A Generalized Explanation Framework for Visualization of Deep Learning Model Predictions

Pei Wang, *Student Member, IEEE*, and Nuno Vasconcelos, *Fellow, IEEE*

***Abstract*—Attribution-based explanations are popular in computer vision but of limited use for fine-grained classification problems typical of expert domains, where classes differ by subtle details. In these domains, users also seek understanding of "why" a class was chosen and "why not" an alternative class. A new *GenerAlized expLanatiOn fRamEwork* (GALORE) is proposed to satisfy all these requirements, by unifying attributive explanations with explanations of two other types. The first is a new class of explanations, denoted *deliberative*, proposed to address the "why" question, by exposing the network insecurities about a prediction. The second is the class of counterfactual explanations, which have been shown to address the "why not" question but are now more efficiently computed. GALORE unifies these explanations by defining them as combinations of attribution maps with respect to various classifier predictions and a confidence score. An evaluation protocol that leverages object recognition (CUB200) and scene classification (ADE20 K) datasets combining part and attribute annotations is also proposed. Experiments show that confidence scores can improve explanation accuracy, deliberative explanations provide insight into the network deliberation process, the latter correlates with that performed by humans, and counterfactual explanations enhance the performance of human students in machine teaching experiments.**

***Index Terms*—Attribution, confidence scores, counterfactual explanations, deep learning, deliberative explanations, explainable AI.**

I. Introduction

While deep learning systems enabled significant advances in computer vision, their black-box nature creates difficulties for many applications. In general, it is difficult to *trust* a system that cannot justify its decisions. This motivated a large literature on explainable AI (XAI) methods, which complement network predictions with human-understandable explanations [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. In computer vision, the dominant XAI paradigm is that of visual explanations computed by *attribution* functions, which generate heatmaps localizing the image pixels [8], [11], [12], [13] or regions [14], [15], [16], [17] responsible for network predictions. Fig. 1 (center) shows the heatmap produced for a bird image by a deep learning system that predicts the label 'Cardinal' with confidence value 0.76.

While attributive explanations provide a *coarse* justification for the predictions, e.g., localizing the object within a larger background or highlighting one among distinct objects in the field of view, they are not sufficient for applications that require fine-grained classification. This can be seen in Fig. 1, where it is clear that the highlighted pixels belong to the bird but unclear which regions of the bird are responsible for the 'Cardinal' prediction. While the explanation would be satisfactory for a classification problem opposing 'Birds' to 'Dogs,' it is not helpful for one opposing 'Cardinals' to 'Summer Tanagers' or other bird species. In this case, the attributive explanation selects the entire bird and it is hard to know what differentiates one class from the other.

Fine-grained classification problems are prevalent in expert domains, such as medical imaging or biology, where there is a need to distinguish objects that differ in subtle details, and even for everyday applications that involve a large number of classes. For such problems, users are likely to demand more from the explanation system. As Fig. 1 illustrates, given the relatively low confidence value of 0.76, a user may want to know exactly why the system chose the 'Cardinal' label. Beyond the post-hoc analysis of classification results, where the user is passive, explanations also play a critical role in interactive applications, such as machine teaching systems where users are taught to annotate images [18], [19], [20]. In this case, users naturally ask counterfactual questions, such as "why is this a Cardinal and not a Summer Tanager?" where an alternative or counter-class ('Summer Tanager') is provided. None of these questions can be satisfied by existing attribution-based visual explanations.

In this work, we propose a *GenerAlized expLanatiOn fRamEwork* (GALORE) for the solution of all these problems. Beyond the popular attributive explanations, GALORE includes a new class of explanations, denoted as *deliberative*, and a new version of *counterfactual explanations*[1] that are easier to compute than those previously available in the literature. Deliberative explanations, illustrated in the left of Fig. 1, address the "why?" question by visualizing insecurities about model predictions. These are the regions that the model considered most ambiguous, together

Manuscript received 8 June 2022; revised 20 December 2022; accepted 23 January 2023. Date of publication 1 February 2023; date of current version 30 June 2023. This work was supported in part by NSF under Awards IIS-1924937 and IIS-2041009, and in part by NVIDIA GPU donations. Recommended for acceptance by M. Cheng. *(Corresponding author: Pei Wang.)*

The authors are with the Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093 USA (e-mail: pew062@eng.ucsd.edu; nvasconcelos@ucsd.edu).

Digital Object Identifier 10.1109/TPAMI.2023.3241106

1.As discussed in [21] and defined in [22], [23], counterfactual explanations are similar to contrastive explanations. Both aim to answer questions "Why P and not Q," although some literature emphasizes that counterfactual explanations should generate alternative examples, illustrating how objects change for the alternative decision [24], [25]. We make no distinction and use the two terms interchangeably.

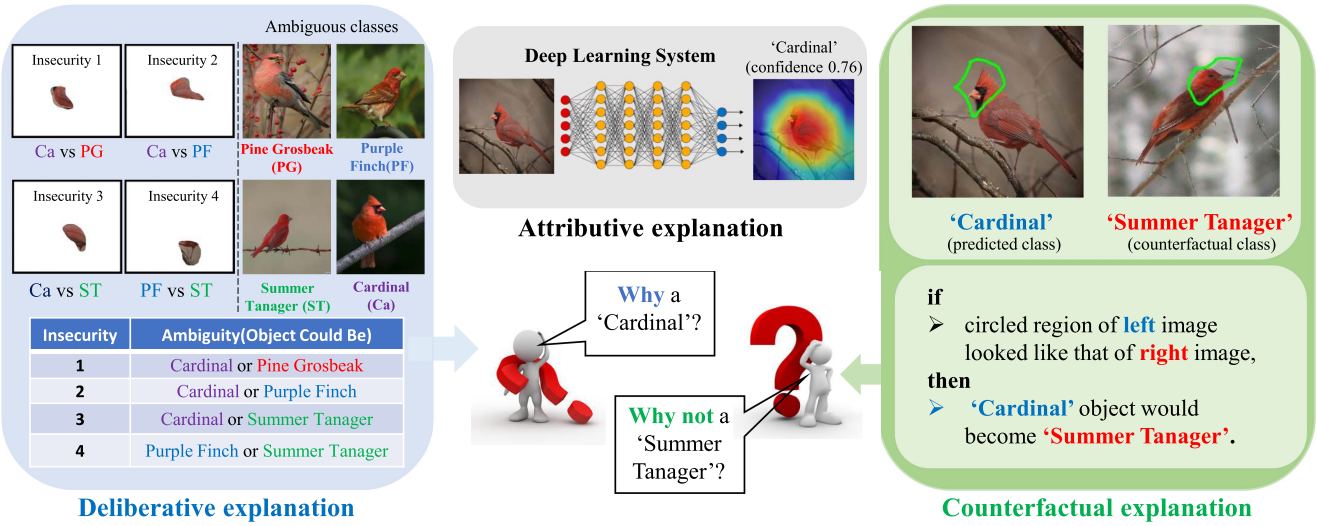

Fig. 1. An ideal explainable deep learning system should produce various explanations to satisfy different user requirements. GALORE addresses this problem by unifying attributive (center), deliberative (left), and counterfactual (right) explanations. Attributive explanations highlight the pixels responsible for the prediction of the label 'Cardinal' for the image shown. Deliberative explanations address the *why* question, producing a set of insecurities, which are image regions deemed ambiguous, together with the classes that define the ambiguity. Counterfactual explanations address the *why not* question, visualizing the input changes needed to elicit the prediction of a user-provided counter class ('Summer Tanager').

with the classes that define the ambiguity. In the example of the figure, insecurities refer to body parts of bird classes that are confusable with 'Cardinal,' such as 'Pine Grosbeak,' 'Purple Finch,' or 'Summer Tanager'. Counterfactual explanations, illustrated in the right of the Figure, address the "why not?" question by visualizing the input changes needed to elicit the prediction of a user-provided counter class. In the example of the figure, the explanation shows that the two classes differ mostly in terms of the bird head. The unification is based on the definition of all explanations as combinations of multiple attribution maps, which vary according to the explanation type. Since attributions are very efficient to compute, the proposed framework establishes a family of low-complexity explanations that can be used in various applications, ranging from naive to expert domains, and supporting both passive post-hoc analysis of predictions or interactive applications such as machine teaching.

A core requirement of deliberative and counterfactual explanations is the ability to reason in terms of the *difficulty* posed to the classification by different image regions. Understanding why the classifier chose a class requires knowing what other classes could have been plausibly selected, and what image regions made those alternatives plausible, i.e., what image regions the classifier found ambiguous for the decision. This is the essence of deliberative explanations, which produce a list of such regions, denoted as *insecurities*, as illustrated in the left of Fig. 1. On the other hand, counterfactual explanations require the identification of regions that discriminate the predicted from the counterfactual class, i.e., which have high probability under the predicted class and low probability under the counterfactual. These regions can then be shown to the user, as illustrated in right of Fig. 1, to identify corresponding parts in objects from predicted and counter class.

Reasoning about ambiguities or class probabilities requires the classifier to produce confidence scores [26], [27], [28], [29], i.e., measure the confidence with which the image belongs to each of the possible classes. From these scores, it is possible to derive how difficult the classification is (the probability of the ground-truth class), how ambiguous it is (similarity between the probabilities of the top classes), or how much the image discriminates between two classes (large probability for one and small for another). We refer to the ability to measure these quantities as *self-awareness,* since it allows a classifier to quantify the confidence in its decisions. One of the insights of this work is that attributions of confidence scores allow the extension of these measures to image regions, so as to identify which regions are ambiguous, discriminant, or difficult to classify. This is naturally integrated in the GALORE framework, by simply combining the attribution maps for self-awareness with the attributions for class predictions required to compute the different explanations.

Beyond explanations, a significant challenge to XAI is the lack of explanation ground truth for performance evaluation. Besides user-based evaluations [30], whose results are difficult to replicate, we propose a quantitative metric based on a proxy localization task. This relies on standard metrics from the object detection literature and attribute annotations for different object parts or scene components. We show that these metrics can be adapted to the evaluation of the different types of explanations proposed with minor specializations. Compared to human experiments, the proposed proxy evaluation has the advantages of being substantially easier to perform and fully replicable.

Overall, the paper makes three contributions. First, it proposes the unified GALORE framework to generate attributive, deliberative, and counterfactual explanations. Deliberative explanations are a newly proposed family of explanations that visualize the deliberations made by a network to reach its predictions. GALORE also redefines counterfactual explanations as combinations of attributive explanations, significantly increasing their computational efficiency. Second, the paper shows how to leverage self-awareness to improve explanation accuracy, for different types of explanations. Third, it proposes a new experimental protocol for quantitative evaluation of deliberative and counterfactual explanations. Experimental results, using

both this protocol and human experiments, show that the proposed deliberative explanations are intuitive, suggesting that the deliberative process of modern networks correlates with human reasoning, and that counterfactual explanations can substantially benefit applications like machine teaching.

II. Related Work

XAI for Computer Vision. Many variants of XAI have been proposed in the literature. For computer vision, explanations can be based on concepts [31], [32], [33], examples [1], [34], [35], [36], image transformations [30], [37], language [38], [39], [40], etc. Among these, the visualization of saliency maps is a widely used approach [16], [17], [41], [42], [43], which we pursue in this work. XAI methods can also be divided into two groups that depend on the design stage where predictions and explanations are performed. One possibility is to design models to be interpretable [2], [43], [44], [45], another to perform post-hoc analysis on pre-trained models [16], [17], [46]. In this paper, we mainly discuss post-hoc methods. Several survey papers [47], [48], [49], [50] provide a more comprehensive review of the field.

Attributive Explanations. The most successful post-hoc XAI approach to create saliency maps is to rely on attribution functions [8], [12], [14], [17], [51]. While many attribution functions have been proposed [8], [11], [12], [13], [17], [51], [52], the most popular approach is to compute some variant of the gradient of the classifier prediction with respect to a chosen network layer and then backproject to the input [15], [16]. Other popular attribution methods include SHAP [53], score-CAM [17], LIME [54], and RISE [55]. The proposed GALORE framework is compatible with any attribution function.

Contrastive and Counterfactual Explanations. Counterfactual visual explanations transform an image of class A so as to elicit its classification into the counter class B [38], [56], [57], [58], [59], [60]. The simplest example are adversarial attacks [23], [56], which optimize perturbations to map an image of class A into class B. However, these perturbations usually push the perturbed image outside the boundaries of the space of natural images. Generative methods have been proposed to address this problem, computing large perturbations that generate realistic images [57], [61], [62], [63]. This is guaranteed by the introduction of regularization constraints, auto-encoders, or GANs [64]. However, because realistic images are difficult to synthesize, these approaches have only been applied to simple, MNIST or CelebA [65] style, datasets and domains that do not require expertise [37], [61], [63]. StylEx [59] and C3LT [66] are two recent methods, leveraging a GAN to produce the explanations. They, however, require training on large-scale data, which is not a necessity for other methods. A more plausible alternative is to exhaustively search the space of features extracted from a large collection of images, to find replacement features that map the image from class A to B [30]. While this has been shown to perform well on fine-grained datasets, exhaustive search is too complex for interactive applications.

XAI Evaluation. Explanations are frequently evaluated though human-in-the-loop experiments that measure their consistency with human intuition [16], [23], [53], [67] or evaluate if explanations improve user performance on some task [30]. It is also possible to assemble a dataset to generate human-driven ground-truth explanations [68]. An alternative approach is automated evaluation, using a proxy task without human participation. A typical example is to erase or add features and observe how the model predictions change [69], [70], [71], [72]. Another is localization, where regions of features deemed important by the explanation are compared to regions deemed intuitive for classification by humans [16], [73]. Another component of the evaluation of explanations is to test their robustness via sanity checks [74], [75], [76], [77]. In this work, we introduce a quantitative protocol for the evaluation of both deliberative and counterfactual visual explanations, which includes sanity checks.

Self-Awareness. Self-aware systems have some ability to measure their limitations or predict failures. This includes out-of-distribution detection [78], [79], [80], [81] or open set recognition [82], [83], [84], [85], where classifiers are trained to reject non-sensical images, adversarial attacks, or images from classes on which they were not trained. All these problems require the classifier to produce a confidence score for image rejection. The most popular solution is to guarantee that the posterior class distribution is uniform, or has high entropy, outside the space covered by training images [86], [87]. This, however, is not sufficient for deliberative explanations, which have to precisely characterize the ambiguity of image regions, or counterfactual explanations, which require precise confidence scores for classes A and B. These explanations are more closely related to realistic classification [88], where a classifier must identify and reject examples that it deems too difficult to classify.

