DUE: Dynamic Uncertainty-Aware Explanation Supervision via 3D Imputation

Qilong Zhao
Emory University
Atlanta, GA, USA
qzhao31@emory.edu

Yifei Zhang
Emory University
Atlanta, GA, USA
yifei.zhang2@emory.edu

Mengdan Zhu
Emory University
Atlanta, GA, USA
mengdan.zhu@emory.edu

Siyi Gu
Stanford University
Palo Alto, CA, USA
sgu33@stanford.edu

Yuyang Gao
The Home Depot
Atlanta, GA, USA
yuyang_gao@homedepot.com

Xiaofeng Yang
Emory University
Atlanta, GA, USA
xyang43@emory.edu

Liang Zhao*
Emory University
Atlanta, GA, USA
liang.zhao@emory

Abstract

Explanation supervision aims to enhance deep learning models by integrating additional signals to guide the generation of model explanations, showcasing notable improvements in both the predictability and explainability of the model. However, the application of explanation supervision to higher-dimensional data, such as 3D medical images, remains an under-explored domain. Challenges associated with supervising visual explanations in the presence of an additional dimension include: 1) spatial correlation changed, 2) lack of direct 3D annotations, and 3) uncertainty varies across different parts of the explanation. To address these challenges, we propose a Dynamic Uncertainty-aware Explanation supervision (DUE[1]) framework for 3D explanation supervision that ensures uncertainty-aware explanation guidance when dealing with sparsely annotated 3D data with diffusion-based 3D interpolation. Our proposed framework is validated through comprehensive experiments on diverse real-world medical imaging datasets. The results demonstrate the effectiveness of our framework in enhancing the predictability and explainability of deep learning models in the context of medical imaging diagnosis applications.

CCS Concepts

• **Computing methodologies** → **Supervised learning**; **Computer vision**.

*Corresponding author
[1]Code available at: https://github.com/AlexQilong/DUE.

KDD '24, August 25–29, 2024, Barcelona, Spain.

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671641

Keywords

Explanation Supervision, 3D Data, Uncertainty Quantification, Visual Explanation

ACM Reference Format:
Qilong Zhao, Yifei Zhang, Mengdan Zhu, Siyi Gu, Yuyang Gao, Xiaofeng Yang, and Liang Zhao. 2024. DUE: Dynamic Uncertainty-Aware Explanation Supervision via 3D Imputation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3637528.3671641

1 Introduction

While deep learning models demonstrate exceptional performance in computer vision, their "black box" feature raises concerns about their application in high-risk areas. In constructing transparent and trustworthy models, Explainable AI (XAI) has become a critical focus, especially in medical imaging. Existing works have introduced various XAI techniques, such as saliency maps [19, 24, 30, 33], which elucidate the features responsible for model predictions. However, there has been limited attention devoted to the quality of model explanations, including their fidelity to predictions and strategies for enhancing explainability when ground truth explanations are absent or inaccurate. Beyond traditional XAI techniques, an emerging research direction known as *explanation supervision* aims to incorporate additional supervision signals during the learning process of a model, in improving both the generalizability and explainability of deep learning models.

Current methods for explanation supervision have been extensively examined across tabular data, natural language, and two-dimensional (2D) image data. For tabular and natural language data, existing studies [1, 3, 13] have leveraged techniques such as attribution and feature regularization as means of supervision to enhance models. For 2D image data, recent studies [12, 31**?**] focus on jointly optimizing explanation loss and prediction loss, by comparing human explanation annotations with model-generated saliency maps from post-hoc [25, 32] or intrinsic explainers [9, 28] along with

comparing the predicted label and ground-truth label. However, the research into the direct application of explanation supervision techniques to 3D data remains an under-explored domain. This gap is notable given the abundance of 3D data in real-world applications, particularly in the field of medical imaging, where images such as computed tomography (CT) scans and magnetic resonance imaging (MRI) intrinsically present data in a 3D format.

This lack of exploration can be attributed to various fundamental challenges associated with supervising visual explanations when an additional dimension is involved: **1) Spatial correlations change from 2D to 3D.** The shift from 2D to 3D image data induces a significant change in spatial correlation, as the additional dimension introduces depth. Unlike 2D image data, where spatial information is confined to width and height dimensions, and each annotation slice is treated as independently distributed, the intricacy of 3D image data necessitates modifications to both the model architecture and the explanation supervision paradigm. The model must capture spatial features and correlations in the third dimension, potentially causing misalignment between data patterns and the learning capabilities of the paradigm. **2) Gaps between 2D explanation annotations and 3D images.** Humans usually cannot directly delineate a precise curvature surface for 3D complex objects in 3D space. Instead, it is intuitive to label annotations on a few 2D slices (usually with a limited number of them to constrain the labor cost). Hence, such 2D slices cannot fully represent 3D explanation annotations, and leads to a gap when being used for explanation supervision on 3D images. This gap impedes effective explanation-guided learning in fields like medical imaging, where training samples are often limited. **3) The quality of annotations varies in 3D space.** The *curse of dimensionality* tells us that 3D space is "much larger" than 2D space, making it almost impossible to maintain the explanation annotations to have the same quality at any point in 3D space. Therefore, it is very important to identify the quality of explanation annotations to customize the strength of supervision accordingly. However, automatic estimation of the quality of explanation annotation is extremely difficult.

To address the above challenges, we propose a **D**ynamic **U**ncertainty-aware **E**xplanation supervision (**DUE**) framework for 3D explanation supervision that ensures uncertainty-aware explanation guidance when handling sparsely annotated 3D data. The uncertainty-aware guidance is achieved by integrating a 3D explanation loss term, a diffusion-based distance-sensitive interpolation method, and a post hoc weighting module. This module dynamically fine-tunes the weights assigned to the smallest individual units in the interpolated annotation slices based on their respective levels of uncertainty.

