Brainomaly: Unsupervised Neurologic Disease Detection Utilizing Unannotated T1-weighted Brain MR Images

Md Mahfuzur Rahman Siddiquee[1,2], Jay Shah[1,2], Teresa Wu[1,2], Catherine Chong[2,3], Todd J. Schwedt[2,3], Gina Dumkrieger[3], Simona Nikolova[3], and Baoxin Li[1,2]
[1]Arizona State University; [2]ASU-Mayo Center for Innovative Imaging; [3]Mayo Clinic

Abstract

Harnessing the power of deep neural networks in the medical imaging domain is challenging due to the difficulties in acquiring large annotated datasets, especially for rare diseases, which involve high costs, time, and effort for annotation. Unsupervised disease detection methods, such as anomaly detection, can significantly reduce human effort in these scenarios. While anomaly detection typically focuses on learning from images of healthy subjects only, real-world situations often present unannotated datasets with a mixture of healthy and diseased subjects. Recent studies have demonstrated that utilizing such unannotated images can improve unsupervised disease and anomaly detection. However, these methods do not utilize knowledge specific to registered neuroimages, resulting in a subpar performance in neurologic disease detection. To address this limitation, we propose Brainomaly, a GAN-based image-to-image translation method specifically designed for neurologic disease detection. Brainomaly not only offers tailored image-to-image translation suitable for neuroimages but also leverages unannotated mixed images to achieve superior neurologic disease detection. Additionally, we address the issue of model selection for inference without annotated samples by proposing a pseudo-AUC metric, further enhancing Brainomaly's detection performance. Extensive experiments and ablation studies demonstrate that Brainomaly outperforms existing state-of-the-art unsupervised disease and anomaly detection methods by significant margins in Alzheimer's disease detection using a publicly available dataset and headache detection using an institutional dataset. The code is available from https://github.com/mahfuzmohammad/Brainomaly.

1. Introduction

Deep neural networks have facilitated supervised learning from annotated datasets [22, 16], but acquiring large annotated medical imaging datasets, particularly for rare diseases, is challenging. Even when enough imaging data are available, manual annotation of such datasets is expensive, laborious, and time-consuming as it requires domain expert knowledge. In such scenarios, unsupervised disease detection methods like anomaly detection can help reduce the

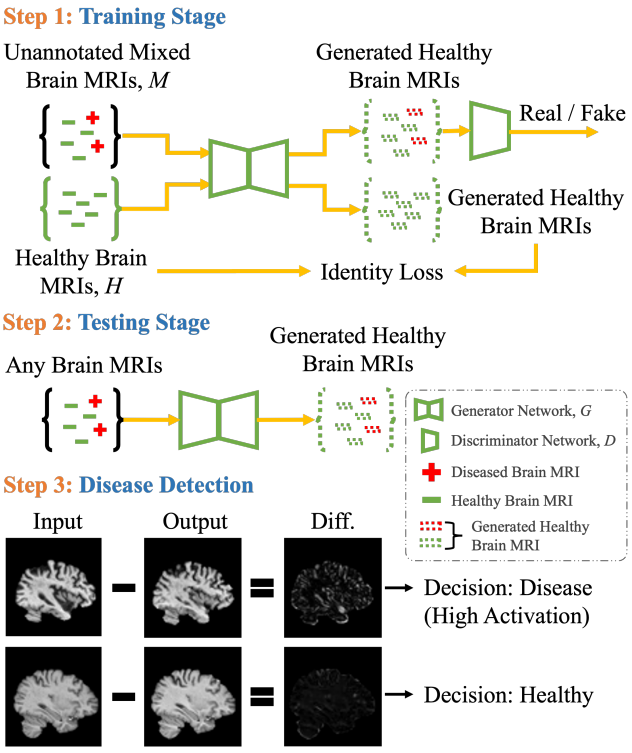

Figure 1. Overview of the proposed method, Brainomaly, for unsupervised neurologic disease detection using unannotated mixed T1-weighted brain MRIs. Brainomaly is a GAN-based image-to-image translation method that is trained (Step 1 in the figure) to remove the diseased regions from any input brain MRI and generate MRI of the corresponding healthy brain using (1) a set of "unannotated mixed brain MRIs" containing T1-weighted brain MRIs from individuals with neurologic disease as well as healthy subjects and (2) another set containing T1-weighted brain MRIs only from healthy subjects. Once trained, the generator turns any brain MRI into the MRI of the corresponding healthy brain (Step 2 in the figure). Hence, subtracting (Step 3 in the figure) the generated MRI of the healthy brain from its input would reveal structural changes if the input MRI is of an abnormal brain. We use the average value of the resultant difference map as the disease detection score, where higher values indicate a higher likelihood that the brain MRI is from someone with a neurologic disease.

annotation burden and save significant human effort. Many prior works in this area have focused on developing diagnostic models that learn to reconstruct images from healthy subjects [11, 37, 38, 3, 43, 44, 1, 17, 36]. These methods rely on poor reconstructions of images from individuals with diseases during inference for detection. However, in practice, *unannotated* images are often available from individuals with diseases (mixed with healthy subject images) in clinical databases or even as test sets where the trained model will be applied, and leveraging the additional information contained in these *unannotated mixed* images could enhance disease detection.

