Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Signature:

Siyi Gu April 10, 2023

Explanation Iterative Supervision via Saliency-guided Data Augmentation

By

Siyi Gu

Liang Zhao, Ph.D.
Advisor

Computer Science

Liang Zhao, Ph.D.
Advisor

Joyce C. Ho, Ph.D.
Committee Member

Bree Ettinger, Ph.D.
Committee Member

2023

Explanation Iterative Supervision via Saliency-guided Data Augmentation

By

Siyi Gu

Liang Zhao, Ph.D.
Advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences of
Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Computer Science
2023

Abstract

Explanation Iterative Supervision via Saliency-guided Data Augmentation
By Siyi Gu

Explanation supervision is a method that involves using human-generated explanations during training to guide the model. Its goal is to enhance both the interpretability and predictability of the model by integrating human understanding into the training process. Since explanation supervision necessitates a vast amount of training data, data augmentation is indispensable to increase the size and diversity of the original dataset. However, data augmentation for complex data like medical images is particularly difficult due to the following: 1) inadequate training data for the learning-based data augmenter, 2) complexity in producing sophisticated and realistic images, and 3) difficulty in ensuring that the augmented data truly enhances the performance of explanation-guided learning. To address these challenges, we propose the Explanation Iterative Supervision via Saliency-guided Data Augmentation (ESSA) framework for conducting explanation supervision and adversarial-trained image data augmentation using a synergistic iterative loop that handles the conversion from annotation to sophisticated images and the creation of synthetic image-annotation pairs with an alternating training strategy. Comprehensive experiments on three medical imaging datasets demonstrate the effectiveness of our proposed framework in enhancing both the predictability and explainability of the model.

Explanation Iterative Supervision via Saliency-guided Data Augmentation

By

Siyi Gu

Liang Zhao, Ph.D.
Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Computer Science

2023

Acknowledgments

First and foremost, I would lie to express my heartfelt gratitude to my advisor, Dr. Liang Zhao, for his unwavering support and guidance throughout my study. I'm incredibly fortunate to learn from his invaluable insights which have enabled me to conduct research with confidence. Dr. Zhao's extensive knowledge and experience have been a constant source of inspiration for me. I am grateful for his tireless efforts in refining my research papers and slides. Without his guidance and constant support, this thesis would not be conceivable.

I am very thankful to all my committee members, who devoted their valuable time to attend my defense and provided invaluable feedback on my thesis. Moreover, Dr. Ho is my mentor in academics and also daily life, providing me extensive expertise as well as offering life suggestions when I'm struggling with graduate school selection. I'm also fortunate to be Dr. Ettinger's academic advisee in the department of applied math and statistics throughout my undergraduate study.

I would like to also express appreciation to my friends and collaborators during my undergraduate study: Yuyang Gao, Yifei Zhang, Eric Lee, Lizhe Zhang, Leisheng Yu, Minxing Zhang, to name a few. Finally, I would like to dedicate this dissertation with special thanks to my family in China, who emotionally and financially support me throughout my college years.

Contents

List of Figures

List of Algorithms

Chapter 1

Introduction

Deep learning has shown remarkable performance in computer vision and has been extensively used in medical imaging [8]. However, due to the "black box" nature of deep learning models, it can be challenging to ensure the validity of AI's decisions in high-stakes domains like medicine [1]. This has led to a growing interest in Explainable AI (XAI), particularly in the medical imaging domain [3, 14, 21, 9]. Several techniques have been proposed to provide saliency maps that identify the most relevant features or sub-parts of an instance for a model's prediction [10]. However, the quality of these explanations has not been thoroughly examined, including whether the explanation accurately reflects the model's prediction and how to improve the model's explainability when the explanation is incorrect [10].

Explanation supervision is a field that has shown promise in enhancing both the task performance and interpretability of models [11]. While it has been well-explored in NLP and tabular data, its applications in imaging and geometric data domains, such as graphs, are relatively under-explored. This is because the geometric patterns in these domains need to be recognized before supervision can be performed, unlike NLP and tabular data where patterns are in the form of words and hand-crafted features. Despite the benefits of explanation-guided learning, it faces a significant

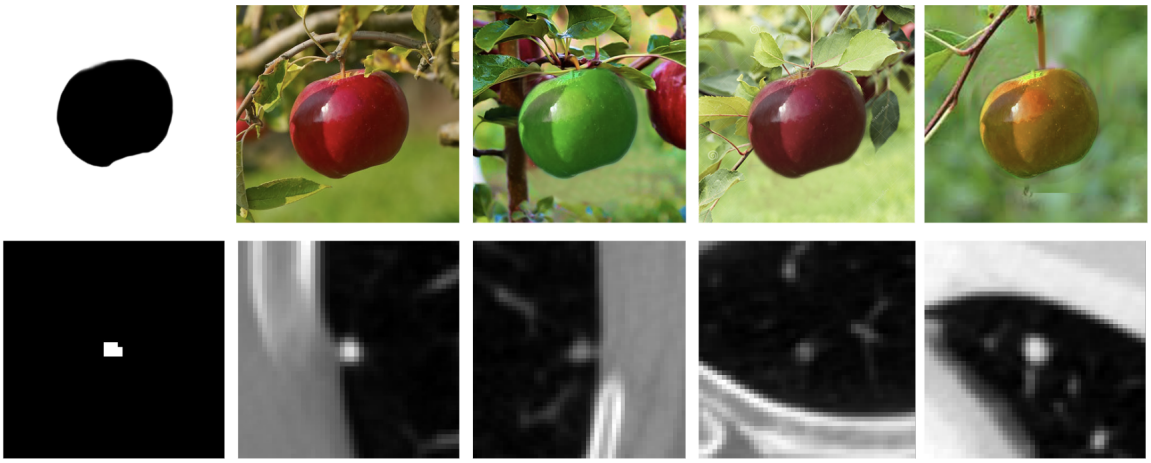

Figure 1.1: Annotation and alternative corresponding images for apple and nodule. The first column is examples of explanation annotations for apples and nodules. The right four columns are possible corresponding images for left annotations.

challenge in terms of the cost of annotating explanations. For example, in the medical imaging domain, radiologists are required to manually annotate explanation masks, using their extensive medical domain knowledge. The manual annotation of volumetric data can take up to fifteen minutes per study, which exacerbates the cost of explanation supervision [31]. As a result, researchers are exploring ways to reduce the cost of annotation and improve the efficiency of explanation supervision in these domains.

One possible solution to address the annotation scarcity issue in machine learning is data augmentation, a technique that artificially increases the size and diversity of the training dataset by generating new samples from existing ones [20]. While traditional data augmentation methods, such as flipping, rotation, cropping, and scaling, have been widely used in the field, they may not capture the true invariants of the patterns and fail to produce diverse and expressive samples. As illustrated in the first row of Figure 1.1, different images that correspond to the same salient area can vary widely in appearance and characteristics, which highlights the importance of more sophisticated and intelligent data augmentation methods that can capture the underlying semantic patterns relevant to the prediction tasks. One such approach

is deep generative models that can learn to generate new samples that are similar to the training data in terms of their features and semantics, while introducing diversity and variations. These models are particularly important in critical areas like medical imaging, where data is often unique and complex, as shown in the second row of Figure 1.1. Therefore, developing powerful data augmenters that can learn sophisticated patterns and generate realistic and diverse samples is a crucial step towards improving the performance and interpretability of machine learning models.

