Neural Computing and Applications (2023) 35:8035–8050
https://doi.org/10.1007/s00521-022-06969-0

S.I. : TAM-LHR

MDVA-GAN: multi-domain visual attribution generative adversarial networks

Muhammad Nawaz[1] · Feras Al-Obeidat[2] · Abdallah Tubaishat[2] · Tehseen Zia[3] · Fahad Maqbool[4] · Alvaro Rocha[5]

Received: 11 June 2021 / Accepted: 14 January 2022 / Published online: 31 January 2022

Abstract

Some pixels of an input image have thick information and give insights about a particular category during classification decisions. Visualization of these pixels is a well-studied problem in computer vision, called visual attribution (VA), which helps radiologists to recognize abnormalities and identify a particular disease in the medical image. In recent years, several classification-based techniques for domain-specific attribute visualization have been proposed, but these techniques can only highlight a small subset of most discriminative features. Therefore, their generated VA maps are inadequate to visualize all effects in an input image. Due to recent advancements in generative models, generative model-based VA techniques are introduced which generate efficient VA maps and visualize all affected regions. To deal the issue, generative adversarial network-based VA techniques are recently proposed, where the researchers leverage the advances in domain adaption techniques to learn a map for abnormal-to-normal medical image translation. As these approaches rely on a two-domain translation model, it would require training as many models as number of diseases in a medical dataset, which is a tedious and compute-intensive task. In this work, we introduce a unified multi-domain VA model that generates a VA map of more than one disease at a time. The proposed unified model gets images from a particular domain and its domain label as input to generate VA map and visualize all the affected regions by that particular disease. Experiments on the CheXpert dataset, which is a publicly available multi-disease chest radiograph dataset, and the TBX11K dataset show that the proposed model generates identical results.

Keywords Visual attribution · Generative adversarial network · Tuberculosis · Chest X-ray · Change map · Abnormal-to-normal translation

✉ Tehseen Zia
tehseen.zia@comsats.edu.pk

Muhammad Nawaz
nawazkhan.cui2018@gmail.com

Alvaro Rocha
amr@iseg.ulisboa.pt

[1] Medical Imaging and Diagnostic Lab, National Center of Artificial Intelligence, COMSATS University Islamabad, Islamabad, Pakistan

[2] College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates

[3] Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan

[4] Department of Computer Science, University of Sargodha, Sargodha, Pakistan

[5] University of Lisbon, ISEG, Rua do Quelhas, No 6, 1200-781 Lisboa, Portugal

1 Introduction

Visual attribution deals with efficient detection and visualization of specific class information in an image that is important for its classification. This task is very important with a vast real-world scope such as weakly supervised segmentation [1], visualization of disease effects for better understanding, and physiological processes in medical images [2, 3] [4]. Several methodologies have been proposed to efficiently tackle VA problem, but the most recent and frequently used approaches apply neural network classifier with two strategies: 1) analyze the gradient of predicted output for its input image [5, 6] and 2) examine the activation of feature maps in order to find out which part of image plays a vital role for making this prediction [1]. Neural network classifiers classify an image based on some salient regions which is not an efficient approach and sometimes produce undesired results because the classifier does not consider the whole object of interest during the decision. It means neural network approaches use certain information for decision, not all, and it is because this approach reduces the mutual information between input and output layers during its training [7]. These approaches perform ill if there are multiple pieces of evidence of a particular class (disease) and all these evidence are at different locations of an image. This is because their classifier just considers strong features and ignores those which have low discriminative power. But it is desirable and necessary in VA problem to visualize all evidence of particular categories, which are located at different locations, at a fine-grained level. To deal the issue, generative adversarial network-based VA techniques are recently proposed, where the researchers leverage the advances in domain adaption techniques to learn a map for abnormal-to-normal medical image translation. As these approaches rely on a two-domain translation model, it would require training as many models as number of diseases in a medical dataset, which is a tedious and compute-intensive task. Most of the existing techniques [8, 9] use different models of generative adversarial network (GAN) [10–12] to efficiently solve the VA problem, but all these techniques are inefficient for tasks which require attribution of multiple classes.

Existing VA models apply domain translation concepts for attribute visualization between just two domains. In simple words, these models are trained to visualize domain-specific features and translate just one abnormal domain to a normal domain. But if there are more than two domains (diseases), two generators and discriminator models will be required to visualize two domains' effects. If the number of domains increases, then model complexity and computation increase with the increasing number of generator and discriminator models. Suppose there are 'n' domains, then there will be a need to train one VA model for 'n' time or combine the number of 'n' VA models for each domain as shown in Fig. 1.

Motivation for this work comes while working on VA of ChesXpert which is a multiple-disease medical image dataset with no availability of pixel-level disease labels. Existing GAN-based VA approaches, since they rely on image-to-image translation, require training as many models as the number of diseases. Therefore, they are tedious and compute-intensive. CycleGAN [11] was the first unpaired image-to-image translator model for two domains translation. In the literature, StarGAN [13], StarGAN v2 [14], AttGAN [15], Fashion-AttGAN [16], and STGAN [17] are non-VA, but multi-domain generative models that are used for human face attributes (e.g., hair color, age, gender, and face expression) editing and cloth attributes (e.g., clothing colors, shapes, logo, texture, and sleeves lengths) editing.

We proposed a GAN-based multi-domain visual attribution (MDVA-GAN) network to tackle the aforementioned problems. The conceptual model of the proposed unified network is shown in Fig. 2. Our main contributions are as follows:

- Perceive and demonstrate the limited VA scalability problem in existing visual attribution methodologies.

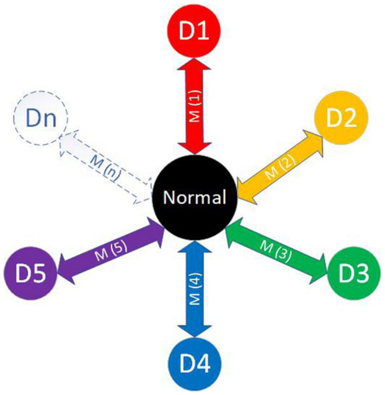

Fig. 1 Conceptual model of single-domain VA

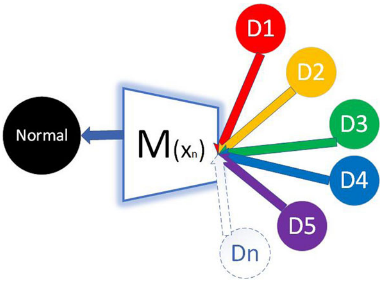

Fig. 2 Conceptual model of multi-domain VA

- We propose a novel unified generative adversarial network that learns multi-domain mapping and visualizes class-specific regions for multi-domain using single architecture.
- We exemplify how we can efficiently generate pixel-level labels from class-level labels.
- We evaluate the proposed model and provide qualitative as well as quantitative results on the multi-domain medical imaging dataset, single-domain medical imaging dataset, and multi-domain handwritten digit dataset that are CheXpert, TBX11K, and MNIST, respectively.

2 Literature review

2.1 Generative adversarial networks (GANs)

GANs have achieved remarkable success in several applications such as image generation [18, 19], image translation [11, 20, 21], super-resolution imaging [22], and images of face synthesis [23, 24]. Its architecture consists of two modules such as a generator model that is trained for generating new images and a discriminator that is trained on real and fake data. Both the models are trained together where the generator tries to generate such a plausible example so that discriminator should get fooled and classify them as real images.

In medical imaging, GANs have been used for medical image synthesizing [25–28], cardiac segmentation [29], liver segmentation [30], disease detection [31], retinal blood vessel segmentation [32], and disease effect visualization [8, 9]. GAN-based existing methodologies for disease effect visualization are capable of visualizing the disease effect of just one specific disease. However, our proposed methodology is a unified visual approach that works with multi-domain (multi-disease) at a time, translates anomalous image of a particular disease to normal image, and visualizes the anomaly of that particular disease in input abnormal image. To our knowledge, this is the first and novel application of GAN in medical images.