With the same inspiration, [34] and [10] have recently proposed methods that utilize *such* unannotated datasets of mixed images during training for improved unsupervised patient-level disease and anomaly detection. [10] trained a set of autoencoders to reconstruct chest X-ray images only from healthy subjects and another set of autoencoders to reconstruct unannotated mixed X-rays from individuals with thoracic diseases and healthy subjects. Anomaly detection scores were obtained by comparing the discrepancies between these two sets of autoencoders. Conversely, [34] employed a GAN-based image-to-image translation approach [19, 18, 25, 13, 14, 35] to remove diseased areas from input images and generate corresponding healthy-looking images. The disease detection scores for each subject were calculated by subtracting the generated healthy-looking images from their corresponding input images. These methods, however, performed suboptimally for neurologic disease detection (Tab. 1) and lacked a reliable inference model selection criterion. [10] used the model from the last training iteration for inference while [34] selected model based on the realism of the generated images using Fréchet inception distance (FID) [24]. We found that the FID metric has a weak correlation with the underlying classification performance of the model (see Sec. 6.3). To address these issues, we propose *Brainomaly*, a GAN-based image-to-image translation method specifically designed for neurologic disease detection. Brainomaly learns to remove neurologic disease from T1-weighted brain MRIs and generates corresponding healthy MRIs. During training, it utilizes an unannotated set of mixed MRIs from diseased and healthy individuals where traditional cycle-consistency-based image translation is not applicable [34]. Since neuroimages are usually registered, we design Brainomaly to predict an *additive map* to transform input images into a healthy appearance instead of directly generating healthy images. The additive map contains voxel-wise values representing the estimated changes required to transform the input MRI into a healthy brain. We hypothesize that this additive map-based translation, combined with *identity loss* (Eq. 4) regularization, relaxes the need for cycle-consistency-based image translation. For inference model selection, we introduce a *pseudo-AUC* (AUCp) metric that further boosts the detection performance. Fig. 1 depicts an overview of the Brainomaly framework.

Through extensive experiments and ablation studies, we demonstrate that Brainomaly outperforms existing state-of-the-art unsupervised disease and anomaly detection methods by significant margins on one public dataset for Alzheimer's disease detection and one institutional dataset for headache detection. Its superior performance is due to the additive map-based image translation technique, leveraging unannotated images during training and employing improved inference model selection using AUCp. In summary, we make the following contributions:

- We introduce a novel neurologic disease detection method that utilizes unannotated T1-weighted brain MRIs from individuals with neurologic disease and healthy subjects.
- We propose a new metric, AUCp, for selecting a suitable model for inference when an annotated validation dataset is unavailable.
- With two neuroimaging datasets, we perform extensive experiments comparing the proposed method, Brainomaly, against the conventional state-of-the-art unsupervised patient-level disease and anomaly detection methods. Our detailed analysis proves the superiority of the proposed method over existing methods.
- We evaluate the proposed method in both transductive and inductive settings to match real-world scenarios.
- We empirically show that our proposed metric has a higher correlation with the models' underlying disease detection performances and selects a higher-performing model than a model selected by FID, which is commonly used in GAN model development.

2. Related Work

Our work is closely related to image-to-image translation, GAN-based anomaly detection, and neurologic disease detection. Hence, we review relevant existing efforts on these tasks and contrast them with our proposed neurologic disease detection method, Brainomaly.

2.1. Image-to-Image Translation

Plenty of work has been done on GAN-based image-to-image translation [25, 27, 28, 31, 40, 42, 47, 48, 13, 2, 45, 23, 30, 32]. Pix2Pix [25] and CycleGAN [47] are pioneer methods in this area. While Pix2Pix requires paired input-output images for training, CycleGAN introduces the concept of cycle consistency for unpaired image-to-image translation. However, we cannot directly use CycleGAN in

our work on unsupervised disease and anomaly detection due to the lack of annotated diseased images.

Recent unpaired image-to-image translation methods [31, 40, 42, 47, 48, 13, 2, 45, 23, 30, 46], regardless of cycle-consistency, also rely on image annotations. For example, [40] utilizes two generators for translating images of human faces between a pair of facial attributes. [2] proposes an attention-based approach that performs image-to-image translation like CycleGAN with two additional networks for generating attention maps. Alternatively, StarGAN [13], AttGAN [23], STGAN [30], and Fixed-Point GAN [35] utilize one generator network that requires target image annotations. A recent ensemble-based method [32] offers an alternative to cycle consistency but demands multiple generator and discriminator networks, making it computationally expensive and difficult to train.

In contrast, our Brainomaly method, which employs only one generator and one discriminator network, overcomes the need for cycle consistency by generating additive maps instead of images. Furthermore, Brainomaly outperforms existing anomaly detection methods, as shown in Sec. 6.

2.2. Anomaly Detection

In general, GAN-based anomaly detection methods [11, 37, 38, 3, 43, 44, 1, 17, 10] in the existing literature primarily focus on learning from *healthy* images. These methods aim to capture the underlying manifold of healthy images, enabling their decoders to reconstruct only healthy images during testing. Consequently, when *diseased* images are reconstructed as healthy, the disparity between the input and output images indicates the presence of anomalies. We elaborate on a few examples below.

Chen *et al*. [11] employ an adversarial autoencoder to learn the distribution of healthy data. They identify anomalies by subtracting the reconstructed diseased image from the input image. In a similar vein, [37] proposes a method that adversarially learns a decoder model to generate healthy images from random noise vectors in the latent space. During testing, this approach maps new images to the latent space through iterative updates of the latent vector. If a new image is healthy, the method is expected to find the actual latent vector that reconstructs the input image, resulting in a negligible difference between the input and reconstructed images. Conversely, for diseased images, the method should find a latent vector that produces the closest healthy image, leading to a higher difference between the input and reconstructed images. The authors propose an anomaly score that combines the reconstruction error and the discrimination score from the discriminator network.

