AI

MDPI

Review

Explainable Image Classification: The Journey So Far and the Road Ahead

Vidhya Kamakshi [1] **and Narayanan C. Krishnan** [1,2,*]

[1] Department of Computer Science and Engineering, Indian Institute of Technology Ropar, Rupnagar 140001, India; 2017csz0005@iitrpr.ac.in
[2] Department of Data Science, Indian Institute of Technology Palakkad, Palakkad 678557, India
* Correspondence: ckn@iitpkd.ac.in

Abstract: Explainable Artificial Intelligence (XAI) has emerged as a crucial research area to address the interpretability challenges posed by complex machine learning models. In this survey paper, we provide a comprehensive analysis of existing approaches in the field of XAI, focusing on the tradeoff between model accuracy and interpretability. Motivated by the need to address this tradeoff, we conduct an extensive review of the literature, presenting a multi-view taxonomy that offers a new perspective on XAI methodologies. We analyze various sub-categories of XAI methods, considering their strengths, weaknesses, and practical challenges. Moreover, we explore causal relationships in model explanations and discuss approaches dedicated to explaining cross-domain classifiers. The latter is particularly important in scenarios where training and test data are sampled from different distributions. Drawing insights from our analysis, we propose future research directions, including exploring explainable allied learning paradigms, developing evaluation metrics for both traditionally trained and allied learning-based classifiers, and applying neural architectural search techniques to minimize the accuracy–interpretability tradeoff. This survey paper provides a comprehensive overview of the state-of-the-art in XAI, serving as a valuable resource for researchers and practitioners interested in understanding and advancing the field.

Keywords: explainable AI survey; interpretable image classification; cross-domain explainers; causal explanations; posthoc explanations; antehoc explanations; concept-based explanations; natural language explanations; counterfactual explanations; model-agnostic explanations

Citation: Kamakshi, V.; Krishnan, N.C. Explainable Image Classification: The Journey So Far and the Road Ahead. *AI* **2023**, *4*, 620–651. https://doi.org/10.3390/ai4030033

Academic Editors: Mobyen Uddin Ahmed and Rosina O Weber

Received: 5 May 2023
Revised: 27 June 2023
Accepted: 20 July 2023
Published: 1 August 2023

1. Introduction

Image classification has undergone significant advancements, transitioning from simple hand-crafted feature extractors [1,2] to the use of deep models [3,4] that can automatically extract relevant features, resulting in improved classification performance. However, as models become more complex, they also become more opaque, hindering interpretability. Unlike earlier models that were transparent in their working mechanism, state-of-the-art deep models achieve high accuracy at the cost of interpretability. The field of explainable AI (XAI) has emerged to address this trade-off between performance and interpretability. XAI aims to unravel the decision-making process of complex black box models in a human-interpretable manner [5]. By providing explanations, XAI techniques can enhance user trust and enable the adoption of opaque models in safety-critical domains such as healthcare [6] and finance [7], where transparency is essential.

This survey paper focuses on organizing XAI approaches that explain the working of Convolutional Neural Networks (CNNs), which are state-of-the-art models for image classification. While there exist various surveys in the literature [6–10] with different aims and scopes, our paper aims to provide a multi-view taxonomy of XAI approaches by carefully analyzing the existing literature. The taxonomy considers the incorporation of explainability during the training phase (antehoc) and approximating the black box's working mechanism without disturbing the deployed model (posthoc). We also discuss other

bases of categorization where applicable to provide a comprehensive understanding of the taxonomy. For instance, we discuss XAI approaches that consider the causal relationships between the input and output. Explaining image classifiers based on causal relationships poses a challenge due to the limited knowledge of the underlying causal structure between image features. Some approaches [11–14] aim to extract the causal relationships exhibited by model representations and compare them with available proxy domain knowledge in a posthoc manner, without disrupting the black box classifier being explained. Another category of approaches [15–17] enforces the classifiers to adhere to the existing causal relationships, similar to the antehoc or explainability-by-design approaches, in order to make predictions.

While the XAI community has primarily focused on explaining traditionally learned in-domain classifiers trained on a single data distribution, deep models trained using such traditional paradigms require large amounts of labeled data to achieve better generalization. However, collecting extensive labeled data is challenging in real-world scenarios. To leverage the power of deep models in data-sparse scenarios, cross-domain classification paradigms [18,19] have emerged, where the model is expected to handle the distribution from which the sparse data is sampled, leveraging knowledge acquired from publicly available voluminous data sampled from a different distribution. This phenomenon, a crucial factor in the widespread adoption of deep models, necessitates explanation, and initial efforts [20,21] have been made in this direction. Given the limited applicability and the complexity of explaining the black box introduced in explaining the cross-domain classifiers, recent approaches [22,23] have leveraged antehoc explainability to build self-explaining cross-domain classifiers. Analyzing this direction opens up future research possibilities, such as devising suitable evaluation metrics and extending the benefits of explainability to classifiers utilizing allied learning paradigms.

The organization of this paper is as follows. Section 2 lists the common terms prevalent in XAI literature. Section 3 discusses the methodology adopted to collect relevant articles surveyed in this paper. The discussion on the survey is begun with a walk-through of the object recognition models (Section 4), shedding light on the accuracy–interpretability tradeoff observed as the community marched from leveraging low-performing interpretable models to high-performing opaque models for the task. After motivating the need to address this tradeoff (Section 5), and a thorough review of the existing literature (Section 6), a multi-view taxonomy (Section 7) has been proposed, and the relevance of the categorization with respect to existing general XAI taxonomies has been put forth. This is followed by a detailed discussion of the different sub-categories of methods analyzing their strengths, weaknesses, and challenges associated with the practical realization of the approaches. While most approaches highlight the correlational aspects between the input variables, some approaches bring in the aspect of causal relationships to be verified or enforced in the model's representation. Section 8 is dedicated to the discussion of such approaches. Section 9 aims to shift the readers' attention towards the under-explored sub-direction of explaining cross-domain models whose training and test data are sampled from different distributions unlike a traditionally learned classifier where all data points are sampled from a single distribution. Based on a thorough analysis carried out in these sections, the final section of the paper (Section 10) discusses open research directions for future XAI research. Specifically, three directions have been envisioned pertaining to explainable allied learning, devising evaluation metrics to validate the approaches dealing with classifiers trained traditionally as well as using allied learning paradigms, and ideas to automatically search for architectures that minimize the long-existing problem of the XAI community regarding mitigating the accuracy–interpretability tradeoff.

2. Basic Definitions

This section lists a glossary of various terms necessary to understand the language prevalent in the Explainable AI research community.

- **Black Box** refers to the model whose working needs to be explained. This is also called the Explanandum in XAI literature. In this survey focussing on the image classification task, the Convolutional Neural Networks (CNNs), among the state-of-the-art image classifiers, are the black boxes whose explanation is sought.
- **Explainer** refers to the approximator or the algorithmic procedure that explains the working mechanism of the black box.
- **Classifier** refers to the model that maps the instance to one of the pre-defined categories called classes.
- **In-domain** classifier refers to a classifier that is trained and tested on data sampled from the same distribution, while **Cross-domain** classifiers would be trained and tested on data sampled from different distributions.
- **Explanation** refers to a simplified illustration of the working mechanism the black box model under consideration employs.
- **Inherently Interpretable Models** refer to the family of Machine Learning models whose working mechanism can be summarized in a user-friendly manner. For example, Decision trees whose working can be viewed as a disjunction of conjunctions of various constraints on the input variables, Linear Regressors whose linear combination weights provide an assessment of the priority the model gives to each input variable, are among the inherently interpretable models.
- **Faithfulness** refers to the extent to which the explainer mimics the working mechanism of the black box it explains.
- **Local Explanations** refers to the category of explanations whose reliability is limited to a small neighborhood around the instance of interest to be explained. On the other hand, **Global Explanations** are reliable anywhere in the entire instance space.
- **Posthoc Explanations** refer to the category of explanations that approximate the working mechanism of the black box without making any modifications to its architecture or parameters. On the contrary, the other family of explanations called the **Antehoc Explanations** enforce changes to the black box under consideration so that it gains the ability to explain itself analogous to that of the inherently interpretable models.
- **Counterfactuals** refers to the hypothetical instances that steer the prediction of the black box towards the desired class of interest
- **Counterfactual explanations** refer to the family of explanatory methods that aim to generate hypothetical counterfactuals that alter the prediction to a desired class.
- **Deliberative explanations** aim to extract input features that help justify a given prediction.
- **Visual Explanations** bring out the working mechanism of the black box through visual cues in a human-understandable format, while **Textual** explanations leverage natural language phrases to bring out the working mechanism of the classifier.
- **Concepts** refer to an abstract vector representation that can be mapped to interpretable input regions.
- **Relevance** refers to an estimate of the importance of a concept towards predicting a given class.

3. Survey Methodology

The methodology employed to conduct the survey was systematic and comprehensive, ensuring a thorough review of the literature on explainable image classification. The survey encompassed various search engines, digital libraries, and reputable conferences in the field of AI and Computer Vision. Search engines such as Google Scholar and digital libraries such as IEEE Xplore, ACM, Springer, MDPI, and Elsevier were extensively searched to gather relevant literature. Additionally, conference proceedings from NeurIPS, CVPR, AAAI, ICML, and other relevant conferences were reviewed to identify suitable research contributions.

Carefully selected keywords and their combinations were used to construct effective search queries. These queries underwent an iterative refinement process, consider-

ing variations in terminology and synonyms to ensure a comprehensive search strategy. General and specific keywords related to explainable AI, interpretability, interpretable ML, and explainable image classification, among others, were utilized.

The literature search and review process focused on publications from 2015 to the present, allowing for the inclusion of recent and up-to-date research. To enhance coverage, relevant articles spawned by the initial search results were accessed and added to the review list.

During the review process, the retrieved papers' titles, abstracts, and keywords were assessed for their relevance to the survey's objectives. Selected papers were then thoroughly read to extract valuable insights, methodologies, and findings. This methodology ensured a systematic approach and the inclusion of high-quality papers that contributed significantly to the field of explainable image classification. The selection of papers was based on their relevance to the survey objectives and the quality of their contributions. By employing this methodology, the survey aimed to provide a comprehensive overview of the literature, presenting valuable insights into the field of explainable image classification.

4. Trajectory Traversed by Object Recognition Models

Object recognition is the task of identifying objects present in an image, for instance, a computer, an animal, or a bird. This is a task that a human can easily accomplish. However, it is challenging to automatically recognize an object through computers. Computer Vision techniques realize the object classification task as choosing the category of the object contained in an image from a given set of object categories [24]. A typical image classification model has two major steps: feature extraction and classification. Feature extraction is the process of extracting relevant attributes called features from the image that contain traces enabling identification of the object class. The classifier combines the extracted features to predict the object class.

Traditional Computer Vision techniques focused on developing hand-crafted features such as the Scale Invariant Feature Transform [1], Histogram of Oriented Gradients [2], etc., to extract features whose aggregation would yield the prediction. However, using hand-crafted features yielded limited results with the growing complexity of data [25]. The advent of deep Convolutional Neural Networks (CNN), since the AlexNet [26], brought in a paradigm shift in the community's notion of feature extraction, as these models were able to extract the discriminative features automatically from the data. With time, deeper architectures with more hidden layers [3,4] demonstrated higher performances.

However, the increased accuracy with the increased number of parameters comes at the cost of decreased transparency. Traditional machine learning models, for instance, a decision tree, are interpretable by nature as their working mechanism can be summarized by means of if–then–else rules. Summing up the working mechanism of a CNN in a similar manner is not trivial. It is well recognized that the initial layers closer to the input detect rudimentary features such as edges or contours while the latter layers closer to the output layer process complex image components such as object parts [27,28]. Gaps persist in the community's understanding of how an image is decomposed, and the extracted features are aggregated to deduce an instance's class.

This opacity of CNNs can limit their widespread use in many safety-critical paradigms such as medicine [29], judiciary [30], where transparency regarding the working mechanism of the deployed model is sought. Hence, it becomes important to develop mechanisms to explain the working of these deep black boxes. Moreover, The Right to Explanation Act by the European Union (EU) [31] has made it mandatory for businesses leveraging Artificial Intelligence (AI) in their work processing pipeline to explain why certain decisions made by the AI model was carried out. This has led to a spurt in the development of Explainable AI.

5. Need for Explaining the CNNs

Firstly, it is important to clarify what it means to explain a CNN. A CNN takes an image as an input, extracts features using the convolutional and pooling layers, and

combines them using the fully connected layers to classify the instance into one of the several categories. Considering the example of the previously stated classical machine learning model, namely the decision tree, translating the path traversed from the root to the leaf node into if–then–else rules yields the features that led to the prediction. Similarly, an explanation that unravels a CNN's working mechanism to classify a given image is expected to highlight the significant image features the CNN uses to arrive at the prediction. We illustrate the benefits of an explanation through a motivating example.

Consider a CNN model that recognizes birds in images and classifies them as either an *albatross*, *hummingbird*, or *pelican*. If a test image of an *albatross* is misclassified as a *pelican*, one may be curious to know why the instance is misclassified. One may turn to XAI algorithms to analyze the feature in the *albatross* image that is misjudged as that of a *pelican*. A good explanation that can justify the misclassification may be that in the given image, the beak of the *albatross* looks similar to that of the *pelican*, resulting in the *albatross* instance being misclassified as a *pelican*.

While misclassification is one scenario where understanding the CNN's working mechanism is sought, explanations may also be needed for correct classifications. Such explanations can reveal what features the model relies on to make predictions and enlighten the correctness of the model's working. Moreover, explanations can highlight spurious biases [32–34] that the model relies on, helping to assess the model's deployability in real-world scenarios. For example, in the bird classification task, a model may rely on the presence of a *water* background to distinguish *pelicans* from other birds. This correlation may enhance accuracy in the given dataset, but the model cannot be deployed in a real-world bird recognition task, where the background need not always contain water, as the model had encountered during the training time.

To improve the user's trust in the deep model and to ensure their ethical deployment for real-world tasks, the XAI research community aims to develop methods that explain the internal working mechanism of the learned CNN models, which are essentially black boxes.

