Multilevel Explainable Artificial Intelligence: Visual and Linguistic Bonded Explanations

Halil Ibrahim Aysel, Xiaohao Cai, and Adam Prugel-Bennett

***Abstract*—Applications of deep neural networks (DNNs) are booming in more and more fields but lack transparency due to their black-box nature. Explainable artificial intelligence (XAI) is, therefore, of paramount importance, where strategies are proposed to understand how these black-box models function. The research so far mainly focuses on producing, for example, class-wise saliency maps, highlighting parts of a given image that affect the prediction the most. However, this method does not fully represent the way humans explain their reasoning, and awkwardly, validating these maps is quite complex and generally requires subjective interpretation. In this article, we conduct XAI differently by proposing a new XAI methodology in a multilevel (i.e., visual and linguistic) manner. By leveraging the interplay between the learned representations, i.e., image features and linguistic attributes, the proposed approach can provide salient attributes and attribute-wise saliency maps, which are far more intuitive than the class-wise maps, without requiring per-image ground-truth human explanations. It introduces self-interpretable attributes to overcome the current limitations in XAI and bring the XAI closer to a human-like explanation. The proposed architecture is simple in use and can reach surprisingly good performance in both prediction and explainability for deep neural networks thanks to the low-cost per-class attributes.**

***Impact Statement*—Our work has the potential of gaining end users? trust in deep neural networks and making it possible to answer ?why? by creating human-like explanations. Future applications could include sensitive fields where practitioners are desperate to understand how black-box models decide on a specific prediction before their deployment. A prominent example is medical imaging where it is sine qua non to see how DNNs make decisions. Our technique could help domain experts trust the automated system they get help from. This is achieved differently from currently available techniques that can only highlight the part of an image that DNNs seem to rely on. We argue that self-explainable DNNs are the future of machine learning applications. As DNNs are now currently the most preferred techniques and their most apparent limitation is the complicated decision process, we bring about a novel and cheap technique that, to the best of our knowledge, has never been proposed before.**

***Index Terms*—Black box, deep neural networks (DNNs), explainable artificial intelligence (XAI), saliency maps.**

Manuscript received 2 March 2023; revised 23 June 2023 and 18 August 2023; accepted 20 August 2023. Date of publication 25 August 2023; date of current version 14 May 2024. The work of H. I. Aysel was supported by the Republic of Turkiye Ministry of National Education. This paper was recommended for publication by Associate Editor Gustavo Olague upon evaluation of the reviewers' comments. *(Corresponding author: Halil Ibrahim Aysel.)*

The authors are with the University of Southampton, SO17 1BJ Southampton, U.K. (e-mail: hia1v20@soton.ac.uk; x.cai@soton.ac.uk; apb@ecs.soton.ac.uk).

Our code: https://github.com/Halilibrahimaysel/Multilevel_XAI.

This article has supplementary downloadable material available at https://doi.org/10.1109/TAI.2023.3308555, provided by the authors.

Digital Object Identifier 10.1109/TAI.2023.3308555

I. Introduction

RECENT developments in computational resources with a significant rise in data size have led deep neural networks (DNNs), such as multilayer perceptron (MLP) and convolutional neural network (CNN), to be widely used in various tasks, for example, image classification. Despite their excellent performance in prediction, DNNs are seen as black boxes as their decision process generally includes a huge number of parameters and nonlinearities [1], [2], [3]. The lack of explanation in these black boxes hinders their direct implementation in important and sensitive domains such as medicine and autonomous driving, where human life may directly be affected [4], [5], [6].

An example would be the DNNs trained to detect coronavirus. Although many works have been conducted and claimed to have a high predictive performance in detecting COVID-19 cases, a Turing Institute's recent report [7] disappointingly found that artificial intelligence (AI) used to detect coronavirus had little to no benefit and may even be harmful, mainly due to unnoticed biases in the data and its inherent black-box nature (also, see, e.g., [8]). Another example is a woman who was hit and killed by an autonomous car. An investigation showed that the death was caused by the incapability of the car in detecting a human unless they are near a crosswalk [9]. In addition to these life-related examples, there are plenty of others where bias in training data or the model itself causes unwanted discrimination that may immensely affect people's lives. Amazon's AI-enabled recruitment tool is an example of how discriminative these models could be by only recommending men and directly eliminating resumes including the word "woman"; the company later announced that this tool had never been used to recruit people due to the detected bias [10]. These examples clearly show that for machine learning models to gain acceptance, it is critical to be able to reason why a certain decision has been made to prevent any unwanted consequences.

