RES: A Robust Framework for Guiding Visual Explanation

Yuyang Gao
yuyang.gao@emory.edu
Emory University
Atlanta, GA, USA

Tong Sun
tsun8@gmu.edu
George Mason University
Fairfax, VA, USA

Guangji Bai
guangji.bai@emory.edu
Emory University
Atlanta, GA, USA

Carrie Gu
carrie.gu@emory.edu
Emory University
Atlanta, GA, USA

Sungsoo Hong
shong31@gmu.edu
George Mason University
Fairfax, VA, USA

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

ABSTRACT

Despite the fast development of explanation techniques where the main focus is handling "how to generate the explanations", research questions like "whether the explanations are accurate", "what if the explanations are inaccurate", and "how to adjust the model to generate more accurate explanations" are still relatively under-explored by the research body. To guide the model toward better explanations, explanation supervision techniques which add supervision signals on the model explanation has started to show promising effects on improving both the generalizability as well as intrinsic interpretability of Deep Neural Networks (DNNs), especially on the text and attributed data where the human annotation labels can be assigned accurately on each feature of the data. However, the research on supervising visual explanations (e.g., explanation represented by saliency maps) is still under-explored and in its nascent stage, due to several inherent challenges: 1) inaccuracy of the human explanation annotation boundary, 2) incompleteness of the human explanation annotation region, and 3) inconsistency of the data distribution between human annotation and model explanation maps. To address the challenges, we propose a generic RES[1] framework for guiding visual explanation by developing a novel objective that handles inaccurate boundary, incomplete region, and inconsistent distribution of human annotations, with a theoretical justification on model generalizability. Extensive experiments on two real-world image datasets demonstrate the effectiveness of the proposed framework on enhancing both the reasonability of the explanation and the performance of the backbone DNNs model.

CCS CONCEPTS

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

KEYWORDS

Explainability, Interpretability, Robustness, Visual Explanation

ACM Reference Format:
Yuyang Gao, Tong Sun, Guangji Bai, Carrie Gu, Sungsoo Hong, and Liang Zhao. 2022. RES: A Robust Framework for Guiding Visual Explanation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539419

1 INTRODUCTION

As DNNs become widely deployed in border areas of applications including sensitive areas, the study on explainability that aims at understanding how DNNs work is currently attracting considerable attention from researchers [1, 2, 16]. To open the "black box" of DNNs, many explainability techniques have been proposed that try to provide the "local explanation" of the DNNs prediction for a specific instance [16], such as methods that provide the saliency maps for understanding which sub-parts (i.e., features) in an instance are most responsible for the model prediction [3, 4, 23, 24, 30, 37].

Despite the fast development of explanation techniques where the main focus is handling "how to generate the explanations", research questions like "whether the explanations are accurate", "what if the explanations are inaccurate", and "how to adjust the model to generate more accurate explanations" are still relatively under-explored by the research body.

Recently, explanation supervision techniques which add supervision signals on the model explanation has started to show promising effects on improving both the generalizability as well as intrinsic interpretability of DNNs in many application domains, especially on text data [18, 29] and attributed data [33] where the human annotation labels can be assigned accurately on each feature of the data.

However, the research on supervising visual explanations (e.g., explanation represented by saliency maps) on image data is still under-explored and in its nascent stage. This is because of several inherent challenges in supervising visual explanations: **1) Inaccuracy of the human explanation annotation boundary.** It is very difficult and costly for the boundary of human annotation to be perfectly accurate, which could lead the model to falsely assign positive explanation value to irrelevant features (i.e., pixels in image data). For example, as shown by the yellow arrows in Figure 1 (b), the coarsely drawn boundary of the human annotation falsely excluded a non-trivial region of the boundary of the wildflowers that could also be important to the prediction. **2) Incompleteness**

[*] Corresponding author
[1] Code available at: https://github.com/YuyangGao/RES.

KDD '22, August 14–18, 2022, Washington, DC, USA.

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539419

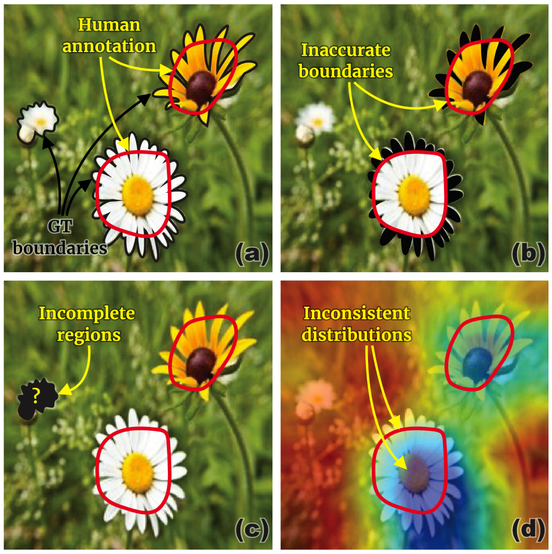


Figure 1: An example showing the challenges present in the human annotation labels: (a) human annotations are represented with red lines while ground-truth boundaries are shown with black lines. (b) Error caused by "inaccurate boundaries" are presented with black regions, (c) Error caused by "incomplete regions" are shown with a black region, and (d) the discrepancies between the "binary" human annotation and the "continuous" model-generated explanation maps. The explanation is queried based on predicting the scene as 'wild nature'.

of the human explanation annotation region. When labeling the explanation for image data, people usually tend to provide only a few regions as long as they are sufficient to convince people about the decision and do not bother to comprehensively find all the possible regions that cover all different ways of explanations. Such incompleteness can mislead the model to wrongly penalize all the regions as long as they are not selected by annotators. An example of such a situation can be illustrated by Figure 1 (c), where the human annotation clearly missed one wildflower as shown in the black region. **3) Inconsistency of the data distribution between human annotation and model visual explanations.** The saliency maps generated by model explainers and those labeled by human annotators reside in different data domains and distributions: the former takes the form of continuous real values while the latter is often in discrete binary values. Figure 1 (d) clearly show the huge difference between "binary" human annotation (i.e., red circled areas as marked as positive, and the rest areas as negative) and the continuous model-generated explanation map (i.e., more importance is given to the area with a warmer color) in distribution. Therefore, human-annotated explanations cannot be directly used to supervise the model and its explanations without significant efforts to fill the gap between the data domain and distributions.

To address the above challenges, beyond merely applying human annotation labels directly as the supervision signals to train the model, this paper focuses on proposing a generic robust explanation supervision framework for learning to explain DNNs under the assumptions that the human annotation labels can be inaccurate in the boundary, incomplete in the region, as well as inconsistent with the distribution of the model explanation. Specifically, we propose a novel robust explanation loss that addresses all three aforementioned challenges present in the noisy human annotation labels. In addition, we give a theoretical justification of the benefits of having the proposed explanation loss to the generalizability power of the backbone DNN model.

Specifically, the main contributions of our study are as follows:

(1) **Proposing a generic framework for learning to explain DNNs with explanation supervision.** We propose a unified framework that enables explanation supervision on DNNs with both positive and negative explanation annotation labels and is generalizable to the existing differentiable explanation methods.
(2) **Developing a robust model objective that can handle the noisy human annotation labels as the supervision signal.** We propose a novel robust explanation loss that can handle the inaccurate boundary, incomplete region, as well as inconsistent distribution challenges in applying the noisy human annotation labels as the supervision signal.
(3) **Providing a theoretical justification on the generalizability power of the proposed framework.** We formally derive a theorem that provides an upper bound for the generalization error of applying the proposed robust explanation loss when training the backbone DNN models.
(4) **Conducting comprehensive quantitative and qualitative experimental analysis to validate the effectiveness of the proposed model.** Extensive experiments on two real-world image datasets, gender classification and scene recognition, demonstrate that the proposed framework improved the backbone DNNs both in terms of prediction power and explainability. In addition, qualitative analyses, including case studies and user studies of the model explanation, are provided to demonstrate the effectiveness of the proposed framework.

