformation from normal-to-abnormal is performed because the VA map has already been generated. By incorporating this cycle loss term, we are able to constrain the optimization process and improve the robustness and accuracy of our results.

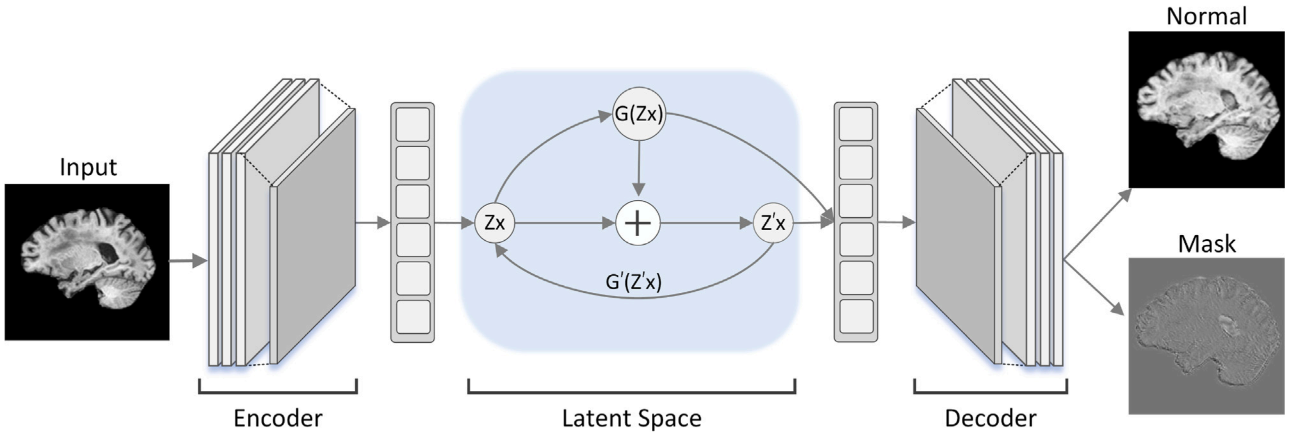

Fig. 1. VA2LT model diagram with an example image from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. The encoder is used to encode the input image into latent vector z_x, which is then transformed into the CF latent vector $z_{\acute{x}}$ using visual attribution (VA) vector $G(z_x)$ as $G(z_x) + z_x$. The map and CF latent vectors are then decoded into the VA map and CF image using the decoder.

Finally, we formalize the main objective of our method, Visual Attribution using Adversarial Latent Transformations (VA2LT). For cycle consistency loss, VA2LT requires access to samples from both the abnormal and normal classes. Given an image $x \in X_c$ from a training set of abnormal images, an image $\acute{x} \in X_{\acute{c}}$ from training set of normal images, VA2LT learns transformations g^* and h^*,

$$g^*, h^* = \text{argmin}_{g,h}\, \mathbb{E}_x[\mathcal{L}_{va2lt}(x, g, h)] + \mathbb{E}_{\acute{x}}[\mathcal{L}_{va2lt}(\acute{x}, h, g)] \tag{3}$$

Where

$$\begin{aligned}\mathcal{L}_{va2lt}(x, g, h) \;=\; & \mathcal{L}_{cls}(D(\acute{x})) + \mathcal{L}_{prx}(\acute{x}, x) + \mathcal{L}_{cyc}(x^{cyc}, x) \\ & + \mathcal{L}_{adv}(\acute{x}, x^{cyc})\end{aligned}$$

$$s.t.\ \acute{x} = D(z_{\acute{x}}),\ \ z_{\acute{x}} = G(z_x) + z_x \qquad z_x = En(x),$$

$$x^{cyc} = De(z_x^{cyc}), z_x^{cyc} = h(z_{\acute{x}})$$

VA2LT learns to transform between the abnormal and normal classes simultaneously; hence, the abnormal and normal notations are interchangeable. For the sake of conciseness, we omit the formal definition of $\mathcal{L}_{va2lt}(\acute{x}, g, h)$. The architecture of our proposed VA2LT is shown in Fig. 1. Details of the loss function in Eq. (3) are given below.

- Classification loss ($\mathcal{L}_{cls}$) promotes the classification of the generated normal examples as belonging to the normal class. We utilize the Negative Log-Likelihood loss for this purpose,

  $$\mathcal{L}_{cls} = -log(D(\acute{x}))$$

  where $D(\acute{x})$ is output of the discriminator for the normal image x́.
- Proximity loss ($\mathcal{L}_{prx}$) encourages the generated normal example to be similar to the input abnormal image according to some distance metric, specifically by promoting normal images that are proximal to the query abnormal image. To achieve sparsity in the changes between the abnormal image and the normal, we utilize an L1 loss term for the proximity loss. Additionally, we use entropy and smoothness losses ($\mathcal{L}_{entr}$and$\mathcal{L}_{smth}$) on the absolute difference between the abnormal and generated normal images to encourage more localized and sparse changes.

