

Going Beyond XAI: A Systematic Survey for Explanation-Guided Learning

YUYANG GAO, Emory University, USA

SIYI GU, Emory University, USA

JUNJI JIANG, Emory University, USA

SUNGSOO RAY HONG, George Mason University, USA

DAZHOU YU, Emory University, USA

LIANG ZHAO*, Emory University, USA

As the societal impact of Deep Neural Networks (DNNs) grows, the goals for advancing DNNs become more complex and diverse, ranging from improving a conventional model accuracy metric to infusing advanced human virtues such as fairness, accountability, transparency (FaccT), and unbiasedness. Recently, techniques in Explainable Artificial Intelligence (XAI) are attracting considerable attention, and have tremendously helped Machine Learning (ML) engineers in understanding AI models. However, at the same time, we started to witness the emerging need beyond XAI among AI communities; based on the insights learned from XAI, how can we better empower ML engineers in steering their DNNs so that the model's reasonableness and performance can be improved as intended? This article provides a timely and extensive literature overview of the field **Explanation-Guided Learning** (EGL), a domain of techniques that steer the DNNs' reasoning process by adding regularization, supervision, or intervention on model explanations. In doing so, we first provide a formal definition of EGL and its general learning paradigm. Secondly, an overview of the key factors for EGL evaluation, as well as summarization and categorization of existing evaluation procedures and metrics for EGL are provided. Finally, the current and potential future application areas and directions of EGL are discussed, and an extensive experimental study is presented aiming at providing comprehensive comparative studies among existing EGL models in various popular application domains, such as Computer Vision (CV) and Natural Language Processing (NLP) domains.

Additional Key Words and Phrases: Explainable AI (XAI), Explainability, Faithfulness, Trustworthiness, Bias, FaccT, Deep Neural Networks, Deep Learning, Explanation-Guided Learning (EGL), Explanation Supervision, Attention Supervision, Explanation Alignment, Learning from Explanation,

ACM Reference Format:

Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. 2022. Going Beyond XAI: A Systematic Survey for Explanation-Guided Learning. 1, 1 (December 2022), 39 pages. <https://doi.org/XXXXXXX.XXXXXXX>

*Corresponding author

Authors' addresses: Yuyang Gao, yuyang.gao@emory.edu, Emory University, Atlanta, Georgia, USA, 30322; Siyi Gu, carrie.gu@emory.edu, Emory University, Atlanta, GA, USA; Junji Jiang, jjian50@emory.edu, Emory University, Atlanta, GA, USA; Sungsoo Ray Hong, shong31@gmu.edu, George Mason University, Fairfax, VA, USA; Dazhou Yu, dazhou.yu@emory.edu, Emory University, Atlanta, GA, USA; Liang Zhao, liang.zhao@emory.edu, Emory University, Atlanta, GA, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

XXXX-XXXX/2022/12-ART \$15.00

<https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

In recent years, techniques in Explainable Artificial Intelligence (XAI) are attracting considerable attention [2, 9, 58], and have gradually become the dominating ways that connect the way Deep Neural Networks (DNNs) work and human reasoning [63, 86]. As DNNs cannot provide human sensible “global structure” of how the model works unlike white-box models, XAI has become an imperative tool that Machine Learning (ML) engineers always use to “make sense” of the way their models work [52]. In recent years, many XAI techniques have been proposed in an effort to open the “black box” of DNNs [58], such as techniques that provide saliency maps for understanding which sub-parts (i.e., features) in an instance are most responsible for the model prediction [11, 101, 102, 127, 173]. Despite the recent fast progress on XAI techniques for DNNs, the majority of the research body in XAI put focus on handling “how to generate the explanations” while showing less attention to advanced questions like “whether the explanations are reasonable/accurate”, “what if the explanations are unreasonable/inaccurate”, and most importantly, “how to adjust the model to generate more reasonable/accurate explanations in the future”. We are starting to witness the emerging need beyond XAI; based on the insights learned from XAI, how can we better steer DNNs such that their future behavior can be improved from the insights learned from XAI techniques? We argue that understanding how to convert insights learned from XAI-driven techniques to steer DNNs would be the key to realizing the DNNs to be more powerful, fair, accountable, transparent, unbiased, and trustworthy, unraveling many real-world application areas.

In recent years, several new areas have emerged which aim at gaining a thorough grasp of the model behavior through the model explanation. Explanatory Debugging is one area of research that has gained popularity [75, 83, 153]. Interactive techniques and systems were developed to enable human users to interactively select features of interest and then investigate how the model behaves in the resulting subspaces for debugging purposes. Another interesting area of research compared the explanation provided by DNNs and the explanation provided by humans to gain a better understanding of the models’ behavior [33, 142]. Although the aforementioned studies are capable of providing more insights about whether the explanations are accurate or reasonable, they are yet to be sufficient for further handling how we can learn from those mistakes, and consequently adjust the model to get better quality explanations and enhance the model performance.

Recently, a new line of research that aims to intervene ML model’s behavior through XAI techniques has started to emerge. In particular, the approaches jointly improve DNNs in terms of both their explainability and generalizability by applying additional supervision signals or prior knowledge onto the model reasoning process to direct the model explanation derived from established XAI techniques. This direction is generally named Explanation Guided Learning (EGL) [65, 84, 123, 137], while several other terms such as *Explanation Supervision* [50–52], *Attention Supervision* [116, 169], *Explanation Alignment* [119, 165], as well as *Learning from Explanation* [22, 124, 162] are also frequently used under the same umbrella.

Recently, there has been a surge of research that both proposes and applies new approaches in numerous application areas, including Computer Vision (CV), Natural Language Processing (NLP), and Visual Question Answering (VQA). Despite the fact that EGL techniques are generally still in their early stage, the majority of existing studies have produced encouraging results, showing that the main DNNs can generally benefit from the additional explanation objective in terms of both model explainability and generalizability to unseen data across various application domains. However, developing EGL frameworks can be difficult due to significant technical challenges caused by its unique characteristics, including:

- (1) **Gap between the pattern of model explanation and human explanation:** The explanation generated by model explainers is typically continuous values, whereas human annotations are

typically binary. Therefore, it is difficult to align the human explanation directly with the model explanation without significant efforts to fill the gap between the data domain and distributions.

- (2) **Difficulty in comprehensively evaluating the EGL models:** unlike the conventional model where the task performance is the main focus, the quality of EGL outcomes generally needs sophisticated and carefully devised evaluation procedures that are often naturally subjective. For example, human participants can be involved in the evaluation to assess the quality of the model explanation. Moreover, beyond XAI explanation, EGL further requires the joint evaluation of the accuracy of prediction, explanation, and their mutual relation towards the reflection of model generalizability. Thus, we lack systematic standardization and comprehensive summarization approaches with which to evaluate the various EGL methodologies that have been proposed.
- (3) **Noisiness in human annotation labels:** Unlike predictive task labels, it is much more likely for human annotators to unintentionally create noisy annotation labels where either the real important features are missed or irrelevant features are mistakenly included in the explanation annotation. For instance, when annotation the image data, some important object parts or even the entire objects may be missed by the coarsely drawn boundary from human annotators. Thus, applying naive supervision directly to train the model can lead to falsely excluded non-trivial features from the input space that are important to the prediction [50].
- (4) **Difficulty in explicitly measuring the faithfulness of the explanation quality with respect to the model generalizability:** Due to the fact that EGL techniques are generally still in their infancy, most existing works still primarily focus on merely evaluating the explanation quality of the EGL model independently of the model task performance. The faithfulness of the improved explanation quality with respect to the model prediction is yet to be explored explicitly and can be a key research question to be answered for EGL techniques to further advance and enhance the model performance and generalizability.

1.1 Contributions

As the majority of existing EGL approaches were built for a specific application domain, cross-referencing these techniques across application domains serving different communities is problematic and challenging. Moreover, the lack of a comprehensive review and taxonomy of existing techniques and applications in EGL creates substantial challenges for researchers working in the related field, since they lack clear information on existing bottlenecks, pitfalls, open-ended questions, and potentially fruitful future research directions.

To this end, this paper provides a systematic survey of EGL models across various application domains, including Computer Vision (CV) [37, 44, 84, 99, 119, 121, 139, 152], Natural Language Processing (NLP) [8, 10, 15, 21, 22, 24, 26, 27, 36, 53, 55, 66, 82, 90, 91, 137, 138, 140, 162, 163, 168, 170, 172], Visual Question Answering (VQA) [25, 48, 109, 116, 155, 169], and more in Section 4. The goal of the survey is to help interdisciplinary researchers build a better understanding of the existing EGL techniques, and develop appropriate frameworks to solve the problems in their applications domains. Besides, this survey also aims at helping researchers outside the AI communities to understand the basic principles as well as identify interdisciplinary open research opportunities in the EGL domain. As far as we know, this is the first comprehensive survey on explanation-guided learning. This work's contributions are as follows:

- We summarize a general learning paradigm of EGL based on existing works in this field to provide overall guidance on identifying and designing new EGL techniques.
- We identify the key factors in terms of comprehensively evaluating the EGL model's performance, and then provided a summarization and categorization of the existing evaluation procedures and metrics.

- We propose a taxonomy of explanation-guided learning categorized by the level of guidance and methodologies. The advantages, drawbacks, as well as relations among different subcategories of EGL techniques, are also introduced and compared.
- We introduce the broader application of EGL and detail the unique benefits and future opportunities for each application domain.
- We conduct a comprehensive experimental analysis and comparative study among existing EGL models in CV and NLP domains.
- We summarize the existing literature on EGL at the current stage, and then provide a set of open problems and potential promising future research directions of EGL.

1.2 Relationship with Related Surveys

This section outlines previously published surveys that have some relevance to Explanation-Guided Learning. These surveys can be classified into three topics: (1) XAI technique and evaluation, (2) AI ethics, and (3) interactive machine learning, as introduced in detail below.

Explainability Technique and Evaluation: The related surveys of interpretability techniques provide a technical review and categorization of existing explanation techniques that can explain the machine learning model, especially for the sophisticated ‘black box’ DNN models. Several related surveys provide an in-depth classification of machine learning interpretability methods in general [9, 58, 89, 120], while others focus on more specific fields of study. Specifically, Burkart et al. [20] review the explainability methods of supervised machine learning models. Montavon et al. [103] provide a survey that specifically focuses on the interpretability techniques designed for explaining DNNs. Zhang et al. [166] research interpretability techniques for Convolutional Neural Networks (CNN) and visual explanation. Tjoa et al. [146] summarized the XAI techniques that have been adopted for explaining medical data. Along the line of interpretability techniques, many recent surveys also review the methods and metrics for comprehensively validating the quality of the explanation generated by the XAI techniques [62, 100, 175].

AI Ethics: As the societal impact of AI grows, the goals for revising AI become more complex and diverse, ranging from improving a conventional model accuracy metric to infusing advanced human virtues such as fairness, accountability, transparency (FaccT), and unbiasedness [92]. Aligning to such direction, recent surveys started to collect, synthesize, and structuralize the existing approaches meant to be designed to handle several types of bias in AI [23, 95]. The most noteworthy finding in our survey for the landmark surveys is that the approaches for detecting bias in ML are more than the ways to mitigate the bias [17, 39]. The second important finding is that even though several studies focus on showing the ways to detect bias, they also present a hint of how we can mitigate them by showing some typical bias cases [95, 106]. Lastly, the existing survey also provides a pressing field needs explaining why we need to improve the ways to steer models in the case of witnessing the evidence of bias [63].

Interactive Machine Learning: Since Fails et al. [45] proposed the idea of interactive ML, the HCI community has put a high priority on applying XAI techniques in developing interactive techniques and systems meant to help ML engineers to better understand their models’ weaknesses and strengths. Landmark surveys related to human factor and interaction can be categorized into 1) the interactive design—emphasizing how to design the feedback loop between humans and ML models through system [40, 68] that are widely proposed in the human factor research communities, such as SIGCHI, CSCW, and UIST, and 2) visual analytic—focusing on how to apply visualization techniques to help ML engineers understanding complex ML model behavior [42, 161].

1.3 Outline of the survey

The remaining part of the survey is organized as follows. In Section 2, we introduce the problem formulation and performance evaluations of EGL models, as illustrated in Figure 1 and Table 1. In Section 3, we provide the taxonomy of EGL categorized by the level of guidance and methodologies, as illustrated in Figure 2. Moreover, the details of each EGL technique, along with their corresponding advantages, drawbacks, and relations to other techniques in the same subcategories are provided. In Sections 4 and 5, we first introduce the broader application of EGL and then conduct a comprehensive experimental analysis and comparative study among existing EGL models in both CV and NLP domains. Lastly, we conclude the current development of EGL techniques and suggest several open problems and potential future research directions in Section 6.

2 PROBLEM FORMULATION AND PERFORMANCE EVALUATIONS

This section begins by introducing the generic denotation and formulation of the Explanation-Guided Learning problem (Section 2.1) and then considers ways to categorize the performance evaluation measures of Explanation-Guided Learning (Section 2.2).

2.1 Problem formulation

Consider a differentiable model f parameterized by θ that learns to fit inputs $X \in \mathbb{R}^{N \times D}$ and the corresponding one-hot class labels $Y \in \mathbb{R}^{N \times K}$, where N denotes the total number of data samples, D denotes the input dimension and K denotes the number of classes. An explainer g is considered to extract the explanation M from the model f given its parameter θ and a set of data points $\langle X, Y \rangle$. Generally speaking, the model explanation M represents the marginal contribution of each input feature to the model's decision after all possible combinations have been considered. Notice that in this paper we use the terms *rationale*, *attention*, and *saliency maps* interchangeably as the specific form of M that is frequently used by the corresponding application domains. Depending on the way the explanation is calculated, M can be generally represented by either local explanation $M^{(L)}$ where $M_i^{(L)}$ is the local explanation of model f with respect to sample $\langle X_i, Y_i \rangle$, or a single global explanation $M^{(G)}$ of the model f .

The Explanation-Guided Learning (EGL) paradigm. The general goal for Explanation-Guided Learning is to boost both the task performance as well as the interpretability of the backbone model by jointly optimizing model prediction as well as the explanation. Based on the earlier exploration of explanation supervision frameworks design [50, 51, 99, 122], we introduce the key objective function of Explanation-Guided Learning as follows:

$$\min \underbrace{\mathcal{L}_{\text{Pred}}(f(X), Y)}_{\text{task supervision}} + \underbrace{\alpha \mathcal{L}_{\text{Exp}}(g(f, \langle X, Y \rangle), \hat{M})}_{\text{explanation supervision}} + \underbrace{\beta \Omega(g(f, \langle X, Y \rangle))}_{\text{explanation regularization}} \quad (1)$$

where \hat{M} explicitly incorporates the 'right' explanation, which can be typically realized by human annotation masks [36, 52].

As shown in Equation (1), the key objective function of Explanation-Guided Learning mainly consists of three terms, namely 1) task supervision term for the typical prediction loss (such as the cross-entropy loss), 2) explanation supervision term for supervising the model explanation with some explicit knowledge of what the 'right' explanation should be, and 3) explanation regularization term for enforcing some general properties about the 'right' explanation (such as maintaining the sparsity nature of the explanation). Notice that all three terms above can be defined and implemented differently depending on each particular explanation-guided learning method.