2.2 Visual attribution (VA)

Term VA is the detection and visualization of particular evidence in the image that elaborates the image class. There are several techniques in the literature, but commonly used approaches for weakly supervised segmentation or localization use feature maps of neural network classifiers [33]. Class activation mapping (CAM) [1] produce class-specific activation maps by reducing the network feature maps and using the global average pooling layer in the network. CAM technique is prominent in medical images with its various applications for the detection of pulmonary nodules form CT scans [2], skin diseases recognition [4], diabetic retinopathy lesions localization [34], and tuberculosis visualization [35]. Using an activation map for VA does not generate a smooth map because the output of each value is determined independently. Creating saliency maps is another class of VA techniques that backpropagate the gradients back to the input image. This class includes different methodologies such as integrated gradients [6], excitation backprop [36], and meaningful perturbation [3]. Similar techniques in the medical imaging domain have been utilized for fatal anatomy localization [5] and the fetal heart localization [37].

CAM-based VA technique has been used in medical imaging, but it relies on classification and does not visualize the full detail of the object of interest. This approach deals with the last feature map of network and therefore requires post-processing of the network prediction. To tackle all these issues, a GAN-based VA technique, called VA-GAN [9], was proposed which generates VA map without relying on classification. The VA-GAN model uses image translation concept where the mapping function maps the image of a specific class to any image of baseline class that leads to undesired results. Therefore, generated normal image contains random noise in it and does not considered a pair of input abnormal image. Another VA technique, ANT-GAN [8], is proposed as an abnormal-to-normal translation methodology which addresses aforementioned problems and generates normal image that preserve the contents of input abnormal image.

For image translation and domain-specific attribute visualization problem, generally, techniques use one generator and one discriminator model for single-domain translation; similarly, VA-GAN and ANT-GAN follow the same concept and use two generators and discriminators models for a pair of domains translation (e.g., abnormal-to-normal and normal-to-abnormal). Existing GAN-based approaches limit the VA scalability in handling multiple domains attributes visualization. If there is more than one disease, existing models fail to handle multi-domains or requires to train model and apply one by one for each pair of diseases. Unlike CycleGAN, VA-GAN, and ANT-GAN models, our proposed MDVA-GAN learns mapping among multiple domains and visualize anomalies of the particular disease.

3 MDVA generative adversarial network

To curb training of multiple GAN-based VA models for multi-disease VA, we aim to develop onetime training model for multi-disease VA. Specifically, we want to learn a disease-specific VA map, that translate a medical image with a disease d_i into an image without d_j. Primarily, this has been accomplished for a single-disease datasets. Hence, this work can be seen as a generalized form of previous works. The objective of this work is to develop a generative visual attribution approach to attribute instances of multiple classes with onetime trained model in contrast to existing approaches where we train m models for m classes. We assume a problem setting where there are M classes of interest and a baseline class. Also, an instance of each class of interest (COI) differs from corresponding instance of baseline class in class-specific features. Further, a number of instances are available for each COI, but there is no accessibility of baseline instance corresponding to an instance of each COI. Example of such a problem settings is readily available is medical domains, e.g., ChesXpert. Within this setting, when given an instance of COI as input, we seek to produce a visual attribution map that contains all of the features that distinguish instance of COI from its counterpart baseline instance. In other words, we aim to produce a map that, when subtracted from the COI instance, generates an image indistinguishable from its counterpart baseline instance. Mathematically,

$$x_i^b = x_i^c - M(x_i^c) \tag{1}$$

where x_i^c is an instance of c^{th} class, x_i^b is counterpart baseline instance of x_i^c, and $M(x_i^c)$ is visual attribution map. Also, x_i^b, x_i^c and $M(x_i^c)$ have same dimensions. Ideally, we need a dataset of (x_i^c, x_i^b) pairs to learn $M(x_i^c)$. However, it is impractical to such pairs particular for medical imaging domains. Hence, previous studies as well as this work, leverage domain adaptation using generative adversarial networks, to achieve this objective. In particular, we follow [11], and use cyclic consistency adversarial function to translate x_i^c into counterpart x_i^b. However, whereas previous works learn a two-domain translation function, we enable the model to perform conditional translation given the COI. Hence, we enable our model to translate multiple classes into a baseline class, by generating their visual attribution maps. Block diagram of the proposed model is shown in Fig. 3. The model consists of two GAN to accomplish the cycle of the cyclic consistency function: upper row of the diagram show forward-cycle GAN (shortly, fGAN) and bottom backward-cycle GAN (shortly, bGAN). We aim that the generator G_{C2M} of fGAN takes an instance x_i^c and label of COI as input and returns the required map $M(x_i^c)$.