2 RELATED WORK

Our work draws inspiration from the research fields of local explainability techniques of DNNs that provide the model-generated explanation, and explanation supervision on DNNs which enables the design of pipelines for the human-in-the-loop adjustment on the DNNs based on their explanations to enhance both explainability and performance of DNN models.

2.1 Local Explainability Techniques of DNNs

As DNNs become widely deployed in border areas of applications including sensitive areas, recent years have seen an explosion of research in understanding how DNNs work under the hood (e.g., explainable AI, or XAI) [1, 2, 12, 16, 34]. Due to the "black box" nature of DNNs, most of the existing and well-received explainability methods focus on providing a "local explanation" that aims at explaining the prediction in understandable terms for humans for a specific instance or record [16]. One popular direction is to compute saliency maps as the local explanation, which provide the saliency values regarding which input features are most responsible for the prediction of the model [3, 23, 24, 30, 37]. For example, for image input, a saliency map is able to summarize where the model is "paying attention to" when performing a certain image recognition task. In this direction, one set of works incorporates network activations into their visualizations, such as Class Activation Mapping (CAM) [37] and Grad-CAM [30]. Another set of approaches take a backward pass and assign a relevance score for each layer backpropagating the effect of a decision up to the input level, existing works such as LRP [3, 23], and DTD [24] belong to this category. In addition, some model inspection methods such as VisualBackProp (VBP) [7] can also provide a local explanation similar to the LRP approaches. Besides the above techniques that are more specifically designed for interpreting image data, there are also several existing techniques that aim at providing more model-agnostic explanations on different types of data, such as LIME [27], Anchors [28], MES [32], and LORE [15]. Please refer to the survey papers [2, 16] for a more comprehensive review of the existing works.

2.2 Explanation Supervision on DNNs

The potential of using *explanation*–methods devised for understanding which sub-parts in an instance are important for making a prediction–in improving DNNs has been studied in many domains across different applications [13]. In particular, explanation supervision techniques have been widely explored on image data by the computer vision community [9, 10, 21, 22, 25, 26, 35]. Existing studies have shown the benefit of using stronger supervisory signals by teaching networks where to attend [21]. Following this line of study, several explanation supervision frameworks have been proposed. Mitsuhara et al. [22] proposed a post hoc fine-tuning strategy, where an end-user is asked to manually edit the model's explanation to interactively adjust its output. However, the proposed framework is only applicable to a specific type of DNN called Attention Branch Network [11]. In addition, several frameworks designed for the Visual Question Answering (VQA) domain have been proposed, where the goal is to obtain the improved explanation on both the text data and the image data [10, 25, 26, 35].

More recently, several more generic frameworks have been proposed for explanation supervision on image data. One existing work proposed a conceptual framework HAICS [31], and the authors further implement it in an image classification application with human annotation in the form of scribble annotations as explanation supervision signals. Another work proposed the GRADIA pipeline [14], which improves the performance and explainability of the backbone DNN by jointly optimizing the task performance as well as the alignment between the human-annotated attention maps and model explanation maps. Besides image data, the explanation supervision has also been studied on other data types, such as texts [18, 29], attributed data [33], and more recently on graph-structured data [13]. However, most of the existing works typically assume the human annotation labels are clean and accurate, while in practice they are prone to be inexact, inaccurate, and incomplete when directly used as the supervision signal for supervising the model explanation. To our best knowledge, we are the first to propose a robust explanation supervision framework that aims at handling this open research problem.

3 MODEL

In this section, we first introduce the proposed RES framework that enables explanation supervision on DNNs with both positive and negative explanation annotation labels. We then move on to propose a novel robust explanation loss that is designed to handle the inaccurate boundary, incomplete region, as well as inconsistent distribution challenges in applying the noisy human annotation labels as the supervision signal. Finally, we give the theoretical justification of the benefits of having the proposed explanation loss to the generalizability power of the backbone DNN model.

Problem formulation: Let $x \in \mathbb{R}^{C\times H\times W}$ be the input image data with C channels, H as height, and W as width. Let y be the class label for input x, the general goal for a DNN model is to learn the mapping function f for each input x to its corresponding label, $f : x \rightarrow y$.

3.1 The RES Framework

The general goal for the RES framework is to boost the model explainability via robust explanation supervision such that the model can robustly learn to assign more importance to the right input features even given noisy human explanation annotation labels, and consequently boost the task performance as well as the interpretability of the backbone DNN model. Here, we present the general learning objective of the RES framework to be a joint optimization of the model prediction loss and the robust explanation loss. Concretely, we propose the objective function as:

$$\min \sum_{i}^{N} \underbrace{\mathcal{L}_{\text{Pred}}(f(x^{(i)}), y^{(i)})}_{\text{prediction loss}} + \underbrace{\mathcal{L}_{\text{Exp}}(\langle M^{(i)}, F^{(i)}, C^{(i)}\rangle)}_{\text{robust explanation loss}} \qquad (1)$$

where $M^{(i)} \in \mathbb{R}^{H\times W}$ denotes the model-generated explanations for ith sample using a given explanation method; $F^{(i)} \in \{0,1\}^{H\times W}$ and $C^{(i)} \in \{0,1\}^{H\times W}$ denote the corresponding binary labels for positive (i.e., $F^{(i)}_{j,k} = 1$ if the pixel at coordinate (j,k) of sample image i should be assigned with high importance, and 0 otherwise) and negative (i.e., $C^{(i)}_{j,k} = 1$ if the pixel at coordinate (j,k) of image i should be assigned with low importance value, and 0 otherwise) explanation marked by the human annotators. $\mathcal{L}_{\text{Pred}}(f(x^{(i)}), y^{(i)})$ is the typical prediction loss (such as the cross-entropy loss).

3.2 Robust Explanation Supervision for Noisy Explanation Annotation labels

To address the challenges presented in the noisy human annotation labels, we propose a robust explanation loss $\mathcal{L}_{\text{Exp}}$ that measures the discrepancies between model and human explanations regarding both the positive and negative explanation and taking into consideration the noisy nature of human annotation labels. Without loss of generality, let us assume $\tilde{M^{(i)}} = \tilde{F}^{(i)} - \tilde{C}^{(i)}$ in range $[-1, 1]$ be the ground truth ideal explanation value for input image $x^{(i)}$, given the ideal positive explanation $\tilde{F}^{(i)} \in [0, 1]$ and negative explanation $\tilde{C}^{(i)} \in [0, 1]$; the binary human annotation as $F^{(i)}$ and $C^{(i)}$; and the model explanation as $M^{(i)} = g(f_\theta((x^{(i)}))$, where function

$g(\cdot)$ specify the explanation method. We have $\mathbb{E}[\|M^{(i)} - (F^{(i)} - C^{(i)})\| - \|(F^{(i)} - C^{(i)}) - \tilde{M}^{(i)}\|] \leq \max\{0, \mathbb{E}[\|M^{(i)} - (F^{(i)} - C^{(i)})\|] - \mathbb{E}[\|(F^{(i)} - C^{(i)}) - \tilde{M}^{(i)}\|]\} \leq \mathbb{E}[\max\{0, \|M^{(i)} - (F^{(i)} - C^{(i)})\| - \|(F^{(i)} - C^{(i)}) - \tilde{M}^{(i)}\|\}] \leq \mathbb{E}[\|M^{(i)} - \tilde{M}^{(i)}\|]$ according triangle inequality. We define $\alpha = \mathbb{E}[\|(F^{(i)} - \tilde{F}^{(i)}) - (C^{(i)} - \tilde{C}^{(i)})\|]$. Therefore, to minimize $\|M^{(i)} - \tilde{M}^{(i)}\|$, we can have a tighter surrogate loss based on the annotated labels as follows:

$$\max\{0, \|M^{(i)} - (\tilde{F}^{(i)} - \tilde{C}^{(i)})\| - \alpha\}$$

Since the ground truth $\tilde{F}$ and $\tilde{C}$ are unknown, estimating α can be difficult. In practice, we can assume their distributions are positively correlated with the distribution of F and C, which can therefore be estimated by a slack variable α. To keep it simple and without loss of generality, in this work, we define α as a hyper-parameter of the framework assuming no additional knowledge about the ideal distribution.