Specifically, our main contributions are summarized as follows:

(1) **Proposing a DUE framework for explanation supervision in 3D.** We propose a novel framework that extends the application of explanation supervision to the 3D domain, thereby improving the predictability and explainability of 3D models.
(2) **Introducing a module for uncertainty-aware guidance.** Our approach introduces an uncertainty quantification module that dynamically estimates uncertainty levels. These estimations are utilized to weight the interpolated annotation slices, providing uncertainty-aware explanation guidance.
(3) **Proposing an objective for incomplete and uncertain 3D annotations.** We propose an explanation loss term to handle the challenges introduced by incomplete 3D annotations and noisy interpolation. The computation of this loss term involves interpolated annotation slices and their weights.
(4) **Conducting comprehensive experiments to evaluate our proposed approach.** We conduct comprehensive experiments on various real-world datasets and employ diverse evaluation metrics, demonstrating the effectiveness of our approach. Additionally, we present a thorough analysis of the generated explanations, showing their consistency and informativeness.

The rest of the paper is organized as follows: Section 2 reviews the background and related work, and Section 3 presents the problem formulation. Section 4 describes the proposed DUE framework. The experiments on 2 real-world tasks are provided in Section 5, and the paper concludes with a summary of the research in Section 6

2 Related Work

2.1 Images Interpolation

Image interpolation is a fundamental task of image processing to enlarge an image's size or resolution. Traditional techniques like linear and cubic interpolation [22] have been foundational but often insufficient for capturing the complex details necessary for accurate medical diagnoses. Methods such as Neighbor Mean Interpolation [18], Interpolation by Neighboring Pixels [37], and New Interpolation Expansion [17] take one step forward in improving detail preservation and accuracy by leveraging local pixel relationships more effectively. While works like Marching Cubes [21] and Volume Rendering [8] focus on interpolating 3D data. Recent years have witnessed a shift towards machine learning approaches, particularly with Convolutional Neural Networks (CNNs) making notable advancements in preserving anatomical structures more effectively [23]. Oring et al. [26] proposed a regularization method that molds the latent space into a smooth, locally convex manifold consistent with training images. [27] presents a method for interpolating between generative models of the StyleGAN architecture in a resolution-dependent manner. However, these approaches fail to capture the substantial alterations in spatial correlation present in 3D image data.

2.2 Explanation Supervision

Incorporating human knowledge into explainable models has been a central focus of research in natural language and tabular data, utilizing methods like attribution and feature regularization [1]. XAI-Class [3] utilize highlighted words from the input as additional signals for training the Transformer model. Commonsenseqa [34] proposes to train language models in a multi-task manner, supervised by both labels and rationales [39]. Recently, there has been a growing recognition of the value of visual explanations. A leading method to achieving this involves the use of saliency maps, which identify the input features most influential to a model's predictions [25, 30]. The HAICS framework [31] represents a notable advancement in image classification, utilizing human-generated scribble annotations as the explanation supervision signal. RES [10] introduced an innovative objective designed to accommodate the

inaccurate, incomplete, and inconsistent nature of human annotations. SRDML [5] utilizes saliency to regularize the input gradients across different tasks, enhancing interpretability of task relationships. MAGI [38] proposes a generative model to solve the multiple noisy and incomplete annotations for supervision. Despite this advancement, research on applying explanation supervision to 3D data remains under-explored.

3 Problem Formulation

This section presents problem formulation regarding explanation supervision in the context of image classification. We first define the general paradigm of explanation supervision, and then explore the extension of this paradigm from 2D to 3D data, introducing unique challenges posed by the additional dimensionality. Problem formulation for 3D explanation supervision is presented as follows:

Given a set of inputs $X = \{x_i\}_{i=1}^N$ with class labels $Y = \{y_i\}_{i=1}^N$, and their corresponding human explanation annotation $M = \{m_i\}_{i=1}^N$, where N denotes the training sample size, the model aims to learn the mapping function $f(\cdot)$ for each input image x_i to its class label as $f : x \rightarrow y$ and provide a model explanation via an explainer $g(\cdot)$ as $g : (f, \langle x, y\rangle) \rightarrow m$. The general paradigm of *explanation supervision* can be formulated as the objective function below:

$$\min \sum_{i=1}^{N} (\underbrace{\mathcal{L}_{\text{Pred}}(f(x_i, y_i)}_{\text{prediction loss}} + \lambda \underbrace{\mathcal{L}_{\text{Exp}}(g(f, \langle x_i, y_i\rangle), m_i))}_{\text{explanation loss}} \quad (1)$$

where the first term measures the prediction loss of the model's predicted class labels, the second term measures the explanation loss of the model-generated explanation, and λ is a hyper-parameter used to balance the two loss terms. Here, $\mathcal{L}_{\text{Pred}}$ represents a common prediction loss (e.g., cross-entropy loss), while $\mathcal{L}_{\text{Exp}}$ is tailored to the characteristics of individual datasets.

In our study, we want to extend the above *explanation supervision* paradigm to 3D data as $X = \{x_i \in \mathbb{R}^{C\times D\times H\times W}\}_{i=1}^N$ with class labels $Y = \{y_i\}_{i=1}^N$, and their corresponding binary annotations $M = \{m_i \in \mathbb{R}^{D\times H\times W}\}_{i=1}^N$, where N denotes the training sample size, C denotes the number of channels, D denotes depth, H denotes height, and W denotes width. However, various challenges arise when extending this paradigm to 3D data: **1) Spatial correlation changed.** Spatial correlation has changed significantly from 2D to 3D image data. The complexity of 3D data requires adjustments to capture features and correlations in the third dimension, potentially causing misalignment between data patterns and the paradigm's capacity. **2) Absence of direct 3D labeling.** The absence of direct 3D labeling poses a challenge because human labeling is initially 2D. Manual labeling of volumetric data is costly and leads to sparse labeling, especially in the depth dimension. Limited training samples in domains such as medical imaging exacerbate the challenge of effective generalization. **3) Uncertainty varies across different parts of the explanation.** Medical imaging data are shaped by anatomical structures and lesions, which exhibit a diverse range of shapes. The inconsistent human labeling further contributes to varying distances between consecutive slices along the depth dimension, introducing dynamic uncertainty. Addressing these uncertainties poses a challenge to the model's capacity to provide consistent explanations across different regions of 3D data.

