# Learning Fixed Points in Generative Adversarial Networks: From Image-to-Image Translation to Disease Detection and Localization

Md Mahfuzur Rahman Siddiquee[1], Zongwei Zhou[1,3], Nima Tajbakhsh[1], Ruibin Feng[1],
Michael B. Gotway[2], Yoshua Bengio[3], and Jianming Liang[1,3]
[1]Arizona State University; [2]Mayo Clinic; [3]Mila – Quebec Artificial Intelligence Institute

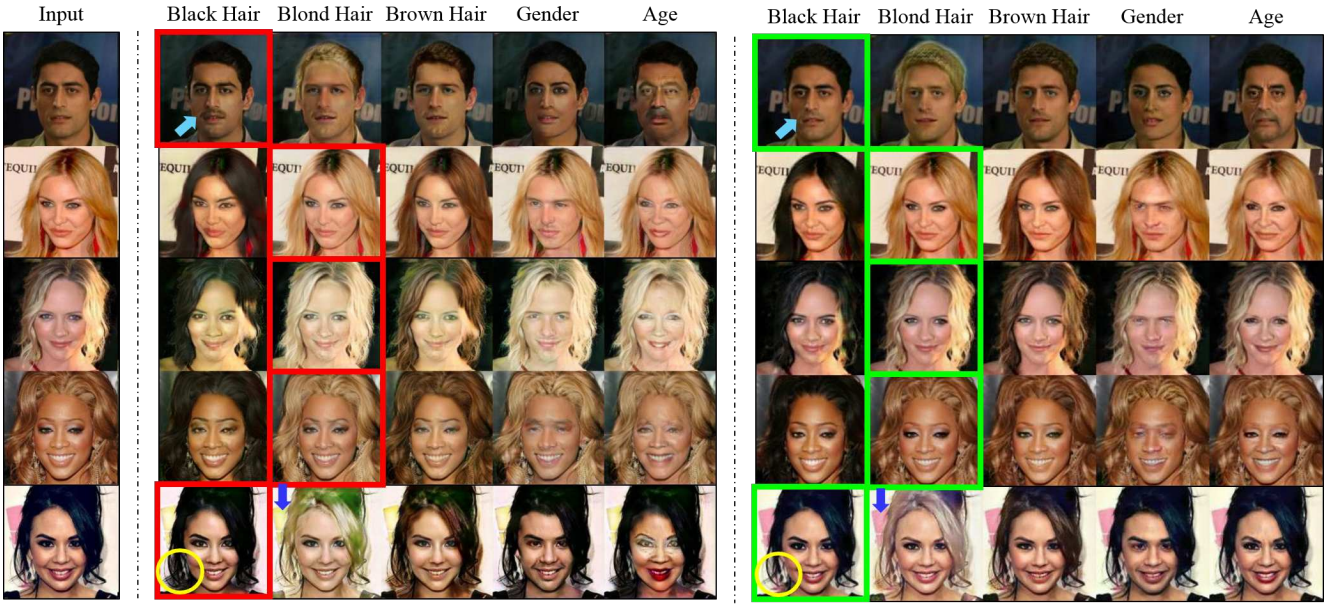


Fig. 1: [Better viewed on-line in color and zoomed in for details] Comparing our Fixed-Point GAN with StarGAN [8], the state of the art in multi-domain image-to-image translation, by translating images into five domains. Combining the domains may yield a same-domain (*e.g.*, black to black hair) or cross-domain (*e.g.*, black to blond hair) translation. For clarity, same-domain translations are framed in red for StarGAN and in green for Fixed-Point GAN. As illustrated, during cross-domain translations, and especially during same-domain translations, StarGAN generates artifacts: introducing a mustache (Row 1, Col. 2; light blue arrow), changing the face colors (Rows 2–5, Cols. 2–6), adding more hair (Row 5, Col. 2; yellow circle), and altering the background (Row 5, Col. 3; blue arrow). Our Fixed-Point GAN overcomes these drawbacks via fixed-point translation learning (see Sec. 3) and provides a framework for disease detection and localization with only image-level annotation (see Fig. 2).

## Abstract

*Generative adversarial networks (GANs) have ushered in a revolution in image-to-image translation. The development and proliferation of GANs raises an interesting question: can we train a GAN to remove an object, if present, from an image while otherwise preserving the image? Specifically, can a GAN "virtually heal" anyone by turning his medical image, with an unknown health status (diseased or healthy), into a healthy one, so that diseased regions could be revealed by subtracting those two images? Such a task requires a GAN to identify a minimal subset of target pixels for domain translation, an ability that we call fixed-point translation, which no GAN is equipped with yet. Therefore, we propose a new GAN, called Fixed-Point GAN, trained by (1) supervising same-domain translation through a conditional identity loss, and (2) regularizing cross-domain translation through revised adversarial, domain classification, and cycle consistency loss. Based on fixed-point translation, we further derive a novel framework for disease detection and localization using only image-level annotation. Qualitative and quantitative evaluations demonstrate that the proposed method outperforms the state of the art in multi-domain image-to-image translation and that it surpasses predominant weakly-supervised localization methods in both disease detection and localization. Implementation is available at* https://github.com/jlianglab/Fixed-Point-GAN.