Explainable AI refers to the set of techniques and methodologies used to make AI systems more transparent, interpretable, and understandable to humans [35]. These techniques can be used to help humans understand how an AI system makes a decision, what factors are considered, and how confident the system is in its decision. In traditional models such as linear regressors, the coefficients reveal the importance the model gives to a certain input feature. Similarly, the working logic encoded in a decision tree can be translated into a set of if–then–else rules. XAI algorithms are developed to unravel the working mechanisms of complex, accurate models such as random forests, neural networks, etc., whose working is difficult to summarize in a similar human-interpretable manner.

6. A Brief Overview of the Previous Attempts in Explainable AI

There have been several notable attempts to explain the workings of various types of black box models employed across different data modalities, including tabular data [36], text [37], and images [5]. Explainable AI (XAI) methods that unveil the internal mechanisms of these black box models offer several benefits for both users and developers of AI systems.

For users, XAI can foster trust in AI systems by providing a clear understanding of their functioning and the rationale behind specific decisions. This aspect is particularly critical in high-stakes domains such as healthcare and finance, where the decisions made by AI systems can have significant consequences. Chaddad et al. [6] reviewed the use of XAI in healthcare and discussed the threats with respect to privacy, confidentiality, and bias that need careful examination in a medical AI system. They analyze the satisfaction rate recorded by medical professionals for posthoc explanations and re-ascertain the need for developing models that can explain themselves [38]. Salahuddin et al. [39] review the evaluation metrics tailored to validate a proposed transparent medical XAI system. The authors identify a void in the approaches to process the multimodal data widely prevalent in healthcare applications, in an interpretable manner and call for collaboration with the sub-community in XAI research that deals

with explaining Graph Neural Networks, capable of processing multimodal data [40]. Weber et al. [7] review the application of XAI methods in finance systems and identify the skewedness in the distribution of leveragement of different XAI approaches. They call for applying techniques to extract causal explanations in a privacy-preserving manner, a crucial requirement that must be satisfied in finance. On the other hand, Owens et al. [41] focus exclusively on the sub-domain of finance dealing with insurance and advocate the need for attention of XAI researchers to tailor XAI techniques to cater to the needs of the insurance sub-domain. In addition to fulfilling the critical desiderata of financial systems, the approaches have to be vigilant about the imbalanced data [42] concerning fraudulent practices, which need to be addressed promptly when encountered. For developers, XAI techniques can aid in system debugging and improvement by revealing insights into decision-making processes and identifying areas for enhancement [43,44]. Furthermore, Clement et al. [8] present a comprehensive survey that positions various XAI methods with respect to software development principles. Researchers interested in applying XAI techniques to these application domains are encouraged to refer to these surveys [6,7,39,41,45], which provide detailed reviews of methods tailored to specific applications.

Our survey aims to shed light on the explanations of CNNs [46], as they represent state-of-the-art deep architectures for image classification. This area demands a thorough survey due to the rapid emergence of numerous techniques. The emphasis on explaining Convolutional Neural Networks (CNNs) and thus focussing the scope of the survey on this major sub-direction within XAI research sets our survey apart. While existing surveys have primarily covered a broader range of black box models and data modalities, our survey delves deeply into the explanations of CNNs, addressing their unique challenges and advancements. By focusing on this specific area, we offer readers a comprehensive tour of seminal contributions, identification of existing shortcomings, and potential future directions to overcome current limitations. Moreover, existing surveys have predominantly focused on classifiers trained using the traditional supervised learning paradigm, where extensive labeled data are available to feed data-hungry deep models. In contrast, our survey also explores the relatively under-explored terrain of explaining classifiers trained with sparse data, leveraging allied learning paradigms such as transfer learning, few-shot learning, incremental learning, and others.

The analysis of XAI approaches organically leads to the development of evaluation metrics for these approaches, which is currently an active research area. Our survey acknowledges this ongoing work and envisions the potential for devising metrics that cater to the needs of explainable allied learning paradigms. Furthermore, we suggest cross-pollinating ideas from the Neural Architectural Search [47] community, which aims to identify the best architecture to model a given data distribution. This collaboration can result in optimized, explainable-by-design architectures that achieve both accuracy and interpretability objectives.

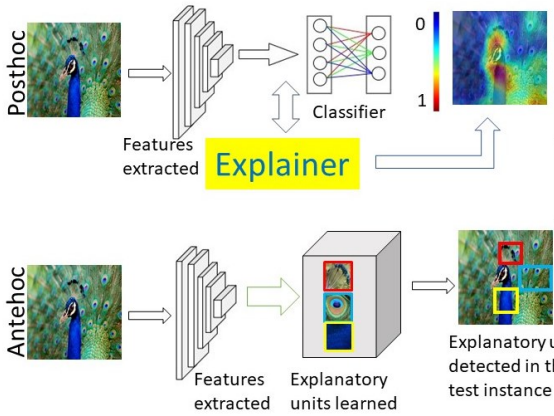
Overall, our survey fills a crucial gap in the literature by providing an in-depth analysis of explaining Convolutional Neural Networks (CNNs), examining the unique challenges they present, and exploring the potential of allied learning paradigms, evaluation metrics, and mechanisms to automatically devise architectures that minimize the accuracy–interpretability tradeoff. We aim to provide researchers and practitioners with valuable insights and perspectives, ultimately advancing the field of explainable AI in image classification.

7. Taxonomy of XAI Methods

This section presents a condensed review of the state-of-the-art contributions in the field of Explainable AI (XAI). It provides an overview of the underlying principles, limitations, and improvements made to these seminal contributions. The XAI methods are categorized into two broad families based on the stage at which the explanations are incorporated: posthoc and antehoc.

Posthoc techniques generate explanations without modifying the underlying Convolutional Neural Network (CNN) architecture. The method may [48,49] or may not assume access [32,50] to the intermediate layers of the CNN. Since the black box, i.e., the CNN, remains undisturbed, there is no need to retrain the model to incorporate explainability. This makes posthoc methods preferred for generating explanations from an already deployed model. However, ensuring the faithfulness of the generated explanation to the working mechanism of the CNN is a key challenge when employing posthoc methods to explain a CNN. Specifically, ensuring the consistency between the explanation's ranking of the features based on their significance to the prediction and the ranking by the black box CNN being explained is a non-trivial requirement that the posthoc explanation method must fulfill.

On the other hand, Antehoc methods incorporate the aspect of explainability and maximize classification accuracy within the learning pipeline. They achieve this by either modifying existing black-box architectures [51] or proposing novel architectures where explainable artifacts are detected. These detections then guide the prediction [52,53]. Since explainability is integrated into the training pipeline, the generated explanations are faithful to the CNN. In other words, the explanations reveal the true underlying mechanism used by the CNN to arrive at its prediction. However, retraining the CNN or modifying its architecture to extract faithful explanations comes at the cost of reduced accuracy. It is challenging to achieve the classification accuracy of an unrestricted CNN in the modified version with explanatory bottlenecks incorporated by design. Thus, as illustrated in Figure 1, Explainable AI methods entail an accuracy–interpretability tradeoff.



Aspect	Posthoc	Antehoc
Modifications to black-box architecture	Not Necessary	Necessary
Training/ Retraining	Not Necessary	Necessary
Faithfulness	Challenging	Guaranteed
Retaining black-box accuracy	Guaranteed	Challenging

Figure 1. Comparison of antehoc and posthoc explainability methods.

Based on the scope of their explanations, XAI methods can be categorized as either local or global, depending on whether the generated explanation unveils the entire working mechanism of the model or focuses on explaining the model's behavior within a limited neighborhood surrounding an instance of interest. Global methods [54–58] explain the CNN across the complete instance space and can be used to construct interpretable proxies that mimic the workings of the CNN. Such proxies are valuable in safety-critical applications where explainability is crucial. However, generating a global explanation that faithfully captures the non-linear manifolds learned by the CNN can be challenging. To address this challenge, local explanations [32,50,59–61] leverage the local linearity of the data manifold to explain the CNN within a specific vicinity around an instance of interest. An approximate global explanation can be obtained by aggregating local explanations over a set of instances.

Methods in the field of XAI can be categorized into model-specific and model-agnostic, based on the assumptions they make regarding the type of black box they query to gen-

erate explanations. Model-specific methods assume architectural constraints to generate explanations specifically tailored to the underlying model. In contrast, model-agnostic methods generate explanations by analyzing the input and output interactions without making any assumptions about the specific black box they aim to explain. Model-agnostic explanations are particularly useful when the black box model is not publicly available and can only be accessed through an API that allows input provision and output retrieval. However, model-agnostic methods rely on certain underlying principles, such as the existence of interpretable features whose aggregation reveals the workings of the black box. These principles may not always hold true, highlighting the importance of using model-specific methods whenever complete access to the black box being explained is available.

Explanations can also be categorized based on the class label for which the explanation is queried. Deliberative explanations provide justifications for the predictions made by the black box model. They help identify any biases present in the learned model and offer insights into the decision-making process. On the other hand, counterfactual explanations support the ability to edit a given instance in order to change the predicted label. These explanations are particularly useful in the context of Machine Teaching, where generating counterfactual instances helps humans better understand the distinctions between different classes. By finding the closest instance belonging to an alternate class of interest, counterfactual explanations provide valuable insights. A summary of the categorization of explainable AI approaches, and the seminal contributions falling under each category can be found in Table 1. Additionally, a visual representation of the taxonomy is provided in Figure 2.

Table 1. Overall summary of various XAI methods as per the proposed taxonomy has been tabulated, and the different situations where these categories of explanations are most suitable has been discussed.

Categorization Basis	Categories	Suitability	References
Incorporation Stage	Posthoc	Suitable to explain an already deployed model	[11–14,23,32,33,48–50,55–93]
	Antehoc	Suitable when an application specifies the need to build models that have interpretability built into its design	[15–17,22,38,51–54,94–125]
Explanation Scope	Local	Useful in privacy-preserving applications as only information around the vicinity of the instance is explored	[32,33,48–51,59–61,65–70,73,75–81,83,84,96,107–111,126]
	Global	Useful to explain the complete working logic of the AI system to business stakeholders who decide to adopt the AI system into the business pipeline	[20,52–58,71,72,74,81,82,85,87,89,91,94,97–99,106,112,118–125,127]
Aim of the Method	Deliberative	Justify the given prediction	[32,50,53,59–61,65,71–73,76–78,90,91,96,109,111,128]
	Counterfactual	Useful to create close looking hypothetical Machine Teaching examples so that learners understand looking at minute discriminant features	[17,85–89,129,130]
Explanation Modality	Visual	Quickly summarize the CNNs' working using visual cues	[33,48,49,51–53,63,64,66–70,77–80,84,94,95,99–106]
	Textual	Useful to explain users with special needs through leveraging text modality	[96,107–112,131]
Training Distribution	In-domain	Explain CNNs trained on a single large dataset	[32,50,51,59–61,65,71–73,76–78,91,96,107–112]
	Cross-Domain	Explain CNNs generalizable to multiple datasets	[21–23,116,117,132]

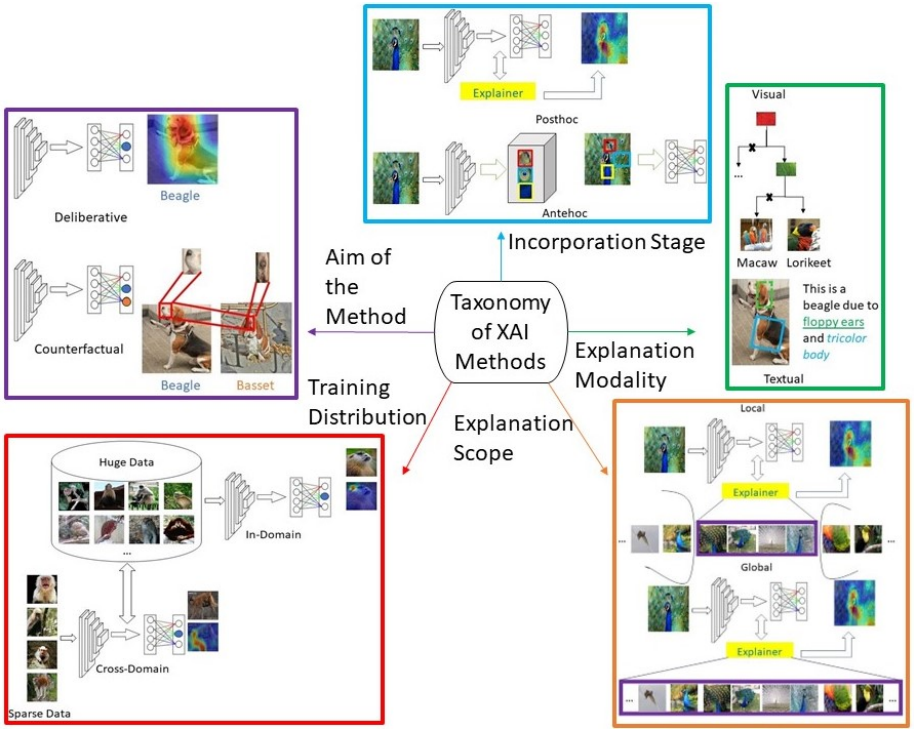


Figure 2. [Best viewed in color] Proposed Taxonomy of XAI Methods. The categorization as antehoc and posthoc methods based on the stage of incorporating the explanation is bounded inside a blue rectangle. The green rectangle bounds the categorization as visual and textual explanations based on the explanation modality. The purple box bounds the categorization based on the aim of the XAI method as to deliberative explanations that justify a prediction and counterfactual explanations that produce hypothetical instances that flips the prediction. The categorization based on the exposed training distribution is bounded by the red bounding box, where in-domain classifiers are trained on a huge data pool belonging to a single distribution, while cross-domain classifiers are exposed to different distributions from which a varied number of samples are presented to it. The proposed taxonomy considers this perspective to drive home the attention of the XAI community to propose techniques to explain the relatively underexplored black boxes, namely the cross-domain classifiers. The orange box bounds the categorization based on explanation scope where Global methods are applicable in the entire instance space, while Local methods are applicable only in a small neighborhood.