III. A Unified View of Explainable AI

In this section, we discuss the different types of explanations implemented by the proposed GALORE framework. The detailed computations required to produce the explanations are discussed in Section IV.

A. Attributive Explanations

Attributive explanations identify pixels responsible for a classifier prediction. This is intuitive but prone to generate explanations that are too generic. For example, when asked "why is an object a truck?" an attributive system would answer "because it has wheels, a hood, seats, a steering wheel, a flatbed, head and tail lights, and rearview mirrors," i.e., generate a list of all the truck parts. After all, all parts are responsible for the 'truck' label. The problem is that, while insightful, the explanation does not inform on what distinguishes the truck from, for example, a car. The explanation for 'car' would share all components other than the flatbed.

Similarly, visual attributive explanations tend to highlight all pixels of objects in the predicted class. This is sensible for coarse grained classification, e.g., 'birds' versus 'cats,' but not for fine-grained, e.g., the CUB birds dataset [89] from which the images of Figs. 1, 2 and 3 were taken. On this dataset, where most images contain a single bird, methods like Grad-CAM [16] (used in these

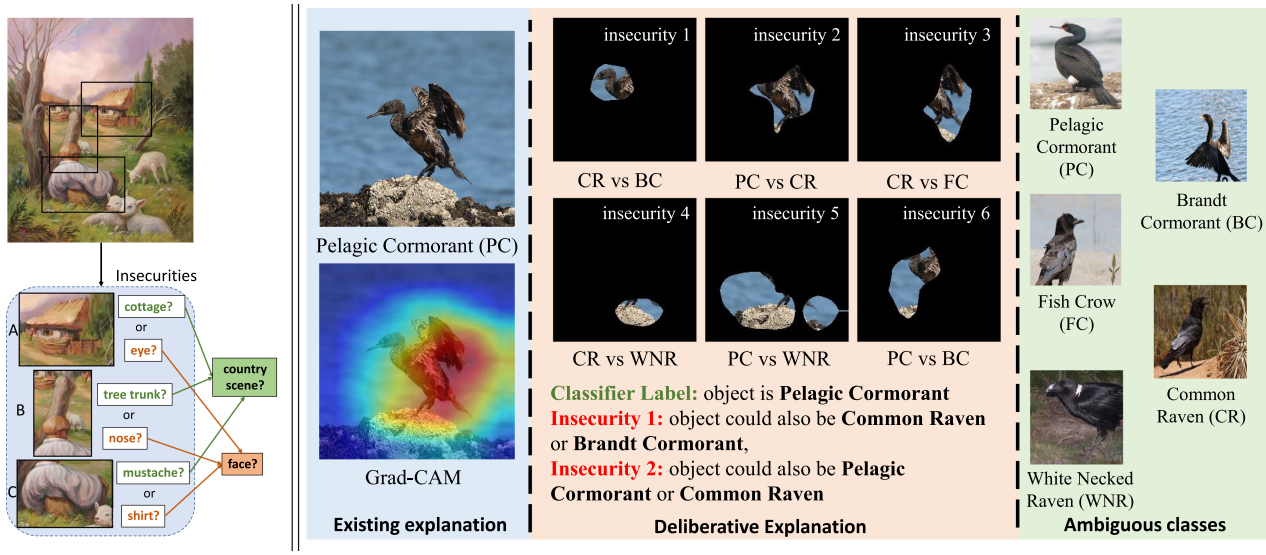

Fig. 2. Left: Illustration of the deliberations made by a human to categorize an ambiguous image. Insecurities are ambiguous regions. Right: Deliberative explanations expose this deliberative process. Unlike attributions, which simply attribute the prediction to image regions (left), they expose the insecurities experienced by the classifier while reaching that prediction (center). Each insecurity consists of an image region and an ambiguity, expressed as a pair of classes to which the region appears to belong to. Examples from the confusing classes are shown in the right.

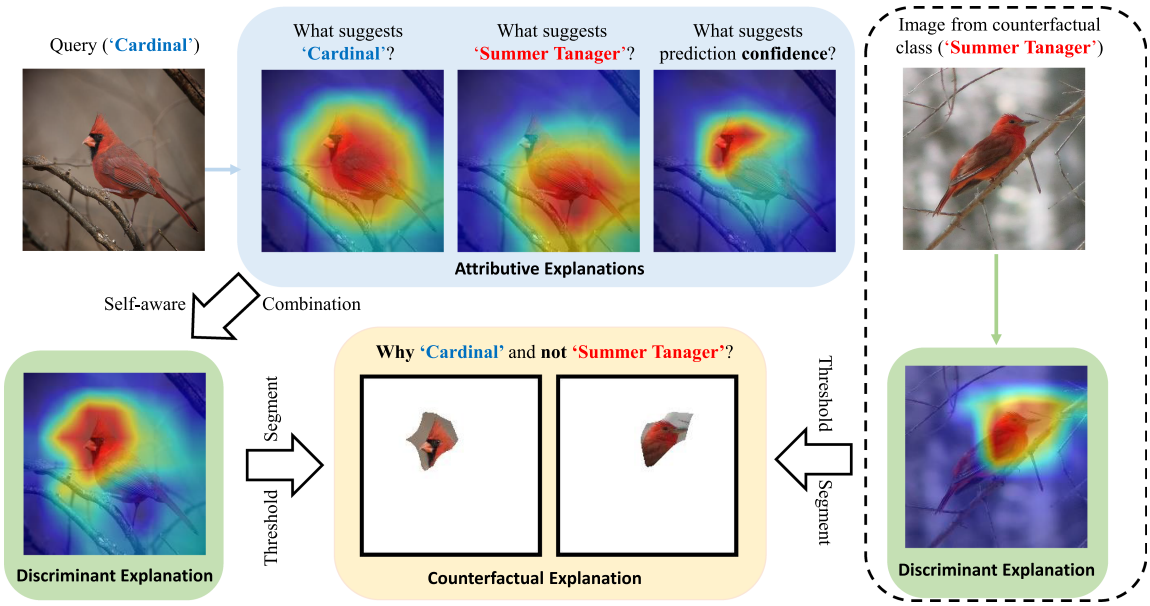

Fig. 3. A counterfactual explanation is derived from a pair of discriminant explanations. Given a query image (Cardinal) and a counterfactual class (Summer Tanager), discriminant explanations are obtained by combining attribution maps for each of the two classes and the confidence score. In this way, they bridge the gap between attributive and counterfactual explanations, enabling fast optimization-free computation of the latter.

examples) produce heatmaps that 1) cover most of the bird, and 2) vary little across classes of largest posterior probabilities, leading to very uninformative explanations. In this work, we seek better explanations for the fine-grained setting.

B. Deliberative Explanations

In this setting, visual concepts differ in subtle ways. There are frequently two or more classes of very similar appearance, and the classification can be quite ambiguous. This is illustrated in both Figs. 1 and 2, which present several similar birds, difficult to differentiate for a layperson. Due to this ambiguity, even an expert could reasonably oscillate between different interpretations while deliberating about the class to predict. An extreme example of this process are visual illusions such as that depicted in the left of Fig. 2, where different image regions provide support for conflicting image interpretations. In this example, the image could depict a 'country scene' or a 'face.' Most humans would consider the two interpretations while deliberating on a final prediction. When asked to explain the latter, they would say something like: "I see a cottage in region A, but region B could be a tree trunk or a nose, and region C looks like a mustache, but

could also be a shirt. Since there are sheep in the background, I am going with country scene." More generally, different regions can provide evidence for two or more distinct predictions and there may be a need to deliberate between multiple classes.

Having access to this deliberative process is important to trust an AI system. For example, in medical diagnosis, a single prediction can appear unintuitive to a doctor, even if accompanied by a heatmap. The doctor's natural reaction would be to ask "why did you reach that conclusion?" Ideally, instead of simply outputting a predicted label and a heat map, the AI system should visualize its *deliberations,* producing a list of image regions that support other plausible predictions. For example, when categorizing medical images with respect to interstitial lung diseases [90], [91], the AI system should explain a prediction of 'emphysema' by highlighting the regions of greatest uncertainty between this and alternative predictions, such as 'normal' or 'fibrosis'. We denote these regions as *insecurities*, since they cast doubt on the validity of the predicted label. To accomplish this, we propose a new type of explanations based on heatmaps of network insecurities. These are denoted as *deliberative explanations,* since they visualize the network deliberations.

As illustrated in the right of Fig. 2, the deliberative explanation provides a list of insecurities (center inset), each consisting of 1) an image region and 2) an *ambiguity*, formed by the pair of classes that led the network to be uncertain about the region. Example images from the ambiguous classes can also be displayed, as shown in the right inset. For example, the first insecurity of Fig. 2 reflects the fact that the head of the Pelagic Cormorant is similar to those of the Brandt Cormorant and the Common Raven. Hence, this region raises uncertainty about the 'Pelagic Cormorant' label predicted by the classifier. The detailed implementation of deliberative explanations is discussed in Section IV-D.

C. Counterfactual Explanations

Returning to the 'truck' example, domain experts will likely not be satisfied by the simply listing of all truck parts. Instead, they are likely to request more *precise* explanations, for instance asking the question "Why is it a truck and not a car?" The answer "because it has a flatbed. If it did not have a flatbed it would be a car," is known as a *counterfactual explanation* [23], [30], [38], [92]. Counterfactual explanations, by supporting a specific query with respect to a *counterfactual* class (B), allow expert users to zero-in on a specific ambiguity between two classes, which they already *know* to be plausible predictions. Unlike attributions, these explanations scale naturally with user expertise. As the latter increases, the class and counterfactual class simply become more *fine-grained.* In computer vision, counterfactual explanations are usually implemented as "correct class is A. Class B would require changing the image as follows," where "as follows" is some visual transformation. Possible transformations include image perturbations akin to those used in adversarial attacks [23], image synthesis [37], [60], or replacing image regions by regions of some images in the counter class B, found by the exhaustive search of a large feature pool [30]. However, image perturbations and synthesis frequently leave the space of natural images, only working on simple non-expert domains, and feature search is too complex for interactive applications.

In this work, we propose the computation of counterfactual explanations by a simple and robust procedure, based on attributions. We start by introducing *discriminant explanations* that, as shown in Fig. 3, connect attributive to counterfactual explanations. Like attributive explanations, they consist of a single heatmap. This, however, is an attribution map for the *discrimination* of classes A and B, attributing high scores to image regions that are informative of A but not of B, and high classification confidence, indicating that the discrimination between the two classes is clear and easy to identify. The detailed generation of discriminant explanations is discussed in Section IV-E. The final *counterfactual explanation* is then composed by two discriminant explanations, with the roles of A and B reversed. It identifies the image regions informative of A but not B and the regions informative of B but not A.

As illustrated in Figs. 1 and 3, the presentation of these regions side by side allows the user to visualize how the image of A would need to be changed in order to be classified as B (and vice-versa). This shows that counterfactual explanations can be seen as a *generalization* of attributive explanations, computed by a *combination* of attribution and confidence prediction methods that is much more efficient to compute than previous methods. In fact, our experiments show that their computation is $50\times$ to $1000\times$ faster for popular networks. This is quite important for applications such as machine teaching, where explanation algorithms should operate in real-time, ideally in low-complexity platforms such as mobile devices.

IV. Implementation of GALORE

In this section, we discuss a unified framework for implementation of the explanations discussed above.

A. Explanation Framework

Consider an object recognition system $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$, mapping images $\mathbf{x} \in \mathcal{X}$ into classes $y \in \mathcal{Y} = \{1, \dots, C\}$, according to a classifier

$$y^* = \arg\max_y h_y(\mathbf{x}), \tag{1}$$

where $\mathbf{h}(\mathbf{x}) : \mathcal{X} \to [0,1]^C$ is a C-dimensional probability distribution with $\sum_{y=1}^{C} h_y(\mathbf{x}) = 1$, usually computed by a convolutional neural network (CNN). The classifier is denoted self-aware if it produces a *confidence scores*$(\mathbf{x}) \in [0,1]$, encoding the strength of its belief that the image $\mathbf{x}$ belongs to the predicted class y^*. The confidence score can be generated by the classifier itself, in which case it is denoted as *self-referential*, or by a complementary network, in which case it is *non-self-referential*. Both the classifier and the confidence score generator are learned from a training set $\mathcal{D}$ of $N^{\mathcal{D}}$ i.i.d. samples $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N^{\mathcal{D}}}$, where $y_n \in \mathcal{Y}$ is the label of image $\mathbf{x}_n \in \mathcal{X}$. Classification performance is evaluated on a disjoint test set $\mathcal{T} = \{(\mathbf{x}_m, y_m)\}_{m=1}^{M^{\mathcal{T}}}$.

TABLE I
IMPLEMENTATION OF DIFFERENT EXPLANATION STRATEGIES UNDER THE GALORE FRAMEWORK

explanation	heatmap	$m^\alpha(\mathbf{a})$	$m^\beta(\mathbf{a})$	$m^\gamma(\mathbf{a})$	$\mathcal{C}$	$s(\mathbf{x})$
Attributive	$\mathcal{A}(\mathbf{x}, y^*)$	$\mathbf{a}$	1	1	None	None
Self-aware attributive	$\mathcal{A}(\mathbf{x}, y^*)$	$\mathbf{a}$	1	$\mathbf{a}$	None	$s(\mathbf{x})$
Deliberative	$\mathcal{I}(\mathbf{x}, \mathcal{C})$	1	$\mathbf{a}$	$\mathbf{a}$	$\{a, b\}$	$s(\mathbf{x}) \leftarrow 1 - s(\mathbf{x})$
Counterfactual	$\mathcal{R}(\mathbf{x}, y^*, y^c, \mathbf{x}^c)$	$\mathbf{a}$	$\max_{i,j} \mathbf{a}_{i,j} - \mathbf{a}_{i,j}$	$\mathbf{a}$	$\{y^c\}$	$s(\mathbf{x})$
Multiclass deliberative	$\mathcal{I}(\mathbf{x}, \mathcal{C})$	1	$\mathbf{a}$	$\mathbf{a}$	$\{a_1, ...a_V\}$	$s(\mathbf{x}) \leftarrow 1 - s(\mathbf{x})$
Multiclass counterfactual	$\mathcal{R}(\mathbf{x}, y^*, \mathcal{C}, \bigcup_{v=1}^{V} \mathbf{x}^v)$	$\mathbf{a}$	$\max_{i,j} \mathbf{a}_{i,j} - \mathbf{a}_{i,j}$	$\mathbf{a}$	$\{y^1, ...y^V\}$	$s(\mathbf{x})$

In this work, we propose a *GenerAlized expLanatiOn fRamEwork* (GALORE) to unify various visualization-based explanations, accounting for both confidence scores and a set $\mathcal{C}$ of class labels of interest beyond the prediction y^*. All GALORE explanations are implemented with a heat map

$$\mathcal{M}_{i,j}(\mathbf{x}, h_{y^*}, \mathcal{C}) = m^\alpha(\mathbf{a}_{i,j}(h_{y^*}(\mathbf{x}))) \cdot \prod_{c \in \mathcal{C}} m^\beta(\mathbf{a}_{i,j}(h_{y^c}(\mathbf{x}))) \cdot m^\gamma(\mathbf{a}_{i,j}(s(\mathbf{x}))), \tag{2}$$

where $\cdot$ denotes multiplication, $\mathbf{a}_{i,j}(.)$ is an attribution function, which measures how the spatial feature of $\mathbf{x}$ at location (i, j) contributes to a prediction. m^α, m^β and m^γ are three functions that depend on the explanation. The detailed implementation of these functions for each type of explanation is discussed in the following sections and summarized in Table I.