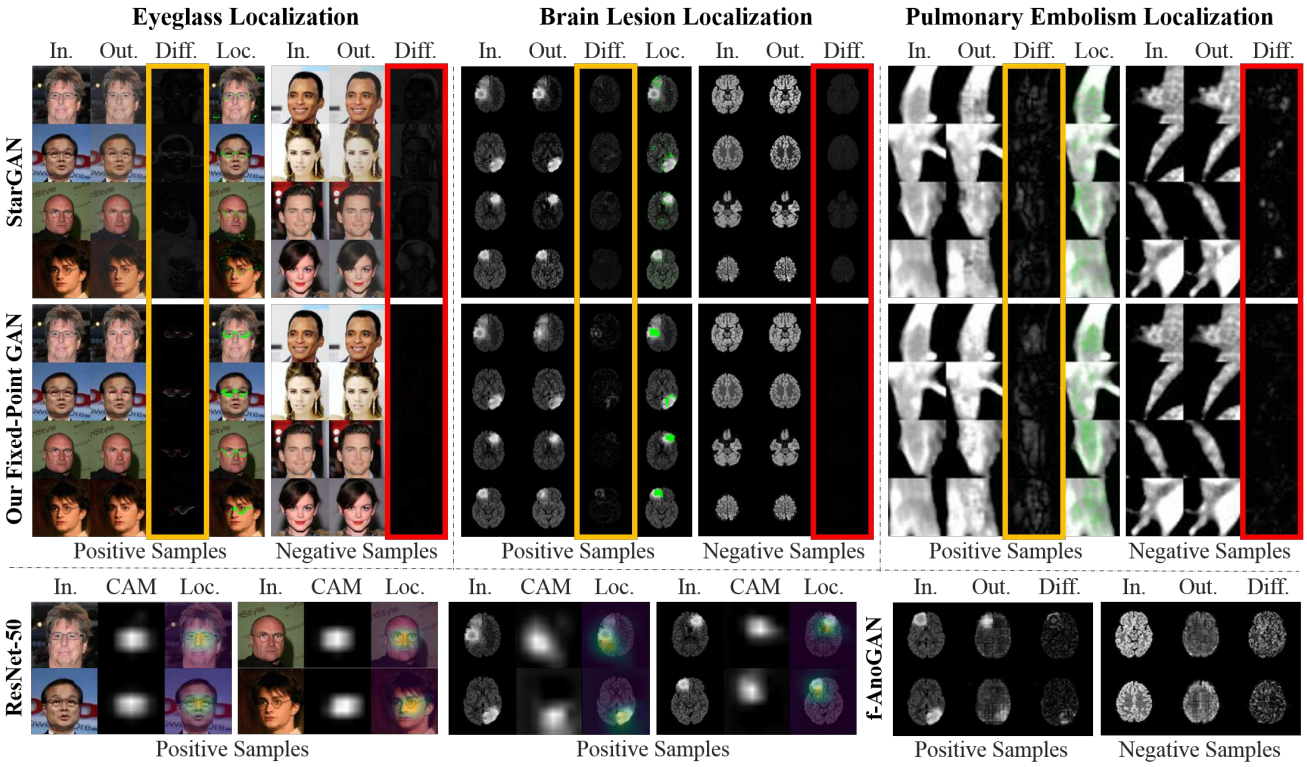


Fig. 2: [Better viewed on-line in color and zoomed in for details] Comparing Fixed-Point GAN with the state-of-the-art image-to-image translation [8], weakly-supervised localization [37], and anomaly detection [24] for detecting and localizing eyeglasses and diseases using only image-level annotation. Using disease detection as an example, our approach is to translate any image, diseased or healthy, into a healthy image, allowing diseased regions to be revealed by subtracting those two images. Through fixed-point translation learning, our Fixed-Point GAN aims to preserve healthy images during the translation, thereby few differences between the generated (healthy) images and the original (healthy) images are observed in the difference maps (columns framed in red). For diseased images, owing to the transformation learning from diseased images to healthy ones, disease locations are revealed in the difference maps (columns framed in yellow). For comparison, the localized diseased regions are superimposed on the original images (Loc. Columns), showing that Fixed-Point GAN is more precise than CAM [37] and f-AnoGAN [24] for localizing eyeglasses and diseases (bottom row; detailed in Sec. 4).

## 1. Introduction

Generative adversarial networks (GANs) [9] have proven to be powerful for image-to-image translation, such as changing the hair color, facial expression, and makeup of a person [8, 6], and converting MRI scans to CT scans for radiotherapy planning [34]. Now, the development and proliferation of GANs raises an interesting question: *Can GANs remove an object, if present, from an image while otherwise preserving the image content?* Specifically, can we train a GAN to remove eyeglasses from any image of a face with eyeglasses while keeping unchanged those without eyeglasses? Or, can a GAN "heal" a patient on his medical image virtually[1]? Such a task appears simple, but it actually demands the following four stringent requirements:

- **Req. 1:** The GAN must handle unpaired images. It may be too arduous to collect a perfect pair of photos of the same person with and without eyeglasses, and it would be too late to acquire a healthy image for a patient with an illness undergoing medical imaging.
- **Req. 2:** The GAN must require no source domain label when translating an image into a target domain (*i.e.*, source-domain-independent translation). For instance, a GAN trained for virtual healing aims to turn any image, with unknown health status, into a healthy one.
- **Req. 3:** The GAN must conduct an identity transformation for same-domain translation. For "virtual healing", the GAN should leave a healthy image intact, injecting neither artifacts nor new information into the image.
- **Req. 4:** The GAN must perform a minimal image transformation for cross-domain translation. Changes should be applied only to the image attributes directly relevant to the translation task, with no impact on unrelated attributes. For instance, removing eyeglasses should not affect the remainder of the image (*e.g.*, the hair, face color, and background), or removing diseases from a dis-

[1] Virtual healing (see Fig. 6 in Appendix) turns an image (diseased or healthy) into a healthy image, thereby subtracting the two images reveals diseased regions.

eased image should not impact the region of the image labeled as normal.

