IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 December 2022, accepted 2 January 2023, date of publication 12 January 2023, date of current version 19 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3236575

RESEARCH ARTICLE

Feature Interpretation Using Generative Adversarial Networks (FIGAN): A Framework for Visualizing a CNN's Learned Features

KYLE A. HASENSTAB[1,2], **JUSTIN HUYNH**[2], **SAMIRA MASOUDI**[3], **GUILHERME M. CUNHA**[4], **MICHAEL PAZZANI**[3], **AND ALBERT HSIAO**[2,3]

[1]Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, USA
[2]Department of Radiology, University of California at San Diego, San Diego, CA 92093, USA
[3]Halıcıoğlu Data Science Institute, University of California at San Diego, San Diego, CA 92093, USA
[4]Department of Radiology, University of Washington, Seattle, WA 98195, USA

Corresponding author: Kyle A. Hasenstab (kahasenstab@sdsu.edu)

This work was supported in part by the National Science Foundation under Grant IIS 2026809.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the UC San Diego Human Research Protections Program.

ABSTRACT Convolutional neural networks (CNNs) are increasingly being explored and used for a variety of classification tasks in medical imaging, but current methods for post hoc explainability are limited. Most commonly used methods highlight portions of the input image that contribute to classification. While this provides a form of spatial localization relevant for focal disease processes, it may not be sufficient for co-localized or diffuse disease processes such as pulmonary edema or fibrosis. For the latter, new methods are required to isolate diffuse texture features employed by the CNN where localization alone is ambiguous. We therefore propose a novel strategy for eliciting explainability, called Feature Interpretation using Generative Adversarial Networks (FIGAN), which provides visualization of features used by a CNN for classification or regression. FIGAN uses a conditional generative adversarial network to synthesize images that span the range of a CNN's principal embedded features. We apply FIGAN to two previously developed CNNs and show that the resulting feature interpretations can clarify ambiguities within attention areas highlighted by existing explainability methods. In addition, we perform a series of experiments to study the effect of auxiliary segmentations, training sample size, and image resolution on FIGAN's ability to provide consistent and interpretable synthetic images.

INDEX TERMS Convolutional neural networks, explainable AI, feature interpretation, generative adversarial networks.

I. INTRODUCTION

Artificial intelligence (AI) systems, particularly convolutional neural networks (CNNs), have become increasingly popular in biomedical imaging for their ability to perform and automate a variety of complex imaging tasks, including classification, segmentation, image registration, modality translation, and synthetic image generation [1], [2], [3], [4]. Their performance over traditional approaches is attributed to their increasingly complex architectures, which model highly nonlinear relationships in imaging tasks. However, their architectural complexity also makes it difficult to explain the imaging features used in each image evaluated by the CNN. Understanding the meaning of these underlying features in the context of a CNN task would both improve the transparency of these models for clinical application and potentially uncover new biomarkers for disease. Multiple post hoc methods for CNN explainability have been proposed but provide only partial visibility into the characteristics used by CNNs during inference, limiting the translation of these methods into clinical radiology practice [5].

A. LIMITATIONS OF COMMONLY USED EXPLAINABILITY METHODS

Several methods have been proposed to interpret network decisions [6], [7], [8], the majority of which being attribution

The associate editor coordinating the review of this manuscript and approving it for publication was Kaustubh Raosaheb Patil.

methods, which highlight salient areas on an input image important for CNN prediction. These *static* visualizations each provide *localization* of the portions of the image utilized by the CNN in each inference, though not the specific characteristics within the active region of the image. It is perhaps for this reason that their effectiveness for explainability has been subject of debate [9], [10], [11], [12]. Importantly, current methods are not able to convey relevant texture characteristics within the region of activation. This problem is particularly relevant for diffuse disease processes where large portions of the image are potentially relevant for classification.

For example, Fig. 1 shows several commonly used attribution (i.e. saliency) maps from two CNNs designed to evaluate two diffuse disease processes, pulmonary edema and hepatic fibrosis. The first is a regression CNN designed to infer log NT-pro B-type natriuretic peptide (BNPP) from chest radiographs—a task related to assessing the severity of pulmonary edema [13]. The second is a classification CNN designed to classify liver magnetic resonance images (MRI) as having adequate or suboptimal contrast uptake for cancer screening and surveillance—a task related to assessing the severity of liver fibrosis [14].

In both cases, the attribution maps highlight areas of attention, but it is not clear what specific characteristics within those areas are relevant to diagnosis. Any of the anatomic or texture characteristics within the highlighted regions could be responsible. In the case of the chest radiograph algorithm, attention could be attributed to any of the following—the vasculature, the interstitium, air spaces, or the edge of the cardiomediastinal silhouette. Similarly, for the liver MRI algorithm, attention is centered on the portal or heptic veins, but could be attributed to any of the following—the veins themselves, contrast with hepatic parenchyma, or texture. With current methods, the relative importance of individual characteristics within the localized region is unknown. For CNN explainability, the "what" is likely as important as the "where". Current methods provide localization information but their effectiveness is limited in medical imaging, where often multiple objects are spatially co-localized.

With these limitations in mind, we therefore propose Feature Interpretation using Generative Adversarial Networks (FIGAN), a framework for the *dynamic* visualization of the features used by a CNN. FIGAN generates synthetic images that smoothly change with a CNN's features. The evolution of these synthetic images is then used to elicit the features' meanings. We apply FIGAN to two independently developed source CNNs and show that the resulting feature interpretations can clarify ambiguities within the attention areas highlighted by the approaches in Fig 1. In addition, we perform a series of experiments to study the effect of auxiliary segmentations, training sample size, and image resolution on FIGAN's ability to provide consistent and interpretable synthetic images.

This paper is organized into the following sections. Section II describes existing explainable AI approaches, with an emphasis on medical imaging. Section III introduces the proposed FIGAN framework. Sections IV and V evaluate the performance of FIGAN when applied to two separate source CNNs and Section VI presents experimental results. We discuss the results and conclude in Section VII.

II. EXPLAINABLE AI IN MEDICAL IMAGING

There are several surveys detailing a variety of explainable AI methods for CNNs, specifically applied to medical imaging [6], [7], [8]. We provide a brief summary of commonly used explainable AI methods to motivate the advantages of our proposed FIGAN framework for CNN feature interpretation.

A. ATTRIBUTION METHODS

Most explainable AI algorithms are attribution methods, a broad class of algorithms that highlight salient areas of an input image important for a CNN's prediction. Attribution methods are widely used since they are model-agnostic and often readily available as open-source implementations in a variety of neural network packages.

1) GRADIENT-BASED SALIENCY MAPS

The most popular form of attribution is gradient-based saliency, which computes the gradient of a prediction with respect to the pixels of the input image. Visualization of these gradients can then be interpreted as the influence of input pixels or regions on the CNN's prediction. Many gradient-based approaches have been proposed [15], [16], [17], [18], [19], [20], [21], [22], mostly differing in how the gradient is computed, however, vanilla gradient (VG) maps [15] and gradient weighted class activation mapping (Grad-CAM) [18] are most popular.

2) PERTURBATION-BASED SALIENCY MAPS

Perturbation maps are another form of saliency that visualize the effect of input feature perturbations on a CNN's prediction. Perturbed areas of the input image showing a relatively large effect on the CNN predictions are highlighted for interpretation. Popular approaches include Shapley values [23] and local interpretable model-agnostic explanations (LIME) [24].

Although saliency maps are widely used, studies have shown that the highlighted areas in these images may not have clinically relevant interpretations or be repeatable across aspects of the model training process, such as weight initialization [11], [12]. In addition, assuming a saliency map highlights clinically meaningful anatomical structures, *the interpretation of the highlighted areas may remain unclear, especially if referring to imaging features such as texture or morphology*. Saliency maps are also *local* interpretability methods and do not provide a global understanding of the CNN features used for prediction since the maps are generated at the image level [25].

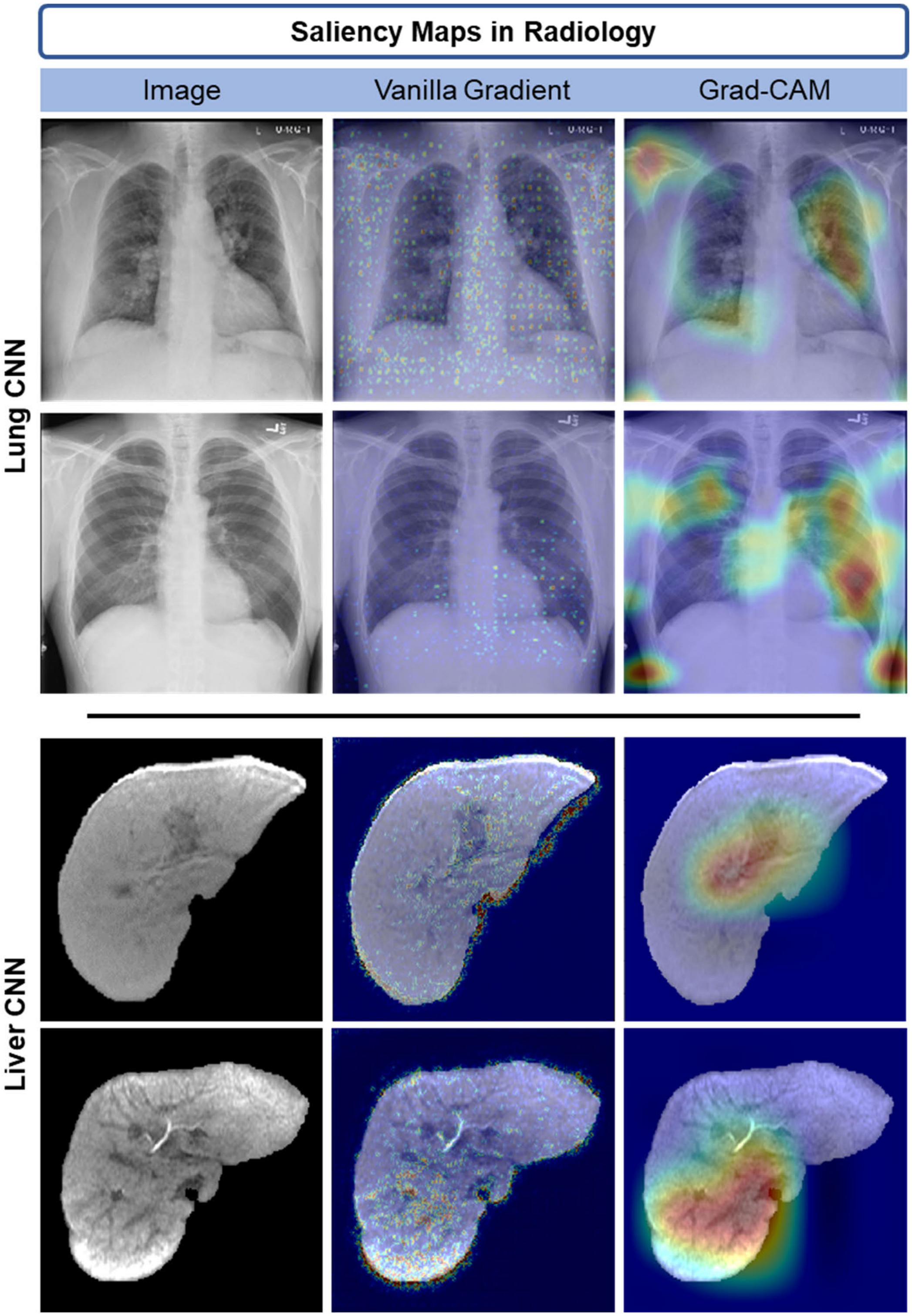


FIGURE 1. Example saliency maps from *a* lung regression CNN designed to infer log BNPP from chest radiographs and *a* liver classification CNN designed to classify liver MRI images as having adequate or suboptimal contrast uptake. Although attribution methods are useful for identifying the location of attention, they do not necessarily uncover the underlying anatomical or pathophysiological nature of this attention.