Augmenting explanation annotation for sophisticated data like medical images is a complex and under-explored task that involves several key challenges. One of the most significant challenges is the scarcity of data required to train a learning-based data augmenter. As human annotators need to mark the corresponding explanation for each image, obtaining a large amount of explanation annotation for medical images can be a costly and time-consuming process. This is especially challenging for medical images, which often require domain experts to annotate the images manually.

Another critical challenge is the difficulty in generating realistic sophisticated images that accurately represent the patterns found in the data. Medical images are unique and sophisticated, and it can be challenging for a generative model to learn the mapping "from explanation to image." This requires a large-scale training dataset to ensure the stability of generative models, making it even more challenging to produce realistic images. Additionally, ensuring that the augmented data indeed improves the performance of explanation-guided learning is another crucial challenge. While data augmentation techniques can increase the number of training samples, it does not always result in an increase in data variety and explainability. Traditional data augmentation methods such as rotation, scaling, and cropping are not effective in explanation supervision tasks. While increasing the number of training samples may be useful, it does not alter the 1-to-1 mapping relationship between an image and its corresponding explanation annotation. To address these challenges, researchers

are exploring new methods for generating sophisticated image data augmentation that can improve the performance of explanation-guided learning. These methods include developing expressive models such as deep generative models that can learn sophisticated semantic patterns and identify the relevant and irrelevant patterns for the prediction task. These methods can be especially useful for critical areas such as medical imaging, where data is always unique and sophisticated, making it challenging to create large-scale training datasets for learning-based data augmentation.

In response to the challenges discussed earlier, we propose an innovative framework called Explanation Semi-Supervision via Saliency-guided Data Augmentation (ESSA) that aims to address the scarcity of annotated data by generating new data samples while ensuring that the generated data is realistic and relevant to the target task. One of the primary goals of our proposed framework is to synergize explanation-guided learning and explanation annotation augmentation through an iterative loop between them.

To tackle the issue of the scarcity of annotated data, we develop a novel explanation-guided image generator that can build the mapping between "explanation to image" and leverage post-hoc explainer to achieve the "image to explanation" mapping. By iterating between these two mappings, we can generate a sufficient amount of data for training both the image generator and the image classifier. Moreover, to address the difficulty in generating realistic and sophisticated images, especially for data like medical images, we design a new stratified image generator guided by the given explanations. Our generator has two generation modules, one for the foreground and the other for the background, due to their different patterns. This approach helps generate diverse and representative samples, which can significantly improve the generalization ability of the model. Finally, we propose an innovative algorithm that alternately and iteratively trains the generator and image classifier until saturation is achieved. This algorithm helps ensure that the augmented data indeed boosts the per-

formance of the explanation-guided learning and helps improve the interpretability of the model. In summary, our proposed ESSA framework addresses the key challenges of explanation-guided learning and explanation annotation augmentation through a synergized iterative loop between them. By developing a novel explanation-guided image generator, a stratified image generator, and an innovative training algorithm, our framework can generate realistic and diverse images while improving the interpretability and generalization ability of the model. Specifically, the major contributions of this paper are summarized as follows:

1. **A novel framework for explanation supervision with adversarial-trained data augmenter.** The framework can jointly achieve explanation supervision and image data augmentation via a synergized iterative loop between them.

2. **A new proposed generator for "annotation-to-image" translation.** The generator includes two decoders to capture different patterns for both foreground and background of the explanation annotation to generate realistic and diverse images, and achieve "1-to-many" mapping relationship from annotation to image.

3. **A new proposed algorithm for training data augmenter and classifier.** The data augmenter and classifier can be trained alternately for multiple iterations. This training strategy can avoid error back-propagation and generate a sufficient amount of data for explanation supervision.

4. **Extensive experiments are conducted to evaluate our proposed approach.** We use a variety of datasets and evaluation metrics to demonstrate that our approach can effectively improve both model predictability and explainability. Furthermore, we conduct a thorough analysis of the generated explanations to show that they are coherent and informative.

Chapter 2

Related Work

2.1 Data Augmentation and applications in Medical imaging

Data augmentation has become a widely used technique in deep learning to increase the size of training datasets, particularly in challenging domains such as medical imaging where obtaining large datasets can be difficult [33]. One popular method to generate synthetic image-label pairs is to use semantic labels of anatomical structures to create synthetic medical images. However, traditional augmentation methods such as cropping, Gaussian noise, and elastic transformation provide limited variability [7]. To address this issue, researchers have developed more sophisticated approaches based on adversarial training. These methods have been applied for a variety of purposes, such as semantic image synthesis, image classification, and image-to-image translation. For instance, Xu et al. proposed a semi-supervised attention-guided CycleGAN for brain tumor classification, which augments tumor images from normal images [35]. Other researchers have utilized label image translation techniques to generate synthetic image-mask pairs for semantic image synthesis. Shin et al. and Cao et al. have applied pix2pixGAN to generate synthetic CT and PET images, and abnormal

brain MRI images, respectively [32, 5, 17]. However, these approaches mainly focus on either generating synthetic images from labels or improving the accuracy of segmentation, whereas our focus is on using image and mask pairs as a joint input to downstream classification tasks.

2.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) have played a significant role in advancing computer vision research and have had a notable impact on various tasks like image-to-image translation, semantic segmentation, and image editing [13, 18, 17]. GANs have become a prevalent technique for generating synthetic data to expand the size of training datasets and avoid overfitting deep neural networks, particularly in fields like medical imaging, where acquiring large datasets can be challenging [33]. However, traditional data augmentation methods like cropping, Gaussian noise, and elastic transformation are limited in terms of variability [7]. Contemporary augmentation approaches are primarily based on adversarial training [7]. In the case of image-to-image translation, the aim is to learn the mapping between images from one domain to another. Conditional generative adversarial networks (cGANs)[22] have been proposed to tackle this problem, either with paired data[17, 6, 25] or unpaired data [36]. The popular pix2pixGAN [17] model utilizes an encoder-decoder generator that takes in semantic label maps as input and a PatchGAN discriminator [16]. Modifications to the generator and discriminator have been made in subsequent works, such as pix2pixHD [34], SPADE [25], and SESAME [24]. However, these approaches require precise semantic labels, which can be difficult to obtain in medical imaging. or example, Xu et al.[35] proposed a semi-supervised attention-guided CycleGAN to augment tumor images from normal images for brain tumor classification. Label image translation technique is also frequently used for semantic image synthesis, where

the purpose is to generate synthetic image-mask pairs instead of synthetic images. Shin et al. and Cao et al. applied pix2pixGAN to generate synthetic CT and PET images, and abnormal brain MRI images, respectively[32, 5, 17].

