Appendix

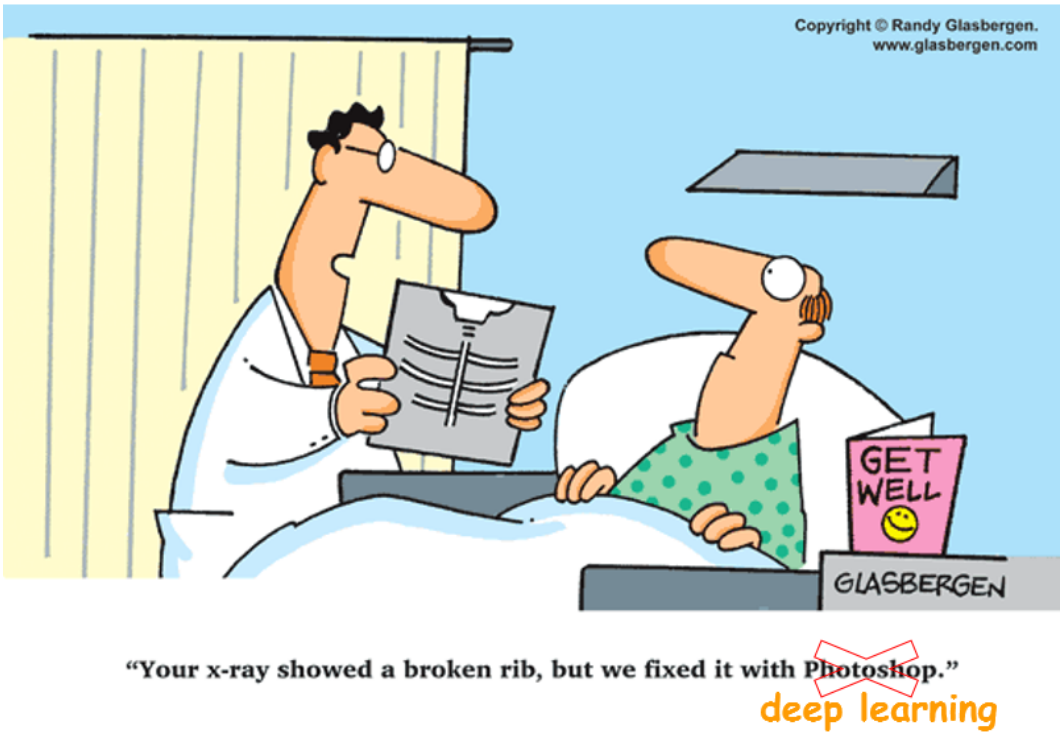

Fig. 6: Now, all humor aside, if a GAN can remove diseases *completely* from images, it will offer an *ideal* method for segmenting *all* diseases, an ambitious goal that has yet to be achieved. Nevertheless, our method is still of great clinical significance in computer-aided diagnosis for medical imaging, because it can be used for disease detection and localization by subtracting the generated healthy image from the original image to reveal diseased regions, that is, **detection and localization by removal**. More importantly, our Fixed-Point GAN is trained using *only* image-level annotation. It is much easier to obtain image-level annotation than lesion-level annotation, because a large number of diseased and healthy images can be collected from PACS (picture archiving and communication systems), and labeled at the image level by analyzing their radiological reports through NLP (natural language processing). With the availability of large *well-organized* databases of medical images and their corresponding radiological reports in the future, we envision that Fixed-Point GAN will be able to detect and localize diseases more accurately—and eventually to segment diseases—using only image-level annotation. [The cartoon was provided courtesy of Karen Glasbergen with permission for adaptation and modification]

All figures and images (including those in the main paper) better viewed on-line in color and magnified for details

Eyeglass Detection and Localization by Removal Using Only Image-Level Annotation of the CelebA Dataset