The definition of (2) as a multiplication of attribution maps strengthens the heat map $\mathcal{M}$ at the locations where all the attributions are large and attenuates it when at least one of them is low. This can be seen as a measure of agreement of the different attributions that drastically penalizes disagreements. In this way, only locations that receive significant attribution from the different components are identified as salient, resulting in sharp heat maps that are informative of object details, as illustrated in Figs. 2 and 3. The process can also be seen as equating attribution maps to probability density functions of independent random variables and $\mathcal{M}$ to the resulting joint distribution. While this is not exact, since the attributions of $h_{y^*}(\mathbf{x})$, $h_{y^c}(\mathbf{x})$, and $s(\mathbf{x})$ are not independent, it provides a computationally efficient approximation. Explanations are provided in the form of collections image segments [54], [93], [94] obtained by thresholding the heat map. We next discuss how (2) is used to implement different visualization strategies.

B. Attributive Explanations

Attributive explanations visualize how strongly the prediction y^* is attributed to different regions of image $\mathbf{x}$ [8], [11], [12], [13], [51]. They are obtained from (2) by setting $m^\alpha(x) = x$, $m^\beta(x) = m^\gamma(x) = 1$, leading to heat map (for brevity, we omit location subscript in the rest of the paper)

$$\mathcal{A}(\mathbf{x}, y^*) = \mathbf{a}(h_{y^*}(\mathbf{x})). \tag{3}$$

The attribution function $\mathbf{a}(.)$ is usually applied to a tensor of activations $\mathbf{F} \in \mathbb{R}^{W \times H \times D}$ of spatial dimensions $W \times H$ and D channels, extracted at some layer of a deep network with $\mathbf{x}$ at the input. While many attribution functions have been proposed, they are usually some variant of the gradient of $h_{y^*}(\mathbf{x})$ with respect to $\mathbf{F}$. This results in an attribution map where the amplitude of $\mathcal{A}_{ij}(.)$ encodes the attribution of the prediction to each entry i, j along the spatial dimensions of $\mathbf{F}$. Two attributive heatmaps of an image of a "Cardinal" with respect to predictions "Cardinal" and "Summer Tanager," are shown in the top row of Fig. 3.

C. Self-Aware Attributive Explanations

Attributive explanations can be extended to account for confidence scores by setting $m^\gamma(x) = x$. In this case, the attributive explanation becomes

$$\mathcal{A}(\mathbf{x}, y^*) = \mathbf{a}(h_{y^*}(\mathbf{x})) \cdot \mathbf{a}(s(\mathbf{x})). \tag{4}$$

GALORE is compatible with any classification confidence score $s(\mathbf{x})$. A few examples that we compare in our experiments are discussed in Section V-B. Large heatmap entries indicate regions that not only contribute to the prediction but also make the classifier confident about it. When compared to standard attributive explanations, the self-aware version emphasizes more class-specific regions. In experiments, we will see that these regions usually cover the attributes discriminant for the predicted classes, providing a sharper and more convincing explanation for the classifier prediction.

D. Deliberative Explanations

A deliberative explanation consists of a set of Q insecurities $\{(\mathbf{r}_q, a_q, b_q)\}_{q=1}^{Q}$ that provide insight on the reasoning performed by the classifier to reach prediction y^*. Each insecurity is a triplet $(\mathbf{r}, a, b)$, where $\mathbf{r}$ is the segmentation mask of a region responsible for classifier uncertainty, and (a, b) an ambiguity composed by a pair of class labels. Altogether, the insecurity shows that the network is insecure as to whether the image region defined by $\mathbf{r}$ should be attributed to class a or b. Note that none of a or b has to be the prediction y^*, although this could happen for one of them. In Fig. 2, y^* is the label "Pelagic Cormorant," and appears in insecurities 2, 5, and 6, but not on the remaining. This reflects the fact that certain parts of the bird could actually be shared by many classes.

Insecurities are generated by first identifying the set $\mathbb{C} = \{y^1, \ldots, y^E\}$ of the E classes y of largest posterior probability $h_y(\mathbf{x})$. A *candidate class ambiguity set* $\mathbb{A} = \binom{\mathbb{C}}{2}$ is then created with all class pairs in $\mathbb{C}$. For each ambiguity $(a, b) \in \mathbb{A}$, an *ambiguity map* is computed using (2) with $\mathcal{C} = \{a, b\}$, $m^\alpha(x) = 1$,

$m^\beta(x) = m^\gamma(x) = x$, and $s(\mathbf{x})$ replaced with $1 - s(\mathbf{x})$,

$$\mathcal{I}(\mathbf{x}, \mathcal{C}) = \mathbf{a}(h_a(\mathbf{x})) \cdot \mathbf{a}(h_b(\mathbf{x})) \cdot \mathbf{a}(1 - s(\mathbf{x})). \quad (5)$$

Using as self-awareness score the complement of the belief in the prediction assigns larger scores to regions where the prediction is most ambiguous, reflecting the difficulty of the classifier decision. $\mathcal{I}_{i,j}$ is large only when location (i, j) is deemed difficult to classify (large difficulty attribution $\mathbf{a}(1 - s(\mathbf{x}))_{i,j}$) *and* this difficulty is due to large attributions to *both* classes a and b. The ambiguity map is thresholded to obtain the segmentation mask

$$\mathbf{r}\{a, b\}(\mathbf{x}) = \mathbb{1}_{\mathcal{I}>T}, \quad (6)$$

where $\mathbb{1}_{\mathcal{S}}$ is the indicator function of set $\mathcal{S}$ and T a threshold. The ambiguity (a, b) and the mask $\mathbf{r}\{a, b\}(\mathbf{x})$ form an *insecurity*.

E. Counterfactual Explanations

Counterfactual explanations assume an image $\mathbf{x}$, a prediction y^*, and a user provided counterfactual class $y^c \neq y^*$. A popular approach is to highlight the differences between $\mathbf{x}$ and an image $\mathbf{x}^c$ from class y^c by displaying matched bounding boxes on the two images. [30] showed that explanation performance is nearly independent of the choice of $\mathbf{x}^c$, i.e., it suffices to use a random image $\mathbf{x}^c$ from class y^c. We adopt a similar strategy in this work, implementing counterfactual explanations as

$$\mathcal{R}(\mathbf{x}, y^*, y^c, \mathbf{x}^c) = (\mathcal{D}(\mathbf{x}, y^*, y^c), \mathcal{D}(\mathbf{x}^c, y^c, y^*)), \quad (7)$$

where $\mathcal{D}(\mathbf{x}, y^*, y^c)$ and $\mathcal{D}(\mathbf{x}^c, y^c, y^*)$ are *discriminant heatmaps* for images $\mathbf{x}$ and $\mathbf{x}^c$, respectively. The first map identifies the regions of $\mathbf{x}$ that are informative of the predicted class but not the counter class while the second identifies the regions of $\mathbf{x}^c$ informative of the counter class but not of the predicted class. Altogether, the explanation shows that the regions highlighted in the two images are matched: the region of the first image depicts features that *only* appear in the predicted class while that of the second depicts features that *only* appear in the counterfactual class. The discriminant map of $\mathbf{x}$ is thresholded to obtain the segmentation mask

$$\mathbf{r}\{y^*, y^c\}(\mathbf{x}) = \mathbb{1}_{\mathcal{D}(\mathbf{x}, y^*, y^c)>T}. \quad (8)$$

Similarly, a segmentation mask is generated for $\mathbf{x}^c$ using

$$\mathbf{r}\{y^c, y^*\}(\mathbf{x}^c) = \mathbb{1}_{\mathcal{D}(\mathbf{x}^c, y^c, y^*)>T}. \quad (9)$$

Fig. 3 illustrates the construction of a counterfactual explanation with two discriminant explanations.

To compute the heatmaps of (7), [30] proposed to exhaustively compare all combinations of features in $\mathbf{x}$ and $\mathbf{x}^c$, which is expensive. We propose a much simpler and more effective procedure that leverages a new class of attributive explanations, denoted as *discriminant* and defined as in (2), with $m^\alpha(x) = m^\gamma(x) = x$, $\mathcal{C} = \{y^c\}$, and $m^\beta(\mathbf{a}(.))$ the complement of $\mathbf{a}(.)$. i.e.,

$$m^\beta(\mathbf{a}(.))_{i,j} = \max_{i,j} \mathbf{a}_{i,j} - \mathbf{a}_{i,j}, \quad (10)$$

leading to heatmap

$$\mathcal{D}(\mathbf{x}, y^*, y^c) = \mathbf{a}(h_{y^*}(\mathbf{x})) \cdot m^\beta(\mathbf{a}(h_{y^c}(\mathbf{x}))) \cdot \mathbf{a}(s(\mathbf{x})). \quad (11)$$

This is large only at locations (i, j) that contribute strongly to the prediction of class y^* but little to that of class y^c, and where the discrimination between the two classes is easy, i.e., the classifier is confident. This, in turn, implies that location (i, j) is strongly specific to class y^* but not specific to class y^c, which is the essence of the counterfactual explanation.

Discriminant explanations have commonalities with both attributive and counterfactual explanations. Like counterfactual explanations, they consider both the prediction y^* and counterfactual class y^c. Like attributive explanations, they compute a single attribution map $\mathcal{D}$. The difference is that this map *attributes the discrimination between the prediction y^* and counter y^c class* to regions of $\mathbf{x}$, identifying pixels strongly informative of class y^* but uninformative of class y^c. Fig. 3 shows how these explanations benefit from the fact that the self-awareness attribution map is usually much sharper than the other two maps. This is critical to identify the object details that differentiate the two classes.

F. Multi-Class Extensions

So far, we considered explanations involving single classes or class pairs. More generally, explanations may require, or benefit from, considering multiple classes. For example, deliberative explanations may involve ambiguities between several classes, such as a region compatible with the "Brandt Cormorant," "Fish Crow" and "Common Raven" classes in Fig. 2. In the extreme, the class posterior distribution $\mathbf{h}(\mathbf{x})$ could be approximately uniform for certain image regions. Similarly, for counterfactual explanations, a user could have more than a single counterfactual class in mind. We now consider the multi-class extension of GALORE, for both deliberative and counterfactual explanations. We define the *dimension of ambiguity* V as the number of classes involved.

For deliberative explanations of dimension V, the *candidate class ambiguity set* is first assembled by finding all class V-tuples $\mathbb{A} = \binom{\mathbb{C}}{V}$ in the candidate class list $\mathbb{C}$. This is illustrated in Fig. 4, where $V = 3$, $\mathbb{C}$ contains the five classes shown on the left (green) and the set $\mathbb{A}$ includes the five ambiguities composed by 3-tuples of these classes, as shown in the right. For each ambiguity $(a_1, a_2, \ldots, a_V)$ in $\mathbb{A}$, an *ambiguity map* is then computed using (2) with $\mathcal{C} = \{a_1, a_2, \ldots, a_V\}$, $m^\alpha(x) = 1$, $m^\beta(x) = m^\gamma(x) = x$, and $s(\mathbf{x})$ replaced by $1 - s(\mathbf{x})$, i.e.,

$$\mathcal{I}(\mathbf{x}, \mathcal{C}) = \prod_{v=1}^{V} \mathbf{a}(h_{a_v}(\mathbf{x})) \cdot \mathbf{a}(1 - s(\mathbf{x})). \quad (12)$$

This leads to large $\mathcal{I}_{i,j}$ only when location (i, j) has strong attributions for all classes in $\mathcal{C}$ and is deemed difficult to classify by the self-awareness predictor. The thresholding of (6) is finally used to create a segmentation mask.