B. NON-ATTRIBUTION METHODS

1) ATTENTION NETWORKS

In contrast to attribution methods, which are applied after CNN training, attention networks incorporate attention modules directly within the CNN architecture to facilitate explainability of a CNN's predictions while improving CNN performance. Specifically, attention modules have been incorporated into a variety of architectures for classification [26] and segmentation [27]. Attention modules within the intermediate layers of a CNN function act as feature selectors, which enhance features important for prediction and suppress features that are not. These enhanced features maps can then be visualized to determine the types of features important for the CNN's prediction.

However, this approach is not model-agnostic, and similar to attribution methods, provides only a localized understanding of the features used for prediction. Attention networks, and CNNs in general, also contain a large number of feature maps within each layer of the network, many of which showing little or no activation, making exploration of this feature space less tractable.

2) FEATURE ANALYTIC METHODS

Visualization of the CNN feature space using low-dimensional embeddings has also been used to improve understanding of CNN decisions. Feature embeddings are often visualized across the output class distributions and can be used to identify cases or clusters of cases that might be difficult for automated assessment. Methods such as principal components analysis, t-SNE, and UMAP are commonly used [8].

Concept methods [28], [29] enforce networks to learn features representing ''human-friendly'' high-level concepts by incorporating user-defined concepts during the training process in a supervised manner. These methods achieve competitive accuracy with conventional end-to-end models while enabling human-friendly interpretation of the model features.

3) GENERATIVE METHODS

A creative approach to improve CNN explainability is to generate synthetic images that maximize the activation of an output neuron or to augment the appearance of existing images to appear as a different output class. Early activation maximization methods used gradient descent to generate an image that maximizes the activation of a specific neuron, thereby visualizing the imaging features important for a CNN's prediction [15]. However, the synthetic images produced by these methods do not appear realistic, which make their interpretation quite difficult, especially in the context of medical images. Extending this idea, Nguyen et al. [30] proposed an end-to-end activation maximization approach that considerably improved the quality of synthetic images by prepending a deep generator network to the input layer of a given CNN. An alternative approach proposed by Seah et al. [31] involved permuting the CNN feature vectors corresponding to an input image until the CNN prediction changed. A generative network was then used to reconstruct an image from these permuted features. Changes in image appearance resulting from the feature permutation provided insight into the imaging features useful for the CNN decision.

Other generative adversarial network (GAN)-based methods have been proposed to ''disentangle'' a set of latent features that describe high-level concepts (e.g., pose, intensity) across a distribution of images in an unsupervised manner. Following GAN training, synthetic images are generated across the range of disentangled feature values to elicit feature meaning. Methods such as Information Maximizing GAN (InfoGAN) [32], Self-attention Conditional GAN (SCGAN) [33], and Improved Information Maximizing GAN (IInfoGAN) [34], incorporate additional information-theoretic constraints on the GAN minimax loss to learn these disentangled feature representations. Alternatively, Karras et al. [35] proposed StyleGAN, which included an entirely new generator architecture designed to perform the unsupervised task of disentangled feature learning. The quality of StyleGAN-generated images was then improved through changes in architecture and training in a later work [36]. The ability to embed outside images into a StyleGAN latent space was also studied [37].

We build on these generative methods to propose a model agnostic GAN-based explainability method that facilitates *global* interpretation of embedded features important for CNN prediction and show the resulting embedded feature interpretations can clarify the ambiguities observed in commonly used attribution maps.

III. METHOD

This retrospective study is Health Insurance Portability and Accountability Act (HIPPA)-compliant and was approved by the institutional review boards of the participating institutions with waived requirement for written informed consent.

Let C be a previously developed regression or classification CNN trained to map an input image array $\boldsymbol{A}$ to an output vector $\boldsymbol{y}$ (i.e. $C : \boldsymbol{A} \rightarrow \boldsymbol{y}$), and let $\boldsymbol{x}$ represent a vector comprising a subset of features from the intermediate layers of C. We propose a framework that uses a conditional generative CNN to elicit meaning of the features $\boldsymbol{x}$ in the context of the task performed by source network C. Our proposed FIGAN framework, outlined in Fig. 2, is organized into three steps: 1) feature extraction, 2) generative model training, and 3) synthetic image analysis for feature interpretation.

In the feature extraction step, a set of images are propagated through C and features of interest, $\boldsymbol{x}$, are extracted. Since intermediate layers of a network C are often large in dimension, we apply a supervised dimension reduction transformation, f, to $\boldsymbol{x}$, resulting in $\tilde{\boldsymbol{x}} = f(\boldsymbol{x})$. In a second step, a conditional generative CNN G is trained to map $\tilde{\boldsymbol{x}}$ back to the input image space of C, producing a synthetic image $\tilde{\boldsymbol{A}}$ visually representing the embedded features $\tilde{\boldsymbol{x}}$ (i.e. $G : \tilde{\boldsymbol{x}} \rightarrow \tilde{\boldsymbol{A}}$). In a final step, G is used to generate synthetic images across the range of observed $\tilde{\boldsymbol{x}}$ values, which are then assessed using image analysis techniques to provide feature interpretation.

A. FEATURE EXTRACTION AND DIMENSION REDUCTION

1) FEATURE EXTRACTION

Let $\boldsymbol{A}_i$, $i = 1, \ldots, I$, represent a set of I two-dimensional input images to network C, each with dimension $J \times K \times$

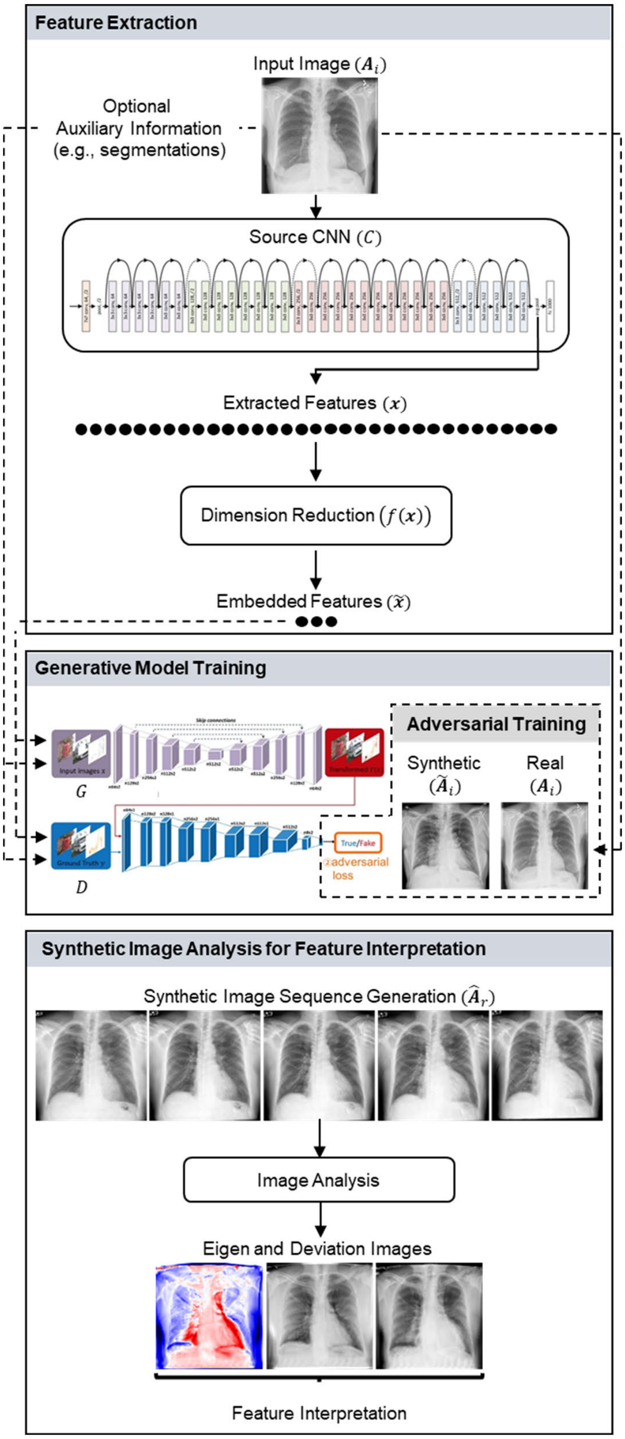

Feature Extraction
Input Image (A_i)
Optional
Auxiliary Information
(e.g., segmentations)
Source CNN (C)
Extracted Features (x)
Dimension Reduction ($f(x)$)
Embedded Features ($\tilde{x}$)
Generative Model Training
G
D
Adversarial Training
Synthetic ($\tilde{A}_i$)
Real (A_i)
True/Fake
adversarial loss
Synthetic Image Analysis for Feature Interpretation
Synthetic Image Sequence Generation ($\hat{A}_r$)
Image Analysis
Eigen and Deviation Images
Feature Interpretation

FIGURE 2. **Proposed framework for Feature Interpretation using Generative Adversarial Networks (FIGAN). FIGAN is organized into three steps: 1) feature extraction, 2) generative model training, and 3) synthetic image analysis for feature interpretation. Images are propagated through a source CNN and features are extracted and reduced in dimension. A conditional generative CNN G is then trained to map the embedded features back to input image space. G is then used to generate a synthetic image sequence across the range of embedded feature values, which are then assessed using image analysis techniques to provide feature interpretation.**

L, where L represents the channel axis. Note that images $\boldsymbol{A}_i$ are not necessarily the same images used to train C. Each $\boldsymbol{A}_i$ are propagated through C and feature vectors of interest $\boldsymbol{x}_i$, each of dimension U, are extracted. Although features in each $\boldsymbol{x}_i$ can conceivably be extracted from any intermediate layer within C, for the remainder of the study, we focus only on the high level features from the final fully-connected layer of C since these are typically the features of interest used to explain a CNN's decisions.

2) DIMENSION REDUCTION

The number of features, U, in the final fully-connected layer is often quite large ($\sim$2048 features) and sparse. We therefore reduce the dimension of this feature space using one of the many dimension reduction techniques available. In this study, we select partial least squares (PLS) [38], [39] for its ability to project $\boldsymbol{x}_i$, using a learned linear transformation f, onto an orthogonal subspace that explains the majority of variability in the source network outcome vector $\boldsymbol{y}_i$. In effect, the majority of information contained within features $\boldsymbol{x}_i$ that is useful for completing the task performed by C is preserved within the low-dimensional embedding, $\tilde{\boldsymbol{x}}_i = f(\boldsymbol{x}_i)$, with dimension $\tilde{U} < U$. Note that other supervised dimension reduction techniques, both linear and nonlinear, may also apply, but the exact nature of the dimension reduction is not the focus of this study.