Currently, no single image-to-image translation method satisfies all aforementioned requirements. The conventional GANs for image-to-image translation [13], although successful, require paired images. CycleGAN [39] mitigates this limitation through cycle consistency, but it still requires two dedicated generators for each pair of image domains resulting a scalability issue due to a requirement for dedicated generators. CycleGAN also fails to support source-domain-independent translation: selecting the suitable generator requires labels for both the source and target domain. StarGAN [8] overcomes both limitations by learning one single generator for all domain pairs of interest. However, StarGAN has its own shortcomings. First, StarGAN tends to make unnecessary changes during cross-domain translation. As illustrated in Fig. 1, StarGAN tends to alter the face color, although the goal of domain translation is to change the gender, age, or hair color in images from the CelebFaces dataset [20]. Second, StarGAN fails to competently handle same-domain translation. Referring to examples framed with red boxes in Fig. 1, StarGAN needlessly adds a mustache to the face in Row 1, and unnecessarily alters the hair color in Rows 2–5, where only a simple identity transformation is desired. These shortcomings may be acceptable for image-to-image translation in natural images, but in sensitive domains, such as medical imaging, they may lead to dire consequences—unnecessary changes and artifacts introduction may result in misdiagnosis. Furthermore, overcoming the above limitations is essential for adapting GANs for object/disease detection, localization, segmentation—and removal.

Therefore, we propose a novel GAN. We call it Fixed-Point GAN for its new fixed-point[2] translation ability, which allows the GAN to identify a minimal subset of pixels for domain translation. To achieve this capability, we have devised a new training scheme to promote the fixed-point translation during training (Fig. 3-3) by (1) supervising same-domain translation through an additional conditional identity loss (Fig. 3-3B), and (2) regularizing cross-domain translation through revised adversarial (Fig. 3-3A), domain classification (Fig. 3-3A), and cycle consistency (Fig. 3-3C) loss. Owing to its fixed-point translation ability, Fixed-Point GAN performs a minimal transformation for cross-domain translation and strives for an identity transformation for same-domain translation. Consequently, Fixed-Point GAN not only achieves better image-to-image translation for natural images but also offers a novel framework for disease detection and localization with only image-level annotation. Our experiments demonstrate that Fixed-Point GAN significantly outperforms StarGAN over multiple datasets for the tasks of image-to-image translation and predominant anomaly detection and weakly-supervised localization methods for disease detection and localization. Formally, we make the following contributions:

1. We introduce a new concept: fixed-point translation, leading to a new GAN: Fixed-Point GAN.
2. We devise a new scheme to train fixed-point translation by supervising same-domain translation and regularizing cross-domain translation.
3. We show that Fixed-Point GAN outperforms the state-of-the-art method in image-to-image translation for both natural and medical images.
4. We derive a novel method for disease detection and localization using image-level annotation based on fixed-point translation learning.
5. We demonstrate that our disease detection and localization method based on Fixed-Point GAN is superior to not only its counterpart based on the state-of-the-art image-to-image translation method but also superior to predominant weakly-supervised localization and anomaly detection methods.

Our Fixed-Point GAN has the potential to exert important clinical impact on computer-aided diagnosis in medical imaging, because it requires only image-level annotation for training. Obtaining image-level annotation is far more feasible and practical than manual lesion-level annotation, as a large number of diseased and healthy images can be collected from the picture archiving and communication systems, and labeled at the image level by analyzing their radiological reports with NLP. With the availability of large databases of medical images and their corresponding radiological reports, we envision not only that Fixed-Point GAN will detect and localize diseases more accurately, but also that it may eventually be able to "cure"[1], thus segment diseases in the future.

## 2. Related Work

Fixed-Point GAN can be used for image-to-image translation as well as disease detection and localization with only image-level annotation. Hence, we first compare our Fixed-Point GAN with other image-to-image translation methods, and then explain how Fixed-Point GAN differs from the weakly-supervised lesion localization and anomaly detection methods suggested in medical imaging.

**Image-to-image translation:** The literature surrounding GANs [9] for image-to-image translation is extensive [13, 39, 14, 40, 19, 35, 8, 16]; therefore we limit our discussion to only the most relevant works. CycleGAN [39] has made a breakthrough in *unpaired* image-to-image translation via cycle consistency. Cycle consistency has proven to be effective in preserving object shapes in translated images, but

[2]Mathematically, $x$ is a fixed point of function $f(\cdot)$ if $f(x) = x$. We borrow the term to describe the pixels to be preserved when applying the GAN translation function.

it may not preserve other image attributes, such as color; therefore, when converting Monet's painting to photos (a cross-domain translation), Zhu *et al*. [39] imposes an extra identity loss to preserve the colors of input images. However, identity loss cannot be used for cross-domain translation in general, as it would limit the transformation power. For instance, it would make it impossible to translate black hair to blond hair. Therefore, unlike CycleGAN, we conditionally incorporate the identity loss only during fixed-point translation learning for same-domain translations. Moreover, during inference, CycleGAN requires that the source domain be provided, thereby violating our Req. 2 as discussed in Sec. 1 and rendering CycleGAN unsuitable for our purpose. StarGAN [8] empowers a single generator with the capability for *multi-domain* image-to-image translation, and does not require the source domain of the input image at inference time. However, StarGAN has its own shortcomings, which violate Reqs. 3 and 4 as discussed in Sec. 1. Our Fixed-Point GAN overcomes StarGAN's shortcomings, not only dramatically improving image-to-image translation but also opening the door to an innovative use of the generator as a disease detector and localizer (Figs.1-2).

**Weakly-supervised localization:** Our work is also closely related to weakly-supervised localization, which, in natural imaging, is commonly tackled by saliency map [27], global max pooling [22], and class activation map (CAM) based on global average pooling (GAP) [37]. In particular, the CAM technique has recently been the subject of further research, resulting in several extensions with improved localization power. Pinheiro and Collobert [23] replaced the original GAP with a log-sum-exponential pooling layer, while other works [28, 36] aim to force the CAM to discover the complementary parts rather than just the most discriminative parts of the objects. Selvaraju et al. [25] proposed GradCAM where the weights used to generate the CAM come from gradient backpropagation; that is, the weights depend on the input image as opposed to the fixed pretrained weights used in the original CAM.