Schlegl *et al*. [38] enhance the speed of [37] by introducing an encoder network capable of mapping input images to the latent space in a single pass. Likewise, [3] employs a GAN to learn a generative model of healthy data. It involves scanning images pixel-by-pixel and feeding the cropped regions to a trained GAN discriminator. An anomaly map is then constructed by combining the anomaly scores provided by the discriminator. Zenati *et al*. [43, 44] utilize BiGAN [15] to jointly train an encoder and a decoder network to learn the mapping of normal images. Like most methods, they use the reconstruction error as the anomaly score.

Akcay *et al*. [1] train an autoencoder supervised with both image-level L_1 distance and adversarial loss using only normal images. Additionally, they train an extra encoder to map the images reconstructed by the autoencoder back to their latent space. In a different approach, [17] trains an encoder network to map normal images to a Gaussian distribution and abnormal images to an out-of-distribution region using adversarial learning. Anomalies are then detected using the Mahalanobis distance in the latent space. It is important to note that this method requires annotated anomalous images during training.

In contrast, [34] and [10] learn from unannotated images of both diseased and healthy individuals, similar to our proposed method as discussed in Sec. 1. However, in Sec. 6, we demonstrate that our method significantly outperforms these existing methods by a large margin.

2.3. Neurologic Disease Detection

Numerous studies have explored deep learning techniques for automated Alzheimer's disease diagnosis using raw imaging data. Most of these studies focus on supervised classification tasks, while a few employ unsupervised anomaly detection methods [9, 21, 7, 26, 5, 12]. For instance, Cabreza *et al*. [9] train a GAN on healthy images, followed by an encoder that returns a vector for input images like [38]. MADGAN [21] leverages MRI slice continuity in reconstruction and uses high reconstruction loss for anomalous image classification. Baydargil *et al*. [7] incorporate a parallel feature extractor within a GAN using PET images, while Choi *et al*. [12] employ a variational autoencoder on PET images for abnormality scoring based on reconstruction error. Jin *et al*. [26] use an adversarial autoencoder for unsupervised data characterization of healthy controls and subsequent Alzheimer's disease vs. healthy control classification. Bai *et al*. [5] combine a classifier with GAN training, incorporating high-level feature extraction and posterior class probabilities.

Except for [5], the aforementioned approaches rely solely on healthy images for GAN training and utilize high reconstruction loss to detect anomalies. In contrast, our method leverages both unannotated mixed images and healthy images during training, leading to superior Alzheimer's disease detection than state-of-the-art methods.

Regarding headache detection from structural MRI scans, we found no unsupervised approaches in the literature. Only a few studies employ deep learning tech-

niques for this task. Rahman Siddiquee *et al*. [33] develop a ResNet-based binary classification model for automated biomarker extraction of headache sub-types. Yang *et al*. [41] proposes a deep convolutional neural network using pre-processed resting-state fMRI data to distinguish between migraine and healthy controls. However, both studies are supervised classification tasks and suffer from limited datasets, which is common in headache classification using deep learning. Thus, our method enhances unsupervised headache detection by utilizing unannotated MRIs from both headache and healthy subjects.

3. The Proposed Method: Brainomaly

Fig. 1 depicts the overview of the proposed method. This section presents details about the neural network models, their training process, selecting the inference model with the proposed AUCp metric, and neurologic disease detection using the outputs from these models.

3.1. The Networks

Brainomaly consists of a discriminator network and a generator network. The discriminator network follows PatchGAN [25, 29, 47] architecture and is similar to the ones used in [13, 35, 34]. Our discriminator distinguishes whether a T1-weighted brain MRI is real or generated.

The generator network is an encoder-decoder type architecture similar to the generator networks used in [13, 35, 34]. As input, the network takes a T1-weighted brain MRI of any subject without knowing whether the subject is healthy or has a neurologic disease. As output, it generates an *additive map* where each voxel contains the value of an estimated required change to turn the brain in the input MRI into a healthy brain. The final healthy-brain MRI is generated by first summing the input MRI and the additive map voxel-wise, then applying $tanh$ activation on the resultant.

Both our generator and discriminator networks operate on 2D MRI slices. The generated healthy-brain MRI is constructed by stacking the generated MRI slices as they appeared in the input MRI. The architecture details for both these networks are provided in the Appx. A.2.

3.2. Training the Networks

Fig. 2 provides a detailed schema for training the generator and discriminator networks of Brainomaly. We train the generator and the discriminator network alternately, like any GAN model. At each training step, we update the generator's weights once for every two weight updates of the discriminator network and repeat them until convergence.

The role of the discriminator network is to improve the generator network by providing iterative feedback during training. At each iteration (Step 1 in Fig. 2), the discriminator network learns to distinguish the real MRIs of healthy brains and the generated MRIs from the previous iteration's generator. During the generator training, it provides feedback so that the generator can improve the quality of the generated MRIs. To be able to perform this role, the discriminator is trained on a set of T1-weighted healthy brain MRIs, H (Fig. 1), to classify them as *real*, as well as on generated MRI slices to classify them as *fake*. During discriminator training, the generator solely uses MRIs from the unannotated mixed set M (Fig. 1) to generate MRIs of healthy brains, excluding any MRIs from set H as MRIs of healthy brains are already present in M. The training objective is achieved using an *adversarial loss* (Eq. 1).