4 Proposed Framework

This section describes the proposed DUE framework in detail. We begin with an overview of the framework and then describe its key components.

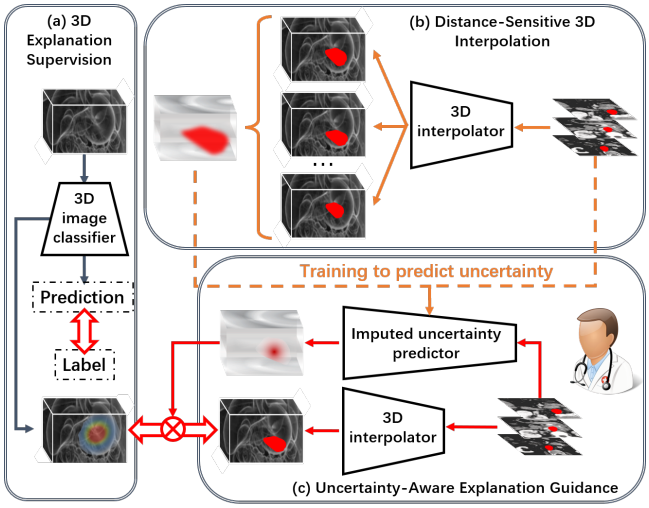

Figure 1: Overview of the DUE framework: (a) presents the 3D explanation supervision, (b) demonstrates the Distance-Sensitive 3D interpolation, and (c) illustrates the Uncertain-Aware Explanation Guidance.

4.1 Framework Overview

The proposed DUE framework, as shown in Figure 1, consists of three modules: *3D explanation supervision*, *distance-sensitive 3D interpolation*, and *uncertain-aware explanation guidance*.

The 3D explanation supervision module aims to enhance both the predictability and explainability of models by introducing 3D explanation supervision. This approach incorporates 3D explanation annotations as additional supervision, alongside traditional prediction label supervision, as depicted in Figure 1(a). While ideal explanation guidance requires comprehensive and precise explanation annotations specifying areas of focus at the pixel (or voxel) level, the sparsity of annotation slices serves as incomplete guides, limiting the effectiveness of explanation supervision. In this scenario, addressing missing slices becomes crucial for effective 3D visual explanation guidance. Traditional interpolation methods such as linear interpolation [22] can fill in missing slices but may introduce bias since they ignore the proportional relationship between uncertainty and the distance from conditional slices. Furthermore, they only provide deterministic predictions, neglecting randomness observed in real-world scenarios [4], making them less suitable compensatory approaches.

To address this issue, we propose a distance-sensitive 3D interpolation module, which comprises a 3D interpolator to better

account for the proportional relationship between uncertainty and the distance from the conditional slices when interpolating missing slices. The 3D interpolator is extended from a conditional diffusion model, detailed in Section 4.2.

Subsequently, we introduce an uncertain-aware explanation guidance module to estimate the uncertainty of the interpolated slices and determine the weights assigned to the interpolated slices based on their associated uncertainty. Specifically, we utilize an imputed uncertainty predictor to accelerate the uncertainty estimation procedure via neural processes, which will be elaborated in Section 4.3. These uncertainties are translated into weights for each voxel contributing to the final 3D explanation, tuning their influence as uncertain-aware explanation guidance, as shown in Figure 1(c). Subsequently, the above two modules are synergistically incorporated into the 3D explanation supervision framework, enabling effective handling of challenges arising from the absence of direct 3D annotations.

Based on the above statement, the proposed DUE framework achieves 3D explanation supervision through the integration of 3D interpolation uncertainty prediction and uncertainty-aware explanation annotation guidance, as depicted in Figure 1. Formally, the overall objective of the DUE framework is expressed as follows:

$$\min_{g,f} \sum_{i=1}^{N_{all}} \mathcal{L}_{\text{Pred}}(f(x_i), y_i) + \lambda \sum_{i=1}^{N_{pos}} \mathcal{L}_{\text{Exp}}\left(g(f, \langle x_i, y_i \rangle), G_{\text{imp}}(m_i) \cdot g_{\text{interp}}(m_i)\right), \tag{2}$$

where prediction loss is computed for all samples, while explanation loss is computed only for samples with manual labels (i.e., positive samples). The function $G_{\text{imp}}(\cdot)$ represents the imputed uncertainty predictor, tasked with adjusting the influence of annotations on the explanation loss based on their imputed uncertainty. Additionally, the 3D interpolator is denoted as $g_{\text{interp}}(\cdot)$.

4.2 Distance-Sensitive 3D Interpolation

To enhance the proportional relationship between uncertainty and distance from conditional slices, we involve a 3D interpolation method that extends a conditional diffusion model to perform distance-sensitive interpolation for the missing slices [35]. When interpolating missing annotation slices, consider a set $\mathcal{A} = \{A_i \in \mathbb{R}^{H\times W}\}_{i=1}^{N}$ representing the annotation slices of an incomplete annotation, where N is the total number of slices. Each annotation slice A_i contains the contour line of the region of interest. Let $D = \{d_1, \in \mathbb{R}\}_{i=1}^{N-1}$ be the corresponding set of distances between adjacent annotation slices, where d_i denotes the distance between slices A_i and A_{i+1}. The goal is to interpolate missing slices $\mathcal{A}_{\text{missing}} = \{A_{\text{missing},i} \in \mathbb{R}^{H\times W}\}_{i=1}^{M}$ within this set based on the set of existing slices $\mathcal{A}$, where M is the number of missing slices.