Counterfactual explanations of dimension V and counterfactual class set $\mathcal{C} = \{y^1, \ldots, y^V\}$ are implemented as

$$\mathcal{R}(\mathbf{x}, y^*, \mathcal{C}, \bigcup_{v=1}^{V} \mathbf{x}^v) = (\mathcal{D}(\mathbf{x}, y^*, \mathcal{C}), \oplus_{v=1}^{V} \mathcal{D}(\mathbf{x}^v, y^v, \mathcal{C}'_v)), \quad (13)$$

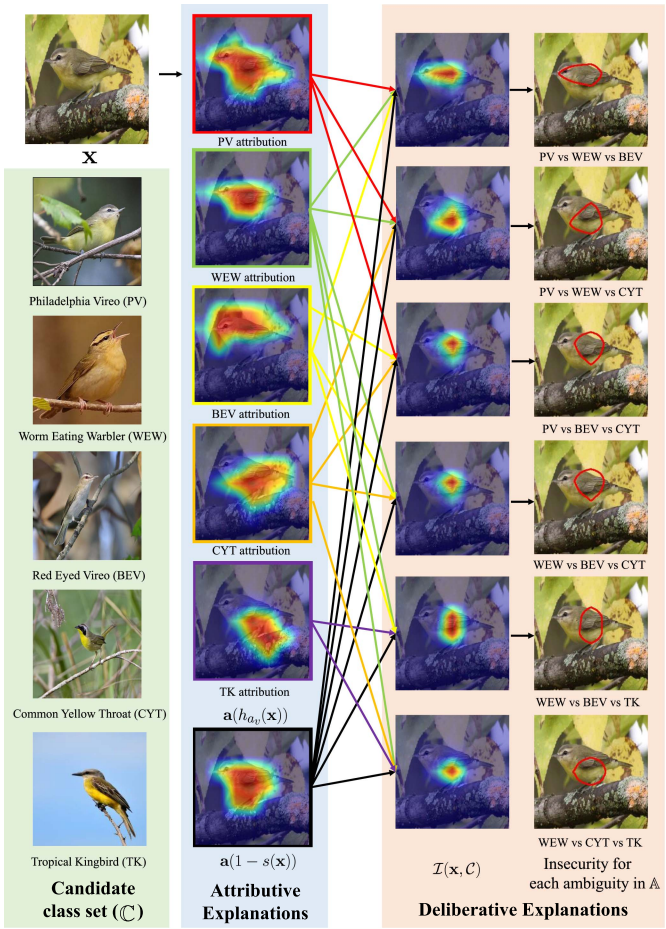

Fig. 4. Multiclass deliberative explanation of dimension $V = 3$. Left (top and green): image $\mathbf{x}$ and candidate class set $\mathbb{C} = \{$'Philadelphia Vireo' (PV), 'Worm Eating Warbler' (WEW), 'Red Eyed Vireo' (BEV), 'Common Yellow Throat' (CYT)), 'Tropical Kingbird (TK)'$\}$. Center: attributions to each of the classes in $\mathbb{C}$. Right: ambiguity map computed for each 3-tuples ambiguous classes in the candidate class ambiguity set $\mathbb{A}$ with (12) and resulting segmentation. For brevity, only six of the ten 3-tuples are randomly selected for display in the figure.

where $\mathcal{D}(\mathbf{x}, y^*, \mathcal{C})$ is the discriminant explanation for counterfactual class set $\mathcal{C}$, $\mathcal{C}'_v = \mathcal{C} \setminus \{y^v\} \cup \{y^*\}$, and $\oplus_v$ represents the side-by-side concatenation of explanations, as illustrated in Fig. 5. Similarly to (11), discriminant explanations are heat maps computed using (2) with the $m^\alpha(\cdot), m^\beta(\cdot)$, and $m^\gamma(\cdot)$ definitions of (11), i.e.,

$$\mathcal{D}(\mathbf{x}, y^*, \mathcal{C}) = \mathbf{a}(h_{y^*}(\mathbf{x})) \cdot \prod_{c \in \mathcal{C}} m^\beta(\mathbf{a}(h_{y^v}(\mathbf{x}))) \cdot \mathbf{a}(s(\mathbf{x})). \tag{14}$$

As shown in the top row of Fig. 5, attributions are first computed with respect to the prediction $h_{y^*}(\mathbf{x})$, the predictions $h_{y^v}(\mathbf{x})$ of all the other classes $y^v \in \mathcal{C}$, and the confidence score $s(\mathbf{x})$, for image $\mathbf{x}$. This is then repeated for images $\mathbf{x}^v$, replacing $\mathbf{x}$ by each $\mathbf{x}^v$, as shown in the remaining rows. The discriminant maps $\mathcal{D}(\mathbf{x}, y^*, \mathcal{C})$ and $\mathcal{D}(\mathbf{x}, y^v, \mathcal{C}'_v)$ are then computed with (14), as shown in the green box. These maps emphasize regions that are predictive of class y^* but unpredictive of all other classes in $\mathcal{C}$, highlighting the class-specific features of y^* that are discriminant with regard to $\mathcal{C}$. The explanations are finally thresholded using (8) and (9) to obtain $\mathbf{r}\{y^*, \mathcal{C}\}(\mathbf{x})$ and $\mathbf{r}\{y^v, \mathcal{C}'_v\}(\mathbf{x})$ for $\forall v \in \{1, \ldots, V\}$.

G. Explanation Strength

The clarity of explanations that involve several regions and several classes, such as deliberative or counterfactual, can benefit from a quantitative score, which we denote as the *explanation strength*, summarizing the relative importance of the different components. For example, ordering insecurities by degree of ambiguity helps guide user attention to the most important ones. To allow this type of manipulation, we define the strength of insecurity $\mathbf{r}$[2] as the average intensity of the ambiguity map of (5) or (12) within the associated image segment

$$\rho(\mathbf{r}) = \frac{1}{|\mathbf{r}|} \sum_{(x,y) \in \mathbf{r}} \mathcal{I}_{x,y}. \tag{15}$$

Similarly, adding strengths to counterfactual explanations informs how much the explanation differentiates the prediction from each counter class. We define the strength of explanation $\mathcal{R}(\mathbf{x}, y^*, y^c, \mathbf{x}^c)$ as

$$\begin{aligned} &\rho(\mathcal{R}(\mathbf{x}, y^*, y^c, \mathbf{x}^c)) \\ &= \frac{1}{2|\mathbf{r}\{y^*, y^c\}(\mathbf{x})|} \sum_{(x,y) \in \mathbf{r}\{y^*, y^c\}(\mathbf{x})} \mathcal{D}_{x,y}(\mathbf{x}, y^*, y^c) \\ &\quad + \frac{1}{2|\mathbf{r}\{y^c, y^*\}(\mathbf{x})|} \sum_{(x,y) \in \mathbf{r}\{y^c, y^*\}(\mathbf{x})} \mathcal{D}_{x,y}(\mathbf{x}^c, y^c, y^*) \end{aligned} \tag{16}$$

Note that we make sure the segment size of two discriminant explanations are equal, i.e., $|\mathbf{r}\{y^*, y^c\}(\mathbf{x})| = |\mathbf{r}\{y^c, y^*\}(\mathbf{x})|$, by tuning the thresholds T in (8) and (9). This follows [30], and works well when the objects have roughly the same size in each image, which is the case for the datasets that we consider in our experiments. A different strategy may be needed in other cases. We leave the optimal threshold tuning strategy as a topic for future research. Similarly, multi-class counterfactual explanations have strength

$$\begin{aligned} &\rho(\mathcal{R}(\mathbf{x}, y^*, \mathcal{C}, \bigcup_{v=1}^{V} \mathbf{x}^v)) \\ &= \frac{1}{(V+1)|\mathbf{r}|} \sum_{(x,y) \in \mathbf{r}\{y^*, \mathcal{C}\}(\mathbf{x})} \mathcal{D}_{x,y}(\mathbf{x}, y^*, \mathcal{C}) \\ &\quad + \sum_v \frac{1}{(V+1)|\mathbf{r}|} \sum_{(x,y) \in \mathbf{r}\{y^v, \mathcal{C}'_v\}(\mathbf{x})} \mathcal{D}_{x,y}(\mathbf{x}^v, y^v, \mathcal{C}'_v). \end{aligned} \tag{17}$$

V. Implementation

Table I summarizes how GALORE produces different visualization-based explanations, including different types of attributive, deliberative, and counterfactual explanations. All

2.Here we omit ambiguity (a, b) or $\mathcal{C}$ for brevity.

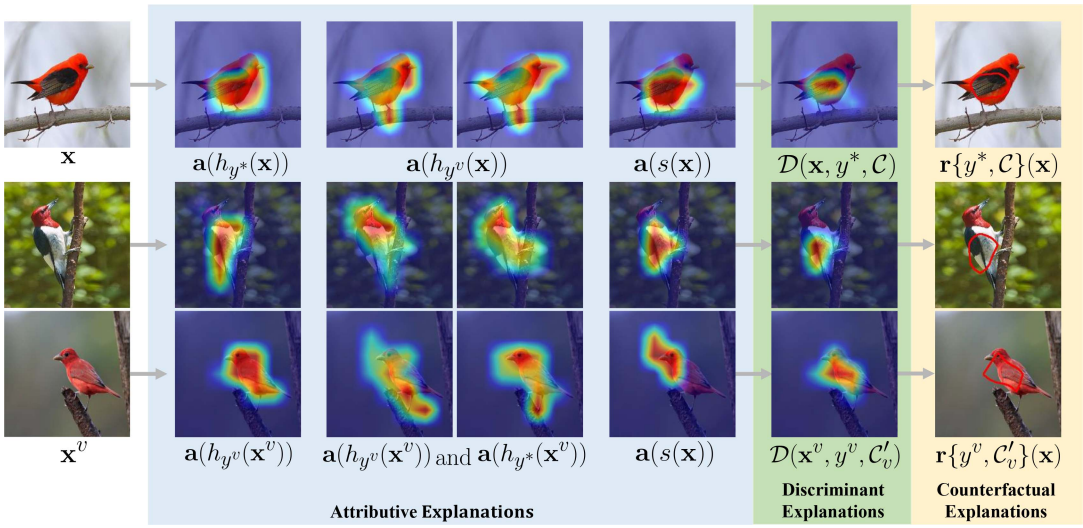

Fig. 5. Multi-class counterfactual explanation of dimension $V = 2$. Left: query image of a 'Scarlet Tanager' (upper left) and two images randomly selected from the counterfactual class set $\mathcal{C} = \{$'Red Headed Woodpecker,' 'Summer Tanager'$\}$. Middle-left: attributive maps are computed for the query and each of counter images, with respect to class predictions h_{y^*} and $h_{y^v}, y^v \in \mathcal{C}$ and confidence predictor s. Middle-right: attributive maps are combined with (14) to generate discriminant explanations. Right: discriminant explanations are thresholded to generate a multi-class counterfactual explanation.

explanations are obtained by combinations of attribution maps and classification confidence scores using (2). In this section, we discuss how these are computed.

A. Attribution Maps

Given a feature tensor $\mathbf{F}(\mathbf{x})$ in some deep network layer, attribution map $\mathbf{a}_{i,j}(h_y(\mathbf{x}))$ quantifies how the activations $\mathbf{F}_{i,j}(\mathbf{x})$ at locations (i, j) contribute to prediction y. This could be either a class prediction or the prediction of a confidence score. In this section, we make no distinction between the two, simply denoting $p(\mathbf{x}) = g_p(\mathbf{F}(\mathbf{x}))$, where g is the mapping from activation tensor $\mathbf{F}$ into prediction vector $g(\mathbf{F}) \in [0, 1]^P$. For class predictions $P = C$, the prediction p is a class y, and $g_p(\mathbf{F}(\mathbf{x})) = h_y(\mathbf{x})$. For confidence predictions $P = 1$, the prediction is a confidence score, and $g_p(\mathbf{F}(\mathbf{x})) = s(\mathbf{x})$.

GALORE is compatible with any attribution function in the literature [8], [11], [16], [17], [41], [51]. One of the most popular class of such functions is that of gradient-based attributions [11], [16], [51], which are derived from $\nabla g_p(\mathbf{F}(\mathbf{x}))$ and $\mathbf{F}(\mathbf{x})$, i.e., have the form $q([\nabla g_p(\mathbf{F}(\mathbf{x}))]_{i,j}, \mathbf{F}_{i,j}(\mathbf{x}))$ for some function q. Our implementation uses the vanilla gradient based function of [11], which computes the dot-product of the partial derivatives of prediction p with respect to activations $\mathbf{F}(\mathbf{x})$ by these activations,

$$\mathbf{a}^p_{i,j} = [\nabla g_p(\mathbf{F})]^T_{i,j} \mathbf{F}_{i,j}. \tag{18}$$

Here we omit the dependency on $\mathbf{x}$ for simplicity.

This is compared to two more complex attribution functions, integrated gradient (InteGrad) [51] and GradCAM [16]. InteGrad is based on the Riemman approximation of the integral of the gradient ∇g_p along a linear path from a reference $\mathbf{F}^0$ to the observed activation tensor $\mathbf{F}$,

$$\mathbf{a}^p_{i,j} = \left(\sum_{k=1}^{\Omega} [\nabla g_p(\mathbf{F})\Big|_{\mathbf{F}^0 + \frac{k}{\Omega} \times (\mathbf{F} - \mathbf{F}^0)}]_{i,j} \cdot \frac{1}{\Omega} \right)^T (\mathbf{F}_{i,j} - \mathbf{F}^0_{i,j}), \tag{19}$$

where Ω is the number of steps in the approximation and set to 50. The reference $\mathbf{F}^0$ is defined by the user and often chosen to be the image that induces zero activation. Unlike (18), which only uses the partial derivative at activation $\mathbf{F}_{i,j}(\mathbf{x})$, InteGrad computes the average gradient along the linear path from $\mathbf{F}^0$ to $\mathbf{F}$. Grad-CAM [16] assigns a unique weight per activation channel k, which is the spatial mean of the activations of this channel

$$\mathbf{a}^p_{i,j} = \text{ReLU}\left(\sum_k w_k \mathbf{F}_{i,j,k} \right), \tag{20}$$

where $w_k = \frac{1}{W \times H} \sum_{i,j} \frac{\partial g_p(\mathbf{F})}{\partial \mathbf{F}_{i,j,k}}$. In our implementation, the attribution maps of (18), (19), (20) are normalized to [0,1] by min-max normalization, i.e., subtracting the minimum value and dividing by the maximum.

GALORE is also compatible with non gradient-based attribution functions [17], [53], [54], [55]. In experiments, we present results for score-CAM [17] and SHAP [53], two representatives of these methods. Like Grad-CAM, the attribution map of score-CAM is a weighted sum of channel activation maps but the weight w_k of (20) is not derived from gradients, involving forward computations only. SHAP quantifies the element strength of an attribution map by its Shapley value. We omit the details for brevity.

