A2B-GAN: UTILIZING UNANNOTATED ANOMALOUS IMAGES FOR ANOMALY DETECTION IN MEDICAL IMAGE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated anomaly detection in medical images can significantly reduce human effort in disease diagnosis. Owing to the complexity in modeling anomalies and the high cost of manual annotation by domain experts, a typical technique in the current literature is to employ only data from healthy subjects to derive the model for normal images and then to detect anomalies as outliers to this model. In many real applications, mixed datasets with both normal and potential abnormal images (e.g., images of patients with confirmed diseases) are abundant. This paper poses the research question of how to improve anomaly detection by using an unannotated set of mixed images of both normal and anomalous samples (in addition to a set of normal images from healthy subjects). We propose a novel one-directional image-to-image translation method named A2B-GAN, which learns to translate any images to only normal images (hence "one-directional"). This alleviates the requirement of direct cycle consistency of existing unpaired image-to-image translation methods, which is unattainable with unannotated data. Once the translation is learned, we generate a difference map for any given image by subtracting its translated output. Regions of significant responses in the difference map correspond to potential anomalies (if any). In terms of average AUC, our A2B-GAN outperforms the state-of-the-art methods by 0.1 points (approximately 16.25%) on two medical imaging datasets: COVID-19 detection and Cardiomegaly detection by utilizing an unannotated set mixed with anomalies. Our code is available for public release upon the paper decision.

1 INTRODUCTION

Supervised learning from a large annotated dataset is becoming easier (He et al., 2015; Esteva et al., 2017), thanks to deep neural networks. For problems like anomaly detection (e.g., rare disease detection in medical images), however, it may often be very difficult to obtain a large enough set of annotated anomalous samples, making it impractical to rely on supervised learning for the task. Therefore, many recent anomaly detection methods typically learn only from the normal images of healthy patients (Chen & Konukoglu, 2018; Schlegl et al., 2017; 2019; Alex et al., 2017; Zenati et al., 2018a;b; Akcay et al., 2018; Gherbi et al., 2019; Roth et al., 2021; Defard et al., 2021). In practice, *unannotated* anomalous samples are usually available and what is missing is the elaborated annotation. For example, we may easily obtain a dataset that contains many anomalous samples because of the underlying patients have the confirmed pathology (although typically it is unknown which images are anomalous and where the anomalies are in an image). In other words, besides the set of normal images, we may assume the availability of a mixed dataset with both normal and potential abnormal images. In this research, we seek the answer to the question: *How can we utilize such an unannotated mixed dataset, in addition to the set of normal images, to improve the performance of anomaly detection?*

Answer to such question has been explored in the distant past for lesion detection in vascular CT images using SVM (Zuluaga et al., 2011). Differently, aiming to achieve a more generalized solution, we have developed a novel one-directional unpaired image-to-image translation network, termed A2B-GAN, based on Generative Adversarial Network (Goodfellow et al., 2014; 2020) (GAN). Existing unpaired image-to-image translation methods (Liu et al., 2017; Shen & Liu, 2017; Yi et al.,

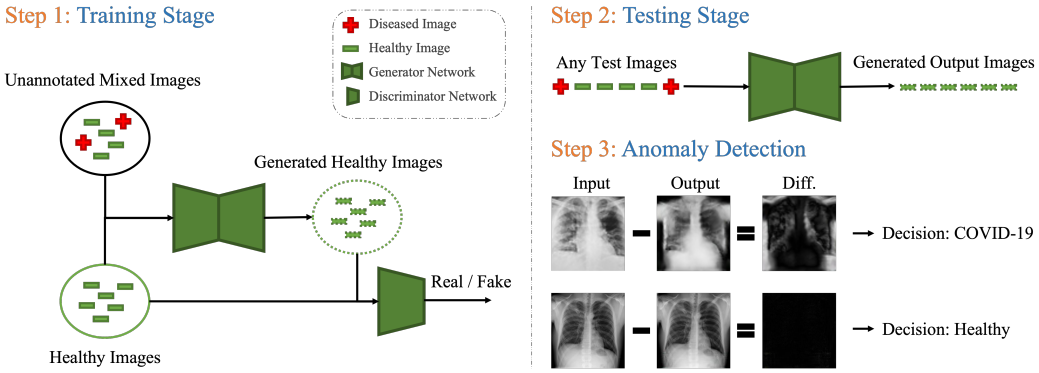

Figure 1: Overview of the proposed anomaly detection method. At training stage, our proposed A2B-GAN learns to generate healthy images utilizing an unannotated dataset mixed with both healthy and potential diseased/anomalous images, in addition to, a set of healthy images. At testing stage, we translate any given image to a corresponding healthy image using the trained A2B-GAN. Then we subtract the output image from the input image, which reveals the presence of an anomaly.