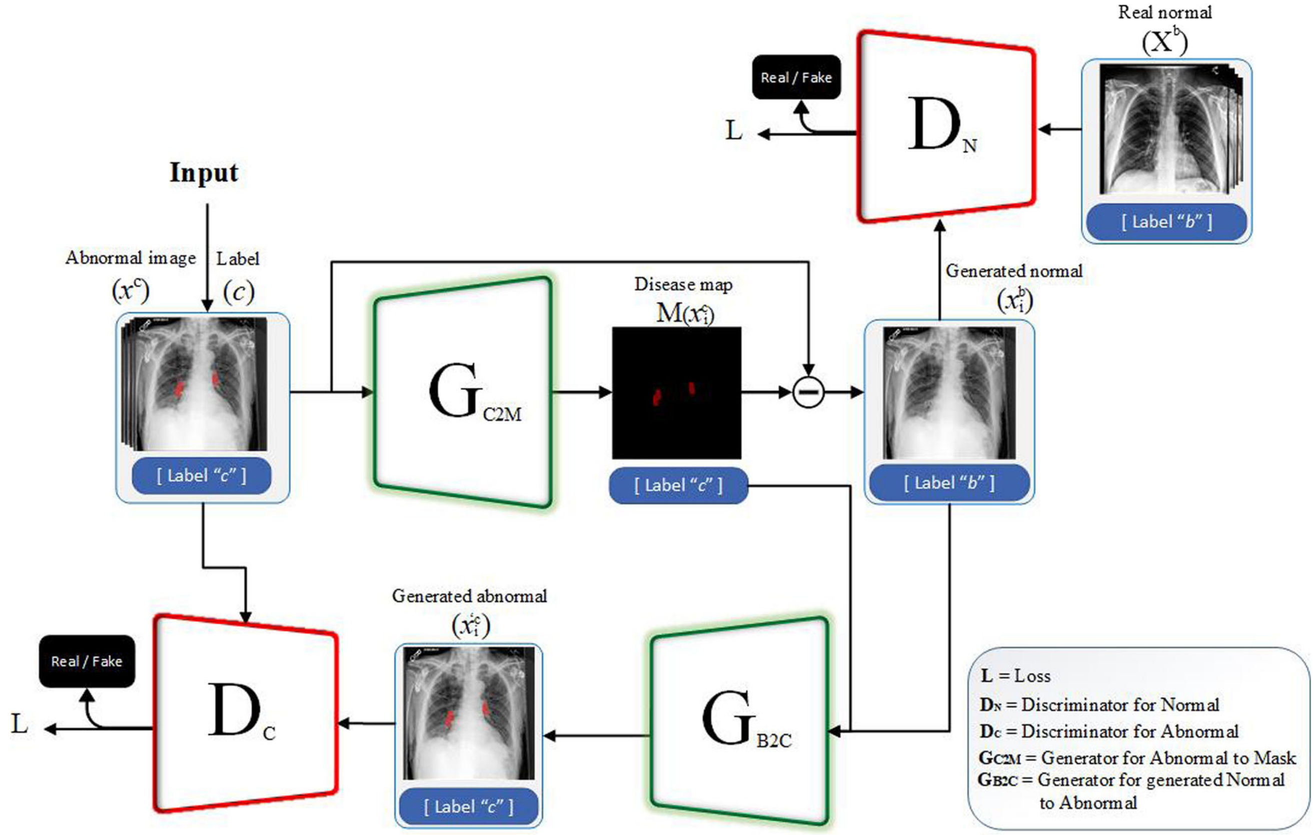

Fig. 3 Proposed MDVA-GAN

Table 1 Generator model architecture

Part	Layer Details
Downsampling	Convolution layer with number of output channel = 64, kernel size = 7 × 7, stride size = 1, padding size = 3, instance normalization, and activation function = ReLU.
	Convolution layer with number of output channel = 128, kernel size = 4 × 4, stride size = 2, padding size = 1, instance normalization, and activation function = ReLU.
	Convolution layer with number of output channel = 256, kernel size = 4 × 4, stride size = 2, padding size = 1, instance normalization, and activation function = ReLU.
Bottleneck residual block	Convolution layer with number of output channel = 256, kernel size = 3 × 3, stride size = 1, padding size = 1, instance normalization, and activation function = ReLU.
	Convolution layer with number of output channel = 256, kernel size = 3 × 3, stride size = 1, padding size = 1, instance normalization, and activation function = ReLU.
	Convolution layer with number of output channel = 256, kernel size = 3 × 3, stride size = 1, padding size = 1, instance normalization, and activation function = ReLU.
	Convolution layer with number of output channel = 256, kernel size = 3 × 3, stride size = 1, padding size = 1, instance normalization, and activation function = ReLU.
	Convolution layer with number of output channel = 256, kernel size = 3 × 3, stride size = 1, padding size = 1, instance normalization, and activation function = ReLU.
	Convolution layer with number of output channel = 256, kernel size = 3 × 3, stride size = 1, padding size = 1, instance normalization, and activation function = ReLU.
Upsampling	Deconvolution layer with number of output channel = 128, kernel size = 4 × 4, stride size = 2, padding size = 1, instance normalization, and activation function = ReLU.
	Deconvolution layer with number of output channel = 64, kernel size = 4 × 4, stride size = 2, padding size = 1, instance normalization, and activation function = ReLU.
	Convolution layer with number of output channel = 1, kernel size = 7 × 7, stride size = 1, padding size = 3, instance normalization, and activation function = ReLU.

Table 2 Discriminator model architecture

Part	Layer Details
Input Layer	Convolution layer with number of output channel = 64, kernel size = 4 × 4, stride size = 2, padding size = 1, and activation function = Leaky ReLU.
Hidden Layer	Convolution layer with number of output channel = 128, kernel size = 4 × 4, stride size = 2, padding size = 1, and activation function = Leaky ReLU.
	Convolution layer with number of output channel = 256, kernel size = 4 × 4, stride size = 2, padding size = 1, and activation function = Leaky ReLU.
	Convolution layer with number of output channel = 512, kernel size = 4 × 4, stride size = 2, padding size = 1, and activation function = Leaky ReLU.
	Convolution layer with number of output channel = 1024, kernel size = 4 × 4, stride size = 2, padding size = 1, and activation function = Leaky ReLU.
	Convolution layer with number of output channel = 2048, kernel size = 4 × 4, stride size = 2, padding size = 1, and activation function = Leaky ReLU.
Output Layer	Convolution layer with number of output channel = 1, kernel size = 3 × 3, stride size = 1, padding size = 1, and activation function = Leaky ReLU.

To train G_{C2M}, we use a discriminator D_B which aim to classify $x_i^c - M(x_i^c)$ instance as baseline instance. We use the following adversarial loss function to train the fGAN:

$$L_{fGAN} = \mathbb{E}_{x^b \sim p_{data(x^b)}}\left[\log(D_N(x^b))\right] + \mathbb{E}_{x^c \sim p_{data(x^c)}} \left[\log(1 - D_N(x_i^c - M(x_i^c)))\right] \quad (2)$$

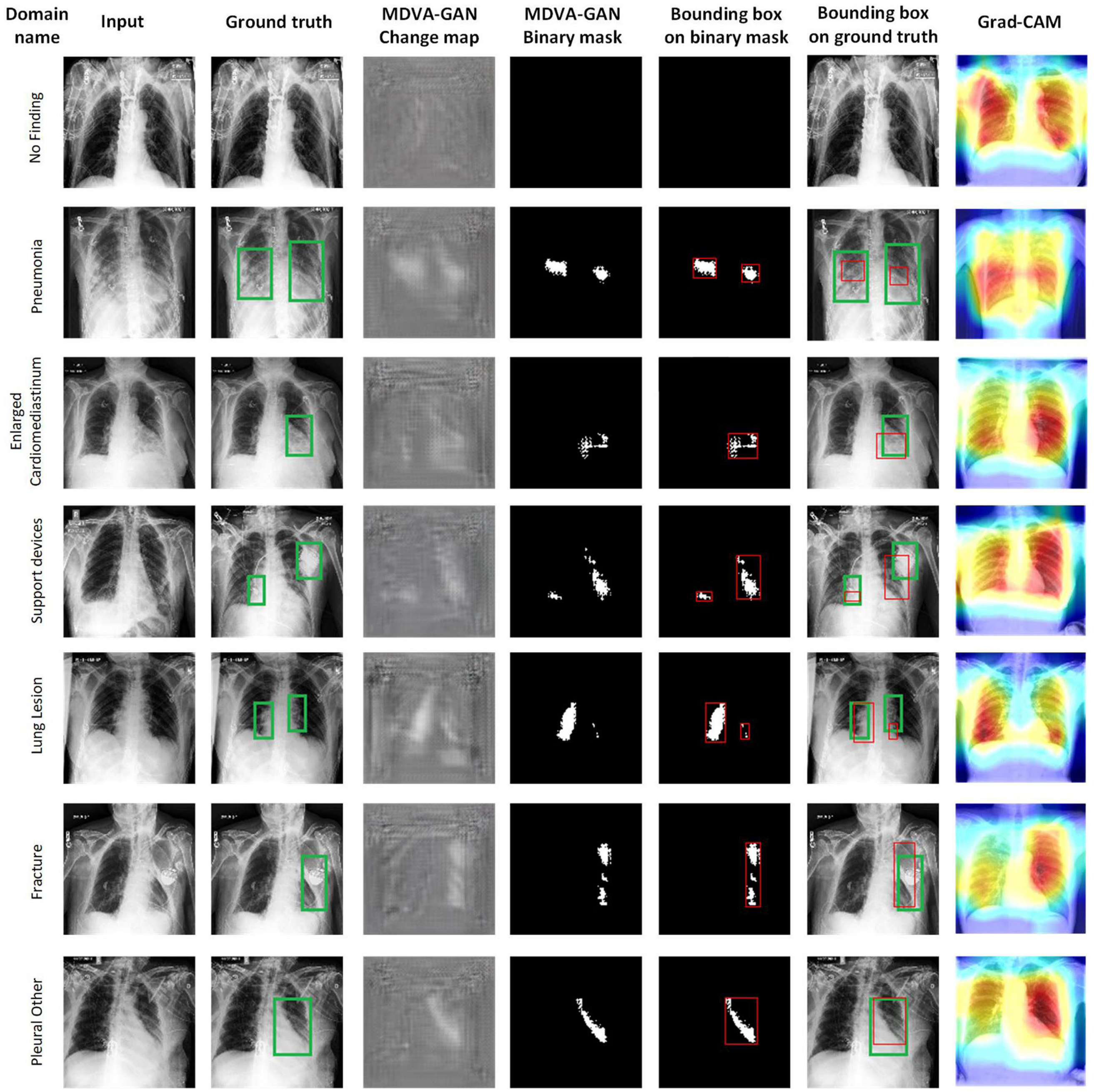

Fig. 4 Visual attribution results on 7 classes of the CheXpert dataset. The first column is domain name, the second column contains input images from each domain, the third column shows radiologist bounding boxes on input image, the fourth column contains change map images generated by MDVA-GAN, generated binary mask from change map are shown in the fifth column, and in the six and the seven columns, we draw bounding boxes on binary masks and ground truth, respectively. The last column shows the results of Grad-CAM methodology

As the discriminator D_B only ensure that the translated instance $x_i^c - M(x_i^c)$ is a baseline instance, to enable the model to translate an instance x_i^c into a counterpart x_i^b rather than any baseline instance, we use bGAN. The main objective of bGAN is to translate back the generated baseline instance $x_i^c - M(x_i^c)$ into the original x_i^c. The generator of bGAN G_{B2C} takes $x_i^c - M(x_i^c)$ instance and baseline label as input and generates x_i^c as output. To train G_{B2C}, we use the discriminator D_C which aim to classify the generates instance as instance of the COI. The following adversarial loss function is used to train the bGAB:

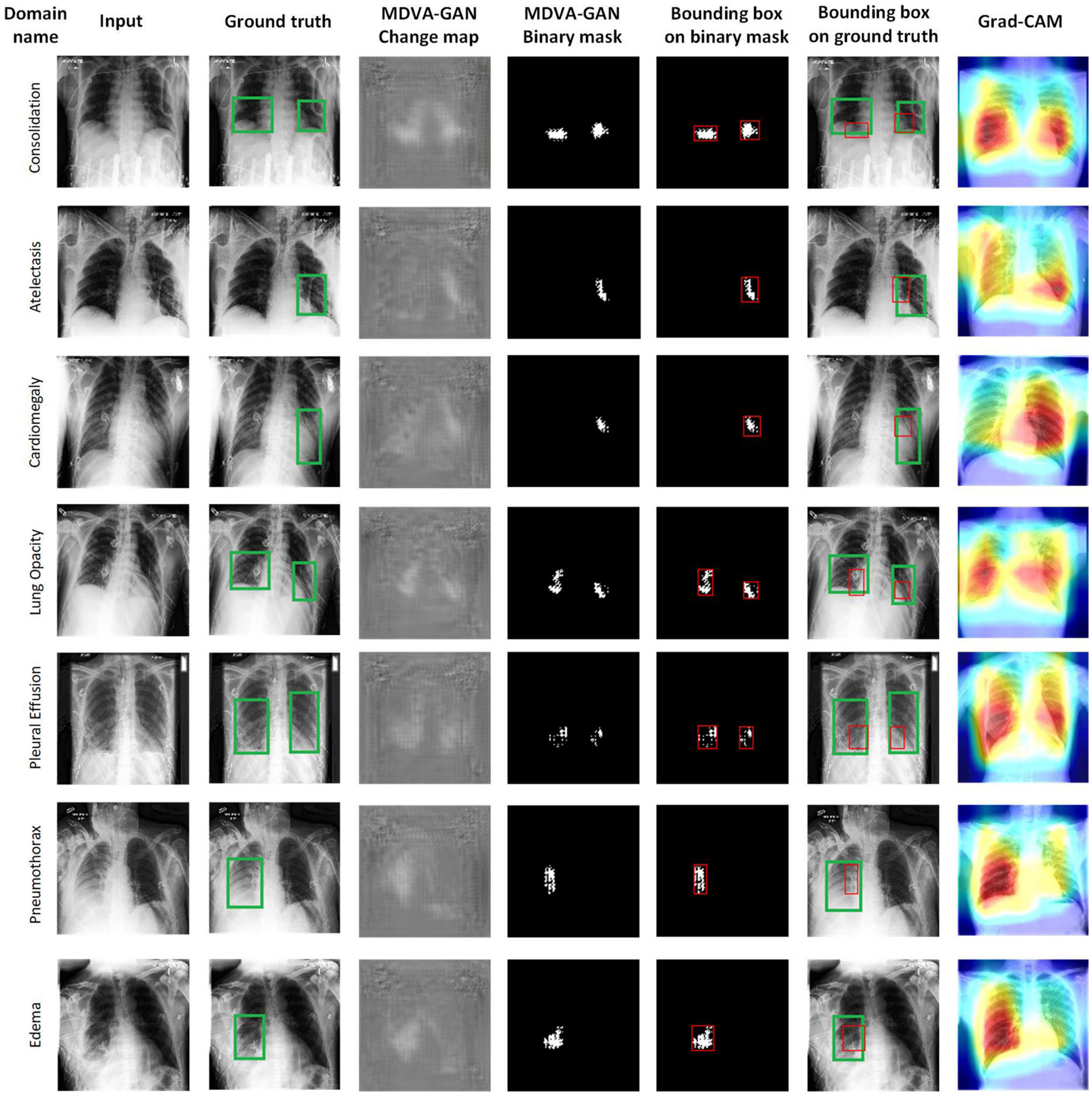

Fig. 5 Visual attribution results on next 7 classes of the CheXpert dataset. Similar to Fig. 4, The first column is domain name, the second column contains input images from each domain, the third column shows radiologist bounding boxes on input image, the fourth column contains change map images generated by MDVA-GAN, generated binary mask from change map is shown in the fifth column, and in the six and the seven columns, we draw bounding boxes on binary masks and ground truth, respectively. The last column shows the results of Grad-CAM methodology

$$L_{bGAN} = \mathbb{E}_{x^c \sim p_{data(x^c)}}[\log(D_C(x^c))] + \mathbb{E}_{x^c \sim p_{data(x^c)}} \\ [\log(1 - D_C(G_{B2C}(x_i^c - G_{C2M}(x_i^c, c)), b)] \tag{3}$$

We adopted cyclic consistency loss function to map x_i^c onto its counterpart x_i^b, as follows:

$$L_{cyc}(G_{C2M}, G_{B2C}) = \mathbb{E}_{x^c_{data(x^c)}}[||G_{B2C}(x_i^c - G_{C2M}(x_i^c, c)) \\ -x^c||_1] + \mathbb{E}_{x^b_{data(x^b)}}[||G_{C2M}(G_{B2C}(x_i^b, b)) - x^b||_1] \tag{4}$$

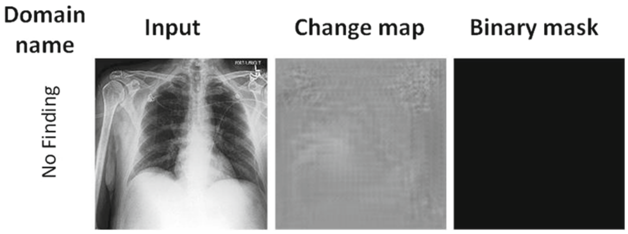

Fig. 6 Visual attribution: Healthy image with disease label

Equation (5) is the finally objective function which optimize the G and D models.

$$L(G,D) = L_{fGAN} + L_{bGAN} + L_{cyc} \quad (5)$$

4 Implementation

4.1 Baseline model

We acquire CycleGAN [11] architecture as our baseline model which is unpaired image-to-image translation architecture for two different domains at a time. This architecture introduced adversarial loss and cycle consistency loss to learn the mapping between two different domains and regularize this mapping, respectively. This architecture requires four models, two generators and two discriminators, for each pair of domain translation and has been used as the baseline model for domain-specific attribute visualization by ANT-GAN [8] model. CycleGAN architecture accepts the image as input to learn mapping between just two different domains, but we modify this architecture and introduce a novel unified model which is capable of learning the mapping between more than two domains at a time.

4.2 Network architecture

The proposed architecture is adopted from the CycleGAN [11] framework; therefore, the MDVA-GAN network consists of pair of generator and discriminator models. The generator network is composed of convolution layers for upsampling and downsampling, residual block [38] as the bottleneck, and instance normalization [39]. It is build with one convolution layer with one stride size and two convolution layers with two stride size in the downsampling part, six residual blocks in the bottleneck part, and two transposed convolution layers with two stride size in upsampling part. Instance normalization is used just for generator networks, not for discriminator networks. Similarly, discriminator network is consist of one convolution layer with two stride side as input layer, five convolution layers with two stride size as hidden layers, and one convolution layer with one stride size as output layer. Detail of architectures is comprised in descriptive Tables 1 and 2.

5 Experiments

We perform experiments by models on medical imaging datasets and include results of both, the generative and discriminative, models. Then compare MDVA-GAN against recent attribute visualization methods on pixel-level disease effect visualization. Next, we compare the qualitative and quantitative results of algorithms. Lastly, we verify the results of included techniques from the radiologists. Selected datasets, models training, and the results are explained in this section.

5.1 Datasets

CheXpert. The **Ch**est e**X**pert (CheXpert) [40] is one of the large, publicly available multi-class chest X-rays (CXRs) datasets which contain 224,316 chest radiographs of 65,240 different disease patients. All the chest radiographs are annotated with class-level label for the presence of 14 classes (e.g., 13 different diseases and 1 normal class) as 1 for the presence of disease, 0 for uncertain, and -1 for the absence of disease. First thing in our experiment, we categorize uncertain labels in the absence of disease. Furthermore, dataset contains 390 × 320 size images of male and female gender with frontal and lateral view. For experiment, we included male CXRs with frontal view and resize them as 256 × 256.

TBX11K. The Tuberculosis X-ray (TBX11K) [41] is latest, larger, and better annotated than existing Tuberculosis (TB) datasets which contains 512 × 512 sized samples of 11,200 CXRs images. Dataset consists of 4 different classes, such as Healthy, Active TB, Sick Non-TB, Latent TB, with bounding box labels as ground truth. This is considered two classes dataset *(TB and normal)*; therefore, we included Healthy and Active TB classes in our experiment.