It is to be noted that an XAI method can fall under multiple categories based on the aspect used for categorization. Our forthcoming discussions are structured by considering the stage of incorporation of explainability as the basis for sub-categorization. We would like to note that this is just a choice we have made for coherent discussion of sub-categories. A reader may view the methods categorized based on any of the aspects and the discussions consider highlighting the differences/sub-categorizations based on other aspects whenever feasible.

We also acknowledge the other taxonomies prevalent in the literature. For instance, Schwalbe & Finzel [10] propose a multi-level taxonomy covering XAI methods for different black boxes and data modalities. Cabitza et al. [9] propose a taxonomy based on the explanandum–explanatory relationship–explanans triad and justify their categorizations to be in accordance with the existing taxonomies. One may see that the categorization of Global and Local in terms of the explanandum in these taxonomies has been considered in our taxonomy as the categorization based on the Explanation Scope. Similarly, the categorization based on the explanans (explainer) has been subcategorized in our proposed taxonomy under broad categories of Posthoc and Antehoc methods based on the stage at which explainability is incorporated into the black box. The sub-categories discussed

as pertaining to the explanatory relationship, namely epistemological methods, which are derived from the principles of knowledge theory (neuro-symbolic methods [125]) and cognitive methods drawing inspiration from historical ideas of explanations in the cognitive science sub-domain have been discussed in our survey article, driving home the supremacy of a specific category of explanations (concept-based explanations) in representing the working mechanism of a model such that it is easier for humans to diagnose the black box [71]. In a nutshell, the proposed taxonomy covers the approaches to explaining image classifiers covered in the prevalent taxonomies for general XAI approaches. Given that the focus is specifically on image classification, detailed analyses of the strengths and weaknesses of the methods have been carried out, inferring the suitability of the methods to different situations.

7.1. Posthoc Methods

Posthoc XAI methods refer to techniques and methodologies used to explain the behavior of an AI system after it has been trained to make a decision. These methods do not necessarily modify the AI system itself but rather analyze the output generated by the system to provide explanations for the decision-making process. A major advantage of using these methods is that they do not require any architectural modification or black-box retraining. They probe the trained black-box model to understand its working. The posthoc methods can be subcategorized under four major heads: Saliency Map, Model-agnostic, Counterfactual, and Concept-based approaches, as discussed in the following subsections. Table 2 provides a quick summary of the key strengths, weaknesses, and complexity involved in training the explainer for various sub-categories.

Table 2. A summary of contributions explaining the CNNs in a posthoc manner is presented with a discussion on the strengths, weaknesses, and challenges in training the explainer corresponding to the individual categories.

Sub-Category	Strengths	Weaknesses	Training Complexity	References
CAMs	These mechanisms can be used as a Plug & Play module to an already deployed model due to simpler definition of an explanation being a linear combination of intermediate activation maps	The heatmaps exhibited are almost always coarse (Figure 3), rendering them unable to provide finer explanations	Low	[33,48,49,66–70,83,84]
Model-agnostic	These explanations are interpretable when applied to images since the images are segmented using a human-friendly mechanism	It is not necessary that the CNN also employs a similar segmentation mechanism to process images	Moderate	[32,50,55–61,65,133]
Counterfactual	These explanations are pedagogical in nature since hypothetical counterfactual instances which are closer to the data in hand govern the explanation so that the human learners look at finer discriminative features to better distinguish related classes	Realistic image generation is challenging	High	[85–89,134]
Concept-based	The concepts extracted are based on examples provided by humans and hence interpretable	To obtain faithful explanations, the examples provided have to be sampled from the same distribution on which the CNN is modelled	Moderate	[63,64,71–74,90,91,97,135]

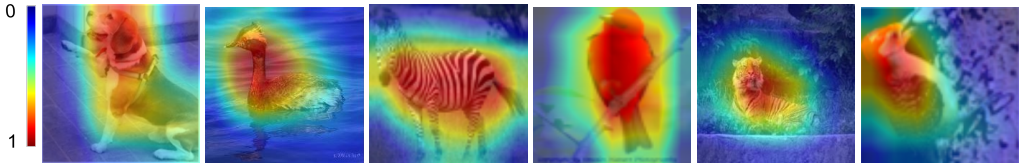


Figure 3. [Best viewed in color] CNN explanation in the form of a saliency map localizing the image region contributing to the prediction.

7.1.1. Class Activation Maps

The most common outlook for explaining a CNN is identifying the key image regions contributing to the predictions [33,48,49,66,67]. These key regions are often displayed using a saliency map, where the image regions are color-coded based on their importance. A few examples of these saliency maps are shown in Figure 3.

Class Activation Maps assume that the region salient toward the prediction of a class can be obtained from a weighted combination of the activation maps from the convolutional layer filters. Inspired by the observation that the latter layers encode complex parts [28], most saliency estimation approaches extract activation maps from the last convolutional layer closest to the output. Let the convolutional layer of interest have n filters. Let A_i be the activation map from the i^{th} filter. The explanation algorithms assess the salient regions that the CNN focuses on by means of a saliency map S that can be expressed as a weighted combination of the activation maps from each of the n filters, i.e., $S = \sum_{i=1}^{n} w_i A_i$. This formulation stems from the understanding that the features extracted are combined to arrive at the prediction. The low dimensional saliency map obtained through the weighted combination of the activation maps from the individual filters is then upsampled to the full image size to generate an explanation showing the image region that the CNN focuses on to arrive at the prediction. Various mechanisms have been proposed to estimate the weights $\{w_i\}_{i=1}^{n}$ that combine the activation maps from the filters. These approaches can be bifurcated based on leveragement of gradients, as will be discussed below.

Gradients capture the direction along which the value of a function increases. Thus gradients propagated back to the convolutional layers from the output layer carry a signal indicating the features whose presence steers the model towards making a desired prediction. This signal is leveraged to estimate the weights and combine the activation maps using gradient-based saliency approaches. Grad-CAM (Gradient-weighted Class Activation Mapping) [33] generates a saliency map highlighting the regions of the input image that were most relevant for the neural network's prediction. It works by computing gradients of the output prediction with respect to the activations of the final convolutional layer. The activation maps are combined based on the weights obtained by averaging the gradients with respect to the corresponding filter over all the spatial locations. No additional modifications to the neural network architecture are needed to generate explanations and thus can be leveraged to explain any CNN. The following year Chattopadhyay et al. [48] observed that having the averaged gradients as weights to combine the activation maps does not localize well in images where multiple instances of the same class are present. They proposed applying different weights to gradients observed at each spatial location to uncover all regions steering the prediction; thereby, the observed limitation of Grad-CAM [33] in localizing more than one instance of the class can be overcome. The weights to these spatial locations were deduced to be obtained from higher-order derivatives whose computation could be demanding in complex architectures. Integrated gradients [75] consider a reference input and traverses the instance space across the path from a reference input to reach the given instance. The attributions with respect to the intermediate instances along the path are integrated to obtain a robust saliency map depicting the salient pixels in the given instance. Excitation backpropagation [76] utilizes a probabilistic winner take all strategy where the attribution being propagated to a downstream neuron is probabilistically determined. Guided backpropagation [77] proposes propagating attribution only to those neurons which were active during the forward pass,

thereby generating finer pixel-level saliency maps compared to the vanilla backpropagation [78] that propagated gradients as attribution irrespective of the contribution of the neuron until arriving at the output layer.

Various quantitative metrics [48] have been proposed to assess the faithfulness of the generated explanations. The proposed metrics are based on the requirement that removing a salient region must lower the model's prediction confidence while its presence has to amplify the confidence. Viewed differently, these metrics observe the effect of perturbing the regions deemed salient on the model's prediction probability. The proposal of these metrics is inspired by the first principles of generating explanations that a region whose perturbation impacts the prediction is salient. Instead of going through the voluminous possibilities of all image perturbations, Chattopadhyay et al. [48] propose to use derivatives to localize salient regions and verify if the regions localized to be salient are truly salient by observing the effect of perturbing those regions on the CNN's prediction probability.

Wang et al. [136] empirically showed that the gradient-based saliency maps obtained do not vary with respect to the queried class, thereby questioning the faithfulness of these explanations. Adebayo et al. [137] proposed litmus tests that a posthoc XAI method has to pass towards its proof of faithfulness to the underlying black box model. There are two basic tests that an explanation method has to pass, namely the parameter randomization, which observes the change in explanations when the model weights are randomized, and label randomization, which observes the change in explanations when the labels are randomized, and the CNN model is retrained to model the altered distribution. It has been observed that most of the gradient-based techniques fail to satisfy these proposed litmus tests. The theoretical analysis by Sixt et al. [138] attributes the invariance in the saliency map for the model parameters and query labels to the restriction of the explanatory model to the positive subspace of the activations.

Following the issues found with using gradients to determine saliency, the XAI community has proposed other methods to generate saliency maps. There have been attempts [49,66,67] to incorporate the effect of perturbation at the level of filters to assess the importance of the activation maps, which will, in turn, be the weights w_i combining the activation maps A_i. It is easier to manage the possible perturbations [79,80] with n filters of the convolutional layer of interest than that of the input image of much higher dimensions. Wang et al. [66] associate the importance weight w_i to combine the activation map A_i based on its effect on obtaining the prediction for the desired class. In other words, the prediction probability obtained when the activation map A_i is present, and the other activation maps are nullified, is the weight w_i that combines the activation map A_i. Desai & Ramaswamy [49] take a complementary route by considering the drop in prediction probability when the activation map of interest A_i is ablated while forward propagating other activation maps without any modification to determine the weight w_i. A limitation of these approaches is the need for multiple forward propagations to obtain a single saliency map. In contrast, the previously proposed gradient-based approaches can generate the saliency map in a single backward pass. To mitigate this issue, Salama et al. [67] propose clustering similar activation maps and obtaining the ablation score for a cluster from which the weights w_i for each activation map A_i can be recursively determined. There have been attempts to propagate a special signal called relevance [81,82] from the output layer back to the input to determine the pixels salient to the prediction. However, the fact that these pixel-level saliency maps are not class-discriminative has led to the cross-pollination of ideas from these techniques to estimate the combination weights w_i of CAM [68–70,83]. Layerwise Relevance Propagation [81] propagates the neural network output back through the different layers to assign relevance scores to these input features. The forward pass propagates the activation from the input layer and reaches the output layer. Relevance propagation starts in the opposite direction from the output layer, and gradually the relevance signal reaches the individual input pixels. The relevance propagation is based on the idea of conservation, i.e., the relevance signal from a neuron is distributed across all neurons that have contributed to it during the forward pass proportional to their contribution.

Lee et al. [68] apply the idea of relevance propagation [81] to estimate the relevance of the filters, which can act as the weights w_i to combine the activation maps A_i. Deep-LIFT [82] is a modified form of relevance propagation where differences between activations with respect to a reference input are propagated to obtain the relevance of the different input features. The input with zero in all its dimensions is mostly taken as the reference input. Extending the idea from Deep-LIFT [82], Jung & Oh [69] estimate the filter weights w_i to combine the activation maps A_i through the differences of the combination weights obtained with respect to a reference input. Sattarzadeh et al. [70] extend the idea of integrated gradients [75] to integrate the attribution maps obtained across the path from the reference input to the given input. Wang et al. [83] generate image patches [139] and use an attention mechanism to estimate the salient regions in a given image. However, a major limitation of these saliency map approaches is that they almost always highlight the region containing the entire object to be salient [38,63,64]. While these explanations can ascertain whether the model looks at the object to arrive at its prediction or relies on any non-object spurious correlations [32,34], finer explanations depicting the contributions of image primitives such as colors, textures, and parts cannot be obtained from the Class Activation Maps.

7.1.2. Model-Agnostic Explanations

Model-agnostic methods refer to the family of XAI methods, which explain the working of a black box model by observing input-output interactions. They can be applied to any machine learning model, regardless of its type or architecture, and can work to explain data of any modality such as text, images, tabular data, etc. The scope of these methods' explanations can be local to a given instance or can globally explain the overall working of the black box. These methods aim to construct an inherently interpretable pseudo classifier that approximates the working mechanism of the black box classifier to be explained either locally around a small neighborhood of an instance for which the explanation is sought or globally, spanning the complete instance space of the classifier.

Local Interpretable Model-agnostic Explanations (LIME) [32] generates a simpler, more interpretable model, for instance, a linear regressor or a decision tree whose complexity is optimized such that the determined approximator mimics the behavior of the original model in the local vicinity of the input space around the instance to be explained. This simpler model can then be used to provide local explanations for individual predictions. It can be observed that different explanations can be generated for the same instance depending on the sampled neighbors based on which the local neighborhood is estimated. Zafar & Khan [59] propose a deterministic approach to sampling neighbors utilizing agglomerative hierarchical clustering and sampling k-nearest neighbors, using which an interpretable approximator is constructed. Collaris et al. [65] hint at the possibility of sampling fewer neighbors when sampling is performed independent of the queried instance to be explained and propose to sample from a hypersphere around the instance to obtain a robust local explanation. Anchors [50] generate explanations for individual predictions using if–then rules constructed in a bottom-up fashion such that the rule precisely covers the local neighbors of the instance to be explained. MAIRE [60] extends Anchors [50] to handle continuous-valued attributes by learning to construct an optimal orthotope automatically, unlike the prior approach [61] that requires the range of values to construct the orthotope. Local explanation methods aim to extract explanations that are faithful in a local neighborhood by means of special metrics such as coverage which estimates the fraction of instances that lie within the explainer's vicinity, and precision which denotes the fraction of covered instances whose prediction by the explainer matches with the prediction by the black box CNN. Constructing a MAIRE [60] explainer maximizes the coverage, ensuring faithfulness to the underlying black box by satisfying a precision level set by the user. Though these methods offer local explanations, a global understanding of the model can only be obtained by aggregating the local explanations over a set of instances.