$$\begin{aligned}\mathcal{L}^{D}_{adv} = \mathbb{E}_{x_M \in M}[& \\ & log(1 - D_{real/fake}(tanh(x_M + G(x_M))))] \\ & + \mathbb{E}_{x_H \in H}[log(D_{real/fake}(x_H))]\end{aligned} \quad (1)$$

Here, x_M and x_H are MRI of random subjects from M and H, respectively. The generator's output, which is the generated MR images of a healthy brain, is denoted as $G(.)$. The discriminator network's output is represented by $D_{real/fake}(.)$. We revised Eq. 1 based on the Wasserstein GAN [4] and added a gradient penalty [20] with weight λ_{gp} to enhance training stability. The revised training objective is shown in Eq. 2 where $\hat{x}$ is a random weighted average of a batch of real and generated MRIs.

$$\begin{aligned}\mathcal{L}^{D}_{adv} = & \ \mathbb{E}_{x_M \in M}[D_{real/fake}(tanh(x_M + G(x_M)))] \\ & - \mathbb{E}_{x_H \in H}[D_{real/fake}(x_H)] \\ & + \lambda_{gp}\, \mathbb{E}_{\hat{x}}[(||\nabla_{\hat{x}} D_{real/fake}(\hat{x})||_2 - 1)^2]\end{aligned} \quad (2)$$

On the other hand, the generator aims to generate realistic MRIs of healthy brains by utilizing the discriminator's iterative feedback (Step 2 in Fig. 2). For training, it translates MRI from both M and H sets and updates the generated MRIs of healthy brains iteratively and gradually so that the discriminator fails to distinguish the generated MRIs from the real ones. For the set of unannotated mixed MRIs (M), if the input is an MRI containing a neurologic disease, then the generator is expected to remove the diseased regions and generate the MRI of a corresponding healthy brain. If the input is already an MRI of a healthy brain, the generator is expected to behave like an autoencoder; that is, it is expected to generate exactly the same input MRI in output. As MRIs in set M are unannotated, we used the *adversarial loss* for generating the corresponding MRIs of healthy brains. The loss is defined as in Eq. 3.

$$\mathcal{L}^{G}_{adv} = -\mathbb{E}_{x_M \in M}[D_{real/fake}(tanh(x_M + G(x_M)))] \quad (3)$$

For the set of healthy-brain MRIs (H), we explicitly train the generator to be an autoencoder using the *identity loss* (defined in Eq. 4) for these MRI slices.

$$\mathcal{L}_{id} = \mathbb{E}_{x_H \in H}[||tanh(x_H + G(x_H)) - x_H||_1] \quad (4)$$

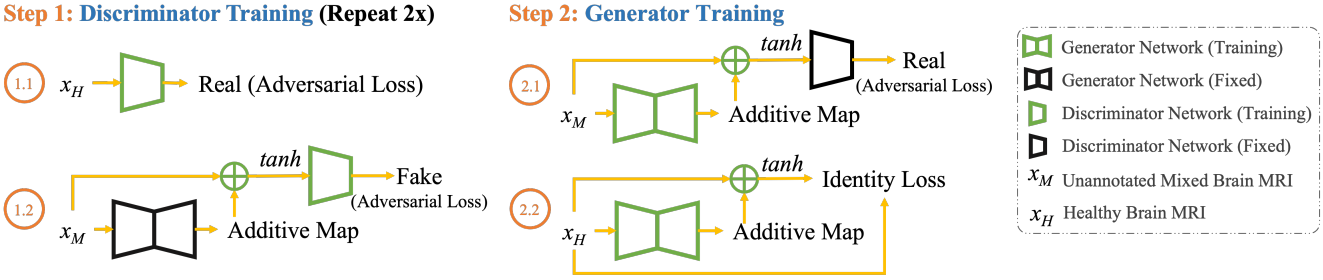

Figure 2. A training iteration in Brainomaly involves two steps: (1) the training of the discriminator and (2) the training of the generator. The discriminator is taught to identify the difference between real images and those generated by the generator. Meanwhile, the generator attempts to create indistinguishable images from real images to trick the discriminator. The iteration is repeated multiple times during the training phase to improve the realism of the generated images.

Combining all these losses, the final full objective function for the discriminator and generator can be described by Eq. 5 and Eq. 6, respectively.

$$\mathcal{L}_D = \mathcal{L}_{adv}^D \tag{5}$$

$$\mathcal{L}_G = \mathcal{L}_{adv}^G + \lambda_{id}\mathcal{L}_{id} \tag{6}$$

where λ_{id} is the relative importance of the *identity loss*.

3.3. Detecting the Diseases

We detect diseases from the MRIs using *disease detection scores*. To get these scores, we translate all the given brain MRIs to MRIs of healthy brains using a trained generator model of Brainomaly (Step 2 in Fig. 1). Then, we subtract the generated MRIs of healthy brains from their corresponding input MRIs (Step 3 in Fig. 1). If the brain in the input MRI is diseased (*i.e.* abnormal), the resultant difference map would reveal structural changes. The difference map should reveal less or no structural changes for an input MRI of a healthy brain. We call the voxels showing the structural changes as *activations*. The average of all the activations in the difference map of an input MRI is its *disease detection score*, where a higher score indicates a higher likelihood of the input brain being diseased.