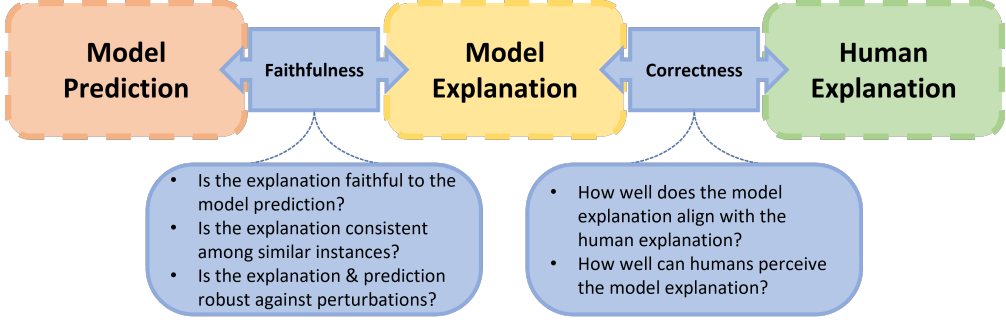


Fig. 1. The gaps in Explanation-Guided Learning performance evaluations.

Table 1. Detailed evaluation measures categorized by the gaps.

Category	Evaluation Measure
Explanation Faithfulness	Perturbation based [10, 27, 36, 38, 65, 99, 136, 165]
	Explanation Consistency based [10, 136]
Explanation Correctness	Case study based [26, 38, 109, 168]
	Human annotation based ¹ [22, 25, 36, 38, 48, 55, 60, 66, 84, 91, 99, 109, 136, 159] User study based (user-perceived understandability) [27, 66, 112, 144]

2.2 Performance evaluations

Unlike the evaluation of conventional machine learning models that typically only focus on the goodness of performance of the model, and the evaluation of traditional explainable AI models that only focus on the quality of the generated model explanation, Explanation-Guided Learning essentially jointly investigates the model prediction performance, the quality of model explanation, and their relation. As illustrated in Figure 1, we identify and categorize two types of evaluations that are essential in measuring the performance of EGL models, namely the *faithfulness* and *correctness* of the model explanation. Here we summarize the existing evaluation metrics into the two categories in Table 1 and introduce each type of metric in great detail in the following two subsections.

2.2.1 Metrics on evaluating explanation faithfulness. Here we introduce the metrics for explanation faithfulness (model prediction v.s. model explanation) evaluation, which aims at evaluating how the model-generated explanation influences the corresponding model’s prediction.

Perturbation-based evaluations: To evaluate the faithfulness of the model explanation, the study of how different types of perturbations on the input space influence the model prediction has become a very common and well-received approach in the literature [10, 27, 36, 38, 65, 99, 136, 165]. Existing measures can be mainly categorized into three groups, depending on the type of perturbation as follows:

- **Occlusion-based perturbation:** These metrics basically study how much influence on the model’s prediction if the important feature or rationale identified by the model explanation are occluded or masked from the original sample [27, 36, 65, 99, 165]. One commonly used occlusion-based metric is *comprehensiveness* [36], where the difference of the predicted probability from the model $f(\cdot)$ for the same class Y_i is compared between the original input X_i and $X_i \setminus g(f, \langle X_i, Y_i \rangle)$, where the operation ‘ \setminus ’ represents the exclusion of the supporting rationales $g(f, \langle X_i, Y_i \rangle)$ from input X_i . Mathematically, Comprehensiveness can be defined as follows:

$$\text{Comprehensiveness} = f(X_i)_{Y_i} - f(X_i \setminus g(f, \langle X_i, Y_i \rangle))_{Y_i} \quad (2)$$

Besides the comprehensiveness score, many other intuitive methods are also used to evaluate the quality of the explanation. Inspired by previous works [104, 128], a common intuitive strategy to measure the faithfulness of the explanation used by existing works [27, 99, 165] is to track the degradation of model performance by removing importance features (often in decreasing order) from the input.

- **Insertion-based perturbation:** These metrics study how well the prediction aligns between the original sample and an artificially generated sample where only the important feature/rationales are included [10, 36, 99, 136]. One popular metric is Sufficiency [36], which captures the degree to which the snippets within the extracted rationales $g(f, \langle X_i, Y_i \rangle)$ are adequate for a model to make a prediction. Concretely, it can be defined as follows:

$$\text{Sufficiency} = f(X_i)_{Y_i} - f(g(f, \langle X_i, Y_i \rangle))_{Y_i} \quad (3)$$

Similarly, many other intuitive methods are also used following the insertion idea. A common strategy used by existing works [10, 99] is to track the increase in model performance by gradually inserting the important features (often in decreasing order) from the input.

- **Adversarial perturbation:** These metrics in general check whether the model explanation is still faithful to the model prediction under adversarial attacks [38, 165]. For instance, [165] leveraged the sanity check method originally proposed by [3] to check if attribution maps look different when the deep network being explained is extremely perturbed or under adversarial attacks. The intuition behind this measure is that a faithful attribution method should yield different explanations for the randomized model.

Consistency-based evaluations: Besides the perturbation-based metrics which only focus on evaluating each instance locally at a time, existing works also propose consistency-based evaluation, where more global evaluation metrics have been proposed to validate how well the explanation aligns across similar instances [10, 136]. More specifically, [10] proposed a metric called *Data Consistency* that measures how similar the explanations for similar instances are. Although the specific equation of the measurement in the paper is specifically designed for NLP and generative explanation, the basic idea can be generally expressed as follows:

$$\text{Data Consistency} = |g(f, \langle X_i, Y_i \rangle) - g(f, \langle X_i \setminus M, Y_i \rangle)| \quad (4)$$

where M is a random mask that masks out K input features from X_i ; K will be treated as a hyper-parameter depending on the dataset. In short, the general assumption behind this is that the model explanation between very similar samples should also be close to each other, so higher values represent better performance. Besides, the authors also suggested that it can also serve as an additional regularisation term during training for the model to be consistent in the generated explanations.

Similar to the above idea, another work employed Intersection over Union (IoU) score to measure explanation stability across similar instances [136]. Specifically, they proposed to find similar instances by searching for the nearest neighbors of X_i in the dataset based on both the semantic similarity – cosine of their BERT representations; and the lexical similarity – the ratio of overlapping n-grams.

2.2.2 Metrics on evaluating explanation correctness. Here we introduce the metrics for explanation correctness evaluation, which aims at evaluating how well the model-generated explanation aligns with the human explanation annotation or how well can humans perceive the model-generated explanation.

Case study: Case study has been widely used as a conventional method for qualitatively evaluating the explanation generated by the model [26, 38, 109, 168], where a set of instances and their

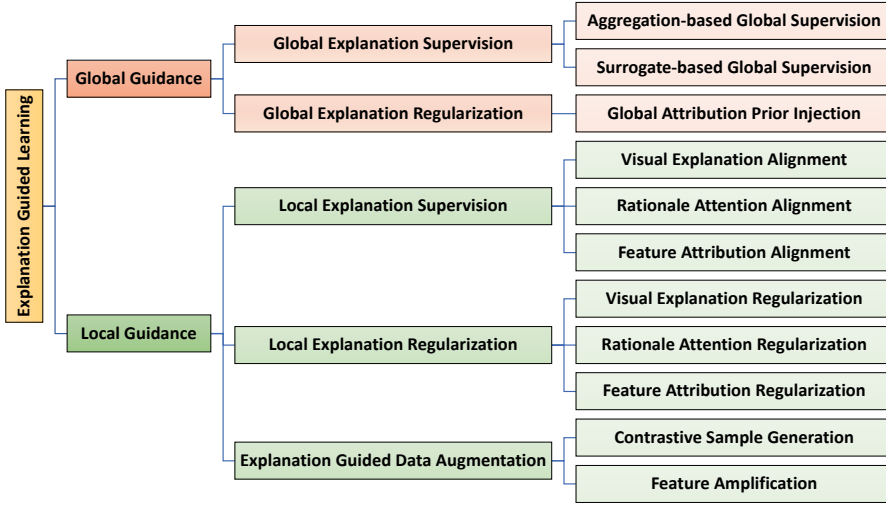


Fig. 2. Taxonomy of Explanation-Guided Learning problems and techniques.

corresponding model explanations are selected and investigated qualitatively. Although making qualitative assessments and detailed analyses of just a few samples can be easily achieved, it is in general less scientifically rigorous and the claims or conclusions are prone to be biased due to the author's subjectivity.

User study (user-perceived understandability): User study, specifically user-perceived understandability, has been commonly used as a qualitative evaluation method to assess how humans can understand the explanation generated by the model [24, 66, 112, 144]. The user-perceived understandability methods are typically achieved by developing a user interface to show the model explanations to human subjects, and collecting the rating of how likely the important features identified by the model explanation can lead to the correct prediction of the underlying ground-truth label.

Human annotation-based evaluation: Explanation alignment is a unique yet commonly used quantitative metric in Explanation-Guided Learning which measures how the human-annotated ground truth explanation is aligned with the model generated explanation [22, 25, 36, 38, 48, 55, 60, 66, 84, 91, 99, 109, 136, 159]. The distance is commonly measured by the Intersection over Union (IoU) score [36, 91, 136], precision, recall, and F-1 scores [55, 136].

2.2.3 Other general metrics. Besides measuring the faithfulness and correctness of model explanation, most of the papers also included the conventional model task performance metrics to verify if the Explanation-Guided Learning actually helped the generalizability of the backbone DNN models. Like most papers working on classification tasks, the common metrics used to evaluate model performance are accuracy, AUC (Area Under the ROC Curve) score, and F1 score.

3 EXPLANATION-GUIDED LEARNING TECHNIQUES

This section focuses on the taxonomy and representative techniques utilized for each category and subcategory. According to the level at which the model explanation is obtained and supervised, the technique types for EGL can be divided into global guidance and local guidance, as shown in Figure 2. Specifically, global guidance focuses on the model's global explanation and refines the model's overall decision-making process, while local guidance guides the model with each

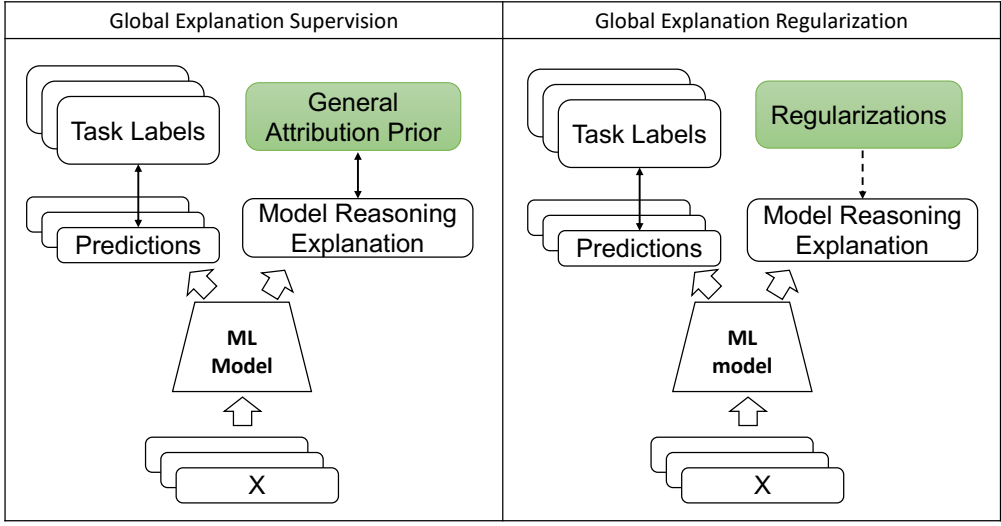


Fig. 3. Illustration of Global Guidance techniques. Specifically, global explanation supervision techniques (left) aim at providing supervision in terms of model attribution, while global explanation regularization techniques (right) aim at confining the model reasoning process with prior knowledge.

sample-specific explanation. The aforementioned techniques are then further categorized in terms of the way explanation guidance is injected during the course of model training.

3.1 Global Guidance

Global explanation guidance focuses on injecting prior knowledge or adding supervision signals to improve the model's global explanation that explains the decision-making process of the model in general. Based on the way explanation guidance is injected, global explanation guidance methods can be categorized into two types: 1) Global Explanation Supervision: The ground truth explanation labels are provided as an additional supervision signal to train the feature-wise explanation of the model; and 2) Global Explanation Regularization: in which some regularization terms that represent some general prior knowledge about the model explanation are added to regularize the feature-wise explanation of the model, as illustrated in Figure 3.

3.1.1 Global Explanation Supervision. The techniques proposed in global explanation supervision [44, 90, 152] aim at providing a single feature-wise explanation of the model globally. Compared with instance-level local explanation supervision where the explanation ground truth is provided for each instance [50, 52, 99], global explanation aims to provide a more effective global guide to the model's behavior as a whole. Depending on the strategies to compute the global explanation of the model, current literature can be mainly categorized in two directions: 1) aggregation-based [44, 90] and 2) surrogate-based [32, 114, 148].

Aggregation-based Global Supervision: This type of method typically achieves explanation supervision by first estimating the global feature attribution via aggregating local feature attribution of each sample and aligning it with a single ground truth feature attribution vector \hat{m} as the additional supervision signal to train the model jointly with the conventional task loss. The common techniques used to calculate each sample's feature attribution are integrated gradient [90] and the expected gradient proposed by Erion et al. [44]. Specifically, the objective function for

aggregation-based global supervision can be summarized as follows:

$$\min \mathcal{L}_{\text{Pred}}(f(X), Y) + \alpha \mathcal{L}_{\text{Exp}}\left(\frac{1}{N} \sum_{i=1}^N g(f, \langle X_i, Y_i \rangle), \hat{M}\right) \quad (5)$$

This type of method has been utilized and shown promising results in many application domains, such as text classification [90] and image classification tasks [44]. The advantage of the aggregation-based global supervision methods is that they can easily adapt existing techniques developed for local explanation with little to no extra effort. Besides, the acquisition of only one single feature attribution vector as a class-wise explanation signal is much more affordable, as compared with instance-wise supervision methods which require much more labor from human annotators. However, the drawbacks of this type of technique also come from the aggregation of local explanation, as the aggregated explanation is sensitive to the samples used to calculate, and thus could bring the sample bias into the global explanation of the model estimated.

Surrogate-based Global Supervision: This branch of work achieves explanation supervision by first estimating the global explanation of the target model via a surrogate model where the model-level explanation is easy to obtain, and then human knowledge can be leveraged to guide the global explanation and consequently supervise the model behavior. In this branch, the rule-based explanation is commonly used as it can be easily understood and edited by practitioners [32, 114, 148].

Rule-based explanation supervision can be achieved from many different angles. For instance, Vojř et al. [148] proposed the editable rule-based models that enable the users to edit rules and replace the underlying machine learning model; Popordanoska et al. [114] proposed the Explanatory Guided Learning (XGL) framework that creates simple rules capturing the prediction of the target model and allows the user to correct instances that are incorrect and the model is retrained. Besides, the rule-based explanation is also used as a mechanism for feedback that supports user adjustments without retraining the model [32]; Cornec et al. [31] developed the AI Model Explorer and Editor tool (AIMEE) that provides visualization of model decision boundaries using interpretable surrogates, and allows for the real-time modification of the decision boundaries. More recently, Lee et al. [81] proposed SELOR, a framework for upgrading a deep model with a Self-Explainable version with LOGic rule Reasoning capability, inspired by neuro-symbolic reasoning [34] that integrates deep learning with logic rule reasoning to inherit advantages from both. SELOR provides high human precision by explaining logic rules while also maintaining high prediction performance, and does not require predefined rule sets and can be learned in a differentiable way.