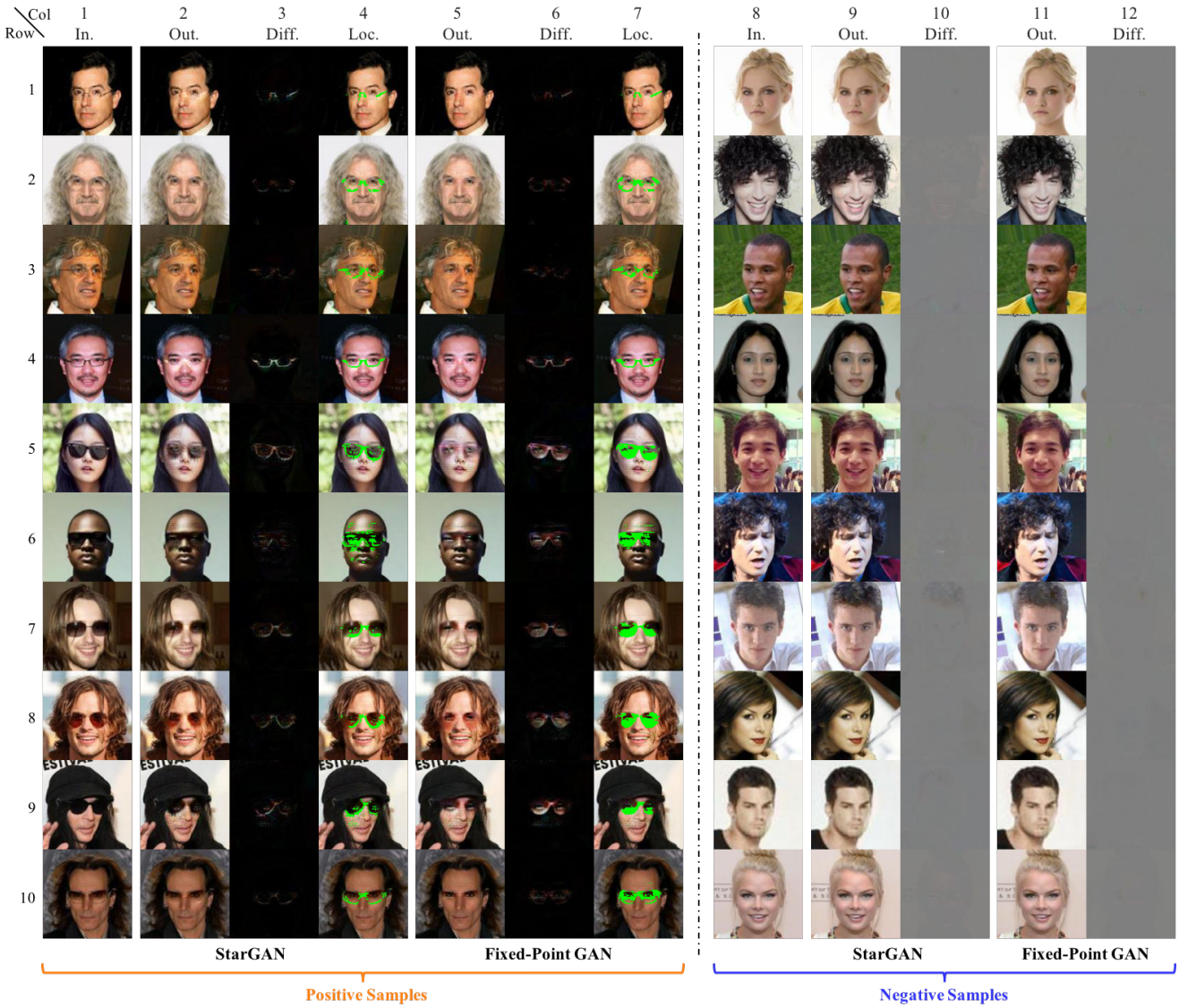

Fig. 7: [Continued from Fig. 2] Additional test results in eyeglass detection and localization by removal. The difference map (Column 3 for StarGan; Column 6 for Fixed-Point GAN) shows the absolute difference between the input (Column 1) and output (Column 2 for StarGAN; Column 5 for Fixed-Point GAN). Applying the k-means clustering algorithm on the difference map yields a localization map, which is then superimposed on the original image (Column 4 for StarGAN; Column 7 for Fixed-Point GAN), showing both StarGAN and Fixed-Point GAN attempt to remove eyeglasses. However, the former leaves noticeable white "inks" along eyeglass frames (Rows 1 and 4, Column 2), while our method better preserves the face color. Removing sunglasses (Rows 5–9) has proven to be challenging: both methods suffer from partial removal and artifacts. Nevertheless, Fixed-Point GAN tends to recover the face under the glasses and frames, but StarGAN only changes regions around the frames. More importantly, our method can "insert" eyes at proper positions, as revealed in the difference maps (Rows 5–9, Column 6), while StarGAN can hardly do so. To better visualize the subtle changes for negative samples (Column 8), instead of the absolute difference, we show the difference directly, where the gray color (*i.e.*, 0) means "no change". In this way, it can be observed more easily that StarGAN does some unnecessary small changes on hair (Rows 7 and 9, Column 10) and eyes (Rows 7 and 10, Column 10), while Fixed-Point GAN generates smooth gray images (*i.e.*, close to 0 everywhere; Column 12). Please note that the CelebA Dataset currently does not have ground truth on the location and segmentation of glasses; therefore, a quantitative performance evaluation of eyeglass localization cannot be conducted. However, our quantitative performance evaluations of brain lesion localization and pulmonary embolism localization are included in Sec. 4.

Multi-Domain Image-to-Image Translation

Fig. 8: [Continued from Fig. 1] More test results in multi-domain image-to-image translation on CelebA dataset. Visually, Fixed-Point GAN outperforms StarGAN: Fixed-Point GAN (Columns 7-11) better preserves the background (Rows 1, 3, 4, 6, 8, and 9), face color (Rows 2–7), and facial features (Rows 7 and 9), whereas StarGAN (Columns 2-6) makes unnecessary changes. Furthermore, for same-domain translation, StarGAN introduces noticeable artifacts (outlined in red), while Fixed-Point GAN can leave all the details intact (outlined in green). It is worthy noting that the hair color of the facial image in the last input row (*i.e.*, Row 9, Column 1) belongs to Domain `gray hair`, which is not included in the training phase. As can be seen, Fixed-Point GAN successfully translates the input image to target domains by changing the unseen hair color to desired colors and maintaining the original hair color (gray) in hair-color-unrelated translations (Row 9, Columns 10–11). However, StarGAN produces unnatural images with artifacts (Row 9, Columns 2–4) and inconsistent white hair colors (Row 9, Columns 5–6). This example shows that Fixed-Point GAN outperforms StarGAN in generalization.