There have also been attempts to build an explainer that approximates the global behavior of the model as a whole. SHAP [55] uses the principles from game theory (Shapley

values) to assign an importance score to each input feature, indicating how much each feature contributes to the system's output. These importance scores can be used to identify the most relevant features and understand their influence on the system's decisions. Computing Shapley values requires considering all possible subsets of the feature space and assessing each subset's perturbation effect on the output. This is computationally exhaustive due to the exponential time complexity. Many approaches have been proposed based on Shapley values approximated by considering only the perturbation of one feature at a time. Permutation feature importance [57] calculates the importance of each input feature by randomly permuting its values and measuring the decrease in the model's performance. Partial dependence plots [56] visualize the relationship between an input feature and the model's prediction while holding all other features constant. Despite approximations [58] to compute Shapley values efficiently, there has been a recent observation [140] highlighting their inadequacy in faithfully capturing the global behavior of the black box being explained. Huang & Marques-Silva [140] construct a boolean dataset where a set of features relevant to determine the output are known. A global explanation is ideal if it assigns zero importance to irrelevant features and non-zero importance to features that correlate with the output. It was observed that there existed features that were truly irrelevant to the prediction but had non-zero Shapley values. Also, when pairs of features were analyzed such that one was actually a relevant feature and the other was irrelevant to the prediction, Shapley values of the irrelevant features were higher compared to that of the relevant features. Sometimes Shapley values for truly relevant features turned out to be zero, contrary to the basic requirement that a global explanation must accurately capture the feature importances. Huang & Marques-Silva [140] conclude that the Shapley values are not always correlated with the actual relevance of features for the black box predictions.

Another important observation is that the model-agnostic methods are developed to generate explanations for any black box model, and hence no assumption regarding its architecture is made. The explanation is given in terms of input features that are significant towards the prediction. In images, the pixels constitute the input features. As pixel-level explanations are not easily interpretable for humans, a workaround suggested using a collection of spatially closer pixels called the superpixels. These superpixels serve as complex input features for the model-agnostic methods to generate explanations. An example of how such an explanation would appear can be seen in Figure 4, where the segments covering the ears, muzzle, legs, and black body are highlighted to be significant to the prediction. For this, the existing model agnostic approaches [50,60] use different predefined image segmentation algorithms [141,142] to obtain segments constituting the superpixels on which model agnostic explanations are sought. On the surface, it may seem that this workaround achieves satisfactory human interpretability when model-agnostic explanations are sought on images. However, it is to be noted that CNN need not process the image by segmenting it similar to that of the model-agnostic explainer [84]. This refutes the preliminary necessity of the proposed approximator, aka the explainer, to be faithful to the underlying black box, aka, the CNN being explained.

7.1.3. Counterfactual Explanations

The viewpoints on explaining a CNN discussed so far are deliberative, meaning they intend to explain a given prediction. They are mostly used to justify a classifier's predictions and diagnose any spurious correlations it relies on. On the other hand, the misclassification scenario, as discussed previously, involves comparing pairs of images of different categories to justify/diagnose the misclassification. To address this concern, the counterfactual perspective on explanations was introduced.

Counterfactual explanations involve generating alternative scenarios to explain the behavior of an AI system. For example, if an AI system for processing loan applications denies a loan application, a counterfactual explanation might involve generating a set of hypothetical inputs that would have resulted in an approved application [143]. These counterfactual explanations can help users understand the decision-making process

and identify potential biases or errors in the system. They differ from deliberative explanations that aim to justify why a certain prediction was made. Counterfactual explanations go a step further to analyze the changes to the input to get another desired prediction. This explanation can be applied to analyze a classifier that works with any data modality, be it tabular, text, or image. The methods try to perform minimal edits to the given query instance such that the prediction is steered towards an alternate desired class. This can be thought of as perturbations intending to flip the prediction. In the case of tabular data, where the efficacy of the counterfactual approaches has been mostly demonstrated [134], the perturbations are manageable as the range of values the tabular features can take is known, and the instance can be perturbed to generate another realistic instance that lies within the manifold on which the classifier was trained. Determining this realistic manifold is non-trivial in the case of images whose constituents, aka, the pixels, can theoretically assume any real value. The objective of explaining using a perturbed instance is common in adversarial learning, except that it does not have a target class towards which the prediction has to be steered. The objective in generating an adversarial example is that prediction on the generated instance must not be the same as that of the unperturbed instance. Caution has to be observed as a random perturbation can generate an adversarial example [144], which may flip a prediction towards the target class of interest but may not be an ideal candidate to extract counterfactual explanations as the instance may be an outlier with respect to the realistic training images' manifold, thereby questioning the faithfulness of the generated counterfactual explanation to the underlying model and data. To circumvent this challenge, the existing approaches [85,86] either maintain an image bank from which the closest counterfactual image is chosen, or a generative model [87] is used to sample the counterfactual neighbors of the query instance from the distribution on which the CNN is trained.

There have also been some deliberative explanation approaches that allow querying explanation with respect to another class of interest [33], harnessable to generate a counterfactual explanation for the alternate target class of interest. However, these approaches do not generate explanations that vary significantly with respect to the alternate queried class [38].

The preliminary approach to generating counterfactual explanations through realistic instances is by maintaining an image bank from which the closest counterfactual instance to a given test instance is chosen. Various approaches have considered different ways to estimate the closest instance. Wang & Vasconcelos[88] generate deliberate explanations for the given test instance and all instances in the counterfactual image bank and chooses the instance containing features supporting the counterfactual class and no information of the predicted class as the closest counterfactual instance. Goyal et al. [85] simulate permuting feature maps to obtain features closer to that of the counterfactual instances that steer prediction towards the desired class. A main limitation of these approaches is the necessity to skim through the image bank for every test instance to be explained. Additionally, the image bank must be sampled from the same distribution as the data on which the CNN is trained.

To maintain the distribution, an alternate set of approaches employed variants of Generative Adversarial Networks (GAN) [145] to learn the underlying distribution. Singla & Pollack [87] sample instances that vary the prediction probability to navigate through the manifold of the counterfactuals. Zhao [89] proposes using a Star-GAN [146] to generate robust counterfactuals faster. However, it is to be noted that the generative models employed to learn the underlying distribution are, again, black boxes whose working is unknown. This complicates the problem at hand as techniques to interpret GAN [147] need to be employed on top of the existing counterfactual explainers.

7.1.4. Concept-Based Explanations

Humans process images through the lens of concepts [148], which can be abstract textures, colors, parts, etc. For instance, a *zebra* can be thought of as a *horse* having alternate

black and white *stripes* throughout the body. Concept-based explanations have been proposed to align the explanation algorithms closer to human-like thinking [52,91,148–150], i.e., the explanations are generated in terms of abstract vector representations that can be mapped to such human-interpretable concepts. Typically, a set of examples where the concept is present (termed positive examples) and absent (termed negative examples) is provided, from which the abstract vector representations are learned. Koh et al. [97] proposed a family of classifiers called the concept bottleneck models, which forces the classification to be completed through the set of known concepts, which act as a bottleneck through which the processing pipeline has to pass. The basic idea behind the concept bottleneck models is to insert a bottleneck layer between the feature extractor and the classifier of the original model and then train the bottleneck layer to capture the most important concepts from the features of the input data. This approach allows for extracting the salient concepts from the original model, which can be used to create a more interpretable approximator. The training of the concept bottleneck models can be sequential, where the bottleneck layer that detects concepts enables the classifier to use the detected concepts to arrive at its prediction or joint where weighted optimization of the concept detection and classification objectives is carried out, or independent where the training of concept detectors and classifier occurs independently of each other using the available ground truth. At the test time, the model mimics the pipeline of a sequential model. While the model proposed by Koh et al. [97] may require retraining, Yuksekgonul et al. [74] suggest the usage of a dimensionality reducer as the bottleneck layer that can faithfully map the space of the CNN features to an interpretable low-dimensional concept space, keeping the CNN untouched. Kim et al. [71] leverage the given positive and negative examples to extract representations from the CNN layer of interest. The boundary that separates the positive examples containing a concept from the rest is learned using these representations. The vector in the direction of the positive examples and orthogonal to the learned decision boundary is chosen to be the representative vector denoting the concept. This is illustrated in Figure 4 by means of a red bounding box in the middle column, where the vector color-coded in blue, orthogonal to the linear decision boundary separating the white-colored instances from others is chosen to denote the CNN's representation of the concept white. Once the concept representation is extracted, its relevance is estimated by inducing perturbation of the concept captured by the directional derivatives. As directional derivatives approximate the inherent non-linearity in the CNN being explained, Pfau et al. [72] propose propagating the perturbed concept through the rest of the CNN and observing the impact of the perturbation on the probability as this could be a more faithful measure due to accounting of the non-linearity of the CNN. However, a key challenge associated with generating such concept-based explanations is the need for annotated examples denoting the presence and absence of concepts. Ramaswamy et al. [135] observed that the curated examples have to be sampled from the same distribution as that of the data on which the CNN is trained so that the extracted concept representations faithfully capture the internals learned by the CNN. Ghorbani et al. [73] propose to use segmentation to subdivide the images at different granularities and curate them to extract examples depicting the presence and absence of concepts automatically. This reintroduces the issue associated with model-agnostic approaches for explaining a CNN regarding the questionable guarantee of the CNN processing images in terms of segments [84], thereby raising a question on the faithfulness of the generated explanation. Arendsen et al. [90] propose leveraging natural language word vectors to learn additional concepts automatically. However, this approach leverages another black box whose working mechanism needs to be unearthed [151]. Yeh et al. [91] propose automatically extracting the complete set of concepts from the data, thereby preventing a possible loss of faithfulness due to leveraging concepts sampled from a different distribution [135]. Kumar et al. [63] extend the capability of this framework [91] to unravel the complete blueprint of a class by formulating the concepts to be clustered in a class-specific fashion [52]. However, while extracting the explanations, these frameworks use multilayer nonlinear networks, which are also black boxes whose working could not

be unraveled. Kamakshi et al. [64] propose demystifying the black boxes involved in the automatic concept extraction pipeline by proposing the use of interpretable autoencoders. The relevance estimation is tied to the concept extraction objective, so the extracted concepts highly steer the prediction towards its class. However, this framework does not scale with the number of classes into which the CNN categorizes an instance.

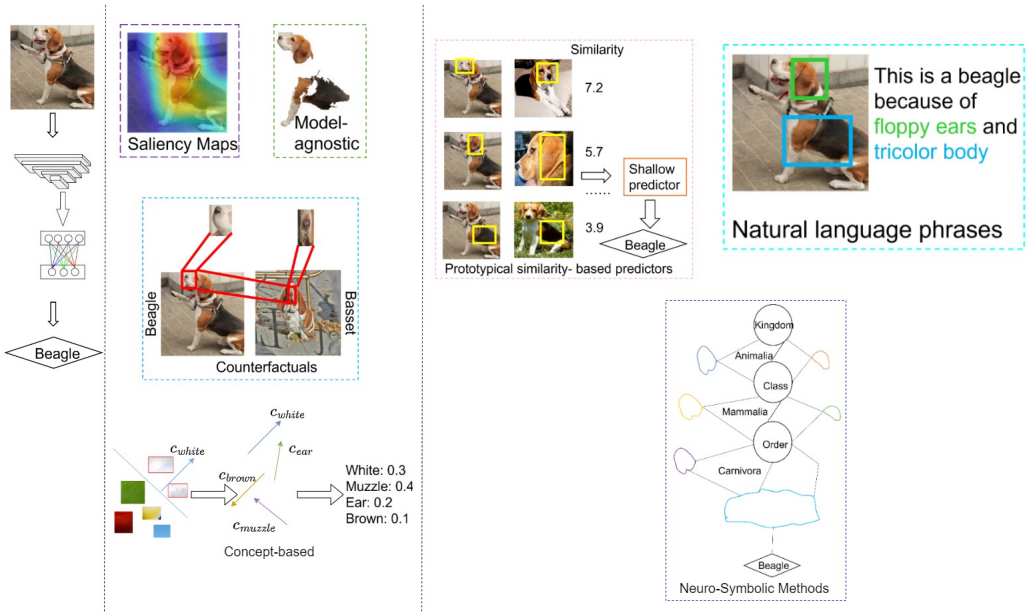


Figure 4. [Best viewed in color] Perspectives of explanations—an illustration. The **leftmost column** shows the general processing pipeline of a CNN, which processes the given input image and predicts its class. The **middle column** shows an illustration of different types of posthoc explanations, which leaves the CNN undisturbed. The explanation labeled Heatmap bounded by a purple dashed box localizes the regions the CNN focuses on to predict the given instance. The green dashed box shows model-agnostic explanations, which divide the image into predefined segments and highlight segments significant to the prediction. The cyan dashed box shows counterfactual explanations where a hypothetical scenario of a minimally edited image flips the prediction to an alternate class of interest. The red bounding box depicts concept-based explanations where from the given concept examples, the concept representation from the lens of the CNN is extracted, and the relevance is estimated by perturbing this representation. The **right column** shows the different types of antehoc explanations. The pink bounding box shows models modified to incorporate explainability by basing its predictions on the learned characteristic concepts detected in the given test image. The blue bounding box shows the explanations which justify a prediction by generating natural language phrases describing the detected characteristic concepts. The blue-dashed box encompasses the Neuro-Symbolic Explanations, which leverage a knowledge graph to map the representations learned by a CNN to existing knowledge.

7.2. Antehoc Explanations

Antehoc explainability, or explainability by design as it is popularly called, refers to the practice of building AI systems with explainability and interpretability in mind from the outset rather than as an afterthought. By incorporating explainability into the design process, these methods aim to create AI systems that are inherently transparent, interpretable, and trustworthy. Despite the advantages such as inherent interpretability and trustworthiness that antehoc explanations can offer, designing such models can be challenging and may require domain-specific knowledge and expertise. Additionally, some interpretability methods may come at the cost of model performance, limiting their usefulness in certain applications. To incorporate explainability, the architecture of existing CNN architectures may be modified [51], or novel components may be devised that are

interpretable by design. The explanation may be highlighting visual artifacts leading to the prediction or providing textual descriptions justifying the predictions. Alternately one may look up to existing knowledge bases to learn models whose working reflects the real-world application requirements. Table 3 briefly summarizes the key strengths, weaknesses, and complexity involved in training the antehoc models positioned under various sub-categories.