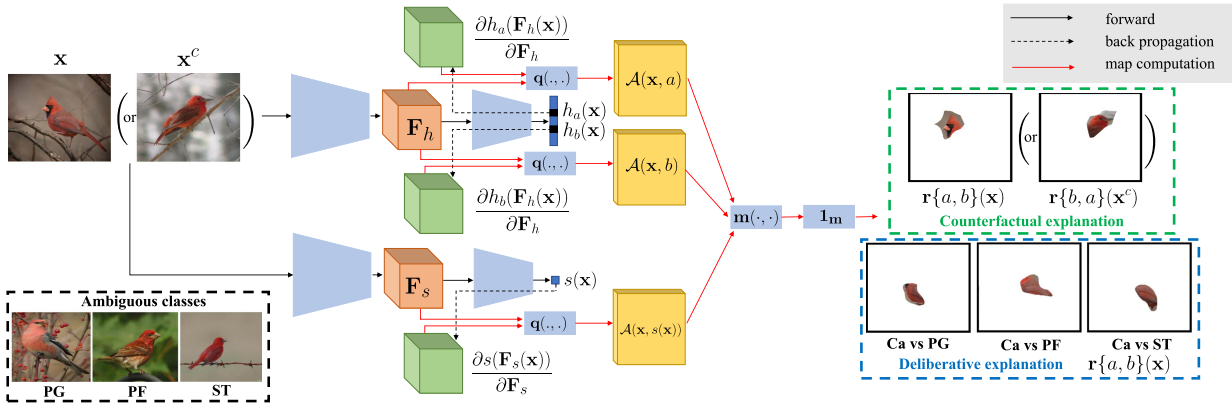

Fig. 6. GALORE explanation architecture ($\mathbf{x}$: Cardinal, $\mathbf{x}^c$: Summer Tanager.). Feature activations $\mathbf{F}_h$ and $\mathbf{F}_s$ are computed for pre-determined layers of the classifier (upper branch) and confidence predictor (lower branch), respectively. Attributions for prediction h_a, ambiguous or counter class h_b, and confidence score s are computed by attribution functions $q(.,.)$ according to (18), (19), (20). a, b are a class pair of candidate class ambiguities set for deliberative explanations, and y^*, y^c for counterfactual explanations. These attributions are combined with (5) or (11) to obtain the final map, which is thresholded to produce explanations. Multiple pairs (a, b) are shown for deliberative explanations, where a is Cardinal (Ca), and b Pine Grosbeak (PG), Purple Finch (PF) or Summer Tanager (ST). Counterfactual explanations are obtained by additionally reversing the roles of $\mathbf{x}$ and $\mathbf{x}^c$ and thresholding the discriminant heat maps.

B. Confidence Scores

Beyond attribution maps, GALORE is compatible with many classification confidence scores. We consider three scores of different characteristics. The *softmax score* [28] is the largest class posterior probability

$$s^s(\mathbf{x}) = \max_y h_y(\mathbf{x}). \tag{21}$$

It is computed by adding a max pooling layer to the network output. The *certainty score* is the complement of the normalized entropy of the softmax distribution [29],

$$s^c(\mathbf{x}) = 1 + \frac{1}{\log C} \sum_y h_y(\mathbf{x}) \log h_y(\mathbf{x}). \tag{22}$$

Its computation requires an additional layer of log non-linearities and average pooling. These two scores are self-referential. We also consider the non-self-referential *easiness score* of [88],

$$s^e(\mathbf{x}) = 1 - s^{hp}(\mathbf{x}) \tag{23}$$

where $s^{hp}(\mathbf{x})$ is computed by an external predictor $\mathcal{S}$, which predicts the difficulty of classifying each example and is trained jointly with the classifier. $\mathcal{S}$ is implemented by a network $s^{hp}(\mathbf{x}) : \mathcal{X} \to [0, 1]$ whose output is a sigmoid unit.

C. Network Implementation

Fig. 6 shows a network implementation of (2). Given a query image $\mathbf{x}$ of class y^*, a user-selected counter class $y^c \neq y^*$, a predictor $h_y(\mathbf{x})$, and a confidence predictor $s(\mathbf{x})$ are used to produce the explanation. Note that $s(\mathbf{x})$ can share weights with $h_y(\mathbf{x})$ (self-referential) or be separate (non-self-referential). $\mathbf{x}$ is forwarded through the network, generating activation tensors $\mathbf{F}_h(\mathbf{x})$, $\mathbf{F}_s(\mathbf{x})$ in pre-chosen network layers and predictions $h_a(\mathbf{x})$, $h_b(\mathbf{x})$, $s(\mathbf{x})$, which depend on the explanation strategy. For deliberative explanations, the predictions are classes a, b from the candidate ambiguities set. For counterfactual explanations, they are $h_{y^*}(\mathbf{x})$, $h_{y^c}(\mathbf{x})$, $s(\mathbf{x})$. The attributions of a, b and $s(\mathbf{x})$ to $\mathbf{x}$, i.e., $\mathcal{A}(\mathbf{x}, a)$, $\mathcal{A}(\mathbf{x}, b)$, $\mathcal{A}(\mathbf{x}, s(\mathbf{x}))$ are then computed with (18), (19), or (20), which reduce to a back-propagation step with respect to the desired layer activations and a few additional operations. These attributions can also be computed by other non-gradient-based functions. Finally, the attributions are combined with (5) or (11). Thresholding the resulting heatmap with (6) or (8) produces the deliberative explanation $\mathbf{r}\{a, b\}(\mathbf{x})$ or discriminant explanation $\mathbf{r}\{y^*, y^c\}(\mathbf{x})$. For counterfactual explanations, the network is simply applied to $\mathbf{x}^c$ to compute $\mathbf{r}\{y^c, y^*\}(\mathbf{x}^c)$. Multi-class deliberative extensions simply require a larger set of classes and replace (5) by (12). For multi-class counterfactual explanations, (11) is replaced by (14) and the process repeated for each counterfactual image $\mathbf{x}^v$.

VI. Evaluation

Explanations can be difficult to evaluate, since ground truth is usually not available. Two major classes of evaluation strategies have been proposed.

A. User Experiments

One possibility is to perform Turk experiments, e.g., measuring whether humans can predict a class label given a visualization, or identify the most trustworthy of two models that make identical predictions from their explanations [16]. We use a similar strategy for deliberative explanations, by measuring whether, given an insecurity produced by the explanation algorithm, humans can predict the associated ambiguities. For counterfactual explanations, we use instead a machine teaching setting, testing whether the explanation helps humans distinguish different classes. While these strategies directly measure how intuitive the explanations appear to humans, they require subject experiments that are somewhat cumbersome to perform and difficult to replicate.

B. Proxy Tasks

A second evaluation strategy uses a proxy task, such as localization [15], [16] on datasets with object bounding boxes. While

this is much easier to implement, there is usually no groundtruth for regions of importantance to the classification of an image. We overcome this problem by leveraging datasets annotated[3] with parts and attributes. Specifically, where the k^{th} part of an object of class c is annotated with a semantic descriptor ϕ_c^k containing the attributes present in this class. For example, in a bird dataset, the "eye" part can have color attribute values "green," "blue," "brown," etc. The descriptor is a probability distribution over these values, characterizing the variability of attribute values of the part per class. Explanation ground-truth is derived from attribute distributions, as described next.

1) Deliberative Explanations: For deliberative explanations, we define insecurities as *ambiguous parts,* namely object parts common to multiple object classes or scene parts (e.g., objects) shared by scene classes. This reduces evaluation to insecurity localization.

For binary explanations, the similarity between classes a and b according to part k is defined as $\alpha_{a,b}^k = \gamma(\phi_a^k, \phi_b^k)$, where γ is a dataset dependent similarity measure. This reflects the *strength of the ambiguity* between classes a and b, declaring as ambiguous parts that have similar attribute distributions under the two classes. To generate ground-truth, the values of $\alpha_{a,b}^k$ are computed for all parts $\mathbf{p}_k$ and class pairs (a, b). The M triplets $\mathcal{G}^d = \{(\mathbf{p}_i, a_i, b_i)\}_{i=1}^M$ of largest similarity in $\mathcal{G} = \{(\mathbf{p}_i, a_i, b_i)|a_i \neq b_i\}_{i=1}^{C*C*K}$ are selected as insecurity ground-truth, where K is the total number of parts. For multi-class explanations, given an ambiguity class set $\mathcal{V} = \{a_1, \ldots a_V\}$, the similarity of the V classes, according to part k, is defined as $\alpha_{\mathcal{V}}^k = \eta(\phi_{a_1}^k, \ldots, \phi_{a_V}^k)$, where η is a dataset-dependent function. The similarities $\alpha_{\mathcal{V}}^k$ are computed for all $\mathbf{p}_k$ and $\mathcal{V}$, and the M tuples $\mathcal{G}^m = \{(\mathbf{p}_i, \mathcal{V})\}_{i=1}^M$ of largest similarity selected as insecurity ground-truth.

Given this groundtruth, two metrics are used to evaluate the quality of the explanations, depending on the nature of part annotations. For datasets where parts are labelled with a single location (usually the geometric center of the part), i.e., $\mathbf{p}_i$ is a point, the quality of segment $\mathbf{r}\{a, b\}(\mathbf{x})$ is computed by precision (P) and recall (R). Here, $P = \frac{J}{|\{k|\mathbf{p}_k \in \mathbf{r}\}|}$, $R = \frac{J}{|\{i|(\mathbf{p}_i, a_i, b_i) \in \mathcal{G}, a_i = a, b_i = b\}|}$ and $J = |\{i|\mathbf{p}_i \in \mathbf{r}, a_i = a, b_i = b\}|$ is the number of ground-truth parts included in the insecurities that compose the explanation. Precision-recall curves are produced by varing the threshold T of (6). For datasets where parts have segmentation masks, the quality of $\mathbf{r}\{a, b\}(\mathbf{x})$ is computed by the intersection over union (IoU) metric IoU $= \frac{|\mathbf{r} \cap \mathbf{p}|}{|\mathbf{r} \cup \mathbf{p}|}$, where $\mathbf{p} = \{\mathbf{p}_i|(\mathbf{p}_i, a_i, b_i) \in \mathcal{G}^d, a_i = a, b_i = b\}$.

2) Counterfactual Explanations: For counterfactual explanations, where the goal is to localize a region predictive of class A but unpredictive of class B, groundtruth is assembled by identifying parts with attributes specific to A that do not appear in B. This enables the evaluation of counterfactual explanations as a class-specific part localization problem.

For two-class explanations, where $\alpha_{a,b}^k$ measures the similarity between two classes according to part k, a small $\alpha_{a,b}^k$ indicates that part k discriminates between the two classes. To generate ground-truth, the N parts of smallest similarity in $\mathcal{G}$, $\mathcal{G}^c = \{(\mathbf{p}_i, a_i, b_i)\}_{i=1}^N$ are selected as counterfactual ground-truth. For multiple counterfactual classes $\mathcal{V} = \{y_1, \ldots, y_V\}$, ground-truth consists of a set of parts that discriminates class a from those in V, which is defined as $\mathcal{G}_a^{\mathcal{V}} = \bigcap_{v=1}^V \{\mathbf{p}_i|(\mathbf{p}_i, a, b_i) \in \mathcal{G}^c, b_i = y_v \in \mathcal{V}\}$.

For two-class counterfactual explanations, evaluation is based on the precision-recall and IoU metrics used for deliberative explanations. For multi-class explanations, the definitions are generalized to account for the multiple counterfactual classes. Given a region $\mathbf{r}\{a, \mathcal{V}\}$, $R = \frac{J}{|\{i|\mathbf{p}_i \in \mathcal{G}_a^{\mathcal{V}}\}|}$, where $J = |\{i|\mathbf{p}_i \in \mathbf{r}, \mathbf{p}_i \in \mathcal{G}_a^{\mathcal{V}}\}|$ and IoU $= \frac{|\mathbf{r} \cap \mathbf{p}|}{|\mathbf{r} \cup \mathbf{p}|}$, where $\mathbf{p} = \{\mathbf{p}_i|\mathbf{p}_i \in \mathcal{G}_a^{\mathcal{V}}\}$. On datasets with point-based ground truth, evaluation is based on precision and recall of the generated counterfactual regions. On datasets with mask-based ground truth, the IoU is used.

We also define a metric that captures the semantic consistency of two segments, $\mathbf{r}\{a, b\}(\mathbf{x})$ and $\mathbf{r}\{b, a\}(\mathbf{x}^c)$, by calculating the consistency of the parts included in them. This is denoted as the part IoU (PIoU),

$$\text{PIoU} = \frac{|\{k|(\mathbf{p}_k, a, b) \in \mathbf{r}\{a, b\}(\mathbf{x})\} \cap \{k|(\mathbf{p}_k, b, a) \in \mathbf{r}\{b, a\}(\mathbf{x}^c)\}|}{|\{k|(\mathbf{p}_k, a, b) \in \mathbf{r}\{a, b\}(\mathbf{x})\} \cup \{k|(\mathbf{p}_k, b, a) \in \mathbf{r}\{b, a\}(\mathbf{x}^c)\}|}. \tag{24}$$

This metric provides a fair comparison of different explanations if their counterfactual regions have the same size. Region size is controlled by T in (8) and (9).

User expertise has an impact on counterfactual explanations. Beginner users tend to choose random counterfactual classes, while experts tend to pick counterfactual classes similar to the true class. Hence, explanation performance should be measured for the two user types. In this work, users are simulated by choosing a random counterfactual class b for beginners and the class predicted by a small CNN for advanced users. Class a is the prediction of the classifier used to generate the explanation, which is a larger CNN.

3) Attributive Explanations: For attributive explanations, ground-truth consists of parts with unique attributes, present in the ground truth class and lacking in all other classes. This is similar to the ground truth of multi-class counterfactual explanations but $\mathcal{V}$ now contains all dataset classes other than y^*. However, it is frequently impossible to find a part whose attributes appear in a single class. Hence, we randomly select L classes from $\mathcal{Y} \setminus \{y^*\}$, to create a label set $\mathcal{L} = \{y_1, \ldots, y_L\}$ and use the evaluation metrics discussed for multi-class counterfactual explanations with $\mathcal{V} = \mathcal{L}$. The difference is that, in the counterfactual setting, $\mathcal{V}$ is selected by the user.

VII. Experiments

In this section we discuss an experimental evaluation of the explanations generated by GALORE.

3.Note that part and attribute annotations are only required to evaluate the accuracy of insecurities, not to compute the visualizations. These require no annotation.

A. Experimental Setup

Datasets. Experiments were performed on the CUB200 [89] and ADE20K [95] datasets. CUB200 [89] is a densely-labeled dataset of fine-grained bird classes, annotated with parts. 15 part locations (points) are annotated including back, beak, belly, breast, crown, forehead, left/right eye, left/right leg, left/right wing, nape, tail and throat. Attributes are defined and assigned to each part according to [89]. ADE20K [95] is a fine-grained scene image dataset with more than 1000 scene categories and segmentation masks for 150 objects. In this case, objects are seen as scene parts and each object has a single attribute, which is its probability of appearance in a scene. Both datasets were subject to standard normalizations. All results are presented on the standard CUB200 test set and the official validation set of ADE20 K.

Networks. VGG16 [96] is the most popular architecture in the explanation literature. Unless otherwise noted, it is used for all visualizations. It is also compared to the ResNet-50 [97] and AlexNet [98]. All predictors are trained by standard strategies [29], [88], [96], [97], [98]. The last convolutional layer output, widely used in the visualization literature [15], [16], [99], is used to create all explanations.