  $$\mathcal{L}_{prx} = \|x - \acute{x}\|_1 + \mathcal{L}_{entr}(x, \acute{x}) + \mathcal{L}_{smth}(x, \acute{x})$$
- Cycle-Consistency Loss ($\mathcal{L}_{cyc}$) ensures that the latent codes for the abnormal and normal classes are consistent with each other,

  $$\mathcal{L}_{cyc} = \|z_x - z_{\acute{x}}\|_1$$
- Adversarial loss ($\mathcal{L}_{adv}$) helps the generated normal images and cycled images to be similar to the original data by using the discriminator to guide them towards the manifold of the original data,

  $$\mathcal{L}_{adv} = log(1 - D(z_x^{cyc})) + log(1 - D(z_{\acute{x}}))$$

3.3. *Inference using VA2LT*

To use the model at inference time, the following steps are taken:

1. An abnormal image x is input into the pre-trained encoder to obtain its latent vector $z_x = En(x)$. We employed distinct autoencoder models for each specific application, such as one for BraTS and another for ADNI, whereby the comprehensive details regarding the model architecture and training process can be found in Section 3.1. Ablation studies concerning the autoencoder model are presented in Section 5. In this particular task, we solely utilized the encoder component of the autoencoder model. Specifically, we input an image of dimensions 256 × 256 (for BraTS and ADNI) and 128 × 128 (for Synthetic dataset) into the model, which subsequently generates a vector of size 256.
2. The latent vector is input into the generator G to obtain latent vectors for the VA map $G(z_x)$ and latent vector for normal image $G(z_x) + z_x$. The details of this model (VA2LT) are given in Section 3.2. This is essentially a Cycle-GAN model optimized to generate latent of normal images, given the latent vectors of abnormal images. More specifically, the model takes a vector of length 256 and generates a vector of the same length.
3. The latent vectors for the VA map and normal image are decoded using a decoder to obtain the VA map $De(G(z_x))$ and normal image $De(G(z_x){+}z_x)$. In this case, the decoder component of the autoencoder model is responsible for reconstructing the generated latent normal images from the latent space representation.

4. Experiments

We compare the performance of the proposed VA2LT method with several other visual explanation methods (CAM, gradCAM, VA-GAN, VANT-GAN, iGOS++ and C3LT) on three datasets: a synthetic dataset and two medical imaging datasets (ADNI and BraTS). Most of the tested methods are based on GAN-based VA map generation, except for CAM, gradCAM and iGOS++ which utilize classification networks. The performance of the methods is evaluated using various evaluation

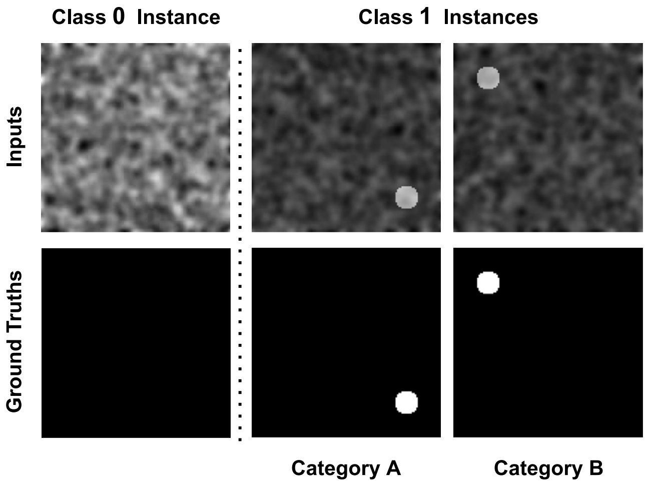

Fig. 2. Synthetic data examples: left of the dotted line are samples of Class 0 (i.e., the normal class) and right of the dotted line are samples of Class 1 (i.e., the disease class). The upper row shows the input and the bottom row shows the ground truth.

metrics, including the Dice Coefficient, Intersection over Union (IoU), normalized cross correlation (NCC), and FID scores. For the synthetic and BraTS datasets, these metrics are calculated using available ground truths, while for the ADNI dataset, the NCC score is used to evaluate the models as ground truths are not available. The discriminator architecture used in all of the tested methods is similar, with the exception of CAM, gradCAM and iGOS++, where the last two layers are replaced with a global average pooling layer and dense prediction layer to create class-specific activation maps for visual explanation.

4.1. Evaluation on synthetic dataset

4.1.1. Dataset and evaluation protocol

In this study, we evaluated the proposed and benchmark approaches on a synthetic dataset consisting of 10,000 128 × 128 images, which were divided into two label classes. The first class represents the healthy control group and the second class represents the patient group. The images in the healthy control group were generated by convolving random IID Gaussian noise with a Gaussian blurring filter, while the images in the patient group were produced using the same noise generation process but also included effects due to one of two distinct disease processes. These effects were visualized through the insertion of a circle on either the top left or bottom right side of the image, with a maximum 5-pixel offset in each direction. The resulting images are shown in Fig. 2.