Brain Lesion Detection and Localization by Removal Using Only Image-Level Annotation

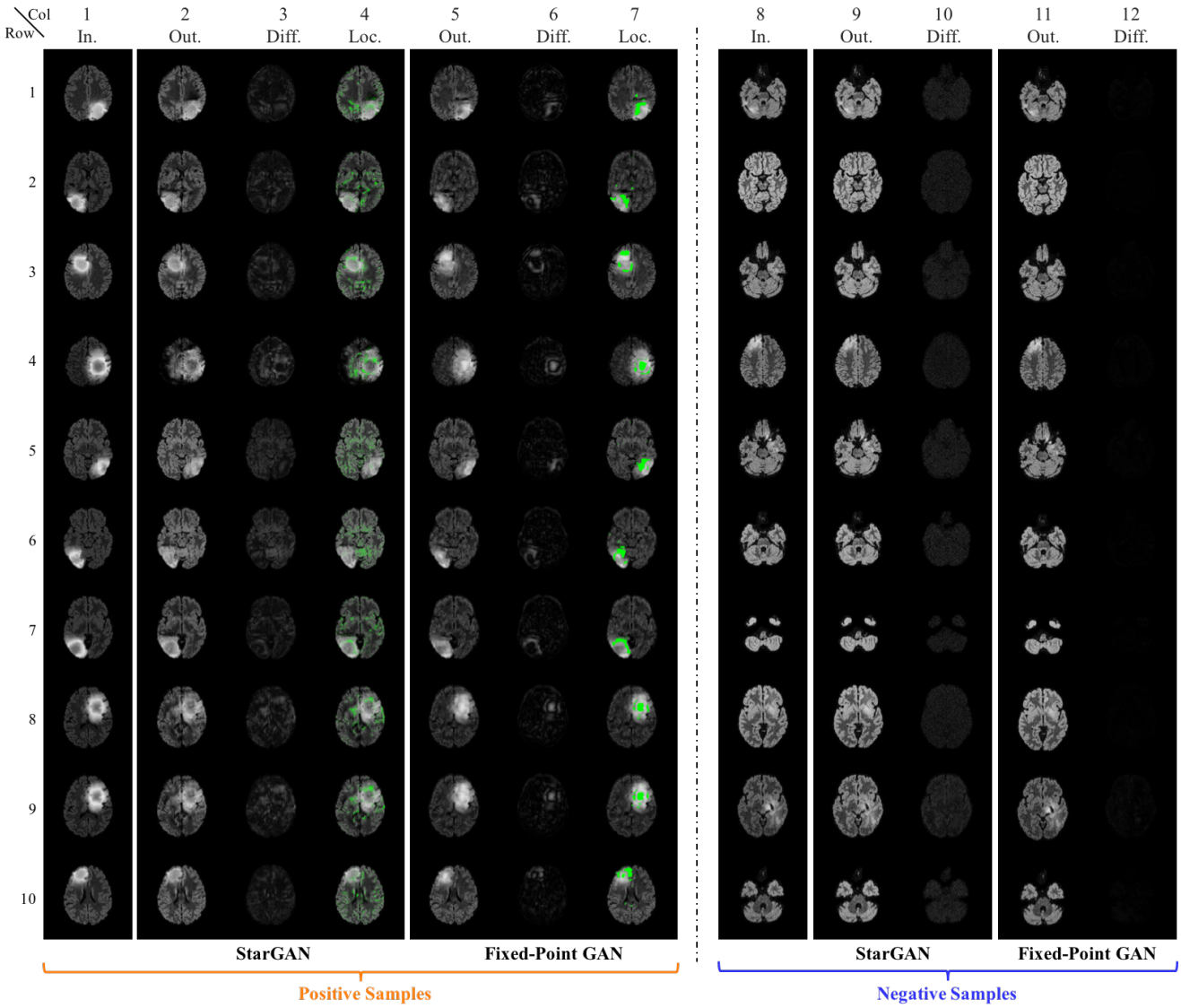

Fig. 9: [Continued from Fig. 2] Brain lesion detection and localization tested on additional positive samples (*i.e.* brain images with lesions; Column 1) and negative samples (*i.e.* brain images without lesions; Column 8). Fixed-Point GAN achieves superior detection performance, benefiting from the cleaner difference maps of negative samples (Column 12), while StarGAN highlights the brain regions in all cases, thereby rendering the difference maps of positive and negative samples indistinguishable (comparing Column 3 with Column 10). Although both methods fail to remove lesions completely, our method focuses on the lesion regions, and consequently, it produces higher localization accuracy. In contrast, the StarGAN localization map (Column 4) is very noisy and unsuitable for lesion localization. These comparisons demonstrate the superiority of Fixed-Point GAN in lesion detection and localization. For quantitative performance evaluations, please refer to Fig. 4 and Sec. 4.2.