3.2.1 Bridging the distribution between human labels and model explanation maps. To bridge the continuous model explanation $M^{(i)}$ with binary human labels C and F, we propose to split the above objective into two terms with bidirectional projections, as follows:

$$\min_{\theta,a} \sum_{i}^{N} \max\{0, \|[\hat{M}^{(i)} - (F^{(i)} - C^{(i)})]\| - \alpha\}$$
$$+ d(M^{(i)}, h(F^{(i)}, C^{(i)})) \tag{2}$$

where $d(\cdot)$ is a distance function, $h(\cdot)$ is a mapping function that maps the binary masks $F^{(i)}$ and $C^{(i)}$ to continuous value in range $[0, 1]$, and $\hat{M}^{(i)}$ is a binary projection of $M^{(i)}$ by a threshold a, as:

$$\hat{M}^{(i)} = \begin{cases} 1 & M^{(i)} \geq a \\ -1 & M^{(i)} < a \end{cases} \tag{3}$$

Basically, the above equation takes both the absolute difference (measured by the first term) and relative distance (measured by the second term) into consideration when comparing the continuous model explanation and the binary human explanation masks.

3.2.2 Mitigating the Inaccurate Boundary via Label Imputation. To realize the mapping function $h(\cdot)$ in Equation (3) which aims at projecting the binary human labels into continuous value domain, an intuitive way is to define $h(\cdot)$ as applying a $k \times k$ Gaussian kernel on the binary annotation labels F and C such that the pixels that close to the boundary of the manual label will also obtain slack values to boost the robustness and deal with the inexact and inaccurate boundary from human annotation.

However, a predefined kernel matrix might not be suitable for every data sample, and the discrepancy and inconsistency among annotators can also influence the accuracy of such a pre-defined estimation on handling the inaccurate boundary issue. Therefore, we further extend this idea and define a learnable imputation function $h_\phi(\cdot)$ with multiple learnable kernel transformations as the parameter set ϕ, such that the kernels' weights can be adjusted and learned to make better estimations of the ground truth explanation values and provide better mitigation to the inaccurate boundary problem. Specifically, the explanation loss with a learnable imputation function as follows:

$$\min_{\theta,a,\phi} \sum_{i}^{N} \max\{0, \|[\hat{M}^{(i)} - (F^{(i)} - C^{(i)})]\| - \alpha\}$$
$$+ d(M^{(i)}, h_\phi(F^{(i)}, C^{(i)})) \tag{4}$$

where ϕ is the parameter set of the imputation function $h_\phi(\cdot)$. The imputation function can be realized by applying multiple layers of convolution operations with learnable kernels over the raw annotation label F and C.

3.2.3 Handling the Incomplete Region by Selective Penalization. Finally, due to the incompleteness of human annotation labels, and to avoid falsely penalizing the model from assigning importance to the relevant features missed by the human labels, we propose to only selectively apply the explanation supervision signal onto the features with either positive or negative annotation labels. Concretely, we define the robust explanation loss $\mathcal{L}_{Exp}$ as follows:

$$\min_{\theta,a,\phi} \sum_{i}^{N} \max\{0, \|[\hat{M}^{(i)} - (F^{(i)} - C^{(i)})] \cdot \mathbf{1}(F^{(i)} - C^{(i)} \neq 0)\| - \alpha\}$$
$$+ d(M^{(i)} \cdot \mathbf{1}(F^{(i)} - C^{(i)} \neq 0), h_\phi(F^{(i)}, C^{(i)}) \cdot \mathbf{1}(F^{(i)} - C^{(i)} \neq 0)) \tag{5}$$

where $\mathbf{1}(\cdot)$ is the indicator function, and $\cdot$ represents the elemental-wise multiplication operation. This formulation also gives the model a certain degree of flexibility on deciding the importance of unlabeled features based on data and downstream task, thus could yield a more generalizable and reasonable explanation that enhance both explainability as well as task performance of the model.

3.3 Optimization of Robust Explanation Loss

The indicator function for calculating $\hat{M}^{(i)}$ (as shown in Equation (3)) prevents us from directly optimizing our model objective with conventional gradient descent algorithms such as Adam [19]. Concretely, the optimization problem presented in Equation (5) involves optimizing both the adaptive threshold a and the model-generated explanation $M^{(i)} = g(f_\theta(x^{(i)}))$. Here, we propose to first find the optimal threshold a given model parameter θ, and then optimize θ with conventional gradient descent algorithm by proposing a differentiable approximation to the indicator function.

First, to find the optimal a given θ, we need to solve the following objective:

$$\min_{a} \sum_{i}^{N} \|[\hat{M}^{(i)} - (F^{(i)} - C^{(i)})] \cdot \mathbf{1}(F^{(i)} - C^{(i)} \neq 0)\| \tag{6}$$

Which is mathematically equivalent to the following equation by expanding $\hat{M}^{(i)}$:

$$\min_{a} \sum_{i}^{N} \|[\mathbf{1}(M^{(i)} \geq a) - F^{(i)}] \cdot F^{(i)}\| +$$
$$\|[\mathbf{1}(M^{(i)} < a) - C^{(i)}] \cdot C^{(i)}\| \tag{7}$$

If we treat each entry of $M^{(i)}$ as having two inequality constraints on a, we can efficiently solve the above formula in $O(m \log m)$ by our proposed algorithm by treating this optimization problem as finding a a that satisfies the maximum number of inequality constraints, where $m = max(|F|, |C|)$. The details of the proposed searching algorithm can be found in Appendix A.3.

To further enable gradient calculation of $M^{(i)}$ in Equation (5), we propose a surrogate loss using the hyperbolic tangent function $tanh(\cdot)$ to approximate the indicator function, as follows:

$$\min_{\theta,a,\phi} \sum_{i}^{N} \max\{0, \|[tanh(\gamma(M^{(i)} - a)) - H^{(i)}] \cdot \mathbf{1}(H^{(i)} \neq 0)\| - \alpha\} + \\ + d(M^{(i)} \cdot \mathbf{1}(H^{(i)} \neq 0), h_{\phi}(F^{(i)}, C^{(i)}) \cdot \mathbf{1}(H^{(i)} \neq 0)) \quad (8)$$

where $H^{(i)} = F^{(i)} - C^{(i)}$; γ controls the slop of the hyperbolic tangent function. Moreover, when $\gamma \to \infty$, we can ensure such a approximation can be mathematically equivalent to the original indicator function in Equation (4) as shown in the following lemma.

Lemma 1. *Equation* (8) *is mathematically equivalent to Equation* (5) *when* $\gamma \to \infty$.

Proof. Please refer to Appendix A.2 for the proof. □

3.4 Theoretical Analysis of Generalizablity

In this subsection, we theoretically justify the generalizability power of the proposed explanation loss, as shown in Theorem 1 below.