Table 3. A summary of contributions explaining the CNNs in an antehoc manner is presented with a discussion on the strengths and weaknesses and challenges in training the explainer corresponding to the individual categories.

Sub-Category	Strengths	Weaknesses	Training Complexity	References
Visual	The complete model pipeline from training till testing only relies on processing cues of a single modality	Possibility of misinterpretations due to subjectivity associated with the human analysis of visual cues	Moderate	[51–53,94,95,98–103,105,106]
Textual	Since visual cues are accompanied by natural language phrases, ambiguity is managed	Training language models, which are also black boxes and are introduced to make the CNNs transparent, is hard and time-consuming	High	[96,107–112,130,152,153]
Neuro-Symbolic	Since domain knowledge is referenced to make inferences; there is a high chance that the systems developed in this paradigm reflect the business requirements	It is difficult to devise such explainers when domain knowledge is unavailable	Moderate	[118–125]

7.2.1. Visual Explanations

Similar to how CNNs learned to extract features automatically from the data, the XAI community proposed enforcing the CNNs to learn interpretable concepts automatically from the data and use them to predict the object category [52,53,98,99]. An illustration can be seen in Figure 4, where the characteristic regions similar to that of the muzzle, ears, body, etc., of a *beagle*, guide the shallow predictor to predict the given test instance as a *beagle*. The discriminative interpretable concepts are learned automatically from the data, and the detection of these concepts in test instances guides the prediction using an inherently interpretable predictor such as a linear regressor or decision tree, allowing the complete reasoning pipeline of the modified CNN to be unearthed. In such models, the ability to provide explanations is incorporated in the training phase by design.

The earliest visual explanatory approaches used attention [101–104], which is a selective retainment of features to classify the test instance. Attention can be hard or soft in the sense that the selection of regions from the features may be deterministic or probabilistic. The regions attended would be turned in as an explanation. However, there have been observations [100,154] that an attention map visualized need not be an ideal explanation. Extending the analyses of Jain & Wallace [100] unearthing the limitations of attention-based approaches to explain natural language models, Akula & Zhu [154] conduct extensive human subject experiments, which reveal the usefulness of non-attention based approaches [71,85] compared to attention-based approaches [32,33,75,81] that explain an image classifier. The authors conduct quantitative tests, which reveal the supremacy of non-attention-based explanations in facilitating the user to think like the CNN as well as qualitative analyses where the users are asked to rate the quality of explanations on various parameters such as satisfaction, completeness, etc., as defined by Hoffman et al. [155] on a 10-point Likert scale, show that attention-based approaches are not suitable explanations.

Zhang et al. [105] propose to use mutual information to explicitly enforce the CNN filters to encode distinct parts so that the filters can be visualized to understand the impact of each part of the image. To facilitate the explanation generation, Zhou et al. [51] propose to change the architecture of the CNN to replace the series of fully connected layers incorporating non-linearity by means of a single linear layer, which accumulates the average pooled features to get a prediction. The weights that combine these average features are used to

combine the feature maps and visualize the salient regions contributing to the prediction. Li et al. [95] propose an autoencoder-based case-based reasoning [156] architecture that looks at characteristic prototypical examples learned from the distribution of instances whose proximity determines the class the test instance belongs to. Chen et al. [52] extend this architecture to automatically learn class-specific concepts called prototypes from data such that the learned concepts are class-discriminant and guide the interpretable classifier following it to do the prediction. Many extensions to this approach have been proposed. Hase et al. [106] propose to perform interpretable hierarchical classification by applying the explainable ProtoPNet [52] at every level of the hierarchy. Wang et al. [94] propose modeling instances as a member of class-specific orthogonal subspaces in the feature space. Hoffman et al. [157] and Huang et al. [158] analyze the prospective shortcomings of the ProtoPNet variants. The assumption of class discriminativeness need not be completely true, as concepts may be shared across classes. This idea of sharedness is exploited after training by encouraging sharing of connections to different classes [99]. Nauta et al. [53] construct a decision tree based on learned concepts that implement sharedness by design. However, using decision trees induces negative reasoning, which is overcome by Protopool [98], which enforces a Gumbel-Softmax distribution across prototypes to enforce sharedness closer to real-world sharedness.

As the ability to explain has been incorporated during the training phase, and the CNN is guided to use these explainable components to make predictions, the faithfulness of these explanations is guaranteed. In other words, whatever information the explanation reveals is truly what the model uses to arrive at the prediction. However, it needs to be retrained from scratch to incorporate such explainability into a CNN. This perspective can be leveraged when the model is yet to be deployed, and it is desirable to deploy a model that can explain itself but cannot be employed for an already deployed model.

7.2.2. Natural Language Explanations

Natural language explanation approaches [96,107–109,111,112] aim to generate textual descriptions that provide insight into how an image classifier makes its predictions. The key idea behind this approach is to leverage the vast amounts of linguistic knowledge that has been accumulated over centuries of language use and incorporate it into the model. This can help the model generate more coherent and natural-sounding explanations that humans can interpret.

This approach assumes the availability of natural language description for the classes under consideration and for individual instances from which the mapping between visual aspects and natural language phrases can be learned. A trained language model is incorporated to act as an explainer into the classification pipeline to construct a CNN that can justify it's working through natural language phrases. The visual features extracted from the feature extractor of the CNN are fed into the language model, which is trained to generate captions describing the image's content. A critic module then assesses the correctness of the generated caption to the image content. To train the critic module, the ground truth (image, caption) pairs are randomized, and the model is trained to provide a low score for a randomized instance where the image and caption do not agree and a high score on true instances where image and captions agree. The visual features and generated captions from the test image are fed to the critic module, which outputs a score denoting the goodness of the generated caption. To avoid multiple back-and-forth passes through the CNN and caption generator based on the feedback from the critic module, the top- k captions from the caption generator are considered, and the top-ranked caption from the critic is passed into a localization module to localize the corresponding image region contributing to the generation of the caption. This can be seen in Figure 4, where a given test instance classified as beagle is justified by localizing the characteristic floppy ears and tricolor body through similarly color-coded bounding boxes.

The approach is mostly used to justify the predictions made in related computer vision tasks, specifically vision-language tasks such as image captioning [159], visual question

answering [152,153], etc. where the task involves understanding both visual and linguistic aspects and can be preferably explained when the explanation mechanism also incorporates both vision and language features. Wickramanayake et al. [110] incorporate the textual embedding of the language model to guide the detection of characteristic concepts that drive predictions. This is an explainable-by-design model that leverages both the vision and language aspects.

However, designing effective natural language explanation approaches can be challenging and may require domain-specific knowledge and expertise. Additionally, the quality and effectiveness of the generated explanations can vary depending on the complexity and accuracy of the underlying image classifier and the quality of the available linguistic annotations. Another key challenge to be addressed when incorporating natural language explanations is that the language model which facilitates justifying the prediction is another black box whose working mechanism needs to be unearthed [151].

7.2.3. Neuro-Symbolic Methods

An alternative family of approaches, known as neuro-symbolic approaches [125], leverages existing knowledge bases or ontologies to acquire the necessary concepts for predicting a given instance, akin to utilizing domain knowledge curated by experts. This phenomenon was initiated with the proposal by Maillot & Thonnat [118], who advocated for collecting knowledge from domain experts and using it to train machine learning models that can base their predictions on the domain experts' knowledge. Marino et al. [122] propose a few-shot classification task by harnessing knowledge encoded in a graphical format. The classifier is trained to traverse different nodes of the knowledge graph and search for image features that match the descriptions associated with the investigated node. As the model navigates through the knowledge graph, the explanation is generated by identifying the localized image regions with the highest degree of match. Alirezaie et al. [123] aim to alleviate the problem of uninterpretable misclassifications by leveraging symbolic knowledge. Daniels et al. [124] propose the design of a bottleneck model [97], which compels the classifier to explore the available knowledge repository and base its predictions on the acquired knowledge. The authors hypothesize that such a design, which enforces the prediction to pass through the knowledge repository bottleneck, enhances the robustness of the learned model. Liao & Poggio [120] investigate the reasons why machine learning models lack the generalizability exhibited by humans. They hypothesize that models adopt a feature-oriented perspective, processing images as a sequence of tensor operations, which leads to variations in representation as objects manifest differently. In contrast, human knowledge processes images in terms of objects and concepts [148–150], exhibiting invariance to modifications in image manifestations. The authors propose mechanisms to transform the operations performed by feature-oriented models into an object-centric view, aiming to incorporate human-like processing. Ordonez et al. [119] propose a multimodal neuro-symbolic model that combines textual and visual knowledge to predict the entry-level categories to which an image belongs. For example, a neuro-symbolic classifier may have learned encyclopedic categories such as *Trachypithecus johnii* from the knowledge base, which refers to a species of monkey commonly known as a langur among wildlife enthusiasts. Ordonez et al. [119] address the challenge of mapping from encyclopedic categories to common categories, initially approaching it as an instance of hypernym search in a textual knowledge graph. Acknowledging the potential errors associated with visual cues in the knowledge base due to images of different categories appearing visually similar to humans, the authors propose a learning objective that combines cues from the visual and textual knowledge base to predict the appropriate entry-level category for an image. Icarte et al. [121] demonstrate the utility of a general-purpose ontology in retrieving realistic images that are closest to a given natural language query.

8. Causal Explanations

This section discusses the various attempts of the XAI community to generate causal explanations. We discuss this explanation category separately because there have been mechanisms proposed to extract causal explanations both during and after the deployment of the black box. Furthermore, the idea behind these explanations is to unravel the causal relationships modelled in the black box classifiers unlike traditional approaches that fall under one of the two categories based on the stage at which explainability has been incorporated which mostly unravelled the correlation between the different features.

In real-world data, the features are rarely independent, which can be observed by a corresponding change in another feature when a feature is perturbed. This relationship may be a mere correlation or causal, i.e., the features have a cause–effect relationship. For instance, if the sales of pens increase with an increase in the temperature of the city, this relationship is just a correlation, as there is no known relationship between a pen and temperature. However, an increase in sales of an umbrella with an increase in temperature has a causal relationship, as it is well-known that people tend to look for umbrellas with increasing temperatures. Viewed differently, an increase in temperature causes an increase in sales of umbrellas, where the increase in temperature is a cause, and the higher sales of umbrellas as an aftermath is a result. Many such cause–effect relationships exist in nature. It is of interest to the research community to see if the machine learning models capture such causal relationships [11–14] and design models which work based on causal relationships so that the spurious correlations [34] are not picked up to arrive at the prediction [15–17].

Frye et al. [113] leverage a causal graph depicting the causal relationship between features to assign Shapley values respecting the causal order where source variables are attributed more than the effects. While relationships may be intuitive in simpler tabular datasets, such causal relationships are unclear to humans in images [160]. For instance, the proposal by Kancheti et al. [16] to build models whose reasoning is aligned with the prior knowledge of the underlying causal structure obtained from the domain experts based on a specialized regularization scheme could not be demonstrated in any image dataset due to non-availability of causal knowledge on image pixels. In the absence of a complete causal structure existing between the pixels, which are the input features of images, Watson et al. [114] suggest using eye-gaze data as a proxy for ground truth causal structure, which can guide the model training to avoid picking up spurious correlations. Though the inter-dependencies between image pixels are less intuitive to humans, inter-dependencies at the level of concepts are known. For example, the presence of a car can be ascertained only when it has wheels. The detection of a concept car causes an increase in confidence in the detection of the concept of wheels [93]. Qin et al. [115] propose a causal interventional training to incorporate such causal concept relationships. Bahadori & Heckerman [15] propose using instrumental variables to debias concept representations learned by Concept Bottleneck Models [97]; thereby, the effect of confounding or correlational concepts on the prediction is mitigated. Dash et al. [17] propose leveraging the causal structure to uncover biases learned by a CNN by generating suitable counterfactuals, which can then be used to retrain the CNN in a regularized manner to debias the CNN. Singla et al. [54] leverage vision-language models to associate concept descriptions to image regions and estimate the causal relationships captured by the trained model by observing the effect of intervening the concept. Yang et al. [11] and Goyal et al. [92] propose a specialized variational autoencoder to facilitate concept-level intervention. Panda et al. [12] hypothesized that the most sparse and class discriminant features are causal and that they leverage a neural network to determine those causal superpixels that maximize the mutual information. However, it is to be noted that these architectures are, again, black boxes whose working needs to be explained, adding up to the problem at hand of explaining the CNN of interest. To eliminate the introduction of another black box to provide a causal explanation, Causal CAM [13] echoes the hypothesis of Panda et al. [12] that the class discriminant features are causal by eliminating the context features that are salient for other classes from the saliency maps generated by Grad-CAM [33], thereby yielding a saliency map highlighting

the causal features. However, as noted in the paper, this approach cannot be scaled to a multi-class classification scenario as it involves enumerating all possible subsets of the set of all class labels except the class of interest to estimate the context features, whose computation grows exponentially.

9. Explaining Cross-Domain Classification

Much effort of the XAI research community is towards explaining classifiers trained and tested on the data sampled from the same underlying distribution, called the in-domain classifiers. Cross-domain classification also plays an important role in extending the fruits of the data-hungry deep models to be reaped for data-scarce applications by adapting the models trained using large amounts of other related data to work on the scarce data sampled from a different distribution. Specifically, domain adaptation refers to the process of adapting a model trained on a data-rich source domain to a data-scarce target domain where the distributions of the data may be different [18,19]. In this context, explainability can help understand how the model adapts to the differences in the source and target domains. Similar to the discussion in the previous section, there have been methods proposed to explain cross-domain classifiers where explainability has been incorporated both during and after training of the black box, leading us to dedicate a separate section for this important under-explored research direction.