Pulmonary Embolism Detection and Localization by Removal Using Only Image-Level Annotation

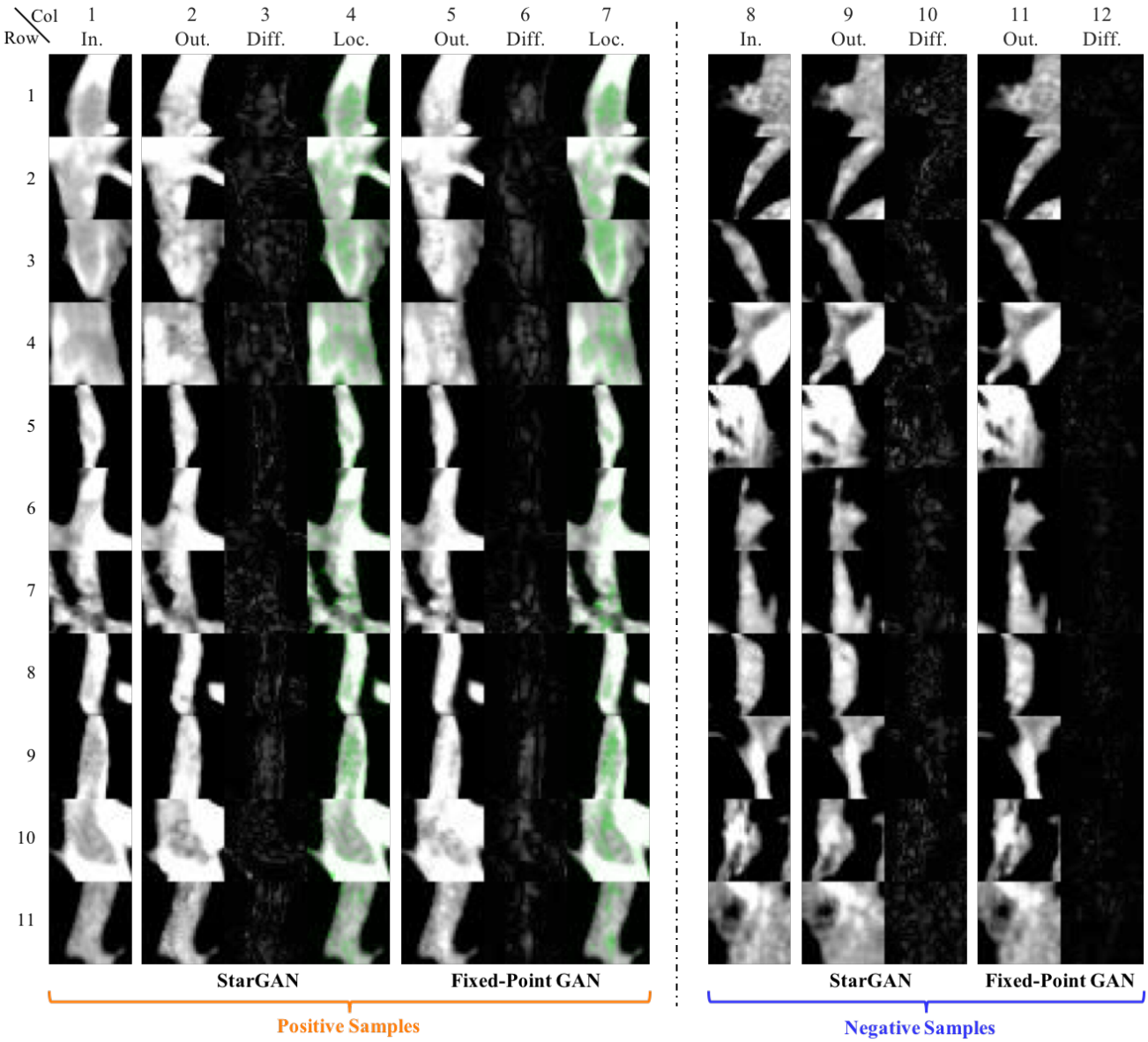

Fig. 10: [Continued from Fig. 2] Pulmonary Embolism (PE) detection and localization (longitudinal view) tested on additional positive samples (*i.e.* images with PEs; Column 1) and negative samples (*i.e.* images without PEs; Column 8). PE is a blood clot that creates blockage (appearing dark and centered within the image) in pulmonary arteries (which appear white). The current candidate generator (*e.g.* [18]) produces many false positive results (negative samples) during localization; therefore, our goal in this application is to reduce false positives through StarGAN and Fixed-Point GAN. Compared with StarGAN, the difference maps of negative samples from Fixed-Point GAN is clean and easy to be separated from the difference maps of positive samples, yielding better detection performance. For quantitative performance evaluations, please refer to Fig. 5 and Sec. 4.3.