2017; Zhu et al., 2017a;b; Choi et al., 2018; Mejjati et al., 2018; Zhang et al., 2018; He et al., 2019; Liu et al., 2019; Zhao et al., 2020) usually require performing both anomalous-to-normal and normal-to-anomalous translation to ensure cycle-consistency (Zhu et al., 2017a); or the anomalous and normal images to be known as a prior during training. The challenge of the anomaly detection scenario in this study is that it is impossible to translate a normal image back to an anomalous one because annotated anomalous images are unavailable. To address this challenge, A2B-GAN employs two important properties for improving anomaly detection: (1) unpaired image-to-image translation; and (2) one-directional image-to-image translation. To achieve these two properties, we introduce a novel reconstruction loss that ensures effective cycle-consistency. Unlike traditional cycle-consistency loss (Zhu et al., 2017a), our reconstruction loss utilizes learned attention-masks to generate the reconstructed images for cycle-consistency. Since all the image manipulation for backward-cycle occurs using basic mathematical operations, there is no need for image annotation. An overview of the proposed approach is illustrated in Figure 1.

Through extensive experiments, we demonstrate that A2B-GAN on average outperforms existing state-of-the-art anomaly detection methods ALAD (Zenati et al., 2018b), f-AnoGAN (Schlegl et al., 2019), Ganomaly (Akcay et al., 2018), PatchCore (Roth et al., 2021), and PaDiM (Defard et al., 2021) by significant margin on two medical imaging datasets: COVID-19 and Cardiomegaly detection. This performance is attributed to A2B-GAN's capability of utilizing unannoted diseased/anomalous images at training time. In summary, we make the following contributions:

- We introduce a novel one-directional unpaired image-to-image translation method for anomaly detection.
- We propose a novel anomaly detection method to utilize unannotated anomalous images.
- We develop a novel reconstruction loss for ensuring cycle-consistency without requiring annotated inputs.
- With two challenging medical datasets, we perform extensive experiments comparing the proposed method against the state-of-the-art anomaly detection methods, and we report significant performance improvements and provide detailed analysis.
- We provide quantitative and qualitative results on a simulated anomaly detection method to ease the readers' analysis without medical expertise.

2 Related Work

Our work is closely related to GAN-based anomaly detection and image-to-image translation. Hence, we review and contrast relevant existing efforts on these tasks with the proposed A2B-GAN.

2.1 ANOMALY DETECTION

In general, the existing GAN-based anomaly detection methods (Chen & Konukoglu, 2018; Schlegl et al., 2017; 2019; Alex et al., 2017; Zenati et al., 2018a;b; Akcay et al., 2018; Gherbi et al., 2019) explore various strategies for learning from only *healthy/normal* images. These methods try to learn the healthy images' manifold so that their decoder can reconstruct healthy images only at test time. Hence, the *diseased/anomalous* images are reconstructed as healthy images. The difference between the input-output images reveals the presence of anomalies. We elaborate a few examples below.

Chen & Konukoglu (2018) use an adversarial autoencoder to learn healthy data distribution. The anomalies are identified by feeding a diseased image to the trained autoencoder, followed by subtracting the reconstructed diseased image from the input image.

The method proposed by Schlegl et al. (2017) adversarially learns a decoder model to generate healthy images from random noise vectors in the latent space. At test time, the method maps a new image to the latent space by iteratively updating the latent vector. If the new image is healthy, then the method is expected to find the exact latent vector that reconstructs the input image. As a result, the difference between the input and the reconstructed image is negligible. If the new image is diseased, then the method is expected to find a latent vector that produces a healthy image closest to the diseased image. This leads to a higher difference between the input and the reconstructed images, indicating an anomaly. The authors propose an anomaly score, which is a weighted average of the reconstruction error and the discrimination score from the discriminator network.

The above method has been made faster in Schlegl et al. (2019) via an encoder network for mapping the input images to the latent space in a single pass. Similarly, Alex et al. (2017) use a GAN to learn a generative model of healthy data. To identify anomalies, they scan an image pixel-by-pixel and feed the scanned crops to the trained GAN discriminator. An anomaly map is then constructed by putting together the anomaly scores given by the discriminator. Zenati et al. (2018a;b) utilize BiGAN (Donahue et al., 2016) to learn the mapping of normal images by training an encoder and a decoder network jointly. Like most methods, they also utilize the reconstruction error as the anomaly score. In the same spirit, Akcay et al. (2018) train an autoencoder using only normal images. The autoencoder is supervised using both image-level L_1 distance and adversarial loss. An additional encoder is also trained to map the images reconstructed by the autoencoder back to its latent space. A different approach proposed by Gherbi et al. (2019) trains an encoder network to map normal images to a Gaussian distribution and abnormal images to out-of-distribution using adversarial learning. Then the anomalies are detected using Mahalanobis distance in the latent space. Please note that this method requires annotated anomalous images at training time.

Though methodologically dissimilar, we have incorporated PatchCore (Roth et al., 2021) and PaDiM (Defard et al., 2021) into our list of baselines since they are top performing anomaly detection methods in natural imaging datasets like MVTec AD (Bergmann et al., 2019; 2021). Both of these methods utilize a memory bank of nominal features from an ImageNet (Deng et al., 2009) pre-trained model. PaDiM converts these features into a matrix of Gaussian parameters (mean and covariance). At the test time, PaDiM extracts feature vectors of each test images as done at training stage. Then anomaly detections are performed based-on the Mahalanobis distance of the feature vectors from the pre-computed Gaussian parameters' matrix. On the other hand, PatchCore downsamples its memory bank of neighborhood-aware patch-level features using greedy coreset subsampling. Finally, an image is predicted as anomalous if any one of the patches in the image is anomalous according to the scoring system.