Evaluation. On CUB200, where all semantic descriptors ϕ_c^k are multidimensional, similarities $\alpha_{a,b}^k$ are computed with $\gamma(\phi_a^k, \phi_b^k) = e^{-\{\mathrm{KL}(\phi_a^k||\phi_b^k)+\mathrm{KL}(\phi_b^k||\phi_a^k)\}}$ [100], where $\mathrm{KL}(.||.)$ is the Kullback–Leibler divergence. $\alpha_{\mathcal{V}}^k$ is computed with $\eta(\phi_{a_1}^k, \ldots, \phi_{a_V}^k) = \min_{i,j\in\mathcal{V}, i\neq j} \gamma(\phi_i^k, \phi_j^k)$, i.e., the minimum similarity $\gamma(\phi_a^k, \phi_b^k)$ between all class pairs in $\mathcal{V}$. To generate groundtruth for insecurities and discriminant regions, the set $\mathcal{G}$ of region and class tuples was divided into two subsets. The size M of the set of groundtruth insecurities was set to the 20% insecurities $(\mathbf{p}_i, a_i, b_i)$ or $(\mathbf{p}_i, \mathcal{V})$ of strongest ambiguity. The size N of the set of discriminant groundtruth regions was set to the remaining 80% parts $(\mathbf{p}_i, a_i, b_i)$ or $(\mathbf{p}_i, \mathcal{V})$ of smallest similarity. This division reflects the fact that dissimilar parts dominate $\mathcal{G}$. Since parts are labelled with points, accuracy is measured with precision and recall.

On ADE20 K, the semantic descriptors ϕ_c^k are scalar (where $k \in \{1, \ldots, 150\}$) namely the probability of occurrence of part (object) k in scenes of class c. This is estimated by the relative frequency with which the part appears in scenes of the class. Only parts such that $\phi_c^k > 0.3$ are considered. For deliberative explanations, ambiguity strengths are computed with $\gamma(\phi_a^k, \phi_b^k) = \frac{1}{2}(\phi_a^k + \phi_b^k)$. This is large when object k appears very frequently in both classes, i.e., the object adds ambiguity. Due to the sparsity of the matrix of ambiguity strengths $\alpha_{a,b}^k$, the number M of ground-truth insecurities is set to the 1% triplets of strongest ambiguity. On the other hand, counterfactual ground truth consists of the triplets $(\mathbf{p}_i, a_i, b_i)$ with $\phi_a^k > 0$ and $\phi_b^k = 0$, i.e., where object k appears in class a but not in class b.

Since deliberative explanations aim to explain examples that are difficult to classify, explanations are produced only for the 100 test images of largest difficulty score on each dataset. The $W = 5$ top classes are used to produce the class ambiguity set (see Section IV-D). In counterfactual explanations, AlexNet predictions [98] are used to mimic advanced users. For multi-class explanations, V is set to $V = 3$ for deliberative and $V = 2$ for counterfactual. This reflects the fact that users typically do not pose counterfactuals with large numbers of classes.

B. Ablation Study

Self-Awareness Scores. Fig. 7 shows the impact of the confidence scores of (21)–(23) on precision-recall curves (on CUB200) and IoU (on ADE20 K) for three explanation strategies. Some conclusions can be drawn. First, self-awareness is useful for all explanations. For attributive explanations, self-awareness attribution functions highlight more class-specific features. For counterfactual explanations, the gains are larger for expert users than for beginners. This is because the counter and predicted classes are more similar for the former, producing attribution maps that overlap. Second, the easiness score substantially outperforms the remaining scores, for all but counterfactual explanations with beginner users, where counter classes are easy to distinguish. Third, for deliberative explanations, only the easiness score $s^e(\mathbf{x})$ improves on the baseline. This suggests that self-referential difficulty scores are not always reliable. For this reason, the easiness score is used in the remaining experiments.

Attribution Function. [4] GALORE is compatible with any attribution function. Fig. 8 (left) compares different functions: baseline gradient ('Grad'), the integrated gradient of [51] ('InteGrad'), Grad-CAM [16], score-CAM [17], and SHAP [53]. For brevity, we only present deliberative and counterfactual results for advanced users. A few conclusions are possible. First, while the four more complex functions always outperform Grad, the differences are small, especially on ADE20 K. This is probably because ADE20 K is more difficult (more than 1000 categories and only about 16 examples per category) than CUB200 (200 categories and 26 examples per category). Second, while GALORE benefits from advanced attribution functions, there is little difference between InteGrad, Grad-CAM, SHAP and score-CAM. No attribution function is consistently better than all others.

Network Architectures. Fig. 8 (right) compares the explanations produced by ResNet-50, VGG16 and AlexNet. For counterfactual explanations, only the former two are compared because AlexNet is used to simulate the users. On CUB200, ResNet-50 has the best performance. Interestingly, although ResNet-50 and VGG16 have similar classification performance on these two datasets, the ResNet segments are much more accurate than those of VGG16. This suggests that the ResNet architecture uses more intuitive, i.e., human-like, deliberations. On ADE20 K, where the classification task is harder ($< 60\%$ mean accuracy), there is no clear difference between the three architectures.

C. Multi-Class Explanations

Fig. 9 summarizes the performance of multi-class deliberative counterfactual explanations. These results are similar to those

4.Since no new algorithm is proposed for attributive explanations, ablations are restricted to deliberative and counterfactual explanations in the remainder of the paper.

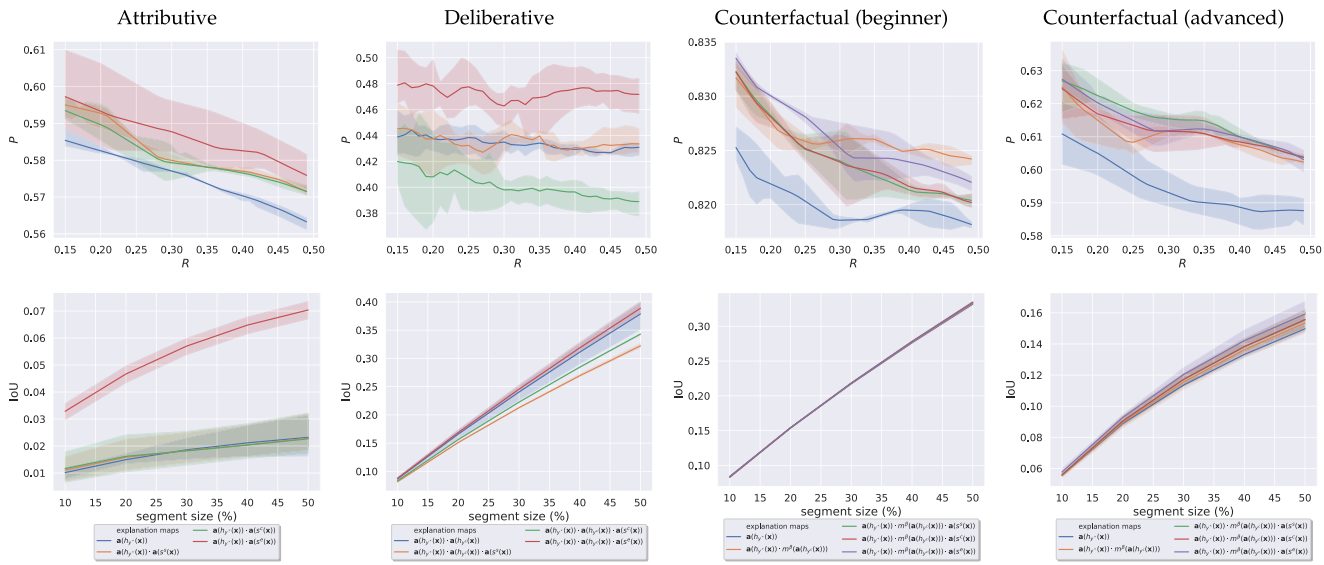

Fig. 7. Effect of confidence scores on precision-recall curves and IoU of different GALORE explanations. Top: on CUB200. Bottom: ADE20 K.

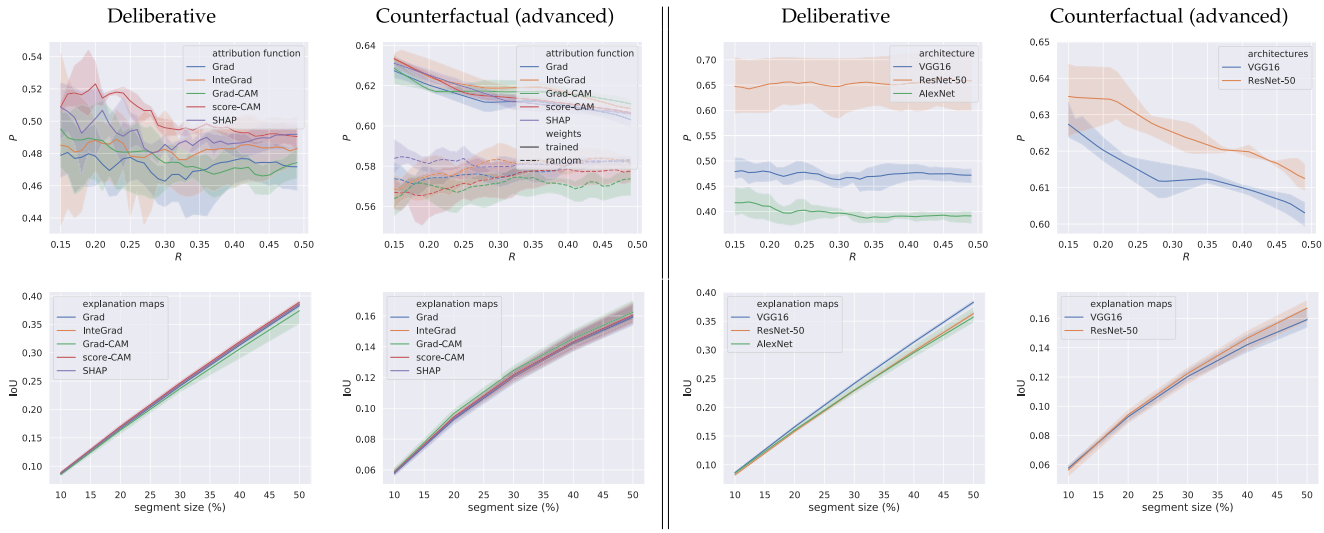

Fig. 8. Impact of attribution function (left) and network architecture (right) on GALORE explanation performance. Top: precision-recall on CUB200. Bottom: IoU on ADE20 K.

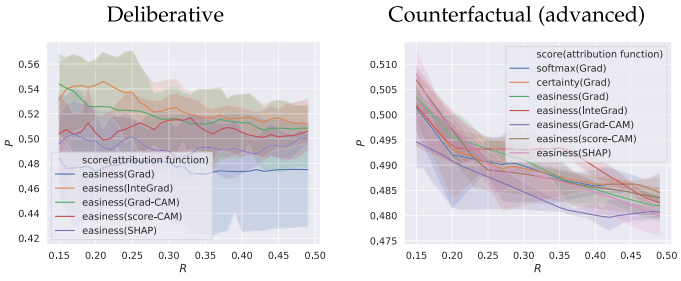

Fig. 9. Precision-recall of multi-class explanations on CUB200.

obtained with binary explanations. An interesting observation is that, for a given recall level, the precision of deliberative explanations is even higher than for binary insecurities. This is seemingly counter intuitive, since more classes should increase the difficulty of the explanation. We hypothesize it happens because the three classes are very similar, having many attributes in common. Combining three attribution maps decreases the risk of missing common attributes. Another observation is that, similarly to binary deliberative and counterfactual explanations, the differences between attribution functions are small.

D. Segment Strength

The accuracy of segment strengths was evaluated by the Pearson correlation coefficient between strength and quality of the explanation, measured by segment precision. Table II shows a strong positive correlation for all explanations. This is sensible because strength is defined as the average intensity

TABLE II
PEARSON CORRELATION COEFFICIENT (ρ) AND P-VALUE BETWEEN SEGMENT STRENGTH AND QUALITY OF THE EXPLANATION ON CUB200

explanation	ρ	p-value
Deliberative		
Binary	0.62	0.01
Multiclass	0.57	0.02
Counterfactual		
Binary	0.63	8e-3
Multiclass	0.59	0.03

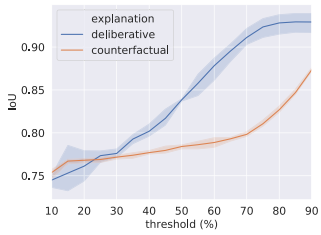

Fig. 10. Robustness of GALORE to image shifts on CUB200.

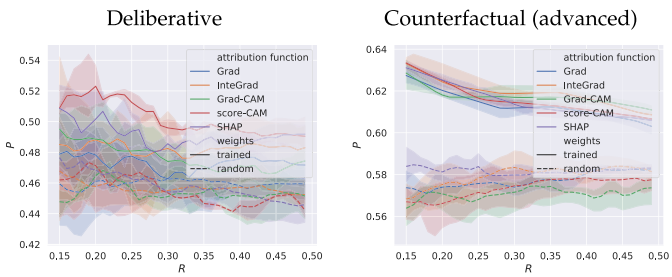

Fig. 11. Precision-recall of GALORE explanations obtained with pre-trained and random weights on CUB200.

of the attribution map inside the segment. Hence, the explanation should be more class-specific for larger strengths, corresponding to segments of higher quality.

E. Sanity Checks

Recent works have shown that attribution maps can be sensitive to data shifts and model variance [76], [77]. Data shift checks [76] test the robustness of the explanation to input shifts. For this, test images were randomly translated by 1 to 10 pixels along four directions. The resulting insecurities and counterfactual segments were compared to those obtained without translations, by measuring the similarity (IoU) between segments. The average IoU across all segments and examples is shown in Fig. 10 as a function of the threshold T. While these are plots for the 'easiness-Grad-VGG' configuration, they are typical. The average IoU is almost always above 75% showing that the explanations of GALORE are robust to image shifts. Parameter randomization tests [77] compare the explanation of well-trained and random initialized models. Similar outputs indicate that the explanation method is insensitive to model parameters, which is undesirable. Fig. 11 shows that all attribution functions passed the sanity check, since pre-trained models always outperformed random initialization. This was especially true for score-CAM and the differences were larger for counterfactual explanations.

F. Comparison to State of the Art

GALORE was compared with state of the art explanation methods, with the results of Table III. The left side of the table presents a counterfactual explanation comparison between GALORE, the method of [30], and CounteRGAN [63], for the two user types considered in this work. To the best of our knowledge there have been no other attempts in the literature to produce deliberative explanations. The right side of the table compares the deliberative explanations of GALORE to a baseline that we have designed, inspired by the method of [30] for counterfactual explanations.