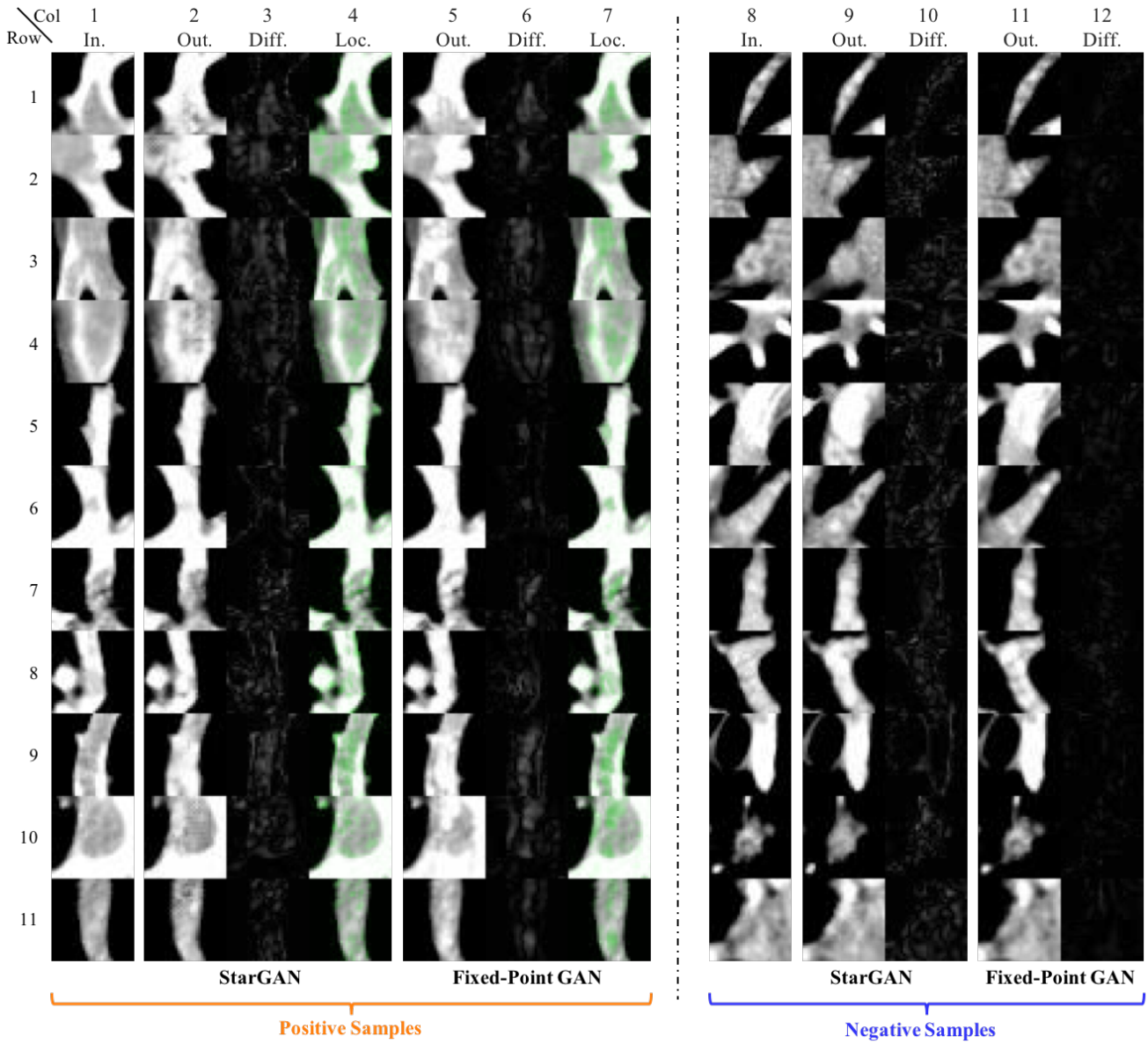

Fig. 11: Pulmonary Embolism (PE) detection and localization (longitudinal view). Notice the images are from the same candidates as Fig. 10 but the view direction is orthogonal to the angle used in Fig. 10.