We consider the regularized expected loss:

$$\mathcal{L}(f_\theta) = \mathbb{E}\left[\mathcal{L}_{\text{Pred}}(f_\theta(x), y) + \mathcal{L}_{\text{Exp}}(\nabla f_\theta(x))\right] \quad (9)$$

where f_θ is any learnable function with parameter $\theta \in \Theta$. In addition, denote the empirical loss as

$$\hat{\mathcal{L}}(f_\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{L}_{\text{Pred}}(f_\theta(x^{(i)}), y^{(i)}) + \mathcal{L}_{\text{Exp}}(\nabla f_\theta(x^{(i)}))\right) \quad (10)$$

where N denotes the training sample size. $\nabla f_\theta(x)$ denotes the gradient of f_θ on input x, which can be used to generate any explanation. We omit the label (namely, $F^{(i)}$ and $C^{(i)}$) in $\mathcal{L}_{\text{Exp}}$ here for more compact notation. Also, we assume that $\mathcal{L}_{\text{Pred}}$ is L_1-Lipschitz and $\mathcal{L}_{\text{Exp}}$ is L_2-Lipschitz continuous w.r.t its first input, respectively.

Definition 1 (δ-minimizer). *A function $f_{\hat{\theta}}$ is said to be a δ-minimizer of $\mathcal{L}(\cdot)$ if*

$$\mathcal{L}(f_{\hat{\theta}}) \leq \inf_{\theta \in \Theta} \mathcal{L}(f_\theta) + \delta \quad (11)$$

Assumption 1. *Let f_{θ^*} be the solution to Eq.* (9)*. There exists a neural network f_τ with $\tau \in \Theta$ such that*

$$\|f_\tau - f_{\theta^*}\|^2 := \mathbb{E}\left[|f_\tau - f_{\theta^*}|^2 + |\nabla f_\tau - \nabla f_{\theta^*}|^2\right] \leq C_1^2 \frac{\|\theta^*\|^2}{m^\gamma} \quad (12)$$

where C_1 is some constant, m is a constant related to the number of parameters in f, and γ is a constant order.

Assumption 2. *Given any neural network f_θ from $\theta \in \Theta$ and i.i.d sample $\{x^{(i)}\}_{i=1}^{N}$. Given any $0 < \epsilon < 1$, we assume that*

$$\sup_{\theta \in \Theta} |\mathcal{L}(f_\theta) - \hat{\mathcal{L}}(f_\theta)| \leq \frac{C_2(V, m, \epsilon)}{\sqrt{N}} \quad (13)$$

with probability at least $1 - \epsilon$. C_2 relies on set Θ, m and ϵ.

Such an inequality can be ontained using some statistical learning theories like Rademacher complexity.

Now we provide our generalization error bound as follow:

Theorem 1 (Generalizability of Equation (1)). *Let f_{θ^*} be the minimizer of $\mathcal{L}(\cdot)$, $f_{\hat{\theta}}$ be a δ-minimizer of $\hat{\mathcal{L}}$, then given $0 < \epsilon < 1$, with probability at least $1 - \epsilon$ over the choiec of $x^{(i)}$, we have*

$$0 \leq \mathcal{L}(f_{\hat{\theta}}) - \mathcal{L}(f_{\theta^*}) \leq (L_1 + L_2)\frac{C_1\|\theta^*\|}{m^{\gamma/2}} + \frac{2C_2(V, m, \epsilon)}{\sqrt{N}} + 2\delta \quad (14)$$

Proof. Please refer to Appendix A.1 for the formal proof. □

Our Theorem 1 provides an upper bound for the generalization error between the numerical optimal solution $\hat{\theta}$ and the theoretical optimal solution θ^*. The first term in the bound corresponds to the approximation error given in the first assumption, the second term corresponds to the quadrature error given in the second assumption, and the last term corresponds to the training error. To reduce the generalization error, we need to increase both the number of parameters and training samples. Meanwhile, the empirical loss is needed to be solved sufficiently well.

4 EXPERIMENTS

We test our RES framework on two application domains, gender classification and scene recognition. We first describe the detailed settings for the experiments and then present the quantitative studies on both model prediction as well as the explanation. In addition, we include several qualitative studies, including case studies and user studies, to make a better qualitative assessment of how the proposed model has enhanced the explainability of the backbone DNN models.

4.1 Experimental Settings

Gender Classification Dataset: The gender classification[2] is one of the widely used tasks in the research of fairness in broader machine learning communities [5, 8, 36]. We constructed the dataset from the Microsoft COCO dataset[3] [20] by extracting images that had the word "men" or "women" in their captions. We then filtered out instances that 1) contain both words, 2) include more than two people, or 3) humans appear in the figure is nearly not recognizable from human eyes. We collected a total of 1,600 images that satisfied our criterion and obtained the human annotation labels for all the image samples with our human annotation UI (please refer to Appendix A.4 for more details). For data splitting, we only randomly sampled 100 samples out of the 1,600 images as the training set to better simulate a more practical situation where we only have limited assess to the human explanation labels. The rest 1,500 data samples were then evenly split as the validation set and test set.

Scene Recognition Dataset: We obtained the scene images from the Places365 dataset[4] [38]. The original dataset contains more than 10 million images comprising 400+ unique scene categories. Following the macro-class defined by [38], we constructed a binary scene recognition task: nature vs. urban. The data samples for the two classes were randomly sampled from a set of pre-defined categories under macro-class "nature" and "urban", respectively. Specifically, the categories we used to sample the data are listed below:

- *Nature*: mountain, pond, waterfall, field wild, forest broadleaf, rainforest
- *Urban*: house, bridge, campus, tower, street, driveway

Notice that the categories are non-comprehensive and the generated datasets are just for the purpose of studying the quality of model

[2] We are aware that using a binary classification in gender does not reflect on the diverse viewpoint of gender in the real world, and we emphasize that the binary "gender classification" task here does not represent our viewpoint on gender.

[3] Available online at: https://cocodataset.org/

[4] Available online at: http://places2.csail.mit.edu/index.html

explanation. We balanced the sample size for each category and collected a total of 1,600 images. Again, we obtained the human annotation labels for all the samples with the human annotation UI, and split the data randomly with sample sizes of 100/750/750 for training, validation, and testing.

Evaluation Metrics: We evaluate the model in terms of task performance as well as in terms of explainability. For model performance, we use the conventional prediction accuracy to measure the prediction power of the backbone DNN models as the datasets studied are well imbalanced. For explainability assessment, we leverage the human-labeled explanation on the test set to assess the quality of the model explanation. Specifically, we use the Intersection over Union (IoU) score [6], which is calculated by taking the bit-wise intersection and union operations between the ground truth explanation and the binarized model explanation to measure how well the two explanation masks overlap. In addition, since the IoU score only assesses the quality of positive explanation, we further compute the precision, recall, and F1-score as additional metrics which provide a more comprehensive evaluation of the model-generated explanation by taking into consideration the alignment of both positive and negative explanation.

Comparison methods: We compare the performance of the RES framework with the vanilla backbone model as the baseline as well as two existing explanation supervision methods, GRAIDA [14] and HAICS [31]. For the proposed framework, we show two variations: RES-G and RES-L, with different implementations of the imputation function. Concretely, we studied the following methods:

- **Baseline**: The conventional DNN model that is trained with only the prediction loss.
- **GRADIA** [14]: A framework that trains the DNN model with both the prediction loss as well as a conventional L1 loss that directly minimizes the distance between the continuous model explanation and the binary positive explanation labels.
- **HAICS** [31]: A framework that trains the DNN model with both the prediction loss as well as a conventional Binary Cross-Entropy (BCE) loss that directly minimizes the distance between the continuous model explanation and the combination of positive and negative binary explanation labels.
- **RES-G**: The proposed RES framework with the imputation function $g(\cdot)$ as a fixed value Gaussian convolution filter.
- **RES-L**:The proposed RES framework with the learnable imputation function $g_\phi(\cdot)$ via multiple layers of learnable kernels.