We divided the data into a training set and a testing set using an 80–20 split, following the protocol of [8]. To evaluate the performance of our approach quantitatively, we calculated the Intersection over Union (IoU), Dice and FID scores between the disease maps and the visual explanation maps. We used the maximum pixel value as a threshold to convert the visual explanation maps into binary masks. In addition, we employed the normalized cross correlation (NCC) measure between the ground-truth maps and the predicted visual explanation maps, as described in [8].

4.1.2. Results

The results of the experiments on the synthetic data are reported in Table 1 for all of the tested methods. These results indicate that the proposed method outperforms all of the benchmark methods. Examples of the visual explanation maps produced by all of the methods are shown in Fig. 3.

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

Method	IoU (%)	Dice (%)	FID
CAM	10.4	18.8	225.76
Grad-CAM	30.7	47.0	138.57
VA-GAN	87.2	92.8	103.78
iGOS++	52.0	59.3	59.47
C3LT	67.2	73.7	56.38
VANT-GAN	89.4	93.5	52.32
VA2LT	91.7	96.0	27.85

It is notably apparent that our method pays attention to both the foreground and background of the image. For instance, the visualization of synthetic disease is smoother and more precise compared to the other methods. Moreover, our proposed method generates high-quality explanations in an optimal way by using a meaningful nonlinear transformation in the latent space. Visual results show that the CAM-based methods tend to focus on areas where the circles are distributed uniformly and are not able to provide detailed visual explanation maps. The VA-GAN method produces noisy visual explanation maps due to the under-constrained mapping from unaligned noisy images, which leads to many false positives and degraded performance. In contrast, the proposed method produces more accurate visual explanation maps due to the constrained CycleGAN-based mapping. Compared to VANT-GAN, which deals with image-to-image translation, latent space mapping enables the model to learn a more general function as it is not tied to specific pixel values. This results in better generalization performance and the ability to translate images (both synthetic and actual medical images) realistically. Another advantage of working in latent space is that it typically uses a smaller and lower-dimensional representation of the data compared to image space. This makes it easier to learn a function that maps between the image space and the latent space. VA2LT produces far more reasonable explanations, primarily due to the meaningful latent vector and, secondarily, due to constrained CycleGAN-based mapping in the latent space.

4.2. Evaluation on brats dataset

4.2.1. Dataset and evaluation protocol

The brain tumor dataset was collected as part of the Multimodal Brain Tumor Segmentation (BRATS) 2017 challenge [F, G]. This data consists of both abnormal (tumorous) and normal (non-tumorous) images as well as ground truth. The dataset comprises 463 normal and 3174 abnormal images. To increase performance, run-time data augmentation is performed by resizing images to 286 × 286, which are then randomly cropped to 256 × 256 size. Following the 80/20 rule, images are divided into train (2538) and test (636) sets. Further run-time augmentation is carried out through random jittering and mirroring.

In both the ADNI and BraTS datasets, there is a significant amount of black space around the brain. While cropping, we made sure that no part of the brain was cut off. Since our region of interest is always located within the brain, it is impossible to lose any lesion part. It is also important to note that the Alzheimer's images obtained from the ADNI dataset include the skull. We used the ROBEX algorithm to precisely remove the skull from the brain. This resulted in more black space around the brain, after which we cropped the images to the desired size of 256 x 256 pixels.

4.2.2. Results

Table 2 represents the quantitative results of the experiments conducted on the BraTS dataset, and the visual explanation of these quantitative results can be seen in Fig. 4. It can be seen that the

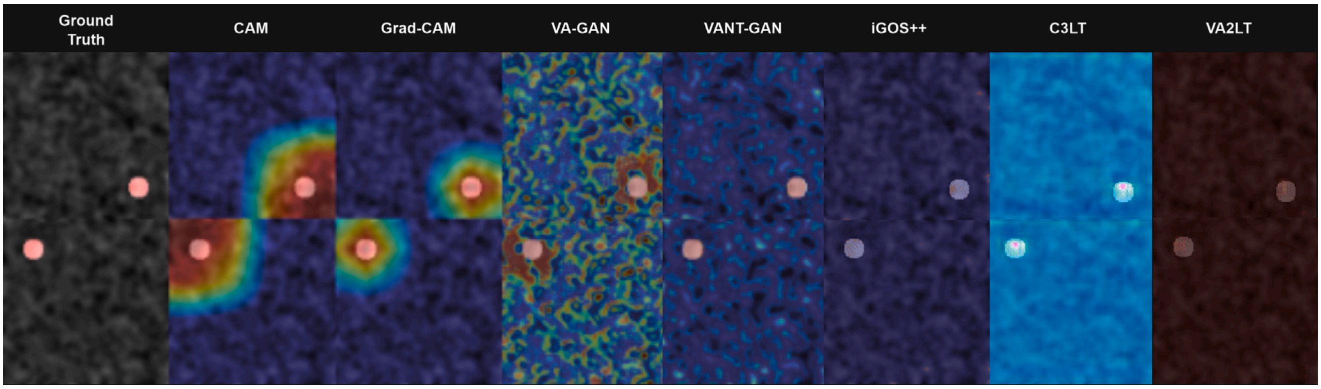

Fig. 3. Example visualization maps of the compared methods with the synthetic dataset.