2.3 Explanation Supervision

In the fields of NLP and tabular data, there has been significant research on incorporating human knowledge into interpretable models through methods such as attribution and feature regularization [23, 28]. More recently, there has been a growing awareness of the importance of visual explanations, with saliency maps being a popular approach to generating local explanations that highlight the input features responsible for a model's prediction [23, 28]. The incorporation of network activations into visualizations has further improved the effectiveness of this approach, as demonstrated in Grad-CAM [28]. A conceptual framework for image classification with human annotation in the form of scribble annotations as the explanation supervision signal is HAICS [29]. However, reliance on the accuracy of ground truth annotations poses a significant challenge in practice. Inaccurate, incomplete, and inconsistent distribution of human annotation can lead to errors when directly used as supervision signals for model explanation [12]. To address this challenge, Gao et al.[12] develop a novel objective that handles these issues. Despite these recent advances, the acquisition of a large volume of explanation labels is still a significant challenge due to their high cost[23, 28, 29].

Chapter 3

Problem Formulation

Given a set of images $X = \left\{x^{(i)} \in \mathbb{R}^{C\times H\times W}\right\}_{i=1}^{N}$ with their class labels $y = \left\{y^{(i)}\right\}_{i=1}^{N}$ and corresponding explanation annotations $M = \left\{m^{(i)} \in \mathbb{R}^{H\times W}\right\}_{i=1}^{N}$ for the label, where N is the sample size, C denotes number of channels, H denotes height, and W denotes width.

The problem of explanation supervision is to predict the class label $\hat{y}^{(i)}$ and give model explanation $\hat{m}^{(i)}$ of the input image $x^{(i)}$ supervised by both ground truth class label $y^{(i)}$ and annotation $m^{(i)}$ such that both model prediction loss $-\sum_{i=1}^{N} y^{(i)} \cdot \log\left(\hat{y}^{(i)}\right)$ and model explanation loss $\sum_{i=1}^{N} \|m^{(i)} - \hat{m}^{(i)}\|_1$ are minimized. However, getting usable X and M is difficult due to the following challenges.

Challenge 1 (Scarcity of annotation): A learning-based data augmenter requires a large amount of annotated images which are usually marked by human annotators. This process can be costly, especially for medical images that need to be annotated by professionals with specialized knowledge.

Challenge 2 (Generation of sophisticated image): Generating realistic and sophisticated images is a challenge. This is particularly true for medical images which are highly unique and sophisticated. It is difficult for a generative model to learn the mapping from explanation annotation to image due to the diverse and intricate

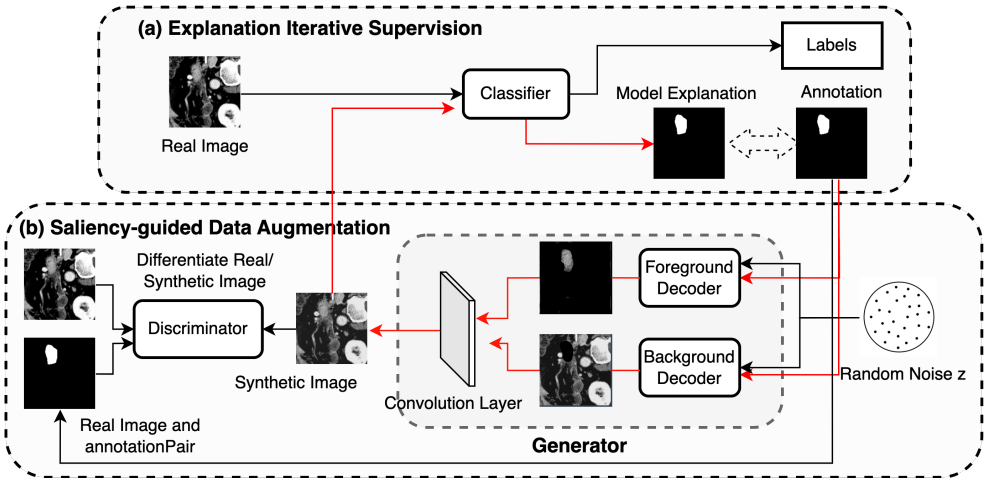

Figure 3.1: Illustration of our proposed ESSA Framework. ESSA consists of an Explanation Iterative Supervision module (a) and a Saliency-guided Data Augmentation module (b). In particular, Saliency-guided Data Augmentation devotes to training a learning-based data augmenter to achieve "annotation-to-image" translation and generate synthetic images. The Explanation Iterative Supervision devotes to the prediction task supervised by annotations and labels iteratively.

patterns in these images.

Challenge 3 (Effectiveness of data augmentation): Ensuring augmented data can effectively improves explanation-guided learning performance is a challenge. Data augmentation increases samples but not always diversity or explainability. Traditional methods such as rotation, scaling, and cropping are ineffective in explanation-guided tasks since they do not alter the 1-to-1 mapping relationship between image and annotation.

Chapter 4

Model

In this section, we will begin by introducing our framework, which we have named ESSA. Following this, we will present our recently developed adversarial-trained data augmenter. This augmenter features a specially designed generator that is capable of addressing the challenge of translating annotations to images. To conclude, we will introduce a novel algorithm that allows for the iterative and alternate training of both the image classifier and the generator.

4.1 Proposed Framework

The primary objective of our proposed framework, ESSA, is to enhance both the predictability and interpretability of the model by utilizing explanation supervision, where the model is supervised based on both class label and explanation annotation. However, the scarcity of image-annotation pairs makes it challenging to train the model effectively. To overcome this challenge, data augmentation is necessary to generate additional input images, explanation annotations, and prediction label pairs.

While traditional non-learnable data augmentation techniques like rotation, scaling, and cropping exist, they are inadequate to achieve our goal as they do not change the one-to-one mapping relationship between the image and annotation. In the real-

world scenario, the relationship between a mask and an image is usually one-to-many, implying that one annotation can correspond to several images with different patterns and backgrounds, as illustrated in Fig. 1.1. Therefore, to ensure that the classifier can recognize the reasonable and predictable explanation annotation from the diverse and noisy images, we require a model that can perform the annotation-to-image translation task to augment the data.

The issue of scarcity of image-annotation pairs can be solved by our proposed ESSA framework, as depicted in Figure 3.1. The framework takes an explanation annotation and a random noise vector as input and generates a synthetic image through a generator. A discriminator is then employed to distinguish between the synthetic and real images. Once the generator is adequately trained, it is used to generate synthetic images that are fed, along with their corresponding annotations and labels, to the classifier for prediction with supervision from both labels and annotations. This framework achieves annotation-to-image translation, enabling the creation of multiple images corresponding to a single annotation, providing adequate training samples for explanation supervision. Traditional data augmentation methods, such as rotation, scaling, and cropping, are inadequate for this purpose since they do not alter the 1-to-1 mapping relationship between the image and annotation, whereas in the real-world setting, the mask-to-image relationship is always 1-to-many, meaning a single annotation can correspond to several images. Therefore, our framework is designed to achieve annotation-to-image translation for data augmentation, ensuring that the classifier can identify the reasonable and predictable explanation annotation from the diverse, noisy, and realistic images.