Implementation Details: For all the methods studied in this work, the backbone DNN model is based on the pretrained ResNet50 architecture [17]. All models were trained for 50 epochs using the ADAM optimizer [19] with a learning rate of 0.0001. To make a fair comparison on explainability, the model explanations were all generated by the well-recognized explanation technique Grad-CAM [30], although other local explanation techniques can also be applied in our framework. The generated explanation maps are normalized in the range of $(0, 1]$ by dividing the maximum saliency value on each sample for model training as well as visualization. When calculating the explanation evaluation metrics, the explanation maps were further binarized by a fixed threshold of 0.5. The hyper-parameter α of the proposed RES framework was set to 0.001 for the gender classification task, and 0.01 for the scene recognition task, based on grid research via prediction accuracy on the validation set. The detailed implementation of the imputation layers for RES-L can be found in the Appendix A.5.

4.2 Performance

Table 1 shows the model performance and model-generated explanation quality for gender classification and scene recognition datasets. The results are obtained from 5 individual runs for every setting. The best results for each dataset are highlighted with boldface font and the second bests are underlined. In general, our proposed framework variations, i.e., RES-G and RES-L, outperformed all other comparison methods in terms of both prediction accuracy as well as explainability on both datasets. Specifically, regarding prediction power, the RES-G with a pre-defined Gaussian transformation kernel as the imputation function achieved the best performance, outperforming the baseline DNN model by 4% and 3% on prediction accuracy on gender classification and scene recognition datasets, respectively. In addition, the proposed RES framework enhanced the explainability of the backbone DNNs by a significant margin as compared with the baseline DNN model as well as other explanation supervision methods. The proposed RES-L with learnable kernels as the imputation function achieved the highest improvement on model explainability in terms of both IoU and F1 scores on both datasets, out-performing other comparison methods by 8%-72% and 16%-36% on IoU and explanation F1 scores, respectively. The comparison methods GRADIA and HAICS also improved the model performance by leveraging the additional human attention labels, but are generally much less effective than the proposed RES framework. Those results demonstrated the effectiveness of the proposed framework on enhancing the model explainability robustly under noisy annotation labels, and consequently improved the model performance and prediction power on the prediction tasks.

Next, we further studied how the DNN models can benefit from the RES framework to gain a better generalization power under different training sample size scenarios. Specifically, we studied four training sample scenarios with training sample sizes of 10, 20, 50, and 100 on the Gender Classification Dataset. As shown in Figure 2, we present the test prediction accuracy, IoU score, and explanation F1 score of each method under the four training sample size scenarios. The data point represents the mean value over 5 runs, and the error bar here corresponds to the standard deviation. We can see that the proposed RES framework outperformed all other comparison methods by a significant margin under all scenarios studied, especially on boosting the explainability of the backbone DNNs as reflected by IoU and explanation F1 scores. Specifically, RES was able to improve the model prediction accuracy by 2% - 5%, and boosted the quality of the model explanation by 60%-80% and 36%-40% in terms of IoU and explanation F1 scores, respectively. Interestingly, we also observed degradation in model performance when applying GRADIA and HAICS when the sample size is extremely limited, such as in 10 and 20 training sample sizes scenarios. This could be due to the fact that GRADIA and HAICS simply treat the raw human annotation as clear data and thus suffer significantly from learning directly from the noisy labels and consequently prone to over-fitting badly. In contrast, with the robust learning objective,

Table 1: The performance and model-generated explanation evaluation among the proposed models and the comparison methods on both gender classification and scenes recognition tasks. The results are obtained from 5 individual runs for every setting. The best results for each task are highlighted with boldface font and the second bests are underlined.

Dataset	Model	Accuracy	IoU	Precision	Recall	F1
Gender Classification	Baseline	68.35 ± 1.00	13.68 ± 0.89	52.68 ± 0.61	56.34 ± 1.63	47.77 ± 1.14
	GRADIA	70.01 ± 1.47	16.66 ± 1.10	64.07 ± 2.07	51.84 ± 3.55	53.35 ± 3.08
	HAICS	69.29 ± 0.50	17.56 ± 0.79	60.06 ± 2.17	56.48 ± 2.13	54.90 ± 2.14
	RES-G	**71.33 ± 0.53**	22.97 ± 0.44	**76.47 ± 0.45**	63.90 ± 3.64	63.54 ± 2.29
	RES-L	70.39 ± 0.35	**23.60 ± 0.36**	76.32 ± 0.77	**65.75 ± 1.20**	**65.24 ± 0.74**
Scene Recognition	Baseline	93.42 ± 0.43	38.55 ± 0.22	**89.67 ± 0.07**	60.96 ± 0.56	68.47 ± 0.46
	GRADIA	95.03 ± 0.35	39.60 ± 1.13	87.98 ± 0.19	63.47 ± 2.24	70.80 ± 1.84
	HAICS	94.89 ± 0.20	41.29 ± 0.91	88.47 ± 0.53	66.23 ± 1.00	72.95 ± 0.87
	RES-G	**95.91 ± 0.31**	**45.97 ± 0.12**	87.54 ± 0.30	82.88 ± 1.14	82.90 ± 0.33
	RES-L	95.53 ± 0.54	44.64 ± 0.31	86.37 ± 0.08	**88.01 ± 0.39**	**84.78 ± 0.29**

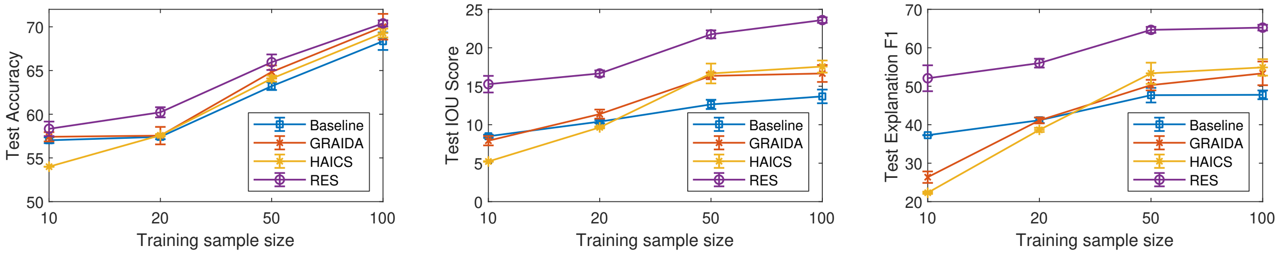


Figure 2: Model performance under different training sample size scenarios on gender classification dataset. The data point represents the mean value over 5 runs, and the error bar here corresponds to the standard deviation. (Left) The test prediction accuracy comparison. (Middle) The test IoU score comparison. (Right) The test explanation F1 score comparison.

the proposed RES framework is able to cope with the noisy label pretty well even under a very limited sample size, and consequently boosted the model performance in terms of prediction power as well as explainability robustly in all scenarios studied.

4.3 Qualitative Analysis of the Explanation

4.3.1 Case Studies. Here we provide some case studies about the model-generated explanation comparison for both gender classification and scene recognition datasets, as illustrated in Figure 3. Here we present the model-generated explanations as the heatmaps overlaid on the original image samples, where more importance is given to the area with a warmer color.

Gender Classification: As shown in the left four rows of Figure 3, we studied two 'male' class instances (top 2 rows) and two 'female' class instances (bottom 2 rows). As can be seen, in general, the explanation generated by the proposed RES models can more accurately focus on the important areas (e.g. the human face areas) for identifying the gender of the person in the image. In contrast, both the baseline model as well as the two comparison methods failed to generate reasonable explanation, as the models' 'attention' was distracted by some other objects presented in the images that are irrelevant to the gender classification task. For example, as shown in the first row on the left in Figure 3, where both a dog and a person are presented in the image sample. The explanation generated by the baseline and comparison methods assigned importance to the areas in between the dog and the person, and therefore could not focus properly on the person. On the other hand, both RES-G and RES-L learned to focus only on the person, and more specifically focused on human face area. Similar patterns could also be observed in the rest three rows on the left, demonstrating the powerful effect of the proposed RES framework on learning to generate more accurate explanations, and consequently enhances the explainability of the DNN models.