In contrast to these approaches, our method learns from unannotated diseased/anomalous images mixed with healthy/normal images and performs better anomaly detection (results in section 4).

2.2 IMAGE-TO-IMAGE TRANSLATION

Plenty works have been done on GAN-based image-to-image translation (Isola et al., 2017; Kim et al., 2017; Ledig et al., 2017; Liu et al., 2017; Shen & Liu, 2017; Yi et al., 2017; Zhu et al., 2017a;b; Choi et al., 2018; Mejjati et al., 2018; Zhang et al., 2018; He et al., 2019; Liu et al., 2019; Nizan & Tal, 2020). While Pix2Pix (Isola et al., 2017) is among the first to do so, it requires input-output image pairs to train. CycleGAN (Zhu et al., 2017a) revolutionizes unpaired image-to-image translation by introducing cycle-consistency. It suggests that if an image of a horse is forward-

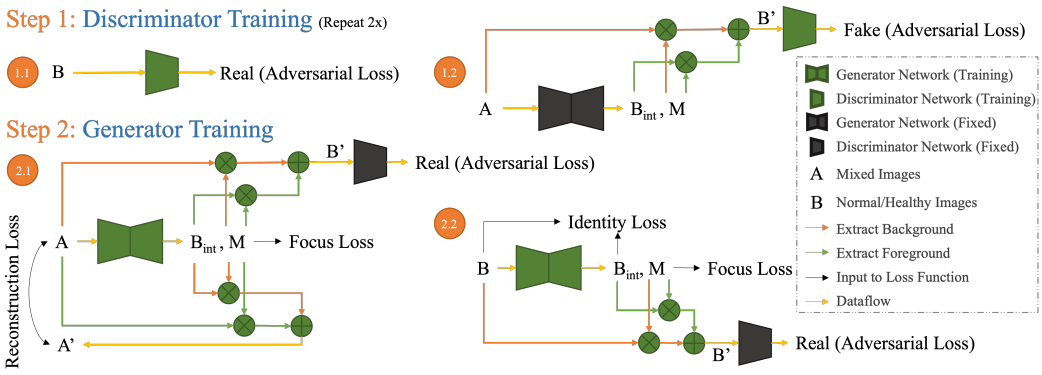

Figure 2: Overview of a single training iteration of the proposed A2B-GAN. Detailed training process is described in section 3 and hyper-parameters for all applications are provided in section 4.4.

translated to an image of a zebra, then backward-translating the image of that zebra should result in the original horse image. Utilizing this concept at training time, CycleGAN tries to keep the appearance of the horse and the zebra the same except the skin color. CycleGAN utilizes two generator networks to realize the concept. At the training, first, one generator is used to translate input images (horses) to the output images (zebras), and then another generator is used to translate the output images (zebras) back to input images (horses). Note that we are required to know the annotations of the images (horse/zebra) in either case. Though the cycle-consistency concept is important for unpaired image-to-image translation between diseased and healthy images, we cannot directly use CycleGAN in our work as we do not have annotated diseased images in anomaly detection.

Most recent unpaired image-to-image translation methods (Liu et al., 2017; Shen & Liu, 2017; Yi et al., 2017; Zhu et al., 2017a;b; Choi et al., 2018; Mejjati et al., 2018; Zhang et al., 2018; He et al., 2019; Liu et al., 2019; Zhao et al., 2020), irrespective of utilization of cycle-consistency, require image annotations. For example, Shen & Liu (2017) utilize two generators for translating images of human faces between a pair of facial attributes. Mejjati et al. (2018) propose an attention-based approach that performs image-to-image translation like CycleGAN with additional two networks for generating attention maps. Instead of using multiple networks for each domain translation pair, methods like StarGAN (Choi et al., 2018), AttGAN (He et al., 2019), STGAN (Liu et al., 2019), and Fixed-Point GAN (Rahman Siddiquee et al., 2019) utilize one generator network that takes the target images' annotation as input. Therefore, this line of works is also unsuitable for our purpose. A recent ensemble-based method (Nizan & Tal, 2020) proposes an alternative to cycle-consistency (Zhu et al., 2017a) for improved unpaired image-to-image translation. This method can perform image-to-image translation without knowing the annotations for diseased/anomalous images. However, this method requires at least two generator networks and four discriminator networks, and thus it is computationally expensive while being difficult to train.

In contrast, the proposed A2B-GAN method utilizes only one generator and one discriminator network. It satisfies the cycle-consistency requirement through a novel reconstruction loss that does not need annotation for the diseased/anomalous images. We will show that our A2B-GAN outperforms existing leading anomaly detection methods in section 4 while achieving image-to-image translation performance comparable to existing methods (more in Appendix A).