Based on the statement above, there are three different losses in our framework: the model's prediction loss, the model's explanation loss, and the data augmenter's training loss. The overall objective function of our ESSA framework can be expressed

as:

$$\begin{aligned}
&\min_{f,G,D} \sum_{i=0}^{N} \left(\mathcal{L}_{\text{Pred}}(f(x^{(i)}), y^{(i)}) + \sum_{t=0}^{T} \mathcal{L}_{\text{Pred}}(f(\tilde{x}^{(i,t)}, y^{(i)})) \right) + \\
&\quad \sum_{i=0}^{N} \left(\mathcal{L}_{\text{Exp}}(g(f(x^{(i)})), m^{(i)}) + \sum_{t=0}^{T} \mathcal{L}_{\text{Exp}}(g(f(\tilde{x}^{(i,t)})), m^{(i)}) \right) + \\
&\quad \mathcal{L}_{\text{Reg}}(G, D) \\
&\quad s.t. \quad \tilde{x}^{(i,t)} = G(m^{(i)}, z^{(t)})
\end{aligned} \tag{4.1}$$

The first term is the prediction loss for both real and synthetic images; The second term is the explanation loss for explanation annotations generated by the model explanation method from both real and synthetic images; The last term is the loss for the data augmenter's adversarial training; $g(\cdot)$ denotes the model explanation method; $G(\cdot)$ denotes the generator for data augmentation; $D(\cdot)$ denotes the discriminator for data augmentation; $\mathcal{L}_{\text{Pred}}$ is the prediction loss (such as the cross-entropy loss); $\mathcal{L}_{\text{Exp}}$ is the explanation loss to measure the difference between model generated explanation and ground truth explanation annotation (such as the L1 loss); $\mathcal{L}_{\text{Aug}}$ is the data augmenter's training loss (refer to 4.2 for details), and T denotes the number of augmentation iterations required to reach the optimum.

4.2 Data Augmentation via Adversarial Training

To generate realistic and sophisticated images from annotations, we proposed a novelty data augmenter consisting of a stratified generator and a discriminator trained by adversarial strategy.

Generator The generator $G(\cdot)$ learns a mapping from annotations $m^{(i)}$ and random noise $z^{(t)}$ to synthetic image $\tilde{x}^{(i,t)}$:

$$G : \left\{ m^{(i)}, z^{(t)} \right\} \rightarrow \left\{ \tilde{x}^{(i,t)} \in \mathbb{R}^{C \times H \times W} \right\} \tag{4.2}$$

where t denotes the t-th adversarial training iteration. Realistic and sophisticated synthetic images should have both reasonable saliency area and background. However, the distribution of saliency area and background are always highly different. For example, as shown in row 2 of Fig. 1.1, the patterns in the nodule and those outside show a big difference. To solve this issue, we establish two decoders $\{Dec_1, Dec_2\}$: $\mathbb{R}^{H\times W} \rightarrow \mathbb{R}^{H\times W}$ that is designed to generate foreground and background areas respectively. The architecture of the decoder is a U-Net [26]. Decoded input annotation's foreground and background are then combined together and fed into a convolution layer F_{Conv} to get the final output. Formally, the generation of synthetic images by stratified generator $G = \{Dec_1, Dec_2, F_{Conv}\}$ can be written as:

$$\begin{aligned}\tilde{x}^{(i,t)} = F_{Conv}(m^{(i)} \odot Dec_1(m^{(i)}, z^{(t)}) + \\ (\mathbb{J} - m^{(i)}) \odot Dec_2(m^{(i)}, z^{(t)}))\end{aligned} \tag{4.3}$$

where $\odot$ denotes element-wise multiplication, $\mathbb{J}_{H\times W}$ denotes a matrix whose all elements are 1, and $\mathbb{J} - m$ denotes the background area of the annotation $m^{(i)}$ by reversing its bit pixels.

Discriminator To train such stratified generator $G(\cdot)$, we use an adversarial training strategy where a conditional discriminator $D(\cdot)$ is leveraged. The conventional way to discriminate the real and synthesized domains is by using a binary classifier [30]. However, it is infeasible for image generation because of its sparse supervision. To achieve pixel-level supervision, we use a discriminator architecture named PatchGAN [16]. The convolution layers of discriminator architecture first map both real and synthetic image $\left\{x^{(i)}, \tilde{x}^{(i,t)}\right\}$ to a patch domain $\{p, \tilde{p}\} \in \mathbb{R}^{N\times N}$, where N is the patch size, to classify if each $N \times N$ patch in an image is real or fake. To be more specific, the objective of the discriminator $D(\cdot)$ is to

$$\min_D \|p - \mathbb{J}_{N\times N}\|_1 + \|\tilde{p} - \mathbb{O}_{N\times N}\|_1 \tag{4.4}$$

where $\mathbb{O}_{N\times N}$ and $\mathbb{J}_{N\times N}$ are matrices with all 0 and 1 elements and with sizes of $N\times N$.

Objective Based on the statement above, the objective of the generator and discriminator is:

$$\mathcal{L}_{Aug}(\theta_G,\theta_D)=\sum_{i=1}^{k}\Big[\log D(m^{(i)},x^{(i)})+\log\left(1-D(m^{(i)},G(m^{(i)},c^{(t)}))\right)\Big] \tag{4.5}$$

where θ_G denotes the parameters of generator; θ_D denotes the parameters for discriminator; k denotes number of training samples. We also introduce the L1 regularization term to the objective function to make the synthetic images generated by the generator can not only fool the discriminator but also be near the real images.

$$\mathcal{L}_{L1}(\theta_G)=\sum_{i=1}^{k}\|x^{(i)}-G(m^{(i)},c^{(t)})\|_1 \tag{4.6}$$

The final objective of the proposed Saliency-guided data augmentation model is

$$G*=arg\min_{G}\max_{D}=\mathcal{L}_{Gen}(\theta_G,\theta_D)+\lambda\mathcal{L}_{L1}(\theta_G)) \tag{4.7}$$

where λ is the weight hyper-parameter for the L1 regularization term.