3) SELECTION OF $\tilde{U}$

Traditionally, PLS selection of $\tilde{U}$, the dimension of the space spanned by $\tilde{\boldsymbol{x}}_i$, is determined empirically by selecting the number of components that minimizes cross validation error for predicting $\boldsymbol{y}_i$ [38]. However, preliminary experiments showed these approaches tend to select a larger number of components explaining minimal variation in $\boldsymbol{y}_i$, and subsequently produced synthetic images with a feature distribution dissimilar to that of real images. We therefore determined an *a priori* selection of $\tilde{U}$ that produces the most "realistic" synthetic images as measured by the Fréchet Inception Distance (FID) criterion [40] in Eq. 1,

$$FID = \left|\left|\boldsymbol{\mu}_{\boldsymbol{x}^{(A)}} - \boldsymbol{\mu}_{\boldsymbol{x}^{(\tilde{A})}}\right|\right| + tr\left[\boldsymbol{\Sigma}_{\boldsymbol{x}^{(\tilde{A})}} + \boldsymbol{\Sigma}_{\boldsymbol{x}^{(A)}} - 2\sqrt{\boldsymbol{\Sigma}^{1/2}_{\boldsymbol{x}^{(\tilde{A})}}\boldsymbol{\Sigma}_{\boldsymbol{x}^{(A)}}\boldsymbol{\Sigma}^{1/2}_{\boldsymbol{x}^{(\tilde{A})}}}\right], \quad (1)$$

where $\boldsymbol{x}^{(A)}, \boldsymbol{x}^{(\tilde{A})}$ represent the features vectors from C corresponding to real image $\boldsymbol{A}$ and synthetic image $\tilde{\boldsymbol{A}}$, and $\boldsymbol{\mu}_{\boldsymbol{x}^{(A)}}, \boldsymbol{\mu}_{\boldsymbol{x}^{(\tilde{A})}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}^{(A)}}, \boldsymbol{\Sigma}_{\boldsymbol{x}^{(\tilde{A})}}$ represent the means and covariances of the source network feature vectors, respectively. A smaller FID indicates the feature distribution of the synthetic images is similar to the feature distribution of real images.

B. CONDITIONAL GENERATIVE MODEL

Following feature reduction to $\tilde{\boldsymbol{x}}_i$ with preselected dimension $\tilde{U}$, we train a conditional generative CNN to predict $\boldsymbol{A}_i$ (i.e. the input images of C) using information in $\tilde{\boldsymbol{x}}_i$ as input. To accomplish this task, we select the pix2pix conditional

GAN framework proposed by Isola et al. [41], given its proven ability to synthetize realistic images from minimally informative generator inputs (e.g., masks) across a variety of applications.

Briefly, the pix2pix GAN comprises a generator network G and discriminator network D, which are adversarially trained through a minimax loss (Fig. 3). The generator G is trained to synthetize "fake" images that are indistinguishable from "real" images when evaluated by the discriminator D. In turn, the discriminator D is trained to determine if an image provided by G is "real" or "fake". We use the architectures for network G and patch network D with receptive field 70×70 as described in [41], with exception to the dropout layers in the generator, which are turned off during test time to enforce deterministic predictions for $\tilde{\boldsymbol{A}}_i$.

Let $\tilde{\boldsymbol{x}}_i$ represent the input to generator G. $\tilde{\boldsymbol{x}}_i$ is a $J \times K \times \tilde{U}$ array where the $\tilde{u}^{th}$ channel ($\tilde{u} = 1, \ldots, \tilde{U}$) is the product $x_{i\tilde{u}} \times \mathbf{1}_{J\times K}$ for the embedded feature $x_{i\tilde{u}}$ and ones matrix $\mathbf{1}_{J\times K}$.The generator network maps $\tilde{\boldsymbol{x}}_i$ back to the input image space of C, producing synthetic image $\tilde{\boldsymbol{A}}_i = G(\tilde{\boldsymbol{x}}_i)$. Input to D is $\tilde{\boldsymbol{A}}_i||\tilde{\boldsymbol{x}}_i$ in the "fake" case or $\boldsymbol{A}_i||\tilde{\boldsymbol{x}}_i$ in the "real" case, where $a||b$ represents concatenation along the channel axis for two arrays a and b.

Networks G and D are adversarially trained using gradient descent in the same manner described in [41] with GAN loss

$$\mathcal{L}_{FIGAN} = \arg\min_{\mathrm{G}}\max_{\mathrm{D}} \mathcal{L}_{cGAN}(G, D) + \lambda\mathcal{L}_{L1}(G), \tag{2}$$

such that $\mathcal{L}_{cGAN}(G, D)$ (Eq. 2) is the average binary cross entropy loss of the discriminator output neurons corresponding to each 70×70 patch and $\mathcal{L}_{L1}(G)$ is the L1-norm between the "real" $\boldsymbol{A}_i$ and "fake" $\tilde{\boldsymbol{A}}_i$ images with some mixing parameter $\lambda > 0$. The result of the training process is a generator G capable of synthesizing realistic images $\tilde{\boldsymbol{A}}_i$ that visually represent the meaning of the feature information contained within $\tilde{\boldsymbol{x}}_i$, or equivalently $\tilde{\boldsymbol{x}}_i$.

Auxiliary information: Auxiliary information (e.g., segmentations, demographics, clinical data) relevant to the task performed by C, but not included in the development of C, may also be incorporated as additional channels in the input array $\tilde{\boldsymbol{x}}_i$ to improve GAN performance. In this study, we incorporate segmentations of the structures of interest to control for the spatial and morphological variation of structures across images in the training set. We later show this regularization approach can improve GAN performance and further facilitate feature interpretation.

C. SYNTHETIC IMAGE ANALYSIS

1) IMAGE INTERPOLATION

Following generator training, we propagate each $\tilde{\boldsymbol{x}}_i$ through G and apply image analysis techniques to the resulting set of synthetic images $\tilde{\boldsymbol{A}}_i$ for visual feature interpretation. Since the feature information $\tilde{\boldsymbol{x}}_i$ is not observed on a regular grid, we first interpolate the synthetic images across the range of feature values (illustrated in Fig. 4).

Let $\tilde{\boldsymbol{a}}_{ijkl}$ represent the pixel in the j^{th} row, k^{th} column, and l^{th} channel of synthetic image $\tilde{\boldsymbol{A}}_i$. For each pixel j, k, l and each feature $\tilde{u}$, we estimate a separate univariate smoothing function $g_{jkl}^{(\tilde{u})}$ (Eq. 3) such that

$$\tilde{\boldsymbol{a}}_{ijkl}^{(\tilde{u})} = g_{jkl}^{(\tilde{u})}(x_{i\tilde{u}}) + \varepsilon_{ijkl}^{(\tilde{u})}, \tag{3}$$

where $\varepsilon_{ijkl}^{(\tilde{u})}$ represents error. Note the superscript $\tilde{u}$ is included in each of the terms to emphasize separate functions for each feature. The function estimate, $\hat{g}_{jkl}^{(\tilde{u})}$, is then determined using one of many univariate smoothing methods available. We found a computationally efficient 1D smoother with Gaussian kernel to suffice (Scipy v1.1.0) [42]. The function estimate $\hat{g}_{jkl}^{(\tilde{u})}$ (Eq. 4) is then used to interpolate the synthetic images across a regular dense grid of values $\tilde{p}_r$, $r = 0, \ldots, R$, that partition the range of $\tilde{\boldsymbol{x}}_{\tilde{u}}$ for each $\tilde{u}$,

$$\hat{a}_{rjkl}^{(\tilde{u})} = \hat{g}_{jkl}^{(\tilde{u})}(\tilde{p}_r), \tag{4}$$

where $\tilde{p}_r$ is the value of the r^{th} grid point and $\hat{a}_{rjkl}^{(\tilde{u})}$ is the interpolated synthetic value at pixel j, k, l for feature $\tilde{u}$ at the r^{th} grid point. The interpolated synthetic 2D image for feature $\tilde{u}$ at the r^{th} grid point is represented by $\hat{A}_r^{(\tilde{u})}$. Note that we assume $\tilde{p}_r$ is the same for each $\tilde{u}$.

2) FEATURE INTERPRETATION

The interpolated synthetic images $\hat{A}_r^{(\tilde{u})}$ can be viewed as a sequence that visualizes the "main effect" of each feature $\tilde{u}$. That is, the evolution of $\hat{A}_r^{(\tilde{u})}$ as r increases, which can be explored using a gif or medical image viewer, elicits the features' interpretations. However, since viewing the sequence in this manner can render subtle changes imperceptible, we further facilitate visualization of the most salient changes in $\hat{A}_r^{(\tilde{u})}$(Eqs. 5 and 6) by identifying the dominant mode of variation in the sequence using the following M-dimensional principal components decomposition,

$$\hat{A}_r^{(\tilde{u})} \approx \boldsymbol{\mu}^{(\tilde{u})} + \sum\nolimits_{m=1}^{M} \xi_{rm}^{(\tilde{u})} \boldsymbol{\phi}_m^{(\tilde{u})}, \tag{5}$$

$$\mu_{jkl}^{(\tilde{u})} = \frac{1}{R}\sum\nolimits_{r=1}^{R} \hat{a}_{rjkl}^{(\tilde{u})}, \tag{6}$$

where $\boldsymbol{\mu}^{(\tilde{u})}$ is a $J \times K \times L$ mean image across r, $\boldsymbol{\phi}_m^{(\tilde{u})}$ is the eigen image for the m^{th} component, and $\xi_{rm}^{(\tilde{u})}$ is the m^{th} principal component score corresponding to $\hat{A}_r^{(\tilde{u})}$ such that $Var\left(\xi_{rm}^{(\tilde{u})}\right) = \lambda_m^{(\tilde{u})}$. In our investigations, visualizing the scaled eigen image corresponding to the leading component $\left(\sqrt{\lambda_1^{(\tilde{u})}}\boldsymbol{\phi}_1^{(\tilde{u})}\right)$ is sufficient for identifying the most salient changes in $\hat{A}_r^{(\tilde{u})}$ for feature interpretation.

D. FIGAN IN PRACTICE

Unlike attribution maps, which are local interpretability methods that are applied to individual images, FIGAN is a global interpretability method, providing a holistic interpretation of the features used for CNN prediction [43]. FIGAN output comprises a single synthetic image sequence and

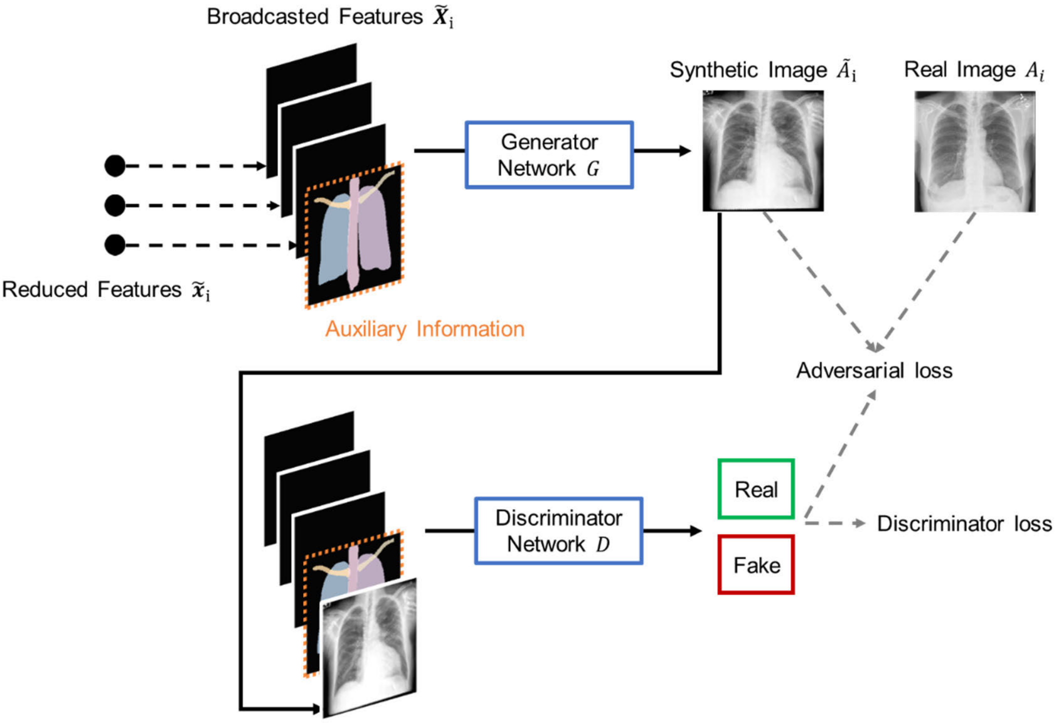

Broadcasted Features $\tilde{X}_i$
Reduced Features $\tilde{x}_i$
Auxiliary Information
Generator Network G
Synthetic Image $\tilde{A}_i$
Real Image A_i
Adversarial loss
Discriminator Network D
Real
Fake
Discriminator loss

FIGURE 3. Conditional generative model. Following the partial least squares dimension reduction, the reduced features $\tilde{x}_i$ are then broadcasted to 2D arrays $\tilde{x}_i$ and concatenated along the channel dimension as input to the GAN. Optional auxiliary information in the form of segmentations can be included as additional channels. The conditional generative model is a pix2pix GAN comprising a generator network and discriminator network. The generator is trained to synthetize "fake" images that are indistinguishable from "real" images when evaluated by the discriminator. The discriminator is trained to determine if an image provided by G is "real" or "fake". Both networks are trained in the conventional adversarial manner.