3 A2B-GAN: The Proposed Approach

3.1 Network Architecture

The proposed A2B-GAN consists of a discriminator network and a generator network. The discriminator network follows PatchGAN (Isola et al., 2017; Li & Wand, 2016; Zhu et al., 2017a) architecture and is similar to the ones used in Choi et al. (2018); Rahman Siddiquee et al. (2019). Our discriminator distinguish whether the input image is a real or a fake (generated) healthy image.

The generator network takes both diseased and healthy images without knowing their labels and translates them to only healthy images. We denote the mixed dataset containing both diseased and health images as A and the dataset containing only healthy images as B. We denote the generated images mimicking the distribution of set A and B as A' and B', respectively. The generator network does not generate the A' and B' images directly; rather, it generates intermediate healthy images B_{int} and masks M. The masks' values are in the range [0 – 1], where 0 denotes a background pixel, and 1 denotes a foreground pixel. Then we produce the final generated image B' following Equation 1.

$$B' = B_{int} \times M + A \times (1 - M) \tag{1}$$

Similarly, A' is generated following Equation 2.

$$A' = A \times M + B_{int} \times (1 - M) \tag{2}$$

Note that an image in the set A can be either diseased or healthy. If it is diseased, we expect the mask M to activate the diseased region as foreground; otherwise, we expect M to be empty. It is worth noting the similarity between Equation 2 and the cycle-consistency concept introduced in (Zhu et al., 2017a). Since the proposed method is controlling the image generation, partially, by the mask M, it neither requires a label nor an additional generator network for the image in A to generate back to A' (Figure 2). As the generator network translates the input images to a single direction, we call the proposed method a one-directional image-to-image translation method and thus named A2B-GAN. This particular property lets us utilize an unannotated mixed dataset during the training stage.

3.2 Training

Figure 2 depicts the detailed training methodology of A2B-GAN. We train the generator and the discriminator network, alternatively, like any GAN models. At each training step, we update the weights of the generator once for every two weight updates of the discriminator network. We repeat this for many iterations until convergence.

We train the discriminator to treat the real healthy images, B, as real and any images generated by the generator to be fake. Therefore, the adversarial loss for the discriminator is defined in Equation 3.

$$\mathcal{L}^D_{adv} = \mathbb{E}_{x \in A}[D_{real/fake}(G(x))] - \mathbb{E}_{x \in B}[D_{real/fake}(x)] \tag{3}$$

Here, $G(x)$ denotes the output of the generator and is obtained by Equation 1. $D_{real/fake}(x)$ denotes the output of the discriminator network. We have revised adversarial loss (Equation 3) based on the Wasserstein GAN (Arjovsky et al., 2017) objective by adding a gradient penalty (Gulrajani et al., 2017) with weight λ_{gp} to stabilize the training, which is defined as

$$\begin{aligned} \mathcal{L}^D_{adv} = \; & \mathbb{E}_{x \in A}[D_{real/fake}(G(x))] - \mathbb{E}_{x \in B}[D_{real/fake}(x)] \\ & + \lambda_{gp}\, \mathbb{E}_{\hat{x}}[(||\nabla_{\hat{x}} D_{real/fake}(\hat{x})||_2 - 1)^2] \end{aligned} \tag{4}$$

The objective of the generator is to take any image as input and generate a healthy image corresponding to the input image. To be specific, if the input image is a healthy image, the generator is expected to behave like an autoencoder and produce the same input image as output. If the input is a diseased image, the generator should remove anomalous parts and produce a healthy image in the output. The adversarial loss for the generator is defined in Equation 5.

$$\mathcal{L}^G_{adv} = - \sum_{x \in \{A,B\}} \mathbb{E}_x[D_{real/fake}(G(x))] \tag{5}$$

For the known healthy image set, B, the generator should behave like an autoencoder. Therefore, we apply an *identity loss* for these images. It is defined in Equation 6.

$$\mathcal{L}_{id} = \mathbb{E}_{x \in B}[||G_{int}(x) - x||_1] \tag{6}$$

Here, G_{int} denotes the generated image before applying the mask (B_{int} in Figure 2). Since we train A2B-GAN using unpaired images we add a reconstruction loss (Equation 7) to ensure that the generated images are close to the input images.

$$\mathcal{L}_{rec} = \mathbb{E}_{x \in A, y \in A'}[||x - y||_1] \tag{7}$$

We have adopted the focus loss from (Nizan & Tal, 2020) to control the size of the mask. The focus loss is defined by Equation 8.

$$\mathcal{L}_f = \lambda_{fs}(\sum_{i=1}^{n} M_i/n)^2 + \lambda_{fz}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|M_i - 0.5| + \epsilon} \tag{8}$$

Here, n denotes the number of pixels in the mask M and M_i denotes a pixel in it. The first component controls the size of the mask and the second component forces the values to be close to 0/1. λ_{fs} and λ_{fz} are relative weights of these components, respectively.