Scene Recognition: For the scene recognition dataset, as shown in the right four rows in Figure 3, we studied two instances of 'urban' scene (top 2 rows) and two instances of 'nature' scene (bottom 2 rows). Once again, we found that compared with the baseline model and other comparison methods, the explanations generated by RES models are more accurate and close to the ground truth for identifying whether the scene is taken from the urban areas or from wild nature. For instance, as shown in the third row on the right in Figure 3, the explanation generated by both the baseline and comparison methods focuses more on the water surface while RES focuses more on the wild animal itself. Similarly, as shown in the fourth row, the explanation generated by RES focuses more on the wildflowers rather than the grass-fields background. Although in those situations the prediction can be correct for all the models studied, we argue that the model trained with the RES framework can be more robust and have a batter generalizability power to the downstream predictive tasks by learning to assign importance more accurately to the most distinguishable features/patterns presented in the data samples.

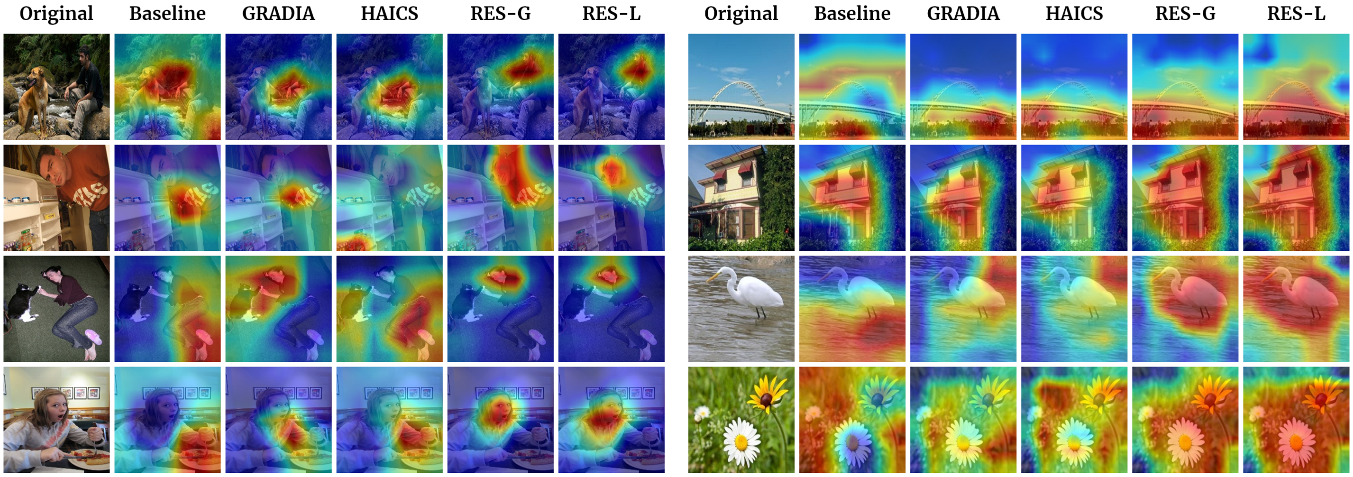


Figure 3: Selected explanation visualization results on gender classification dataset (left) and scene recognition dataset (right). The model-generated explanations are represented by the heatmaps overlaid on the original image samples, where more importance is given to the area with a warmer color.

Model Pairs	Perceived Quality (p-values)
Baseline vs. GRADIA	**2.68e-03**‡
Baseline vs. HAICS	**2.33e-04**‡‡
Baseline vs. RES-G	**4.98e-37**‡‡
Baseline vs. RES-L	**4.96e-28**‡‡
GRADIA vs. HAICS	0.4980
GRADIA vs. RES-G	**2.71e-22**‡‡
GRADIA vs. RES-L	**1.54e-15**‡‡
HAICS vs. RES-G	**1.67e-19**‡‡
HAICS vs. RES-L	**2.96e-13**‡‡
RES-G vs. RES-L	0.0824

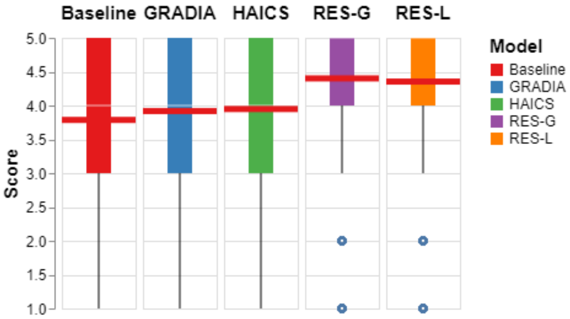


Figure 4: Top: results for pairwise comparison of five conditions. †: $p < 0.05$, ‡: $p < 0.01$, ‡‡: $p < 0.001$. Bottom: Distributions of human users' perceived attention quality ratings. 5-level Likert scale is used (5: Excellent, 4: Good, 3: Fair, 2: Bad, 1: Inferior).

4.3.2 Human Assessment. To evaluate the quality of explanations for the five comparison methods, we developed a web-based user interface (UI) where a human annotator can go over all the model-generated explanations and make qualitative evaluation on both datasets. We distributed the model-generated explanations from the test set to three separate human annotators. We asked annotators to assess the perceived quality of explanations with the five-level Likert scale. "5-Excellent" when explanations show positive attention very clearly while don't contain negative attention at all, and "4-Good" when positive attention is clearly presented with negligible negative attention. "3-Fair" meant that positive attention is partially seen while negative attention is clearly visible. "2-Bad" in case positive attention can be barely seen while negative can be found evidently. "1-Inferior" is assigned when a human annotator can only find negative attention. After performing the Shapiro-Wilk normality test, we found participants' ratings don't follow a normal distribution. Therefore, we applied Kruskal-Wallis H-test for identifying the differences between the five conditions.

According to Kruskal-Wallis H-test, the quality ratings of five models are significantly different, with a p-value of 7.82e-51 (< 0.05). For post-hoc pairwise comparisons using Dunn's test, all pairs are significantly different, with the exception of GRADIA vs. HAICS and RES-G vs. RES-L. This means that the ranking among the five conditions is that RES-G (M = 4.40, SD = 0.91) and RES-L (M = 4.35, SD = 0.89) are rated notably higher than the rest, followed by GRADIA (M = 3.92, SD = 1.24) and HAICS (M = 3.95, SD = 1.23). The least performing condition was Baseline (M = 3.79, SD = 1.25). Specific pair-wise testing results and visual representation between conditions are explained in Figure 4.

4.4 Sensitivity Analysis of Hyper-parameter α

Here we further provide a sensitivity analysis of the hyper-parameter α introduced in the proposed RES framework, as shown in Equation (5) which measures the tolerance level we give to the discrepancies between human annotation labels and the model explanation. Figure 5 shows the prediction accuracy, IoU, and explanation F1-score of the RES-L model for various values of α on the gender classification dataset. The scene recognition dataset follows a similar trend. The red dashed lines represent the baseline model's performance. In general, the model performance is not too sensitive to the value of α within the range studied, as all models outperformed the baseline model by a significant margin in terms of both prediction accuracy

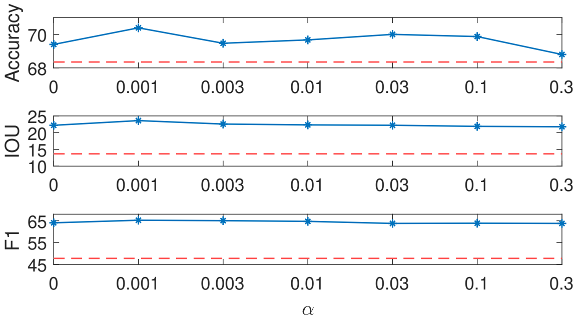


Figure 5: The sensitivity study of hyper-parameter α in RES framework (RES-L) on gender classification dataset. The red dashed lines represent the baseline model's performance.

as well as explainability. As we developed our models based on the accuracy of the validation set, we indeed observed a concave curvature on test accuracy, peaking at a α value between 0.001 and 0.1. While the specific best value of α can vary depending on the dataset as well as the degree of nosiness of the human annotation labels (such as the granularity of the annotation), in general, the proposed framework can perform well when α is relatively small (e.g. less than 0.1).