Explanations delivered by explainable AI (XAI) can help machine learning practitioners debug their models by, for example, investigating the misclassification cases [11] and detecting bias in data [12]. There have been a number of works in this context to reveal the reasoning of the black-box models [13], [14], [15], [16], [17], [18], [19], [20]. However, the most widely used techniques, creating class-wise saliency maps (e.g., see left of Fig. 1) to indicate the areas that contribute to the prediction the most, have severe innate limitations. The first is the validation process of these maps, which is mostly qualitative

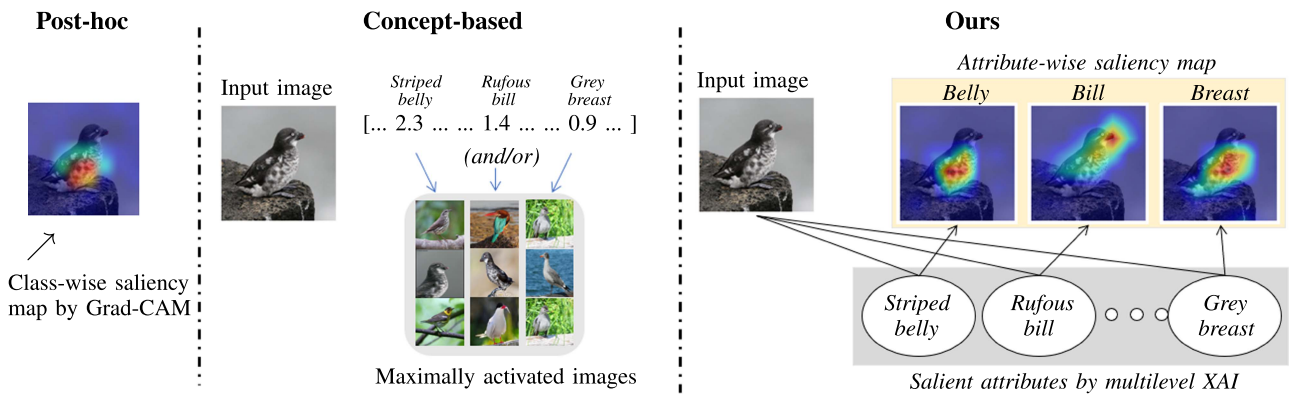

Fig. 1. Explainability of the proposed multilevel XAI model. A bird image from the Least Auklet class is predicted correctly by our approach, with human-like multilevel explanations via salient attributes (e.g., "striped belly") and the corresponding attribute-wise saliency maps (*right*). Result by Grad-CAM [29] (*left*) and concept-based models such as [21] and [22] (*middle*) are also given for comparison.

or requires labor intensive object-wise annotations [23], [24]. A recent study [25] showed that a full supervision of object segmentation by humans takes around 78 s per instance, while higher error rate bounding boxes take 10 s per instance to produce, which are much more expensive than 1 s per instance image level annotations. Moreover, requiring a higher level of annotation by experts is rather impractical. Another limitation stems from the discrepancy between these maps and human-like explanations. Humans naturally explain their reasoning using discriminative words (e.g., domestic versus wild or weak versus strong to differentiate a cat from a lion) together with pointing to where those words lie in the given image if visually permitting [23], [24] (*cf.*, our results on the right of Fig. 1). To produce human-like explanations, this multilevel (i.e., visual and linguistic) manner is crucial, which also inspires the work in this article.