4.3 Alternating and Iterative Training

From Equation 4.1 we can see that our objective function can be regarded as a constrained optimization problem. Since the input images of the classifier are generated by fixed trained generator G, we can not optimize generator G together with classifier f when doing back-propagation via $\mathcal{L}_{Pred}$ and $\mathcal{L}_{Exp}$. To solve this problem, we design an innovative algorithm to train the classifier and data augmenter alternately and iteratively. To be more specific, our algorithm is first to fix the classifier's parameter

Algorithm 1: Alternating and Iterative Training Algorithm

Require: X, M, y
Ensure: f, G, D

initialize: $best\_acc = 0, val\_acc = 0$
for $t = 1 : T$ **do**
  Sample k pairs (X', M', y') from (X, M, y)
  if best_acc $<$ val_acc **then**
    Sample v pairs from $(X, M, y) \setminus (X', M', y')$
    for $q = 1 : T_G$ **do**
      Compute $\mathcal{L}_{Aug}$ based on Equation 4.7
      Compute ∇_{θ_D} and ∇_{θ_G} based on Equation 4.8, 4.9
      $\theta_D \leftarrow \theta_D - \eta_D \nabla_{\theta_D}$
      $\theta_G \leftarrow \theta_G - \eta_G \nabla_{\theta_G}$
    end for
    for $q = 1 : T_C$ **do**
      Generate v images using (X', M', y') via Equation 4.3
      Compute $\mathcal{L}_{Pred}$ and $\mathcal{L}_{Exp}$ based on Equation 4.10, 4.11
      Compute $\mathcal{L}_{Reg}$ based on Equation 4.13
      Compute ∇_{θ_f} and ∇_{θ_G} based on Equation 4.12, 4.14
      $\theta_f \leftarrow \theta_f - \eta_f \nabla_{\theta_f}$
      $\theta_G \leftarrow \theta_G - \eta_g \nabla_{\theta_G}$
      Compute val_acc and update $best\_acc$
    end for
  end if
end for

θ_f while optimizing generator G and discriminator D, and then fix the generator's parameter θ_G and discriminator's parameter θ_D while optimizing classifier f. We repeat this alternating training process iteratively until the model reaches the termination criteria.

The overall algorithm is summarized in Algorithm 4.2. (X, M, y) are pre-given real image, annotation, and label pairs. We first initialize the classifier's best prediction accuracy to be 0 on Line 1. Then we repeat the whole alternating training process for T iteration from Line 2-22 which consists of two training modules in each iteration. Within one iteration, we first randomly sample k image, annotation, and label pairs (X', M', y') from real dataset (X, M, y) on Line 3 for synthetic image generation and explanation supervision. Then we sample v image annotation pairs from (X, M, y) and not in (X', M', y') on Line 5 for data augmenter training. From Lines 6-11, we train the data augmenter for T_G iterations. Specifically, we first use the randomly sampled v image-annotation pairs to compute the data augmenter loss $\mathcal{L}_{Aug}$ and its gradient w.r.t. θ_G and θ_D, respectively. Gradients are computed as follows:

$$\nabla_{\theta_D} = \frac{\partial}{\partial \theta_D} \mathcal{L}_{Aug}(\theta_G, \theta_D) \tag{4.8}$$

$$\nabla_{\theta_G} = \frac{\partial}{\partial \theta_G} \mathcal{L}_{Aug}(\theta_G, \theta_D) \tag{4.9}$$

and then update parameters θ_G and θ_D with a learning rate η_G and η_D, respectively, from Line 9-10.

From Lines 12-20, we train the classifier for T_C iterations. We first use trained generators G from Line 6-11 to generate v synthetic images corresponding to annotations from v sampled pairs on Line 13. Then we use the generated synthetic images to do predictions with the supervision of labels and annotations. We first compute

$\mathcal{L}_{Pred}$ and $\mathcal{L}_{Exp}$ as follows:

$$\mathcal{L}_{Pred}(\theta_f) = -\sum_{i=1}^{v}(y^{(i)} \cdot \log f(\tilde{x}^{(i,t)})) \tag{4.10}$$

$$\mathcal{L}_{Exp}(\theta_f) = \sum_{i=1}^{v}\|g(f(\tilde{x}^{(i,t)})) - m^{(i)}\|_1 \tag{4.11}$$

and then compute their gradient w.r.t. θ_G as follows:

$$\nabla_{\theta_f} = \frac{\partial}{\partial \theta_f}(\mathcal{L}_{Pred}(\theta_f) + \lambda \mathcal{L}_{Exp}(\theta_f)) \tag{4.12}$$

where λ is the hyper-parameter to balance prediction and explanation loss. In this training process, we also include a regularization term $\mathcal{L}_{Reg}$ to make the generated images to be near the real images:

$$\mathcal{L}_{Reg}(\theta_G) = \sum_{i=1}^{k}\|x^{(i)} - G(m^{(i)}, z^{(t)})\|_1 \tag{4.13}$$

The gradient of regularization term $\mathcal{L}_{Reg}$ w.r.t. θ_G is compute as follows:

$$\nabla_{\theta_G} = \frac{\partial}{\partial \theta_G}(\mathcal{L}_{Reg}(\theta_G)) \tag{4.14}$$

We then update generator G while updating classifier f with a learning rate η_g and η_f, respectively, from Line 17-18.

On Line 19, we compute the model's prediction accuracy on the validation set and update the best accuracy. We repeat the overall alternating training process stated above for T iterations until the training reaches optimum when the classifier's prediction accuracy does not increase on the validation set.

Chapter 5

Experiment

We test our framework on three datasets in the healthcare domain: LIDC-IDRI [2], Pancreas-CT [27] and Medical Segmentation Decathlon for two different tasks: pulmonary nodule classification and pancreatic tumor classification. We first outline the specific configurations for the experiments and then showcase the quantitative evaluations of both the model's predictions and its explanations. Furthermore, we incorporate various qualitative evaluations, such as case studies of model-generated explanations and synthetic images, to have a comprehensive assessment of the proposed model.

5.1 Experimental Settings

5.1.1 Pancreatic tumor classification

We acquired normal pancreas images from the Cancer Imaging Archive [27][1], and abnormal images from the Medical Segmentation Decathlon dataset (MSD) [2]. The dataset comprises 281 CT scans with tumors and 80 CT scans without tumors, and its objective is pancreatic tumor classification. Since the CT scans are in 3D, we

[1]The dataset is available at: https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT
[2]The dataset is available at: http://medicaldecathlon.com/

transformed them into 2D by randomly slicing along the z-axis while preserving the tumor's presence, resulting in images of size 224 x 224. We balanced the sample size for training and kept the original ratio for validation and test sets. The MSD dataset includes two types of annotations, namely, tumor lesions and pancreas segmentation. We considered the tumor lesions as our explanation labels. To simulate a more practical scenario where we have limited access to human explanation labels, we randomly selected 20, 50, and 100 images for training. We split the dataset into 70% for training, 15% for validation, and 15% for testing.

5.1.2 Pulmonary nodule classification

We used the LIDC-IDRI dataset [2] [3] which contains thoracic computed tomography (CT) scans for lung cancer screening, with annotated lesions. We preprocessed the 3D nodule images into 2D images by slicing them along the z-axis at the middle. The dimensions of the resulting images were 224 x 224. The annotations were provided in XML format by four experienced thoracic radiologists. We computed the consensus volume among the four annotations for each image at a 50% consensus level, which was used as the explanation annotation. We further used the surrounding areas of the nodules as negative samples. After preprocessing, the dataset contained a total of 2625 nodules and 65505 non-nodule images. The objective of the task was to classify images as containing nodules or not. We split the dataset into 20% for training, 20% for validation, and 60% for testing.