3.4. Inference Model Selection using AUCp

As discussed in Sec. 3.2 and Fig. 2, Brainomaly learns iteratively from the given T1-weighted brain MRIs, generating multiple model checkpoints each after a fixed number of iterations. In a supervised learning setting, these models would be evaluated on a small validation dataset, and the best-performing model would be selected for inference and disease detection. However, annotated validation dataset is unavailable in our problem setting. Therefore, we use our proposed AUCp metric for model selection. To calculate AUCp, we first generate the *disease detection scores* for each model, as discussed in Sec. 3.3. As we already know that the set H contains only healthy-brain MRIs, we assume the labels for MRIs in the unannotated mixed brain MRI set, M, to be diseased brains. Then, we use these *imperfect* annotations as ground truths along with the *disease detection scores* in the traditional AUC calculation resulting in AUCp scores. Once the AUCp scores are available for all the models, we select a model with the highest AUCp score for inference. In Sec. 6.3, we have also shown that AUCp sets a better-performing model for inference compared to FID, commonly used in existing works [34]. Appx. A.1 shows a schematic diagram for the AUCp calculation.

4. Datasets

4.1. Alzheimer's Disease Dataset

The Alzheimer's disease dataset was obtained from the ADNI database (adni.loni.usc.edu), which is a large-scale public repository of clinical, neuropsychological, behavioral, genetic, and neuroimaging data to track the progression of Alzheimer's disease dementia. Using data from 3 studies that ADNI offers, ADNI-1, ADNI-2, and ADNI-GO, we collected and processed T1 MRI scans of 536 Alzheimer's disease patients and 1271 healthy controls.

We randomly selected 501 MRIs from healthy controls for our experiments for the healthy brain MRI set (H). We created two unannotated mixed brain MRI sets (M): *AD DS1* and *AD DS2*. Each contains 268 MRIs from patients with Alzheimer's disease and 385 MRIs from healthy controls. Splitting the dataset helps us evaluate the proposed Brainomaly for Alzheimer's disease detection in both transductive and inductive settings in Sec. 6.3.

All 3D MRIs in this dataset were registered to the *MNI152 1mm* template and skull stripped. We converted the 3D MRIs and saved them as 2D sagittal slices. The proposed Brainomaly method performs a prediction for each 2D slice. We aggregated the slice-level predictions by averaging them for patient-level predictions during evaluation.

4.2. Headache Dataset

We collected MRIs of 96 individuals with migraine, 48 with acute post-traumatic headache (APTH), 49 with persistent post-traumatic headache (PPTH), diagnosed according to the International Classification of Headache Disorders (ICHD) diagnostic criteria, and 104 healthy controls from

Mayo Clinic. We extended our dataset with MRIs of 428 healthy controls from the publicly available IXI dataset [8].

For our experiments, we trained our model by combining all headache types into one group first and then investigated each subgroup's performance separately in the post-analysis. We randomly selected 232 MRIs of healthy controls for the healthy brain MRI set (H). We created two unannotated mixed brain MRI sets (M): *HEAD DS1* and *HEAD DS2*. Each contains an equal number of MRIs for migraine (n = 48), APTH (n = 24), and healthy controls (n = 150). 24 out of 49 MRIs of those with PPTH were included in *HEAD DS1* and the rest in *HEAD DS2*. Similar to the experiment on Alzheimer's disease's dataset, such splitting helps evaluate the proposed Brainomaly for headache detection in both transductive and inductive settings in Sec. 6.3.

All 3D MRIs in this dataset were registered to the *MNI152 1mm* template and skull stripped. We converted the 3D MRIs and saved them as 2D sagittal slices. We aggregated the slice-level predictions by averaging them for patient-level predictions during evaluation.

5. Experiments

We evaluated our proposed Brainomaly on Alzheimer's disease (Sec. 6.1) and headache (Sec. 6.2) detection comparing with six state-of-the-art unsupervised disease/anomaly detection methods. Among these, DDAD [10] and HealthyGAN [34] also utilize unannotated mixed images like Brainomaly. On the other hand, ALAD [44], ALOOC [36], f-AnoGAN [38], and Ganomaly [1] learn only from images of healthy subjects. In addition, we analyzed Brainomaly's performance in transductive and inductive learning settings, provided an ablation study of the object functions, and compared our proposed AUCp metric with FID (Sec. 6.3).

All of our models operate on 2D MRI slices. We used 2D sagittal slices for all experiments. We performed a central crop to remove empty regions outside the brain, resulting in 192×192 sagittal slices for both datasets. We used $\lambda_{id} = 1$ and a batch size of 16. We trained the models for 400,000 iterations and saved a model for AUCp calculation after every 10,000 iterations. We have used Adam optimizer with a $1e^{-4}$ learning rate. The learning rate has been decayed for the last 100,000 iterations.

6. Results and Analyses

6.1. Alzheimer's Disease Detection

Tab. 1 compares Brainomaly's Alzheimer's disease detection performance on both *AD DS1* and *AD DS2* data with six state-of-the-art methods. *AD DS1* and *AD DS2* columns report the numbers when these data were used as an unseen test set. Brainomlay outperforms the existing methods by a large margin, achieving an average Alzheimer's disease detection AUC of 0.6550. Among the competing methods, f-AnoGAN performed the best. It achieved an average AUC of 0.6020, which is *8.09% less* than the proposed Brainomaly. Ganomaly and DDAD performed close to f-AnoGAN, achieving average AUC of 0.5956 and 0.5926, respectively. The rest of the competing methods performed like random guesses. Fig. 3 (left) shows Brainomaly's qualitative results for Alzheimer's disease detection. As seen, the difference maps for subjects with Alzheimer's disease have higher activation than that of healthy subjects. Receiver operating characteristics (ROC) curve analyses are provided in Appx. A.3.