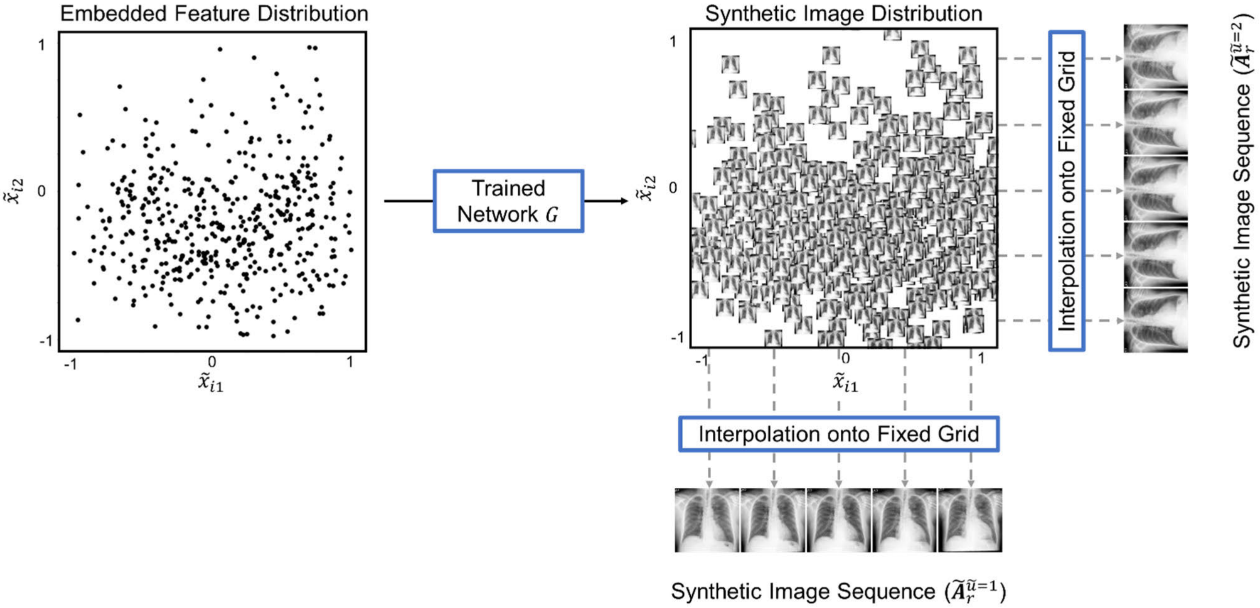

Embedded Feature Distribution
$\tilde{x}_{i2}$
$\tilde{x}_{i1}$
1
0
-1
Trained Network G
Synthetic Image Distribution
Interpolation onto Fixed Grid
Synthetic Image Sequence ($\tilde{A}_r^{\tilde{u}=2}$)
Interpolation onto Fixed Grid
Synthetic Image Sequence ($\tilde{A}_r^{\tilde{u}=1}$)

FIGURE 4. Process to generate synthetic image sequence for each embedded feature. The entire set of broadcasted embedded features are propagated one-by-one through the trained generator network to generate a set of corresponding synthetic images. The synthetic images are then interpolated across a fixed grid for each embedded feature ($\tilde{u}$), producing synthetic image sequences for interpretation.

eigen image for each embedded feature, which are visually interpreted for a global understanding of the source CNN. In practice, attribution maps can be used to identify the location of attention on an individual image, and FIGAN can be used to uncover the underlying anatomical or pathophysiological nature of this attention. Python code for the FIGAN implementation can be found at github.com/khasenst/figan.

IV. APPLICATION TO A LUNG CNN

We study the ability of the proposed FIGAN framework to facilitate global interpretation of features from a CNN developed to infer severity of pulmonary edema from chest radiographs, trained with concurrent serum NT-proBNP measurements (C).

A. SOURCE CNN ARCHITECTURE

The source network C is a ResNet152 regression CNN developed to predict log NT-pro B-type natriuretic peptide (BNPP), a biomarker for pulmonary edema, using 256 × 256 × 1 chest radiographs as input [13]. Note the radiographs were expanded to three channels to accommodate the ResNet152 pretrained weights.

B. IMAGING DATA AND PREPROCESSING

We collected $I = 21,374$ radiographs ($\boldsymbol{A}_i$) and log BNPP values ($\boldsymbol{y}_i$) used to train the source network and an additional 500 for validation. The lungs, spine, and clavicles were then segmented using a 2D U-net CNN independently developed at our institution as part of another ongoing study [44]. Source radiographs were then used to extract the $U = 2048$ features ($\boldsymbol{x}_i$) for each image from the final fully connected layer of C. A PLS linear transformation, f, was then estimated using $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ $(i = 1, \ldots, I)$, and applied to $\boldsymbol{x}_i$, producing the embedded features $\tilde{\boldsymbol{x}}_i$. The embedded features $\tilde{\boldsymbol{x}}_i$ were then broadcasted to the generator input array $\tilde{\boldsymbol{x}}_i$ of dimension $256 \times 256 \times \left(\tilde{U} + 5\right)$, where the five additional input channels represent binary segmentations of the left and right lung, left and right clavicle, and spine as auxiliary information. Radiographs, segmentations, and each embedded feature were then scaled between -1 and 1 for GAN training.

C. GENERATIVE CNN TRAINING

We trained five separate pix2pix GANs with $\tilde{U} = 1, \ldots, 5$, respectively, where $\tilde{U} = 1$ refers to the GAN using only the leading PLS component as input, $\tilde{U} = 2$ uses the first two leading components as input, and so on. For each GAN, generator and discriminator networks were trained with a batch size of one using $\lambda = 100$ and Adam optimizer with learning rate 0.0002 and momentum decay 0.5, which are identical to the hyperparameters specified in [41]. Each GAN was trained for 500,000 steps, and generator weights were saved every 10,000 steps. Total training time for each FIGAN instance required ∼48 hours of processing time on a NVIDIA Titan V graphics card.

To prevent generator overfitting, all inputs were aggressively augmented dynamically during training using random rotations (±25 degrees), horizontal and vertical shifts (±25 pixels), horizontal and vertical flips, and zoom (90%-110%). Although the augmentation process produces images outside of the native data distribution, information on the spatial orientation of the augmented images is implicitly contained within the auxiliary channels. This enabled sufficient regularization of the generator network while avoiding the diminished performance associated with spatial transformations outside the data distribution.

The generator weights producing the minimum FID on the validation set across the number of PLS components $\tilde{U}$ and training steps were used for synthetic image analysis and feature interpretation. Since FIGAN is a global interpretability method that facilitates a holistic understanding of embedded features $\tilde{\boldsymbol{x}}_i$ using the synthetic image sequence [43], FIGAN is not applied to individual images during a testing phase. Therefore, generalizability assessment using a leave-out testing set is not necessary to evaluate FIGAN's utility for explainability.

D. FEATURE INTERPRETATION

Following GAN training, features $\tilde{\boldsymbol{x}}_i$ from the validation set and a randomly selected segmentation were used to generate synthetic images $\tilde{\boldsymbol{A}}_i$. Synthetic images were then used to calculate the interpolated synthetic image sequence $\hat{A}_r^{(\tilde{u})}$, for $r = 0, 1, \ldots, 20$, where $\tilde{p}_r$ equally partitioned the feature range $[-1, 1]$. Principal components analysis was then performed (Eqs. 5 and 6). A cardiothoracic radiologist (A.H.) then reviewed the leading eigen image $\phi_1^{(\tilde{u})}$ and its effect on deviations from the mean image $\boldsymbol{\mu}^{(\tilde{u})} \pm 2\sqrt{\lambda_1^{(\tilde{u})}}\boldsymbol{\phi}_1^{(\tilde{u})}$ for feature interpretation.

E. RESULTS FOR THE LUNG CNN

1) FEATURE EXTRACTION AND SELECTION

FID traces for the validation images (Fig. 5) across the number of training steps indicated $\tilde{U} = 1$ to 3 to be the optimal number of features for realistic image generation. The generator weights for $\tilde{U} = 3$ at training step 470,000 produced the minimum FID and was selected for subsequent analysis. The first three PLS components explained 77.74%, 5.14%, and 1.6% of the variation in log BNPP, respectively.

2) FEATURE INTERPRETATION

Instances of the synthetic image sequence, $\hat{A}_r^{(\tilde{u})}$, and the leading eigen and deviation images for each feature are shown in Figs. 6 and 7. The leading principal components for $\tilde{u} = 1, 2, 3$ explained 85.27%, 92.30%, and 58.98% of variation in the synthetic image sequence $\hat{A}_r^{(\tilde{u})}$, respectively. The eigen and deviation images for both $\tilde{u} = 1$ and $\tilde{u} = 2$ indicate cardiomegaly (enlarged heart) and increased perihilar vascularity as important imaging features for prediction. Feature two ($\tilde{u} = 2$) additionally emphasizes the importance of the chest wall soft tissues. Both $\tilde{u} = 1$ and $\tilde{u} = 2$ show negative correlations with body habitus. Feature 3 ($\tilde{u} = 3$) exhibits similar salient areas to feature 1 but with increased upper lobe cephalization and peripheral vascularity. However note that $\tilde{u} = 3$ only explained 1.6% of variability in log BNPP and must be interpreted with caution.

3) COMPARISON TO VG AND GRAD-CAM

The VG map (Fig. 7) shows diffuse network attention ambiguously scattered throughout the radiograph.