Despite the extensive literature in natural imaging, weakly supervised localization in medical imaging has taken off only recently. Wang et al. [33] used the CAM technique for the first time for lesion localization in chest X-rays. The following research works, however, either combined the original CAM with extra information (*e.g*., limited fine-grained annotation [17, 26, 3] and disease severity-level [32]), or slightly extended the original CAM with no significant localization gain. Noteworthy, as evidenced by [5], the adoption of more advanced versions of the CAM such as the complementary-discovery algorithm [28, 36] has not proved promising for weakly-supervised lesion localization in medical imaging. Different from the previous works, Baumgartner *et al*. [4] propose VA-GAN to learn the difference between a healthy brain and the one affected by Alzheimer's disease. Although unpaired, VA-GAN requires that all images be registered; otherwise, it fails to preserve the normal brain structures (see the appendix for illustrations). Furthermore, VA-GAN requires the source-domain label at inference time (input image being healthy or diseased), thus violating our Req. 2 as listed in Sec. 1. Therefore, the vanilla CAM remains as a strong performance baseline for weakly-supervised lesion localization in medical imaging.

To our knowledge, we are among the first to develop GANs based on image-to-image translation for disease detection and localization with image-level annotation only. Both qualitative and quantitative results suggest that our image-translation-based approach provides more precise localization than the CAM-based method [37].

**Anomaly detection:** Our work may seem related to anomaly detection [7, 24, 1] where the task is to detect *rare* diseases by learning from only *healthy* images. Chen *et al*. [7] 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 diseased image. The method suggested by Schlegl *et al*. [24] learns a generative model of healthy training data through a GAN, which receives a random latent vector as input and then attempts to distinguish between real and generated fake healthy images. They further propose a fast mapping that can identify anomalies of the diseased images by projecting the diseased data into the GAN's latent space. Similar to [24], Alex *et al*. [1] 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 discriminator of the trained GAN. An anomaly map is then constructed by putting together the anomaly scores by the discriminator.

However, Fixed-Point GAN is different from anomaly detectors in both training and functionality. Trained using only the healthy images, anomaly detectors cannot distinguish between different types of anomalies, as they treat all anomalies as "a single category". In contrast, our Fixed-Point GAN can take advantage of anomaly labels, if available, enabling both localization and recognition of all anomalies. Nevertheless, for a comprehensive analysis, we have compared Fixed-Point GAN against [24] and [1].

## 3. Method

In the following, we present a high-level overview of Fixed-Point GAN, followed by a detailed mathematical description of each individual loss function.

Like StarGAN, our discriminator is trained to classify an image as real/fake and its associated domain (Fig. 3-1). Using our new training scheme, the generator learns both cross- and same-domain translation, which differs from

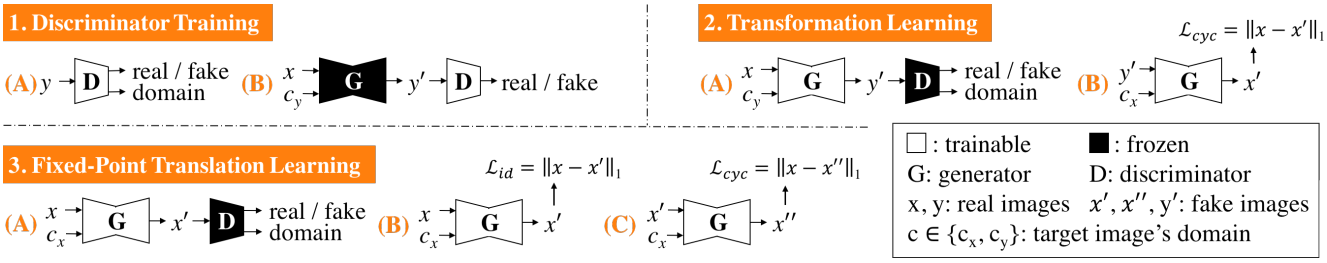


Fig. 3: Fixed-Point GAN training scheme. Similar to StarGAN, our discriminator learns to distinguish real/fake images and classify the domains of input images (1A–B). However, unlike StarGAN, our generator learns to perform not only cross-domain translations via transformation learning (2A–B), but also same-domain translations via fixed-point translation learning (3A–C), which is essential for mitigating the limitations of StarGAN (Fig. 1) and realizing disease detection and localization using only image-level annotation (Fig. 2).

StarGAN, wherein the generator only learns the former. Mathematically, for any input $x$ from domain $c_x$ and target domain $c_y$, the StarGAN generator learns to perform cross-domain translation ($c_x \neq c_y$), $G(x, c_y) \rightarrow y'$, where $y'$ is the image in domain $c_y$. Since $c_y$ is selected randomly during training of StarGAN, there is a slender chance that $c_y$ and $c_x$ turn out identical, but StarGAN is not designed to learn same-domain translation explicitly. The Fixed-Point GAN generator, in addition to learning the cross-domain translation, learns to perform the same-domain translation as $G(x, c_x) \rightarrow x'$.