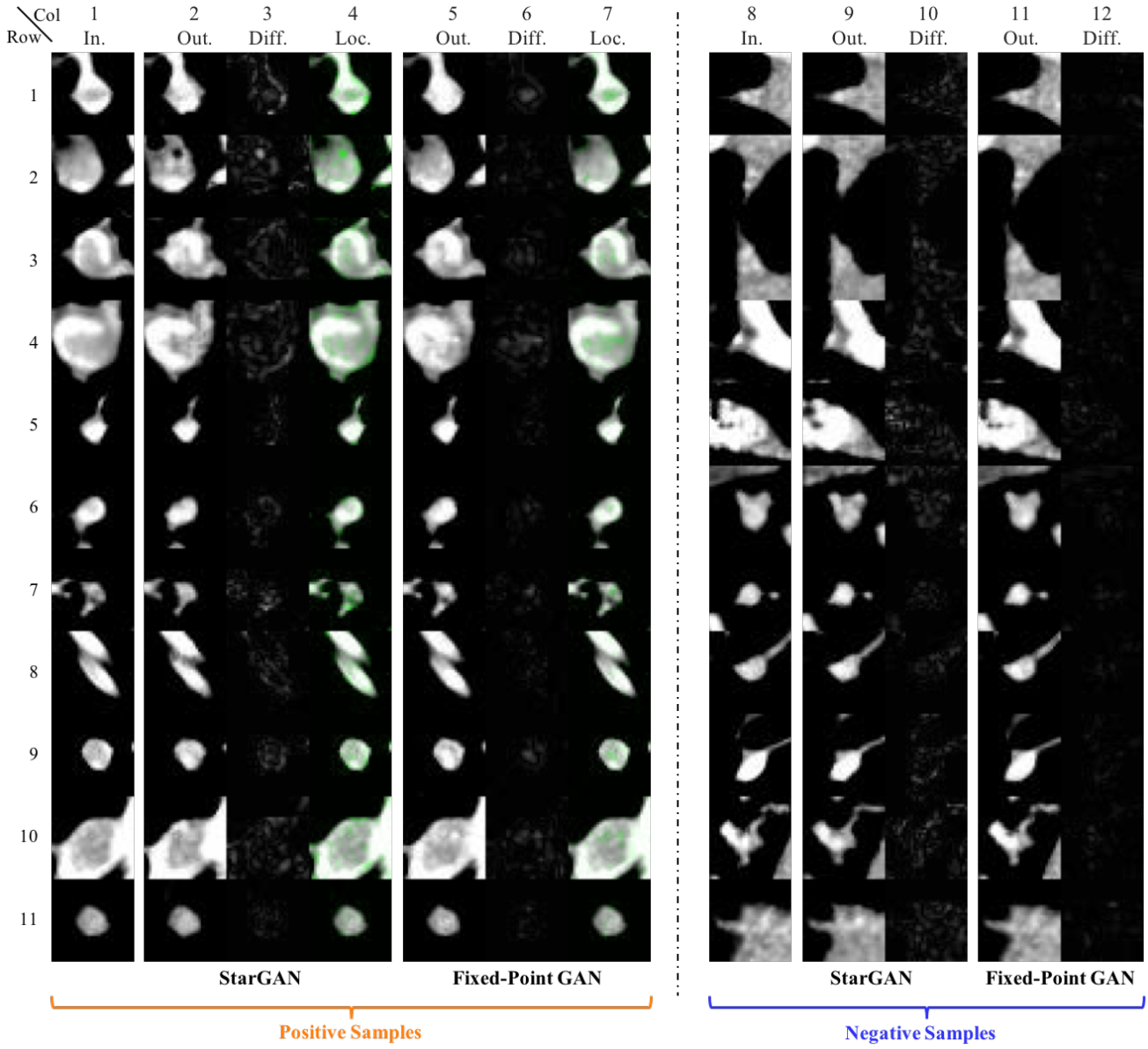

Fig. 12: Pulmonary Embolism (PE) detection and localization (cross-sectional view). Notice the images are from the same candidates as Fig. 10 but the orientation is cross-sectional.

Localization Using Class Activation Maps (CAMs)

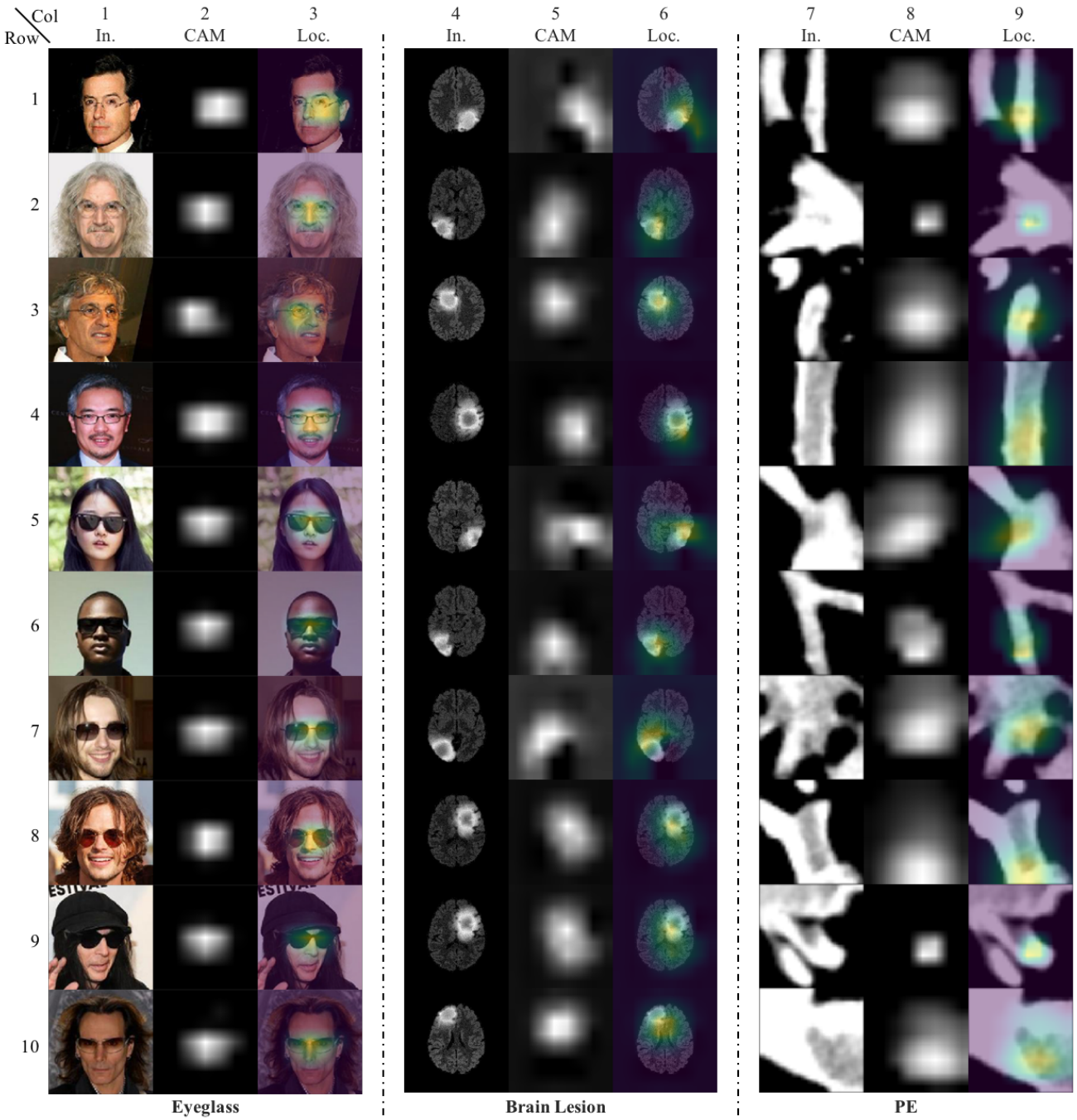

Fig. 13: [Continued from Fig. 2] Additional test results of localization using class activation maps (CAMs). CAMs for localizing glasses, brain lesion, and PE are obtained from ResNet-50 classifiers trained with corresponding datasets. Localization using CAMs is not as precise as Fixed-Point GAN, as discussed in Sec. 4.2.