3.1.2 Global Explanation Regularization. Global explanation regularization is the method where some regularization terms that incorporate general prior knowledge about the global explanation are applied to the model. A good example of a preferred property of the model explanation is the sparseness, as it can provide a better understanding of the model behavior by humans, and in the meantime, serve as a regularizer of the explanation space to enhance model generalizability [49, 111, 129]. Concretely, the objective function for global explanation regularization can be summarized as follows:

$$\min \mathcal{L}_{\text{Pred}}(f(X), Y) + \beta \Omega(M^{(G)}) \quad (6)$$

where function $\Omega(\cdot)$ represents the specific regularization function for regulating the model's global explanation, and $M^{(G)}$ represents the model's global explanation vector calculated based on intrinsic parameters of the model f .

A commonly used prior knowledge to define $\Omega(\cdot)$ is to ensure the sparseness of the explanation, where a regularization term is proposed to penalize small magnitude weights of f that connect to the input features [49, 111, 129]. The existing studies suggest that this can result in a feature

selection effect, and greatly enhance the model's computational efficiency as well as generalizability. In addition, Burkart et al. [19] proposed a batch-wise regularization technique to enhance the interpretability of DNN models by means of a global surrogate rule list with a novel regularization approach that yields a differentiable penalty term. In Wu et al. [156], the authors proposed the regional tree regularization that encourages a DNN model to be well-approximated by several separate decision trees specific to predefined regions of the input space, yielding simpler explanations without compromising model accuracy.

3.2 Local Guidance

Local explanation guidance focuses on applying supervision signals or regularization terms to the model explanation of each local sample to guide the model learning. As shown in Figure 4, compared with the global explanation guidance, local guidance is more commonly used and explored in the current research communities thanks to the development of local explanation techniques, such as GradCAM [127] and attention mechanism [12, 147]. Based on the way explanation guidance is injected, local explanation guidance techniques can be categorized into three types: 1) Local Explanation Supervision: The ground truth explanation labels for each individual sample are provided as additional supervision signals to train the corresponding model explanation; 2) Local Explanation Regularization: in which some regularization terms that represent some general prior knowledge about the local model explanation are added to regularize all the local explanation of the model; and 3) Explanation Guided Data Augmentation: where the local model explanations are used to construct additional data samples for model training.

3.2.1 Local Explanation Supervision. Just as we supervise the model prediction via ground truth labels, local explanation supervision methods add additional supervision signals to align the model explanation with ground truth explanation labels (e.g. human annotation masks) during model training. The explanation loss and the conventional prediction loss are typically jointly optimized during model training. The general assumption behind this approach is that the model can benefit from the explanation labels by learning to focus on the right features and consequently lead to better generalizability to unseen instances. Depending on the data representation and application domains, we further narrow down the techniques into three subcategories: 1) visual explanation alignment, 2) rationale attention alignment, and 3) feature attribution alignment.

Visual Explanation Alignment: The visual explanation of image data is typically represented by a heat map overlaid on top of the original image, and the ground truth explanation labels \hat{M} are typically obtained by human annotation in the form of bounding boxes or fine-grained contours.

The first framework that can be applied to visual explanation alignment was proposed by Ross et al. [122], where the authors defined a very generic explanation-guided learning loss called "Right for the Right Reasons" loss (RRR) as follows:

$$\min \sum_{i=1}^N -Y_i \log(f(X_i)) + \alpha \sum_{n=1}^N (\hat{M}_i \frac{\partial}{\partial X_i} \log(f(X_i)))^2 + \beta \|\theta\|_2^2 \quad (7)$$

where \hat{M}_i denotes the ground truth explanation mask of a sample i ; the task supervision loss is implemented as the conventional cross-entropy loss, and the explanation supervision loss is designed to enforce the alignment of the ground truth explanation mask \hat{M} and the gradient maps via inner product operations.

Later on, the RRR loss is further extended by Schramowski et al. [125] and Dharma et al. [37] regarding the definition of the explanation losses. Specifically, instead of regularizing the gradients with respect to input X , Schramowski et al. [125] proposed to regularize the gradients of the final convolutional layer of the model that corresponds to GradCAM explanation and add a rescaling

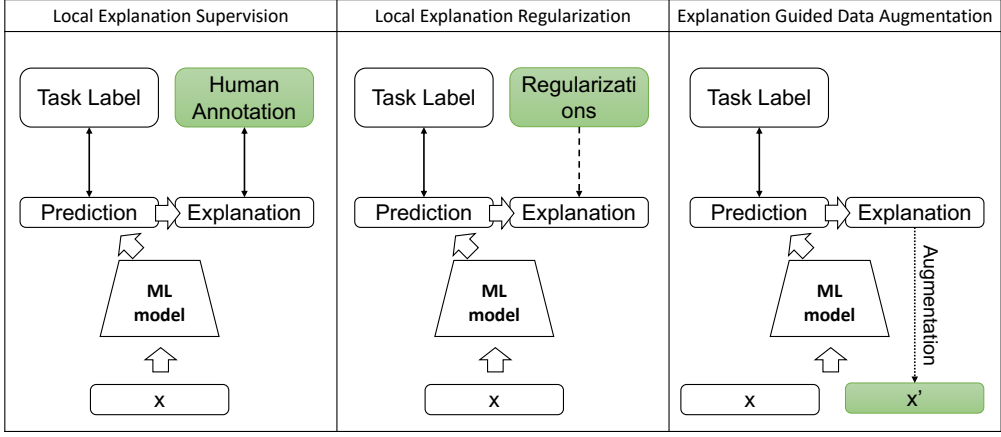


Fig. 4. Illustration of Local Guidance techniques. (left) Local Explanation Supervision: The ground truth explanation labels for each individual sample are provided as additional supervision signals to train the corresponding model explanation; (middle) Local Explanation Regularization: in which some regularization terms that represent some general prior knowledge about the local model explanation are added to regularize all the local explanation of the model; (right) Explanation Guided Data Augmentation: where the local model explanations are used to construct additional data samples for model training.

weight c_k to each class k for handling the unbalanced dataset issue. In Dharma et al. [37], the explanation loss is broken down into two terms to characterize the sensitivity of the gradient maps differently based on the relationship between each pixel of input and the ground truth mask, as shown below:

$$\mathcal{L}_{\text{Exp}} = \alpha_1 \sum_{j \in \hat{M}_i} \frac{\partial \mathcal{L}_{\text{Pred}}(f(X_i), Y_i)}{\partial X_{i,j}} + \alpha_2 \sum_{j \in [d] \setminus \hat{M}_i} \frac{\partial \mathcal{L}_{\text{Pred}}(f(X_i), Y_i)}{\partial X_{i,j}} \quad (8)$$

where $[d] \setminus \hat{M}_i$ represent the complement subset of the explanation \hat{M}_i of the whole feature set.

Later, many more models that are designed for visual explanation alignment are proposed [50, 52, 99, 139, 160]. Specifically, Stammer et al. [139] proposed a visual explanation alignment model based on symbolic (concept) alignment, where The symbolic (concept) explanation is modeled by a set transformer module. Mitsuahara et al. [99] proposed a visual explanation alignment objective specifically designed for the Attention Branch Network (ABN) [47], where the attention branch outputs are used as the model explanation. The limitation of this work is that it can only work under ABN architecture. Ying et al. [160] proposed the Visual Feature Importance Supervision (VISFIS) framework that optimizes four key model objectives: (1) accurate predictions given limited but sufficient information (Sufficiency); (2) max-entropy predictions given no important information (Uncertainty); (3) invariance of predictions to changes in unimportant features (Invariance); and (4) alignment between model explanations and human explanations (Plausibility) to improve model accuracy as well as performance. Nguyen et al. [105] proposed two novel architectures of self-interpretable image classifiers that first explain, and then predict by harnessing the visual correspondences between a query image and exemplars, and demonstrated the improvement on out-of-distribution (OOD) dataset scenarios. Gao et al. [52] proposed a more generic visual explanation alignment framework called GRADIA based on the GradCAM explanation. In addition, they proposed the Reasonability Matrix that can better determine what samples need to be adjusted to improve the model performance and explanation quality. More recently, Gao et al. [50] proposed

a robust visual explanation alignment framework that can better handle the nosiness of human annotation on image data. Specifically, the explanation loss is defined as follows:

$$\mathcal{L}_{\text{Exp}} = \min_{\theta, \lambda, \phi} \sum_i^N \max\{0, \|\tilde{M}_i - \hat{M}_i\| - \lambda\} + (g(f, \langle X_i, Y_i \rangle) - h_{\phi}(\hat{M}_i))^2 \quad (9)$$

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 labels; \tilde{M}_i is a binary projection of the explanation $g(f, \langle X_i, Y_i \rangle)$ by a threshold λ , as:

$$\tilde{M}_i = \begin{cases} 1 & g(f, \langle X_i, Y_i \rangle) \geq \lambda \\ -1 & g(f, \langle X_i, Y_i \rangle) < \lambda \end{cases} \quad (10)$$

Besides the application to general images, visual explanation alignment techniques have also been applied to medical image domains [132, 133, 176]. Please refer to Applications Section for more details.

Rationale Attention Alignment: The explanation of natural language data is typically represented by the rationales (e.g. word tokens) that highlight the most significant part of the data for making specific task predictions. The ground truth explanation labels are typically obtained by human annotation in the form of rationales or natural language format (sentences). In this domain, the datasets collected by the ERASER benchmark [36] are commonly used as the datasets come with ground truth rationales obtained from human annotators.

Many existing works have proposed to supervise the rationale attention of the model to improve the model performance and quality of attention [22, 26, 55, 70, 138, 168, 172]. The explanation loss is commonly realized by conventional losses, such as cross-entropy loss [26, 70], Mean Squared Error (MSE) [137, 138], and KL-divergence loss [172]. Besides, Atanasova et al. [10] proposed several novel ways to enforce the alignment, such as Data Consistency, Confidence Indication, and Faithfulness.

In addition to leveraging the model attention value itself, many existing works have also proposed to directly generate rationales [15, 66, 82, 137, 162, 172] or natural language [21, 91] as the ‘explanation’ of the model to be aligned with ground truth labels via additional decoders, such as Conditional Random Field (CRF) [76, 162], Gated Recurrent Unit (GRU) [28, 172], Transformer-based models [35, 91, 137, 147].

Feature Attribution Alignment: Besides the specific domain of applications, local explanation supervision can be generally applied to any dataset where the input feature importance can be computed. Such feature importance is typically referred to as ‘Feature Attribution’, and can be also treated as the model explanation to be aligned with human explanation ground Truth. For instance, in Balayan et al. [13] the feature attribution is computed by a specific designed semantic layer (an intermediate output that is the importance of each feature) and is aligned with human-labeled feature masks by Cross-Entropy loss; in Singh et al. [134], feature attribution is calculated by Contextual Decomposition, and is aligned with the ground truth human labels by ℓ_1 distance [119].

Overall, the idea of local explanation supervision has been explored extensively in many application domains in recent few years, primarily due to the fact that 1) it is straightforward for human annotators to provide an instance-wise explanation with necessary domain knowledge, and 2) the development and popularity of local explanation techniques, such as GradCAM [127] and attention mechanism [12, 147] to explain complex DNNs in high dimensional problem space (such as image and text data). So far the results seem to be promising, as most existing works suggest that applying the local explanation supervision during training can greatly enhance both the quality of the explanation as well as the performance of the backbone DNNs model. However, as also pointed out by several existing works, the scalability remains the biggest challenge for this kind of approach, as the additional instance-wise human explanation labels may not be easily accessible

and require non-trivial effort from human annotator [50, 51]. Designing effective semi-supervised or weakly-supervised explanation supervision frameworks, or even adapting the idea from active learning can be promising future directions to further overcome this limitation. Nevertheless, existing works also demonstrated the effectiveness of local explanation supervision under very limited training sample sizes [50, 52], which could suggest the potential benefit of applying current techniques to the domains where data samples are limited and hard to acquire, yet both model performance and the explainability are on-demand, such as in medical domains.

3.2.2 Local Explanation Regularization. Local explanation regularization methods add additional regularization terms to regularize each local explanation to ensure the generated model explanations follow some general properties (such as smoothness, stability, and sparsity) or follow the knowledge from other existing well-trained models. The additional explanation of regularization loss is typically jointly optimized with the prediction loss during model training. Depending on the type of regularization terms, we break down the existing techniques into two subcategories: 1) Property-based Regularization and 2) Explanation Distillation Regularization.

Property-based Regularization: The additional regularization terms are injected into the model explanation to enforce some general properties (such as smoothness, stability, and sparsity). Specifically, in Lei et al. [82] the authors proposed the *continuity* and *sparsity* regularization terms; In Erion et al. [44] the authors proposed the *smoothness* Regularization (i.e. Laplace 0-mean prior) on the model Explanation computed by expected gradient; In Halliwell et al. [59] the authors proposed the Prediction-guided sparsity regularization (in Equations (5) and (6)) to penalize the model to have small values in saliency maps (computed by GradCAM and guided BP) if the prediction is incorrect; Alvarez et al. [5] proposed a gradient regularization approach for enforcing explanation *robustness/stability*; Plumb et al. [113] apply the *fidelity* and *stability* regularization on the explanation. Specifically, the explainer $g(\cdot)$ is realized by Local Interpretable Model-Agnostic Explanations (LIME) [118] with a linear function $l(\cdot)$, and the authors applied two regularization terms on the model explanation: 1) neighborhood-fidelity and 2) stability based on the neighborhood of input, as shown below:

$$\Omega = \underbrace{\mathbb{E}_{X' \sim \mathcal{N}_{X_i}} [(l(X') - f(X'))^2]}_{\text{neighborhood-fidelity}} + \underbrace{\mathbb{E}_{X' \sim \mathcal{N}_{X_i}} [\|g(f, \langle X_i, Y_i \rangle) - g(f, \langle X', Y_i \rangle)\|_2^2]}_{\text{stability}} \quad (11)$$

where \mathcal{N}_{X_i} is a neighborhood of sample X_i in the space of probability distributions over the whole input data distribution X , and X' is sampled from the neighborhood \mathcal{N}_{X_i} . Intuitively, the fidelity regularization enhanced the explanation to accurately convey which patterns the model used to make this prediction, while the stability regularization will lead to more stable explanations, which will improve the model's trustworthiness [5, 6].

Explanation Distillation Regularization: Besides enforcing predefined properties of the explanation, this line of work tries to distill explanation knowledge from other well-trained models to guide the explanation of the target model. In Zeng et al. [165], the authors proposed to align the explanation of a target model with another pre-trained adversarially counterpart model generated explanation using ℓ_2 distance loss. Singh et al. [135] proposed to align the target model's Class Activation Maps (CAM) [173] explanation with a pre-trained model's explanation by minimizing the overlap between each classes explanation. More specifically, the explanation loss consists of two terms, 1) regularization loss which measures the distance between the target models' and the corresponding pre-trained model's explanation of class i , and 2) overlapping loss which calculates the similarity between the target model's explanation of different classes. As a result, the model can be trained under the constraints that 1) the target model explanation should be as close as possible to a pre-trained model, and 2) the model explanation of different classes should be as different as

possible, leading to a better explanation quality and higher accuracy. More recently, Fernandes et al. [46] proposed Scaffold-Maximizing Training (SMaT) framework for directly optimizing explanations of the model's predictions to improve the training of a student simulating the said model. The authors found that, across tasks and domains, explanations learned with SMaT both lead to students that simulate the original model more accurately and are more aligned with how people explain similar decisions.

While using the idea from model distillation to extract the knowledge to guide the model explanation is an interesting direction, the potential positive effect is largely dependent on the choice and quality of the pre-trained model and is prone to negative transfer, such as contextual bias in pre-trained model explanation, that can hurt the target model performance. Thus additional validation and guidelines are on demand for this type of technique to be applied to handle real-world problems.