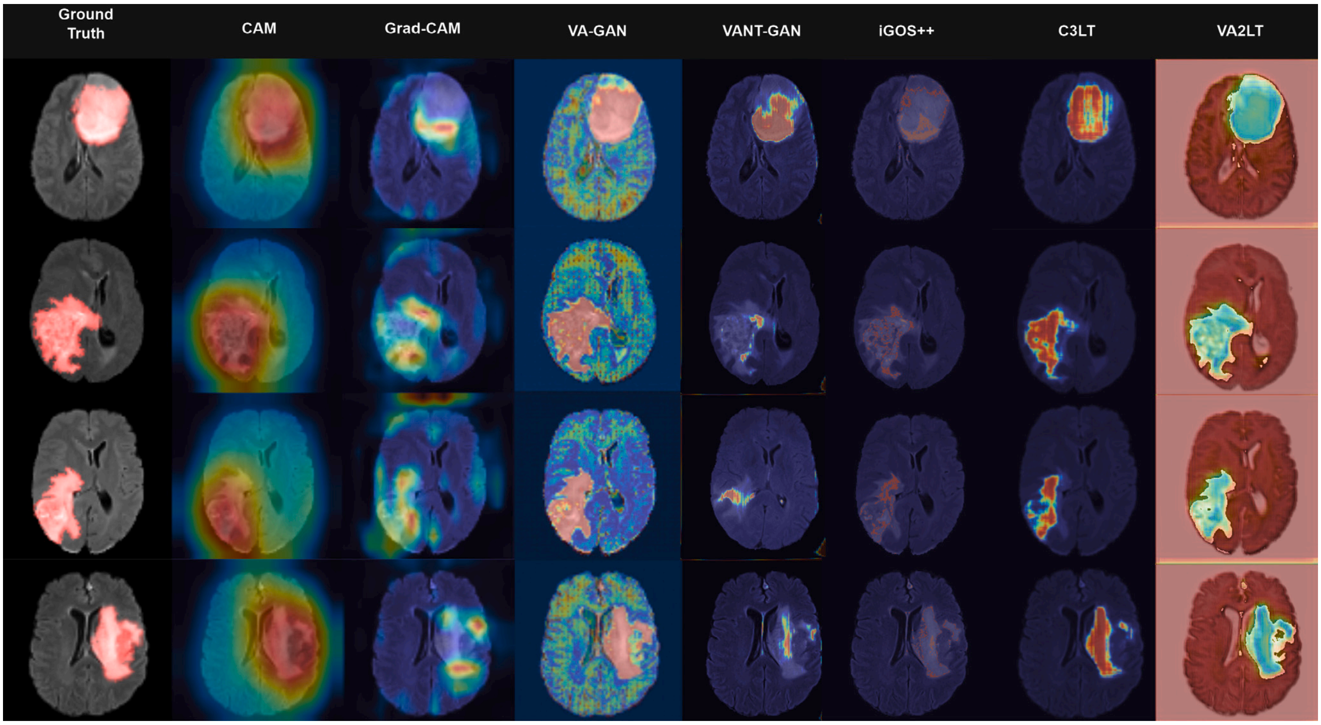

Fig. 4. Example visualization maps of the compared methods with the BraTS dataset.

visual explanation generated on the BraTS dataset is consistent with the findings of the synthetic dataset. Again, we generated the latent vector from the actual brain MRI and fed the translation network in the latent space with those vectors. It can be seen from Fig. 4, that the tumor region is precisely denoted in the final column, which indicates the explicative consistency of VA2LT.

CAM-based methods produce somewhat poor results as they focus only on a subset of the most discriminative features while ignoring the rest, which in turn leads to the generation of low resolution, noisy visual explanations. The explanation generated by GradCam is better than that of CAM; however, it covers a smaller area of the infected region as compared to the actual ground truth.

In terms of covering the whole region, VA-GAN generates a good explanation around the region but includes noise. VANT-GAN, by contrast, outperforms other methods in its exclusive coverage of the affected region; however, the edges are somewhat noisy as compared to our proposed method and have less coverage of the infected regions as compared to the proposed VA2LT model.

Table 2
IoU, Dice and FID Scores of evaluated methods on BraTS data.

Method	IoU (%)	Dice (%)	FID
CAM	30.8	45.1	217.29
Grad-CAM	54.7	60.3	132.86
VA-GAN	89.5	93.2	74.13
VANT-GAN	89.2	92.6	78.26
iGOS++	90.3	94.6	57.04
C3LT	91.7	94.9	46.82
VA2LT	92.0	96.2	28.61

While VA-GAN performs well in generating explanations that cover the entire region, it tends to include noise. On the other hand, VANT-GAN excels in providing coverage of the affected region compared to other methods. However, its edges may contain some noise, and it offers less coverage of the infected regions compared to our proposed VA2LT model.