Qualitative Results of f-AnoGAN and VA-GAN for Brain Lesion Detection

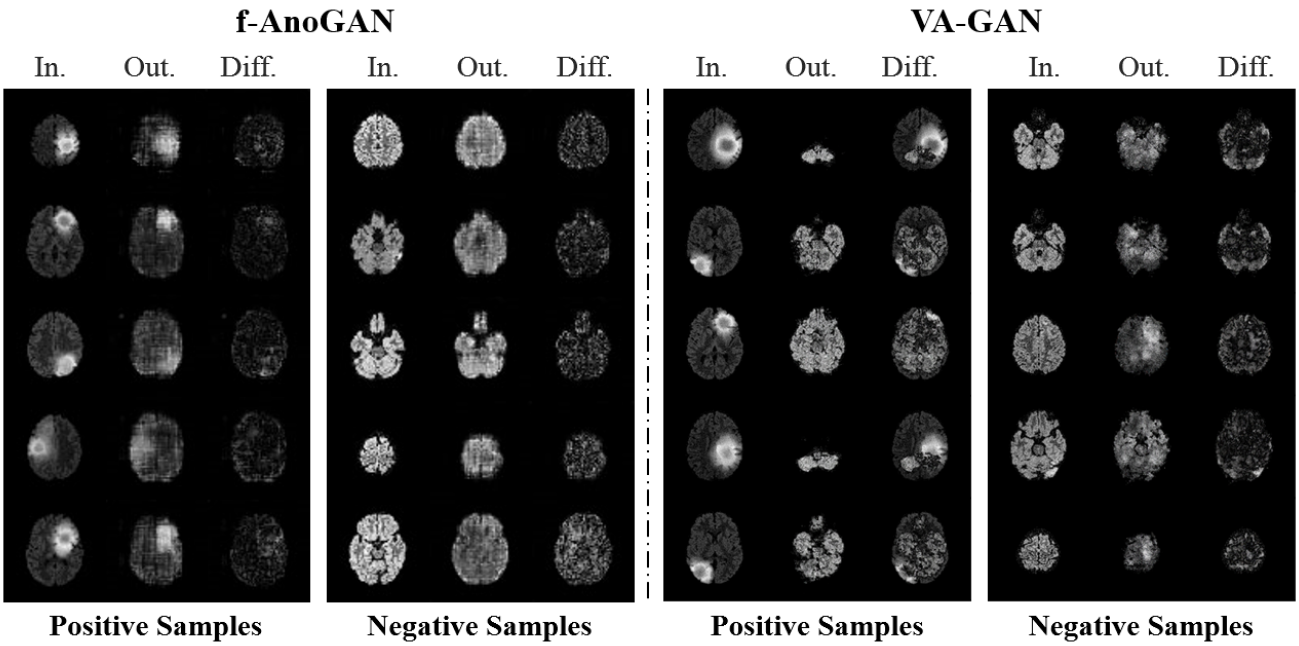

Fig. 14: [Continued from Fig. 2] In spite of learning healthy images only, f-AnoGAN [24] performs competitively on image-level detection task (see Fig. 4a), however, noisy difference maps impede its localization power (see Fig. 4b). On the other hand, VA-GAN [4] fails to preserve anatomical structures when trained with unpaired images, thus violates our Req. 2 and renders unsuitable for our purpose.

Qualitative Results of f-AnoGAN on the PE Dataset

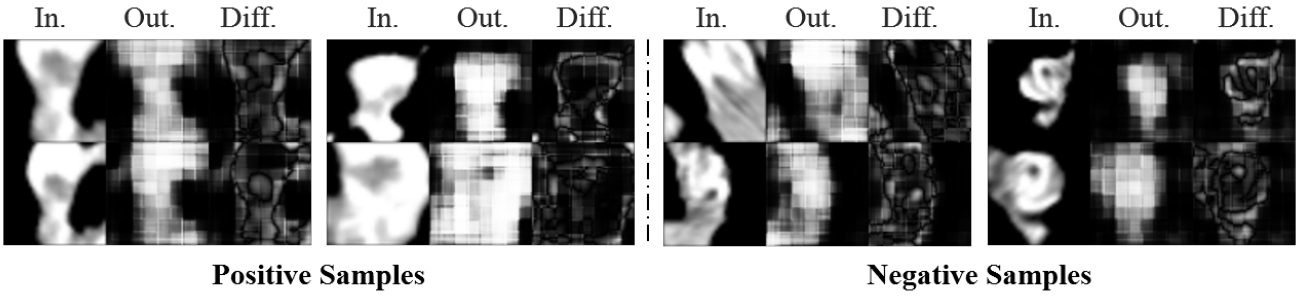

Fig. 15: As discussed in Sec. 4.3, f-AnoGAN [24] fails to produce good quality images for the PE dataset, resulting in noisy difference maps. Hence, performs miserably on detection and localization tasks for the PE dataset.