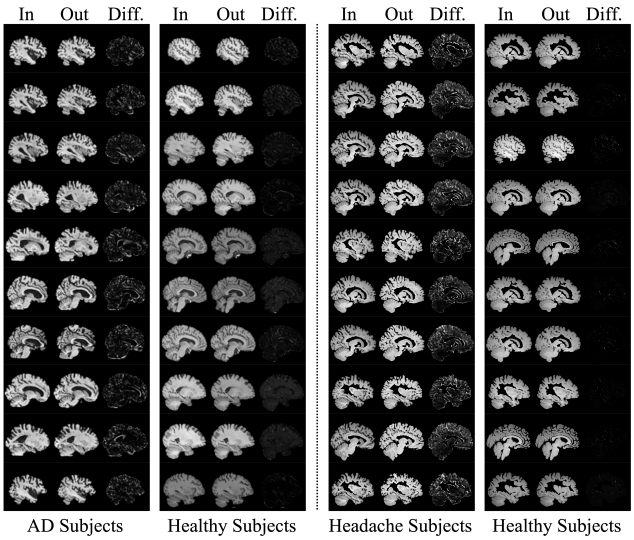

Figure 3. Qualitative results of Alzheimer's disease and headache detection by Brainomaly. The left two columns display the results of Alzheimer's disease detection experiments, while the right two columns depict the outcomes of headache detection experiments. As expected, Brainomaly exhibits higher activations in the difference map for diseased subjects in comparison to healthy subjects, which is the base for its disease detection.

6.2. Headache Detection

Tab. 1 also compares Brainomaly's headache detection performance on both *HEAD DS1* and *HEAD DS2* data with six state-of-the-art methods. Like Alzheimer's disease detection, Brainomaly also outperforms the competing methods in headache detection by a large margin. It achieved an average headache detection AUC of 0.8960. Performing *13.84% less* than Brainomaly, HealthyGAN achieved the second-best average AUC of 0.7720. Other baseline methods like ALAD, Ganomaly, and DDAD achieved even poorer AUCs of 0.0.7653, 0.6913, and 0.6280, respectively. f-AnoGAN and ALOOC just failed in this task.

Brainomaly also performed better in detecting headache sub-types. On *HEAD DS1*, it achieved a precision of 0.9375 (3 incorrect out of 48) in migraine detection, 0.3750 (15

Training Data	Methods	Alzheimer's Disease Dataset			Headache Dataset		
		AD DS1	*AD DS2*	***Average***	*HEAD DS1*	*HEAD DS2*	***Average***
Healthy Only	ALAD [44]	0.5329	0.5239	0.5284	<u>0.7819</u>	0.7486	0.7653
	ALOOC [36]	0.4670	0.4746	0.4708	0.3044	0.6566	0.4805
	f-AnoGAN [38]	<u>0.5946</u>	<u>0.6093</u>	<u>0.6020</u>	0.4354	0.3925	0.4071
	Ganomaly [1]	0.5864	0.6048	0.5956	0.7313	0.6514	0.6913
Healthy + Unannotated Mixed	DDAD [10]	0.5897	0.5955	0.5926	0.6128	0.6431	0.6280
	HealthyGAN [34]	0.4598	0.5468	0.5033	0.7107	<u>0.8333</u>	<u>0.7720</u>
	HealthyGAN (AUCp)	0.5905	0.6034	0.5970	0.8276	0.7899	0.8088
	Brainomaly (FID)	0.6389	0.6453	0.6421	0.9002	0.8589	0.8796
	Brainomaly (AUCp)	**0.6452**	**0.6648**	**0.6550**	**0.9041**	**0.8878**	**0.8960**

Table 1. Comparison of Brainomaly's performance with state-of-the-art anomaly detection methods on Alzheimer's disease and headache detection on unseen test sets using AUC metric. Numbers in **boldface** indicate the best results, and <u>underlined</u> numbers indicate the second-best results. As seen, Brainomaly outperforms all the existing state-of-the-art methods for neurologic disease detection. The rows "HealthyGAN (AUCp)" and "Brainomaly (FID)" are for ablation study purpose only (see Sec. 6.3).

(a) Transductive vs. Inductive Learning

Alzheimer's Disease Dataset				
	AD DS1	*AD DS2*	Avg.	*p*-value
Transduc.	0.6526	0.6825	0.6676	0.555
Inductive	0.6452	0.6648	0.6550	
Headache Dataset				
	HEAD DS1	*HEAD DS2*	Avg.	*p*-value
Transduc.	0.9182	0.8633	0.8908	0.873
Inductive	0.9041	0.8878	0.8960	

(b) Importance of Identity Loss

Alzheimer's Disease Dataset			
		AD DS1	*AD DS2*
Transduc.	**Brainomaly**	**0.6526**	**0.6825**
	$-\mathcal{L}_{id}$	0.6303	0.6815
Inductive	**Brainomaly**	0.6452	**0.6648**
	$-\mathcal{L}_{id}$	**0.6521**	0.6455
Headache Dataset			
		HEAD DS1	*HEAD DS2*
Transduc.	**Brainomaly**	**0.9182**	**0.8633**
	$-\mathcal{L}_{id}$	0.7824	0.8091
Inductive	**Brainomaly**	**0.9041**	**0.8878**
	$-\mathcal{L}_{id}$	0.8073	0.8359

(c) FID vs. AUCp: Correlation with AUC

Alzheimer's Disease Dataset		
	AD DS1	*AD DS2*
FID	0.5701	0.4773
AUCp (Our)	**0.9583**	**0.9656**
Headache Dataset		
	HEAD DS1	*HEAD DS2*
FID	0.5227	0.3187
AUCp (Our)	**0.9528**	**0.5986**