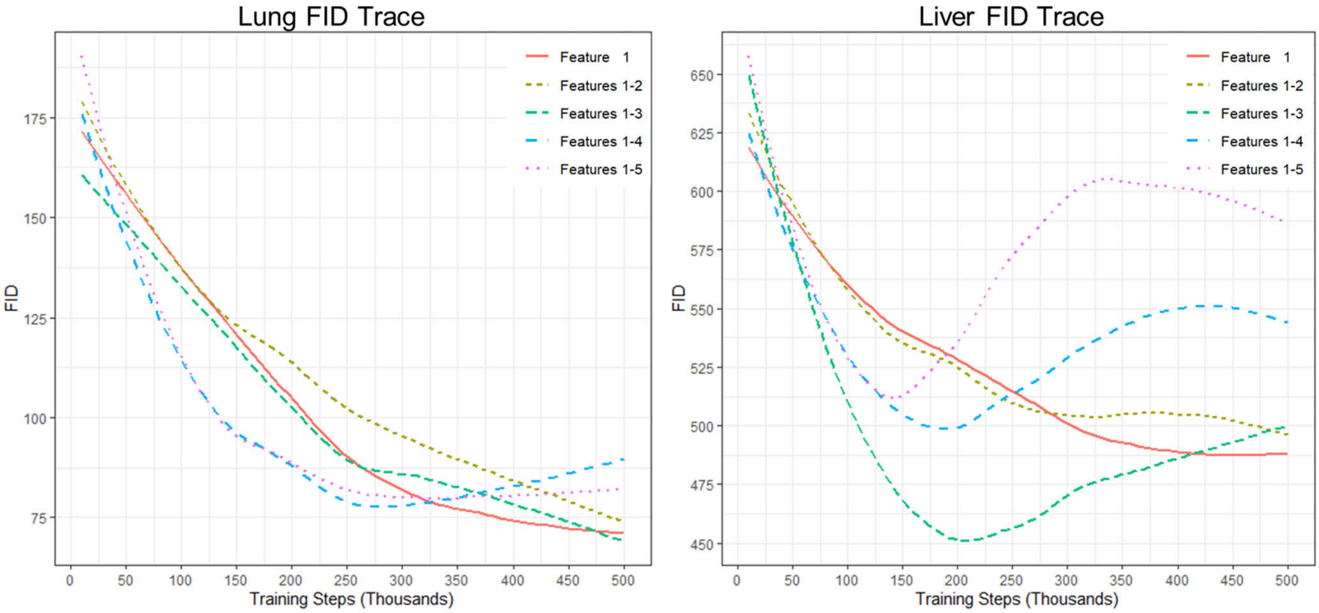


FIGURE 5. Loess smoothed validation FID traces when applying FIGAN to the lung and liver source CNNs. Lung CNN traces indicate FIGAN training using the leading one to three reduced features produces the smallest FID. Liver CNN traces strongly indicate FIGAN training using the first three features produces the smallest FID. Traces for both CNNs exhibit increases in FID when using more than the first three features during FIGAN training, suggesting feature distributions of synthetic images that are dissimilar to that of real images. We selected the first three reduced features for both source CNNs in our implementation.

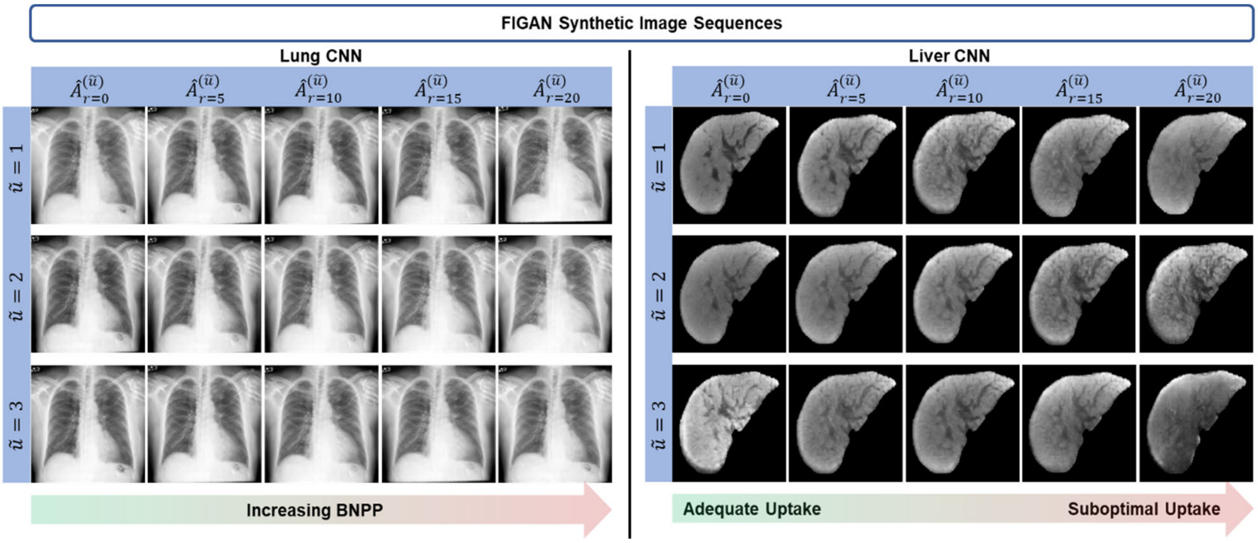

FIGAN Synthetic Image Sequences
Lung CNN
$\hat{A}^{(\tilde{u})}_{r=0}$
$\hat{A}^{(\tilde{u})}_{r=5}$
$\hat{A}^{(\tilde{u})}_{r=10}$
$\hat{A}^{(\tilde{u})}_{r=15}$
$\hat{A}^{(\tilde{u})}_{r=20}$
$\tilde{u} = 1$
$\tilde{u} = 2$
$\tilde{u} = 3$
Increasing BNPP
Liver CNN
$\hat{A}^{(\tilde{u})}_{r=0}$
$\hat{A}^{(\tilde{u})}_{r=5}$
$\hat{A}^{(\tilde{u})}_{r=10}$
$\hat{A}^{(\tilde{u})}_{r=15}$
$\hat{A}^{(\tilde{u})}_{r=20}$
$\tilde{u} = 1$
$\tilde{u} = 2$
$\tilde{u} = 3$
Adequate Uptake
Suboptimal Uptake

FIGURE 6. Synthetic image sequences when applying FIGAN to the lung (left) and liver (right) source CNNs. The chest x-ray synthetic image sequences show variations in heart size and hilar vascular fullness. The liver MRI synthetic image sequences exhibit variations in vessel-liver parenchymal tissue contrast, liver nodularity/texture, and liver signal intensity.

Grad-CAM, shows attention toward the left lung and right hilum, but cannot resolve which co-localized structures in these areas drive the regression of NT-proBNP.

V. APPLICATION TO A LIVER CNN

A. SOURCE CNN ARCHITECTURE

The second source network is a ResNet50 classification CNN with customized feature fusion layer developed to classify hepatobiliary phase liver MRI images as having adequate or suboptimal contrast uptake for cancer screening and surveillance. Input to the liver CNN is a 224 × 224 × 1 masked liver image [14]. Note the liver images were expanded to three channels to accommodate the ResNet50 pretrained weights.

B. IMAGING DATA AND PREPROCESSING

We collected the $I = 826$ liver images ($\boldsymbol{A}_i$), liver segmentations, and uptake classifications ($\boldsymbol{y}_i$) used to train the source network and an additional 375 for validation. $U = 2048$ features ($\boldsymbol{x}_i$) were extracted from the source CNN C for each liver image, and a learned PLS transformation applied to estimate the embedded features $\tilde{\boldsymbol{x}}_i$. The embedded features $\tilde{\boldsymbol{x}}_i$ were then expanded to the generator input array $\tilde{\boldsymbol{x}}_i$

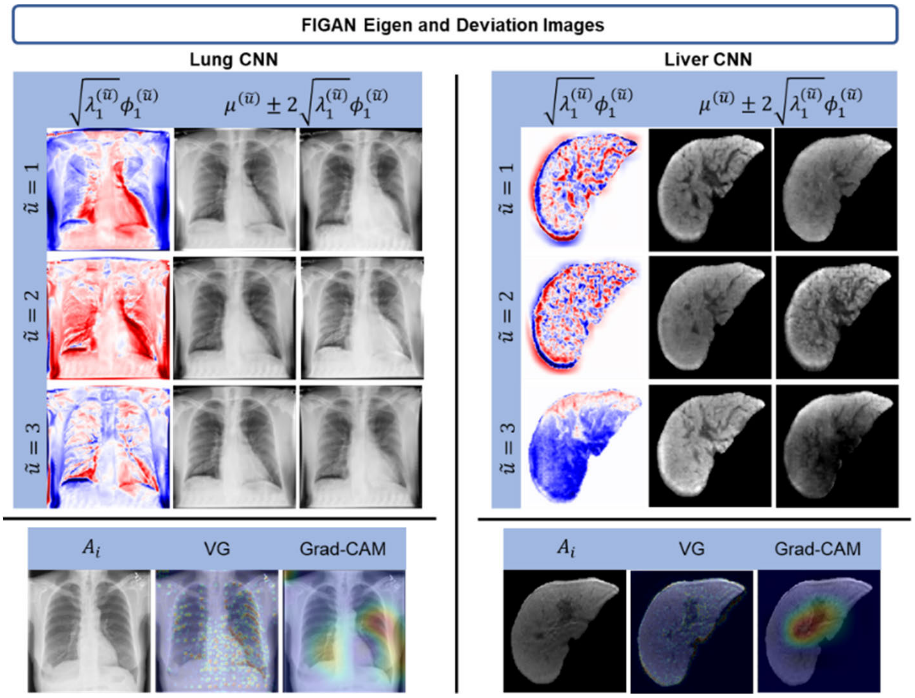

FIGAN Eigen and Deviation Images
Lung CNN
Liver CNN
$\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\mu^{(\tilde{u})} \pm 2\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\tilde{u} = 1$
$\tilde{u} = 2$
$\tilde{u} = 3$
A_i
VG
Grad-CAM

FIGURE 7. Eigen and deviation images for the lung (left) and liver (right) CNNs calculated from the synthetic image sequences generated by FIGAN (top) and corresponding saliency maps and Grad-CAMs (bottom) for comparison. Note the eigen images are scaled by root eigen values to provide an assessment of the variation of principal component scores.

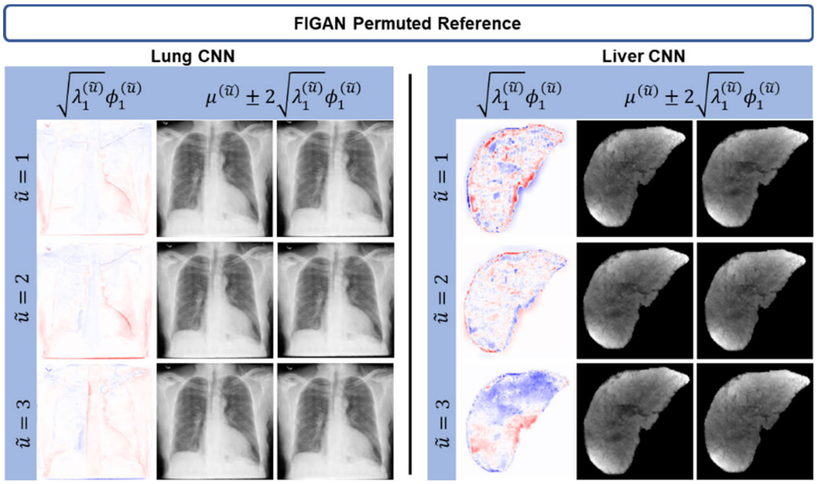

FIGAN Permuted Reference
Lung CNN
Liver CNN
$\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\mu^{(\tilde{u})} \pm 2\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\tilde{u} = 1$
$\tilde{u} = 2$
$\tilde{u} = 3$

FIGURE 8. Eigen and deviation images for the lung (left) and liver (right) CNNs for the FIGAN permuted reference. FIGAN images across features for each CNN indicate little or no meaningful changes across synthetic images, suggesting FIGAN is capturing meaningful signal between features $\tilde{x}_{i\tilde{u}}$ and images A_i.

of dimension 256 × 256 × $\left(\tilde{U}+1\right)$, where the additional input channel represents the liver segmentation as auxiliary information. Liver images, segmentations, and each embedded feature were then scaled between -1 and 1 for GAN training. Note liver images and segmentations were resized to 256 × 256 resolution prior to training.

C. GENERATIVE CNN TRAINING

GAN training was performed in the exact manner as for the FIGAN application to the lung CNN, with identical training parameters, augmentation, and FID criterion for generator weight selection. Total training time for each FIGAN instance required ~36 hours of processing time on a NVIDIA Titan V graphics card.