In this article, we propose a new methodology called *multilevel XAI* to delve into DNNs by leveraging visual and linguistic attributes. Our approach exploits per-class attributes (rather than per-image attributes, which are too expensive and generally impractical) to interpret DNNs in, e.g., classifying raw images. By creating multilevel explanations, i.e., linguistic salient attributes and attribute-wise saliency maps, our method can provide explanations close to those we might expect from humans (e.g., see the right of Fig. 1). This is a big step forward in XAI and this new methodology does not suffer from the aforementioned limitations existing in current XAI solutions. The proposed setting adds a small extra cost to the training set, i.e., per-class attributes, which can be easily obtained if needed using, for example, online search engines or some autonomous tools (e.g., GPT-3 API [26]), and once acquired, they can always be in use since in most cases they are time and image invariant.

Our main contributions include the following:

1) proposing a multilevel XAI methodology, which is easy to use and can achieve near human-like explanations;
2) implementing extensive experiments on both coarse-grained and fine-grained datasets to validate the performance of the proposed approach;
3) conducting insightful discussions in XAI and future paths.

The rest of this article is organized as follows. Section II presents the related work in the XAI field. Section III details the methodology of our proposed multilevel XAI. We provide the specifics of the datasets used in the experiments in Section IV. The experiment details and the results are given in Section V. In Section VI, we raise a number of general research questions regarding XAI and discuss how much our technique, together with its limitations, addresses them. Finally, Section VII concludes this article. Pseudocodes and further experiments are provided in the supplementary material.

II. Related Work

The complexity of machine learning models generally affects the transparency/explainability because of the difficulty of following the model prediction process [11]. One line of research is where researchers employ inherently explainable models and utilize white-box models such as Bayesian rules [27] and linear models [28] to handle complex problems. These models, generally, struggle to reach the prediction ability of DNNs .

A. Post Hoc Approaches

Methodologies in the XAI field mainly aim to propose methods to understand how high performance black-box machine/deep learning models work. The majority of XAI methods introduced ideas to explain pretrained models in a *post hoc* manner, i.e., they are neither interested in the training setting nor in changing any of the models' components. These methods could be model independent requiring only prediction function [16], [17] or model dependent that need additional information of the trained model such as feature maps at a certain layer [20] or gradients [29]. DNNs for visual tasks do not output any textual justification. Modern visual-language models are effective in describing image content but lack outputting discriminative features that cause the prediction [23]. Forcing these models to output more discriminative features is one related work proposed in [24]. It aims to output multilevel explanations for vision-language tasks, e.g., visual question answering and activity recognition. Apart from a completely different focus against the work in this article, this method also requires labor