(d) FID vs. AUCp: Detection Performance

Alzheimer's Disease Dataset			
		AD DS1	*AD DS2*
Transduc.	FID	0.618	0.6771
	AUCp (Our)	**0.6526**	**0.6825**
Inductive	FID	0.6389	0.6453
	AUCp (Our)	**0.6452**	**0.6648**
Headache Dataset			
		HEAD DS1	*HEAD DS2*
Transduc.	FID	0.8807	**0.9120**
	AUCp (Our)	**0.9182**	0.8633
Inductive	FID	0.9002	0.8589
	AUCp (Our)	**0.9041**	**0.8878**

Table 2. These ablation studies of different components of Brainomaly show its (a) generalization ability on both unannotated seen and unseen datasets, (b) effectiveness of the objective function, and (c–d) superiority of the proposed AUCp metric for inference model selection.

incorrect out of 24; see discussion) in APTH detection, and 0.9600 (only 1 incorrect out of 25) in PPTH detection. On *HEAD DS2*, it achieved a precision of 0.9167 (4 incorrect out of 48) in migraine detection, 0.6667 (8 incorrect out of 24) in APTH detection, and 0.9583 (only 1 incorrect out of 24) in PPTH detection.

Fig. 3 (right) shows Brainomaly's qualitative results for headache detection. Similar to Alzheimer's disease detection, the difference maps for subjects with headaches have higher activation than those for healthy subjects. ROC curve analyses are provided in Appx. A.3.

6.3. Ablation Studies

Comparison of Image-to-Image Translation. To better understand the contribution of Brainomlay's additive map-based image-to-image translation, we have compared it with HealthyGAN [34] by keeping the network architecture, data-split, and inference model selection metrics the same. The results summarized in Tab. 1 show that Brainomaly consistently outperformed HealthyGAN across tasks irrespective of inference model selection metrics.

Transductive vs. Inductive Learning. Using an unannotated set of mixed brain MRIs (Fig. 1) allows Brainomaly

to operate in transductive and inductive learning modes. Therefore, we evaluate our proposed Brainomaly in both learning settings. For the transductive learning setting, we evaluate Alzheimer's disease and headache detection on the unannotated mixed brain MRI set used during training. In contrast, for the inductive learning setting, we utilize an additional unseen test set for Alzheimer's disease detection and headache detection evaluation. Please note that the performance reported in Tab. 1 and analyzed in the previous two subsections were in inductive settings.

Tab. 2a summarizes Brainomaly's performance for Alzheimer's disease and headache detection in both transductive and inductive learning settings. The average performance of Brainomaly for Alzheimer's disease and headache detection is statistically the same (p-value > 0.005) in both transductive and inductive learning settings. These results show that Brainomaly generalizes well on unseen test data.

Impact of Identity Loss on Objective Function. Inspired from [35], we incorporated the *identity loss* ($\mathcal{L}_{id}$ in Eq. 4) to balance the image translation. As seen in Tab. 2b, $\mathcal{L}_{id}$ plays an important role in Brainomaly's image translation and significantly improves its performance for both Alzheimer's disease and headache detection.

Inference Model Selection—FID vs. AUCp. In Tab. 2c, we have shown that our AUCp score proposed in Sec. 3.4 has a stronger correlation with the actual (when all the annotations are available) AUC scores. Therefore, the proposed AUCp metric renders itself a better metric than FID for selecting the model for inference. Please note that Tab. 2c reports absolute correlation values. To further validate, we have provided the AUC scores obtained by the best models according to FID and the AUCp scores in both transductive and inductive learning settings for each dataset in Tab. 2d. It is evident from the figure that the models selected by our AUCp metric dominate in detection performance over the models selected by FID.

7. Discussion

The proposed Brainomaly method aims to perform patient-level neurologic disease detection without requiring brain image annotation. Though it generates the difference maps showing structural changes in Alzheimer's disease and headache subjects (Fig. 3), these maps are not precise. They show more structural changes than actual changes performed by the underlying diseases; as a result, they are not useful for precise localization. If needed, weakly-supervised localization methods such as GradCAM [39], Fixed-Point GAN [35], and VAGAN [6] can be utilized for better localization using the patient-level detections from Brainomaly as weak annotations.

The proposed AUCp metric does not guarantee the selection of the best possible model for inference as it uses *imperfect* annotations. However, our empirical analyses show that AUCp generally selects a better inference model than the popular FID metric.

In Sec. 6.2, we have seen that APTH detection using Brainomaly is not as good as detecting other headache subtypes. This might be due to the acuity of the condition and greater heterogeneity in brain structural changes amongst these individuals compared to those who have had long-standing migraine or PTH (i.e., those with PPTH). Among the 15 misclassified APTH subjects in *HEAD DS1*, we found 5 were recovered at a 3-month time point. This improves the APTH detection rate from 0.3750 to 0.5833. Similarly, in *HEAD DS2*, 1 out of 8 misclassified subjects recovered at a 3-month time point, improving the detection rate from 0.6667 to 0.7083. Future studies are needed to explore the heterogeneity amongst those with APTH.

Using images from healthy individuals is common in unsupervised anomaly detection literature [1, 36, 38, 44]. Our method leverages an unannotated mixed dataset without added annotation costs. Besides, images of healthy individuals are readily accessible in **P**icture **A**rchiving and **C**ommunication **S**ystems (PACS), making our proposed method, Brainomaly, virtually annotation cost-free to train.