D. FEATURE INTERPRETATION

FIGAN synthetic images, the leading eigen image $\boldsymbol{\phi}_1^{(\tilde{u})}$, and its effect on deviations from the mean image $\boldsymbol{\mu}^{(\tilde{u})} \pm 2\sqrt{\lambda_1}\boldsymbol{\phi}_1^{(\tilde{u})}$, were then reviewed by an abdominal radiologist (G.M.C) for feature interpretation.

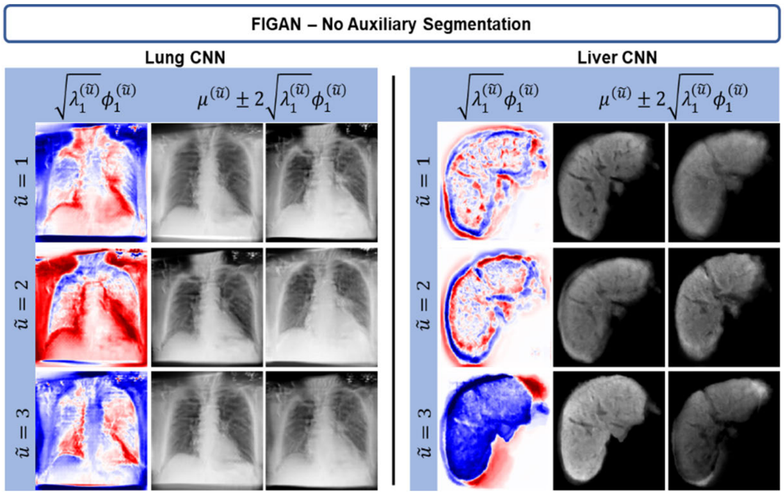

FIGAN – No Auxiliary Segmentation
Lung CNN
Liver CNN
$\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\mu^{(\tilde{u})} \pm 2\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\tilde{u} = 1$
$\tilde{u} = 2$
$\tilde{u} = 3$

FIGURE 9. Eigen and deviation images for the lung (left) and liver (right) CNNs when applying FIGAN without auxiliary segmentations. FIGAN images provided the same feature interpretations but contain boundary artifacts due to changes in anatomical morphology. Subtle features, such as heterogeneous liver texture ($\tilde{u}$= 2), are not as prominent.

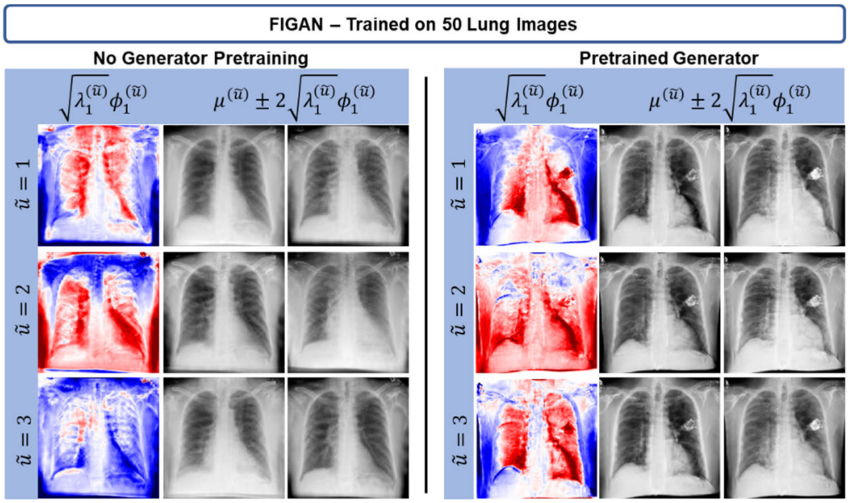

FIGAN – Trained on 50 Lung Images
No Generator Pretraining
Pretrained Generator
$\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\mu^{(\tilde{u})} \pm 2\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\tilde{u} = 1$
$\tilde{u} = 2$
$\tilde{u} = 3$

FIGURE 10. Eigen and deviation images for the lung CNN when training FIGAN on 50 lung images without (left) and with (right) generator pretraining. FIGAN captures signal similar to that found in the complete training runs, with exception to the third lung feature, which showed a decrease in heart size. Generator pretraining produces results consistent with the training runs using the complete datasets.

E. RESULTS FOR THE LIVER CNN

1) FEATURE EXTRACTION AND SELECTION

FID traces for the validation images (Fig. 5) across the number of training steps indicated $\tilde{U} = 3$ to be the optimal number of features for realistic image generation. The first three PLS components explained 35.79%, 19.25%, and 7.5% of the variation in contrast uptake adequacy ($\boldsymbol{y}_i$), respectively. Seven PLS components would have been necessary to exceed 80% of the variation in $\boldsymbol{y}_i$.

2) FEATURE INTERPRETATION

Instances of the synthetic image sequence and the leading eigen and deviation images are shown in Figs. 6 and 7. Leading principal components for $\tilde{u} = 1, 2, 3$ explained 47.97%, 70.58%, 80.56% of variation in the synthetic image sequence $\hat{A}_r^{(\tilde{u})}$, respectively. Note that FIGAN also simulated magnetic inhomogeneity in the synthetic images, which made visualization of subtle features more difficult. We therefore detrended the synthetic image sequence at each grid point r using a 32 × 32 kernel prior to the applying principal components analysis. The original eigen and deviation images without the detrending step can be found in the supplement (Figs. 13-14).

The eigen image for $\tilde{u} = 1$ indicates high vessel-liver contrast as an important predictor for adequate contrast uptake. This is apparent in the deviation images, where changes in the leading score are correlated with the appearance of

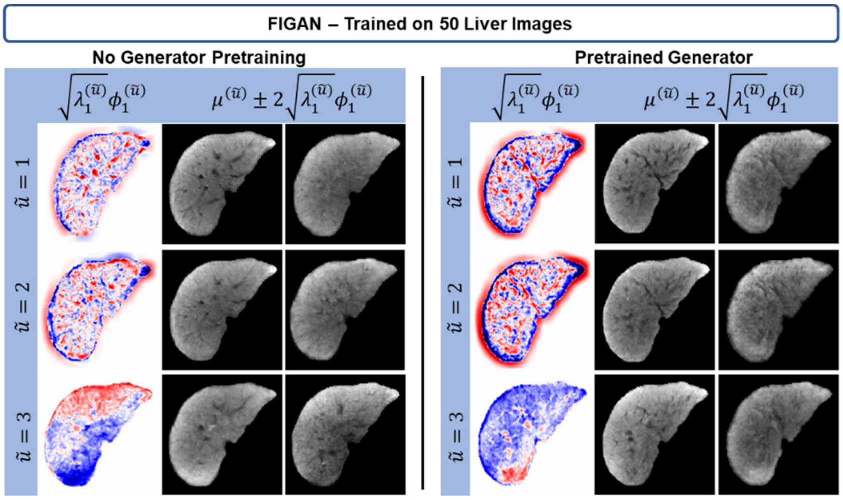

FIGAN – Trained on 50 Liver Images
No Generator Pretraining
Pretrained Generator
$\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\mu^{(\tilde{u})} \pm 2\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\tilde{u} = 1$
$\tilde{u} = 2$
$\tilde{u} = 3$

FIGURE 11. Eigen and deviation images for the liver CNN when training FIGAN on 50 liver images without (left) and with (right) generator pretraining. FIGAN images are similar to the images from the complete training run, with exception to the second liver feature, which did not show obvious changes in texture. The interpretation of texture is preserved when using a pretrained generator.

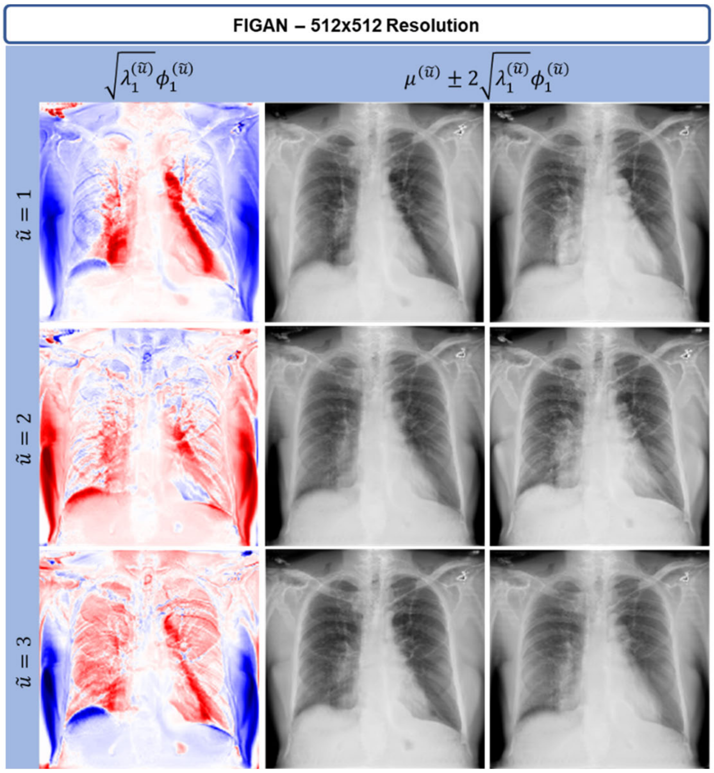

FIGAN – 512x512 Resolution
$\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\mu^{(\tilde{u})} \pm 2\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\tilde{u} = 1$
$\tilde{u} = 2$
$\tilde{u} = 3$

FIGURE 12. Eigen and deviation images for the lung CNN when training FIGAN lung images of 512× 512 resolution. Feature interpretations are similar to the interpretations of the 256×256 implementations but show greater pulmonary vascular detail across the synthetic image sequence.

vessels in the liver. Feature two ($\tilde{u} = 2$) was associated with heterogeneous liver texture, which is negatively associated with adequate contrast uptake, and is often indicative of liver fibrosis. Finally, feature 3 was associated with general liver brightness, and brighter livers typically indicate adequate contrast uptake.

3) COMPARISON TO VG AND GRAD-CAM

Both VG and Grad-CAM (Fig. 7) indicate attention toward the portal vein. However, the VG also shows activations scattered throughout the liver, and the meaning of these activations is unclear. FIGAN images provide additional insight, highlighting the liver texture ($\tilde{u} = 2$) and poor liver

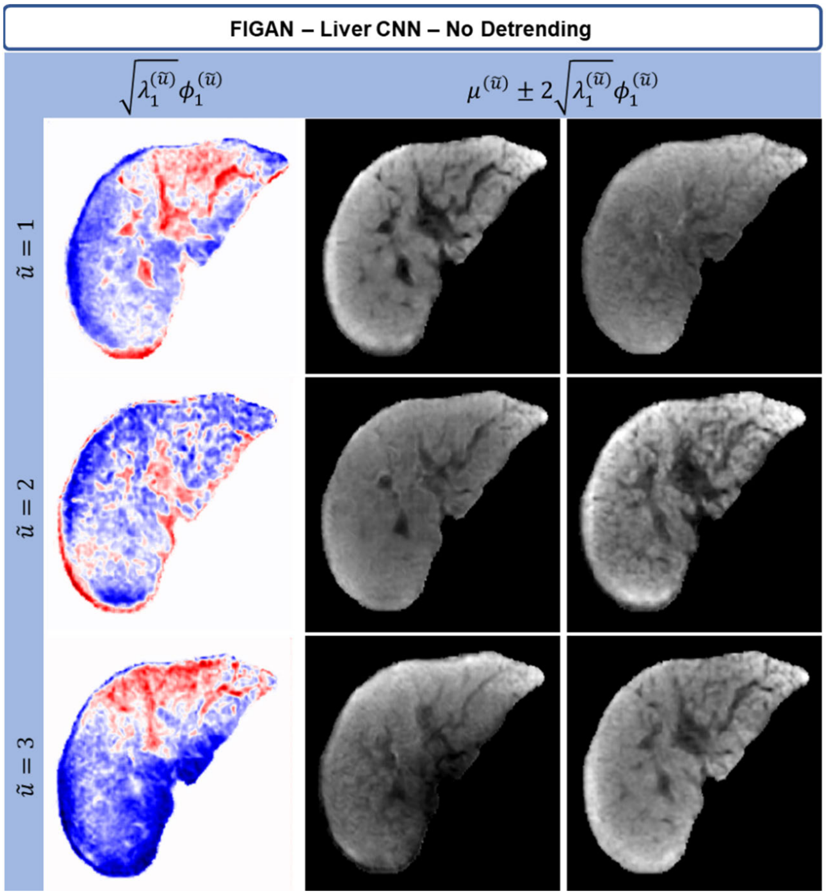

FIGAN – Liver CNN – No Detrending
$\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\mu^{(\tilde{u})} \pm 2\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\tilde{u} = 1$
$\tilde{u} = 2$
$\tilde{u} = 3$

FIGURE 13. Synthetic image sequences when applying FIGAN to the liver source CNN without a detrending step.

enhancement ($\tilde{u} = 3$) to make its decision that this particular liver has suboptimal contrast uptake.