Zunino et al. [116] propose to leverage explainability approaches [33] to identify common features across both domains. Once the domain-invariant features are identified, the CNN is enforced to focus on these features to classify the instances. This, by design, forces the CNN to pay attention to discriminative domain-invariant features; thereby, the model would be accurate on any domain, and hence a domain-generalized classifier is built.

Szabó et al. [20] explores the temporal process of transfer learning. An Imagenet [161] trained model is adapted to perform a face recognition task, and the features encoded by the different filters of the CNN are analyzed using Activation Maximization (AM) [127], which performs gradient ascent in the input image so that the activation of a desired neuron of interest gets maximized. It was observed that the initial layers only adjust trivial features such as color-space to adapt to the target domain, while the latter layer filters undergo significant transformation. However, interpreting the results of AM requires expertise. It may not be suited to explain to people with good domain expertise but limited deep learning expertise, as the optimization process of AM may generate perturbed pixels from which abstracting the underlying concept as similar to how humans process images [148] is challenging. Neyshabur et al. [132] perform a detailed analysis to unearth the role of feature reuse and pre-trained weights during the process of fine-tuning.

Zhang et al. [117] extend the idea of Li et al. [95] to learn characteristic source domain prototypes whose similarity would determine the class of the given test instance. They propose building an unsupervised domain-adapted classifier with case-based reasoning [156] ability incorporated by design. As no labeled target domain instances exist in unsupervised domain adaptation, the classifier is trained using the source domain instances sampled from the same distribution from where prototypes are learned. To instill domain invariance, GAN-based domain adaptation mechanisms [162,163] are employed to generate domain-invariant features so that the target domain test instances may be classified using the same classifier, which was trained to classify the labeled source domain instances based on proximity to prototypes. A main drawback of this approach is that the prototypes are complete images, unlike recent antehoc approaches [52,53] that offer part-level explanations. Hence, this framework needs to use the framework proposed by Nauta et al. [164] as an add-on to obtain finer information regarding the prototypes. Hao & Zheng [21] use a GAN to understand features that help achieve domain invariance. However, using another black box to explain the black box of interest makes the explanation less faithful.

Kamakshi & Krishnan [22] propose building a supervised domain-adapted classifier that can explain itself. Class-specific characteristic prototypes are learned in each domain whose detection guides the prediction of the proposed case-based reasoner [156]. The differ-

ences between the domains are aligned by minimizing the highest intra-class prototypical distance while simultaneously maximizing the least prototypical distance across different classes [165]. This enables the learned explanatory backbone to pick up discriminatory features to identify the class while ignoring the domain from which the instance is sampled. However, the framework could not scale to common domain adaptation datasets with many classes. Xiao et al. [23] attempt to build a posthoc approximator for an unsupervised domain-adapted classifier based on ProtoPNet [52] whose prototypes are learned using the labeled source domain instances, which, when visualized through the unlabelled target domain instances, reveals the mapping between the source and target domain instances leveraged to classify the unlabelled target domain instances. However, this approach has challenges regarding the fidelity of the explanation as there is no consensus regarding assessing the correctness of how the features are aligned across the source and target domains. Furthermore, other frameworks [164] have to be applied to obtain additional information on what is encoded by the class-specific prototypes learned from the source domain instances.

10. Future Work

While several novel frameworks that advance the field of XAI have been discussed with seminal contributions being summarized in Table 1, several open problems are available to be solved collectively by the community. Mainly, three possible research directions are envisioned.

The preliminary direction shall be to extend the fruits of explainability to allied learning paradigms. Traditional deep learning methods were data-hungry as they leveraged voluminous chunks of data. However, various allied learning paradigms have been introduced to reap the fruits of deep learning to data-scarce scenarios. Transfer Learning aims to extend a classifier trained on a related data set containing many instances to work on the scarce data of interest by aligning the feature and label spaces. Kamakshi & Krishnan [22] propose a framework that explains a supervised domain-adapted classifier by design. Similarly, parallel works [23,117] explain an unsupervised domain-adapted classifier. Extensions to explainable classifiers using heterogeneous transfer learning and open-set domain adaptation paradigms can be a possible future avenue to explore. Few Shot Learning [166,167] aims to learn classifiers from fewer examples by leveraging features learned from related classes having a larger number of instances. For instance, a zebra can be considered as an animal with a horse-like body and tiger-like stripes. A motivating example from the medical domain would follow to distinguish it from Transfer Learning. Few Shot Learning aims to leverage features learned by a pneumonia detector to detect a related disease, say COVID-19, from fewer examples. Transfer Learning may leverage COVID-19 data collected from another country where more examples are available to learn a robust classifier that can be adapted to classify instances sampled from a country having fewer positive cases. Wang et al. [168] propose an explainable by design few-shot classifier which classifies an unseen novel test instance by matching the features detected against characteristic patterns learned in the seen categories. Incremental Learning [169,170] mimics how humans learn. For instance, a computer scientist does not learn to build an application in a day. First, the programming principles are learned, then he learns to implement the different data structures needed to manage the various modules and finally learns to assemble the modules to get the end product. While learning an intermediate skill, humans do not forget the preliminary skills acquired. However, this is not the case in AI systems. When new classes are expected to be learned by a classifier trained to classify an instance into a set of classes, they tend to forget the distinctions across older classes already learned [171]. However, the reason for such behavior is unknown. The paper envisions the application of explainability to help unravel the mechanism behind the incrementally learned classifiers, thereby guiding the research community toward building classifiers that can mimic human-level incremental knowledge expansion. Rymarczyk et al. [172] propose building an antehoc model whose explainable components

are learned such that catastrophic forgetting is managed by design. This model is an extension of ProtoPNet architecture [52] where the prototypes corresponding to the novel classes are enforced to be closer to the seen classes so that catastrophic forgetting is minimized. An extension using antehoc frameworks that encourage learning shared concepts [98] similar to how concepts are shared in nature may enable minimizing catastrophic forgetting as, despite sharedness in nature, humans expand their knowledge base without forgetting knowledge they gathered in the past.

The secondary direction shall be to develop quantitative metrics to assess the goodness of the learned concepts. In saliency map-based methods, the goodness of the explanations is quantitatively assessed by simulating the effect on perturbation of the regions deemed salient. Union of regions comprising the concepts may be unfair to assess the goodness of the concepts as the union may cover up the entire image, nullifying the assessment. Recent works [63,64] propose a new metric called agreement accuracy which assesses how well the concept-based explainer approximates the working of the CNN to be explained. Leemann et al. [173] propose using natural language models to assess the goodness of the concepts. However, interpreting the language models [151] is needed on top of the evaluation process to make it transparent. Zarlenga et al. [174] proposed metrics to assess if the learned concept representations are pure with respect to a known oracle and suggested using inter-concept disentanglement to measure if the learned representations capture dissimilar concepts. Zhou et al. [175] analyze the taxonomy proposed by Arya et al. [176] for AIX360, a popular explainability toolkit and focus on the evaluation metrics adopted by the different approaches. The main idea inferred from this analysis is that the subjective metrics such as trust, satisfaction, confidence, etc. [155] have to be evaluated involving human subject experiments. The quantitative metrics [177] evaluate the objective aspects such as fidelity and soundness of the explanations. These can help select the subset of instances, which, using the human subject experiments, can be conducted better. Agarwal et al. [178] provides an open source framework to facilitate the evaluation of popular posthoc methods to enable the stakeholders to choose the explanation that best suits the need of the application. Lopes et al. [179] propose a taxonomy of various evaluation metrics and discuss the aspects that help in designing the tasks to evaluate the XAI approaches better using human subject experiments, as humans ultimately benefit the AI system. Herm et al. [180] bring out the desiderata that XAI approaches have to fulfill for their pervasive utility through human-centered experiments. We foresee that when explainability is reaped to allied learning paradigms, metrics have to be developed to assess the correctness of the peculiar aspects of those paradigms as encoded by the explainer. For instance, if a posthoc explanatory approach is developed to unravel the mapping of features across different domains; an evaluation is needed to assess if what is being unraveled is true.

The tertiary direction suggests the cross-pollination of ideas from Neural Architecture Search [47], which aims to identify the best architecture and parameters to model the distribution from which the dataset of interest is sampled into XAI. There have been recent proposals in this direction. Liu et al. [181] suggest using intrinsically explainable components such as regressors to search for optimal configurations to achieve black-box level accuracy. Hosseini & Xie [182] propose updating the search for a suitable neural architecture based on feedback from posthoc saliency maps [33]. The paper envisions applying the principles of neural architecture search to identify the optimal number of concepts so that the accuracy–interpretability tradeoff inherent to antehoc frameworks can be minimized eventually. This, when possible, shall have a greater impact on recent classifiers employing allied learning paradigms [22,168,172] where explainability is incorporated by design.

While these are the possible future avenues with potential impact on the XAI field, an alternate route that has been started and an active area of research currently [183–185] is using the feedback from the explanation algorithm and introducing humans in the loop [186] to edit the erroneous classifier. This can be an interesting direction one can focus on, especially when working in safety-critical applications, where adhering to the working mechanism laid by domain experts is essential.

Another important direction that is a relatively under-explored area of Explainable AI is in looking into the economic aspects associated with it. Adadi & Berrara [187] hint at the importance of this analysis and discuss the preliminary attempts by the community to quantify the cost associated with bringing transparency into an AI system and establish a link between the well-studied discipline of Structural Econometrics and XAI [188,189]. Beaudouin et al. [190] analyze the possible expenses an organization has to incur when it adopts explainability along with the possible threats of disclosure of private information to cater to the needs of explainability approaches. Langer et al. [191] hint at the possibility of the economic factors acting as a confounder for the user specifying the desiderata expected from an explainable system, thereby giving rise to a possibility of springing up of a tradeoff between the mental model of the users regarding the explainers and the working mechanism of the explainers themselves. Despite being very important in shaping transparent decision support systems, this aspect needs further exploration. This is an area that requires the collaboration of XAI researchers who understand the science of extracting the working mechanism of an accurate black box, HCI researchers who understand the dynamics between humans and computing systems, and Economists who can act on the threats which the introduction of the novel system could bring in the business.

The future avenues, especially incorporating explainability into black boxes leveraging the allied learning paradigms, have great potential for developing models having human-like learning capabilities. These models can be involved in machine teaching tasks where humans and models can symbiotically create novel knowledge. Incorporating the machine teaching feedback back and forth requires a mutual understanding between the model explaining its outcomes faithfully and humans imparting domain knowledge to inculcate real-world requirements better. On a concluding note, the research explored so far in explainability is just the tip of the iceberg. We would like to inspire aspiring researchers that the exploration of the envisioned future research avenues may open up many further research avenues with the goal of marching towards Artificial General Intelligence faster.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable

Data Availability Statement: This is a review, and no new data is created.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2; pp. 1150–1157.
2. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; Volume 1; pp. 886–893.
3. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
5. Saeed, W.; Omlin, C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowl.-Based Syst.* **2023**, *263*, 110273. [CrossRef]
6. Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of Explainable AI Techniques in Healthcare. *Sensors* **2023**, *23*, 634. [CrossRef]
7. Weber, P.; Carl, K.V.; Hinz, O. Applications of Explainable Artificial Intelligence in Finance—A systematic review of Finance, Information Systems, and Computer Science literature. *Manag. Rev. Q.* **2023**. [CrossRef]
8. Clement, T.; Kemmerzell, N.; Abdelaal, M.; Amberg, M. XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 78–108. [CrossRef]
9. Cabitza, F.; Campagner, A.; Malgieri, G.; Natali, C.; Schneeberger, D.; Stoeger, K.; Holzinger, A. Quod erat demonstrandum? —Towards a typology of the concept of explanation for the design of explainable AI. *Expert Syst. Appl.* **2023**, *213*, 118888. [CrossRef]
10. Schwalbe, G.; Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Min. Knowl. Discov.* **2023**. [CrossRef]

11. Yang, C.H.H.; Liu, Y.C.; Chen, P.Y.; Ma, X.; Tsai, Y.C.J. When causal intervention meets adversarial examples and image masking for deep neural networks. In Proceedings of the IEEE International Conference on Image Processing, Taipei, Taiwan, 22–25 September 2019; pp. 3811–3815.
12. Panda, P.; Kancheti, S.S.; Balasubramanian, V.N. Instance-wise causal feature selection for model interpretation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 1756–1759.
13. Prabhushankar, M.; AlRegib, G. Extracting causal visual features for limited label classification. In Proceedings of the IEEE International Conference on Image Processing, Anchorage, AK, USA, 19–22 September 2021; pp. 3697–3701.
14. Ganguly, N.; Fazlija, D.; Badar, M.; Fisichella, M.; Sikdar, S.; Schrader, J.; Wallat, J.; Rudra, K.; Koubarakis, M.; Patro, G.K.; et al. A review of the role of causality in developing trustworthy ai systems. *arXiv* **2023**, arXiv:2302.06975.
15. Bahadori, M.T.; Heckerman, D. Debiasing Concept-based Explanations with Causal Analysis. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
16. Kancheti, S.S.; Reddy, A.G.; Balasubramanian, V.N.; Sharma, A. Matching learned causal effects of neural networks with domain priors. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; p. 10676.
17. Dash, S.; Balasubramanian, V.N.; Sharma, A. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 915–924.
18. Singhal, P.; Walambe, R.; Ramanna, S.; Kotecha, K. Domain Adaptation: Challenges, Methods, Datasets, and Applications. *IEEE Access* **2023**, *11*, 6973–7020. [CrossRef]
19. Iman, M.; Arabnia, H.R.; Rasheed, K. A review of deep transfer learning and recent advancements. *Technologies* **2023**, *11*, 40. [CrossRef]
20. Szabó, R.; Katona, D.; Csillag, M.; Csiszárik, A.; Varga, D. Visualizing Transfer Learning. In Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, Vienna, Austria, 12–18 July 2020.
21. Hou, Y.; Zheng, L. Visualizing Adapted Knowledge in Domain Transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13824–13833.
22. Kamakshi, V.; Krishnan, N.C. Explainable supervised domain adaptation. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–8.
23. Xiao, W.; Ding, Z.; Liu, H. Visualizing Transferred Knowledge: An Interpretive Model of Unsupervised Domain Adaptation. *arXiv* **2023**, arXiv:2303.02302.
24. Sarkar, N.K.; Singh, M.M.; Nandi, U. Recent Researches on Image Classification Using Deep Learning Approach. *Int. J. Comput. Digit. Syst.* **2022**, *12*, 1357–1374. [CrossRef]
25. Zaidi, S.S.A.; Ansari, M.S.; Aslam, A.; Kanwal, N.; Asghar, M.; Lee, B. A survey of modern deep learning based object detection models. *Digit. Signal Process.* **2022**, *126*, 103514. [CrossRef]
26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
27. Gonzalez-Garcia, A. Image Context for Object Detection, Object Context for Part Detection. Ph.D. Thesis, The University of Edinburgh, Edinburgh, UK, 2018.
28. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Object detectors emerge from training CNNs for scene recognition. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–12.
29. Lipton, Z.C. The doctor just won't accept that! Interpretable ML symposium. In Proceedings of the Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
30. Bonicalzi, S. A matter of justice. The opacity of algorithmic decision-making and the trade-off between uniformity and discretion in legal applications of artificial intelligence. *Teor. Riv. Filos.* **2022**, *42*, 131–147.
31. Council of European Union. 2018 Reform of EU Data Protection Rules, 2018. Available online: https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf (accessed on 1 June 2019).
32. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
33. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 618–626.
34. Neuhaus, Y.; Augustin, M.; Boreiko, V.; Hein, M. Spurious Features Everywhere—Large-Scale Detection of Harmful Spurious Features in ImageNet. *arXiv* **2022**, arXiv:2212.04871.
35. Vilone, G.; Longo, L. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 615–661. [CrossRef]
36. Martino, F.D.; Delmastro, F. Explainable AI for clinical and remote health applications: A survey on tabular and time series data. *Artif. Intell. Rev.* **2023**, *56*, 5261–5315. [CrossRef]
37. Zhu, L.; Zhu, Z.; Zhang, C.; Xu, Y.; Kong, X. Multimodal sentiment analysis based on fusion methods: A survey. *Inf. Fusion* **2023**, *95*, 306–325. [CrossRef]

38. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
39. Salahuddin, Z.; Woodruff, H.C.; Chatterjee, A.; Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput. Biol. Med.* **2022**, *140*, 105111. [CrossRef]
40. Holzinger, A.; Malle, B.; Saranti, A.; Pfeifer, B. Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. *Inf. Fusion* **2021**, *71*, 28–37. [CrossRef]
41. Owens, E.; Sheehan, B.; Mullins, M.; Cunneen, M.; Ressel, J.; Castignani, G. Explainable Artificial Intelligence (XAI) in Insurance. *Risks* **2022**, *10*, 230. [CrossRef]
42. Shanthini, M.; Sanmugam, B. A Performance Comparison of State-of-the-Art Imputation and Classification Strategies on Insurance Fraud Detection. In *Micro-Electronics and Telecommunication Engineering: Proceedings of 6th ICMETE, Ghaziabad, India, 2022*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 215–225.
43. Barnett, A.J.; Schwartz, F.R.; Tao, C.; Chen, C.; Ren, Y.; Lo, J.Y.; Rudin, C. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nat. Mach. Intell.* **2021**, *3*, 1061–1070. [CrossRef]
44. Wu, S.; Yuksekgonul, M.; Zhang, L.; Zou, J. Discover and Cure: Concept-aware Mitigation of Spurious Correlation. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
45. Degas, A.; Islam, M.R.; Hurter, C.; Barua, S.; Rahman, H.; Poudel, M.; Ruscio, D.; Ahmed, M.U.; Begum, S.; Rahman, M.A.; et al. A Survey on Artificial Intelligence (AI) and Explainable AI in Air Traffic Management: Current Trends and Development with Future Research Trajectory. *Appl. Sci.* **2022**, *12*, 1295. [CrossRef]
46. Buhrmester, V.; Münch, D.; Arens, M. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 966–989. [CrossRef]
47. Kang, J.S.; Kang, J.; Kim, J.J.; Jeon, K.W.; Chung, H.J.; Park, B.H. Neural Architecture Search Survey: A Computer Vision Perspective. *Sensors* **2023**, *23*, 1713. [CrossRef]
48. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
49. Desai, S.; Ramaswamy, H.G. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 983–991.
50. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
51. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2921–2929.
52. Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; Su, J.K. This looks like that: Deep learning for interpretable image recognition. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 8928–8939.
53. Nauta, M.; van Bree, R.; Seifert, C. Neural prototype trees for interpretable fine-grained image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 14933–14943.
54. Singla, S.; Wallace, S.; Triantafillou, S.; Batmanghelich, K. Using causal analysis for conceptual deep learning explanation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention; Springer: Berlin/Heidelberg, Germany, 2021; pp. 519–528.
55. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
56. Friedman, J.H.; Popescu, B.E. Predictive learning via rule ensembles. *Ann. Appl. Stat.* **2008**, *2*, 916–954. [CrossRef]
57. Lundberg, S.M.; Erion, G.G.; Lee, S.I. Consistent individualized feature attribution for tree ensembles. In Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, Sydney, Australia, 11–15 August 2017.
58. Harris, C.; Pymar, R.; Rowat, C. Joint Shapley values: A measure of joint feature importance. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
59. Zafar, M.R.; Khan, N. Deterministic local interpretable model-agnostic explanations for stable explainability. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 525–541. [CrossRef]
60. Sharma, R.; Reddy, N.; Kamakshi, V.; Krishnan, N.C.; Jain, S. MAIRE-A Model-Agnostic Interpretable Rule Extraction Procedure for Explaining Classifiers. In *Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 329–349.
61. Lakkaraju, H.; Kamar, E.; Caruana, R.; Leskovec, J. Faithful and customizable explanations of black box models. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; pp. 131–138.
62. Ben Zaken, D.; Segal, A.; Cavalier, D.; Shani, G.; Gal, K. Generating Recommendations with Post-Hoc Explanations for Citizen Science. In Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, 4–7 July 2022; pp. 69–78.
63. Kumar, A.; Sehgal, K.; Garg, P.; Kamakshi, V.; Krishnan, N.C. MACE: Model Agnostic Concept Extractor for Explaining Image Classification Networks. *IEEE Trans. Artif. Intell.* **2021**, *2*, 574–583. [CrossRef]

64. Kamakshi, V.; Gupta, U.; Krishnan, N.C. PACE: Posthoc Architecture-Agnostic Concept Extractor for Explaining CNNs. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
65. Collaris, D.; Gajane, P.; Jorritsma, J.; van Wijk, J.J.; Pechenizkiy, M. LEMON: Alternative Sampling for More Faithful Explanation through Local Surrogate Models. In Proceedings of the Advances in Intelligent Data Analysis XXI, Louvain-la-Neuve, Belgium, 12–14 April 2023; pp. 77–90.
66. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 24–25.
67. Salama, A.; Adly, N.; Torki, M. Ablation-CAM++: Grouped Recursive Visual Explanations for Deep Convolutional Networks. In Proceedings of the IEEE International Conference on Image Processing, Bordeaux, France, 16–19 October 2022; pp. 2011–2015.
68. Lee, J.R.; Kim, S.; Park, I.; Eo, T.; Hwang, D. Relevance-cam: Your model already knows where to look. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 14944–14953.
69. Jung, H.; Oh, Y. Towards better explanations of class activation mapping. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1336–1344.
70. Sattarzadeh, S.; Sudhakar, M.; Plataniotis, K.N.; Jang, J.; Jeong, Y.; Kim, H. Integrated grad-CAM: Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6–11 June 2021; pp. 1775–1779.
71. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2668–2677.
72. Pfau, J.; Young, A.T.; Wei, J.; Wei, M.L.; Keiser, M.J. Robust semantic interpretability: Revisiting concept activation vectors. In Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, Vienna, Austria, 12–18 July 2020.
73. Ghorbani, A.; Wexler, J.; Zou, J.Y.; Kim, B. Towards automatic concept-based explanations. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 10–12 December 2019; pp. 9277–9286.
74. Yuksekgonul, M.; Wang, M.; Zou, J. Post-hoc Concept Bottleneck Models. In Proceedings of the ICLR Workshop on PAIR2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data, Virtual, 25–29 April 2022.
75. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 11–15 August 2017; pp. 3319–3328.
76. Zhang, J.; Bargal, S.A.; Lin, Z.; Brandt, J.; Shen, X.; Sclaroff, S. Top-down neural attention by excitation backprop. *Int. J. Comput. Vis.* **2018**, *126*, 1084–1102. [CrossRef]
77. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. In Proceedings of the Workshop at International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
78. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In Proceedings of the Workshop at International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
79. Fong, R.; Patrick, M.; Vedaldi, A. Understanding deep networks via extremal perturbations and smooth masks. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2950–2958.
80. Fong, R.C.; Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3429–3437.
81. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef] [PubMed]
82. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 11–15 August 2017; pp. 3145–3153.
83. Wang, P.; Kong, X.; Guo, W.; Zhang, X. Exclusive Feature Constrained Class Activation Mapping for Better Visual Explanation. *IEEE Access* **2021**, *9*, 61417–61428. [CrossRef]
84. Hartley, T.; Sidorov, K.; Willis, C.; Marshall, D. SWAG: Superpixels weighted by average gradients for explanations of CNNs. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 423–432.
85. Goyal, Y.; Wu, Z.; Ernst, J.; Batra, D.; Parikh, D.; Lee, S. Counterfactual visual explanations. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 11–13 June 2019; pp. 2376–2384.
86. Abid, A.; Yuksekgonul, M.; Zou, J. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 66–88.
87. Singla, S.; Pollack, B. Explanation by Progressive Exaggeration. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
88. Wang, P.; Vasconcelos, N. Scout: Self-aware discriminant counterfactual explanations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8981–8990.
89. Zhao, Y. Fast real-time counterfactual explanations. In Proceedings of the Workshop at International Conference on Machine Learning, Vienna, Austria, 13–18 July 2020.

90. Arendsen, P.; Marcos, D.; Tuia, D. Concept discovery for the interpretation of landscape scenicness. *Mach. Learn. Knowl. Extr.* **2020**, *2*, 22. [CrossRef]
91. Yeh, C.K.; Kim, B.; Arik, S.; Li, C.L.; Pfister, T.; Ravikumar, P. On Completeness-aware Concept-Based Explanations in Deep Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 20554–20565.
92. Goyal, Y.; Feder, A.; Shalit, U.; Kim, B. Explaining classifiers with Causal Concept Effect (CaCE). *arXiv* **2019**, arXiv:1907.07165.
93. Lopez-Paz, D.; Nishihara, R.; Chintala, S.; Scholkopf, B.; Bottou, L. Discovering causal signals in images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, 21–26 July 2017; pp. 6979–6987.
94. Wang, J.; Liu, H.; Wang, X.; Jing, L. Interpretable image recognition by constructing transparent embedding space. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 895–904.
95. Li, O.; Liu, H.; Chen, C.; Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
96. Hendricks, L.A.; Rohrbach, A.; Schiele, B.; Darrell, T.; Akata, Z. Generating visual explanations with natural language. *Appl. AI Lett.* **2021**, *2*, e55. [CrossRef]
97. Koh, P.W.; Nguyen, T.; Tang, Y.S.; Mussmann, S.; Pierson, E.; Kim, B.; Liang, P. Concept bottleneck models. In Proceedings of the International Conference on Machine Learning, PMLR, Vienna, Austria, 12–18 July 2020; pp. 5338–5348.
98. Rymarczyk, D.; Struski, Ł.; Górszczak, M.; Lewandowska, K.; Tabor, J.; Zieliński, B. Interpretable image classification with differentiable prototypes assignment. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 351–368.
99. Rymarczyk, D.; Struski, Ł.; Tabor, J.; Zieliński, B. ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14–18 August 2021; pp. 1420–1430.
100. Jain, S.; Wallace, B.C. Attention is not Explanation. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 3543–3556.
101. Hassanin, M.; Anwar, S.; Radwan, I.; Khan, F.S.; Mian, A. Visual attention methods in deep learning: An in-depth survey. *arXiv* **2022**, arXiv:2204.07756.
102. Mohankumar, A.K.; Nema, P.; Narasimhan, S.; Khapra, M.M.; Srinivasan, B.V.; Ravindran, B. Towards Transparent and Explainable Attention Models. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA, 5–10 July 2020; pp. 4206–4216.
103. Xu, W.; Wang, J.; Wang, Y.; Xu, G.; Lin, D.; Dai, W.; Wu, Y. Where is the Model Looking at—Concentrate and Explain the Network Attention. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 506–516. [CrossRef]
104. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
105. Zhang, Q.; Nian Wu, Y.; Zhu, S.C. Interpretable convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8827–8836.
106. Hase, P.; Chen, C.; Li, O.; Rudin, C. Interpretable Image Recognition with Hierarchical Prototypes. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Honolulu, HI, USA, 27 January–1 February 2019; Volume 7, pp. 32–40.
107. Kim, Y.; Mo, S.; Kim, M.; Lee, K.; Lee, J.; Shin, J. Explaining Visual Biases as Words by Generating Captions. *arXiv* **2023**, arXiv:2301.11104.
108. Yang, Y.; Kim, S.; Joo, J. Explaining deep convolutional neural networks via latent visual-semantic filter attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–23 June 2022; pp. 8333–8343.
109. Hendricks, L.A.; Hu, R.; Darrell, T.; Akata, Z. Grounding visual explanations. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 264–279.
110. Wickramanayake, S.; Hsu, W.; Lee, M.L. Comprehensible convolutional neural networks via guided concept learning. In Proceedings of the International Joint Conference on Neural Networks, Shenzhen, China, 18–22 July 2021; pp. 1–8.
111. Hendricks, L.A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; Darrell, T. Generating visual explanations. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 3–19.
112. Khan, M.A.; Oikarinen, T.; Weng, T.W. Concept-Monitor: Understanding DNN training through individual neurons. *arXiv* **2023**, arXiv:2304.13346.
113. Frye, C.; Rowat, C.; Feige, I. Asymmetric shapley values: Incorporating causal knowledge into model-agnostic explainability. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1229–1239.
114. Watson, M.; Hasan, B.A.S.; Al Moubayed, N. Learning How to MIMIC: Using Model Explanations To Guide Deep Learning Training. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 1461–1470.
115. Qin, W.; Zhang, H.; Hong, R.; Lim, E.P.; Sun, Q. Causal interventional training for image recognition. *IEEE Trans. Multimed.* **2021**, *25*, 1033–1044. [CrossRef]