Dataset Processing Details

Brain Lesion Detection and Localization with Image-Level Annotation: BRATS 2013 consists of synthetic and real images, where each of them is further divided into high-grade gliomas (HG) and low-grade gliomas (LG). There are 25 patients with both synthetic HG and LG images and 20 patients with real HG and 10 patients with real LG images. For each patient, `FLAIR`, `T1`, `T2`, and `post-Gadolinium T1` magnetic resonance (MR) image sequences are available. To ease the analysis, we keep the input features consistent by using only one MR imaging sequence (FLAIR) for all patients in both the HG and LG categories, resulting in a total of 9,050 synthetic MR slices and 5,633 real MR slices. We further pre-process the dataset by removing all slices that are either blank or have very little brain information. Finally, we randomly select 40 patients with 5,827 slices for training and 10 patients with 1,461 slices for testing from synthetic MRI images. For the experiments on real MRI images, we randomly select 24 patients with 3,044 slices for training and 6 patients with 418 slices for testing. During training, we set aside one batch of the random samples from the training dataset for validation. We pad the slices to 300×300 and then center-crop to 256×256, ensuring that the brain regions appear in the center of the images. Each pixel in the dataset is assigned one of the five possible labels: 1 for non-brain, non-tumor, necrosis, cyst, hemorrhage; 2 for surrounding edema; 3 for non-enhancing tumor; 4 for enhancing tumor core; and 0 for everything else. We assign an MR slice to the healthy domain if all contained pixels are labeled as 0; otherwise, the MR slice is assigned to the diseased domain.

Pulmonary Embolism Detection and Localiza-tion with Image-Level Annotation: 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]. A candidate generator [18] is first applied to generate a set of PE candidates, and then by comparing against the ground truth, the PE candidates are labeled as PE or non-PE. Finally, a 2D patch of size 15×15mm is extracted around each PE candidate according to a vessel-aligned image representation [29]. As a result, PE appears at the center of the PE images. The extracted images are rescaled to 128×128. The dataset is divided at the patient-level into a training set with 434 PE images (199 unique PEs) and 3,406 non-PE images, and a test set with 253 PE images (127 unique PEs) and 2,162 non-PE images. To enrich the training set, rotation-based data augmentation is applied for both PE and non-PE images.

Implementation Details

	Image-Level Detection (AUC)				Lesion-Level Loc. Sensitivity at 1 False Positive		
Dataset	**StarGAN**	**w/ Delta**	**w/ Fixed-Point Translation**	**w/ Both**	**StarGAN**	**w/ Fixed-Point Translation**	**w/ Both**
BRATS	0.4611	0.5246	0.9980	**0.9831**	13.6%	81.2%	**84.5%**
PE	0.8832	0.8603	0.9216	**0.9668**	88.9%	94.4%	**97.2%**

Tab. 4: Ablation study of the generator's configuration on brain lesion (BRATS 2013) and pulmonary embolism (PE) detection. Selected combinations are in bold. The columns "w/Delta", "w/Fixed-Point Translation", and "w/Both" mean StarGAN trained with only delta map, only fixed-point translation learning, and both of them combined, respectively. The empirical results show that the performance gain is largely due to fixed-point translation learning—the contribution by the delta map is minor and application-dependent.

We have revised adversarial loss (Eq. 1) based on the Wasserstein GAN [2] objective by adding a gradient penalty [10] to stabilize the training, which is defined as

$$\mathcal{L}_{adv} = \mathbb{E}_x[D_{real/fake}(x)] - \sum_{c\in\{c_x,c_y\}} \mathbb{E}_{x,c}[D_{real/fake}(G(x,c))] - \lambda_{gp}\,\mathbb{E}_{\hat{x}}[(||\nabla_{\hat{x}} D_{real/fake}(\hat{x})||_2 - 1)^2], \tag{8}$$

Here, $\hat{x}$ is uniformly sampled along a straight line between a pair of a real and a fake image. The gradient penalty coefficient (λ_{gp}) is set to 10 for all experiments. Values for λ_{domain} and λ_{cyc}, are set at 1 and 10, respectively, for all experiments. λ_{id} is set to 10 for CelebA, 0.1 for BRATS 2013, and 1 for PE dataset. 200K iteration is found to be sufficient for CelebA and the PE dataset, whereas BRATS 2013 requires 300K iteration for generating good quality images. To facilitate a fair comparison, we use the same generator and discriminator architectures as the public implementation of StarGAN. All models are trained using the Adam optimizer with learning rate $1e^{-4}$ for both the generator and discriminator across all experiments.

Following [4], we slightly change the architecture of the generator to predict a residual (delta) map rather than the desired image directly. Specifically, the generator's output is computed by adding the delta map to the input image, followed by the application of a $tanh$ activation function, $tanh(G(x,c) + x)$. Our ablation study, summarized in Tab. 4, shows the disease detection and localization performance of StarGAN (baseline approach), and the incremental performance improvement using delta map learning, fixed-point translation learning, and the two approaches combined. We find that the major improvement over StarGAN comes from fixed-point translation learning, but the combined approach, for most cases, provides enhanced performance compared to each individual approach (see Tab. 4). We therefore use the combination of delta map learning and fixed-point translation learning in our proposed Fixed-Point GAN, noting that the major improvement over StarGAN is due to the proposed fixed-point translation learning scheme. The implementation is publicly available at http://github.com/jlianglab/Fixed-Point-GAN