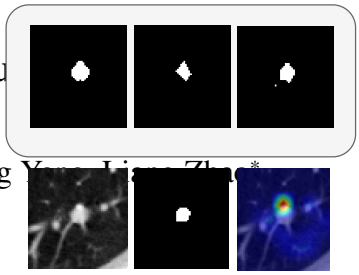
MAGI: Multi-Annotated Explanation-Gu



Yifei Zhang, Siyi Gu, Yuyang Gao, Bo Pan, Xiaofeng Yang, Liang Zhao*
Emory University
Atlanta, GA, United States
{yifei.zhang2, carrie.gu, yuyang.gao, bo.pan, xyang43, liang.zhao}@emory.edu

Abstract

Explanation supervision is a technique in which the model is guided by human-generated explanations during training. This technique aims to improve the predictability of the model by incorporating human understanding of the prediction process into the training phase. This is a challenging task since it relies on the accuracy of human annotation labels. To obtain high-quality explanation annotations, using multiple annotations to do explanation supervision is a reasonable method. However, how to use multiple annotations to improve accuracy is particularly challenging due to the following: 1) The noisiness of annotations from different annotators; 2) The lack of pre-given information about the corresponding relationship between annotations and annotators; 3) Missing annotations since some images are not labeled by all annotators. To solve these challenges, we propose a Multi-annotated explanation-guided learning (MAGI) framework to do explanation supervision with comprehensive and high-quality generated annotations. We first propose a novel generative model to generate annotations from all annotators and infer them using a newly proposed variational inference-based technique by learning the characteristics of each annotator. We also incorporate an alignment mechanism into the generative model to infer the correspondence between annotations and annotators in the training process. Extensive experiments on two datasets from the medical imaging domain demonstrate the effectiveness of our proposed framework in handling noisy annotations while obtaining superior prediction performance compared with previous SOTA.

1. Introduction

In medical imaging, deep learning models are often used to provide predictions for tasks such as diagnosing diseases and they have shown exceptional performance [11, 40, 42, 33, 16, 44, 45]. However, these models can be seen as a "black box", making it difficult for clinicians to understand why a particular diagnosis was made [1]. An important line of research to tackle such limitations is to provide post-hoc explanations of the model behavior [12, 27, 31]. A popular way to achieve this in computer vision is through the use of attention maps, which highlight specific regions of an image that are most relevant for the model's prediction [31]. Despite the attention mechanism being an increasingly prominent component in deep neural networks (DNNs) as a means of explanation, little research has been done to analyze whether attention is trustworthy and how to further improve it until recently.

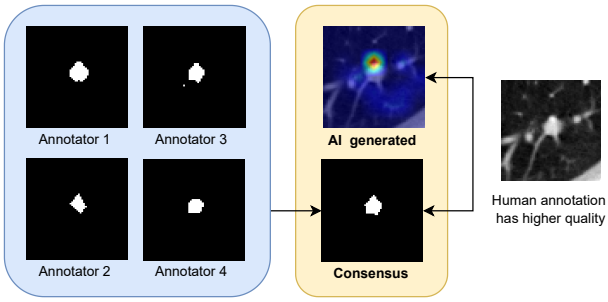


Figure 1: An example showing the challenges present in explanation supervision: (a) Model-generated explanations still lack accuracy compared with human annotation. (b) The model performance heavily relies on the quality of human explanation labels, which is often inconsistent and incomplete among different annotators.

Recently, explanation supervision, a technique that jointly optimizes prediction loss and explanation loss between ground-truth annotations and model-generated explanations, has started to show promising effects in improving both the predictability and interpretability of deep neural networks [9, 13, 34]. By injecting explanation as a supervision signal, models aim to explain the decision-making process at either the local or global level. Global explanations provide a general understanding of how the model works across the entire dataset, while local explanation techniques are applied for each input data and are therefore more com-

*Corresponding author.

monly used [12]. Depending on the type of data, local explanation techniques can be further summarized into several categories: 1) visual explanation, 2) rationale attention, and 3) feature attribution alignment [12]. While most research focuses on text and tabular data, supervising explanations on image data is relatively under-explored [36, 29, 28].