5 CONCLUSION

This paper proposes a generic framework for visual explanation supervision by developing a novel explanation model objective that can handle the noisy human annotation labels as the supervision signal with a theoretical justification of the benefit to model generalizability. Extensive experiments on two real-world image datasets demonstrate the effectiveness of the proposed framework on enhancing both the reasonability of the explanation as well as the performance of the backbone DNNs model.

Although the additional data of human explanation labels may not be easily accessible, our studies have demonstrated the effectiveness of the proposed RES framework under a quite limited amount of training samples, which could benefit application domains where data samples are limited and hard to acquire, yet both model performance as well as the explainability are on-demand, such as in medical domains. Furthermore, designing effective semi-supervised or weakly-supervised explanation supervision frameworks can be promising future directions to further overcome this limitation.

ACKNOWLEDGMENTS

This work was supported by the NSF Grant No. 1755850, No. 1841520, No. 2007716, No. 2007976, No. 1942594, No. 1907805, a Jeffress Memorial Trust Award, Amazon Research Award, NVIDIA GPU Grant, and Design Knowledge Company (subcontract number: 10827.002.120.04).

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] 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.

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[4] Guangji Bai and Liang Zhao. 2022. Saliency-regularized Deep Multi-task Learning. In *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

[5] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. 2021. To" See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-off. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–31.

[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] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry Jackel, Urs Muller, and Karol Zieba. 2016. Visualbackprop: visualizing cnns for autonomous driving. *arXiv preprint arXiv:1611.05418* 2 (2016).

[8] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*. Springer, 793–811.

[9] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. 2020. AiR: Attention with Reasoning Capability. In *European Conference on Computer Vision*. Springer, 91–107.

[10] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* 163 (2017), 90–100.

[11] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2019. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10705–10714.

[12] Yuyang Gao, Giorgio A Ascoli, and Liang Zhao. 2021. BEAN: Interpretable and efficient learning with biologically-enhanced artificial neuronal assembly regularization. *Frontiers in Neurorobotics* 15 (2021), 68.

[13] Yuyang Gao, Tong Sun, Rishab Bhatt, Dazhou Yu, Sungsoo Hong, and Liang Zhao. 2021. GNES: Learning to Explain Graph Neural Networks. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 131–140.

[14] Yuyang Gao, Tong Sun, Liang Zhao, and Sungsoo Hong. 2022. Aligning Eyes between Humans and Deep Neural Network through Interactive Attention Alignment. arXiv:2202.02838 [cs.AI]

[15] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820* (2018).

[16] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[17] 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.

[18] Alon Jacovi and Yoav Goldberg. 2020. Aligning Faithful Interpretations with their Social Attribution. *arXiv preprint arXiv:2006.01067* (2020).

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

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[21] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. 2018. Learning what and where to attend. *arXiv preprint arXiv:1805.08819* (2018).

[22] Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2019. Embedding Human Knowledge into Deep Neural Network via Attention Map. *arXiv preprint arXiv:1905.03540* (2019).

[23] 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.

[24] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition* 65 (2017), 211–222.

[25] Badri Patro, Vinay Namboodiri, et al. 2020. Explanation vs attention: A two-player game to obtain attention for VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11848–11855.

[26] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring human-like attention supervision in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of ACM*

International Conference on Knowledge Discovery and Data Mining (KDD'16). ACM, 1135–1144. https://doi.org/10.1145/2939672.2939778

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

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

[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 the 2021 CHI Conference on Human Factors in Computing Systems*. 1–8.

[32] Ryan Turner. 2016. A model explanation system. In *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)*. IEEE, 1–6.

[33] Roman Visotsky, Yuval Atzmon, and Gal Chechik. 2019. Few-shot learning with per-sample rich supervision. *arXiv preprint arXiv:1906.03859* (2019).

[34] Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, Yanzhi Wang, and Xiang Chen. 2019. Interpreting and evaluating neural network robustness. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 4199–4205.

[35] Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. 2019. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 349–357.

[36] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).

[37] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.

[38] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

A APPENDIX

A.1 Proof of Theorem 1

Proof. Suppose f_ψ is a δ-minimizer of $\mathcal{L}$ with $\psi \in \Theta$. From Assumption 1, we know that there exists a neural network f_τ such that

$$\|f_\tau - f_{\theta^*}\|^2 := \mathbb{E}\left[|f_\tau - f_{\theta^*}|^2 + |\nabla f_\tau - \nabla f_{\theta^*}|^2\right] \leq C_1^2 \frac{\|\theta^*\|^2}{m^\gamma} \tag{15}$$

Then, we have

$$\begin{aligned}\mathcal{L}(f_\psi) - \mathcal{L}(f_{\theta^*}) &\leq \mathcal{L}(f_\tau) - \mathcal{L}(f_{\theta^*}) + \delta \\ &\leq L_1 \mathbb{E}\left[|f_\tau(x) - f_{\theta^*}(x)\right] + L_2 \mathbb{E}\left[|\nabla f_\tau(x) - \nabla f_{\theta^*}(x)\right] + \delta \\ &\leq (L_1 + L_2)\frac{C_1\|\theta^*\|}{m^{\gamma/2}} + \delta\end{aligned} \tag{16}$$

From Assumption 2, given $0 < \epsilon < 1$, we have

$$P(|\mathcal{L}(f_\theta) - \hat{\mathcal{L}}(f_\theta)| \leq \frac{C_2(V, m, \epsilon)}{\sqrt{N}}) \geq 1 - \epsilon, \quad \forall\, \theta \in \Theta \tag{17}$$

Then,

$$\begin{aligned}\mathcal{L}(f_{\hat{\theta}}) - \mathcal{L}(f_{\theta^*}) &\leq \hat{\mathcal{L}}(f_{\hat{\theta}}) - \mathcal{L}(f_{\theta^*}) + \frac{C_2(V, m, \epsilon)}{\sqrt{N}} \\ &\leq \hat{\mathcal{L}}(f_\psi) - \mathcal{L}(f_{\theta^*}) + \frac{C_2(V, m, \epsilon)}{\sqrt{N}} + \delta \\ &\leq \mathcal{L}(f_\psi) - \mathcal{L}(f_{\theta^*}) + \frac{C_2(V, m, \epsilon)}{\sqrt{N}} + \delta \\ &\leq (L_1 + L_2)\frac{C_1\|\theta^*\|}{m^{\gamma/2}} + \frac{2C_2(V, m, \epsilon)}{\sqrt{N}} + 2\delta\end{aligned} \tag{18}$$

□

A.2 Proof of Lemma 1

Proof. Since

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{19}$$

where the last equality follows by multiplying by $\frac{e^{-x}}{e^{-x}} = 1$. And since: $\lim_{x \to \infty} 1 - e^{-2x} = 1$, and $\lim_{x \to \infty} 1 + e^{-2x} = 1$, we have

$$\lim_{x \to \infty} tanh(x) = 1 \tag{20}$$

Similarly, we also have

$$\lim_{x \to -\infty} tanh(x) = \lim_{x \to -\infty} \frac{e^{2x} - 1}{e^{2x} + 1} = -1 \tag{21}$$

Thus we have

$$\lim_{\gamma \to \infty} tanh(\gamma(M^{(i)} - a)) = \begin{cases} 1 & M^{(i)} > a \\ -1 & M^{(i)} < a \end{cases} \tag{22}$$

Thus we have the equivalency of Equation (8) and Equation (5) when $\gamma \to \infty$. □