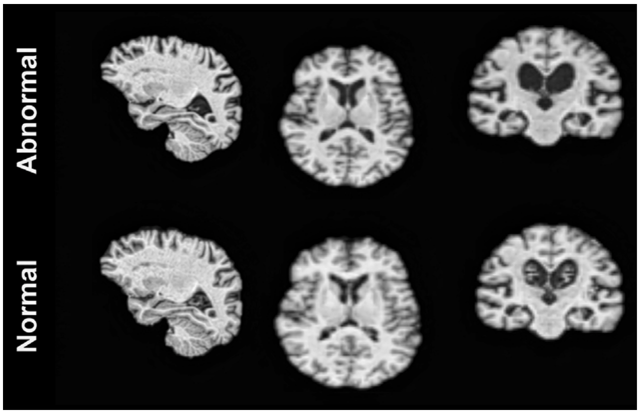

Fig. 5. Example abnormal and generated counterfactual normal images from the ADNI dataset.

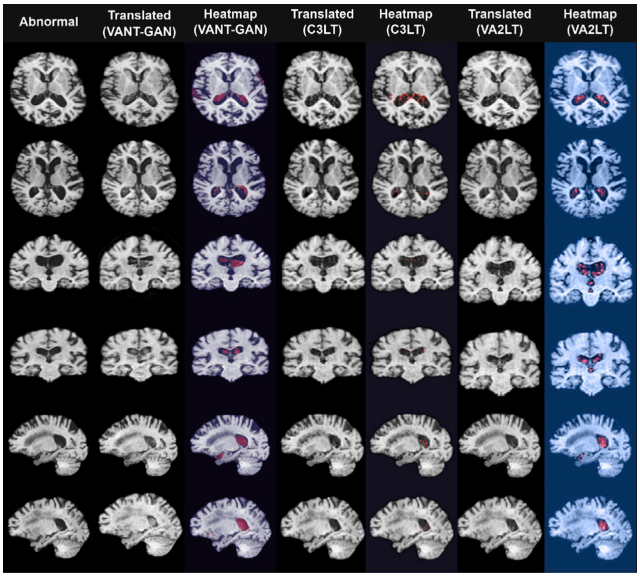

Fig. 6. The figure presents a comparative analysis using ADNI dataset visualizations. First column: factual input; columns 2, 4, and 6: method outcomes (VANT-GAN, C3LT, VA2LT), with accompanying heatmaps in adjacent columns.

Table 3
NCC scores of evaluated methods on ADNI data.

Method	Mean	Std
CAM	0.09	0.07
Grad-CAM	0.14	0.11
VA-GAN	0.27	0.16
C3LT	0.31	0.24
VANT-GAN	0.36	0.35
VA2LT	0.39	0.38

4.3. Evaluation on ADNI dataset

4.3.1. Dataset and evaluation protocol

The ADNI dataset used in this study comprises 5778 3D T-1 weighted MR images of 1288 subjects, labeled as either MCI (label 0) or AD (label 1). The images were obtained using a 1.5T magnet for 2839 of the images and a 3T magnet for the remaining images. The subjects were scanned at regular intervals, with some subjects converting from MCI to AD over time. These correspondences were not utilized for training but were exploited for their advantages. Standard pre-processing techniques, including reorientation, registration to MNI space, cropping, and correction of inhomogeneous fields, were applied to the images using the FSL toolbox. The ROBEX algorithm was then used to skull strip the images, which were subsequently resampled to 1.3 mm3 and normalized to a range between −1 and 1. The final voxel size for the images is 128 × 256 × 256. Examples of normal and abnormal images from the ADNI dataset are shown in Fig. 5.

For the BraTS and synthetic datasets, we used different evaluation protocols and used IoU, dice metrics and FID scores for evaluation. However, for the ADNI dataset, the ground truths are not available, and IoU and dice scores could not be calculated. Instead, we used the NCC score. We followed the same evaluation protocol for the ADNI dataset and computed the NCC score, as described in [1]. By maximizing the NCC score between the generated image and the input image, we ensured that the generated image was similar to the input image.

4.3.2. Results

Table 3 represents the quantitative results of the experiments conducted on the ADNI dataset, and the visual explanation of these quantitative results can be seen in Fig. 6. As discussed above, since ground truth is not available for ADNI, we follow the method outlined in [8], normalized cross correlation (NCC), to quantitatively measure the effectiveness of visual attribution. The results of these experiments, presented in Table 3, indicate that the VA2LT method has a significantly higher NCC score compared to the other evaluated methods. Additionally, the generated visual explanations on ADNI were consistent with those generated using both the synthetic and BraTS datasets.