Human annotations are often subjective, which gives rise to the need for multiple annotations on a single image from different annotators. Learning from multiple annotations from diverse annotators can provide a more diverse range of insights, which can help to improve the accuracy and reliability of the annotations [18]. However, the challenge with multi-annotation is that different annotators may have different levels of expertise, and therefore their annotations are often inconsistent. While previous works aim to improve model performance by incorporating annotations from multiple annotators (e.g., Figure 1), these approaches suffer from several limitations: **1) The noisiness of annotations from different annotators due to the personality attributes**. In medical imaging, the quality of annotations can vary based on the personality of annotators, such as attention to detail and experience. For example, as shown in Figure 1, annotator 1 acts more aggressively than annotator 2 when marking the nodule location. **2) The lack of prior information about the corresponding relationship between annotations and annotators**. Due to privacy concerns and difficulty in data collection, there is often no pre-given information about the relationship between annotations and annotators. This makes it difficult to learn the persistent behavioral characteristics specific to each annotator. **3) Missing annotations as not all the annotators will label each image**. Multiple annotations for a medical image (e.g., CT scan) may be missing due to various factors such as technical limitations, time constraints, or insufficient resources. However, the lack of multiple annotations can limit the accuracy and reliability of medical image interpretation, which may impact patient diagnosis and treatment outcomes. To address the above challenges, we propose a novel explanation supervision framework that deals with the inconsistent quality of annotation labels among different annotators. This work makes several contributions, which can be summarized as follows:

- Introducing a novel framework for explanation supervision that incorporates **multiple explanation annotations.** The proposed framework is supervised by class labels and multiple explanation annotations aggregated by learnable weights for each annotator.
- Proposing a new generative model to **generate missing annotations.** The generative model is inferred based on a newly proposed variational inference technique and can learn the characteristics of each annotator when generating annotations.
- Developing a **novel alignment mechanism** and incorporated into proposed generative model to infer the correspondence between annotations and annotators in the training process. This novel alignment mechanism can convert the correspondence inference problem to a linear sum assignment problem.
- Conducting extensive experiments with a variety of evaluation metrics on two medical datasets demonstrates our effectiveness in improving model predictability and generating robust consensus labels via our generative model to noisy annotations.

2. Related Work

Medical Image Diagnosis Early detection of nodules can lead to more effective treatment and improved patient outcomes. Deep learning-based approaches have shown particularly promising results in terms of medical image classification, particularly convolutional neural networks (CNNs) [40, 42, 33, 16, 44, 6]. [42] presents a novel approach to automatically detect pulmonary nodules in CT scans using convolutional neural networks (CNNs) that are based on maximum intensity projection (MIP). [33] proposes a novel multi-view CNN architecture that uses multiple views of the same nodule as inputs, which achieves higher classification accuracy than single-view methods, demonstrating the potential of multi-view approaches for medical image classification. Despite the impressive results achieved by these studies, a major limitation is the lack of interpretability due to the "black-box" nature of neural networks. This limits the potential clinical utility of these methods, as physicians may require justifications to make informed decisions about patient care.

Explanation Supervision The integration of human knowledge into interpretable models has been extensively studied in NLP and tabular data through techniques such as attribution and feature regularization [3, 1, 5, 4]. In recent years, there has been increasing awareness of the importance of visual explanations. One prominent approach involves obtaining local explanations through saliency maps, which highlight the input features that contribute the most to a model's prediction [23, 31]. HAICS [34] is a conceptual framework for image classification that uses human annotation in the form of scribble annotations as the explanation supervision signal. However, such approaches depend heavily on the annotation quality, which is often inaccurate and prone to biases. To address these problems, [13] proposed a novel objective that can handle inaccurate, incomplete, and inconsistent boundaries of human annotations.

Multi-annotation Multi-annotation occurs when a dataset is labeled by several annotators, aiming to increase the annotations' reliability and accuracy [18]. The challenge with multi-annotation is that different annotators may have different levels of expertise, and therefore their annotations are often inconsistent. In order to build accurate and reli-

able machine learning models, it is essential to first evaluate the reliability of each annotator. Several existing techniques have been proposed to evaluate the annotation quality, such as Expectation Maximization(EM) based approaches and annotator confusion estimation-based methods [26, 39, 19, 38]. However, one major limitation of these approaches is that they require the global identification of annotators, which is not always feasible [37]. When the identity of the annotator is unknown, it's harder to make a judgment of the quality locally, and therefore simple aggregation method is more favorable. The Simple-mean method calculates the mean loss over each annotation of each sample to train the model. The Weighted-minimum loss-based method only uses the annotation with the smallest loss with the model prediction for error propagation [22, 24]. The Aggregation method, which is the most commonly used, uses a majority vote or consensus-based approach [18, 8, 25].