MNIST. We also do experiments on Modified National Institute of Standards and Technology (MNIST) [42] handwritten dataset for understanding model's decision, especially for medical inexpert. It is a multi-class dataset of 60,000 handwritten images of digit 0 to 9. In our experiments, we selected 3000 images of each class from 5

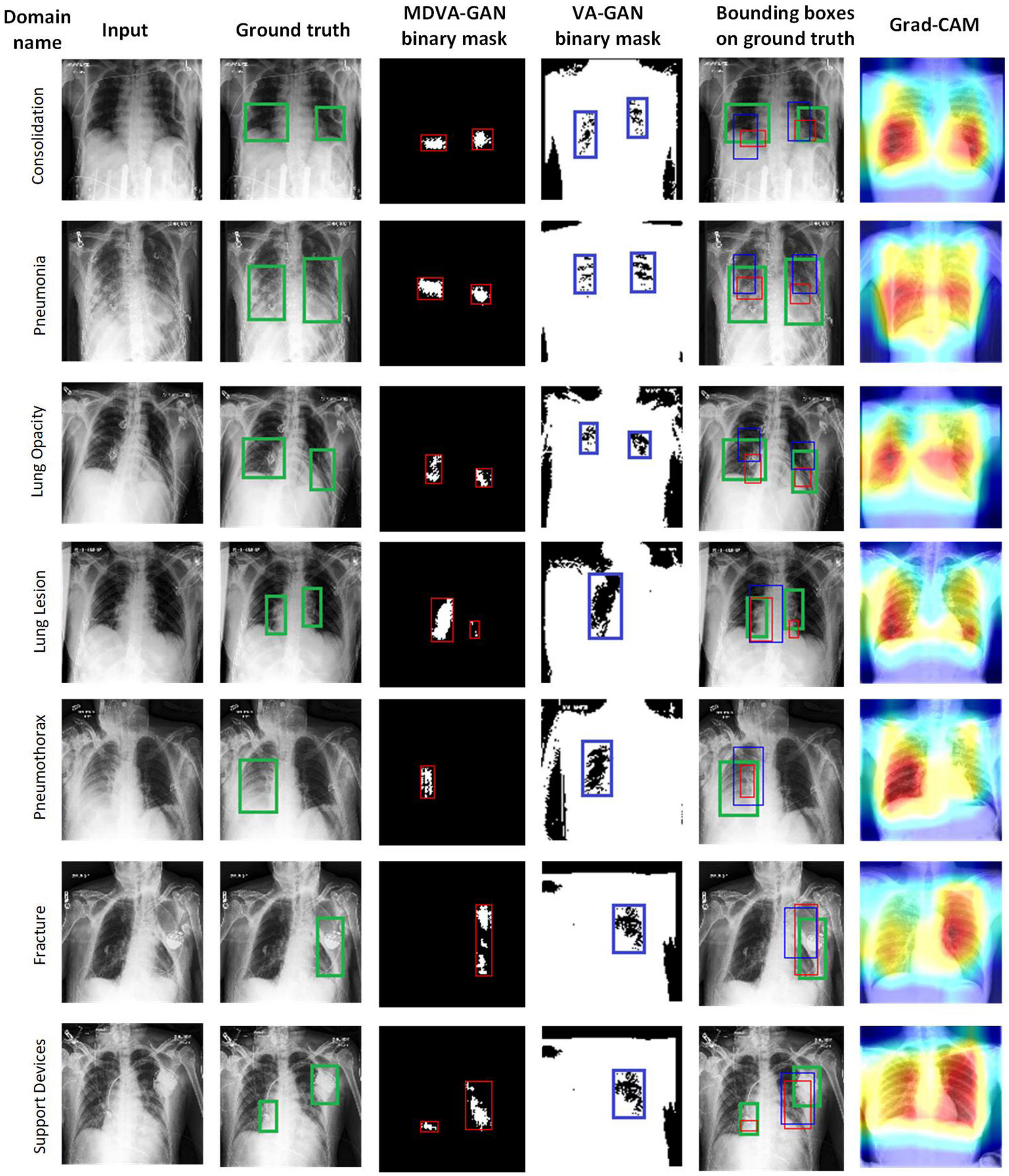

Fig. 7 Comparative visual attribution results on 7 classes of the CheXpert dataset. The first column is domain name, the second shows input images, the third shows radiologist bounding boxes on input images, the fourth shows binary masks by MDVA-GAN, the fifth shows the binary masks from VA-GAN, the six column shows comparative bounding boxes, and the last column shows the results of Grad-CAM methodology

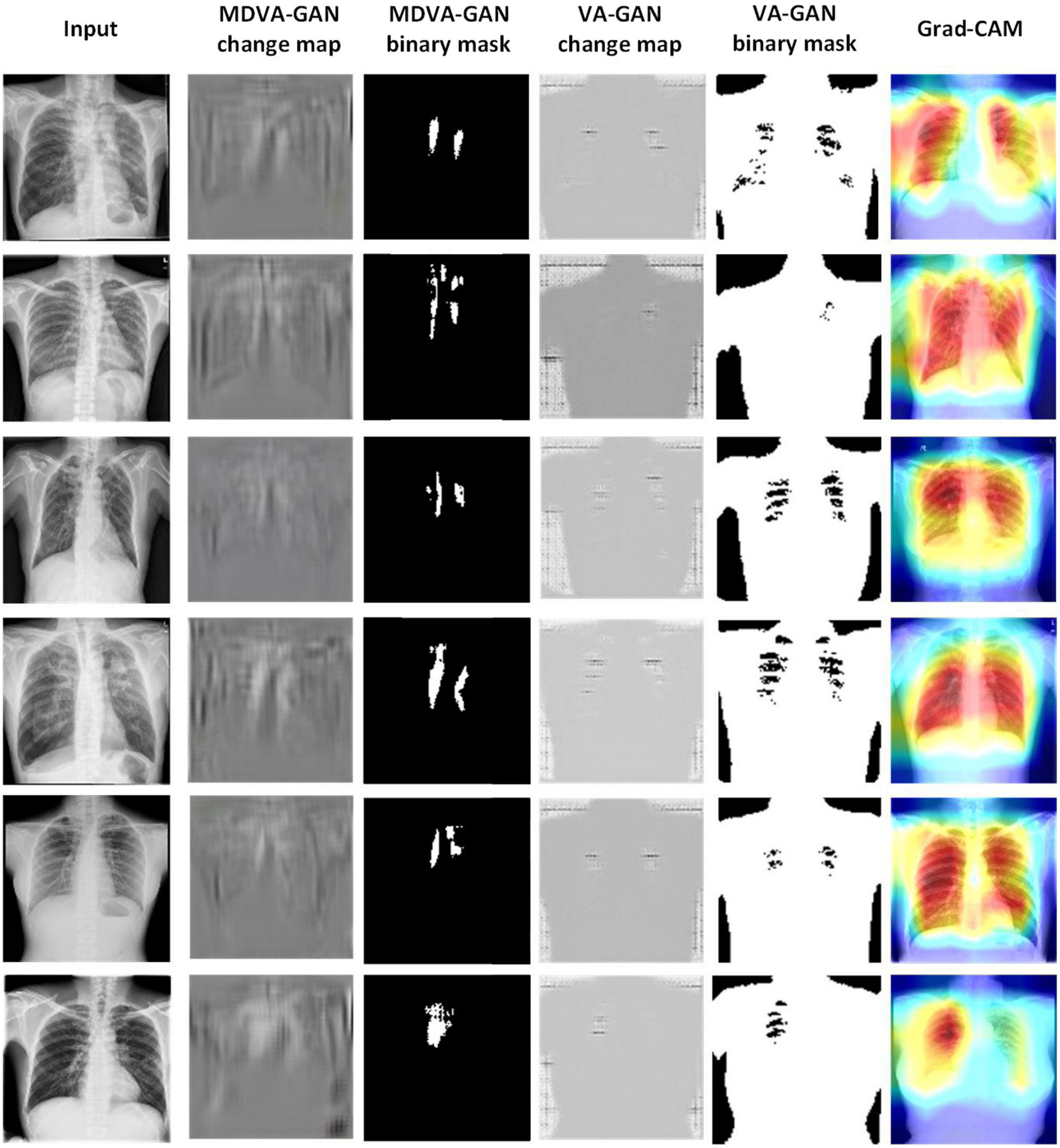

Fig. 8 Visual attribution results of the TBX11K dataset. The first column contains input images from dataset, and the next two columns show the results of proposed model. The fourth and the fifth columns are results from VA-GAN model, whereas the last column is attribute visualization by Grad-CAM method

classes such as one, four, six, seven, and nine. Then generate class-level labels, similar to CheXpert dataset, for each class. During label generation, we assign label 0 to digit one class which means it is a normal class, and assign label 1 to other four classes to consider diseased classes. Actually, this makes sense that there are some regions in 4 classes, other than digit 1 class, that made them abnormal.