3.2.3 Explanation Guided Data Augmentation. Explanation Guided Data Augmentation is an emerging subdomain in the data augmentation domain, where the ground truth explanation (i.e. rationale) of the prediction task is taken into account when building up additional augmented samples for model training. The general formulation for generating explanation-guided data augmentation samples can be summarized as follows:

$$X'_i = \text{aug}(X_i, g(f, \langle X_i, Y_i \rangle)) \quad (12)$$

Where $\text{aug}(\cdot)$ denotes the specific augmentation function based on the original input sample X_i and the model's explanation for the given input-output pair $\langle X_i, Y_i \rangle$.

The underlying assumption is that training the model with the augmented samples X' will encourage the model to better learn to pay attention to the right rationales for the prediction tasks and thus naturally enhances both the explainability as well as the generalizability of the model. Based on the way explanation is used for the data augmentation, existing techniques can be categorized into two directions: 1) Rationale Inclusion/Amplification and 2) Rationale Exclusion/Masking.

Rationale Inclusion/Amplification: This line of works typically emphasizes the right rationales and de-emphasize other irrelevant features. The inclusion/amplification-based augmentation function can be generally defined as follows:

$$\text{aug}_{in}(X_i, g(f, \langle X_i, Y_i \rangle)) = X_i \times (\gamma + \lambda g(f, \langle X_i, Y_i \rangle)) \quad (13)$$

where γ is used to set a default offset value to preserve all the feature values regardless of the importance; λ is the scale factor that controls the degree of amplification of the important features.

Specifically, Sharma et al. [131] proposed to amplify the feature values of the right rationales relatively higher by a certain degree. In their experiments, γ is set to 0.01, and λ is set to 1 to emphasize the rationale features in the augmented samples. The results demonstrated the general effectiveness of the proposed method on several conventional ML models, such as Naive Bayes, logistic regression, and SVM. In Saha et al. [123] only the important part of the image for network prediction is selected using saliency-based explanations and stored in the episodic memory with the corner coordinate for continual learning. Ismail et al. [65] proposed to minimize the KL divergence between $f(X)$ and $f(X')$, where X' is augmented by masking the features with low gradient values. These types of methods can be seen as a special case of Equation (13) where γ is set to 0, and λ is set to 1 to only include the rationale features in the augmented samples.

Besides the simple augmentation of the feature values, several other works have also proposed some novel and unique ways to augment data to best leverage the extra information from the model explanation. In Pillai et al. [112], the saliency map explanation of the original sample and other samples as a composed image is aligned with the model explanation of the original input sample. In Teso et al. [144], the irrelevant features are perturbed while the rationales and task labels are

preserved as new samples to guide the model attending the ground truth rationales. In Schneider et al. [124], the model explanation (generated by GradCAM) is treated as additional input for model prediction and requires a specific change of the model architecture. In summary, the general idea stays the same, which is to build additional samples and inform the model to better learn which features are the right rationales to make the right prediction of the downstream tasks.

Rationale Exclusion/Masking: As opposite to inclusion/amplification, this line of works typically teaches the model not to attend irrelevant rationales by excluding/masking out the right rationales, as summarized by the following equation:

$$\text{aug}_{ex}(X_i, g(f, \langle X_i, Y_i \rangle)) = X_i \times (\gamma - g(f, \langle X_i, Y_i \rangle)) \quad (14)$$

where γ is typically set to be the maximum possible value of the importance, e.g. 1, to exclude the value of the important features from X , and thus serve as a masking function for the data augmentation.

Specifically, in Zaidan et al. [163], the authors propose to construct some additional samples by masking out those important features of some existing samples to simulate the loss of confidence (uncertainty should raise) in predicting the right answer. A similar idea can be also found in Li et al. [84], where the proposed self-guidance is basically using the model explanation as a mask to augment the original image and thus construct an unsupervised loss based on the augmented/masked image.

Overall, the unique advantages of explanation-guided data augmentation techniques can be summarized as follows: 1) it takes the model behavior (i.e. rationale for the prediction) into consideration; 2) it can be model agnostic with respect to the specific explainability techniques used for calculating the model explanation; 3) it can be used in combination with other conventional data augmentation techniques, and in parallel with other EGL techniques for model training. However, the effectiveness of the existing works is mainly supported by intuitions and empirical observations. Thus further development of quantitative evaluation metrics as well as theoretical analysis and justification of the techniques can be essential to further advance this field of research.

4 APPLICATIONS

4.1 Computer Vision (CV)

Applying EGL to solve image classification problems has become a hot and attractive research area in recent years [50, 52, 99], largely thanks to the popularity and advancement of visual explanation techniques [127, 166]. Depending on the nature of the image source, existing works can be further categorized into (1) general image prediction and (2) medical image analysis.

4.1.1 General Image Prediction. The application of EGL on general images typically involves image classification tasks on natural image data such as ImageNet [74], Caltech-UCSD Birds (CUB) [149], Microsoft COCO [88], and Places365 [174], and some synthetic image data such as ToyColor [122], MNIST [78], and many MNIST variants including Fashion-MNIST [158], Decoy-MNIST [122], and Color-MNIST [85]. The typical EGL technique used in this application domain is local explanation supervision and regularization, where the sample level visual explanation of the model is jointly optimized together with the conventional prediction loss [37, 44, 50, 52, 84, 99, 119, 121, 139, 152]. When applying the explanation supervision techniques, the ground truth explanation labels are typically collected from human annotators, and the additional attention loss is typically realized by a distance loss between the ground truth and the model visual explanation at the sample level. For model explanation assessment, case studies are most commonly used for qualitative analysis [50, 52, 99], while IoU score is for quantitative evaluation [50, 52, 84].

4.1.2 Medical Image Analysis. Besides generic image applications, EGL has also been widely studied in the medical domain, thanks to the availability of domain-expert annotation on many medical image datasets [29, 77, 151]. In general, we observed a variety of datasets studied by existing works, including but not limited to ISIC Skin Cancer dataset [29], Iris-Cancer dataset [87], scaphoid fracture detection dataset [77], Fundus image dataset (IDRiD) [115], and the pneumonia detection X-ray dataset [151] for disease identification task [176]. Similar to most EGL frameworks on generic image data, an additional explanation loss is added to the model objective and is typically realized by a distance loss between the ground truth annotation collected from domain experts and the model visual explanation. However, compared with generic image data, several unique challenges have been identified by existing works when applying EGL to medical images, such as 1) difficulty in assessing the quality of the model explanation, and 2) the scalability of the sample size of the annotation labels of the datasets.

4.2 Natural Language Processing (NLP)

Interest has recently grown in applying EGL to designing NLP systems. Based on how the explanation is acquired, we have two categories of the application: (1) using the attention mechanism as the explanation and (2) using a generative model to generate the explanation.

4.2.1 Attention mechanism as the explanation. NLP systems generally use variants of attention mechanisms to get explanations. To evaluate the explanation, the ground truth explanation labels are typically collected from human annotators (Stacey et al. [138] use TextRank to get ground truth labels), and the evaluation metric can be the F1 score and IoU score based on token or snippet level. In addition to the agreement with human rationales, a faithful explanation is related to the downstream task performance, so rationale-level supervision is widely applied [26, 53, 138, 140, 168, 172]. Comprehensiveness and sufficiency are two main metrics regarding the influence of the explanation on the downstream task, Faithfulness, Data Consistency, and Confidence Indication are other diagnostic properties [10]. Attention mechanisms can learn to assign soft weights to token representations so that one can extract highly weighted tokens as rationales [36]. While this is intuitive for most of the NLP systems, the weights can be useless to give a faithful explanation because of the complex interaction of tokens. Another strand of works [22, 55] hard-select tokens or snippets from the input and only uses the selected part for the downstream task to get untangled explanation. This strand can be further divided into pipeline approaches and reinforcement learning approaches according to how the models are trained. Aligning the explanation with human annotators is not necessarily the optimal objective for improving model accuracy, the various loss strategies are proposed [22]. By e.g. masking out important explanation features of existing samples [163], one can augment the data. Liu et al. [90] propose global supervision by adding feature attribution prior to the total loss.

4.2.2 Generative model to generate the explanation. In addition to giving explanations directly by the attention mechanism of NLP systems, a lot of works apply additional generative models to generate natural language explanations [15, 66, 82]. Although the rationales acquired from attention mechanisms provide concise and quick explanations, they may not have the means to provide important details of the reasoning of a model. By using an additional natural language decoder, one can generate a comprehensive description of the decision-making process behind a prediction, some examples of the generative module include a conditional random field (CRF) [162], a natural language decoder [21, 91], a GRU following an MLP [170], BiLSTM and Transformer [137]. The commonly used datasets and corresponding tasks are ComVE [150] for commonsense validation, e-SNLI [21] for natural language inference, COSe [117] for commonsense question answering, e-SNLI-VE [71] for visual entailment, VCR [164] for visual commonsense reasoning. The ground

truth is usually also natural language explanations provided by humans. To evaluate the quality of natural language explanations, one can either use automatic metrics like METEOR [14], BERTScore [167], and BLEURT [126], or use human evaluation with metrics like e-ViL score [71], confidence, and readability. In terms of the faithfulness of the natural language explanations, Wiegrefe et al. [154] provides two necessary conditions: feature importance agreement and robustness equivalence.

4.3 Visual Question Answering (VQA)

Attention and reasoning are two intertwined mechanisms underlying visual question-answering (VQA) tasks. Thanks to the widely used attention mechanism in VQA, applying EGL to help improve both the interpretability and performance of VQA tasks have become a hot and attractive research area in recent years. The typical EGL technique used in VQA tasks is local explanation supervision and regularization, where the sample-level visual explanation of the model is jointly optimized together with the conventional prediction loss. When applying the explanation supervision techniques, the ground truth explanation labels can be collected from human annotators [25, 48, 155], or generated by another model [116, 169]. There have been a lot of VQA datasets with annotations, some are annotated with human-generated questions and answers like MovieQA [143] and VQA v1.0 dataset in [7], while others are developed with synthetic scenes and rule-based templates like GQA [64], Clevr [69], and VCR [164]. VQA-2.0 [56] includes complementary images that lead to different answers, reducing language bias by forcing the model to use visual information. The AiR-D [25] is the first dataset of eye-tracking data collected from humans performing the VQA tasks. The VQA-HAT dataset [33] is a visual explanation dataset that collects human attention maps by giving human experts blurred images and asking them to determine where to deblur in order to answer a given visual question. VQA-CP [4] contains QA pairs whose distribution is significantly different between the training and test set. VQA-X [108] offers human textual explanations which can be used to determine important objects and then are grounded to important regions in the image as the explanation. The additional attention loss can be attention accuracy [25], false sensitivity rate [155] rank correlation loss [116, 169] and IoU loss [48].

4.4 Healthcare

EGL techniques have also been well-explored in general healthcare applications, such as on gene interaction graph [57], Adult Changes in Thought (ACT) [98], Mount Sinai Brain Bank (MSBB), Religious Orders Study/Memory and Aging Project (ROSMAP) [1], and healthcare mortality prediction [97]. Specifically, Erion et al. [43] studied the tissue-specific gene interaction graph for the tissue most closely related to acute myeloid leukemia (AML, a type of blood cancer) in the HumanBase database [57] on how penalizing differences between the attributions of neighbors in an arbitrary graph connecting the features can be used to incorporate prior biological knowledge about the relationships between genes, yield more biologically plausible explanations of drug response predictions, and improve test error. They tested the model performance on a healthcare mortality prediction dataset [97], where the model inputs are 35 features representing patients' demographic information and medical data. Erion et al. [44] then further extended their previous study and proposed to add a graph attribution prior regularization on explanation to a two-layer neural network. Their experimental results show the proposed method can significantly outperform all other methods. In addition, Weinberger et al. [152] extracted prior information from multiple gene expression datasets of the Accelerating Medicines Partnership Alzheimer's Disease Project (AMP-AD) portal, incorporated meta-features in a gene-gene interaction graph and proposed a deep attribution prior framework to Alzheimer's disease biomarker prediction.

4.5 Chemistry

EGL has also started to see emerging applications in the chemistry domain, especially for molecular puppetry prediction tasks [93, 94, 141]. For instance, one recent work proposed an EGL framework for Graph Neural Networks (GNNs) by supervising their node- and edge-level explanation to align with domain expert annotation labels [51]. In this work, the authors studied three binary classification molecular datasets², namely 1) The Blood-brain barrier penetration (BBBP) dataset comes from a recent study [93] on the modeling and prediction of barrier permeability, 2) the BACE dataset provides quantitative (IC50) and qualitative (binary label) binding results for a set of inhibitors of human b-secretase 1 (BACE-1) [141], and the “Toxicology in the 21st Century” (TOX21) initiative created a public database measuring the toxicity of compounds [157]. The general goal for each dataset is identifying functional groups on organic molecules for biological molecular properties. Each dataset contains binary classifications of small organic molecules as determined by the experiment. The experimental results suggest that the proposed GNES framework can effectively improve the reasonability of the explanation while still keeping or even improving the backbone GNNs model performance.

4.6 Crime

EGL has also been studied in the application of risk and crime-related applications, where it is important to check if the model is leveraging reasonable features when predicting crime incidences or assessing the future risk of crime suspects. For instance, several works have studied the Propublica’s COMPAS Recidivism Risk Score datasets³, which contains data for predicting recidivism (i.e. whether a person commits a crime / a violent crime within 2 years) from many attributes [5, 119]. COMPAS dataset is designed for checking whether there exist biases in the model explanation, such as the model’s treatment of the person’s race attribute when making the prediction. Specifically, Rieger et al. [119] proposed contextual decomposition explanation penalization (CDEP), a method that enables practitioners to leverage explanations to improve the performance of a deep learning model. In particular, CDEP enables inserting domain knowledge into a model to ignore spurious correlations, and correct errors, and demonstrates the ability to increase performance on real datasets; Alvarez et al. [5] proposed an EGL framework by explicitly enforcing three basic desiderata for interpretability—explicitness, faithfulness, and stability—during training to enhance the robustness and interpretability of model explanations. Besides, Balayan et al. [13] studied a private online retailer fraud detection dataset with the proposed JOEL framework, a neural network-based framework to jointly learn a decision-making task and associated explanations that convey domain knowledge. Specifically, JOEL is tailored to human-in-the-loop domain experts that lack deep technical ML knowledge, providing high-level insights about the model’s predictions that very much resemble the experts’ own reasoning. Moreover, they collect the domain feedback from a pool of certified experts and use it to ameliorate the model (human teaching), hence promoting seamless and better-suited explanations.

4.7 Potential Future Domains of Applications

Despite the recent attention and major advance of EGL in the aforementioned popular application domains, there are still a number of open problems and potentially fruitful directions for future research and application of EGL, as follows:

4.7.1 FaccT. Fairness, Accountability, and Transparency (FaccT) are becoming as important as—or depending on application areas—more important than model accuracy as an evaluation metric.

²Available online at: <http://moleculenet.ai/datasets-1>

³Available online at: github.com/propublica/compas-analysis/

Since it is nearly not feasible to prepare an impeccable dataset that can equally represent every possible feature related to a model's task, blindly pursuing a model's accuracy cannot exclude the chance of causing "catastrophic consequences" in critical circumstances [63]. One of EGL's crucial application areas is to realize the balance between the model accuracy and FaccT by allowing human users to elicit their perspectives on steering the model. In shaping the balance, one crucial research direction is to understand how to maximize the case where reasonable human reasoning can also cause accurate prediction. There are several arguments discussing when human reasoning can cause a beneficial or detrimental effect on model prediction. While the debate is ongoing, we are gradually seeing more evidence where human involvement can result in a positive effect [30, 54]. For example, Shao et al. find humans "arguing against" unreasonable explanation can benefit the model [130]. At the end of the day, from the perspective of model accuracy and FaccT, a railroad should not be the reason for predicting a train [80], a snowboard cannot be a male class [61], and a shopping cart should not only belong to a woman class [171].