3. Problem Formulation

Suppose we have a dataset $(I, E, y) = \{(I_i, E_i, y_i)\}_{i=1}^{N}$, where N is the sample size, $I_i \in \mathbb{R}^{C\times H\times W}$ represents the original image with C, H, W denoting the number of channels, height, and width, $y_i \in \mathbb{R}$ denotes the class label of the original image. $E_i = \left\{E_{i,u} \in \mathbb{R}^{H\times W}\right\}_{u=1}^{U}$ for $i = 1, \ldots, N$ represents U explanation annotation masks of the class label y_i, and U denotes the number of annotators annotating the explanation annotation.

The goal of explanation supervision is to learn the mapping function f of the backbone classifier for input images I to class labels y: $f : I \rightarrow y$ by the supervision of both class labels and multiple explanation annotations. More formally, the objective function is:

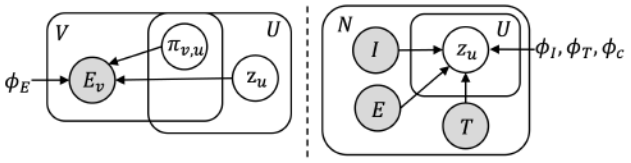


Figure 2: Probabilistic graphical model of the proposed generative model. The left figure represents the generative process and the right figure denotes the inference process.

function of the Gaussian distribution.

4.2. Architecture of MAGI Framework

Based on the above inference for the objective, we incorporate our generative model into explanation supervision and proposed our MAGI framework, as shown in Figure 3. We use the backbone DNN classifier to do prediction with the supervision of both class labels and generated multiple annotations from the proposed generative model. Details will be described in this section.

Visual Encoder We have two visual encoders that take the original images I_i and the aggregated annotation T_i (aggregated from multiple annotations, using methods such as consensus and random selection, serves as a proxy for the ground truth annotation) to model the distribution of $q(z_i^I|I_i)$ and $q(z_i^T|T_i)$, respectively. We introduce two CNNs, $Enc_I(\cdot)$ and $Enc_T(\cdot)$, to infer mean and standard derivation of the representations of I_i and T_i, i.e., $\mu_i^I, \sigma_i^I = Enc_I(I_i; \phi_I)$, $\mu_i^T, \sigma_i^T = Enc_T(T_i; \phi_T)$, where ϕ_I and ϕ_T are the parameters of two visual encoders. Each representation is sampled using its own inferred mean and standard derivation. For example, the representation vectors z_i^I are sampled as $z_i^I = \mu_i^I + \sigma_i^I * \eta$, where η follows a standard normal distribution [15].

Annotation Decoder We also have a decoder to model the conditional distribution $p(E_{i,u}|z_i^I, z_i^T, c^u)$ with a CNN model. With the image feature z_i^I, annotation feature z_i^T, and annotator embedding c^u, the feature composition is done by concatenation to integrate multiple features as the inputs of the decoder $Dec(\cdot)$ to generate multiple annotation maps. The mathematical representation is written as:

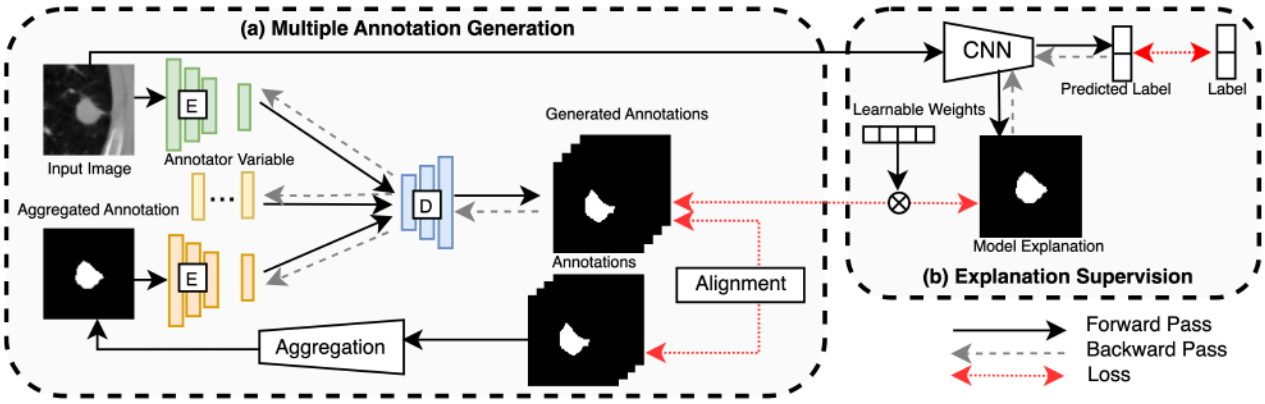


Figure 3: Illustration of proposed MAGI Framework. MAGI consists of a multiple-annotations generation model (a) to generate multiple annotations and an Explanation Supervision module (b) to train the image classifier supervised by both class labels and generated multiple annotations.

Finally, we combine all these three losses as the objective function to optimize our model as follows:

Comparison	Model	Pancreas					LIDC-IDRI				
		Accuracy ↑	AUC ↑	Precision ↑	Recall ↑	F1 ↑	Accuracy ↑	AUC ↑	Precision ↑	Recall ↑	F1 ↑
a) Annotation Aggregation	Baseline	85.089	96.063	98.816	83.693	90.223	66.404	71.790	59.288	69.020	63.466
	Consensus	88.790	95.589	98.038	88.755	93.063	67.435	72.291	65.645	73.184	68.991
	MeanLoss	88.861	96.402	98.884	88.008	93.099	67.090	72.355	65.762	71.765	68.471
	Random	86.441	96.920	**98.994**	85.104	91.239	66.835	72.240	64.947	73.137	68.705
	Proposed	**90.049**	**96.907**	98.905	**89.336**	**93.867**	**68.996**	**74.229**	**66.125**	**78.675**	**71.587**
b) Explanation Supervision	GRADIA	83.132	95.630	**99.042**	81.120	89.103	67.435	72.291	65.645	73.184	68.991
	HAICS	86.441	**96.920**	98.994	85.104	91.239	66.855	71.197	64.709	74.604	69.136
	RES-L	89.786	95.469	98.065	**89.876**	93.784	68.353	73.509	65.966	76.275	70.551
	RES-G	89.893	96.743	98.943	89.170	93.791	68.557	**74.728**	**67.930**	71.459	69.267
	Proposed	**90.049**	96.907	98.905	89.336	**93.867**	**68.996**	74.229	66.125	**78.675**	**71.587**

Table 1: The prediction evaluation on both datasets on a) different multiple annotation aggregation methods and b) different explanation supervision methods. The best results for each task are highlighted in boldface font.

tion generation, we compare the aggregated generated multiple annotations with the ground truth annotation in the test set utilizing the Mean Squared Error (MSE), Binary Cross-Entropy (BCE), and Structural Similarity Index (SSIM).

Comparison Methods We compare our method in two dimensions: 1) Explanation supervision methods:

- **Baseline**: A conventional image classifier is trained only on the prediction loss with the ResNet-18 architecture.
- **HAICS** [34]: A framework that minimizes both the prediction loss and the distance between the model explanation and the scribble annotation labels.
- **GRADIA**: A framework with L1 loss that minimizes the distance between the continuous model explanation and the binary positive explanation labels.
- **RES** [13]: A framework for Guiding Visual Explanation a) with a fixed imputation function via Gaussian convolution filter and b) with a learnable imputation function via multiple layers of learnable kernels.

and 2) Multiple annotations aggregation methods:

- **Consensus**: Each pixel/voxel of the image is considered separately, and a new boolean-valued annotation is generated based on the fraction of the segmentations that agree on the presence of that pixel/voxel.
- **Random**: Randomly sample one annotation among multiple annotations for each sample during training.
- **MeanLoss**: Compute the mean loss over each annotation from multiple annotations against model explanation annotation for each sample to compute the explanation loss during the training process.