This baseline is as follows. Given the query image $\mathbf{x}$ and associated candidate class ambiguity set $\mathbb{A}$, a pair of images is randomly sampled from the training set for each ambiguity $(a, b) \in \mathbb{A}$: $\mathbf{x}^{a,0}$, of class a, and $\mathbf{x}^{b,0}$, of class b. A sliding window is defined over $\mathbf{x}$. For each window $\mathcal{W}$, we exhaustively search matching windows $\mathcal{W}_a$ in $\mathbf{x}^{a,0}$ and $\mathcal{W}_b$ in $\mathbf{x}^{b,0}$. The matching is defined as follows. Let $\mathbf{x}_a$ ($\mathbf{x}_b$) be $\mathbf{x}$ with $\mathcal{W}$ replaced by $\mathcal{W}_a$ ($\mathcal{W}_b$). The matching windows are those that minimize the change of prediction when inserted in $\mathbf{x}$, i.e., $|h_a(\mathbf{x}^a) - h_a(\mathbf{x})| + |h_b(\mathbf{x}^b) - h_b(\mathbf{x})|$. Regions $\mathcal{W}_a$ and $\mathcal{W}_b$ should have features that are common to the two ambiguous classes, and thus be most confusing for the classifier.

For fair comparison, these experiments use the softmax score of (21), so that model sizes are equal for both [30] and the proposed approach. The size of the counterfactual (or deliberative) region is the receptive field size of one unit ($\frac{1}{14*14} \approx 0.005$ of image size for VGG16 and $\frac{1}{7*7} \approx 0.02$ for ResNet-50). This is constrained by the speed of the algorithm of [30], where the counterfactual region is determined by exhaustive feature matching. For CounteRGAN, we guarantee the same region size by thresholding the residual outputs of the generator.

Several conclusions can be drawn from the table. First, GALORE outperforms the counterfactual explanations of [30], [63] and the baseline deliberative explanation for almost all metrics. Second, GALORE is much faster, improving the speed of [30] by 1000+ times on VGG and 50+ times on ResNet. This is because it does not require exhaustive feature matching. These gains increase with the size of the counterfactual (or deliberative) region, since computation time is constant for GALORE but exponential on region size for [30]. Third, due to the small size used in these experiments, PIoU is relatively low for all methods. It is, however, larger for GALORE explanations with large gains in some cases (VGG & advanced). Fig. 14 shows that PIoU can raise to 0.5 for regions of 10% (VGG) or 20% (ResNet) of the image size. This suggests that, for such regions sizes, region pairs have matching semantics.

G. Visualizations

Fig. 12 shows two examples of deliberative explanations of two insecurities each. The left of the figure shows the insecurities of the classifier for an image of a 'Glaucous gull'. For GALORE, the top insecurity covers the leg/belly region, which is a region of

TABLE III
COMPARISON TO THE STATE OF THE ART IN COUNTERFACTUAL EXPLANATIONS. (IPS: IMAGES PER SECOND, IMPLEMENTED ON NVIDIA TITAN XP. RESULTS ARE OMITTED FOR THE COUNTERGAN [63] DUE TO THE VERY LONG TRAINING TIMES IT REQUIRES.) RESULTS ARE SHOWN AS MEAN(STDDEV)

		Counterfactual explanations						Deliberative explanations	
		Beginner User			Advanced User				
Arch.	Metric	Goyal [30]	CounteRGAN [63]	GALORE	Goyal [30]	CounteRGAN [63]	GALORE	Baseline	GALORE
VGG16	R	0.02 (0.01)	0.03 (0.00)	**0.05** (0.01)	**0.05** (0.00)	**0.05** (0.00)	**0.05** (0.00)	0.02 (0.00)	**0.04** (0.00)
	P	0.76 (0.01)	0.78 (0.00)	**0.84** (0.01)	0.56 (0.01)	0.61 (0.00)	**0.64** (0.01)	0.43 (0.03)	**0.48** (0.02)
	PIoU	0.13 (0.00)	0.13 (0.00)	**0.15** (0.00)	0.09 (0.00)	0.12 (0.00)	**0.14** (0.02)	N/A	N/A
	IPS	0.02 (0.00)	N/A	**26.51** (0.71)	N/A	N/A	N/A	<0.01	**3.78** (0.31)
ResNet-50	R	0.03 (0.01)	0.06 (0.00)	**0.09** (0.02)	0.12 (0.01)	**0.17** (0.00)	0.16 (0.00)	0.03 (0.00)	**0.06** (0.00)
	P	0.77 (0.01)	0.74 (0.01)	**0.81** (0.01)	0.57 (0.02)	0.56 (0.00)	**0.60** (0.01)	0.67 (0.03)	**0.72** (0.04)
	PIoU	0.18 (0.01)	**0.20** (0.00)	0.16 (0.01)	**0.15** (0.00)	0.14 (0.00)	**0.15** (0.01)	N/A	N/A
	IPS	1.13 (0.07)	N/A	**78.54** (11.87)	N/A	N/A	N/A	0.12 (0.06)	**8.41** (0.45)

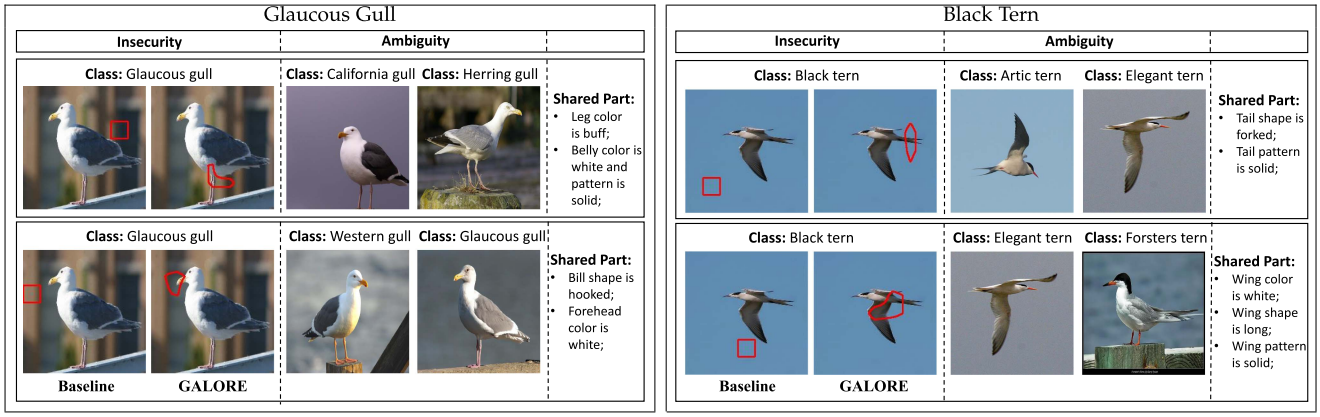

Fig. 12. Deliberative explanation comparisons produced by the baseline method and GALORE for two images from CUB. Left: a Glaucous Gull creates two insecurities. Top: the insecurity shown on the left elicits ambiguity between the California and Herring Gull classes. The attributes of the image region covered by the GALORE insecurity are listed on the right. Bottom: insecurity with ambiguity between Western and Glaucous Gull classes. Right: similar for Black Tern.

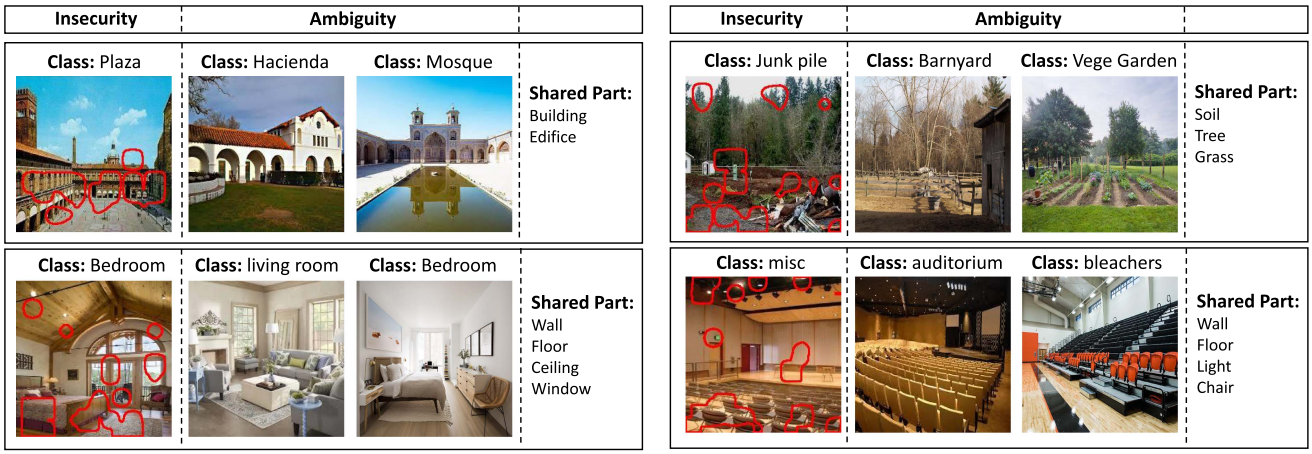

Fig. 13. Deliberative explanations produced by GALORE for four images from ADE20 K.

ambiguity with classes 'California gull' and 'Herring gull' that also have leg color 'buff,' belly color 'white,' and belly pattern 'solid'. The lower insecurity covers the bill/forhead region of the gull, due to an ambiguity between the 'Glaucous gull' and the 'Western gull' with whom the 'Glaucous gull' shares a 'hooked' bill shape and a 'white' colored forehead. The right side of the figure shows insecurities for a 'Black tern,' due to a tail ambiguity with 'Artic' and 'Elegant' terns and a wing ambiguity with 'Elegant' and 'Forsters' terns. These insecurities are much more informative of class ambiguity than those produced by the baseline, which sometimes localizes irrelevant regions, like backgrounds. Fig. 13 shows single GALORE insecurities from four images of ADE20 K. In all cases, the insecurities correlate with regions of attributes shared by different classes. This shows that deliberative explanations unveil truly ambiguous image regions, generating intuitive insecurities that help understand

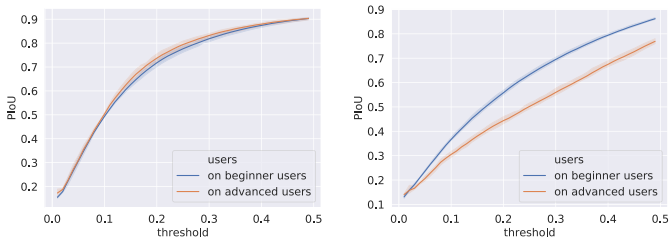

Fig. 14. PIoU of proposed counterfactual explanations as a function of the segmentation threshold on CUB200. Left: VGG16, right: ResNet-50.

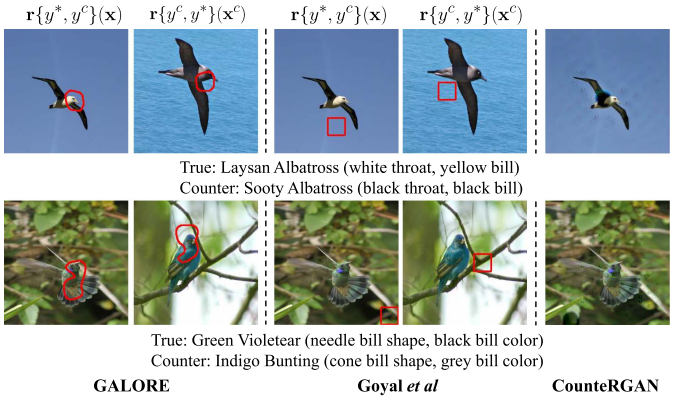

Fig. 15. Counterfactual explanations (true and counter classes shown below each example, ground truth class-specific part attributes in parenthesis). Left: GALORE. Center: [30]. Right: CounteRGAN [63].

network predictions. Note, for example, how the visualization of insecurities tends to highlight classes that are semantically very close, such as the different families of gulls or terns and class subsets such as 'plaza,' 'hacienda,' and 'mosque' or 'bedroom' and 'living room'. All of this suggests that the deliberative process of the network correlates well with human reasoning.

Fig. 15 shows two examples of counterfactual visualizations on CUB200. The regions selected in the query and counter class image are shown in red. For CounteRGAN [63], the generated explanatory images are shown. The true y^* and counter y^c class are shown below the images and followed by the ground truth discriminative attributes for the image pair. Note how GALORE explanations identify semantically matched and class-specific bird parts on both images. For example, the throat and bill that distinguish Laysan from Sooty Albatrosses. This feedback enables a user to learn that Laysans have white throats and yellow bills, while Sootys have black throats and bills. This is unlike the regions produced by [30], also shown in the figure, which sometimes highlight irrelevant cues, such as the background. CounteRGAN, only generates some patterns from the counterfactual classes (zoom in for more detail), but not realistic images. This is consistent with the well known difficulty of GANs to translate images across hundreds of fine-grained classes. Fig. 16 presents similar figures for ADE20 K, where the proposed explanations tend to identify scene-discriminative objects. For example, that a promenade deck contains objects 'floor,' 'ceiling,' 'sea,' while a bridge scene includes 'tree,' 'river' and 'bridge'.

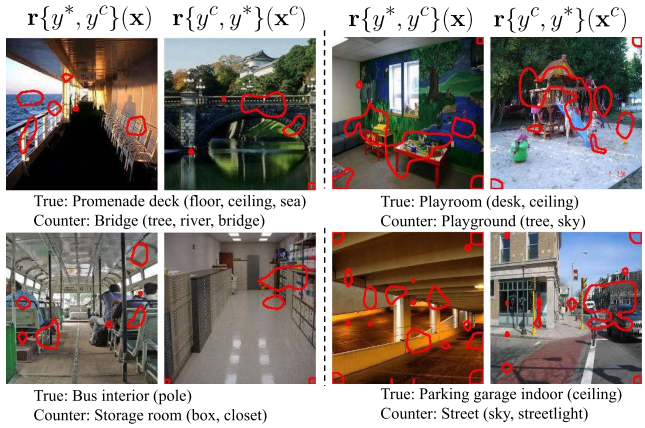

Fig. 16. Counterfactual explanations by GALORE on ADE20 K.