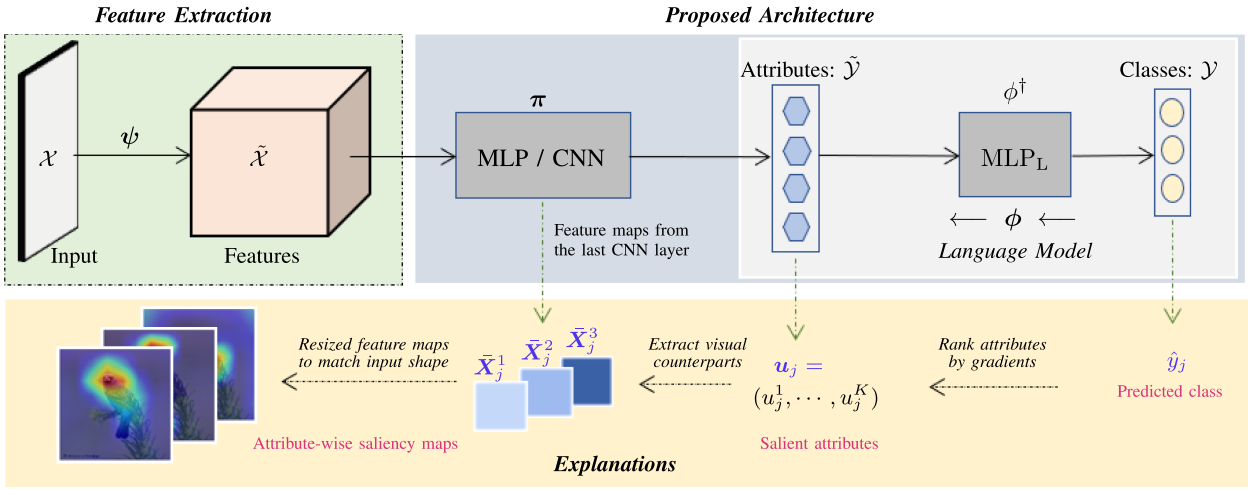

Fig. 2. Multilevel XAI architecture. Image features $\tilde{\mathcal{X}}$ are extracted from images $\mathcal{X}$ using feature extraction model ψ. Class labels $\mathcal{Y}$ are embedded into class attributes $\tilde{\mathcal{Y}}$ using the language model ϕ. DNN π (e.g., MLP and CNN) is then trained to match $\tilde{\mathcal{X}}$ with $\tilde{\mathcal{Y}}$. The explainability of the DNN π for image $\boldsymbol{X}_j$ is by the obtained salient attributes $\boldsymbol{u}_j$ (linguistic) and the attribute-wise saliency maps $\bar{\boldsymbol{X}}_j^i$ (visual).

intensive per-image annotations during training that are avoided in our work.

B. Ante Hoc Approaches

More recently, there have also been attempts to train self-explainable models, also known as *ante hoc* approaches. They can output explanations alongside their predictions, hence eliminating the need for any *post hoc* design. Most of the state-of-the-art methods for visual recognition tasks are based on the parametric softmax function, which projects latent features to the class space. One line of research presents methodologies based on nonparametric distance-based learning in the latent space and eliminates the use of softmax projection, making the decision processes of DNNs more human understandable. These methods cluster training images in the latent space to obtain class centroids, and then, classify test images based on their distances to these centroids [30], [31]. As our classifier is pretrained with attribute-class information and is frozen during the explainable MLP (X-MLP) and explainable CNN (X-CNN) training (see the detail in Section III-B), we share the main aim of these methodologies toward more understandable predictions—putting effort into modeling the latent data structure. Explanations by distance-based methods are achieved by constraining the class centroids to be samples from the training set, and predictions are claimed to be inherently explainable as class centroids (i.e., real observations from the training set) can be displayed as a reason for predictions.

Concept bottleneck models (CBM) [22] share ideas with our work and are analogous to X-MLP, but neither of them is multilevel. In particular, unlike CBM, our X-MLP and X-CNN only require class-wise attributes, ensuring our approach is significantly cheaper to implement. Similar to CBM in terms of being analogous to our X-MLP, Sarkar et al. [21] proposed a framework that can leverage concepts in different levels of

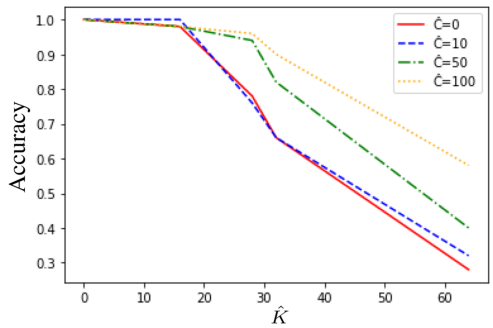

Fig. 3. Classification accuracy of $\mathrm{MLP_L}$ using the generated training datasets $\mathcal{T}$ with different $\hat{C}$ and $\hat{K}$ values on the AwA2 dataset. Note that the set of C samples is perturbed $\hat{C} > 0$ times to obtain $C\hat{C}$ number of samples; and $\hat{K}$ represents the number of attributes whose values are manipulated per class against the ground-truth attribute-class matrix.

supervision scenarios but with more storage and training capacity requirements. Explanations by these existing *ante hoc* approaches are predefined concepts (i.e., meaningful words such as stripes), nondefined concepts (e.g., concepts $1, 2, 3, \ldots$), and/or the images that maximally activate these concepts (see also the middle of Fig. 1). In contrast, our explanations are multilevel, possessing the advantage over the *ante hoc* approaches mentioned previously in terms of being capable of providing a spatial location in individual images associated with each linguistic attribute (see Fig. 11 for an example that presents the significance of our multilevel explanations).