5.3. Comparison with annotation aggregation methods

The effectiveness of the proposed model is confirmed in Table 1 through comparison with both the baseline and other multi-annotation aggregation methods. All aggregation methods show enhanced performance over the baseline, underscoring the efficacy of jointly optimizing prediction loss and explanation loss. Our model consistently outshines other methods in the LIDC-IDRI datasets, achieving the best results in terms of accuracy, AUC, recall, and F1 in the pancreas dataset. This emphasizes the significance of generating annotations that align closely with annotators. The improved performance over the consensus mask approach implies that consensus maps are more susceptible to errors when noise is present. Recall, which measures the proportion of true positives correctly identified by the model, is especially crucial in medical contexts where false negatives are unwelcome. Although we did not attain the highest precision score in the pancreas dataset, our F1 score remains the highest, signifying that our overall performance excels in comparison to all other aggregation methods.

Dataset	Method	MSE ↓	BCE ↓	SSIM ↑
Pancreas	Noisy	22.291	0.004	0.984
	Proposed	**15.043**	**0.002**	**0.991**
LIDC-IDRI	Noisy	28.282	0.009	0.970
	Proposed	**21.318**	**0.007**	**0.976**

Table 2: Result of consensus of generated multiple annotations on LIDC-IDRI dataset. MSE, BEC, and SSIM metrics are reported compared to noisy consensus annotation. "Noisy" denotes using the simple consensus method to aggregate noisy-added multiple annotations.

5.4. Multiple Annotation Generation Analysis

Quantitative Analysis of Generated Annotation. We further compare the model-generated annotations with the ground truth and report performance on MSE, BEC, and SSIM in Table 2. We see 32.5% decrease in MSE, 50% decrease in BCE, and 0.71% improvement in SSIM for the pancreas dataset, and 24.6% decrease in MSE, 22.2% decrease in BCE, and 0.61% improvement in SSIM for the LIDC-IDRI dataset. MAGI-generated explanation consistently achieves the best performance, which means that the model is able to capture the actual shape of annotation even

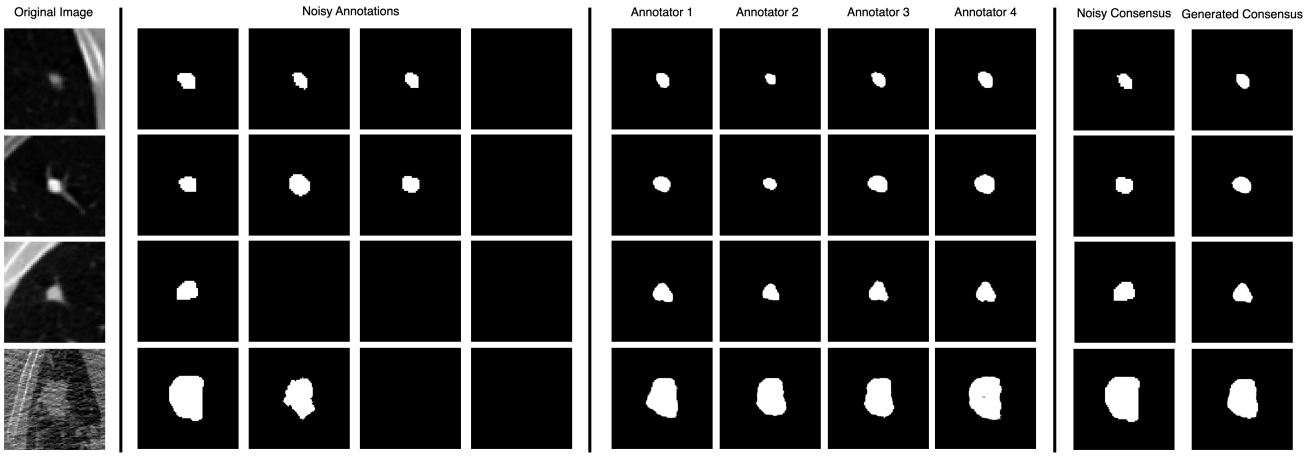


Figure 4: Multiple annotation generation results and comparison with noisy annotation.