5.1.3 Evaluation Metrics

To evaluate the model, we consider both its performance in the task and its level of explainability. For performance evaluation, we use standard metrics like prediction accuracy and the AUC of the Receiver Operating Characteristic curve. To assess the

[3] Available at: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254

quality of the model's explanations, we compare them with human-labeled ground truth in the test set. We use the IoU score described in [4], which is computed by comparing the ground truth explanation with the binary explanation generated by the model. The IoU score measures the positive overlap between the two inputs. In addition, we calculate pixel-wise precision, recall, and F1-score, which provide a more comprehensive evaluation of the model's explanations by considering both positive and negative explanations.

5.1.4 Comparison methods

We compare the performance of the different augmentation methods with RES-G as the classification model. Concretely, we studied the following methods:

- **Baseline**: No augmentation methods are applied. We simply train the model with cross entropy as prediction loss and apply the robust explanation loss from RES-G.

- **Traditional augmentation**: We apply a series of augmentation techniques to the training data: affine, horizontal flip, vertical flip, Gaussian Noise, cropping, rotation, and elastic transformation at random.

- **Pix2pix** [29]: A framework that adds the L1 distance to conditional GAN, comparing the pixel difference between synthetic image and real image. Therefore, the training loss includes an adversarial loss and an L1 loss. To be specific, the pix2pix model employs a U-net as the generator and a PatchGAN as the discriminator.

- **SPADE [25]**: A framework that modulates the activations in normalization layers via a spatially adaptive, learned transformation from low to high scales. Its training loss consists of an adversarial loss, a feature-matching loss, and

the VGG-based perceptual loss, which heavily learns from the framework of pix2pixHD [34].

5.1.5 Implementation Details

For all the methods studied in this work, the classification model is based on the ResNet18 [15] architecture with the addition of a robust explanation loss introduced by RES [12]. RES is implemented according to the standards outlined in [12]. The batch size, slack variable α, regularization factor, and attention weight are set to 16, 0.01, 0, and 1, respectively. The models are trained for 50 epochs using the ADAM optimizer [19] with a learning rate of 0.001. For adversarial training of comparison methods, we follow the publicly available implementations without substantial changes to the architecture. λ, the weight of L1 loss is set to 100. For SPADE, we specifically set instance labels to False. All models are trained for 100 epochs using the ADAM optimizer with a learning rate of 0.0002. The batch size is set to 1. We augment the training data by 100% for all methods besides iterative training, where the number of iteration training is determined by the optimal validation accuracy.

5.2 Performance

Table 5.1 shows the prediction power and explanation performance for two downstream tasks: pulmonary nodule classification and pancreatic tumor classification. The results are obtained from 5 individual runs. The best results are highlighted with boldface font and the second bests are underlined. In general, our proposed framework outperformed all other comparison methods in terms of both prediction power as well as explainability.

To be specific, in the pancreatic tumor classification task, ESSA consistently yields the best performance on all metrics, slightly improving classification performance and

Table 5.1: The prediction and explanation evaluation on both datasets. The best results for each task are highlighted with boldface font and the second bests are underlined. The training sample size for pulmonary nodule classification is 100, and the sample size for pancreatic tumor classification is 20.

Dataset	Model	Accuracy ↑	AUC ↑	IoU ↓	Precision ↑	Recall ↑	F1 ↑
Pancreas	No augmentation	89.09 ± 2.39	94.70 ± 4.78	5.82 ± 0.91	76.52 ± 1.80	62.22 ± 3.96	64.85 ± 3.45
	Traditional	92.72 ± 2.27	98.46 ± 0.06	5.80 ± 0.01	77.78 ± 0.01	63.64 ± 0.02	68.75 ± 0.03
	Pix2pix	93.18 ± 1.23	98.53 ± 0.82	7.32 ± 1.82	70.82 ± 3.21	62.82 ± 4.23	65.93 ± 4.12
	SPADE	89.18 ± 3.23	88.53 ± 4.82	5.39 ± 0.82	66.82 ± 3.81	52.32 ± 3.83	58.21 ± 3.28
	Proposed(1 iter)	95.34 ± 1.91	98.75 ± 0.66	11.58 ± 2.32	81.43 ± 4.74	61.90 ± 4.87	66.38 ± 1.45
	Proposed(multiple iters)	**96.16 ± 1.21**	**98.95 ± 0.83**	**13.21 ± 2.16**	**85.43 ± 4.28**	**71.28 ± 3.21**	**73.82 ± 1.21**
LIDC-IDRI	No augmentation	94.35 ± 1.95	78.69 ±3.17	8.08 ± 1.83	50.79 ± 7.88	31.06 ± 10.60	36.36 ± 3.23
	Traditional	95.33 ± 1.57	81.82 ± 3.71	14.95 ± 2.85	68.58 ± 6.70	41.53 ± 6.74	46.34 ± 51.55
	Pix2pix	93.63 ± 0.32	83.07 ± 0.94	19.04 ± 0.90	61.35 ± 3.09	37.10 ± 3.55	42.32 ± 3.09
	SPADE	88.79 ± 2.39	78.16 ± 2.1	18.64 ± 2.92	67.34 ± 3.21	43.20 ± 4.23	48.55 ± 3.82
	Proposed(1 iter)	**96.29 ± 3.28**	85.11 ± 2.22	21.07 ± 3.19	62.46 ± 6.49	38.88 ± 3.90	43.30 ± 4.14
	Proposed(multiple iters)	95.36 ± 3.95	**87.87 ± 2.01**	**31.31 ± 2.32**	**81.32 ± 2.82**	**50.55 ± 3.88**	**58.18 ± 2.63**

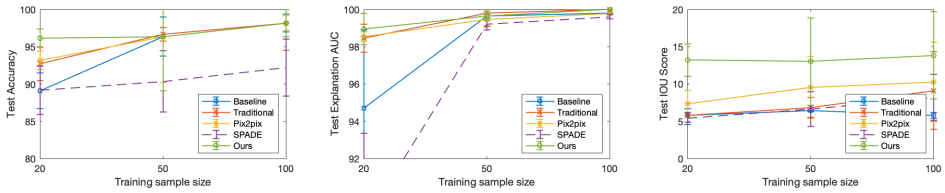

Figure 5.1: Model performance under different training sample size scenarios on pancreatic tumor classification. The data points and error bars represent the mean value and standard deviation over 5 runs respectively. (Left) Test accuracy. (Middle) Test AUC. (Right) Test IoU.

boosting explanation quality significantly, with 1-7.8%, 29-135%, and 8-12% increase in terms of accuracy, IoU, and explanatory F1 scores, respectively, compared with the baseline and other comparison methods.

Moreover, for pulmonary nodule classification, the ESSA framework increases the prediction accuracy and AUC by 1-8.4%, and 5.77-11.6%. ESSA with one iteration achieves the highest accuracy among all models, while ESSA with iterative training obtains the highest AUC. This is attributed to the data imbalance issue as the positive-to-negative ratio is approximately 1:26 for LIDC-IDRI. Moreover, ESSA consistently outperforms other models in terms of model explainability. To be specific, there is a 285% increase in IoU compared with the baseline. ESSA also outperforms SPADE and pix2pix by approximately 64%, 21-33%, 18-36%, and 21-38% in terms of IoU, explanatory precision, recall, and F1.