Our new fixed-point translation learning (Fig. 3-3) not only enables same-domain translation but also regularizes cross-domain translation (Fig. 3-2) by encouraging the generator to find a minimal transformation function, thereby penalizing changes unrelated to the present domain translation task. Trained for only cross-domain image translation, StarGAN cannot benefit from such regularization, resulting in many artifacts as illustrated in Fig. 1. Consequently, our new training scheme offers three advantages: (1) reinforced same-domain translation, (2) regularized cross-domain translation, and (3) source-domain-independent translation. To realize these advantages, we define the loss functions of Fixed-Point GAN as follows:

**Adversarial Loss.** In the proposed method, the generator learns the cross- and same-domain translations. To ensure the generated images appear realistic in both scenarios, the adversarial loss is revised as follows and the modification is highlighted in Tab. 1:

$$\mathcal{L}_{adv} = \sum_{c \in \{c_x, c_y\}} \mathbb{E}_{x,c}[\log\left(1 - D_{real/fake}(G(x, c))\right)] + \mathbb{E}_x[\log D_{real/fake}(x)] \quad (1)$$

**Domain Classification Loss.** The adversarial loss ensures the generated images appear realistic, but it cannot guarantee domain correctness. As a result, the discriminator is trained with an additional domain classification loss, which forces the generated images to be of the correct domain. The domain classification loss for the discriminator is identical to that of StarGAN,

$$\mathcal{L}^r_{domain} = \mathbb{E}_{x,c_x}[-\log D_{domain}(c_x|x)] \quad (2)$$

but we have updated the domain classification loss for the generator to account for both same- and cross-domain translations, ensuring that the generated image is from the correct domain in both scenarios:

$$\mathcal{L}^f_{domain} = \sum_{c \in \{c_x, c_y\}} \mathbb{E}_{x,c}[-\log D_{domain}(c|G(x, c))] \quad (3)$$

**Cycle Consistency Loss.** Optimizing the generator, for unpaired images, with only the adversarial loss has multiple possible, but random, solutions. The additional *cycle consistency loss* (Eq. 4) helps the generator to learn a transformation that can preserve enough input information, such that the generated image can be translated back to original domain. Our modified *cycle consistency loss* ensures that both cross- and same-domain translations are cycle consistent.

$$\mathcal{L}_{cyc} = \mathbb{E}_{x,c_x,c_y}[||G(G(x, c_y), c_x) - x||_1] + \mathbb{E}_{x,c_x}[||G(G(x, c_x), c_x) - x||_1] \quad (4)$$

**Conditional Identity Loss.** During training, StarGAN [8] focuses on translating the input image to different target domains. This strategy cannot penalize the generator when it changes aspects of the input that are irrelevant to match target domains (Fig. 1). In addition to learning a translation to different domains, we force the generator, using the *conditional identity loss* (Eq. 5), to preserve the domain identity while translating the image to the source domain. This also helps the generator learn a minimal transformation for translating the input image to the target domain.

$$\mathcal{L}_{id} = \begin{cases} 0, & c = c_y \\ \mathbb{E}_{x,c}[||G(x, c) - x||_1], & c = c_x \end{cases} \quad (5)$$

**Full Objective.** Combining all losses, the final full objective function for the discriminator and generator can be described by Eq. 6 and Eq. 7, respectively.

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{domain}\mathcal{L}^r_{domain} \quad (6)$$

| | Loss | Definition |
|---|---|---|
| Eq. 1 | $\mathcal{L}_{adv}$ | $= \sum_{c\in\{c_x,c_y\}} \mathbb{E}_{x,c}\left[\log\left(1 - D_{r/f}(G(x,c))\right)\right] + \mathbb{E}_x[\log D_{r/f}(x)]$ |
| Eq. 2 | $\mathcal{L}^r_{domain}$ | $= \mathbb{E}_{x,c_x}[\log D_{domain}(c_x\|x)]$ |
| Eq. 3 | $\mathcal{L}^f_{domain}$ | $= \sum_{c\in\{c_x,c_y\}} \mathbb{E}_{x,c}[-\log D_{domain}(c\|G(x,c))]$ |
| Eq. 4 | $\mathcal{L}_{cyc}$ | $= \sum_{c\in\{c_x,c_y\}} \mathbb{E}_{x,c_x,c}[\|\|G(G(x,c),c_x) - x\|\|_1]$ |
| Eq. 5 | $\mathcal{L}_{id}$ | $= \mathbb{E}_{x,c}\,[\|\|G(x,c) - x\|\|_1]$ **if** $c = c_x$**; 0 otherwise** |
| Eq. 6 | $\mathcal{L}_D$ | $= -\mathcal{L}_{adv} + \lambda_{domain}\mathcal{L}^r_{domain}$ |
| Eq. 7 | $\mathcal{L}_G$ | $= \mathcal{L}_{adv} + \lambda_{domain}\mathcal{L}^f_{domain} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{id}\mathcal{L}_{id}$ |

Tab. 1: Loss functions in Fixed-Point GAN. Terms inherited from StarGAN are in black, while highlighted in blue are our modifications to mitigate StarGAN's limitations (Fig. 1).

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{domain}\mathcal{L}^f_{domain} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{id}\mathcal{L}_{id} \quad (7)$$

where $\lambda_{domain}$, $\lambda_{cyc}$, and $\lambda_{id}$ determine the relative importance of the *domain classification loss*, *cycle consistency loss*, and *conditional identity loss*, respectively. Tab. 1 summarizes the loss functions of Fixed-Point GAN.

# 4. Applications

## 4.1. Multi-Domain Image-to-Image Translation

**Dataset.** To compare the proposed Fixed-Point GAN with StarGAN [8] (the current state of the art), we use the CelebFaces Attributes (CelebA) dataset [20]. This dataset is composed of a total of 202,599 facial images of various celebrities, each with 40 different attributes. Following StarGAN's public implementation [8], we adopt 5 domains (`black hair`, `blond hair`, `brown hair`, `male`, and `young`) for our experiments and pre-process the images by cropping the original 178×218 images into 178×178 and then re-scaling to 128×128. We use a random subset of 2,000 samples for testing and the remainder for training.