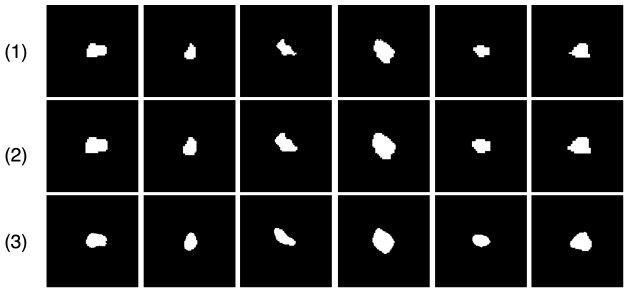


Figure 5: The comparison between 1) ground truth consensus annotation, 2) noisy consensus annotation, and 3) consensus annotation from our generated multiple annotations.

when noise is present. By aligning annotations with annotators, we are able to generate consensus label from de-noised annotations and address the missing annotation issue. The improvement in accuracy shows our annotation is superior to the simple aggregation methods.

Robustness against noisy annotations. Our results imply that our model is robust against noisy annotations and less prone to errors existent in the annotations. As shown in Figure 5, the consensus annotations generated by our model exhibit a similar appearance as the ground truth annotations, implying that our model is able to detect and avoid noisy annotations when constructing consensus labels. Moreover, the ground-truth annotation label is subject to the image size provided by the annotators. When scaling the images to a higher resolution, we see abnormal shapes in the corners of the first two rows, where MAGI-generated annotation has higher similarity with what nodule looks like in reality. The model is able to learn from incomplete and misleading shapes and offer better solutions. On the other hand, the quality of consensus obtained from noisy annotations is heavily subjected to the noisy label, as shown in row 4 of Figure 4. The noisy consensus is highly identical to the first noisy annotation (second column) when the variety of annotations is limited. Our algorithm is able to learn from the annotators and generate multiple annotations even when not all annotators have the time to annotate the image. The generation process offers higher quality and more variations of the annotations and therefore results in consensus labels that capture more details. Row 3 clearly shows the appealing performance of our algorithm in detecting the correct shape of the nodules.

Learning annotator-specific patterns. Our proposed multiple annotation generation model can learn the characteristics of annotators and generate annotations for an image according to that specific annotator. We show samples of the multiple annotations generated by our model in Figure 4. Patterns can be observed between annotations generated corresponding to different annotators. For example, the conservative personality of annotator 2 is learned by our model, so annotator 2 always provides conservative annotations, as shown in the consistently reduced size of annotations (column 7). Conversely, the aggressive personality of annotator 4 is learned, so annotator 4 always provides aggressive annotations (column 9). The comparison between the noisy consensus and our consensus label reveals the appealing performance of our model in learning the variety of annotators' characteristics.

5.5. Comparison with explanation supervision methods

Table 1 shows the quantitative evaluation of two classification tasks and demonstrates the importance of annotation quality for explanation supervision. Overall, our method outperforms all other supervised methods on both LIDC-IDRI and pancreas datasets. In particular, our model achieves the best accuracy and F1 on the pulmonary nodule classification task and the best accuracy, Recall, and F1 on the pancreatic tumor classification task. Our explanation supervision architecture is based on GRADIA, where the

difference is the use of consensus labels. We obtain an accuracy of 90.049% and 68.996% on the two datasets, which is an 8.32% and 2.31% improvement respectively. The results suggest that our annotation label is robust to noises when used as the supervision signal. While the precision or recall scores reported in the pancreas dataset are slightly lower, we achieve the best F1 scores on both datasets, indicating that our overall performance is superior to all other explanation supervision methods. Since our approach does not directly interfere with the supervision process, it can be easily applied to any explanation supervision algorithms for better predictability and interpretability.

6. Conclusion

This paper proposes a novel MAGI:multi-annotated explanation-guided learning framework that generates noise-robust annotations via variational inference-based techniques. We introduce a multiple annotation generative model to tackle the missing, inconsistency, and noisiness of multi-annotation among different annotators. We also incorporate an annotation alignment module to learn the unknown correspondence between annotations and annotators. Results show appealing annotation quality and improved performance on downstream classification tasks. We show that each generated annotation can learn the corresponding characteristics of annotators with the existence of synthetic noise. Extensive experiments on the two medical datasets demonstrate the superiority of our method over existing algorithms.

Broader Impacts. Explaining the decision-making process behind AI systems with high-quality explanations is crucial to the development of trustworthy AI. We envision this work to offer new opportunities to a wide range of high-stakes domains, such as healthcare and finance, and to open a new door to optimize noisy human knowledge when integrating it into deep learning algorithms.