We further investigated the utilization of the EESA framework to enhance the generalization capabilities of DNN models under varying training sample size scenarios. We considered three training sample sizes (20, 50, and 100) using the pancreatic tumor dataset. The results of the prediction accuracy, AUC, IoU score, and explanatory F1, precision, and recall are presented in Figure 5.1. Each data point represents the mean value calculated from five independent runs, with the error bars indicating the standard deviation.

The results demonstrate that the proposed framework outperforms other comparison methods in all the scenarios studied, particularly in terms of the explainability of the DNN models as reflected by IoU, explanatory F1, precision, and recall. Interestingly, our results show that when the sample size is limited, such as when the training sample size is 20, ESSA outperforms all other comparable models. This indicates that the augmented samples are effective at supervising the model performance. On the other hand, when the training size gets larger, our explanation supervision framework copes well with the noisy labels. Therefore, the prediction accuracy and AUC are similar across models when training samples are relatively sufficient.

5.3 Qualitative Analysis of Augmentation

In this section, we compare the effectiveness of different data augmentation techniques for synthesizing medical images. Five representative examples were selected from each dataset, based on differences in nodule size and the variety of anatomical structures in the background. The results of the comparison are presented in Figure 5.2. We observe that their proposed framework was able to capture more semantic structures of the surrounding tissues and offer more data variability. Specifically, when generating pancreas images, our framework is successful in capturing the shape of the tumor while also providing a diversity of background information.

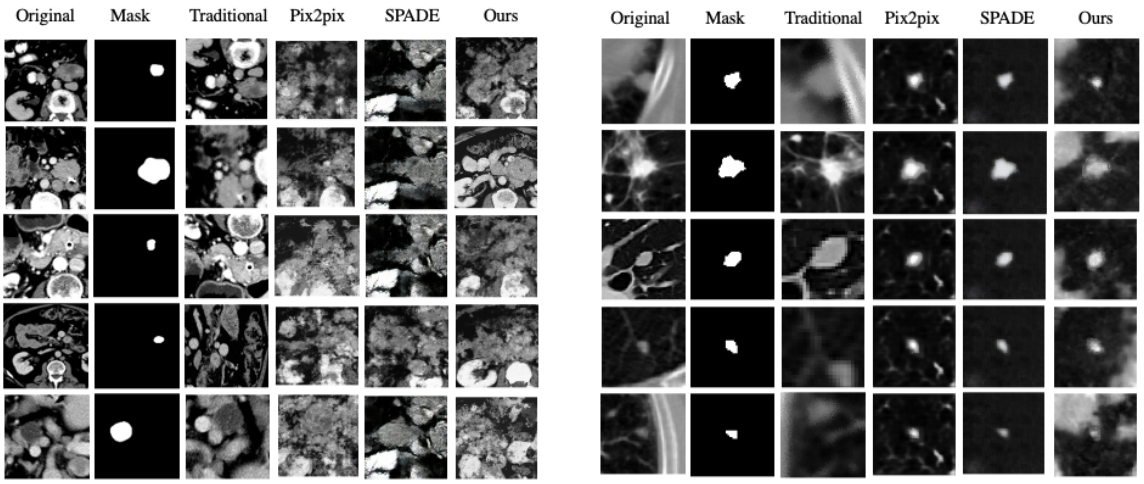

Figure 5.2: Selected visualization of synthetic images on pulmonary nodule classification (left) and pancreatic cancer classification (right). The first two columns represent the real images and corresponding masks. The following columns are synthetic images from different augmentation methods.

However, we note that SPADE generated clearer patterns of anatomical structures due to the application of pix2pixHD. Nonetheless, SPADE and pix2pix exhibit mode collapse, as seen in column 4 and 5, where they generate similar backgrounds regardless of the difference in input masks. We also observe similar patterns in pulmonary nodule synthesis. Pix2pix and SPADE are unable to capture any or very little background when the training sample is highly irregular without semantic labels of all the structures. Although all synthetic images remain differentiable from real images, this was likely due to the small training size. It is possible that when the training set is sufficiently large, generation of indistinguishable synthetic images may occur.

5.4 Qualitative Analysis of the Explanation

In this paper, we present a thorough examination of the comparison between model-generated explanations for both pulmonary nodule classification and pancreatic tumor classification. Our visualization only includes images with positive labels since they have ground-truth annotation. Our results are displayed in Figure 5.3, where

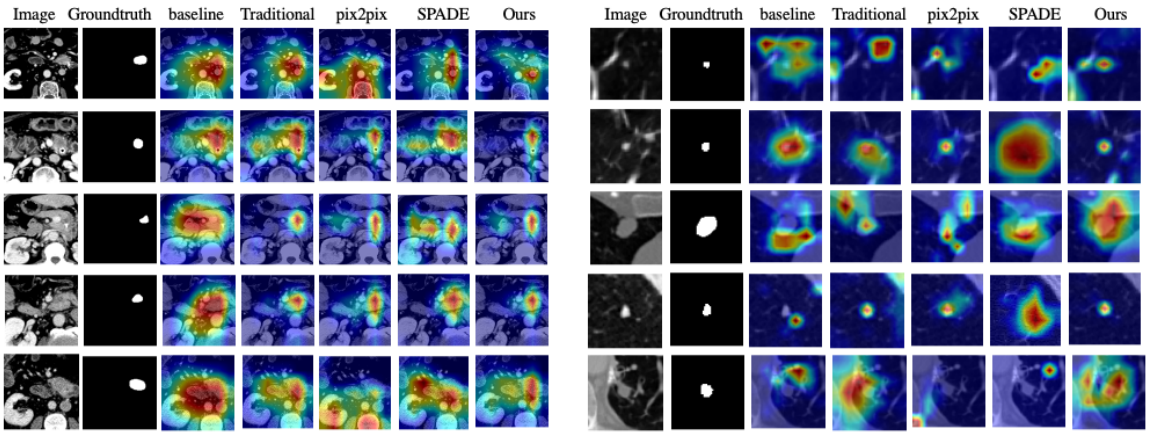

Figure 5.3: Selected explanation visualization results on pulmonary nodule classification (left) and pancreatic cancer classification (right). The model-generated explanations are represented by the heatmaps overlaid on the original image samples, where more importance is given to the area with a warmer color.

we present the model-generated explanations in the form of heatmaps generated by Grad-CAM [28]. The intensity of the color in the heatmaps represents the level of importance attributed to a particular area in the image.

Pancreatic Tumor Classification: For the pancreatic tumor classification, as shown in the left half of Figure 5.3, we selected five examples of model-generated explanations under different models. The results of our study indicate that explanations generated by models using the RES framework outperform the baseline model and other comparison methods in terms of accuracy and alignment with ground truth in categorizing scenes as originating from urban or natural environments. As depicted in Figure 5.3, the explanations generated by the proposed model are more fine-grained, whereas the baseline model focuses on a high proportion of the image. In the third and fourth rows, the explanation generated by pix2pix and ESSA are similar but our model is slightly more fine-grained and more accurate. While all the models provide correct predictions of where the nodule is located, ESSA exhibits improved robustness and generalization capabilities by more precise identification of important areas.