**Method and Evaluation.** We evaluate the cross-domain image translation quantitatively by classification accuracy and qualitatively by changing one attribute (*e.g.* hair color, gender, or age) at a time from the source domain. This stepwise evaluation facilitates tracking changes to image content. We also evaluate the same-domain image translation both qualitatively and quantitatively by measuring image-level $L_1$ distance between the input and translated images.

**Results.** Fig. 1 presents a qualitative comparison between StarGAN and Fixed-Point GAN for multi-domain image-to-image translation. For the cross-domain image translation, StarGAN tends to make unnecessary changes, such as altering the face color when the goal of translation is to change the gender, age, or hair color (Rows 2–5 in Fig. 1). Fixed-Point GAN, however, preserves the face color while successfully translating the images to the target domains. Furthermore, Fixed-Point GAN preserves the image background (marked with a blue arrow in Row 5 of Fig. 1), but StarGAN fails to do so. This capability of Fixed-Point GAN is further supported by our quantitative results in Tab. 2.

| Real Images (Acc.) | Our Fixed-Point GAN | StarGAN |
|---|---|---|
| 94.5% | 92.31% | 90.82% |

Tab. 2: Comparison between the quality of images generated by StarGAN and our method. For this purpose, we have trained a classifier on all 40 attributes of CelebA dataset, which achieves 94.5% accuracy on real images, meaning that the generated images should also have the same classification accuracy to look as realistic as the real images. As seen, the quality of generated images by Fixed-Point GAN is closer to real images, underlining the necessity and effectiveness of fixed-point translation learning in cross-domain translation.

| Autoencoder | Our Fixed-Point GAN | StarGAN |
|---|---|---|
| 0.11 ± 0.09 | 0.36 ± 0.35 | 2.40 ± 1.24 |

Tab. 3: Image-level $L_1$ distance comparison for same-domain translation. Fixed-Point GAN achieves significantly lower same-domain translation error than StarGAN, approximating the lower bound error that can be achieved by a stand-alone autoencoder.

The superiority of Fixed-Point GAN over StarGAN is even more striking for the same-domain image translation. As shown in Fig. 1, Fixed-Point GAN effectively keeps the image content intact (images outlined in green) while StarGAN undesirably changes the image content (images outlined in red). For instance, the input image in the fourth row of Fig. 1 is from the domains of blond hair, female, and young. The same domain translation with StarGAN results in an image in which the hair and face colors are significantly altered. Although this color is closer to the average blond hair color in the dataset, it is far from that in the input image. Fixed-Point GAN, with fixed-point translation ability, handles this problem properly. Further qualitative comparisons between StarGAN and Fixed-Point GAN are provided in the appendix.

Tab. 3 presents a quantitative comparison between StarGAN and Fixed-Point GAN for the task of same-domain image translation. We use the image-level $L_1$ distance between the input and generated images as the performance metric. To gain additional insights into the comparison, we have included a dedicated autoencoder model that has the same architecture as the generator used in StarGAN and Fixed-Point GAN. As seen, the dedicated autoencoder has an image-level $L_1$ reconstruction error of 0.11±0.09, which can be regarded as a technical lower bound for the reconstruction error. Fixed-Point GAN dramatically reduces the reconstruction error of StarGAN from 2.40±1.24 to 0.36±0.35. Our quantitative comparisons are commensurate with the qualitative results shown in Fig. 1.

## 4.2. Brain Lesion Detection and Localization with Image-Level Annotation

**Dataset.** We extend Fixed-Point GAN from an image-to-image translation method to a weakly supervised brain le-

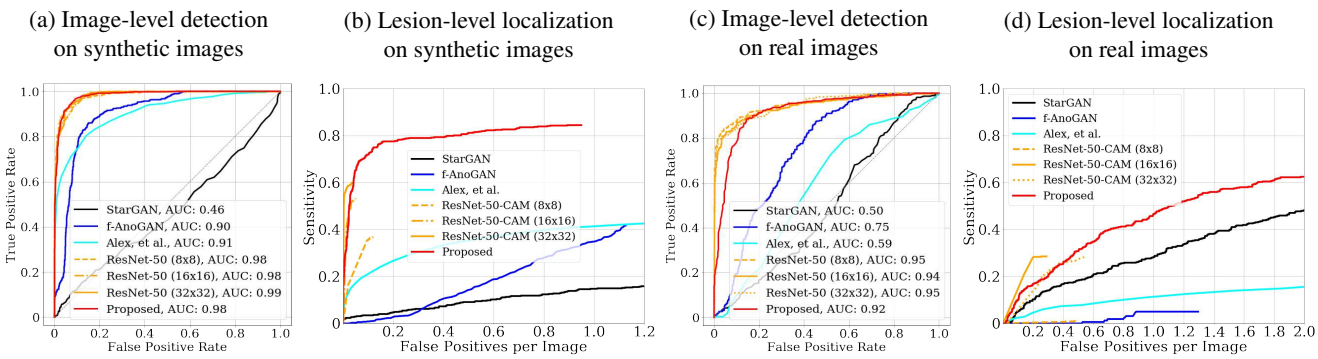


Fig. 4: Comparing Fixed-Point GAN with StarGAN, f-AnoGAN, GAN-based brain lesion detection method by Alex, *et al*. [1], and ResNet-50 on BRATS 2013. ROCs for image-level detection and FROCs for lesion-level localization on synthetic brain images are provided in (a), (b) respectively and on real brain images in (c), (d) respectively.

sion detection and localization method, which requires only image-level annotation. As a proof of concept, we use the BRATS 2013 dataset [21, 15]. BRATS 2013 consists of synthetic and real images. We randomly split the synthetic and real images at the patient-level into 40/10 and 24/6 for training/testing, respectively. More details about the dataset selection are provided in the appendix.