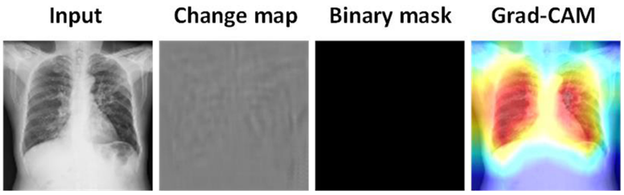

Fig. 9 MDVA-GAN and Grad-CAM on normal image

5.2 Training

The parameters of the generator and discriminator model are updated alternatively. All the networks are optimized with the ADAM optimizer. The learning rate and batch size are kept as 0.0002 and 1, respectively. The stopping criterion is chosen to be the patience of 10 epochs of validating the precision of generating a normal image. We train networks on a GPU-based desktop system with 128 GB RAM, Nvidia TitanX Pascal (12 GB VRAM), and 10 core Intel Xeon processor.

5.3 Experimental results on CheXpert

Figures 4 and 5 show the multi-domain visual attribution results of the proposed MDVA-GAN on the CheXpert dataset. A total of 14 classes are represented in the CheXpert dataset, each with class-level labels. There is a ground truth column next to each input column that has bounding boxes drawn by a medical expert on input images. We consulted with radiologists to verify our model's generated results and obtain diseased areas for quantitative model evaluation because the CheXpert is a multi-domain dataset, but without any ground truth. The fourth and fifth columns are the outputs of the proposed model, which generated a unique change map/disease map for each input domain, that we then converted to a binary mask in order to better comprehend the results, obtain pixel-level labels, and compare the results to ground truth. The change maps in column four are, in fact, pixel-level visualizations of disease impacts in the input image. We can see in the first row and the first column of Fig. 6 where the input image is 'Healthy' from 'No Finding' domain. Model's generated change map from the healthy image does not show any disease effect in it and its binary mask is also blank. However, for images that are not healthy, the model creates a distinct change map to visualize the particular disease regions and derives a pixel-level information from the class-level diseased label. Furthermore, when we input a healthy image with any disease label, our model did not confused by the diseased label and produces a blank change map and binary mask. As shown in Fig. 6, when we input a healthy image labeled with a disease, the model generates a blank disease map and blank binary mask to illustrate that the input image does not contain any disease. In contrast to previous generative-based VA techniques, MDVA-GAN generates a change map and visualizes disease effect using a change map subtraction method from the input image. We draw bounding boxes on the model's binary mask, as shown in column six, and similarly on the ground truth images, as shown in 'bounding box on ground truth' column, to show the MDVA-GAN model's results. The final column depicts Grad-CAM methodology results over each domain.

Furthermore, the comparison of discriminative and generative models is shown in Fig. 7 where the first two columns belong to the dataset, the third column shows the bounding boxes drawn by radiologists which are ground truth. The fourth and fifth columns contain bounding boxes drawn on MDVA-GAN and VA-GAN model's results, respectively. In the sixth column, we draw both the models bounding boxes on ground truth image in order to show that where exactly models results fall in the ground truth. It can clearly be seen that the MDVA-GAN model efficiently generated pixel-level disease effects. Grad-CAM results are also included in the last column which shows a big picture of desired regions.

5.4 Experimental results on TBX11K

TBX11K is a single-disease dataset, but it contains bounding boxes of tuberculosis disease images. We included this dataset for two main reasons: (1) CXRs images and (2) containing ground truth bounding boxes of diseased images. It is two domains (normal and abnormal), but its given bounding boxes are helpful to evaluate the performance of our model. The results of proposed model on TBX11K are shown in Fig. 8. The first two columns are part of the dataset: The first column contains images from the tuberculosis disease class, while the second column contains a ground truth representation of the input images.

We did not use the dataset's bounding boxes to train the MDVA-GAN. Input to the MDVA-GAN model is images and domain labels. We give label 0 to healthy and label 1 to TB. The rest of model is comparable to CheXpert training. The only difference is the number of classes, as this dataset comprises two classes. Like the CheXpert dataset, we fed the TB image to MDVA-GAN, which generated the disease map as shown in the third column of the figure. We then generated a binary mask of the change map and extracted bounding box values from it. Binary mask and bounding boxes represent TB disease location in image. A binary mask is drawn on the ground truth image to show the model's projected pixel-level disease location. The last column is Grad-CAM [43]-based representation of regions that participate most in decision making by model. As can be demonstrated, discriminative-based visualization techniques like Grad-CAM do not depict fine-grained regions. It is the most noticing point explained in [44] that saliency maps just tell about the regions where network looks much while taking a particular decision. Similarly in our scenario, Grad-CAM representation is just the regions where network focuses much during change map generating. This representation does not mean that shown area is a disease. Furthermore, Fig. 9 clears this point where we input a normal image to the MDVA-GAN model. The model generates a blank change map and binary mask which means there is no disease in the input image, but Grad-CAM highlighted the lungs regions which show that

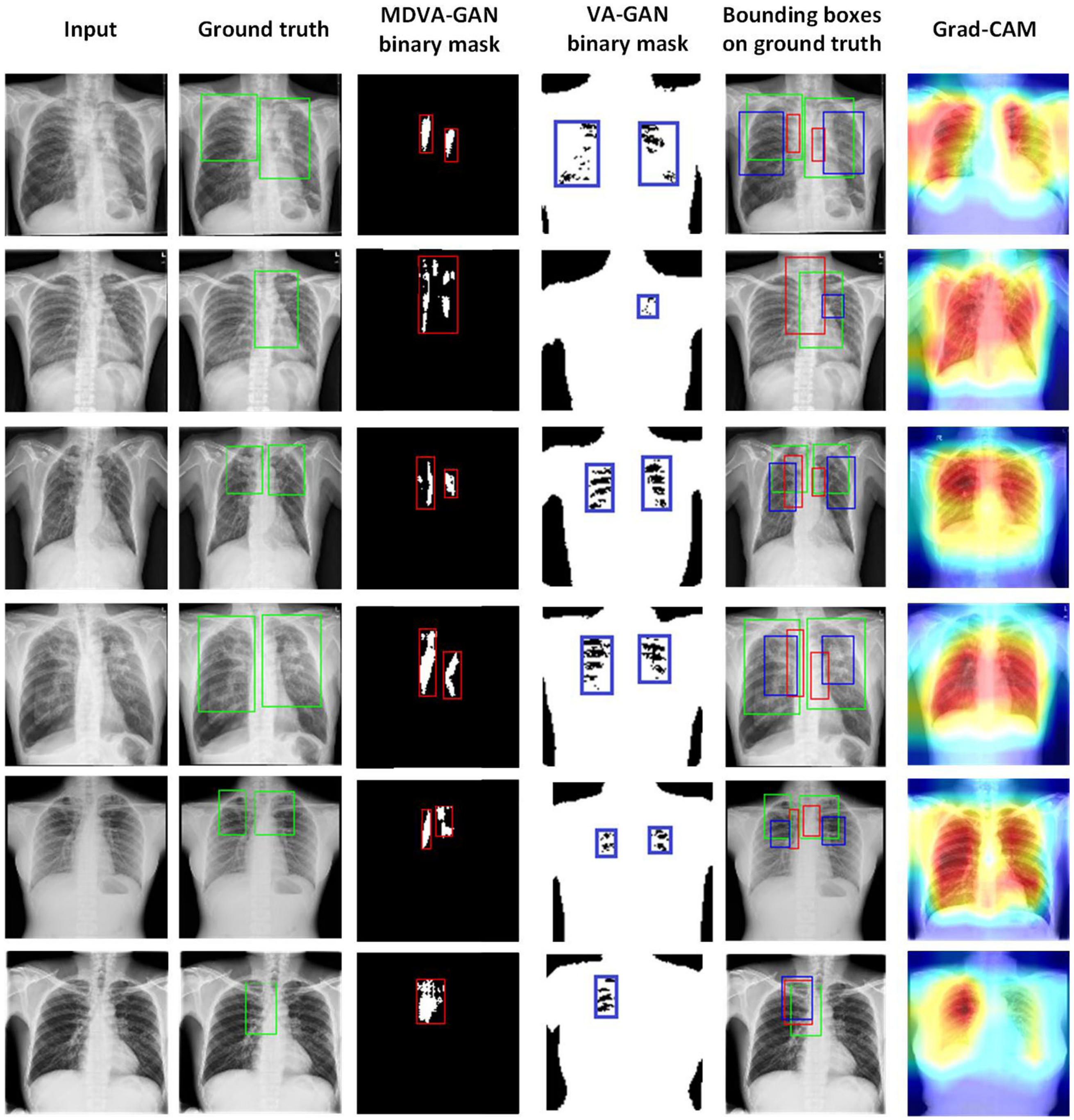

Fig. 10 Comparative visual attribution results of the TBX11K dataset. The first column is input images from dataset, the second column shows the disease location in input image, and the next two columns show the bounding boxes on output binary masks from MDVA-GAN and VA-GAN, respectively. The fourth column shows bounding boxes of both the models on ground truth, and the last column contains visualization results from Grad-CAM method