Given the considerable time consumption and computational complexity associated with 3D interpolation, we adopt a chunking approach to interpolate the entire annotation. Initially, the annotation is segmented into multiple blocks, each comprising two slices. Subsequently, interpolation is applied to fill in the missing slices within each interval, with conditioning on the two slices. The interpolation process adopts an autoregressive approach, allowing each new interpolation conditioned on the preceding interpolations. This generates $j < d_i$ slices iteratively until a cumulative total of d_i slices is attained. By amalgamating all the missing slices within each interval $\sum_{i=1}^{N-1} d_i$, the total number of slices M is obtained.

Conditional diffusion for slice interpolation. Based on the above discussion, consider an annotation block comprising two annotation slices, denoted as A_i and A_{i+1}, utilized as conditions. The diffusion model is tasked with interpolating each intermediate slice A_j, where j ranges from 1 to d_i, as shown in Figure 2. This interpolation process can be formulated as follows:

$$\mathbb{E}_{t,[A_i,A_j,A_{i+1}]\sim p_{\text{annotation}},\epsilon\sim\mathcal{N}(0,I),(m_p,m_f)\sim\text{B}(p_{\text{mask}})} \left[\left|\left|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}A_j + \sqrt{1-\bar{\alpha}_t}\epsilon \,\middle|\, m_p A_i, m_f A_{i+1}, t\right)\right|\right|^2\right], \tag{3}$$

where m_p and m_f are all-zero masks independently sampled from the Bernoulli distribution $\mathcal{B}$ with a probability of $p_{\text{mask}} = 1/2$. The term $p_{\text{annotation}}$ represents the distribution of annotation slices. The function $\epsilon_\theta(A_{j,t}|t)$ estimates ϵ through a time-conditional neural network parameterized by θ, and $\bar{\alpha}_t$ is a parameter that regulates the balance between the contribution of the interpolated slice A_j and the noise term ϵ at step t in the *reverse process* [16].

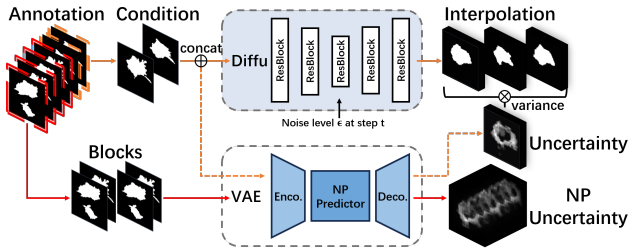

Figure 2: Overview of the *imputed uncertainty predictor* training: A diffusion model is first trained for interpolation and uncertainty generation (solid orange line). Then, a Neural Processes (NP)-based VAE is trained to impute uncertainty for NP representations (dashed orange line). The red line represents the deployment path.

4.3 Uncertainty-Aware Explanation Guidance

After obtaining the interpolated slices and the 3D interpolator, we introduce the uncertain-aware explanation guidance module for generating associated uncertainties. Specifically, we involve an imputed uncertainty predictor to dynamically estimate voxel uncertainty levels, represented as $U = \{u^{(i)} \in \mathbb{R}^{H\times W}\}_{i=1}^{M}$, for intervals within the annotation slices by utilizing spatial correlation and the distribution of annotation features, conditioned on two ground truth slices. After determining the voxel uncertainties of interpolated 3D annotations, we translate the uncertainties into weights, then multiply weights by the interpolated annotations to obtain the final, complete explanatory annotation.

Initially, we use the diffusion model described in Section 4.2 to iteratively perform the interpolation process and compute variance, as illustrated in Algorithm 1. The variance is used as an approximation to uncertainty. The iterative interpolation process and the computation of variance are carried out from lines 4 to 8. The interpolation procedure is detailed from lines 10 to 15, while the function for computing variance spans from lines 16 to 21. This approach

enables a quantitative assessment of uncertainty, offering insights into the randomness inherent in the interpolation results. These estimations are then used to weight the interpolated annotation slices, tuning their influence as explanation guides.

Algorithm 1 Algorithm for Variance Generation

Input: Training data D, number of iterations T
Output: Variance V
Train a diffusion model on D.
for $t = 1$ to T **do**
    Interpolate segments of the annotation using the diffusion model: $A_t \leftarrow$ 3DInterpolation(D).
    Aggregate interpolated segments: $A \leftarrow A \cup A_t$.
end for
Compute the variance of all outcomes: $V \leftarrow$ ComputeVariance(A).
return V.
function 3DInterpolation(D)
    Input: Training data D.
    Output: Interpolated annotation segments A_t.
    Interpolate D using the diffusion model, following the specified loss function.
    return Interpolated segments A_t.
end function
function ComputeVariance(A)
    Input: Aggregate of interpolated annotation segments A.
    Output: Variance V.
    Calculate the variance of outcomes across all A.
    return Calculated variance V.
end function

As in the above statements, we use diffusion models for interpolation and generate the associated uncertainties. However, generating variance with diffusion models is slow and inherently unstable. To address this, we substitute the diffusion model with a Variational Autoencoder (VAE) based on Neural Processes (NP) [36] for more stable and swift variance generation. The VAE model enables a continuous mapping of uncertainty and promptly generates it, alleviating concerns about deployment speed and stability associated with the diffusion model. Additionally, neural processes accommodate varying spacing between annotation slices. To emphasize the purpose of the VAE, we name it *imputed uncertainty predictor* in Figure 1.

The learning process of the imputed uncertainty predictor comprises two stages, as shown in Figure 2. Initially, the diffusion model is trained for interpolation and uncertainty generation (depicted by the solid orange line). Subsequently, we utilize the same condition slices from the diffusion model as conditions for the VAE, with the variances generated by the diffusion model serving as the targets. The VAE is trained to expand the representation of uncertainty for varying spacing (illustrated by the dashed orange line). The reason for this two-stage approach is the complexity associated with training diffusion models and VAEs, which makes simultaneous training a challenging task. During deployment, the VAE is employed directly, as indicated by the red line. Upon obtaining the uncertainty, it is converted into weights using a simple mapping: applying a flipped *Sigmoid* function to the uncertainty, followed by min-max normalization. After the training process, we can generate the NP uncertainty of a given interpolated 3D annotation with the VAE model.