Combining all losses, the final full objective function for the discriminator and generator can be described by Equation 9 and Equation 10, respectively.

$$\mathcal{L}_D = \mathcal{L}^D_{adv} \tag{9}$$

$$\mathcal{L}_G = \mathcal{L}^G_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{id}\mathcal{L}_{id} + \lambda_f\mathcal{L}_f \tag{10}$$

where λ_{rec}, λ_{id}, and λ_f determine the relative importance of the *reconstruction loss*. *identity loss*, and *focus loss*, respectively.

3.3 Detecting Anomalies

Figure 1 provides an overview of the proposed anomaly detection method. Given an unannotated mixed dataset containing a mixture of both diseased and healthy images A and another dataset containing only healthy images B, we train the A2B-GAN as described in section 3.2. Once trained, we first translate each of the test images to healthy images, and then we subtract the generated healthy images from the input images. Ideally, we expect the resultant difference images to show the diseased regions if the input is a diseased image; otherwise, we expect the difference image to be empty. Therefore, we detect the presence of the disease by checking the activations of the difference images. Please note that the difference images indicate the presence of the disease/anomaly, as well as, serve as localization maps of the disease/anomaly.

4 Experiments and Results

Baselines. We have compared the proposed A2B-GAN with 5 state-of-the-art anomaly detection methods. We have selected these methods as they are the most recent. Among them, ALAD (Zenati et al., 2018b), f-AnoGAN (Schlegl et al., 2019), and Ganomaly (Akcay et al., 2018) are methodologically the closest to the proposed A2B-GAN. We have excluded other methodologically similar works such as EGBAD (Zenati et al., 2018a) and AnoGAN (Schlegl et al., 2017) from our baseline list since ALAD and f-AnoGAN are improved versions of these methods, respectively. However, we have included PatchCore (Roth et al., 2021) and PaDiM (Defard et al., 2021), though methodologically different than the proposed A2B-GAN, as they are top-performing methods for novelty detection in natural image dataset like MVTec AD (Bergmann et al., 2019; 2021). A methodological comparison among these methods has been discussed in section 2.1.

Evaluation. We have compared the proposed A2B-GAN for anomaly detection with the baseline methods using AUC score from the receiver operating characteristic (ROC) curve. To get the prediction score for A2B-GAN, we subtract the input image from their translated images first. Then we take the mean value of the resultant difference image. For the baseline methods, we use the anomaly score generation method proposed by their corresponding authors.

Methods	COVID-19	Cardiomegaly	Average
ALAD	0.5802	0.5286	0.5544
PatchCore	0.5200	*0.5999	0.5600
PaDiM	0.5400	*<u>0.6034</u>	0.5717
Ganomaly	0.5840	**0.6300**	0.6070
f-AnoGAN	<u>0.6382</u>	0.5987	<u>0.6185</u>
A2B-GAN (Ours)	**0.8364**	0.6015	**0.7190**

Table 1: Summary of the anomaly detection results. We have compared the proposed A2B-GAN with existing state-of-the-art anomaly detection methods, ALAD (Zenati et al., 2018b), Ganomaly (Akcay et al., 2018), f-AnoGAN (Schlegl et al., 2019), PatchCore (Roth et al., 2021), and PaDiM (Defard et al., 2021) using AUC metric on 2 medical imaging datasets. The best results are shown in bold and the second best results are underlined. PatchCore and PaDiM throw out-of-memory error for Cardiomegaly dataset on our 500GB machine. Therefore, we have run them on a smaller subset of the Cardiomegaly dataset. These results are shown with an asterisk. On average, the proposed A2B-GAN performs better than all the baseline methods.

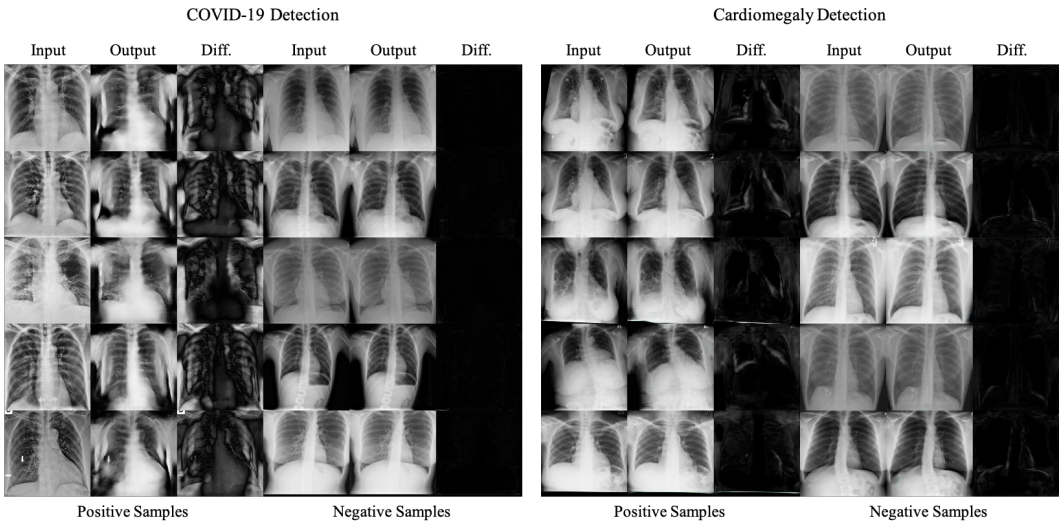

Figure 3: Qualitative results of COVID-19 and Cardiomegaly detection by A2B-GAN. As seen, A2B-GAN has resulted high response in the difference maps both for COVID-19 and Cardiomegaly positive samples. In contrast, the difference maps of the negative samples are almost empty.