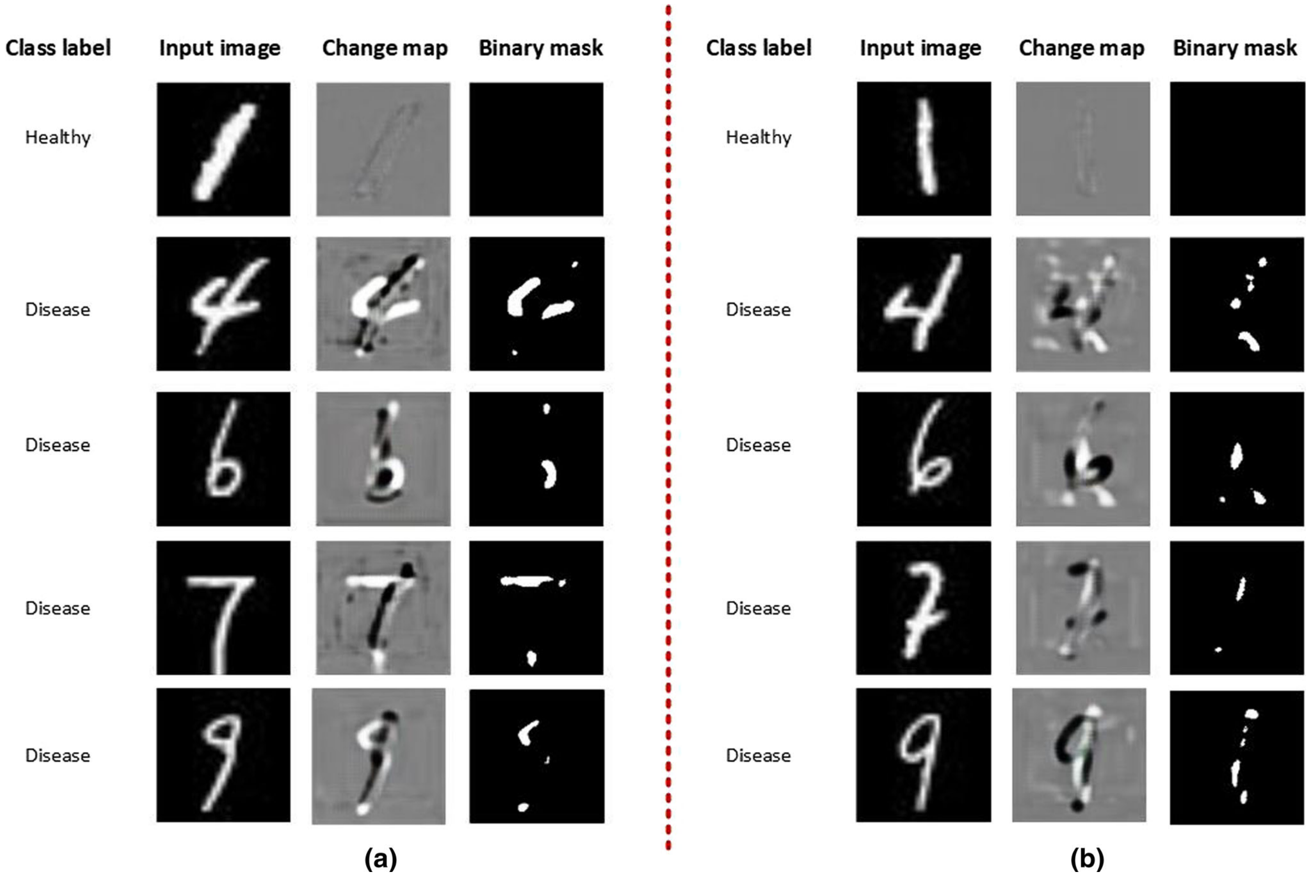

Fig. 11 Visual attribution: **a** Subtract output of generator from the input before applying loss functions, **b** Add output of generator to the input before applying loss functions

Table 3 Quantitative results of models on medical datasets

	CheXpert		TBX11K	
Method	IoU	IP	fIoU	IP
Grad-CAM	41.62	52.15	32.33	47.30
VA-GAN	37.52	72.23	29.15	68.23
MDVA-GAN	32.79	82.53	25.40	73.45

these lungs regions participate most in model's decision. It is not mean that there is a disease in Grad-CAM highlighted areas.

Also, the comparison of models is shown in Fig. 10 where the first two columns belong to dataset, the third column shows bounding boxes drawn on binary mask of proposed model, and the fourth column contains bounding boxes drawn on the VA-GAN model output. In the fifth column, we draw both the models bounding boxes on ground truth in order to show that where exactly models results fall in the ground truth. It can clearly seen that proposed model efficiently generated pixel-level disease effects. Grad-CAM results are also included in the last column which shows a big picture of desired regions unlike MDVA-GAN and VA-GAN models.

5.5 Experimental results on MNIST

To evaluate the performance of models and gain a better knowledge of how systems work, the MNIST dataset was the ideal choice. Indeed, the selected multi-class medical imaging dataset (e.g., CheXpert) comprises only class-level labels and no ground truth or bounding box at the pixel level. As a consequence, it is exceedingly difficult for medical non-experts to comprehend how the proposed model works on such a medical imaging dataset. Therefore, we examined the MNIST dataset and five classes: one, four, six, seven, and nine. First of all, we created class-level labels for each of the five-handwritten-digit classes. During label generation, we assign the label 0 to the digit ‘one’ class, indicating that it is normal and free of any disease. Similarly, we assign label 1 to the remaining four classes, indicating that there is something about the image that distinguishes it from being ‘one.’ For instance, if the head line of digit seven is removed, it transforms into digit 'one'; similarly, if a part of each class is removed, all the digits (except one) turn into class 'one' This makes sense and corresponds to a multi-class medical imaging dataset in which normal and diseased images are nearly identical, but the diseased image has some additional regions. When these excess regions in the diseased image are removed, the image is turned into a normal image. It is critical to note that each anomalous image has a unique label, not simply 1. It is a one-hot vector with unique combination of a single 1 bit and remaining 0 bits.

Results of proposed model on MNIST dataset are shown in Fig. 11 where map generator function M(x) generates a change map which shows the regions of input image that are extra in it. When we subtract this generated change map of Fig. 11a from input image, then input image (diseased image) is transformed into output domain (normal image). Similarly, when we add this generated change map of Fig. 11b from input image then input image (diseased image) is transformed into output domain (normal image).

Existing VA methodologies, [9] and [8], have limited VA scalability and are only capable of learning two domains mapping and both the methodologies add generated change map into the input. Figure 11b shows that the addition causes random noise in change map and cannot efficiently visualize the desired regions.