116. Zunino, A.; Bargal, S.A.; Volpi, R.; Sameki, M.; Zhang, J.; Sclaroff, S.; Murino, V.; Saenko, K. Explainable deep classification models for domain generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 3233–3242.
117. Zhang, Y.; Yao, T.; Qiu, Z.; Mei, T. Explaining Cross-Domain Recognition with Interpretable Deep Classifier. *arXiv* **2022**, arXiv:2211.08249.
118. Maillot, N.E.; Thonnat, M. Ontology based complex object recognition. *Image Vis. Comput.* **2008**, *26*, 102–113. [CrossRef]
119. Ordonez, V.; Liu, W.; Deng, J.; Choi, Y.; Berg, A.C.; Berg, T.L. Predicting entry-level categories. *Int. J. Comput. Vis.* **2015**, *115*, 29–43. [CrossRef]
120. Liao, Q.; Poggio, T. In *Object-Oriented Deep Learning*; Technical Report; Center for Brains, Minds and Machines (CBMM): Cambridge, MA, USA, 2017.
121. Icarte, R.T.; Baier, J.A.; Ruz, C.; Soto, A. How a general-purpose commonsense ontology can improve performance of learning-based image retrieval. In Proceedings of the International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 1283–1289.
122. Marino, K.; Salakhutdinov, R.; Gupta, A. The More You Know: Using Knowledge Graphs for Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2673–2681.
123. Alirezaie, M.; Längkvist, M.; Sioutis, M.; Loutfi, A. A Symbolic Approach for Explaining Errors in Image Classification Tasks. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018.
124. Daniels, Z.A.; Frank, L.D.; Menart, C.J.; Raymer, M.; Hitzler, P. A framework for explainable deep neural models using external knowledge graphs. In *Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*; SPIE: Bellingham, WA, USA, 2020; Volume 11413, pp. 480–499.
125. Tiddi, I.; Schlobach, S. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.* **2022**, *302*, 103627. [CrossRef]
126. Veugen, T.; Kamphorst, B.; Marcus, M. Privacy-preserving contrastive explanations with local foil trees. *Cryptography* **2022**, *6*, 54. [CrossRef]
127. Nguyen, A.; Dosovitskiy, A.; Yosinski, J.; Brox, T.; Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3387–3395.
128. Wang, P.; Nvasconcelos, N. Deliberative explanations: Visualizing network insecurities. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 10–12 December 2019; Volume 32.
129. Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; Fernández-Leal, Á. Human-in-the-loop machine learning: A state of the art. *Artif. Intell. Rev.* **2023**, *56*, 3005–3054. [CrossRef]
130. Hendricks, L.A.; Hu, R.; Darrell, T.; Akata, Z. Generating Counterfactual Explanations with Natural Language. In Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 95–98.
131. Feldhus, N.; Hennig, L.; Nasert, M.D.; Ebert, C.; Schwarzenberg, R.; Möller, S. Saliency Map Verbalization: Comparing Feature Importance Representations from Model-free and Instruction-based Methods. In Proceedings of the First Workshop on Natural Language Reasoning and Structured Explanations (NLRSE), Toronto, ON, Canada, 13 July 2023.
132. Neyshabur, B.; Sedghi, H.; Zhang, C. What is being transferred in transfer learning? *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 512–523.
133. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 93. [CrossRef]
134. Verma, S.; Arthur, A.; Dickerson, J.; Hines, K. Counterfactual Explanations for Machine Learning: A Review. In Proceedings of the NeurIPS Workshop: ML Retrospectives, Surveys & Meta-Analyses, Vancouver, BC, Canada, 6–12 December 2020.
135. Ramaswamy, V.V.; Kim, S.S.; Fong, R.; Russakovsky, O. Overlooked factors in concept-based explanations: Dataset choice, concept salience, and human capability. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 20–22 June 2023.
136. Wang, Y.; Su, H.; Zhang, B.; Hu, X. Learning Reliable Visual Saliency for Model Explanations. *IEEE Trans. Multimed.* **2019**, *22*, 1796–1807. [CrossRef]
137. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity checks for saliency maps. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 2–8 December 2018; pp. 9505–9515.
138. Sixt, L.; Granz, M.; Landgraf, T. When Explanations Lie: Why Modified BP Attribution Fails. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 12–18 July 2020.
139. Dabkowski, P.; Gal, Y. Real time image saliency for black box classifiers. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6967–6976.
140. Huang, X.; Marques-Silva, J. The Inadequacy of Shapley Values for Explainability. *arXiv* **2023**, arXiv:2302.08160.
141. Felzenszwalb, P.F.; Huttenlocher, D.P. Efficient graph-based image segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167–181. [CrossRef]
142. Vedaldi, A.; Soatto, S. Quick shift and kernel methods for mode seeking. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 705–718.
143. Rasouli, P.; Chieh Yu, I. CARE: Coherent actionable recourse based on sound counterfactual explanations. *Int. J. Data Sci. Anal.* **2022**. [CrossRef]

144. Pawelczyk, M.; Agarwal, C.; Joshi, S.; Upadhyay, S.; Lakkaraju, H. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 28–30 March 2022; pp. 4574–4594.
145. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–11 December 2014; Volume 27.
146. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.
147. Lang, O.; Gandelsman, Y.; Yarom, M.; Wald, Y.; Elidan, G.; Hassidim, A.; Freeman, W.T.; Isola, P.; Globerson, A.; Irani, M.; et al. Explaining in style: Training a gan to explain a classifier in stylespace. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 693–702.
148. Lake, B.M.; Ullman, T.D.; Tenenbaum, J.B.; Gershman, S.J. Building machines that learn and think like people. *Behav. Brain Sci.* **2017**, *40*, e253. [CrossRef]
149. Armstrong, S.L.; Gleitman, L.R.; Gleitman, H. What some concepts might not be. *Cognition* **1983**, *13*, 263–308. [CrossRef] [PubMed]
150. Biederman, I. Recognition-by-components: A theory of human image understanding. *Psychol. Rev.* **1987**, *94*, 115. [CrossRef]
151. Gurrapu, S.; Kulkarni, A.; Huang, L.; Lourentzou, I.; Freeman, L.; Batarseh, F.A. Rationalization for Explainable NLP: A Survey. *arXiv* **2023**, arXiv:2301.08912.
152. Wu, J.; Mooney, R. Faithful Multimodal Explanation for Visual Question Answering. In Proceedings of the ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Florence, Italy, 28 July–2 August 2019; pp. 103–112.
153. Park, D.H.; Hendricks, L.A.; Akata, Z.; Rohrbach, A.; Schiele, B.; Darrell, T.; Rohrbach, M. Multimodal explanations: Justifying decisions and pointing to the evidence. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, 18–22 June 2018; pp. 8779–8788.
154. Akula, A.R.; Zhu, S.C. Attention cannot be an Explanation. *arXiv* **2022**, arXiv:2201.11194.
155. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for explainable AI: Challenges and prospects. *arXiv* **2018**, arXiv:1812.04608.
156. Richter, M.M.; Weber, R.O. *Case-Based Reasoning*; Springer: Berlin/Heidelberg, Germany, 2016.
157. Hoffmann, A.; Fanconi, C.; Rade, R.; Kohler, J. This looks like that... does it? shortcomings of latent space prototype interpretability in deep networks. *arXiv* **2021**, arXiv:2105.02968.
158. Huang, Q.; Xue, M.; Zhang, H.; Song, J.; Song, M. Is ProtoPNet Really Explainable? Evaluating and Improving the Interpretability of Prototypes. *arXiv* **2022**, arXiv:2212.05946.
159. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
160. Reimers, C.; Runge, J.; Denzler, J. Determining the Relevance of Features for Deep Neural Networks. In Proceedings of the European Conference on Computer Vision, Glasgow, Scotland, UK, 23–28 August 2020; pp. 330–346.
161. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
162. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial discriminative domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7167–7176.
163. Pei, Z.; Cao, Z.; Long, M.; Wang, J. Multi-adversarial domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
164. Nauta, M.; Jutte, A.; Provoost, J.; Seifert, C. This looks like that, because... explaining prototypes for interpretable image recognition. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 441–456.
165. Zhou, X.; Xu, X.; Venkatesan, R.; Swaminathan, G.; Majumder, O. d-SNE: Domain Adaptation Using Stochastic Neighborhood Embedding. In *Domain Adaptation in Computer Vision with Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 43–56.
166. Köhler, M.; Eisenbach, M.; Gross, H.M. Few-Shot Object Detection: A Comprehensive Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2023.** [CrossRef]
167. Cai, H.; Zhu, X.; Wen, P.; Han, W.; Wu, L. A Survey of Few-Shot Learning for Image Classification of Aerial Objects. In *Proceedings of the China Aeronautical Science and Technology Youth Science Forum*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 570–582.
168. Wang, B.; Li, L.; Verma, M.; Nakashima, Y.; Kawasaki, R.; Nagahara, H. Match them up: Visually explainable few-shot image classification. *Appl. Intell.* **2023**, *53*, 10956–10977. [CrossRef]
169. Menezes, A.G.; de Moura, G.; Alves, C.; de Carvalho, A.C. Continual Object Detection: A review of definitions, strategies, and challenges. *Neural Netw.* **2023**, *161*, 476–493. [CrossRef]
170. Wang, S.; Zhu, L.; Shi, L.; Mo, H.; Tan, S. A Survey of Full-Cycle Cross-Modal Retrieval: From a Representation Learning Perspective. *Appl. Sci.* **2023**, *13*, 4571. [CrossRef]
171. Masana, M.; Liu, X.; Twardowski, B.; Menta, M.; Bagdanov, A.D.; van de Weijer, J. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5513–5533. [CrossRef]

172. Rymarczyk, D.; van de Weijer, J.; Zieliński, B.; Twardowski, B. ICICLE: Interpretable Class Incremental Continual Learning. *arXiv* **2023**, arXiv:2303.07811.
173. Leemann, T.; Rong, Y.; Kraft, S.; Kasneci, E.; Kasneci, G. Coherence Evaluation of Visual Concepts With Objects and Language. In Proceedings of the ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality, Virtual, 29 April 2022.
174. Zarlenga, M.E.; Barbiero, P.; Shams, Z.; Kazhdan, D.; Bhatt, U.; Weller, A.; Jamnik, M. Towards Robust Metrics For Concept Representation Evaluation. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023.
175. Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* **2021**, *10*, 593. [CrossRef]
176. Arya, V.; Bellamy, R.K.; Chen, P.Y.; Dhurandhar, A.; Hind, M.; Hoffman, S.C.; Houde, S.; Liao, Q.V.; Luss, R.; Mojsilovic, A.; et al. One Explanation Does Not Fit All: A Toolkit And Taxonomy Of AI Explainability Techniques. In Proceedings of the INFORMS Annual Meeting, Houston, TX, USA, 3–5 April 2021.
177. Elkhawaga, G.; Elzeki, O.; Abuelkheir, M.; Reichert, M. Evaluating Explainable Artificial Intelligence Methods Based on Feature Elimination: A Functionality-Grounded Approach. *Electronics* **2023**, *12*, 1670. [CrossRef]
178. Agarwal, C.; Krishna, S.; Saxena, E.; Pawelczyk, M.; Johnson, N.; Puri, I.; Zitnik, M.; Lakkaraju, H. Openxai: Towards a transparent evaluation of model explanations. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 15784–15799.
179. Lopes, P.; Silva, E.; Braga, C.; Oliveira, T.; Rosado, L. XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Appl. Sci.* **2022**, *12*, 9423. [CrossRef]
180. Herm, L.V.; Heinrich, K.; Wanner, J.; Janiesch, C. Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *Int. J. Inf. Manag.* **2023**, *69*, 102538. [CrossRef]
181. Liu, C.H.; Han, Y.S.; Sung, Y.Y.; Lee, Y.; Chiang, H.Y.; Wu, K.C. FOX-NAS: Fast, On-device and Explainable Neural Architecture Search. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 789–797.
182. Hosseini, R.; Xie, P. Saliency-Aware Neural Architecture Search. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 14743–14757.
183. Santurkar, S.; Tsipras, D.; Elango, M.; Bau, D.; Torralba, A.; Madry, A. Editing a classifier by rewriting its prediction rules. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 23359–23373.
184. Wang, J.; Hu, R.; Jiang, C.; Hu, R.; Sang, J. Counterexample Contrastive Learning for Spurious Correlation Elimination. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 4930–4938.
185. Tanno, R.; F Pradier, M.; Nori, A.; Li, Y. Repairing Neural Networks by Leaving the Right Past Behind. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 13132–13145.
186. Johs, A.J.; Agosto, D.E.; Weber, R.O. Explainable artificial intelligence and social science: Further insights for qualitative investigation. *Appl. AI Lett.* **2022**, *3*, e64. [CrossRef]
187. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
188. Igami, M. Artificial Intelligence as Structural Estimation: Economic Interpretations of Deep Blue, Bonanza, and AlphaGo. *arXiv* **2018**, arXiv:1710.10967.
189. Akyol, E.; Langbort, C.; Basar, T. Price of transparency in strategic machine learning. *arXiv* **2016**, arXiv:1610.08210.
190. Beaudouin, V.; Bloch, I.; Bounie, D.; Clémençon, S.; d'Alché Buc, F.; Eagan, J.; Maxwell, W.; Mozharovskyi, P.; Parekh, J. Flexible and context-specific AI explainability: A multidisciplinary approach. *arXiv* **2020**, arXiv:2003.07703.
191. Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; Baum, K. What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell.* **2021**, *296*, 103473. [CrossRef]