Furthermore, the visual explanations of these quantitative results as seen in Fig. 6 indicate that the images generated by the VA2LT method exhibit a high level of detail, smooth edges, and clear structuring, all of which are essential for identifying subtle changes. The NCC score of VA2LT also supports this observation, indicating that the latent space in the proposed GAN model is able to effectively embed components corresponding to key VA-relevant features in the image space. Consequently, we believe that VA2LT represents a promising method for generating high-resolution counterfactual explanations. To run the experiments, we used a system that contains an Intel(R) Xeon (R)320 E5-2630 v4 CPU running at 2.2–3.1 GHz, 128 GB of RAM, and an Nvidia Titan X (Pascal) GPU with 12 GB of memory.

5. Ablation studies

Our research utilized an autoencoder framework. The encoder network employed a Convolutional Neural Network (CNN) with two convolutional layers, followed by batch normalization and ReLU activation. The resulting output was flattened and connected to two fully connected layers to compute the latent space representation. On the other hand, the decoder network also adopted a CNN architecture, starting with a fully connected layer, followed by reshaping, batch normalization, and ReLU activation. It further utilized two transposed convolutional layers with ReLU activation, ultimately concluding with the final layer using hyperbolic tangent activation. To optimize the model, we employed stochastic gradient descent (SGD) with mean squared error (MSE) as the loss function for both the encoder and decoder.

We explored the influence of different latent vector sizes on the performance of the autoencoder architecture. The results presented in Fig. 7 indicate that a latent size of 256 achieved the highest structural similarity index measure (SSIM) score. Visual results in Fig. 8 show reconstructed images from the autoencoder model against different

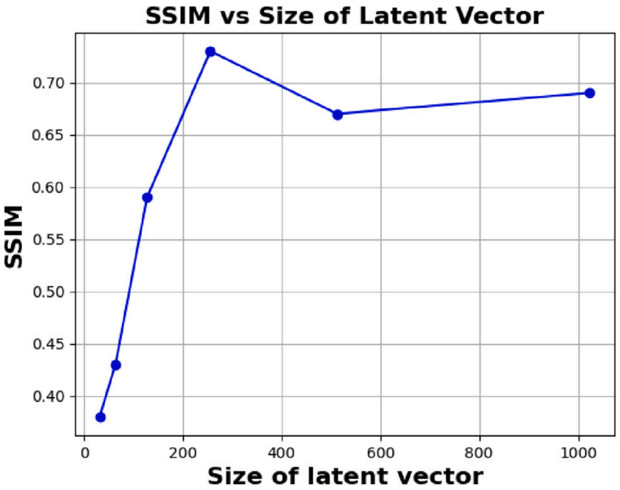

Fig. 7. Graph of the structural similarity index measure (SSIM) against the size of the latent vectors on a synthetic dataset.

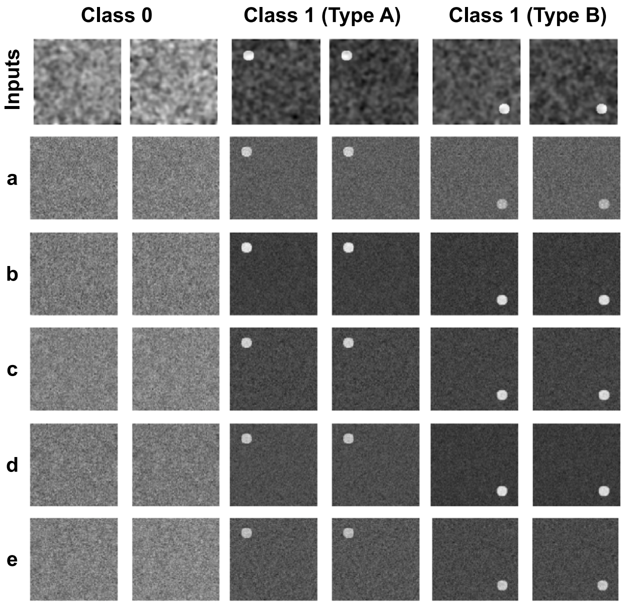

Fig. 8. Example visualizations show the relative magnitudes of latent vectors on the synthetic dataset. The initial row corresponds to the actual input data, whereas the succeeding rows exhibit images that have been reconstructed from latent vectors of different sizes, specifically: row (a) corresponds to a dimensionality of 64, (b) to 128, (c) to 256, (d) to 512, and (e) to 1024.

sizes of latent vectors. Consequently, this latent vector size was selected for adoption in our study, ensuring optimal performance of the autoencoder model.

We performed a comparative evaluation of two prominent GAN-Based approaches (Wasserstein GAN and Cyclic-Consistency GAN) for generating counterfactuals on the BraTS dataset. The experimental analysis involves visual comparisons of counterfactual (CF) instances generated by both methods, as illustrated in Fig. 9. To validate the usefulness of the approach, we used statistical tests like the Shapiro test, which indicated a non-normal distribution with a significance level of p < 0.05. Further, we utilized FID scores to provide additional validation for the effectiveness of our approach; the results are depicted in Table 4.