Pulmonary Nodule Classification: In the context of pulmonary nodule classification, we analyzed the explanations generated by different models and visually

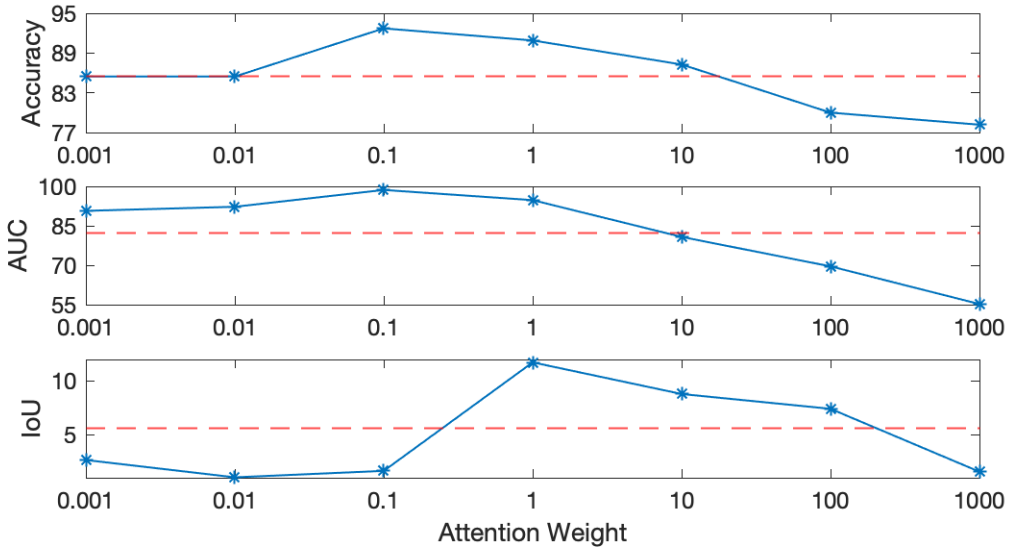

Figure 5.4: The sensitivity study of attention weight on pancreatic tumor classification.

compared their performance in the right half of Figure 5.3. Our proposed RES models were found to generate more accurate and precise heat maps in locating the important areas, such as nodules, compared to the baseline model and traditional augmentation methods. These models were only able to provide a broad range of where the nodule is located, whereas our ESSA model provided a more precise location of the nodule area. Moreover, the attention mechanism used in the baseline model was found to be distracted by irrelevant surrounding tissues, whereas our model focused only on the nodule itself. This was demonstrated in the first row of Figure 5.3, where the baseline model highlighted multiple areas that were important, while our model focused solely on the nodule. The accuracy of the explanations generated by baseline and comparison models was found to be less accurate when there were multiple surrounding tissues. This pattern was observed in the remaining four rows of Figure 5.3, which highlighted the robustness of the proposed ESSA framework in generating more accurate and refined explanations and improving model explainability.

5.5 Sensitivity Analysis of Hyper-parameter

To evaluate the impact of attention weight on the performance of our proposed RES framework for pancreatic tumor classification, we conducted a sensitivity analysis. The attention weight is the factor that determines the weight of the prediction loss and attention loss. We varied the attention weight from 0.001, 0.01, 0.1, 1, 10, 100, 1000 and evaluated the model's prediction accuracy, AUC, and Intersection over Union (IoU) score, which measures the overlap between the model-generated explanation and the ground truth explanation. The baseline model's performance is shown by red dashed lines in Figure 5.4. The results indicate that the model's performance is sensitive to the value of the attention weight, with a concave curvature observed in all metrics. The model achieves the best accuracy and AUC at an attention weight of 0.1 and the best IoU at an attention weight of 1. As the attention weight increases, the model becomes more heavily influenced by attention loss, leading to reduced performance. Conversely, when the attention weight is too small or too large, the model underperforms compared to the baseline. We conclude that the model yields the best overall performance when the attention weight is 1, although this comes at the expense of explainability.

Chapter 6

Concluding Remarks

Explanation supervision is a challenging task for the model to be trained sufficiently and effectively. In this paper, we propose Explanation Iterative Supervision via Saliency-guided Data Augmentation (ESSA) framework. ESSA combines explanation supervision and data augmentation in an iterative loop. Instead of applying the traditional data augmentation methods or generative models, we develop a novel explanation-guided image generator, which has two generation modules specialized for foreground and background due to their different patterns, to build the mapping "from explanation to image". We also propose an innovative algorithm that can alternately and iteratively train the data augmenter and image classifier to do explanation supervision exhaustively. We conduct extensive experiments to evaluate the robustness of the proposed framework on two medical image classification tasks with different training sizes. Experiment results show that ESSA can effectively improve both the predictability and explainability of the model when data samples are limited. Specifically, ESSA outperforms other approaches by approximately 80% and 18%, on average, in terms of explanation IoU and F1.

Bibliography

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[2] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, pages 6541–6549, 2017.

[5] Kaiyi Cao, Lei Bi, Dagan Feng, and Jinman Kim. Improving pet-ct image segmentation via deep multi-modality data augmentation. In *Machine Learning for Medical Image Reconstruction: Third International Workshop, MLMIR 2020,*

Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 3, pages 145–152. Springer, 2020.

[6] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 40–48, 2018.

[7] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021.

[8] Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. Machine learning for medical imaging. *Radiographics*, 37(2):505–515, 2017.

[9] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, et al. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021.

[10] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning, 2022.

[11] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning. *arXiv preprint arXiv:2212.03954*, 2022.

[12] Yuyang Gao, Tong Steven Sun, Guangji Bai, Siyi Gu, Sungsoo Ray Hong, and Zhao Liang. Res: A robust framework for guiding visual explanation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 432–442, 2022.

[13] 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.

[14] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, 47(6):569–576, 2015.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[16] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[17] 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.

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE, 2018.

[21] Henry W Miller. Plan and operation of the health and nutrition examination survey, united states, 1971-1973. *DHEW publication no.(PHS)-Dept. of Health, Education, and Welfare (USA)*, 1973.

[22] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[23] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

[24] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 394–411. Springer, 2020.

[25] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[26] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[27] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and*

Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18, pages 556–564. Springer, 2015.

[28] 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.

[29] Haifeng Shen, Kewen Liao, Zhibin Liao, Job Doornberg, Maoying Qiao, Anton Van Den Hengel, and Johan W Verjans. Human-ai interactive and continuous sensemaking: A case study of image classification using scribble attention maps. In *Extended Abstracts of CHI*, pages 1–8, 2021.

[30] Yujun Shen, Ping Luo, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2018.

[31] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[32] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with*

MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, pages 1–11. Springer, 2018.

[33] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[34] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[35] Zhenghua Xu, Chang Qi, and Guizhi Xu. Semi-supervised attention-guided cyclegan for data augmentation on medical images. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 563–568. IEEE, 2019.

[36] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.