**Method and Evaluation.** For training we use only image-level annotation (healthy/diseased). Fixed-Point GAN is trained for the cross-domain translation (diseased images to healthy images and vice versa) as well as the same-domain translation using the proposed method. At inference time, we focus on translating any images into the healthy domain. The desired GAN behaviour is to translate diseased images to healthy ones while keeping healthy images intact. Having translated the images into the healthy domain, we then detect the presence and location of a lesion in the difference image by subtracting the translated healthy image from the input image. We refer the resultant image as *difference map*.

We evaluate the difference map at two different levels: (1) image-level disease detection and (2) lesion-level localization. For image-level detection, we take the maximum value across all pixels in the difference map as the *detection score*. We then use receiver operating characteristics (ROC) analysis for performance evaluation. For the lesion-level localization task, we first binarize the difference maps using color quantization followed by a connected component analysis. Each connected component with an area larger than 10 pixels is considered as a lesion candidate. A lesion is considered "detected" if the centroid of at least a lesion candidate falls inside the lesion ground truth.

We evaluate Fixed-Point GAN in comparison with StarGAN [8], CAM [37], f-AnoGAN [24], GAN-based brain lesion detection method proposed by Alex, *et al*. [1]. Comparison with StarGAN allows us to study the effect of the proposed fixed-point translation learning. We choose CAM for comparison because it covers an array of weakly-supervised localization works in medical imaging [33, 32, 12], and as discussed in Sec. 2, it is arguably a strong performance baseline for comparison. We train a standard ResNet-50 classifier [11] and compute CAM following [37] for localization, referring as ResNet-50-CAM in the rest of this paper. To get higher resolution CAMs, we truncate ResNet-50 at three levels and report localization performance in $8\times8$, $16\times16$, and $32\times32$ feature maps. Although [24] and [1] stand as state of the art for anomaly detection, we select them for more comparison since they also fulfill the task requirements. We use the official implementation of [24].

**Results.** Fig. 4a compares the ROC curves of Fixed-Point GAN and the competing methods for image-level lesion detection using synthetic MRI images. In terms of the area under the curve (AUC), Fixed-Point GAN achieves comparable performance with ResNet-50 classifier, but substantially outperforms StarGAN, f-AnoGAN, and Alex, *et al*. Note that, for f-AnoGAN, we use the average activation of difference maps as the detection score, because we find it more effective than using the maximum activation of difference maps and also more effective than the anomaly scores proposed in the original work.

Fig. 4b shows the Free-Response ROC (FROC) analysis for synthetic MR images. Our Fixed-Point GAN achieves a sensitivity of 84.5% at 1 false positive per image, outperforming StarGAN, f-AnoGAN, and Alex, *et al*. with the sensitivity levels of 13.6%, 34.6%, 41.3% at the same level of false positive. The ResNet-50-CAM at 32x32 resolution achieves the best sensitivity level of 60% at 0.037 false positives per image. Furthermore, we compare ResNet-50-CAM with Fixed-Point GAN using mean IoU (intersection over union) score, obtaining mean IoU of $0.2609\pm0.1283$ and $0.3483\pm0.2420$, respectively. Similarly, ROC and FROC analysis on real MRI images are provided in Fig. 4c and Fig. 4d, respectively, showing that our method is outperformed at the low false positive range, but achieves a significantly higher sensitivity overall. Qualitative comparisons between StarGAN, Fixed-Point GAN, CAM, and f-AnoGAN for brain lesion detection and localization are pro-

vided in Fig. 2. More qualitative comparisons are available in the appendix.

### 4.3. Pulmonary Embolism Detection and Localization with Image-Level Annotation

**Dataset.** Pulmonary embolism (PE) is a blood clot that travels from a lower extremity source to the lung, where it causes blockage of the pulmonary arteries. It is a major national health problem, but computer-aided PE detection and localization can improve diagnostic capabilities of radiologists for the detection of this disorder, leading to earlier and effective therapy for this potentially deadly disorder. We utilize a database consisting of 121 computed tomography pulmonary angiography (CTPA) scans with a total of 326 emboli. The dataset is pre-processed as suggested in [38, 31, 30], divided at the patient-level into a training set with 3,840 images, and a test set with 2,415 images. Further details are provided in the appendix.

**Method and Evaluation.** As with brain lesion detection and localization (Sec. 4.2), we use only image-level annotations during training. At inference time, we always remove PE from the input image (*i.e.* translating both PE and non-PE images into the non-PE domain) irrespective of whether PE is present or absent in the input image. We follow the same procedure described in Sec. 4.2 to generate the difference maps, detection scores, and ROC curves. Note that, since each PE image has an embolus in its center, an embolus is considered as "detected" if the corresponding PE image is correctly classified; otherwise, the embolus is considered "missed". As such, unlike Sec. 4.2, we do not pursue a connected component analysis for PE localization.

We compare our Fixed-Point GAN with StarGAN and ResNet-50. We have excluded GAN-based method [1] and f-AnoGAN from the quantitative comparisons because, despite our numerous attempts, the former encountered convergence issues and the latter produced poor detection and localization performance. Nevertheless, we have provided images generated by f-AnoGAN in appendix.

**Results.** Fig. 5a shows the ROC curves for image-level PE detection. Fixed-Point GAN achieves an AUC of 0.9668 while StarGAN and ResNet-50 achieve AUC scores of 0.8832 and 0.8879, respectively. Fig. 5b shows FROC curves for PE localization. Fixed-Point GAN achieves a sensitivity of 97.2% at 1 false positive per volume, outperforming StarGAN and ResNet-50 with sensitivity levels of of 88.9% and 80.6% at the same level of false positives per volume. The qualitative comparisons for PE removal between StarGAN and Fixed-Point GAN are given in Fig. 2.