5 Experiments

We present a comprehensive analysis of the experimental results of our proposed framework, focusing on two tasks: pancreatic tumor classification and lung nodule classification. We first introduce our experimental settings, including tasks and datasets, evaluation metrics, and comparative methods. Subsequently, we conduct an extensive quantitative assessment of the model's predictions and explanations and simulate real-world scenarios by restricting training samples. Additionally, we conduct an ablation study, a qualitative assessment featuring case studies, and a sensitivity analysis to provide further insights.

5.1 Experimental Settings

Pancreatic tumor classification: We obtained negative samples (i.e., normal samples) from the Pancreas-CT dataset[2] [29] and positive samples (i.e., abnormal samples) from the Medical Segmentation Decathlon dataset[3], resulting in a dataset of 281 CT scans with tumors and 80 CT scans without tumors. The pancreas region was extracted based on doctors' annotations while retaining the presence of tumors. We kept the original 3D modality for the samples, and extracted the middle slice along the depth dimension for 2D comparative methods, resulting in $128 \times 128 \times 64$ image blocks and 128×128 image slices, respectively. We split the dataset into 30% for training and validation, and 70% for testing, maintaining a balanced data distribution for training while keeping the original ratio for validation and test sets. In our experiments, we only utilized 20 samples during training to simulate a real-world scenario where manual labels are strictly limited.

Lung nodule classification: We obtained both positive samples (i.e., samples with nodules) and negative samples (i.e., non-nodule samples) from the LIDC-IDRI dataset[4] [2]. This dataset includes CT scans collected from 1010 patients, accompanied by annotations provided by four experienced radiologists. Employing a standard 50% consensus consolidation of these annotations, we identified nodule regions as positive samples and the surrounding areas as negative samples, resulting in 2625 positive samples and 68,160 negative samples. We retained the original 3D modality for the samples and extracted the middle slice along the depth dimension for 2D comparative methods. This yielded image blocks of size $128 \times 128 \times 64$ and image slices of size 128×128, respectively. The dataset was first split at patient level, allocating 10% for training, 30% for validation, and 60% for testing. We ensured a balanced distribution of the training data while preserving the original distribution of validation and test sets. To simulate real-world scenarios, we conducted experiments using 20, 50, and 100 training samples.

Evaluation metrics: In assessing the model's performance, we consider both its predictability and explainability. To evaluate its predictive capabilities, common metrics such as prediction accuracy and

[2]Available at: https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT
[3]Available at: http://medicaldecathlon.com/
[4]Available at: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254

Table 1: The experimental results comparing model prediction and generated explanations to various methods for pancreatic tumor and lung nodule classification tasks. Optimal outcomes for each task are highlighted in bold.

Dataset	Method	ROC-AUC (↑)	PR-AUC (↑)	IoU (↑)	Precision (↑)	Recall (↑)	F1 (↑)
Pancreas	HAICS	76.37 ± 10.26	89.35 ± 6.09	33.63 ± 3.95	99.83 ± 0.29	40.29 ± 5.20	54.83 ± 4.40
	GRADIA	76.41 ± 11.62	90.35 ± 6.52	36.40 ± 5.59	99.83 ± 0.29	45.13 ± 7.18	59.84 ± 6.95
	RES	77.90 ± 11.02	90.57 ± 6.52	35.47 ± 8.36	99.66 ± 0.59	42.68 ± 11.31	57.23 ± 11.77
	Baseline	96.83 ± 4.22	99.14 ± 1.12	38.81 ± 16.85	99.83 ± 0.29	49.11 ± 21.55	62.60 ± 20.27
	Baseline$^+$	96.51 ± 5.42	99.17 ± 1.26	36.88 ± 5.39	**100 ± 0**	42.40 ± 6.75	57.33 ± 6.31
	DUE (proposed)	**99.58 ± 0.24**	**99.88 ± 0.07**	**51.20 ± 2.43**	**100 ± 0**	**63.43 ± 4.19**	**75.66 ± 2.76**
LIDC	HAICS	62.29 ± 14.20	8.25 ± 6.63	26.77 ± 4.49	99.74 ± 0.45	47.60 ± 9.36	62.22 ± 9.27
	GRADIA	58.37 ± 12.05	7.72 ± 3.21	29.46 ± 5.30	**100 ± 0**	52.69 ± 8.18	67.27 ± 7.31
	RES	65.84 ± 17.37	14.59 ± 16.28	28.53 ± 8.69	99.48 ± 0.45	50.05 ± 16.83	64.00 ± 16.29
	Baseline	97.60 ± 0.39	81.72 ± 1.94	14.30 ± 6.58	94.32 ± 1.79	30.10 ± 13.76	40.42 ± 14.38
	Baseline$^+$	96.55 ± 1.76	79.81 ± 4.35	31.47 ± 0.41	97.67 ± 2.33	43.84 ± 5.12	56.99 ± 4.18
	DUE (proposed)	**98.58 ± 0.38**	**87.82 ± 1.25**	**33.28 ± 2.69**	90.18 ± 3.82	**64.99 ± 4.39**	**67.66 ± 4.25**

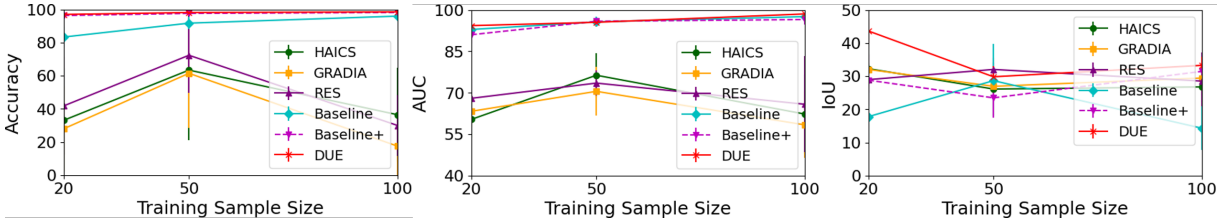

Figure 3: Model performance under varying training sample sizes on the lung nodule classification dataset. (Left) Comparison of test prediction accuracy. (Middle) Comparison of test prediction ROC-AUC. (Right) Comparison of test IoU score.