4.1 COVID-19 Detection

Dataset. We have utilized the COVIDx dataset from Wang et al. (2020). The original dataset contains a train and a test set. The train set has 15,464 Chest X-rays, of which 1,670 are COVID-19 positive, and 13,794 are COVID-19 negative. In contrast, the test set has 200 Chest X-rays, of which 100 are positive, and the rest of the 100 are negative. For our training set, we have randomly taken 10,031 negative images for the known healthy image set. For the mixed unannotated training set, we have randomly taken 3,663 negative images from the rest and 1,570 COVID-19 positive images.

Results. The 2^{nd} column in Table 1 summarizes the COVID-19 detection results. As seen, A2B-GAN achieves COVID-19 detection AUC of 0.8364 outperforming all the baseline methods by a large margin. The best performing baseline method, f-AnoGAN (Schlegl et al., 2019), achieves an AUC score of only 0.6382 which is 0.1982 points lower than the proposed A2B-GAN. Other GAN-based approaches ALAD (Zenati et al., 2018b) and Ganomaly (Akcay et al., 2018) achieve similar AUC scores: 0.5802 and 0.5840, respectively. Interestingly, the top-performing methods in natural image dataset (MVTec AD) perform the worst in COVID-19 detection. PatchCore (Roth et al., 2021) achieves an AUC score of 0.5200 and PaDiM (Defard et al., 2021) achieves 0.5400 only. Note that PatchCore has the highest detection score in MVTec AD dataset at the time of

writing this manuscript yet achieves the lowest COVID-19 detection score. We believe it is due to the fact that PatchCore and PaDiM was designed to achieve best anomaly detection scores in natural image datasets like MVTec AD. Therefore, the necessity of more anomaly detection methods for medical imaging domain such as the proposed A2B-GAN is obvious. Qualitative results of COVID-19 detection by A2B-GAN have been provided in Figure 3.

4.2 Cardiomegaly Detection

Datasets. We have utilized the ChestX-ray8 dataset (Wang et al., 2017) for this experiment. We have used only Posterior Anterior (PA) images from the dataset. There are 39,302 PA healthy (negative samples) X-rays and 1,563 PA X-rays with Cardiomegaly (positive samples) in the dataset. We split these images into train and test set. The test set contains 301 positive and 7,589 negative samples. For the train set, we randomly select 20,657 negative samples for the known healthy image set. From the rest of the X-rays, we randomly select 2,111 negative images and 905 positive images for the mixed unannotated training set. Please note that PaDiM (Defard et al., 2021) and PatchCore (Roth et al., 2021) were initially unable to run on Cardiomegaly dataset due to the out-of-memory error on our 500GB machine. Therefore, we had to reduce the number of negative samples in our healthy image set from 20,657 to 10,000 only for these two methods.

Results. The 3^{rd} column of Table 1 summarizes the Cardiomegaly detection results. Ganomaly (Akcay et al., 2018) achieves the best Cardiomegaly detection AUC score of 0.6300. PaDiM (Defard et al., 2021) secures second place with an AUC score of 0.6034. The proposed A2B-GAN places third with a score of 0.6015 which is only 0.03 points lower than the best performing method, Ganomaly. However, A2B-GAN, compared to Ganomaly, achieved 0.2524 points higher in COVID-19 detection. The difference in AUC between A2B-GAN and PaDiM is insignificant. The other baseline methods ALAD (Zenati et al., 2018b), f-AnoGAN (Schlegl et al., 2019), and PatchCore (Roth et al., 2021) achieve Cardiomegaly detection AUC score of 0.5286, 0.5987, and 0.5999, respectively. Qualitative results of Cardiomegaly detection by A2B-GAN are available in Figure 3.

4.3 Simulated Anomaly Detection

Analysing the qualitative results in medical imaging can be challenging for the readers without the domain knowledge. Realizing the fact, we have incorporated experiments on this simple but instructive simulated anomaly detection dataset.

Datasets. We have utilized the CelebA dataset (Liu et al., 2015) for this experiment. We have made an image anomalous by randomly selecting a 40×40 square area and replacing the color of all the pixels in that area by their mean. The simulated dataset is split into training and test set. The training set contains randomly selected 5,000 normal images which we call the known healthy set. 3,500 images from the rest of the normal images and 1,500 anomalous images have been randomly selected for the mixed unannotated training set. The test set contains 2,500 normal and 2,500 anomalous images.