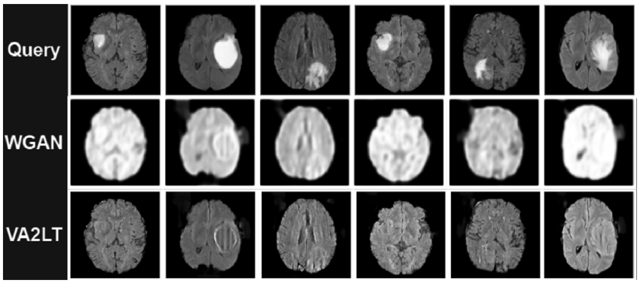

Fig. 9. Visual comparison of CF examples generated by VA2LT and WGAN on the BraTS dataset. CF examples from WGAN exhibit blur and scattered perturbations, while VA2LT produces more realistic CF images closely resembling the original query image.

Table 4
FID scores of evaluated methods on generated CF examples.

Method	FID scores	Shapiro test
WGAN	86.0	0.07
VA2LT(ours)	31.4	0.18

6. Conclusion

In this study, we proposed a novel visual attribution technique for medical images, VA2LT which uses cycle-consistency GANs to learn a transformation map in the latent space to generate a "healthy" counterpart image for an "unhealthy" input image, thus enabling medical practitioners to identify abnormalities more easily. Our experiments on synthetic, BraTS, and ADNI datasets demonstrate that the proposed method outperforms prior work in all metrics. Overall, we conclude that the VA2LT model presents a promising solution for visual attribution in medical imaging.

CRediT authorship contribution statement

Tehseen Zia: Conception and design, or analysis and interpretation of the data, Writing – original draft, Writing – review & editing. **Abdul Wahab:** Conception and design, or analysis and interpretation of the data, Writing – original draft, Writing – review & editing. **David Windridge:** Conception and design, or analysis and interpretation of the data, Writing – original draft, Writing – review & editing. **Santosh Tirunagari:** Conception and design, or analysis and interpretation of the data, Writing – original draft, Writing – review & editing. **Nauman Bashir Bhatti:** Conception and design, or analysis and interpretation of the data, Writing – original draft, Writing – review & editing.

Declaration of competing interest

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

Acknowledgments

All authors approved the version of the manuscript to be published.

References

[1] C.F. Baumgartner, L.M. Koch, K.C. Tezcan, J.X. 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.

[2] W.M. Gondal, J.M. Köhler, R. Grzeszick, G.A. Fink, M. Hirsch, Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images, in: 2017 IEEE International Conference on Image Processing, ICIP, IEEE, 2017, pp. 2069–2073.

[3] X. Feng, J. Yang, A.F. Laine, E.D. Angelini, Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 568–576.

[4] H.-L. Yang, J.J. Kim, J.H. Kim, Y.K. Kang, D.H. Park, H.S. Park, H.K. Kim, M.-S. Kim, Weakly supervised lesion localization for age-related macular degeneration detection using optical coherence tomography images, PLoS One 14 (4) (2019) e0215076.

[5] K.H. Kim, H.-W. Koo, B.-J. Lee, S.-W. Yoon, M.-J. Sohn, Cerebral hemorrhage detection and localization with medical imaging for cerebrovascular disease diagnosis and treatment using explainable deep learning, J. Korean Phys. Soc. 79 (3) (2021) 321–327.

[6] A. Jamaludin, T. Kadir, A. Zisserman, SpineNet: Automatically pinpointing classification evidence in spinal MRIs, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 166–175.

[7] Q. Zhang, A. Bhalerao, C. Hutchinson, Weakly-supervised evidence pinpointing and description, in: International Conference on Information Processing in Medical Imaging, Springer, 2017, pp. 210–222.

[8] C.F. Baumgartner, K. Kamnitsas, J. Matthew, T.P. Fletcher, S. Smith, L.M. Koch, B. Kainz, D. Rueckert, SonoNet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound, IEEE Trans. Med. Imaging 36 (11) (2017) 2204–2215.

[9] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, R. Garnavi, Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 250–258.

[10] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020) 56–67.

[11] M. Pennisi, I. Kavasidis, C. Spampinato, V. Schinina, S. Palazzo, F.P. Salanitri, G. Bellitto, F. Rundo, M. Aldinucci, M. Cristofaro, et al., An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans, Artif. Intell. Med. 118 (2021) 102114.

[12] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Is object localization for free?-weakly-supervised learning with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 685–694.

[13] P.O. Pinheiro, R. Collobert, From image-level to pixel-level labeling with convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1713–1721.

[14] 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.

[15] 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.

[16] S. Khorram, T. Lawson, L. Fuxin, iGOS++ integrated gradient optimized saliency by bilateral perturbations, in: Proceedings of the Conference on Health, Inference, and Learning, 2021, pp. 174–182.

[17] T. Zia, S. Murtaza, N. Bashir, D. Windridge, Z. Nisar, VANT-GAN: Adversarial learning for discrepancy-based visual attribution in medical imaging, Pattern Recognit. Lett. 156 (2022) 112–118.