4.7.2 Adversarial Learning. Adversarial perturbations can significantly drop the model's accuracy. In a dramatic situation, it can reach nearly to 0%. Current ML models are vulnerable to adversarial attacks. Since the majority of adversarial attack shift model's attention, applying EGL in detecting unusual shifts could be one of the solutions for developing a more robust ML model against adversarial attacks. However, in pursuing such a direction, the change of the attention map after the attack can be subtle from human eyes [18]. In order to apply EGL in the area of adversarial learning, we see devising better solutions in the following areas to be crucial. First, providing additional signals other than model attention can help human users effortlessly detect the attacked cases. Second, devising an advanced EGL mechanism that can (1) guide the users to generate effective input (2) and applying such input to improve the model's robustness would be essential. Following this line of thought, very recently, Jeong et al. [67] proposed Generative Noise Injector for Model Explanations (GNIME), a novel defense framework that perturbs model explanations to minimize the risk of model inversion attacks while preserving the interpretabilities of the generated explanations. Thus, we believe future studies on model explanation defense and attack can be one of the key research sub-areas of EGL domain.

4.7.3 Continual & Active Learning. EGL's core principle is motivating ML engineers' iterative training, such as continual learning [41, 123] and active learning [24, 70]; helping them to figure out the vulnerability through explanation and fixing the issue by providing a human-level guideline. In supporting such an iterative training, we believe one of the promising areas is "data iteration", a design that can help ML engineers to fortify the dataset by adding more examples based on detected vulnerabilities through explanation. In such a direction, we believe understanding the pros and cons of retraining and continual learning can be crucial. For example, there can be a case where newly found data points can be stacked up on an existing dataset and be used in retraining. Another case can be to iteratively update the last model through some of the existing techniques in continual learning [107]. In general, in the world of EGL, understanding when to apply retraining or continual learning and what are the pros and cons of each training strategy are not well understood. Understanding which strategy can yield what strengths and weaknesses in the scenario of data iteration would be one of the core future applications of EGL.

4.7.4 Contrastive Learning. Contrastive learning is a powerful self-supervised learning strategy that encourages augmentations of the same input to have more similar representations compared to augmentations of different inputs. In the field of EGL, we have started to see several works that apply the contrastive objective to the model explanation between similar/dissimilar samples to build up the explanation objective [38, 110, 135, 163]. The most significant advantage of leveraging

Table 2. A list of publicly available datasets for EGL with human annotation labels.

Dataset	Type	Link	Annotation Type
Gender Classification	Vision	https://github.com/YuyangGao/RES	Pixel level
Scene Recognition	Vision	https://github.com/YuyangGao/RES	Pixel level
Face Glasses Recognition	Vision	https://github.com/carriegu0818/EGL_benchmark	Pixel level
Prohibited Item Detection	Vision	https://github.com/carriegu0818/EGL_benchmark	Pixel level
ACT-X	Vision	https://github.com/Seth-Park/MultimodalExplanations	Pixel level and Textual
Caltech-UCSD Birds	Vision	https://authors.library.caltech.edu/27452/	Pixel level(bounding box)
The PASCAL VOC Challenge 2007	Vision	http://host.robots.ox.ac.uk/pascal/VOC/voc2007/	Pixel level
The PASCAL VOC Challenge 2012	Vision	http://host.robots.ox.ac.uk/pascal/VOC/voc2012/	Pixel level
ISIC2018 Challenge	Vision	https://challenge.isic-archive.com/landing/2018/	Pixel level(bounding box)
Pneumonia Detection	Vision	https://www.kaggle.com/c/rsna-pneumonia-detection-challenge	Pixel level(bounding box)
Movie Review	NLP	https://github.com/jayded/eraserbenchmark	Span-level rationale
MultiRC	NLP	https://github.com/jayded/eraserbenchmark	Single sentence-level rationale
FEVER	NLP	https://github.com/jayded/eraserbenchmark	Sentence-level rationale
BoolQ	NLP	https://github.com/jayded/eraserbenchmark	Token-level rationale
Evidence inference	NLP	https://github.com/jayded/eraserbenchmark	Sentence-level rationale
e-SNLI	NLP	https://github.com/jayded/eraserbenchmark	Token-level rationale
Commonsense Explanations (CoS-E)	NLP	https://github.com/jayded/eraserbenchmark	Sentence-level rationale
VQA-HAT	VQA	https://computing.ece.vt.edu/~abhshkdz/vqa-hat/	Pixel level
GQA	VQA	https://cs.stanford.edu/people/dorad/gqa/about.html	Pixel level
VQA-X	VQA	https://github.com/Seth-Park/MultimodalExplanations	Pixel level and Textual
VQS	VQA	https://github.com/Cold-Winter/vqs	Pixel level
BBBP	Graph	https://github.com/YuyangGao/GNES	Node- and edge-level
BACE	Graph	https://github.com/YuyangGao/GNES	Node- and edge-level
TOX21	Graph	https://github.com/YuyangGao/GNES	Node- and edge-level

the contrastive learning paradigm for explanation guidance is that no ground truth explanation annotation labels are required for model training. However, designing an appropriate contrastive framework for EGL can be more challenging due to the lack of a standard form of model explanation under different application domains. Besides, how to define and formulate the positive and negative explanation samples to contrast with the anchor sample’s explanation can be challenging without knowing the ground-truth labels. Thus, we believe the further development of the contrastive EGL framework can be one of the core future directions in EGL, and it can lead to a significant leap in the application of EGL to the domains where ground truth explanation labels are generally difficult to obtain in large scale.

5 EXPERIMENTS

This section aims at providing an extensive and comprehensive experimental study among existing EGL models in various popular application domains. Specifically, the comparative studies of four datasets from the Computer Vision (CV) domain, namely 1) Gender Classification, 2) Scene Recognition, 3) Face Glasses Recognition, and 4) Prohibited Item Detection, and three datasets from the Natural Language Processing (NLP), namely 1) Movie Review, 2) MultiRC, and 3) FEVER are provided. The details about each dataset are included in Table 2, where a full list of publicly available datasets for EGL is provided.

5.1 Visual Explanation Guided Learning

5.1.1 Gender Classification [50]. The gender classification task is derived from the Microsoft COCO dataset ⁴ [88] by extracting images that had the word “men” or “women” in their captions. The dataset is further filtered by removing images with 1) both gender in the caption, 2) multiple people present, or 3) not recognizable humans. A subset of the images is further manually annotated by human annotators as factual and counterfactual masks. The dataset in total consists of 1,736 images with human annotations, where the distribution of female to male is even. For data splitting, we only randomly sampled 100 samples as the training set to better simulate a more practical situation

⁴Available at: <https://cocodataset.org/>

where we only have limited access to the human explanation labels. The validation and test set is set to 700.

5.1.2 Scene Recognition Dataset [50]. The scene recognition dataset is originally derived from the Places365 dataset⁵ [174] and manually annotated by Gao et al. [50]. The task for this dataset is a binary classification of scene recognition: nature vs. urban. Specifically, the categories used to sample the data are listed below:

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

The dataset consists of a total of 2086 images with human explanation labels. Similarly, we split the data randomly with a sample size of 100/700/700 for training, validation, and testing.

5.1.3 Face Glasses Recognition. We construct the glasses recognition dataset from the CelebAMask-HQ dataset⁶ [79] by categorizing face images with and without glasses. In CelebAMask-HQ, masks were manually annotated with 19 classes including all facial components and accessories. The rationale of the task is that we are able to obtain factual annotation labels by the segmentation of eyes and glasses directly. While the original dataset is highly imbalanced in the ratio between faces with and without glasses, we randomly select an equal number of images in both classes, with a total of 100/393/392 images for training/validation/testing respectively.

5.1.4 Prohibited Item Detection. The task is constructed from the Sixray dataset⁷ [96] by splitting images based on the presence of prohibited items. Sixray is highly imbalanced with 1,059,231 X-ray images, including 6 classes of 8,929 prohibited items. Merging the 6 prohibited classes, the task of the new dataset is a binary prohibited item detection. Bounding boxes of prohibited items are included in all images. Due to data imbalance, the dataset is further filtered into 100/5296/5298 images for training, validation, and testing respectively.

5.1.5 Evaluation Metrics. We evaluate the model in terms of prediction performance as well as in terms of explanation performance. For prediction performance, we use AUC and accuracy as evaluation metrics. To evaluate explanation faithfulness, we employ the Matrix for comprehensiveness and sufficiency by ERASER [36]. For explanation correctness assessment, we compare the saliency map generated by Grad-CAM with ground-truth annotation masks. Specifically, we use the Intersection over Union (IoU) score [16], the bit-wise intersection and union operations between the ground truth explanation and the binarized model explanation. We further evaluate explanation performance with Explanatory F1, precision, and recall by bit-wise comparison between ground-truth explanation and model explanation.

5.1.6 Comparison methods. : We compare the performance of several models as listed below:

- **Baseline** Baseline 1 and 2 are pre-trained ResNet50 and VGG16 model that trains only on prediction loss without explanation loss.
- **GRADIA** [52]: 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.

⁵Available at: <http://places2.csail.mit.edu/index.html>

⁶Available at: http://mmlab.ie.cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html

⁷Available online at: <https://github.com/MeioJane/SIXray>

Table 3. The classification performance and explanation evaluation on Gender Classification. The results are obtained from 3 individual runs and the best results of each metric are highlighted in boldface font.

Model	Architecture	Prediction		Exp Faithfulness		Exp Correctness			
		Acc. \uparrow	AUC \uparrow	Comp. \uparrow	Suff. \downarrow	IoU \uparrow	F1 \uparrow	Precision \uparrow	Recall \uparrow
Baseline 1	ResNet50	0.680	0.664	0.048	0.105	0.147	0.470	0.554	0.525
Baseline 2	VGG16	0.637	0.671	0.048	0.051	0.058	0.155	0.340	0.132
Supervised									
GRADIA[52]	ResNet50	0.695	0.764	0.107	0.090	0.243	0.625	0.787	0.608
RES [50]	ResNet50	0.690	0.744	0.108	0.097	0.240	0.614	0.742	0.621
RRR[125]	VGG16	0.624	0.628	0.027	0.015	0.091	0.270	0.424	0.257
CDEP [119]	VGG16	0.628	0.635	0.014	0.010	0.100	0.254	0.419	0.231
Unsupervised									
SENN [5]	SENN(CNN)	0.589	0.627	0.005	0.027	0.062	0.186	0.300	0.186
SGT [65]	ResNet50	0.645	0.687	0.012	0.002	0.044	0.352	0.502	0.257

- **RES [50]**: A framework that trains the DNN with both factual and counterfactual annotations with two imputation functions: $g(\cdot)$ as a fixed value Gaussian convolution filter and learnable imputation function $g_\phi(\cdot)$ via multiple layers of learnable kernels.
- **CDEP [119]**: A framework that incorporates Contextual Decomposition (CD) to penalize spurious correlations and therefore correct errors.
- **RRR [125]**: A framework that was initially introduced by [122] and altered by [125], which aims to regularize the model to be right for the right reasons.
- **SGT [65]**: A framework that introduces saliency-guided training for neural networks to reduce noisy gradients in predictions.
- **SENN [5]**: A framework that applied two regularization terms on the model explanation: 1) neighborhood-fidelity and 2) stability based on the neighborhood of input.

5.1.7 Implementation Details. : All models are trained for 50 epochs with the same train/val/test split as mentioned above. We use the ADAM optimizer with a learning rate of 0.0001 [73]. The architecture of each model is listed in Table 3. To better compare the performance on explainability, the model explanations are generated by Grad-CAM [127]. When calculating the explanation evaluation metrics, the explanation maps were further binarized by a fixed threshold of 0.5. We use a batch size of 32 for training and 100 for testing. For GRADIA and RES, we set the slack variable α to 0.1 and 0.01, respectively, and the regularization factor to 0. For RRR, we set the regularization parameter to 1. For CDEP, we set the regularizer rate to 0, 0.1, and 10. For SENN, we set the robust regularization, sparsity regularization, and concept regularization hyperparameters to 0.0001, 0.00002, and 1, respectively. For SGT, we set features dropped to 0.1 and 0.3.

5.1.8 Quantitative analysis. Model prediction performance and explanation quality in the domain of computer vision are presented in Tables 3, 4, 5, and 6. We evaluate 6 models through gender classification, scene recognition, glasses identification, and prohibited item discovery. We evaluate two baseline model performances: ResNet50 and VGG16, as they are employed by the selected paper. Overall, supervised models demonstrate better prediction power and higher explanation quality than unsupervised models. ResNet50 seems to perform slightly better in prediction, and significantly better in explanation quality compared with VGG16.

Table 4. The classification performance and explanation evaluation on Scene Recognition. The results are obtained from 3 individual runs and the best results of each metric are highlighted in boldface font.

Model	Architecture	Prediction		Exp Faithfulness		Exp Correctness			
		Acc. ↑	AUC ↑	Comp. ↑	Suff. ↓	IoU ↑	F1 ↑	Precision ↑	Recall ↑
Baseline 1	ResNet50	0.947	0.965	0.068	0.255	0.397	0.702	0.906	0.628
Baseline 2	VGG16	0.953	0.988	0.134	0.117	0.191	0.324	0.890	0.226
Supervised									
GRADIA [52]	ResNet50	0.952	0.987	0.255	0.073	0.378	0.606	0.912	0.501
RES [50]	ResNet50	0.956	0.988	0.189	0.002	0.435	0.722	0.909	0.647
RRR [125]	VGG16	0.953	0.987	0.014	0.019	0.224	0.364	0.925	0.250
CDEP [65]	VGG16	0.934	0.952	0.026	0.039	0.127	0.232	0.807	0.153
Unsupervised									
SENN [119]	SENN(CNN)	0.733	0.798	0.022	0.042	0.082	0.183	0.721	0.108
SGT [5]	ResNet50	0.937	0.985	0.164	0.039	0.056	0.301	0.796	0.213

Table 5. The classification performance and explanation evaluation on the Face Glasses Recognition. The results are obtained from 3 individual runs and the best results of each metric are highlighted in boldface font.

Model	Architecture	Prediction		Exp Faithfulness		Exp Correctness			
		Acc. ↑	AUC ↑	Comp. ↑	Suff. ↓	IoU ↑	F1 ↑	Precision ↑	Recall ↑
Baseline 1	ResNet50	0.991	0.999	0.302	0.163	0.134	0.971	0.998	0.954
Baseline 2	VGG16	0.996	0.864	0.183	0.039	0.299	0.873	0.987	0.804
Supervised									
GRADIA [52]	ResNet50	0.990	0.999	0.368	0.262	0.375	0.949	0.993	0.917
RES [50]	ResNet50	0.991	0.999	0.396	0.128	0.302	0.932	0.997	0.887
RRR [125]	VGG16	0.994	0.999	0.384	0.160	0.332	0.909	0.992	0.864
CDEP [119]	VGG16	0.996	0.999	0.004	0.003	0.042	0.203	0.540	0.248
Unsupervised									
SENN [5]	SENN(CNN)	0.797	0.873	0.002	0.023	0.045	0.202	0.639	0.134
SGT [65]	ResNet50	0.996	0.999	0.083	0.106	0.292	0.671	0.974	0.556

For gender classification, GRADIA generally has the best prediction performance and presents the highest explanation quality, with the best scores in all metrics besides comprehensiveness and Exp recall, and minimal differences of 0.9% and 2.1% in comprehensiveness and explanatory recall. For models with ResNet50 as the backbone architecture, GRADIA and RES significantly outperform SGT, since SGT is unsupervised. SGT presents a lower accuracy, comprehensiveness, IoU, and explanatory F1 than the baseline model, which implies that the unsupervised model is not improving model performance and explanation quality. Yet SGT achieves the lowest sufficiency, which measures how well the prediction aligns between the original input and an explanation-generated input. Models with VGG16 as the backbone report lower sufficiency than those with ResNet50, while models with ResNet50 generally hold higher accuracy, comprehensiveness, IoU, and explanatory F1. SENN, which develops its own architecture with a set of conceptizer, parametrizer, and aggregator, underperforms in all metrics since the backbone is a simple CNN model.