Acknowledgments

This work was supported by the National Science Foundation (NSF) Grant No. 1755850, No. 1841520, No. 2007716, No. 2007976, No. 1942594, No. 1907805, a Jeffress Memorial Trust Award, Amazon Research Award, Oracle for Research Grant Award, Cisco Faculty Research Award, NVIDIA GPU Grant, Design Knowledge Company (subcontract number: 10827.002.120.04), CIFellowship (2021CIF-Emory-05), and the Department of Homeland Security under Grant No. 17STCIN00001.

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018. 1, 2

[2] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. 6

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020. 2

[4] Guangji Bai, Chen Ling, Yuyang Gao, and Liang Zhao. Saliency-augmented memory completion for continual learning. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 244–252. SIAM, 2023. 2

[5] Guangji Bai and Liang Zhao. Saliency-regularized deep multi-task learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 15–25, 2022. 2

[6] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 541–549. Springer, 2019. 2

[7] Sébastien Bougleux and Luc Brun. Linear sum assignment with edition, 2016. 5

[8] Pete Bridge, Andrew Fielding, Pamela Rowntree, and Andrew Pullar. Intraobserver variability: should we worry?, 2016. 3

[9] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 91–107. Springer, 2020. 1

[10] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of A generative adversarial network. *CoRR*, abs/1611.05644, 2016. 5

[11] Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. Machine learning for medical imaging. *Radiographics*, 37(2):505–515, 2017. 1

[12] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning. *arXiv preprint arXiv:2212.03954*, 2022. 1, 2

[13] Yuyang Gao, Tong Steven Sun, Guangji Bai, Siyi Gu, Sungsoo Ray Hong, and Zhao Liang. Res: A robust framework for guiding visual explanation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 432–442, 2022. 1, 2, 7

[14] Siyi Gu, Yifei Zhang, Yuyang Gao, Xiaofeng Yang, and Liang Zhao. Essa: Explanation iterative supervision via

saliency-guided data augmentation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 567–576, 2023. 6

[15] Xiaojie Guo, Yuanqi Du, and Liang Zhao. Deep generative models for spatial networks. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021. 5

[16] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 137–147. Springer, 2020. 1, 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[18] Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis*, 65:101759, 2020. 2, 3

[19] Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*, 2017. 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[22] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015. 3

[23] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019. 2

[24] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013. 3

[25] Guy Nir, Soheil Hor, Davood Karimi, Ladan Fazli, Brian F Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F Villamil, Gang Wang, R Storey Wilson, et al. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical image analysis*, 50:167–180, 2018. 3

[26] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11(4), 2010. 3

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 1

[28] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR, 2020. 2

[29] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017. 2

[30] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*, pages 556–564. Springer, 2015. 6

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 2

[32] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. 3

[33] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram Van Ginneken. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging*, 35(5):1160–1169, 2016. 1, 2

[34] Haifeng Shen, Kewen Liao, Zhibin Liao, Job Doornberg, Maoying Qiao, Anton Van Den Hengel, and Johan W Verjans. Human-ai interactive and continuous sensemaking: A case study of image classification using scribble attention maps. In *Extended Abstracts of CHI*, pages 1–8, 2021. 1, 2, 7

[35] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *CoRR*, abs/1704.02685, 2017. 3

[36] Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3619–3629, 2021. 2

[37] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020. 3

[38] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from

noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11244–11253, 2019. 3

[39] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004. 3

[40] Panpan Wu, Xuanchao Sun, Ziping Zhao, Haishuai Wang, Shirui Pan, and Björn Schuller. Classification of lung nodules based on deep residual networks and migration learning. *Computational intelligence and neuroscience*, 2020, 2020. 1, 2

[41] Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9214–9223, 2021. 4

[42] Sunyi Zheng, Jiapan Guo, Xiaonan Cui, Raymond NJ Veldhuis, Matthijs Oudkerk, and Peter MA Van Ooijen. Automatic pulmonary nodule detection in ct scans using convolutional neural networks based on maximum intensity projection. *IEEE transactions on medical imaging*, 39(3):797–805, 2019. 1, 2

[43] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015. 3

[44] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 384–393. Springer, 2019. 1, 2

[45] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 420–428. Springer, 2019. 1