[18] M. Nawaz, F. Al-Obeidat, A. Tubaishat, T. Zia, F. Maqbool, A. Rocha, MDVA-GAN: Multi-domain visual attribution generative adversarial networks, Neural Comput. Appl. (2022) 1–16.

[19] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, P. Das, Explanations based on the missing: Towards contrastive explanations with pertinent negatives, Adv. Neural Inf. Process. Syst. 31 (2018).

[20] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, Counterfactual visual explanations, in: International Conference on Machine Learning, PMLR, 2019, pp. 2376–2384.

[21] J. Moore, N. Hammerla, C. Watkins, Explaining deep learning models with constrained adversarial examples, in: Pacific Rim International Conference on Artificial Intelligence, Springer, 2019, pp. 43–56.

[22] R.K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 607–617.

[23] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harv. JL & Tech. 31 (2017) 841.

[24] P. Wang, N. Vasconcelos, Scout: Self-aware discriminant counterfactual explanations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8981–8990.

[25] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38.

[26] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, 2013, arXiv preprint arXiv:1312.6199.

[27] S. Khorram, L. Fuxin, Cycle-consistent counterfactuals by latent transformations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10203–10212.

[28] 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.

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

[30] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: Removing noise by adding noise, 2017, arXiv preprint arXiv:1706.03825.

[31] 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, 2017, arXiv preprint arXiv:1711.05225.

[32] C. Yang, A. Rangarajan, S. Ranka, Visual explanations from deep 3D convolutional neural networks for alzheimer's disease classification, in: AMIA Annual Symposium Proceedings, Vol. 2018, American Medical Informatics Association, 2018, p. 1571.

[33] 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.

[34] C. Dasanayaka, M.B. Dissanayake, Deep learning methods for screening pulmonary tuberculosis using chest X-rays, Comput. Methods Biomech. Biomed. Eng.: Imag. Visual. 9 (1) (2021) 39–49.

[35] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, D. Pfeiffer, Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization, Sci. Rep. 9 (1) (2019) 1–9.

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

[37] P. Xia, H. Niu, Z. Li, B. Li, On the receptive field misalignment in CAM-based visual explanations, Pattern Recognit. Lett. 152 (2021) 275–282.

[38] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, Y. Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI, Inf. Fusion 91 (2023) 376–387.

[39] Y. Xu, X. He, G. Xu, G. Qi, K. Yu, L. Yin, P. Yang, Y. Yin, H. Chen, A medical image segmentation method based on multi-dimensional statistical features, Front. Neurosci. 16 (2022) 1009581.

[40] Y. Li, Z. Wang, L. Yin, Z. Zhu, G. Qi, Y. Liu, X-Net: A dual encoding–decoding method in medical image segmentation, Vis. Comput. (2021) 1–11.

[41] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.

[42] Z. Yu, F. Lee, Q. Chen, HCT-Net: Hybrid CNN-transformer model based on a neural architecture search network for medical image segmentation, Appl. Intell. (2023) 1–17.

[43] L. Zhou, W. Deng, X. Wu, Robust image segmentation quality assessment, 2019, arXiv preprint arXiv:1903.08773.

[44] Z. Sims, L. Strgar, D. Thirumalaisamy, R. Heussner, G. Thibault, Y.H. Chang, SEG: Segmentation evaluation in absence of ground truth labels, 2023, bioRxiv, 2023–02.

[45] P. Samangouei, A. Saeedi, L. Nakagawa, N. Silberman, Explaingan: Model explanation via decision boundary crossing transformations, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 666–681.

[46] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, P. Flach, FACE: Feasible and actionable counterfactual explanations, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 344–350.

[47] L. Sun, J. Wang, Y. Huang, X. Ding, H. Greenspan, J. Paisley, An adversarial learning approach to medical image synthesis for lesion detection, IEEE J. Biomed. Health Inform. 24 (8) (2020) 2303–2314.

[48] S. Mertes, T. Huber, K. Weitz, A. Heimerl, E. André, GANterfactual—Counterfactual explanations for medical non-experts using generative adversarial learning, Front. Artif. Intell. 5 (2022).

[49] A. Qi, D. Zhao, F. Yu, A.A. Heidari, Z. Wu, Z. Cai, F. Alenezi, R.F. Mansour, H. Chen, M. Chen, Directional mutation and crossover boosted ant colony optimization with application to COVID-19 X-ray image segmentation, Comput. Biol. Med. 148 (2022) 105810.

[50] H. Su, D. Zhao, H. Elmannai, A.A. Heidari, S. Bourouis, Z. Wu, Z. Cai, W. Gui, M. Chen, Multilevel threshold image segmentation for COVID-19 chest radiography: A framework using horizontal and vertical multiverse optimization, Comput. Biol. Med. 146 (2022) 105618.

[51] K. Hu, L. Zhao, S. Feng, S. Zhang, Q. Zhou, X. Gao, Y. Guo, Colorectal polyp region extraction using saliency detection network with neutrosophic enhancement, Comput. Biol. Med. 147 (2022) 105760.