Table 6. The classification performance and explanation evaluation on the Prohibited Item Detection. The results are obtained from 3 individual runs and the best results of each metric are highlighted in boldface font.

Model	Architecture	Prediction		Exp Faithfulness		Exp Correctness			
		Acc. ↑	AUC ↑	Comp. ↑	Suff. ↓	IoU ↑	F1 ↑	Precision ↑	Recall ↑
Baseline 1	ResNet50	0.961	0.992	0.161	0.053	0.195	0.823	0.870	0.784
Baseline 2	VGG16	0.917	0.988	-0.026	-0.010	0.147	0.330	0.788	0.241
Supervised									
GRADIA [52]	ResNet50	0.974	0.997	0.176	0.272	0.213	0.703	0.928	0.610
RES [50]	ResNet50	0.962	0.995	0.152	0.295	0.235	0.837	0.964	0.776
RRR [125]	VGG16	0.950	0.995	0.055	0.128	0.155	0.343	0.448	0.320
CDEP [119]	VGG16	0.959	0.992	0.038	0.021	0.061	0.403	0.615	0.298
Unsupervised									
SENN [5]	SENN(CNN)	0.754	0.839	-0.026	0.095	0.042	0.152	0.584	0.098
SGT [65]	ResNet50	0.962	0.993	0.027	0.066	0.062	0.552	0.648	0.481

For scene recognition, the baseline VGG16 achieves better accuracy, comprehensiveness, and sufficiency, compared with the baseline ResNet50. VGG16 has a significantly low explanatory recall and IoU, which results in 53.8% worse performance in explanatory F1. For the selected models, RES yields the best performance on all metrics, slightly improving prediction performance and boosting explanation quality significantly, with 1.0%, 2.4%, 177.9%, -99.2%, 9.6%, 2.8% changes in accuracy, AUC, comprehensiveness, sufficiency, IoU, and explanatory F1, respectively. SENN consistently underperforms in all metrics. Among models with VGG16, RRR is able to maintain a similar accuracy as the baseline while improving explanation correctness (IoU and Exp F1) by 17.3% and 12.3%, while CDEP results in worse explanation correctness. RRR and CDEP both obtain a lower score in comprehensiveness and sufficiency, which suggests that even stripping off the model-generated explanations, the models are able to generate similar predictions. The stripped model-generated explanations are useful in terms of prediction, but not all useful information is covered by the saliency map. Moreover, baseline ResNet50 under-performs in terms of explanation faithfulness but outperforms in terms of explanation correctness. This implies that while ResNet50 is able to generate explanations that exhibit the pattern of human annotation, the generated explanations are not useful for the model in terms of prediction.

In the task of the glasses identification, all models achieve high performance besides SENN, which consistently results in worse performance in all metrics. In terms of explanation comprehensiveness and faithfulness, GRADIA, RES, RRR, and CDEP show (21.9%, 60.7%), (31.1%, -21.5%), (109.8%, 310.3%), and (-97.8%, -92.3%) changes with respect to the baseline. RES holds the highest comprehensiveness score and CDEP holds the lowest sufficiency score. This implies that GRADIA, RES, and RRR are successful at extracting all useful attention for prediction, while GRADIA and RRR sacrifice the prediction power if solely using generated attention as the input. Among supervised models, GRADIA has the highest IoU as well as the highest percentage increase, which implies that the model successfully learns the pattern of human annotation and is able to produce saliency maps that are most aligned with human annotation. However, since GRADIA also has the highest sufficiency score, it further implies that learning from the human annotation may not be sufficient for the model to make correct predictions. Meanwhile, SGT shows high accuracy and IoU even as an unsupervised

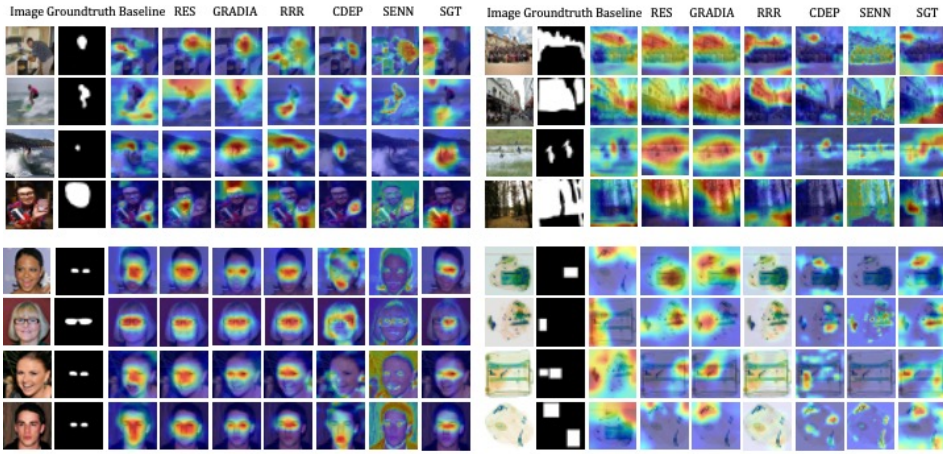


Fig. 5. Selected explanation visualization results on four vision datasets: gender classification, scene recognition, face glasses detection, and prohibited item identification. The model-generated explanations are highlighted.

model. Yet it suffers from low Explanatory recall, low comprehensiveness, and high sufficiency score, indicating that the model-generated attention is not successful in terms of prediction.

In the Prohibited Item Detection task, models generally exhibit a pattern of maintaining high accuracy, better explanation correctness, and comprehensiveness, but worse sufficiency. ResNet50 baseline achieves higher prediction and explanation robustness compared with VGG16. While most models' accuracy ranges from 0.917 to 0.974, SENN yields an accuracy of 0.754, which implies that a CNN model is insufficient for the dataset. All models with VGG16 as the backbone performs well in terms of low sufficiency but poorly in comprehensiveness and explanatory recall. GRADIA yields the highest accuracy (0.974), AUC (0.997) and comprehensiveness score (0.176), but poor sufficiency score of 0.272 when the baseline has a sufficiency of 0.053. RES shows the worst good performance on sufficiency but second to best score on comprehensiveness among EGL models. The baseline model with VGG16 achieves the best sufficiency score of -0.010. This is because the model sufficiency score is highly influenced by the size of generated annotation map, and the VGG16 baseline model sacrifices the sparseness of explanation to achieve a higher sufficiency and consequently leads to the worst comprehensiveness score. To validate the above statement, we further compute the proportion of attention map generated respectively to the entire image of the VGG16 baseline, RES, and GRADIA models. We find that the average explanation map sizes of the VGG16 baseline model are on average 84% and 33% greater than those of RES and GRADIA, respectively. This additional observation provides additional support for our assumption that the baseline model tends to generate larger explanation maps, leading to a much higher sufficiency score but a much worse comprehensiveness score. In terms of explanation correctness, RES and RRR are able to improve IoU from 0.195 to 0.101 and from 0.147 to 0.155, the best among each architecture. RES improves explanatory F1 from 0.827 to 0.837 and CDEP improves explanatory F1 from 0.330 to 0.403. Overall, GRADIA achieves the best prediction accuracy and faithfulness while RES achieves the best explanatory correctness among all models.

5.1.9 Qualitative Case Study. Figure 5 displays the visualization results of the four vision tasks: gender classification, scene recognition, face glasses detection, and prohibited item identification. Overall, RES and GRADIA have more overlap with the ground truth. CDEP is more fine-grained. SENN generates a saliency map that highlights all areas instead of focusing on a particular place. In the gender classification task, RES and GRADIA perform well in identifying the human body, even with distractions. RRR, CDEP, and SGT sometimes highlight areas that are disruptive, such as phones and kitchen stoves. In terms of saliency map size, CDEP generates a saliency map that focuses only on a small area, whereas SENN generates a thin layer of attention all over the image, which explains their low performance in IoU and explanatory F1. In the scene recognition task, RES, GRADIA, and SENN have the most overlapping areas with the ground truth label. RRR and CDEP sometimes are biased. For example, in the last image, RRR considers the floor key elements in scene recognition, whereas the ground truth label is the trees. SGT mostly attends to areas other than the ground truth, which explains its low IoU compared with other models. For face glasses detection, all models generate saliency maps focused on the face area. While the baseline model focus on the entire face, RES, GRADIA, RRR, and SGT generate maps that are more specific to the eye areas. CDEP focuses more on the lower half of the face and SENN attends to all face parts, such as the eyes, mouth, etc. While all models are highly accurate at identifying the presence of prohibited items, model-generated maps do not align with ground truth labels. Most models are able to focus on some small objects, not necessarily the prohibited items. SGT focuses on the white area outside of the baggage, which accounts for its low accuracy and low explanation performance.

5.2 Rationale Attention Guided Learning

To evaluate the performance of EGL models on NLP tasks, three datasets with explanation rationales are selected for the experimental study [36], the details about each dataset are shown as follows:

5.2.1 Movie Review [50]. The movie review dataset includes binary sentiment labels as well as rationale annotations at the span level. The task is to classify movies with positive sentiments from those with negative sentiments. we randomly split the data with a sample size of 1600/150/200 for training, validation, and testing. The explanation label is the sentiment $\in \{\text{positive, negative}\}$.

5.2.2 MultiRC [72]. MultiRC is a reading comprehension dataset originally composed of a series of rationale/question/answer triplets. This is also a binary classification task where the prediction label is True or False. The dataset is divided into 24029/3214/4848 for training, validation, and testing. The ground truth for explanation indicates if the answer is correct.

5.2.3 FEVER [145]. FEVER is a fact verification dataset, where each claim can be classified into supported, refuted, or not enough information. DeYoung et al. [36] further took a subset of the dataset and included only support and refuted claims. The dataset is further separated into 97957/6122/6111 images for training, validation, and testing respectively. For explanation rationales, the model has to predict the veracity of a claim $\in \{\text{support, refuse}\}$.

5.2.4 Evaluation Metrics. : We evaluate the model in two categories: 1) prediction performance and 2) explanation performance. For prediction performance, accuracy and AUC are computed to evaluate the predictive power of the model. For explanation evaluation, we incorporated 6 matrices to fully examine the explanation robustness. Matrix for comprehensiveness and sufficiency are derived from ERASER [36]. In addition, we measure the token level intersection over union (IoU) [16] between ground truth rationale and predicted rationale through IoU, explanatory F1, precision, and recall.

5.2.5 Comparison methods. : We compare the performance of several models listed below:

Table 7. The classification performance and explanation evaluation on the Movie Review dataset. The best results of each metric are highlighted in boldface font.

Model	Architecture	Prediction		Exp Faithfulness		Exp Correctness			
		Acc. \uparrow	AUC \uparrow	Comp. \uparrow	Suff. \downarrow	IoU \uparrow	F1 \uparrow	Precision \uparrow	Recall \uparrow
Baseline 1	BERT+MLP	0.516	0.478	0.086	0.145	0.242	0.365	0.441	0.312
Baseline 2	BERT+LSTM	0.622	0.591	0.027	0.126	0.043	0.112	0.462	0.064
Baseline 3	BERT+BERT	0.756	0.703	0.112	0.113	0.085	0.188	0.411	0.122
ERASER [36]	BERT+LSTM	0.826	0.805	0.128	0.093	0.598	0.749	0.734	0.765
Glockner et al.[55]	BERT+MLP	0.564	0.511	0.114	0.103	0.541	0.702	0.693	0.712
Carton et al. [22]	BERT+BERT	0.834	0.812	0.138	0.084	0.585	0.738	0.726	0.751
Expred [170]	BERT+GRU+MLP	0.794	0.779	0.094	0.076	0.639	0.779	0.781	0.779
FRESH [66]	BERT+LSTM	0.678	0.653	0.144	0.093	0.569	0.726	0.745	0.707

Table 8. The classification performance and explanation evaluation on the MultiRC dataset. The best results of each metric are highlighted in boldface font.

Model	Architecture	Prediction		Exp Faithfulness		Exp Correctness			
		Acc. \uparrow	AUC \uparrow	Comp. \uparrow	Suff. \downarrow	IoU \uparrow	F1 \uparrow	Precision \uparrow	Recall \uparrow
Baseline 1	BERT+MLP	0.564	0.511	0.012	0.188	0.235	0.459	0.534	0.402
Baseline 2	BERT+LSTM	0.593	0.573	0.081	0.205	0.106	0.280	0.471	0.199
Baseline 3	BERT+BERT	0.627	0.580	0.054	0.154	0.076	0.234	0.485	0.154
ERASER [36]	BERT+LSTM	0.639	0.615	0.039	0.132	0.448	0.618	0.615	0.622
Glockner et al. [55]	BERT+MLP	0.587	0.547	0.065	0.136	0.409	0.580	0.576	0.585
Carton et al. [22]	BERT+BERT	0.647	0.613	0.074	0.076	0.473	0.642	0.633	0.651
Expred [170]	BERT+GRU+MLP	0.638	0.622	0.032	0.061	0.447	0.618	0.602	0.635
FRESH [66]	BERT+LSTM	0.607	0.586	0.096	0.113	0.437	0.608	0.613	0.604

- **Baseline:** Baselines 1, 2, and 3 are pre-trained models that train only using the prediction loss without explanation loss. The pre-trained architecture for baselines 1, 2, and 3 are BERT+MLP, BERT+LSTM, and BERT+BERT respectively.
- **ERASER [36]:** A pipeline model that first trains the encoder to extract rationales, and then trains the decoder to perform prediction using only rationales.
- **Glockner et al. [55]:** A differentiable training-framework that aims to output faithful rationales on a sentence level
- **Carton et al. [22]:** A model that applies sentence-level rationale supervision, non-occluding “importance embeddings” on selective rationales with high sufficiency-accuracy.
- **Expred [170]:** A novel explanation generation framework work using multi-task learning that is task-aware and can exploit rationales data for effective explanations.
- **FRESH [66]:** A model that aims to produce faithful rationales for neural text classification by defining independent snippet extraction and prediction modules.

5.2.6 Implementation Details. : The data preprocessing follows the setting of ERASER [36]. We train all the models equally for 20 epochs and Adam is used for optimization with a learning rate of $2e-5$. To evaluate the explanation performance, the threshold for the calculated rationales is set to be 0.5. We follow the hyperparameter settings reported in the papers of the above methods.

Table 9. The classification performance and explanation evaluation on the Fever dataset. The best results of each metric are highlighted in boldface font.