### 4.4. Discussions

In Fig. 4, we show that StarGAN performs poorly for image-level brain lesion detection, because StarGAN is designed to perform general-purpose image translations, rather than an image translation suitable for the task of disease detection. Owing to our new training scheme, Fixed-Point GAN can achieve precise image-level detection.

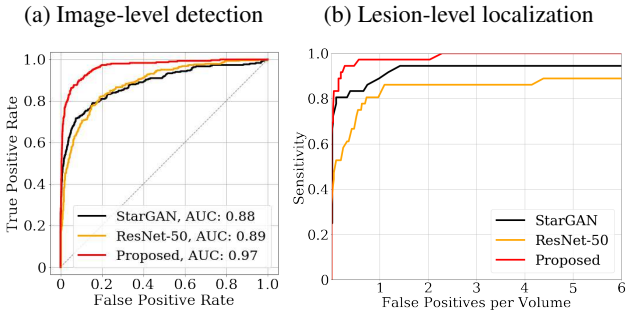


Fig. 5: Comparing Fixed-Point GAN with StarGAN, f-AnoGAN, and ResNet-50 on the PE dataset. (a) ROCs for image-level detection. (b) FROCs for lesion-level localization.

Comparing Fig. 4 and 5, we observe that StarGAN performs far better for PE than brain lesion detection. We believe this is because brain lesions can appear anywhere in the input images, whereas PE always appears in the center of the input images, resulting in a less challenging problem for StarGAN to solve. Nonetheless, Fixed-Point GAN outperforms StarGAN for PE detection, achieving an AUC score of 0.9668 compared to 0.8832 by StarGAN.

Referring to Fig. 2, we further observe that neither StarGAN nor Fixed-Point GAN can completely remove large objects, like sunglasses or brain lesions, from the images. Nevertheless, for image-level detection and lesion-level localization, it is sufficient to remove the objects partially, but precise lesion-level segmentation using an image-to-image translation network requires complete removal of the object. This challenge is the focus for our future work.

## 5. Conclusion

We have introduced a new concept called fixed-point translation, and developed a new GAN called Fixed-Point GAN. Our comprehensive evaluation demonstrates that our Fixed-Point GAN outperforms the state of the art in image-to-image translation and is significantly superior to predominant anomaly detection and weakly-supervised localization methods in both disease detection and localization with only image-level annotation. The superior performance of Fixed-Point GAN is attributed to our new training scheme, realized by supervising same-domain translation and regularizing cross-domain translation.

**Acknowledgments:** This research has been supported partially by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant, and partially by NIH under Award Number R01HL128785. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH. We thank Zuwei Guo for helping us with the implementation of a baseline method.

## References


[1] 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, page 101330G. International Society for Optics and Photonics, 2017. 4, 7, 8

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 21

[3] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac mr image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017. 4

[4] 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. 4, 19, 21

[5] Jinzheng Cai, Le Lu, Adam P Harrison, Xiaoshuang Shi, Pingjun Chen, and Lin Yang. Iterative attention mining for weakly supervised thoracic disease pattern localization in chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 589–598. Springer, 2018. 4

[6] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[7] Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018. 4

[8] 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. The implementation is publicly available at https://github.com/yunjey/StarGAN. 1, 2, 3, 4, 5, 6, 7

[9] 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*, pages 2672–2680, 2014. 2, 3

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017. 21

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[12] Sangheum Hwang and Hyo-Eun Kim. Self-transfer learning for weakly supervised lesion localization. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 239–246, Cham, 2016. Springer International Publishing. 7

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 3

[14] 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, 2017. 3

[15] Michael Kistler, Serena Bonaretti, Marcel Pfahrer, Roman Niklaus, and Philippe Büchler. The virtual skeleton database: An open access repository for biomedical research and collaboration. *J Med Internet Res*, 15(11):e245, Nov 2013. 7

[16] 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, page 4, 2017. 3

[17] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8290–8299, 2018. 4

[18] Jianming Liang and Jinbo Bi. Computer aided detection of pulmonary embolism with tobogganing and mutiple instance classification in ct pulmonary angiography. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 630–641. Springer, 2007. 15, 20

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

[20] 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*, pages 3730–3738, 2015. 3, 6

[21] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993, 2015. 7

[22] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015. 4

[23] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015. 4

[24] 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. The implementation is publicly available at https://github.com/tSchlegl/f-AnoGAN. 2, 4, 7, 19

[25] 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. 4

[26] Seung Yeon Shin, Soochahn Lee, Il Dong Yun, Sun Mi Kim, and Kyoung Mu Lee. Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. *IEEE transactions on medical imaging*, 2018. 4

[27] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 4

[28] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3544–3553. IEEE, 2017. 4

[29] Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 62–69. Springer, 2015. 20

[30] Nima Tajbakhsh, Jae Y Shin, Michael B Gotway, and Jianming Liang. Computer-aided detection and visualization of pulmonary embolism using a novel, compact, and informative image representation. *Medical Image Analysis*, page 101541, 2019. 8, 20

[31] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016. 8, 20

[32] Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald M Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In *International Workshop on Machine Learning in Medical Imaging*, pages 249–258. Springer, 2018. 4, 7

[33] 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*, pages 2097–2106, 2017. 4, 7

[34] Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Išgum. Deep mr to ct synthesis using unpaired data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 14–23. Springer, 2017. 2

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

[36] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 4

[37] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 2, 4, 7

[38] Zongwei Zhou, Jae Y Shin, Lei Zhang, Suryakanth R Gurudu, Michael B Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally. In *CVPR*, pages 4761–4772, 2017. 8, 20

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

[40] 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*, pages 465–476, 2017. 3