A.3 Efficient Adaptive Threshold Searching Algorithm

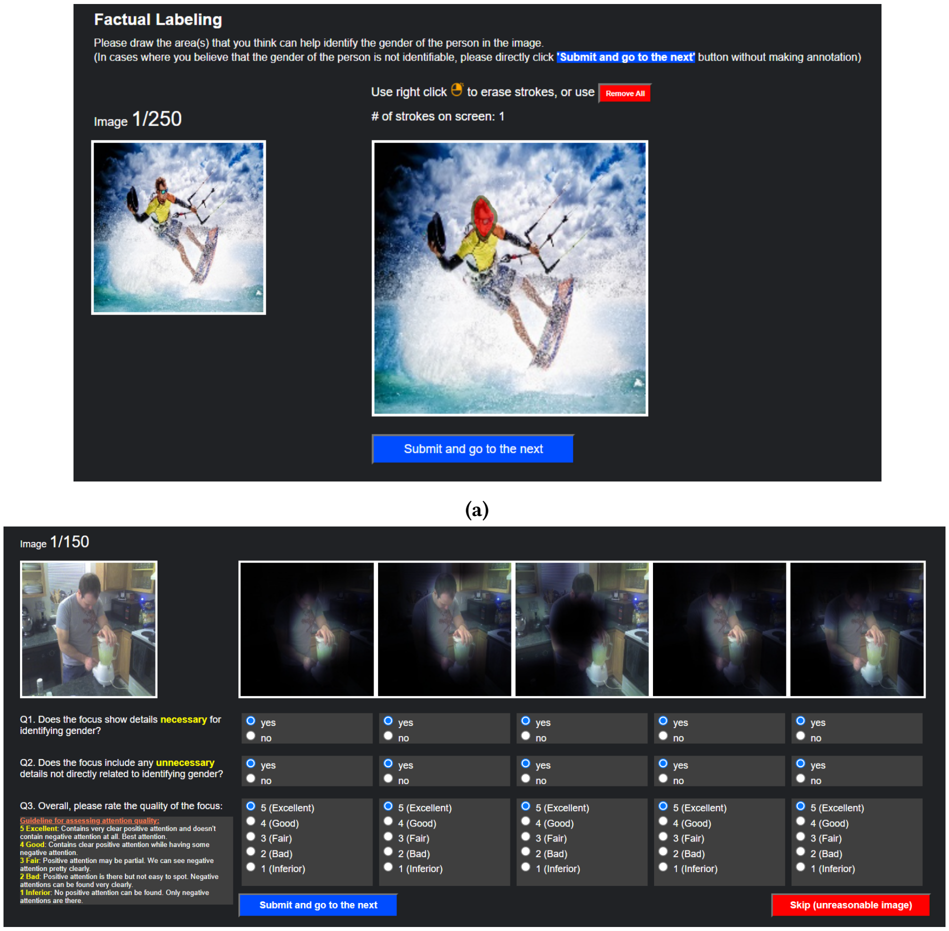
Algorithm 1: Adaptive Threshold Searching Algorithm

```
Require: $M, F, C$
Ensure: solution $a$
initialize: $a = 0, act = 0, v = 0, vct = 0, i = 0, j = 0$
$ge = \{M[find(C > 0)]\}$ % find the set of greater or equal to inequality constraints
$l = \{M[find(F > 0)]\}$ % find the set of less to inequality constraints
$ges = \mathrm{Sort}(ge, \text{'ascend'})$
$ls = \mathrm{Sort}(l, \text{'descend'})$
for $i < |ges|$ do
    $v = ges[i]$
    $vct = i + 1 + \mathrm{BinarySearch}(v, ls)$
    if $vct > act$ then
        $a = v$
        $act = vct$
    end if
    $i = i + 1$
end for
for $j < |ls|$ do
    $v = ls[i]$
    $vct = j + 1 + \mathrm{BinarySearch}(v, ges)$
    if $vct > act$ then
        $a = v$
        $act = vct$
    end if
    $j = j + 1$
end for
```

A.4 Human Annotation and Evaluation UI demonstration

Figure 6 (a) is the interface used to collect attention annotation on the areas people think are relevant to the classification task. For example, for the gender dataset annotation, users first determine whether they can identify the person's gender in the image, then draw the areas that help them for the gender classification. In the back-end, the coordinates of highlighted areas are converted into a binary map, preparing for the modeling step.

Figure 6 (b) is the interface for human assessment on the model-generated explanations. For each image annotation, 5 explanations were presented in random order with 3 questions (Q1 and Q2 are true/false questions, Q3 is a 5-point Likert scale rating question) asked for each explanation. Question 1 asks if the focus on the explanation shows details necessary for identifying the target label



Factual Labeling

Please draw the area(s) that you think can help identify the gender of the person in the image.
(In cases where you believe that the gender of the person is not identifiable, please directly click 'Submit and go to the next' button without making annotation)

Use right click to erase strokes, or use Remove All

# of strokes on screen: 1

Image 1/250

Submit and go to the next

(a)

Image 1/150

Q1. Does the focus show details **necessary** for identifying gender?	● yes ○ no	● yes ○ no	● yes ○ no	● yes ○ no	● yes ○ no
Q2. Does the focus include any **unnecessary** details not directly related to identifying gender?	● yes ○ no	● yes ○ no	● yes ○ no	● yes ○ no	● yes ○ no
Q3. Overall, please rate the quality of the focus:	● 5 (Excellent) ○ 4 (Good) ○ 3 (Fair) ○ 2 (Bad) ○ 1 (Inferior)	● 5 (Excellent) ○ 4 (Good) ○ 3 (Fair) ○ 2 (Bad) ○ 1 (Inferior)	● 5 (Excellent) ○ 4 (Good) ○ 3 (Fair) ○ 2 (Bad) ○ 1 (Inferior)	● 5 (Excellent) ○ 4 (Good) ○ 3 (Fair) ○ 2 (Bad) ○ 1 (Inferior)	● 5 (Excellent) ○ 4 (Good) ○ 3 (Fair) ○ 2 (Bad) ○ 1 (Inferior)

Guideline for assessing attention quality:
5 Excellent: Contains very clear positive attention and doesn't contain negative attention at all. Best attention.
4 Good: Contains clear positive attention while having some negative attention.
3 Fair: Positive attention may be partial. We can see negative attention pretty clearly.
2 Bad: Positive attention is there but not easy to spot. Negative attentions can be found very clearly.
1 Inferior: No positive attention can be found. Only negative attentions are there.

Submit and go to the next

Skip (unreasonable image)

(b)

Figure 6: The screenshots illustrating the two UIs for human annotation and evaluation. (a) The interface for attention annotation where users can draw on the image and generate a binary matrix of the focus area used for improving model explanation quality. (b) The interface for attention quality assessment where 5 model-generated explanations are displayed in random order. Users will answer three questions for each explanation.

(i.e., labels in gender classification or scene recognition), and question 2 asks for the presence of unnecessary details on the image for identifying the target. Question 3 is our main focus of the attention quality assessment, where annotators give 1 to 5 ratings to each model explanation.

A.5 Detailed Implementation of the Learnable Imputation Layers

For the learnable imputation function, we studied both a shallow implementation as well as a deep implementation, as shown in detail below:

Shallow Implementation: We apply one layer of convolution operation to process the raw human annotation label, with a 64×64 convolution kernel with a padding size of 16 and a stride of 32.

Deep Implementation: We apply five layers of convolution operations to process the raw human annotation label, with 7×7, 3×3, 3×3, 3×3, and 3×3 convolution kernel with a padding size of 3 on the first layer and 1 for the rest layer, and a stride 2 for all layers.

We choose the Shallow implementation for the RES-L model as it achieves better performance on the validation set. The reason why the deep version gets inferior performance could be due to the training sample size studied in this work is too small.