5.6 Quantitative results

Table 3 shows the quantitative results for both the medical datasets. One thing to keep in mind is that while both datasets contain bounding boxes as ground truth, the results are generated using model output at the pixel level. In fact, the proposed model is novel in that it accepts images with class-level labels as input and produces pixel-level information about those images. This is the contribution of our study in the context of weakly supervised learning. Because the area of pixel-level outcomes is always small in comparison with bounding boxes, getting a high intersection over union (IoU) score while comparing pixel-level results with bounding boxes ground truth is extremely difficult. The locations where disease can be found in any location are only indicated by ground truth boundary boxes. This does not guarantee that the disease has affected the entire bounding box. However, because both generative models (e.g., MDVA-GAN and VA-GAN) generate pixel-level disease data in this scenario, the IoU score for both models is lower. We further introduce an intersection percentage (IP) evaluation matrix and calculate the IP score of the models. The IP score is the percentage of pixel-level data that belong to a given ground truth. There will be a 100 percent overlap if the pixel-level output falls completely within the bounding box. Similarly, the model's maximum IP score mean pixel-level output belongs entirely to ground truth, while the lowest IoU score mean output region is extremely small in comparison with ground truth. This IP matrix is appropriate for use in weekly supervised experiments where pixel-level information needs to be compared to bounding box ground truths.

The quantitative findings show that the Grad-CAM approach has the highest IoU score since it does not provide pixel-level information, whereas the MDVA-GAN model generates pixel-level information and outperforms other models on both datasets, as demonstrated by the quantitative results.

6 Conclusion

This work proposed the MDVA-GAN model which is a unified image-to-image translation architecture that visualizes multi-domain attributes using two generators and discriminators models. In contrast to existing visual attribution approaches, the proposed model gets image as well as domain label as input to generate a unique change map for each domain. Alike others, our model generates a noise-free change map as it subtracts the change map from input. Also, the proposed technique generates pixel-level labels from class-level labels. The proposed model is built on the architecture of cycle consistent GAN. We showed that how change map of diseased image visualizes the diseased parts and the generator model generates a blank change map of healthy images to visualize that there is not a disease. Finally, we perform experiments on CheXpert, TBX11K, and MNIST datasets and compare the results with existing visual attribution methodologies.

7 Future work

This research aims to generate justifications for a decision in a multi-domain challenge. Our model only accepts one label per input image, yet some images contain many domains. Using multi-label to detect all potential domains in the input image is an exciting future avenue for our work [45].

Declarations

Conflict of Interest Muhammad Nawaz, Feras Al-Obeidat, Abdallah Tubaishat, Tehseen Zia, Fahad Maqbool, and Alvaro Rocha declare that they have no conflict of interest.

References

1. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp 2921–2929)
2. Feng X, Yang J, Laine AF, Angelini ED (2017, September) Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules. In: International conference on medical image computing and computer-assisted intervention (pp 568–576). Springer, Cham
3. Fong RC, Vedaldi A (2017) Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE International Conference on Computer Vision (pp 3429–3437)
4. Ge Z, Demyanov S, Chakravorty R, Bowling A, Garnavi R (2017, September) 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 (pp 250–258). Springer, Cham. (year)
5. Baumgartner CF, Kamnitsas K, Matthew J, Fletcher TP, Smith S, Koch LM, Rueckert D (2017) SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans Med Imaging 36(11):2204–2215
6. Sundararajan M, Taly A, Yan Q (2017, July) Axiomatic attribution for deep networks. In International Conference on Machine Learning (pp. 3319-3328). PMLR
7. Shwartz-Ziv R, Tishby N (2017) Opening the black box of deep neural networks via information. arXiv:1703.00810
8. Sun L, Wang J, Huang Y, Ding X, Greenspan H, Paisley J (2020) An adversarial learning approach to medical image synthesis for lesion detection. IEEE J Biomed Health Informatics 24(8):2303–2314
9. Baumgartner CF, Koch LM, Tezcan KC, Ang JX, Konukoglu E (2018) Visual feature attribution using wasserstein gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp 8309–8319)
10. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Bengio Y (2014) Generative adversarial networks. arXiv:1406.2661
11. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision (pp 2223–2232)
12. Arjovsky M, Chintala S, Bottou L (2017, July) Wasserstein generative adversarial networks. In: International conference on machine learning (pp 214–223). PMLR
13. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp 8789–8797)
14. Choi Y, Uh Y, Yoo J, Ha JW (2020) Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp 8188–8197)
15. He Z, Zuo W, Kan M, Shan S, Chen X (2019) Attgan: facial attribute editing by only changing what you want. IEEE Trans Image Process 28(11):5464–5478
16. Ping Q, Wu B, Ding W, Yuan J (2019) Fashion-AttGAN: Attribute-aware fashion editing with multi-objective GAN. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp 0-0)
17. Liu M, Ding Y, Xia M, Liu X, Ding E, Zuo W, Wen S (2019) STGAN: A unified selective transfer network for arbitrary image attribute editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp 3673–3682)
18. Huang X, Li Y, Poursaeed O, Hopcroft J, Belongie S (2017) Stacked generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp 5077–5086)
19. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196
20. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp 1125–1134)
21. Kim T, Cha M, Kim H, Lee JK, Kim J (2017, July) Learning to discover cross-domain relations with generative adversarial networks. In: International conference on machine learning (pp 1857–1865). PMLR
22. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Shi W (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp 4681–4690)
23. Kim T, Kim B, Cha M, Kim J (2017) Unsupervised visual attribute transfer with reconfigurable generative adversarial networks. arXiv:1707.09798
24. Shen W, Liu R (2017) Learning residual images for face attribute manipulation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp 4030–4038)
25. Nie D, Trullo R, Lian J, Petitjean C, Ruan S, Wang Q, Shen D (2017, September) Medical image synthesis with context-aware generative adversarial networks. In: International conference on medical image computing and computer-assisted intervention (pp 417–425). Springer, Cham
26. Zhang Z, Yang L, Zheng Y (2018) Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern Recognition (pp 9242–9251)
27. Chartsias A, Joyce T, Giuffrida MV, Tsaftaris SA (2017) Multimodal MR synthesis via modality-invariant latent representation. IEEE Trans Med Imaging 37(3):803–814
28. Cao B, Zhang H, Wang N, Gao X, Shen D (2020, April) Auto-GAN: self-supervised collaborative learning for medical image synthesis. In: Proceedings of the AAAI conference on artificial intelligence (Vol 34, No 07, pp 10486-10493)
29. Zhang Y, Yang L, Chen J, Fredericksen M, Hughes DP, Chen DZ (2017, September) Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: International conference on medical image computing and computer-assisted intervention (pp 408–416). Springer, Cham
30. Yang D, Xu D, Zhou SK, Georgescu B, Chen M, Grbic S, Comaniciu D (2017, September) Automatic liver segmentation using an adversarial image-to-image network. In: International conference on medical image computing and computer-assisted intervention (pp 507–515). Springer, Cham
31. Yu K, Wang Y, Cai Y, Xiao C, Zhao E, Glass L, Sun J (2019) Rare disease detection by sequence modeling with generative adversarial networks. arXiv:1907.01022
32. Park KB, Choi SH, Lee JY (2020) M-gan: Retinal blood vessel segmentation by balancing losses through stacked deep fully convolutional networks. IEEE Access 8:146308–146322
33. Oquab M, Bottou L, Laptev I, Sivic J (2015) Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp 685–694)
34. Gondal WM, Köhler JM, Grzeszick R, Fink GA, Hirsch M (2017, September) Weakly-supervised localization of diabetic

retinopathy lesions in retinal fundus images. In: 2017 IEEE international conference on image processing (ICIP) (pp 2069–2073). IEEE

35. Kim HE, Hwang S (2016) Deconvolutional feature stacking for weakly-supervised semantic segmentation. arXiv:1602.04984
36. Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, Sclaroff S (2018) Top-down neural attention by excitation backprop. Int J Comput Vis 126(10):1084–1102
37. Gao Y, Noble JA (2017, September) Detection and characterization of the fetal heartbeat in free-hand ultrasound sweeps with weakly-supervised two-streams convolutional networks. In: International conference on medical image computing and computer-assisted intervention (pp 305–313). Springer, Cham
38. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp 770–778)
39. Ulyanov D, Vedaldi A, Lempitsky V (2016) Instance normalization: The missing ingredient for fast stylization. arXiv:1607.08022. Jul 27
40. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Ng AY (2019, July) Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence (Vol 33, No 01, pp 590–597)
41. Liu Y, Wu YH, Ban Y, Wang H, Cheng MM (2020) Rethinking computer-aided tuberculosis diagnosis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp 2646–2655)
42. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
43. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (pp 618–626)
44. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Mach Intell 1(5):206–215
45. Zintgraf LM, Cohen TS, Adel T, Welling M (2017) Visualizing deep neural network decisions: Prediction difference analysis. arXiv:1702.04595

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.