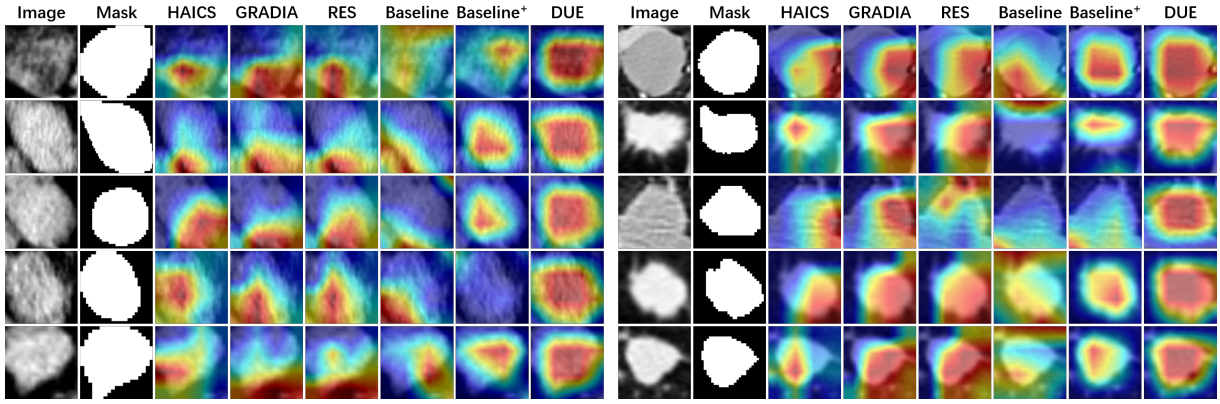

Figure 4: Visualizations display explanations for pancreatic tumor classification (left) and lung nodule classification (right). Human annotations are presented in the Mask columns, while model-generated explanations are depicted using heatmaps overlaid on the original images, highlighting regions of greater importance with warmer color intensities.

the Area Under the Curve of the Receiver Operating Characteristic (ROC-AUC) curves are employed. Due to the pronounced imbalance in the test sets (with a positive-to-negative ratio of approximately 1:26 for LIDC), we additionally employ the Area Under the Curve of the Precision-Recall (PR-AUC) curves as a metric [7]. To assess the quality of the model's explanations, we compare its generated explanations with human annotations. Specifically, we utilize the Intersection over Union (IoU) score, as introduced in [6]. This score is derived from the bit-wise intersection and union operations between the human explanations and the binarized model-generated explanations, providing a measure of overlap between the two inputs. Additionally, we compute pixel-wise precision, recall, and F1 score, offering a comprehensive evaluation of the model-generated explanations.

Comparative methods: We conduct a comparative analysis by evaluating our proposed method against three existing explanation

supervision methods: HAICS [31], GRADIA [11], and RES [10]. Additionally, we include the baseline, which is a 3D model trained solely with the prediction loss. Furthermore, as part of an ablation study, we assess two variants of our proposed method, namely Baseline$^+$ and the standard DUE.

- **HAICS [31]:** This framework employs explanation supervision to train a 2D model, utilizing Binary Cross Entropy (BCE) loss to minimize the discrepancy between the model-generated explanations and provided explanation annotations.
- **GRADIA [11]:** This framework employs explanation supervision to train a 2D model, utilizing L1 loss to minimize the discrepancy between the model-generated explanations and provided explanation annotations.
- **RES [10]:** This framework trains a 2D model with robust explanation supervision, utilizing imputation to bidirectionally minimize the distance between the model's explanations and the provided explanation annotations.
- **Baseline:** The backbone model of our method, which is a bare 3D model trained solely with the prediction loss.
- **Baseline$^+$:** A naive variant of our method, which trains a 3D model using explanation supervision that directly minimizes the distance between the model's explanations and the provided explanation annotations.
- **DUE:** The standard variant of our method, which trains a 3D model using explanation supervision with uncertainty-aware explanation guidance to minimize the distance between the model's explanations and the provided explanation annotations.

Implementation details: The backbone model utilized is a 3D ResNet18 [14]for all 3D methods. For the 2D methods, ResNet18 [15] is employed with customization. This customization involves adjusting the first convolutional layer to possess a kernel size of (7, 7), a stride of (2, 2), and padding of (3, 3). This modification aligns the feature map's view with that of the 3D models, thus mitigating resolution discrepancies and enabling a fair comparison. For RES [10], we use the Gaussian imputation (i.e., RES-G) and set α to 0.001. All explanation supervision methods employ an attention weight λ of 1. All models undergo training for 50 epochs using the Adam optimizer [20] with a learning rate set at 0.001. Model explanations are generated via Grad-CAM [30] and binarized using a threshold of 0.5.

5.2 Performance

Table 1 shows the model prediction and generated explanation performance for the pancreatic tumor and lung nodule classification datasets. The results are obtained from 5 individual runs. The best results for each dataset are highlighted in bold. Overall, our proposed framework DUE outperformed all other comparison methods in both prediction performance and explainability on both datasets. In addition, the huge difference in predictive abilities between 3D and 2D models is due to the underutilized information in 3D data, which further demonstrates the significance of extending the primitive explanation supervision paradigm to 3D models.

For the pancreatic tumor classification task, our proposed DUE consistently yields the best performance on all metrics. Specifically, DUE achieved the highest ROC-AUC and PR-AUC, outperforming the baseline and other comparison methods by 2.75%-23.21%, and 0.71%-10.53% respectively for the predictive capability. The improvements are even more on the model explainability in terms of IoU, recall, and F1 scores, where DUE exceeded baseline and other comparison methods by 12.39%-17.57%, 14.32%-23.14%, and 13.06%-20.83% respectively. Additionally, precision is at a perfect 100% for DUE.