VI. EXPERIMENTS

A. PERMUTED FEATURE REFERENCE

To ensure that the variability across the synthetic image sequence was not attributed to GAN-generated noise, we applied FIGAN to the lung and liver CNNs, but permuted the features $\tilde{\boldsymbol{x}}_{i\tilde{u}}$ at the image (i) level as a reference for comparison (i.e. shuffled $\tilde{\boldsymbol{x}}_{i\tilde{u}}$ over the training set so that training images corresponding to a different set of embedded features). Corresponding scaled eigen and deviation images across features for each CNN on their respective validation sets (Fig. 8) indicate little or no meaningful changes in the synthetic image sequence, suggesting FIGAN is capturing meaningful signal between features $\tilde{\boldsymbol{x}}_{i\tilde{u}}$ and input images $\boldsymbol{A}_i$.

B. NO AUXILIARY INFORMATION

Auxiliary information in the form of anatomical segmentations may not be available. We therefore repeated the analysis without segmentations. Since we initially relied on the segmentations to heavily augment the images during training, in the absence of segmentations, we control spatial orientation by including the plane $z = x + 2y$ as an additional input channel in $\tilde{\boldsymbol{x}}_{i\tilde{u}}$. Eigen and deviation images (Fig. 9) provided the same feature interpretations, but also contained boundary artifacts attributed to changes in anatomical morphology throughout the synthetic image sequence, as observed in the deviation images, especially for the liver. Subtle features, such as heterogeneous texture, were not as prominent.

C. SMALL SAMPLE SIZE

We randomly selected $I = 50$ images from the lung training set and trained FIGAN on these images using the same lung GAN settings, both with and without generator pretraining. The pretrained lung FIGAN generator used the weights from the liver FIGAN application. Similarly, we randomly selected $I = 50$ images from the liver training set and trained FIGAN on these images using the same liver GAN settings, both with and without generator pretraining. The pretrained liver FIGAN generator used the weights from the lung FIGAN application. Since the lung generator required additional channels to accommodate the corresponding anatomical segmentations, we expanded the liver generator input to $256 \times 256 \times \left(\tilde{U} + 5\right)$, repeating the liver mask across the five channels for pretraining.

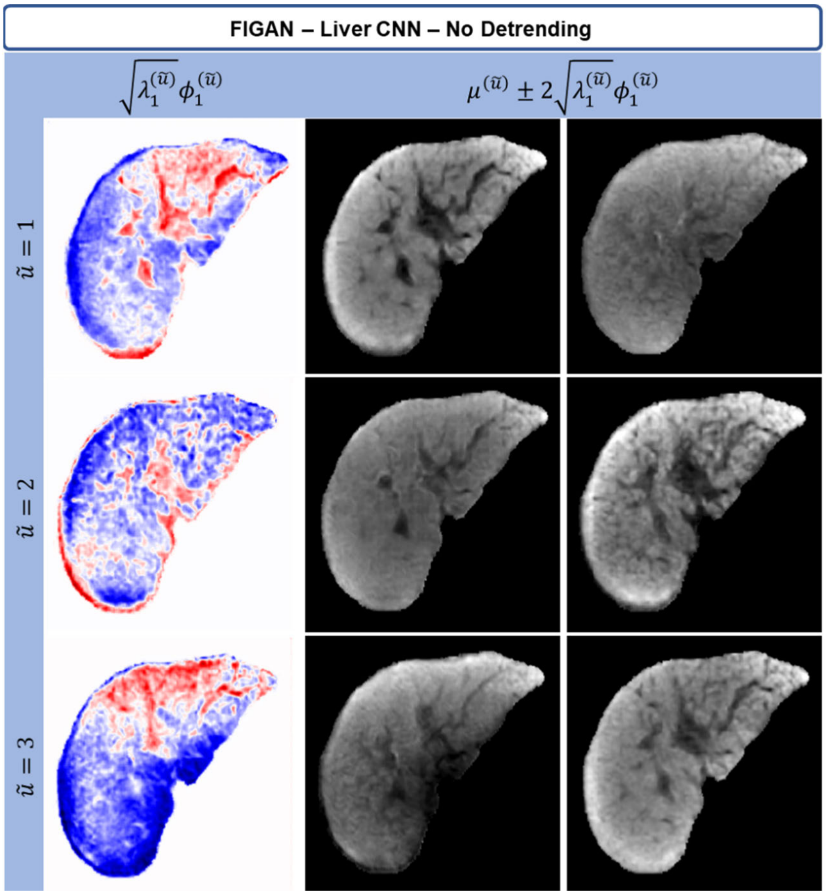

FIGAN – Liver CNN – No Detrending
$\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\mu^{(\tilde{u})} \pm 2\sqrt{\lambda_1^{(\tilde{u})}}\phi_1^{(\tilde{u})}$
$\tilde{u} = 1$
$\tilde{u} = 2$
$\tilde{u} = 3$

FIGURE 14. Eigen and deviation images for the liver CNN without the detrending step.

Eigen and deviation images using the reduced sample with and without generator pretraining are shown in Figs. 10 and 11. For the lung CNN, FIGAN was able to capture signal similar to that found in the complete training runs, with exception to the third lung feature, which showed a decrease in heart size. FIGAN images for the liver CNN were also similar to the complete training run, with exception to the second liver feature, which did not show obvious changes in texture. However, in both cases, generator pretraining produced results consistent with the training runs using the complete datasets, suggesting FIGAN application to smaller datasets is feasible, especially when using pretrained generators.

D. APPLICATION TO HIGHER RESOLUTION IMAGES

We also explored the effects of high-resolution images on FIGAN performance. We applied FIGAN to the lung CNN using input images of size 512 × 512 × 1. The model was trained in the exact manner as done previously, except with a stride of four in the first convolutional layer of the discriminator to maintain a similar receptive field to the 256 × 256 × 1 application. The resulting eigen and deviation images (Fig. 12) showed greater detail in vascular changes across the synthetic image sequence, which is a subtle imaging feature that may be overlooked when using lower resolution networks.

VII. CONCLUSION

In this study, we showed the proposed FIGAN framework can generate synthetic images that elicit meaning of the features used by independently developed regression and classification CNNs. Application to a lung regression CNN revealed cardiomegaly (enlarged heart) and increased perihilar vascularity as the primary imaging features used to predict log BNPP, as well as body habitus, which is a lesser known correlate [45]. Application to a liver classification CNN identified vessel-liver contrast, which is the primary imaging feature used by radiologists to determine contrast uptake adequacy [14]. Although not the primary features for adequacy assessment, FIGAN also identified heterogeneous texture and liver brightness as secondary features of importance, which are also known correlates of liver pathology, and therefore, contrast uptake adequacy. *We also showed that FIGAN images can clarify ambiguities in the interpretation of commonly used explainability visualizations, such as VG maps and Grad-CAMs.*

Prior to GAN training, we used PLS to project features of the source CNN onto a low-dimensional orthogonal basis.

This has the dual purpose of reducing the dimension of the sparse space while mitigating feature confounding. The dimension reduction step was shown to successfully discriminate between features with different interpretations (e.g., vessel-liver contrast vs heterogeneous liver texture), as visualized in the synthetic images. However, we acknowledge that orthogonalization only implies linear independence, and that nonlinear correlations between embedded features may exist. These can appear in the synthetic images as an interaction between the embedded features, where the effect of one feature on synthetic image appearance depends on the value of another feature. For simplicity, we focused only on the ''main effect'' of each embedded feature and reserve the exploration of methods designed to understand the manifestation of feature interactions in synthetic images as a direction for future research.

To facilitate feature interpretation of the GAN-synthesized images, we also proposed an approach for summarizing the salient features of the synthetic images corresponding to the *observed* feature values through interpolation across a regular grid. As an alternative, we considered prediction of synthetic images on a regular grid directly from the generator network. Although useful for the single feature case ($\tilde{U} = 1$), this approach suffers from the curse of dimensionality, exponentially increasing in the number of synthetic images for larger values of $\tilde{U}$, and with no guarantee that all feature combinations are observable for larger dimensions, where the space is increasingly sparse (i.e. images with these feature values may not exist). In addition, following the interpolation step, we summarized the salient features across the synthetic image sequence using the leading principal component. One may alternatively visualize salient areas by subtracting the synthetic images corresponding to the feature extrema, but this approach fails to summarize variability across the entire sequence.

Through a series of experiments, we showed FIGAN identifies legitimate signal relating features $\tilde{\boldsymbol{x}}_i$ and images $\boldsymbol{A}_i$, when compared to a permuted reference. FIGAN is also capable of preserving feature interpretation for small sample sizes ($n = 50$), especially when using generator pretraining, and can be applied to images of higher resolution with minimal changes in architecture. In the absence of segmentation auxiliary information, FIGAN still provided meaningful feature interpretation. However, we recognize this is primarily attributed to the regions of interest being relatively colocalized to the center of the image with minor morphological differences in anatomy across images. Nevertheless, the majority of radiological applications have standardized views of the regions of interest, and colocalization may be further improved using affine or rigid registration against some atlas reference.

In comparison with the existing literature, our approach is most similar to the application and method proposed by Seah et al. [29]. Seah et al. extracted features from a source CNN designed to predict BNP as a marker for congestive heart failure from chest radiographs. These feature vectors were then permuted until the classification of disease was removed. A generative network was then trained to synthesize the appearance of the chest radiographs with the disease removed using the permuted feature vectors as input. Their resulting visualizations of radiographs with and without disease produced interpretations and visualizations consistent with the leading eigen maps found in our study for the lung CNN.

Although similar to the method of Seah et al., there are key differences between our approaches, the primary difference being FIGAN's ability to discriminate between features with different interpretations. This is most apparent in the eigen images corresponding to the liver CNN, where FIGAN was capable of discriminating between vessel-liver contrast, texture, and intensity features through the PLS dimension reduction. In addition, FIGAN provides synthetic image sequences comprising images that smoothly change with a CNN's embedded features. These sequences can help us better understand the functional relationship between the embedded features and the source CNN input images, in contrast to the categorical visualization of images with and without disease in Seah et al.

Our proposed framework also has similarities to GAN-based disentangled feature learning methods [32], [33], [34], [35], [36], [37], which also generate synthetic images across feature values to elicit their conceptual meaning. However, the general goal of these methods is to learn an embedding comprising features that represent high-level concepts across a distribution of images in an unsupervised manner. In contrast, the goal of our proposed method is not to learn such an embedding, but to facilitate interpretation of features from an existing embedding extracted from an independently developed source CNN. That is, we propose FIGAN as a model-agnostic explainability method that facilitates *global* interpretation of embedded features from an existing CNN.