8. Conclusion

In conclusion, the proposed unsupervised neurologic disease detection method, Brainomaly, is highly effective in detecting Alzheimer's disease and headaches from T1-weighted brain MRIs, outperforming existing state-of-the-art methods by a large margin. This performance is attributed to Brainomaly's additive map-based image translation, the capability of utilizing unannotated mixed brain MRIs, and better inference model selection using the proposed AUCp metric. Using an unannotated set of mixed brain MRIs enables Brainomaly to operate in both transductive and inductive learning modes, providing flexibility in its application. In addition, we have shown in Tab. 1 that the AUCp can select better models even for existing methods, for example, HealthyGAN. We believe the proposed Brainomaly method can be generalized for unsupervised disease detection from other organs and modalities, which we aim to study in our future work.

Acknowledgments. This research has been supported by the United States Department of Defense W81XWH-15-1-0286 and W81XWH1910534, National Institutes of Health K23NS070891, National Institutes of Health - National Institute of Neurological Disorders and Stroke, Award Number 1R61NS113315–01, and Amgen Investigator Sponsored Study 20187183. We thank Arizona State University Research Computing (ASURC) for hosting and maintaining our computing resources.

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 622–637. Springer, 2019.

[2] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. *Advances in neural information processing systems*, 31, 2018.

[3] 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, pages 113–121. SPIE, 2017.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[5] Tian Bai, Mingyu Du, Lin Zhang, Lei Ren, Li Ruan, Yuan Yang, Guanghao Qian, Zihao Meng, Li Zhao, and M Jamal Deen. A novel alzheimer's disease detection approach using gan-based brain slice image enhancement. *Neurocomputing*, 492:353–369, 2022.

[6] Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2017.

[7] Husnu Baris Baydargil, Jang-Sik Park, and Do-Young Kang. Anomaly analysis of alzheimer's disease in pet images using an unsupervised adversarial deep learning model. *Applied Sciences*, 11(5):2187, 2021.

[8] Ixi dataset, NA.

[9] Jean Nathan Cabreza, Geoffrey A Solano, Sun Arthur Ojeda, and Vincent Munar. Anomaly detection for alzheimer's disease in brain mris via unsupervised generative adversarial learning. In *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 1–5. IEEE, 2022.

[10] Yu Cai, Hao Chen, Xin Yang, Yu Zhou, and Kwang-Ting Cheng. Dual-distribution discrepancy for anomaly detection in chest x-rays. *arXiv preprint arXiv:2206.03935*, 2022.

[11] Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. *CoRR*, abs/1806.04972, 2018.

[12] Hongyoon Choi, Seunggyun Ha, Hyejin Kang, Hyekyoung Lee, Dong Soo Lee, Alzheimer's Disease Neuroimaging Initiative, et al. Deep learning only by normal brain pet identify unheralded brain anomalies. *EBioMedicine*, 43:447–453, 2019.

[13] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[14] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[15] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *CoRR*, abs/1605.09782, 2016.

[16] 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.

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

[18] 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.

[19] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[21] Changhee Han, Leonardo Rundo, Kohei Murao, Tomoyuki Noguchi, Yuki Shimahara, Zoltán Ádám Milacski, Saori Koshino, Evis Sala, Hideki Nakayama, and Shin'ichi Satoh. Madgan: Unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction. *BMC bioinformatics*, 22(2):1–20, 2021.

[22] 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*, pages 1026–1034, 2015.

[23] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Arbitrary facial attribute editing: Only change what you want. *CoRR*, abs/1711.10678, 2017.

[24] 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.

[25] 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*, pages 1125–1134, 2017.

[26] Shichen Jin, Peini Zou, Ying Han, and Jiehui Jiang. Unsupervised detection of individual atrophy in alzheimer's disease. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2647–2650. IEEE, 2021.

[27] 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, pages 1857–1865. PMLR, 2017.

[28] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

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

[30] 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/CVF conference on computer vision and pattern recognition*, pages 3673–3682, 2019.

[31] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

[32] 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*, pages 7860–7869, 2020.

[33] Md Mahfuzur Rahman Siddiquee, Jay Shah, Catherine Chong, Simona Nikolova, Gina Dumkrieger, Baoxin Li, Teresa Wu, and Todd J Schwedt. Headache classification and automatic biomarker extraction from structural mris using deep learning. *Brain Communications*, 5(1):fcac311, 2023.

[34] Md Mahfuzur Rahman Siddiquee, Jay Shah, Teresa Wu, Catherine Chong, Todd Schwedt, and Baoxin Li. Healthy-gan: Learning from unannotated medical images to detect anomalies associated with human disease. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 43–54. Springer, 2022.

[35] 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/CVF international conference on computer vision*, pages 191–200, 2019.

[36] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. *CoRR*, abs/1802.09088, 2018.

[37] 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 *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*, pages 146–157. Springer, 2017.

[38] 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*, 54:30–44, 2019.

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[40] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4030–4038, 2017.

[41] Hao Yang, Junran Zhang, Qihong Liu, and Yi Wang. Multimodal mri-based classification of migraine: using deep learning convolutional neural network. *Biomedical engineering online*, 17(1):1–14, 2018.

[42] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[43] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018.

[44] 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)*, pages 727–736. IEEE, 2018.

[45] 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)*, September 2018.

[46] Yihao Zhao, Ruihai Wu, and Hao Dong. Unpaired image-to-image translation using adversarial consistency loss. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 800–815. Springer, 2020.

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[48] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.