Model	Architecture	Prediction		Exp Faithfulness		Exp Correctness			
		Acc. ↑	AUC ↑	Comp. ↑	Suff. ↓	IoU ↑	F1 ↑	Precision ↑	Recall ↑
Baseline 1	BERT+MLP	0.822	0.803	0.075	0.126	0.103	0.319	0.513	0.231
Baseline 2	BERT+LSTM	0.851	0.822	0.022	0.099	0.157	0.391	0.454	0.344
Baseline 3	BERT+BERT	0.872	0.856	0.017	0.117	0.036	0.145	0.612	0.082
ERASER [36]	BERT+LSTM	0.874	0.867	0.036	0.053	0.679	0.808	0.805	0.812
Glockner et al. [55]	BERT+MLP	0.835	0.813	0.122	0.066	0.672	0.803	0.833	0.776
Carton et al. [22]	BERT+BERT	0.893	0.876	0.084	0.048	0.707	0.828	0.831	0.826
Expred [170]	BERT+GRU+MLP	0.903	0.889	0.043	0.027	0.696	0.820	0.817	0.824
FRESH [66]	BERT+LSTM	0.862	0.832	0.106	0.053	0.627	0.771	0.732	0.814

5.2.7 Quantitative analysis. Tables 6, 7, and 8 present the model prediction performance and explanation quality of Movie Review, MultiRC, FEVER dataset respectively. The best results for each dataset are highlighted with boldface font. In general, when comparing with baseline, all models achieve a better accuracy and explanation correctness. The sufficiency score also decreases compared with the baseline model, which implies that the model-generated rationale is representative of the entire document.

For the Movie Review dataset, Carton et al. [22] yields the highest classification accuracy and Expred [170] generates the explanations with the highest quality. Compared with the baseline architecture BERT+LSTM, ERASER [36] improve the model accuracy and AUC by 32.8% and 36.2%, and boost the explanation quality by 374.1%, -26.2%, 1290.7%, and 568.8% in terms of comprehensiveness, sufficiency, IoU, and explanatory F1 scores, respectively, while FRESH [66] improves model accuracy, AUC, Sufficiency and Exp F1 by 9.0%, 10.5%, 433.3%, -26.2%, 1223.3%, and 548.2% respectively. ERASER [36] has better performance in terms of both model performance as well as explanation quality. Carton et al. [22], which employs the BERT+BERT architecture, increases accuracy and AUC by 10.3%, and 15.5% and ERASER [36] achieves the second-best result with an architecture of BERT+LSTM. Expred [170] obtain the highest explanation correctness and lowest sufficiency with a model architecture of BERT+GRU+MLP, and FRESH [66](BERT+LSTM) holds the highest comprehensiveness score among the selected models.

For the MultiRC dataset, Carton et al. [22] achieves the highest classification accuracy as well as the highest explanation faithfulness and correctness. It improves the model accuracy and AUC by 3.2%, 15.5%, lowers sufficiency score by 50.6%, and boost IoU and explanation F1 by 522.4%, and 174.4%, respectively, compared with the baseline. For all models with the architecture BERT+LSTM, while they consistently obtain better results than baseline except for comprehensiveness ERASER [36], outperforms FRESH [66] by 5.3%, 4.9%, 2.5%, and 1.6% in terms of model accuracy, AUC, IoU, and explanatory F1. FRESH [66] is more accurate when assessing explanation faithfulness through the sufficiency and comprehensiveness score, with a 14.4% decrease and 146.2% boost compared with ERASER [36].

The performance varies for the FEVER dataset, as FRESH [66] achieves the highest accuracy, AUC, and sufficiency scores, and Expred [170] yields the highest comprehensiveness, IoU, Explanatory F1 and Explanatory Recall. All the models perform generally well in the fact verification task in terms of accuracy, with a range of 0.835 to 0.903. In terms of explanatory faithfulness, FRESH [66] performs worse than the baseline in comprehensiveness but reduces sufficiency by 46.5%. Expred [170] obtain the greatest boost in comprehensive and sufficiency, with a change of 394.1% and -59.0% respectively.

Ground Truth	Baseline	ERASER	Glockner et al.	Carton et al.	Expred	FRESH
... the canadians can make a good movie . the world is coming to an end . we do n't know why or how , but apparently there is no way to stop it . the world has had this information for months , as most of the rioting and other assorted chaos has passed and governments have shut down operations ... these people 's lives however all intersect during their final six hours . writer - director - star don mckellar has crafted a highly unique and emotional film the canadians can make a good movie . the world is coming to an end . we do n't know why or how , but apparently there is no way to stop it . the world has had this information for months , as most of the rioting and other assorted chaos has passed and governments have shut down operations ... these people 's lives however all intersect during their final six hours . writer - director - star don mckellar has crafted a highly unique and emotional film the canadians can make a good movie . the world is coming to an end . we do n't know why or how , but apparently there is no way to stop it . the world has had this information for months , as most of the rioting and other assorted chaos has passed and governments have shut down operations ... these people 's lives however all intersect during their final six hours . writer - director - star don mckellar has crafted a highly unique and emotional film the canadians can make a good movie . the world is coming to an end . we do n't know why or how , but apparently there is no way to stop it . the world has had this information for months , as most of the rioting and other assorted chaos has passed and governments have shut down operations ... these people 's lives however all intersect during their final six hours . writer - director - star don mckellar has crafted a highly unique and emotional film the canadians can make a good movie . the world is coming to an end . we do n't know why or how , but apparently there is no way to stop it . the world has had this information for months , as most of the rioting and other assorted chaos has passed and governments have shut down operations ... these people 's lives however all intersect during their final six hours . writer - director - star don mckellar has crafted a highly unique and emotional film the canadians can make a good movie . the world is coming to an end . we do n't know why or how , but apparently there is no way to stop it . the world has had this information for months , as most of the rioting and other assorted chaos has passed and governments have shut down operations ... these people 's lives however all intersect during their final six hours . writer - director - star don mckellar has crafted a highly unique and emotional film the canadians can make a good movie . the world is coming to an end . we do n't know why or how , but apparently there is no way to stop it . the world has had this information for months , as most of the rioting and other assorted chaos has passed and governments have shut down operations ... these people 's lives however all intersect during their final six hours . writer - director - star don mckellar has crafted a highly unique and emotional film .
Prediction: Positive	Prediction: Negative	Prediction: Positive	Prediction: Positive	Prediction: Positive	Prediction: Positive	Prediction: Negative

Fig. 6. Selected explanation visualization results on FEVER dataset. The model-generated explanations are highlighted.

Ground Truth	Baseline	ERASER	Glockner et al.	Carton et al.	Expred	FRESH
... fossils in younger rocks look like animals and plants that are living today . fossils in older rocks are less like living organisms . fossils can tell us about where the organism lived . was it land or marine ? fossils can even tell us if the water was shallow or deep . fossils can even provide clues to ancient climates . what can we tell about former living organisms from fossils ? how they died	... fossils in younger rocks look like animals and plants that are living today . fossils in older rocks are less like living organisms . fossils can tell us about where the organism lived . was it land or marine ? fossils can even tell us if the water was shallow or deep . fossils can even provide clues to ancient climates . what can we tell about former living organisms from fossils ? how they died	... fossils in younger rocks look like animals and plants that are living today . fossils in older rocks are less like living organisms . fossils can tell us about where the organism lived . was it land or marine ? fossils can even tell us if the water was shallow or deep . fossils can even provide clues to ancient climates . what can we tell about former living organisms from fossils ? how they died	... fossils in younger rocks look like animals and plants that are living today . fossils in older rocks are less like living organisms . fossils can tell us about where the organism lived . was it land or marine ? fossils can even tell us if the water was shallow or deep . fossils can even provide clues to ancient climates . what can we tell about former living organisms from fossils ? how they died	... fossils in younger rocks look like animals and plants that are living today . fossils in older rocks are less like living organisms . fossils can tell us about where the organism lived . was it land or marine ? fossils can even tell us if the water was shallow or deep . fossils can even provide clues to ancient climates . what can we tell about former living organisms from fossils ? how they died	... fossils in younger rocks look like animals and plants that are living today . fossils in older rocks are less like living organisms . fossils can tell us about where the organism lived . was it land or marine ? fossils can even tell us if the water was shallow or deep . fossils can even provide clues to ancient climates . what can we tell about former living organisms from fossils ? how they died	... fossils in younger rocks look like animals and plants that are living today . fossils in older rocks are less like living organisms . fossils can tell us about where the organism lived . was it land or marine ? fossils can even tell us if the water was shallow or deep . fossils can even provide clues to ancient climates . what can we tell about former living organisms from fossils ? how they died
Prediction: False	Prediction: True	Prediction: False	Prediction: True	Prediction: False	Prediction: False	Prediction: True

Fig. 7. Selected explanation visualization results on Movie Review dataset. The model-generated explanations are highlighted.

The baseline models generally show poor performance in explanation faithfulness and correctness, which are improved significantly across all five models. Expred [170] is able to improve IoU by 1863.9% and Explanatory F1 by 471.0%.

5.2.8 Qualitative Case Study. Figures 6, 7, and 8 provide examples of visualization results on FEVER, Movie Review, and MultiRC dataset. The model-generated explanations are highlighted. In general, the baseline model highlights areas that are scattered all around the corpus, whereas trained models generate explanation rationales that are more aggregated. In Figure 6, ERASER [36] and Glockner et al. [55] are highly aligned with ground truth, aligned with their high performance in IoU. While Expred [170] obtains the highest accuracy and comprehensiveness, its generated-explanation does not align with the ground truth annotations, which implies that the ground truth labels may not be sufficient for the model to learn the prediction. FRESH [66] generates explanations that are poorly aligned with the ground truth and outputs a wrong prediction label.

In Figure 7, while Carton et al. [22] aligns well with the ground truth, it focuses on a higher percent of tokens, which explains why it slightly underperforms in explanatory precision and comprehensiveness but outperforms in sufficiency. There exhibits a compromise between high

Ground Truth	Baseline	ERASER	Glockner et al.	Carton et al.	Expred	FRESH
Mount Hood , called Wy'east by the Multnomah tribe , is a potentially active stratovolcano in the Cascade Volcanic Arc of northern Oregon . It was formed by a subduction zone on the Pacific coast and rests in the Pacific Northwest region of the United States . It is located about 50 mi east-southeast of Portland , on the border between Clackamas and Hood River counties ... Mount Hood is in the Pyrenees .	Mount Hood , called Wy'east by the Multnomah tribe , is a potentially active stratovolcano in the Cascade Volcanic Arc of northern Oregon . It was formed by a subduction zone on the Pacific coast and rests in the Pacific Northwest region of the United States . It is located about 50 mi east-southeast of Portland , on the border between Clackamas and Hood River counties ... Mount Hood is in the Pyrenees .	Mount Hood , called Wy'east by the Multnomah tribe , is a potentially active stratovolcano in the Cascade Volcanic Arc of northern Oregon . It was formed by a subduction zone on the Pacific coast and rests in the Pacific Northwest region of the United States . It is located about 50 mi east-southeast of Portland , on the border between Clackamas and Hood River counties ... Mount Hood is in the Pyrenees .	Mount Hood , called Wy'east by the Multnomah tribe , is a potentially active stratovolcano in the Cascade Volcanic Arc of northern Oregon . It was formed by a subduction zone on the Pacific coast and rests in the Pacific Northwest region of the United States . It is located about 50 mi east-southeast of Portland , on the border between Clackamas and Hood River counties ... Mount Hood is in the Pyrenees .	Mount Hood , called Wy'east by the Multnomah tribe , is a potentially active stratovolcano in the Cascade Volcanic Arc of northern Oregon . It was formed by a subduction zone on the Pacific coast and rests in the Pacific Northwest region of the United States . It is located about 50 mi east-southeast of Portland , on the border between Clackamas and Hood River counties ... Mount Hood is in the Pyrenees .	Mount Hood , called Wy'east by the Multnomah tribe , is a potentially active stratovolcano in the Cascade Volcanic Arc of northern Oregon . It was formed by a subduction zone on the Pacific coast and rests in the Pacific Northwest region of the United States . It is located about 50 mi east-southeast of Portland , on the border between Clackamas and Hood River counties ... Mount Hood is in the Pyrenees .	Mount Hood , called Wy'east by the Multnomah tribe , is a potentially active stratovolcano in the Cascade Volcanic Arc of northern Oregon . It was formed by a subduction zone on the Pacific coast and rests in the Pacific Northwest region of the United States . It is located about 50 mi east-southeast of Portland , on the border between Clackamas and Hood River counties ... Mount Hood is in the Pyrenees .
Prediction: REFUTES	Prediction: SUPPORTS	Prediction: REFUTES	Prediction: REFUTES	Prediction: REFUTES	Prediction: REFUTES	Prediction: SUPPORTS

Fig. 8. Selected explanation visualization results on MultiRC dataset. The model-generated explanations are highlighted.

accuracy and high explanation quality, as Carton et al. [22] achieves the highest accuracy but lowest comprehensiveness among the selected models. This examples shows how the amount of attention may manipulate the result of explanation faithfulness. If a high amount of tokens are considered important, sufficiency will be close to 0 and comprehensiveness will be relatively high. Therefore, it's necessary to consider both explanation faithfulness and correctness when analyzing the explanation quality. This example also reveals the importance of a case study, to visualize the quantitative results and understand how attention performs in terms of correctness and faithfulness.

6 CONCLUSION

This survey has presented a comprehensive survey of existing methodologies developed in the field of Explanation-Guided Learning (EGL), a group of techniques that applies XAI-driven insights to steer the DNNs' behavior in realizing iterative model revision. It provides an extensive overview of the EGL challenges, techniques, applications, evaluation procedures, as well as extensive experimental comparison among existing techniques under popular application areas. It summarizes the findings of the research presented in more than 150 publications on EGL, the majority of which were released in the last five years. Concretely, in this survey, the formal definition of EGL and its general learning paradigm is first given, along with an overview of the key factors for EGL evaluation, as well as summarization and categorization of existing evaluation procedures and metrics for EGL are provided. Based upon the numerous historical and state-of-the-art works discussed in this survey, the article concludes by discussing the current and potential future application areas of EGL, and provides an extensive experimental study that aims at providing the first comprehensive comparative study among existing EGL models in various popular application domains, such as Computer Vision (CV) and Natural Language Processing (NLP) domains.

REFERENCES

- [1] David A Bennett, Julie A Schneider, Zoe Arvanitakis, and Robert S Wilson. 2012. Overview and findings from the religious orders study. *Current Alzheimer Research* 9, 6 (2012), 628–645.
- [2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.
- [3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).

- [4] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4971–4980.
- [5] David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems* 31 (2018).
- [6] David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049* (2018).
- [7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- [8] Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. Marta: Leveraging human rationales for explainable text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 5868–5876.
- [9] 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.
- [10] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2021. Diagnostics-Guided Explanation Generation. *arXiv preprint arXiv:2109.03756* (2021).
- [11] 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.
- [12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [13] Vladimir Balayan, Pedro Saleiro, Catarina Belém, Ludwig Krippahl, and Pedro Bizarro. 2020. Teaching the Machine to Explain Itself using Domain Knowledge. *arXiv preprint arXiv:2012.01932* (2020).
- [14] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.
- [15] Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367* (2018).
- [16] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*. 6541–6549.
- [17] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.
- [18] Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, and Luca Daniel. 2020. Proper network interpretability helps adversarial robustness in classification. In *International Conference on Machine Learning*. PMLR, 1014–1023.
- [19] Nadia Burkart, Philipp M Faller, Elisabeth Peinsipp, and Marco F Huber. 2020. Batch-wise regularization of deep neural networks for interpretability. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 216–222.
- [20] Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317.
- [21] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems* 31 (2018).
- [22] Samuel Carton, Surya Kanoria, and Chenhao Tan. 2021. What to learn, and how: Toward effective learning from rationales. *arXiv preprint arXiv:2112.00071* (2021).
- [23] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).
- [24] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2019. A game theoretic approach to class-wise selective rationalization. *Advances in neural information processing systems* 32 (2019).
- [25] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. 2020. Air: Attention with reasoning capability. In *European Conference on Computer Vision*. Springer, 91–107.
- [26] Seungtaek Choi, Haeju Park, Jinyoung Yeo, and Seung-won Hwang. 2020. Less is more: Attention supervision with counterfactuals for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6695–6704.