As with other explainability approaches, we recognize FIGAN has limitations. GANs can be challenging to train, requiring more hyperparameters to tune and several ad-hoc strategies to prevent discriminator overfitting and GAN collapse [46]. However, we found the out-of-box implementation of the pix2pix architecture to be sufficient, owing to its robustness, but this may change with other applications. In addition, FIGAN operates independently on features from a source network, and therefore requires the source network to have satisfactory performance to ensure signal between $\tilde{\boldsymbol{x}}_i$ and $\boldsymbol{A}_i$. In contrast, competing methods operate directly on the source CNN architecture and may be more useful for diagnosing poor CNN performance in this case.

In summary, the FIGAN framework can be a useful tool for interpreting the features used by a CNN to perform medical imaging related tasks, complementing current approaches that provide localizations. This is particularly relevant for diffuse diseases and images where anatomic structures are superimposed and closely adjacent (i.e. attribution methods may identify the ''where'' and FIGAN images may identify the ''what''). Holistic interpretations using these methods can improve the transparency and understanding of how classification and regression CNNs make predictions. This improved understanding can further facilitate the translation and acceptance of these algorithms into radiology clinical practice.

SUPPLEMENTAL FIGURES

See Figures 13 and 14.

REFERENCES

[1] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[2] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "GANs for medical image analysis," *Artif. Intell. Med.*, vol. 109, Sep. 2020, Art. no. 101938, doi: 10.1016/j.artmed.2020.101938.

[3] V. Sorin, Y. Barash, E. Konen, and E. Klang, "Creating artificial images for radiology applications using generative adversarial networks (GANs)—A systematic review," *Academic Radiol.*, vol. 27, no. 8, pp. 1175–1185, Aug. 2020, doi: 10.1016/j.acra.2019.12.024.

[4] K. A. Hasenstab, J. Tabalon, N. Yuan, T. Retson, and A. Hsiao, "CNN-based deformable registration facilitates fast and accurate air trapping measurements at inspiratory and expiratory CT," *Radiol., Artif. Intell.*, vol. 4, no. 1, Jan. 2022, Art. no. e210211, doi: 10.1148/ryai.2021210211.

[5] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. V. Tengg-Kobligk, R. M. Summers, and R. Wiest, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiol., Artif. Intell.*, vol. 2, no. 3, May 2020, Art. no. e190043, doi: 10.1148/ryai.2020190043.

[6] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *J. Imag.*, vol. 6, no. 6, p. 52, Jun. 2020, doi: 10.3390/jimaging6060052.

[7] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.

[8] J. D. Fuhrman, N. Gorre, Q. Hu, H. Li, I. E. Naqa, and M. L. Giger, "A review of explainable and interpretable AI with applications in COVID-19 imaging," *Med. Phys.*, vol. 49, no. 1, pp. 1–14, Jan. 2022, doi: 10.1002/mp.15359.

[9] P. J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (un)reliability of saliency methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. Cham, Switzerland: Springer, 2019, pp. 267–280, doi: 10.1007/978-3-030-28954-6_14.

[10] G. Boccignone, V. Cuculo, and A. D'Amelio, "Problems with saliency maps," in *Image Analysis and Processing—ICIAP 2019*. Cham, Switzerland: Springer, 2019, pp. 35–46, doi: 10.1007/978-3-030-30645-8_4.

[11] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S. Q. Truong, C. D. Nguyen, D. Ngo, J. Seekins, F. G. Blankenberg, A. Y. Ng, M. P. Lungren, and P. Rajpurkar, "Benchmarking saliency methods for chest X-ray interpretation," *Nat. Mach. Intell.*, vol. 4, pp. 867–878, 2022, doi: 10.1038/s42256-022-00536-x.

[12] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, J. Adebayo, M. D. Li, and J. Kalpathy-Cramer, "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," *Radiol., Artif. Intell.*, vol. 3, no. 6, Nov. 2021, Art. no. e200267, doi: 10.1148/ryai.2021200267.

[13] J. Huynh, S. Masoudi, A. Noorbakhsh, A. Mahmoodi, K. Hasenstab, M. Pazzani, and A. Hsiao, "Deep learning radiographic assessment of pulmonary edema: Training with serum biomarkers," in *Proc. Med. Imag. Deep Learn.*, 2022, pp. 1–15. [Online]. Available: https://openreview.net/forum?id=NyxXpTbHUCJ

[14] G. M. Cunha, K. A. Hasenstab, A. Higaki, K. Wang, T. Delgado, R. L. Brunsing, A. Schlein, A. Schwartzman, A. Hsiao, C. B. Sirlin, and K. J. Fowler, "Convolutional neural network-automated hepatobiliary phase adequacy evaluation may optimize examination time," *Eur. J. Radiol.*, vol. 124, Mar. 2020, Art. no. 108837, doi: 10.1016/j.ejrad.2020.108837.

[15] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[16] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," 2013, *arXiv:1311.2901*.

[17] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[19] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[20] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3145–3153. [Online]. Available: https://proceedings.mlr.press/v70/shrikumar17a.html

[21] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328, doi: 10.48550/arXiv.1703.01365.

[22] A. Kapishnikov, T. Bolukbasi, F. Viégas, and M. Terry, "XRAI: Better attributions through regions," 2019, *arXiv:1906.02825*.

[23] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4768–4777, doi: 10.5555/3295222.3295230.

[24] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Demonstrations*, 2016, pp. 97–101, doi: 10.18653/v1/N16-3020.

[25] M. Pazzani, S. Soltani, R. Kaufman, S. Qian, and A. Hsiao, "Expert-informed, user-centric explanations for machine learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 11, pp. 12280–12286, doi: 10.1609/aaai.v36i11.21491.

[26] J. Cheng, S. Tian, L. Yu, C. Gao, X. Kang, X. Ma, W. Wu, S. Liu, and H. Lu, "ResGANet: Residual group attention network for medical image classification and segmentation," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102313, doi: 10.1016/j.media.2021.102313.

[27] O. Petit, N. Thome, C. Rambour, and L. Soler, "U-Net transformer: Self and cross attention for medical image segmentation," 2021, *arXiv:2103.06104*.

[28] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2673–2682.

[29] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5338–5348.

[30] A. M. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 3395–3403, doi: 10.48550/arXiv.1605.09304.

[31] J. C. Y. Seah, J. S. N. Tang, A. Kitchen, F. Gaillard, and A. F. Dixon, "Chest radiographs in congestive heart failure: Visualizing neural network learning," *Radiology*, vol. 290, no. 2, pp. 514–522, Feb. 2019, doi: 10.1148/radiol.2018180887.

[32] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2180–2188, doi: 10.48550/arXiv.1606.03657.

[33] X. Li, L. Chen, L. Wang, P. Wu, and W. Tong, "SCGAN: Disentangled representation learning by adding similarity constraint on generative adversarial nets," *IEEE Access*, vol. 7, pp. 147928–147938, 2019, doi: 10.1109/ACCESS.2018.2872695.

[34] Y. Wang, P. Wang, B. Sun, K. He, and L. Huang, "IInfoGAN: Improved information maximizing generative adversarial networks," in *Proc. 5th Int. Conf. Mech., Control Comput. Eng. (ICMCCE)*, Dec. 2020, pp. 1487–1490, doi: 10.1109/ICMCCE51767.2020.00326.

[35] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405, doi: 10.1109/CVPR.2019.00453.

[36] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116, doi: 10.1109/CVPR42600.2020.00813.

[37] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4431–4440, doi: 10.1109/ICCV.2019.00453.

[38] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemometrics Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, Oct. 2001, doi: 10.1016/S0169-7439(01)00155-1.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[40] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6629–6640, doi: 10.48550/arXiv.1706.08500.

[41] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976, doi: 10.1109/CVPR.2017.632.

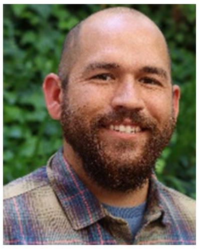
[42] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020, doi: 10.1038/s41592-020-0772-5.

[43] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Munich, Germany, 2022. [Online]. Available: https://www.christophm.github.io/interpretable-ml-book/ and https://www.amazon.com/Interpretable-Machine-Learning-Making-Explainable/dp/B09TMWHVB4

[44] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, vol. 9351. 2015, doi: 10.1007/978-3-319-24574-4_28.

[45] L. B. Daniels, P. Clopton, V. Bhalla, P. Krishnaswamy, R. M. Nowak, J. McCord, J. E. Hollander, P. Duc, T. Omland, A. B. Storrow, W. T. Abraham, A. H. B. Wu, P. G. Steg, A. Westheim, C. W. Knudsen, A. Perez, R. Kazanegra, H. C. Herrmann, P. A. McCullough, and A. S. Maisel, "How obesity affects the cut-points for B-type natriuretic peptide in the diagnosis of acute heart failure," *Amer. Heart J.*, vol. 151, no. 5, pp. 999–1005, May 2006, doi: 10.1016/j.ahj.2005.10.011.

[46] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training GANs," *Proc. 30th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 2234–2242, doi: 10.48550/arXiv.1606.03498.

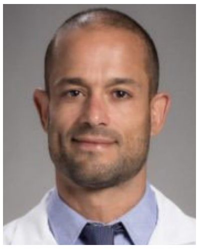


KYLE A. HASENSTAB received the B.S. degree in mathematics and the B.A. degree in quantitative economics from the University of California, Irvine, in 2009, and the M.S. and Ph.D. degrees in statistics from the University of California, Los Angeles, in 2012 and 2015, respectively.

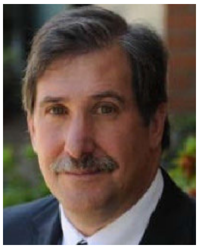
From 2017 to 2020, he was a Postdoctoral Scholar at the Department of Family Medicine and Public Health and the Department of Radiology, University of California, San Diego. He is currently an Assistant Professor with the Department of Mathematics and Statistics, San Diego State University. His research interests include statistical and machine learning methods for image analysis, with applications to medical imaging.

JUSTIN HUYNH received the B.S. degree in computer science from the University of California, San Diego, where he is currently pursuing the M.S. degree in computer science. His research interests include applications of deep learning to radiology and medical imaging.

SAMIRA MASOUDI received the M.S. degree in biomedical engineering from the Sharif University of Technology and the joint Ph.D. degree in electrical engineering from the University of Wyoming in collaboration with the University of Central Florida, in 2019. She is currently a Project Scientist with the Halıcıoğlu Data Science Institute, University of California, San Diego. Her current research interests include the application of deep learning for radiology imaging and image processing.



GUILHERME M. CUNHA received the residency training in radiology and diagnostic imaging from the University of Rio de Janeiro, Brazil. After working as a Clinical Radiologist for a few years in Brazil, he joined as a Senior Research Associate at the Liver Imaging Group, Department of Radiology, University of California, San Diego. He then moved to Seattle, WA, USA, where he is currently an Associate Professor of radiology with the University of Washington. His research interests include oncology imaging, diffuse and focal liver diseases, and application of deep learning to enhance imaging interpretation.



MICHAEL PAZZANI received the Ph.D. degree in computer science from the University of California, Los Angeles. He is currently a Distinguished Scientist with the Halıcıoğlu Data Science Institute, University of California, San Diego. His research interest includes explainable AI for image classification.

ALBERT HSIAO received the dual B.S. degree in engineering and biology from Caltech and the M.D. and Ph.D. degrees in bioengineering with specialization in bioinformatics from the University of California, San Diego (UCSD). He is currently an Associate Professor of radiology and data science with UCSD. He is also a Practicing Cardiothoracic Radiologist with principal areas or research in 4-D flow MRI and the application of imaging deep learning techniques to enhance clinical diagnosis.