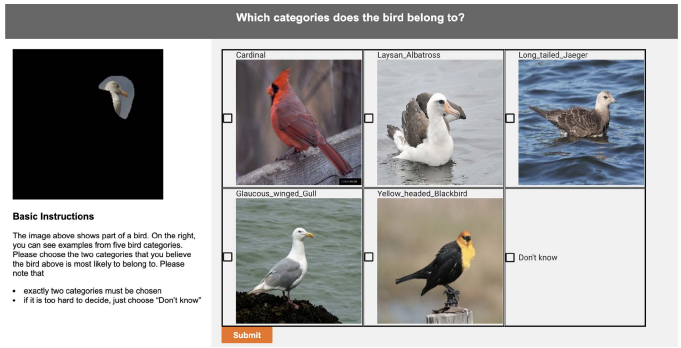

Fig. 17. MTurk interface for human evaluation of deliberative explanations.

VIII. Human Studies

A. Insecurity Evaluation

Fig. 17 shows the interface of the human experiment used to evaluate deliberative explanations on Amazon MTurk. The region of support of the uncertainty is shown on the left and examples from five classes are displayed on the right. These include the two ambiguous classes a and b found by the explanation algorithm, the "Laysan Albatross" and the "Glaucous Winged Gull". The Tuker is asked to select, among the five classes shown, the two to which the segment on the left is most likely to belong. If these two classes match the ambiguities found by the explanation algorithm the insecurity is considered intuitive. Otherwise, it is not. Turker performance was compared for insecurities generated by the explanation algorithm and randomly cropped regions of the same size. Turkers agreed amongst themselves on classes a and b for 59.4% of the insecurities and 33.7% of randomly cropped regions. They agreed with the algorithm for 51.9% of the insecurities and 26.3% of the random crops. This shows that 1) insecurities are much more predictive of the ambiguities sensed by humans, and 2) the algorithm predicts those ambiguities with significant levels of consistency. In both cases, the "Don't know" rate was around 12%.

Fig. 18. Visualization of machine teaching experiment.

B. Application to Machine Teaching

Goyal et al. [30] used counterfactual explanations to design an experiment to teach humans distinguish two bird classes. During a training stage, learners are asked to classify birds. When they make a mistake, they are shown counterfactual feedback of the type of Fig. 15, using the true class as y^* and the class they chose as y^c. This helps them understand why they chose the wrong label, and learn how to better distinguish the classes. In a test stage, learners are then asked to classify a bird without visual aids. Experiments reported in [30] show that this is much more effective than simply telling them whether their answer is correct/incorrect, or other simple training strategies. We made two modifications to this set-up. The first was to replace bounding boxes with highlighting of the counterfactual regions, as shown in Fig. 18. We also instructed learners not to be distracted by the darkened regions. Unlike [30], this guarantees that they do not exploit cues outside the counterfactual regions to learn bird differences. Second, to verify this, we added two experiments where 1) highlighted regions are generated randomly (without telling the learners); 2) the entire images are lighted. If these produce the same results, one can conclude that the explanations do not promote learning.

We also chose two more difficult birds, the Setophaga Citrina and the Kentucky Warbler (see Fig. 18), than [30]. These classes have large intra-class diversity and cannot be distinguished by color alone, unlike those of [30]. The experiment has three steps. The first is a pre-learning test, where humans are asked to classify 20 examples of the two classes, or choose a 'Don't know' option. The second is a learning stage, where counterfactual explanations are provided for 10 bird pairs. The third is a post-learning test, where humans are asked to answer 20 binary classification questions. In this experiment, all students chose 'Don't know' in the pre-learning test. However, after the learning step, they achieved 95% mean accuracy, compared to 60% (random highlighted regions) and 77% (entire images lighted) in the contrast settings. These results suggest that the proposed counterfactual explanations can help teach naive humans distinguish categories from an expert domain.

IX. Conclusion

In this work, we have proposed a new framework, GALORE, for visualization-based explanations of deep neural networks predictions. GALORE unifies attributive, counterfactual, and deliberative explanations, aiming to satisfy the requirements of a diverse set of end-users. Attributive explanations visualize how different pixels contribute to a class prediction, deliberative explanations address the "why?" question, and counterfactual explanations the "why not?" question. All explanations are based on a combination of attributions with respect to class predictions and confidence scores. This makes them very efficient to compute, in some cases orders of magnitude faster than the state of the art. We have also introduced an experimental protocol to evaluate explanation accuracy, which sidesteps the difficulty of replicating user experiments. We believe this will facilitate research in the visualization based XAI problem. Both this protocol and human experiments were used to evaluate GALORE on two fine-grained datasets, demonstrating that its explanations are more accurate than those previously available, intuitive, and correlate with human perception. In this process, we have also validated the importance of self-awareness both to define different explanations and to increase their accuracy. The counterfactual explanation results have shown to be beneficial for machine teaching.

Acknowledgment

The authors acknowledge and thank the use of the Nautilus platform for some of the experiments discussed above.

References

[1] P. C.-H. Lam, L. Chu, M. Torgonskiy, J. Pei, Y. Zhang, and L. Wang, "Finding representative interpretations on convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1345–1354.
[2] J. Wang, H. Liu, X. Wang, and L. Jing, "Interpretable image recognition by constructing transparent embedding space," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 895–904.
[3] M. Nauta, R. van Bree, and C. Seifert, "Neural prototype trees for interpretable fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14933–14943.
[4] B. Carter, S. Jain, J. W. Mueller, and D. Gifford, "Overinterpretation reveals image classification model pathologies," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 15395–15407.
[5] J. Parekh, P. Mozharovskyi, and F. d'Alché Buc, "A framework to learn with interpretation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 24273–24285.
[6] A. A. Ismail, H. C. Bravo, and S. Feizi, "Improving deep learning interpretability by saliency guided training," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 26726–26739.
[7] J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. Srivastava, "How can I explain this to you? An empirical study of deep neural network explanation methods," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 4211–4222.
[8] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.
[9] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 3–19.
[10] D. A. Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7775–7784.
[11] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.
[12] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS One*, vol. 10, no. 7, 2015, Art. no. e0130140.
[13] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "A unified view of gradient-based attribution methods for deep neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst. Workshop*, 2017.

[14] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[17] H. Wang et al., "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 24–25.

[18] P. Wang and N. Vasconcelos, "A machine teaching framework for scalable recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 4945–4954.

[19] X. Zhu, A. Singla, S. Zilles, and A. N. Rafferty, "An overview of machine teaching," 2018, *arXiv:1801.05927*.

[20] O. M. Aodha, S. Su, Y. Chen, P. Perona, and Y. Yue, "Teaching categories to human learners with visual explanations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3820–3828.

[21] A. Dhurandhar and K. Shanmugam, "Counterfactual vs contrastive explanations in artificial intelligence," *Towardsdatascience*, 2020.

[22] A. Korikov, A. Shleyfman, and C. Beck, "Counterfactual explanations for optimization-based decisions in the context of the GDPR," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, 2021, pp. 4097–4103.

[23] A. Dhurandhar et al., "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 592–603.

[24] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence," *IEEE Access*, vol. 9, pp. 11974–12001, 2021.

[25] S. Rathi, "Generating counterfactual and contrastive explanations using SHAP," 2019, *arXiv:1906.09293*.

[26] T. Tsiligkaridis, "Failure prediction by confidence estimation of uncertainty-aware Dirichlet networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 3525–3529.

[27] C. Corbière, N. Thome, A. Saporta, T.-H. Vu, M. Cord, and P. Perez, "Confidence estimation via auxiliary models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6043–6055, Oct. 2022.

[28] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4878–4887.

[29] X. Wang, Y. Luo, D. Crankshaw, A. Tumanov, F. Yu, and J. E. Gonzalez, "IDK cascades: Fast deep learning by learning not to overthink," 2017, *arXiv:1706.00885*.

[30] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2376–2384.

[31] W. Wu et al., "Towards global explanations of convolutional neural networks with concept attribution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8652–8661.

[32] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, "On completeness-aware concept-based explanations in deep neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 20554–20565.

[33] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 832.

[34] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 801.

[35] P. Hase, C. Chen, O. Li, and C. Rudin, "Interpretable image recognition with hierarchical prototypes," in *Proc. 7th AAAI Conf. Hum. Computation Crowdsourcing*, 2019, pp. 32–40.

[36] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! Criticism for interpretability," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2288–2296.

[37] P. Rodríguez et al., "Beyond trivial counterfactual explanations with diverse valuable explanations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1056–1065.

[38] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GPDR," *Harvard J. Law Technol.*, vol. 31, 2017, Art. no. 841.

[39] A. Jacovi, S. Swayamdipta, S. Ravfogel, Y. Elazar, Y. Choi, and Y. Goldberg, "Contrastive explanations for model interpretability," 2021, *arXiv:2103.01378*.

[40] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 269–286.

[41] K. H. Lee, C. Park, J. Oh, and N. Kwak, "LFI-CAM: Learning feature importance for better visual explanation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1355–1363.

[42] D. Lim, H. Lee, and S. Kim, "Building reliable explanations of unreliable neural networks: Locally smoothing perspective of model interpretation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6468–6477.

[43] Y. Wang and X. Wang, "Self-interpretable model with transformation equivariant interpretation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 2359–2372.

[44] M. Bohle, M. Fritz, and B. Schiele, "Convolutional dynamic alignment networks for interpretable classifications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10029–10038.

[45] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8662–8672.

[46] T. Fel, R. Cadène, M. Chalvidal, M. Cord, D. Vigouroux, and T. Serre, "Look at the variance! Efficient black-box explanations with sobol-based sensitivity analysis," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 26005–26014.

[47] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[48] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics*, 2018, pp. 80–89.

[49] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, 2020, Art. no. 18.

[50] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," 2021, *arXiv:2102.13076*.

[51] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328. [Online]. Available: JMLR.org

[52] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 371.

[53] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[54] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.

[55] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*.

[56] A. Van Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," 2019, *arXiv:1907.02584*.

[57] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative counterfactual introspection for explainable deep learning," 2019, *arXiv:1907.03077*.

[58] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *Proc. Int. Conf. Learn. Representations*, 2017.

[59] O. Lang et al., "Explaining in style: Training a GAN to explain a classifier in stylespace," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 693–702.

[60] J. Thiagarajan, V. S. Narayanaswamy, D. Rajan, J. Liang, A. Chaudhari, and A. Spanias, "Designing counterfactual generators using deep model inversion," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 16873–16884.

[61] Y. Zhao, "Fast real-time counterfactual explanations," 2020, *arXiv:2007.05684*.

[62] R. Luss, P.-Y. Chen, A. Dhurandhar, P. Sattigeri, K. Shanmugam, and C.-C. Tu, "Generating contrastive explanations with monotonic attribute functions," 2019, *arXiv:1905.12698*.

[63] D. Nemirovsky, N. Thiebaut, Y. Xu, and A. Gupta, "CounteRGAN: Generating realistic counterfactuals with residual generative adversarial nets," 2020, *arXiv:2009.05199*.

[64] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[65] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3730–3738.
[66] S. Khorram and L. Fuxin, "Cycle-consistent counterfactuals by latent transformations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10203–10212.
[67] I. Lage et al., "An evaluation of the human-interpretability of explanation," 2019, *arXiv:1902.00006*.
[68] M. Yang and B. Kim, "Benchmarking attribution methods with relative feature importance," 2019, *arXiv:1907.09701*.
[69] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017.
[70] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "Evaluating feature importance estimates," 2018.
[71] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," 2017, *arXiv:1711.06104*.
[72] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3429–3437.
[73] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, 2018.
[74] N. Bansal, C. Agarwal, and A. Nguyen, "SAM: The sensitivity of attribution methods to hyperparameters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8673–8683.
[75] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, "On the (in)fidelity and sensitivity of explanations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 984.
[76] P.-J. Kindermans et al., "The (un)reliability of saliency methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Berlin, Germany: Springer, 2019, pp. 267–280.
[77] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 9505–9515.
[78] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Representations*, 2017.
[79] J. Yang et al., "Semantically coherent out-of-distribution detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 8301–8309.
[80] K. Tang et al., "CODEs: Chamfer out-of-distribution examples against overconfidence issue," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1153–1162.
[81] A. Zaeemzadeh, N. Bisagno, Z. Sambugaro, N. Conci, N. Rahnavard, and M. Shah, "Out-of-distribution detection using union of 1-dimensional subspaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9452–9461.
[82] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1757–1772, Jul. 2013.
[83] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1563–1572.
[84] S. Kong and D. Ramanan, "OpenGAN: Open-set recognition via open data generation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 813–822.
[85] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Learning placeholders for open-set recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4401–4410.
[86] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," in *Proc. Int. Conf. Learn. Representations*, 2018.
[87] D. Hendrycks, M. Mazeika, and T. G. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. Int. Conf. Learn. Representations*, 2019.
[88] P. Wang and N. Vasconcelos, "Towards realistic predictors," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 37–53.
[89] P. Welinder et al., "Caltech-UCSD Birds 200," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2010-001, 2010.
[90] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Proc. IEEE 13th Int. Conf. Control Automat. Robot. Vis.*, 2014, pp. 844–848.
[91] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P.-A. Poletti, and H. Müller, "Building a reference multimedia database for interstitial lung diseases," *Computerized Med. Imag. Graph.*, vol. 36, no. 3, pp. 227–238, 2012.
[92] T. Miller, "Contrastive explanation: A structural-model approach," 2018, *arXiv:1811.03163*.
[93] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6541–6549.
[94] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," in *Proc. Int. Conf. Learn. Representations*, 2015.
[95] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5122–5130.
[96] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[97] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[98] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[99] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9215–9223.
[100] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1858–1860, Jul. 2003.

Pei Wang (Student Member, IEEE) received the BS degree from the University of Electronic Science and Technology of China, in 2014, and the MS degree from the Institute of Automation, Chinese Academy of Sciences, in 2017. He is currently working toward the PhD degree with the Electrical and Computer Engineering Department, University of California San Diego. His current research interests include explainable AI and its application to human-machine collaborative learning.

Nuno Vasconcelos (Fellow, IEEE) received the licenciatura degree in electrical engineering and computer science from the Universidade do Porto, Portugal, and the MS and PhD degrees from the Massachusetts Institute of Technology. He is a professor with the Electrical and Computer Engineering Department, University of California, San Diego, where he heads the Statistical Visual Computing Laboratory. He has received a NSF CAREER award, a Hellman Fellowship, several best paper awards, and has authored more than 200 peer-reviewed publications. He has been area chair of multiple computer vision conferences, and is currently an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence*.