- [27] George Chrysostomou and Nikolaos Aletras. 2021. Enjoy the salience: Towards better transformer-based faithful explanations with word salience. *arXiv preprint arXiv:2108.13759* (2021).
- [28] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
- [29] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kallou, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 168–172.
- [30] Dennis Collaris and Jarke J van Wijk. 2020. ExplainExplore: Visual exploration of machine learning explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 26–35.
- [31] Owen Cornec, Rahul Nair, Elizabeth Daly, Dennis Wei, and Oznur Alkan. 2021. AIMEE: Interactive model maintenance with rule-based surrogates. In *Annual Conference on Neural Information Processing Systems*.
- [32] Elizabeth M Daly, Massimiliano Mattetti, Öznur Alkan, and Rahul Nair. 2021. User Driven Model Adjustment via Boolean Rule Explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 5896–5904.
- [33] 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.
- [34] Luc De Raedt, Sebastijan Dumančić, Robin Manhaeve, and Giuseppe Marra. 2020. From statistical relational to neuro-symbolic artificial intelligence. *arXiv preprint arXiv:2003.08316* (2020).
- [35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [36] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. ERASER: A benchmark to evaluate rationalized NLP models. *arXiv preprint arXiv:1911.03429* (2019).
- [37] KC Dharma and Chicheng Zhang. 2021. Improving the trustworthiness of image classification models by utilizing bounding-box annotations. *CoRR* (2021).
- [38] Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. 2021. Learning credible DNNs via incorporating prior knowledge and model local explanation. *Knowledge and Information Systems* 63, 2 (2021), 305–332.
- [39] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems* 36, 4 (2020), 25–34.
- [40] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37.
- [41] Sayna Ebrahimi, Suzanne Petryk, Akash Gokul, William Gan, Joseph E Gonzalez, Marcus Rohrbach, and Trevor Darrell. 2021. Remembering for the right reasons: Explanations reduce catastrophic forgetting. *Applied AI Letters* 2, 4 (2021), e44.
- [42] Alex Endert, William Ribarsky, Cagatay Turkay, BL William Wong, Ian Nabney, I Díaz Blanco, and Fabrice Rossi. 2017. The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 458–486.
- [43] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. 2019. Learning explainable models using attribution priors. (2019).
- [44] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. 2021. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence* 3, 7 (2021), 620–631.
- [45] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.
- [46] Patrick Fernandes, Marcos Treviso, Danish Pruthi, André FT Martins, and Graham Neubig. 2022. Learning to Scaffold: Optimizing Model Explanations for Teaching. *arXiv preprint arXiv:2204.10810* (2022).
- [47] 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.
- [48] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. 2017. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*. 1811–1820.
- [49] Yuyang Gao, Giorgio A Ascoli, and Liang Zhao. 2021. Schematic memory persistence and transience for efficient and robust continual learning. *Neural Networks* 144 (2021), 49–60.
- [50] Yuyang Gao, Tong Sun, Guangji Bai, Siyi Gu, Sungsoo Ray Hong, and Liang Zhao. 2022. RES: A Robust Framework for Guiding Visual Explanation. In *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

- [51] 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.
- [52] 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*
- [53] Reza Ghaeini, Xiaoli Z Fern, Hamed Shahbazi, and Prasad Tadepalli. 2019. Saliency learning: Teaching the model where to pay attention. *arXiv preprint arXiv:1902.08649* (2019).
- [54] Yolanda Gil, James Honaker, Shikhar Gupta, Yibo Ma, Vito D’Orazio, Daniel Garijo, Shruti Gadewar, Qifan Yang, and Neda Jahanshad. 2019. Towards human-guided machine learning. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 614–624.
- [55] Max Glockner, Ivan Habernal, and Iryna Gurevych. 2020. Why do you think that? exploring faithful sentence-level rationales without supervision. *arXiv preprint arXiv:2010.03384* (2020).
- [56] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.
- [57] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, et al. 2015. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* 47, 6 (2015), 569–576.
- [58] 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.
- [59] Nicholas Halliwell and Freddy Lecue. 2020. Trustworthy convolutional neural networks: A gradient penalized-based approach. *arXiv preprint arXiv:2009.14260* (2020).
- [60] Xiaochuang Han and Yulia Tsvetkov. 2021. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. *arXiv preprint arXiv:2110.03212* (2021).
- [61] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 771–787.
- [62] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).
- [63] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction* 4 (2020), 1–26.
- [64] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6700–6709.
- [65] Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. 2021. Improving deep learning interpretability by saliency guided training. *Advances in Neural Information Processing Systems* 34 (2021), 26726–26739.
- [66] Sarthak Jain, Sarah Wiegrefe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. *arXiv preprint arXiv:2005.00115* (2020).
- [67] Hoyong Jeong, Suyoung Lee, Sung Ju Hwang, and Soeul Son. 2022. Learning to Generate Inversion-Resistant Model Explanations. In *Advances in Neural Information Processing Systems*.
- [68] Liu Jiang, Shixia Liu, and Changjian Chen. 2019. Recent research advances on interactive machine learning. *Journal of Visualization* 22, 2 (2019), 401–417.
- [69] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2901–2910.
- [70] Teja Kanchinadam, Keith Westpfahl, Qian You, and Glenn Fung. 2020. Rationale-based Human-in-the-Loop via Supervised Attention.. In *DaSH@ KDD*.
- [71] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1244–1254.
- [72] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *NAACL*.
- [73] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [74] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.
- [75] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

- [76] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).
- [77] David WG Langerhuizen, Anne Eva J Bulstra, Stein J Janssen, David Ring, Gino MMJ Kerkhoffs, Ruurd L Jaarsma, and Job N Doornberg. 2020. Is deep learning on par with human observers for detection of radiographically visible and occult fractures of the scaphoid? *Clinical orthopaedics and related research* 478, 11 (2020), 2653.
- [78] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
- [79] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [80] Seunggho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. 2021. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5495–5505.
- [81] Seungeon Lee, Xiting Wang, Sungwon Han, Xiaoyuan Yi, Xing Xie, and Meeyoung Cha. 2022. Self-explaining deep models with logic rule reasoning. *arXiv preprint arXiv:2210.07024* (2022).
- [82] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155* (2016).
- [83] Piyaawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics* 9 (2021), 1508–1528.
- [84] Kunpeng Li, Ziyang Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. 2018. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9215–9223.
- [85] Yi Li and Nuno Vasconcelos. 2019. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9572–9581.
- [86] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems* (2021).
- [87] Moshe Lichman et al. 2013. UCI machine learning repository.
- [88] 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.
- [89] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2020), 18.
- [90] Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. *arXiv preprint arXiv:1906.08286* (2019).
- [91] Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2021. Rationale-inspired natural language explanations with commonsense. *arXiv preprint arXiv:2106.13876* (2021).
- [92] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How data scientists work together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–23.
- [93] Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. 2012. A Bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling* 52, 6 (2012), 1686–1697.
- [94] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. 2016. DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science* 3 (2016), 80.
- [95] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
- [96] Caijing Miao, Lingxi Xie, Fang Wan, chi Su, Hongye Liu, Jianbin Jiao, and Qixiang Ye. 2019. SIXray: A Large-scale Security Inspection X-ray Benchmark for Prohibited Item Discovery in Overlapping Images. In *CVPR*.
- [97] Henry W Miller. 1973. Plan and operation of the health and nutrition examination survey, United States, 1971-1973. *DHEW publication no.(PHS)-Dept. of Health, Education, and Welfare (USA)* (1973).
- [98] Jeremy A Miller, Angela Guillozet-Bongaarts, Laura E Gibbons, Nadia Postupna, Anne Renz, Allison E Beller, Susan M Sunkin, Lydia Ng, Shannon E Rose, Kimberly A Smith, et al. 2017. Neuropathological and transcriptomic characteristics of the aged brain. *Elife* 6 (2017), e31126.
- [99] 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).
- [100] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.
- [101] 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.
- [102] 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.
 - [103] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital signal processing* 73 (2018), 1–15.
 - [104] Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1069–1078.
 - [105] Giang Nguyen, Mohammad Reza Taesiri, and Anh Nguyen. 2022. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. *arXiv preprint arXiv:2208.00780* (2022).
 - [106] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.
 - [107] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks* 113 (2019), 54–71.
 - [108] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8779–8788.
 - [109] Badri Patro, Vinay Nambodiri, 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.
 - [110] Tejaswini Pedapati, Avinash Balakrishnan, Karthikeyan Shanmugam, and Amit Dhurandhar. 2020. Learning global transparent models consistent with local contrastive explanations. *Advances in neural information processing systems* 33 (2020), 3592–3602.
 - [111] Yan Peng, Zheng Xuefeng, Zhu Jianyong, and Xiao Yumhong. 2009. Lazy learner text categorization algorithm based on embedded feature selection. *Journal of Systems Engineering and Electronics* 20, 3 (2009), 651–659.
 - [112] Vipin Pillai and Hamed Pirsiavash. 2021. Explainable models with consistent interpretations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2431–2439.
 - [113] Gregory Plumb, Maruan Al-Shedivat, Ángel Alexander Cabrera, Adam Perer, Eric Xing, and Ameet Talwalkar. 2020. Regularizing black-box models for improved interpretability. *Advances in Neural Information Processing Systems* 33 (2020), 10526–10536.
 - [114] Teodora Popordanoska, Mohit Kumar, and Stefano Teso. 2020. Machine guides, human supervises: Interactive learning with global explanations. *arXiv preprint arXiv:2009.09723* (2020).
 - [115] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabudde, and Fabrice Meriaudeau. 2018. Indian Diabetic Retinopathy Image Dataset (IDRiD). <https://doi.org/10.21227/H25W98>
 - [116] 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.
 - [117] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361* (2019).
 - [118] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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*. 1135–1144.
 - [119] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. 2020. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*. PMLR, 8116–8126.
 - [120] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. 2020. Explainable machine learning for scientific insights and discoveries. *Ieee Access* 8 (2020), 42200–42216.
 - [121] Andrew Ross and Finale Doshi-Velez. 2018. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
 - [122] 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).
 - [123] Gobinda Saha and Kaushik Roy. 2021. Saliency Guided Experience Packing for Replay in Continual Learning. *arXiv preprint arXiv:2109.04954* (2021).
 - [124] Johannes Schneider and Michalis Vlachos. 2020. Reflective-net: Learning from explanations. *arXiv preprint arXiv:2011.13986* (2020).
 - [125] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence* 2, 8 (2020), 476–486.

- [126] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696* (2020).
- [127] 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.
- [128] Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731* (2019).
- [129] Rudy Setiono and Huan Liu. 1997. Neural-network feature selector. *IEEE transactions on neural networks* 8, 3 (1997), 654–662.
- [130] Xiaoting Shao, Tjitze Rienstra, Matthias Thimm, and Kristian Kersting. 2020. Towards understanding and arguing with classifiers: Recent progress. *Datenbank-Spektrum* 20, 2 (2020), 171–180.
- [131] Manali Sharma, Di Zhuang, and Mustafa Bilgic. 2015. Active learning with rationales for text classification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 441–451.
- [132] 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.
- [133] Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. 2019. Gradmask: Reduce overfitting by regularizing saliency. *arXiv preprint arXiv:1904.07478* (2019).
- [134] Chandan Singh, Wooseok Ha, and Bin Yu. 2022. Interpreting and improving deep-learning models with reality checks. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 229–254.
- [135] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. 2020. Don't judge an object by its context: learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11070–11078.
- [136] Xuelin Situ, Ingrid Zukerman, Cecile Paris, Sameen Maruf, and Gholamreza Haffari. 2021. Learning to explain: Generating stable explanations fast. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5340–5355.
- [137] Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems* 33 (2020), 6327–6341.
- [138] Joe Stacey, Yonatan Belinkov, and Marek Rei. 2021. Supervising model attention with human explanations for robust natural language inference. *arXiv preprint arXiv:2104.08142* (2021).
- [139] Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. 2021. 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*. 3619–3629.
- [140] Julia Strout, Ye Zhang, and Raymond J Mooney. 2019. Do human rationales improve machine explanations? *arXiv preprint arXiv:1905.13714* (2019).
- [141] Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. 2016. Computational modeling of β -secretase 1 (BACE-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling* 56, 10 (2016), 1936–1949.
- [142] Chenhao Tan. 2021. On the diversity and limits of human explanations. *arXiv preprint arXiv:2106.11988* (2021).
- [143] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhofen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4631–4640.
- [144] Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 239–245.
- [145] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. <https://doi.org/10.18653/v1/N18-1074>
- [146] Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems* 32, 11 (2020), 4793–4813.
- [147] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [148] Stanislav Vojř and Tomáš Kliegr. 2020. Editable machine learning models? A rule-based framework for user studies of explainability. *Advances in Data Analysis and Classification* 14, 4 (2020), 785–799.
- [149] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

- [150] Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. *arXiv preprint arXiv:1906.00363* (2019).
- [151] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2097–2106.
- [152] Ethan Weinberger, Joseph Janizek, and Su-In Lee. 2020. Learning deep attribution priors based on prior knowledge. *Advances in Neural Information Processing Systems* 33 (2020), 14034–14045.
- [153] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.
- [154] Sarah Wiegrefe, Ana Marasović, and Noah A Smith. 2020. Measuring association between labels and free-text rationales. *arXiv preprint arXiv:2010.12762* (2020).
- [155] Jialin Wu and Raymond Mooney. 2019. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems* 32 (2019).
- [156] Mike Wu, Sonali Parbhoo, Michael Hughes, Ryan Kindle, Leo Celi, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. 2020. Regional tree regularization for interpretability in deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6413–6421.
- [157] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* 9, 2 (2018), 513–530.
- [158] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).
- [159] Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. 2021. Refining language models with compositional explanations. *Advances in Neural Information Processing Systems* 34 (2021), 8954–8967.
- [160] Zhuofan Ying, Peter Hase, and Mohit Bansal. 2022. VisFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives. *arXiv preprint arXiv:2206.11212* (2022).
- [161] Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia, and Shixia Liu. 2021. A survey of visual analytics techniques for machine learning. *Computational Visual Media* 7, 1 (2021), 3–36.
- [162] Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 conference on Empirical methods in natural language processing*. 31–40.
- [163] Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using “annotator rationales” to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*. 260–267.
- [164] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6720–6731.
- [165] Guohang Zeng, Yousef Kowsar, Sarah Erfani, and James Bailey. 2021. Generating Deep Networks Explanations with Robust Attribution Alignment. In *Asian Conference on Machine Learning*. PMLR, 753–768.
- [166] Quan-shi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 27–39.
- [167] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
- [168] Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, Vol. 2016. NIH Public Access, 795.
- [169] Yundong Zhang, Juan Carlos Nibbles, 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.
- [170] Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and predict, and then predict again. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 418–426.
- [171] 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).
- [172] Ruiqi Zhong, Steven Shao, and Kathleen McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870* (2019).
- [173] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *CVPR*. 2921–2929.
- [174] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *TPAMI* (2017).

- [175] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (2021), 593.
- [176] Jiaxin Zhuang, Jiabin Cai, Ruixuan Wang, Jianguo Zhang, and Weishi Zheng. 2019. Care: Class attention to regions of lesion for classification on imbalanced data. In *International Conference on Medical Imaging with Deep Learning*. PMLR, 588–597.