Furthermore, for the lung nodule classification task, the DUE shows stronger predictive capabilities, where the ROC-AUC and PR-AUC scores are 0.98%-40.21%, and 6.1%-80.1% higher for the DUE compared to baseline and other comparison methods. While precision has a decrease of 4.14% compared to the baseline model, this is offset by a significant improvement in recall with an increase of 34.89%. This balance between precision and recall is reflected in the IoU, with an improvement of 18.98% over the baseline model, as well as an overall increase in the F1 score by 27.24%, indicating enhancing model explainability robustly.

Subsequently, we investigate the performance of the DUE framework to strengthen the generalization capabilities under various training sample sizes. We study three training sample sizes of 20, 50, and 100 using the lung nodule classification dataset. As depicted in Figure 3, we present the results of the prediction accuracy, AUC, and IoU score of each method concerning the training sample size. Each data point represents the mean values of five independent runs and the corresponding error bar stands for the standard deviation. In general, our DUE framework outperformed all other comparison methods, especially in the predictive capabilities, demonstrating the effectiveness of our proposed framework. Specifically, DUE can improve the prediction accuracy and AUC by 40% and 30% respectively on average against other comparison methods. Interestingly, DUE performs the best in the IoU score when the sample size is as small as 20. As the training sample size increases, the IoU score decreases and then stabilizes while still outperforming other comparison methods, indicating a transition from potential overfitting to better generalization.

5.3 Qualitative Analysis of Model Explanation

Here, we present a case study examining the comparison of model-generated explanations for pancreatic tumor and lung nodule classification datasets, as depicted in Figure 4. The model-generated explanations are showcased through heatmaps overlaid on the original image samples, with increased emphasis on areas exhibiting warmer colors to denote higher importance.

Pancreatic tumor classification: In the context of pancreatic tumor classification, illustrated in the left portion of Figure 4, we chose five samples of model-generated explanations from all models. The visualization reveals that explanations produced by models employing the proposed DUE framework exhibit superior performance in both accuracy and alignment with human annotations compared to the comparative methods. Notably, the explanations generated by the baseline model largely fail to concentrate on the annotated region, indicating its inherent lack of explainability. The 2D methods exhibit a recurring focus on the lower region, consistently focusing on the lower region, while the ground truth is distributed in the middle of the view. Generally, Baseline+ manages to anchor on the correct region but with a narrow scope deviating from the central tumor area. DUE achieves optimal visualization by

consistently focusing on the central area, with margin adjustments to accurately encompass the tumor region.

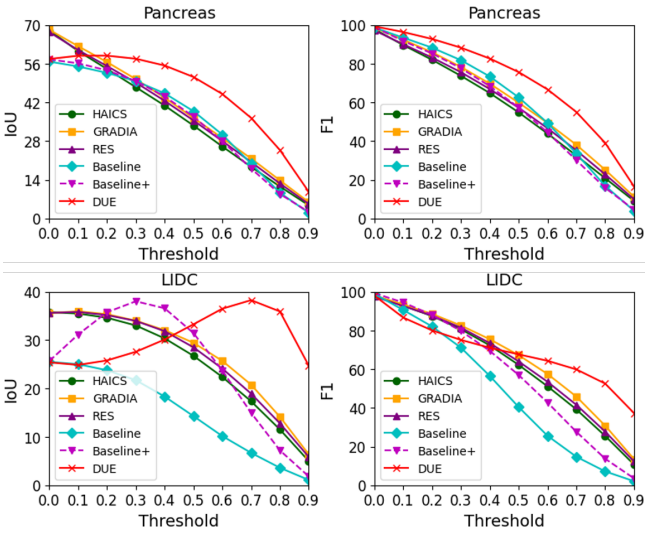

Figure 5: IoU and F1 scores for model explanations at different thresholds on pancreatic tumor and lung nodule classification datasets.

Lung nodule classification: In the right segment of Figure 4, five selected samples of model-generated explanations across all models are presented for lung nodule classification. The contours of the lung nodules exhibit greater variability, with DUE consistently delivering the most effective visualization results. Specifically, the baseline model continues to favor areas outside the annotated region, indicating a questionable basis for its predictions. The 2D methods consistently focus on the lower right part, capturing a significant portion of the ground truth but also encompassing large non-relevant areas. Baseline$^+$ achieves precise concentration; however, it remains incomplete and repeats the same bias observed in the baseline for row 3. In contrast, DUE tends to encompass the annotated region as extensively as possible while also adjusting its margins to prevent overreaching. Furthermore, DUE exhibits strong confidence in its attention area, as indicated by its heatmap being predominantly red with a narrow margin of yellow.

5.4 Quantitative Analysis of Model Explanation

In addition to the qualitative analysis of model explanations presented in Section 5.3, which is conducted with a limited number of samples, we provide IoU and explanation F1 scores calculated using various thresholds for the model explanations (given their continuous nature) for a more comprehensive study. We vary the threshold of attention values in the model explanations and recalculate both IoU and F1 scores. A low threshold (e.g., 0.1) encompasses regions with minimal influence on the model's prediction results when calculating the degree of overlap with human annotations. Conversely, a high threshold (e.g., 0.9) focuses solely on regions with significant influence on the model's prediction results when determining the degree of overlap with human annotations. Using Figure 4 to better illustrate, the yellow-green areas in the model

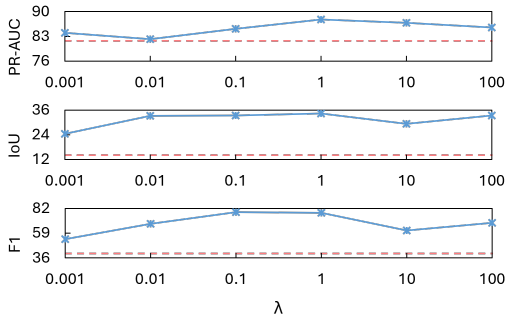

Figure 6: The sensitivity study of λ in our framework, DUE, on the lung nodule classification dataset. The red dashed line denotes the baseline model's performance.

explanation may have an attention value of 0.1, while the red areas in the model explanation may have an attention value of 0.9. Figure 5 displays the IoU and F1 scores achieved by each model across varying attention value thresholds from 0 to 0.9 (where attention values range from 0 to 1) for both Pancreas and LIDC datasets. In general, DUE has the highest IoU and F1 at different thresholds and consistently outperforms others, evident in the red line consistently occupying a higher position in most cases. This superiority becomes more pronounced as the threshold increases, suggesting that DUE exhibits greater confidence in identifying important areas. This observation aligns with our findings in Section 5.3.