Results. Figure 4 provides qualitative results on the simulated anomaly detection by A2B-GAN. As seen, the proposed A2B-GAN tries to interpolate facial attributes in the anomalous region while keeping the normal images unchanged. From the Figure 4, it is also evident that difference maps can be employed to detect, as well as, localize the anomalous regions. Quantitatively, PaDiM (Defard et al., 2021) performs the best on this dataset achieving an AUC score of 0.9609. The proposed A2B-GAN places second with an AUC score of 0.8002 which is 0.1607 lower than PaDiM. Please note that the proposed A2B-GAN scored 0.2964 points higher than PaDiM in COVID-19 detection. It is also noteworthy that PaDiM performs well only on the simulated dataset which confirms its state-of-art performance in natural imaging domain. However, it performs poorly on the two medical imaging datasets. In contrast, the proposed A2B-GAN performs consistently well on all three datasets. The other baseline methods ALAD, Ganomaly, f-AnoGAN, and PatchCore achieved AUC score of 0.5417, 0.5880, 0.5969, and 0.7549, respectively

4.4 Implementation Details

We have resized the input images to 256×256 for the experiments on COVID-19 and Cardiomegaly detection datasets in section 4.1 and section 4.2, respectively. For the Simulated Anomaly dataset

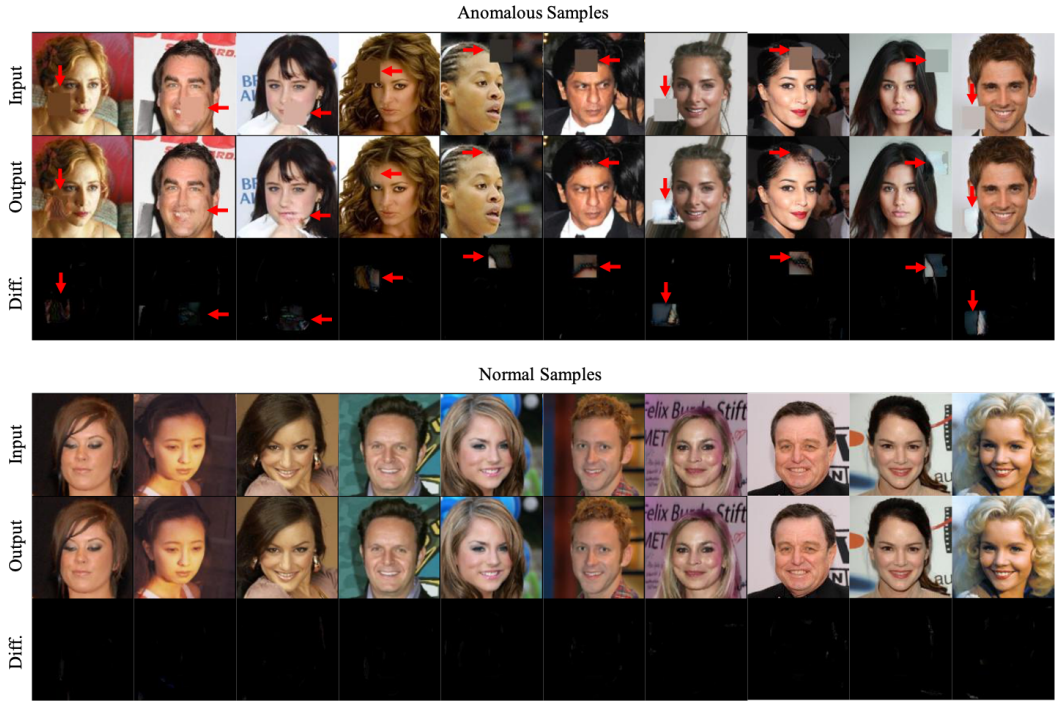

Figure 4: Qualitative results of the simulated anomaly detection by A2B-GAN. The red arrows point the locations of the anomalies. As seen, the difference maps for the anomalous images activate the anomalous locations. In contrast, the difference maps appear empty for the normal images.

in section 4.3, we have resized the images to 128×128. We have set $\lambda_{gp} = 10$, $\lambda_{id} = 1$, $\lambda_{rec} = 1$, $\lambda_f = 0.1$, and $\lambda_{fz} = 1$ for all the experiments. The value of λ_{fs} has been set to 0.001 for the simulated anomaly detection and 1 for COVID-19 and Cardiomegaly detection. For all experiments, we have used a batch-size of 16. We trained the models for 400,000 iterations. We have used Adam optimizer with a learning rate of $1e^{-4}$. The learning rate has been decayed for the last 100,000 iterations in all training settings. Once trained, we have picked the best model using Fréchet inception distance (FID) (Heusel et al., 2017; Seitzer, 2020).

4.5 Discussion

Using FID for selecting model is a bottleneck for the proposed method. We found FID did not pick the model with the best AUC score. If the best models produced by A2B-GAN were selected then the AUC score would be 0.8398 instead of 0.8364 for COVID-19 detection, 0.6383 instead of 0.6015 for Cardiomegaly detection, and 0.8571 instead of 0.8002 for simulated anomaly detection. Therefore, the proposed A2B-GAN has potential to perform better when an improved metric is available in an unsupervised scenario or a tiny annotated validation set is available in a semi-supervised setting.

5 Conclusion

We have introduced a novel one-directional unpaired image-to-image translation method for anomaly detection, named A2B-GAN. We have devised a methodology to utilize an unannotated mixed dataset with both normal and anomalous images during the training of the proposed A2B-GAN. It has been possible due to the proposed novel reconstruction loss that ensures effective cycle-consistency without requiring input image annotations. Our extensive evaluation has demonstrated that the proposed A2B-GAN's superiority over the existing state-of-the-art anomaly detection methods. The superior performance is attributed to A2B-GAN's capability of utilizing unannotated anomalous images at training time.

REFERENCES

Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pp. 622–637. Springer, 2018. 1, 2, 3, 6, 7, 8

Varghese Alex, Mohammed Safwan KP, Sai Saketh Chennamsetty, and Ganapathy Krishnamurthi. Generative adversarial networks for brain lesion detection. In *Medical Imaging 2017: Image Processing*, volume 10133, pp. 101330G. International Society for Optics and Photonics, 2017. 1, 3

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017. 5

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019. 3, 6

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021. 3, 6

Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018. 1, 3

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 4, 13

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pp. 475–489. Springer, 2021. 1, 2, 3, 6, 7, 8

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 3

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 3

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017. 1

Elies Gherbi, Blaise Hanczar, Jean-Christophe Janodet, and Witold Klaudel. An encoding adversarial network for anomaly detection. In *Asian Conference on Machine Learning*, pp. 188–203. PMLR, 2019. 1, 3

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014. 1

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017. 5

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015. 1

Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 2, 3, 4

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 9

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017. 3, 4

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pp. 1857–1865, 2017. 3

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, pp. 4, 2017. 3

Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*, pp. 702–716. Springer, 2016. 4

Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3673–3682, 2019. 2, 3, 4

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pp. 700–708, 2017. 1, 3, 4

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015. 8, 13

Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*, pp. 3693–3703, 2018. 2, 3, 4

Ori Nizan and Ayellet Tal. Breaking the cycle-colleagues are all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7860–7869, 2020. 3, 4, 6, 13

Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B Gotway, Yoshua Bengio, and Jianming Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 191–200, 2019. 4, 13

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. *arXiv preprint arXiv:2106.08265*, 2021. 1, 2, 3, 6, 7, 8

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pp. 146–157. Springer, 2017. 1, 3, 6

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 2019. 1, 2, 3, 6, 7, 8

Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.1.1. 9

Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4030–4038, 2017. 1, 3, 4

Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10 (1):19549, Nov 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-76550-z. URL https://doi.org/10.1038/s41598-020-76550-z. 7

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017. 8

Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pp. 2868–2876, 2017. 1, 3, 4

Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018a. 1, 3, 6

Houssam Zenati, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar. Adversarially learned anomaly detection. In *2018 IEEE International conference on data mining (ICDM)*, pp. 727–736. IEEE, 2018b. 1, 2, 3, 6, 7, 8, 13

Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 417–432, 2018. 2, 3, 4

Yihao Zhao, Ruihai Wu, and Hao Dong. Unpaired image-to-image translation using adversarial consistency loss. *arXiv preprint arXiv:2003.04858*, 2020. 2, 4

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017a. 2, 3, 4, 5

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pp. 465–476, 2017b. 2, 3, 4

Maria A Zuluaga, Don Hush, Edgar JF Delgado Leyton, Marcela Hernández Hoyos, and Maciej Orkisz. Learning from only positive and unlabeled data to detect lesions in vascular ct images. In *International conference on medical image computing and computer-assisted intervention*, pp. 9–16. Springer, 2011. 1

A APPENDIX: A2B-GAN ON IMAGE-TO-IMAGE TRANSLATION

A.1 DATASETS

Eyeglass Removal. We have utilized CelebA (Liu et al., 2015) dataset to create dataset for eyeglass removal task. The original dataset contains 202,599 face images of celebrities with 40 different facial attributes. For eyeglass removal task, we have divided the dataset based-on the "Eyeglasses" attribute. Our training dataset contains 10,523 images with eyeglasses and 152,251 images without eyeglasses. In contrast, our testing dataset contains 2,672 images with eyeglasses.

Male-to-Female Translation. We have also utilized CelebA (Liu et al., 2015) dataset for translating male faces to appears as female. In doing so, we have divided the dataset based-on the "Male" attribute. For this task, our dataset contains 68,261 male images and 94,509 female images. For testing, we have utilized 16,173 male images.

A.2 QUANTITATIVE RESULTS

Method	**# Gen.**	**# Disc.**	**Eyeglasses**	**Male-to-Female**
StarGAN (Choi et al., 2018)	1	1	34.62	–
Fixed-Point GAN (Rahman Siddiquee et al., 2019)	1	1	34.61	86.11
Council GAN (Nizan & Tal, 2020)	4	8	34.65	34.30
ACL GAN (Zenati et al., 2018b)	2	2	38.97	49.21
A2B-GAN	1	1	**34.29**	**34.00**

Table 2: Summary of image-to-image translation results. We have compared the proposed A2B-GAN with existing state-of-the-art image-to-image translation methods. We have used FID score for the evaluation. Our quantitative results show that the proposed A2B-GAN performs competitively with the existing methods even though using only 1 generator and 1 discriminator network.