5.5 Sensitivity Analysis of Hyper-Parameter

We evaluate the robustness of our proposed DUE framework to various changes in hyper-parameter λ as shown in Equation 2, which determines the balance between the predictive loss and explanation loss. Figure 6 shows that the PR-AUC is relatively stable across the range of λ values, with a slight increase as λ approaches 1. The IoU metric shows an initial increase with small values from 0.001 to 0.01 and the explanation F1 score increases notably as λ moves from 0.001 to 0.1, and then gradually declines, indicating a moderate emphasis on explanation loss leads to a better performance. Generally, our model outperformed the baseline model by a significant margin in both prediction accuracy as well as explainability. The optimal range for λ is between 0.1 and 1, with the peak at 1, which suggests the overall performance is the best when the prediction loss and explanation loss are balanced.

6 Conclusion

This paper introduces the Dynamic Uncertainty-aware Explanation supervision (DUE) framework, addressing challenges in applying explanation supervision to 3D medical images. Our approach overcomes issues such as altered spatial correlations, sparse 3D annotations, and varying uncertainty by introducing a diffusion-based 3D interpolation method with uncertainty-aware guidance. Through extensive experiments on diverse medical imaging datasets, we show that the DUE framework significantly improves the predictability and explainability of deep learning models in medical diagnosis, showcasing its potential to advance Explainable AI (XAI) in healthcare diagnostics.

Acknowledgments

This work was supported by the NSF Grant No. 2432418, No. 2414115, No. 2007716, No. 2007976, No. 1942594, No. 1907805, No. 2318831, Cisco Faculty Research Award, Amazon Research Award.

References

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

[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. 2011. 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 (2011), 915–931.

[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. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[4] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. 2017. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252* (2017).

[5] Guangji Bai and Liang Zhao. 2022. Saliency-regularized deep multi-task learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 15–25.

[6] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6541–6549.

[7] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. 233–240.

[8] Robert A Drebin, Loren Carpenter, and Pat Hanrahan. 1988. Volume rendering. *ACM Siggraph Computer Graphics* 22, 4 (1988), 65–74.

[9] Alex A Freitas. 2014. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter* 15, 1 (2014), 1–10.

[10] Yuyang Gao, Tong Steven Sun, Guangji Bai, Siyi Gu, Sungsoo Ray Hong, and Liang Zhao. 2022. RES: A Robust Framework for Guiding Visual Explanation. In *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

[11] Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. 2022. Aligning eyes between humans and deep neural network through interactive attention alignment. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[12] Siyi Gu, Yifei Zhang, Yuyang Gao, Xiaofeng Yang, and Liang Zhao. 2023. Essa: Explanation iterative supervision via saliency-guided data augmentation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 567–576.

[13] Daniel Hajialigol, Hanwen Liu, and Xuan Wang. 2023. XAI-CLASS: Explanation-Enhanced Text Classification with Extremely Weak Supervision. *arXiv preprint arXiv:2311.00189* (2023).

[14] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6546–6555.

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

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[17] Wien Hong and Tung-Shou Chen. 2011. Reversible data embedding for high quality images using interpolation and reference pixel distribution mechanism. *Journal of Visual Communication and Image Representation* 22, 2 (2011), 131–140.

[18] Fangjun Huang, Xiaochao Qu, Hyoung Joong Kim, and Jiwu Huang. 2015. Reversible data hiding in JPEG images. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 9 (2015), 1610–1621.

[19] Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv preprint arXiv:1611.07270* (2016).

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

[21] William E Lorensen and Harvey E Cline. 1998. Marching cubes: A high resolution 3D surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*. 347–353.

[22] Einar Maeland. 1988. On the comparison of interpolation methods. *IEEE transactions on medical imaging* 7, 3 (1988), 213–217.

[23] Jiageng Mao, Xiaogang Wang, and Hongsheng Li. 2019. Interpolated convolutional networks for 3d point cloud understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1578–1587.

[24] Vivek Miglani, Aobo Yang, Aram H Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Using captum to explain generative language models. *arXiv preprint arXiv:2312.05491* (2023).

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

[26] Alon Oring. 2021. *Autoencoder image interpolation by shaping the latent space*. Ph. D. Dissertation. Reichman University (Israel).

[27] Justin NM Pinkney and Doron Adler. 2020. Resolution dependent gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334* (2020).

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).

[29] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. 2015. 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*. Springer, 556–564.

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

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

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

[33] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[34] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937* (2018).

[35] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. 2022. MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation. In *(NeurIPS) Advances in Neural Information Processing Systems*. https://arxiv.org/abs/2205.09853

[36] Xi Ye and Guillaume-Alexandre Bilodeau. 2023. A Unified Model for Continuous Conditional Video Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 3603–3612.

[37] Xinpeng Zhang. 2011. Reversible data hiding in encrypted image. *IEEE signal processing letters* 18, 4 (2011), 255–258.

[38] Yifei Zhang, Siyi Gu, Yuyang Gao, Bo Pan, Xiaofeng Yang, and Liang Zhao. 2023. MAGI: Multi-Annotated Explanation-Guided Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1977–1987.

[39] Yifei Zhang, Bo Pan, Chen Ling, Yuntong Hu, and Liang Zhao. 2024. ELAD: Explanation-Guided Large Language Models Active Distillation. *arXiv preprint arXiv:2402.13098* (2024).