The zero-shot learning regime is where side information (e.g., attributes and class taxonomies) is exploited to classify images of classes that have no labeled samples during training [32]. The aim is to match image features with class attributes, and then, to classify unseen classes thanks to the prior side information. There are various of techniques to find the best match that allow unseen class predictions [33], [34], [35], [36]. Although we integrate side information into our training process similar

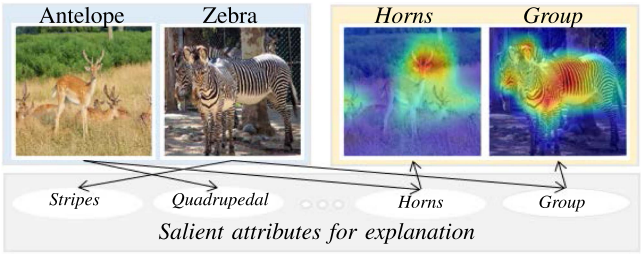

Fig. 4. Explainability of the proposed approach for correct prediction. Human-like multilevel explanations are delivered by the salient attributes and their saliency maps (by X-CNN), which are matched well.

to zero-shot learning, we are not interested in unseen classes; instead, the proposed work aims to train self-explainable models in the many-shot case. Unlike the majority of XAI methods, our explanations are multilevel outputting both linguistic and visual explanations. Finally, different from other extremely limited number of multilevel attempts that specifically work on vision-language models, our training setting is significantly cheaper and does not require per-image annotations.

III. Methodology of Multilevel XAI

In this section, we introduce our multilevel XAI methodology; see Fig. 2 for its main architecture. It consists of the following three main components:

1) a pretrained feature extraction block generating high level image features from input images (left of Fig. 2);
2) a self-explainable DNN block bridging the extracted features with linguistic attributes (middle of Fig. 2);
3) a language model block (being frozen after training) linking the linguistic attributes to the output class labels (right of Fig. 2).

All of these blocks are important and are well studied in various fields, yet in the XAI regime, their study is rather limited. To the best of our knowledge, this is the first time they have been used to explain neural networks in a multilevel (i.e., visual and linguistic) manner particularly when the per-image attributes are unavailable. Further description is given as follows.

Preliminary: Let $\mathcal{X}$ be the set of images and $\mathcal{Y} = \{1, 2, \ldots, C\}$ be the set of C class labels. Let $\mathcal{S} = \{(\boldsymbol{X}_i, y_i) \mid \boldsymbol{X}_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, 2, \ldots, N\}$ be a training set with N image/label pairs, where y_i is the ground-truth label of image $\boldsymbol{X}_i \in \mathbb{R}^{M_1 \times M_2 \times M_3}$ with M_3 set to 1 and 3, respectively, for gray and color images. In our formalism, we used a pretrained embedding $\boldsymbol{\psi} : \mathcal{X} \to \tilde{\mathcal{X}}$ from input images, $\boldsymbol{X}_i$, to high-level visual feature vectors, $\boldsymbol{\psi}(\boldsymbol{X}_i)$. Rather than learn a mapping to classes y_i, we instead learn a mapping to linguistic features describing the classes. This can be viewed as an embedding, $\boldsymbol{\phi} : \mathcal{Y} \to \tilde{\mathcal{Y}}$, of the classes to a linguistic feature space, $\tilde{\mathcal{Y}}$. This is a distributed embedding in that a single class will have many features associated with it and each feature will be associated with many different classes. We are left with the relatively simple task of finding a mapping between visual feature vectors $\boldsymbol{\psi}(\boldsymbol{X}_i)$ and linguistic feature vectors $\boldsymbol{\phi}(y_i)$. To achieve this, we use a neural architecture $\boldsymbol{\pi}(\boldsymbol{\psi}(\boldsymbol{X}_i), \boldsymbol{W})$, where $\boldsymbol{W}$ are trainable weights chosen by minimising the loss (energy) function

$$\sum_{(\boldsymbol{X}_i, y_i) \in \mathcal{S}} \ell\left(\boldsymbol{\phi}(y_i), \boldsymbol{\pi}(\boldsymbol{\psi}(\boldsymbol{X}_i), \boldsymbol{W})\right) \tag{1}$$

where $\ell : \tilde{\mathcal{Y}} \times \tilde{\mathcal{Y}} \to \mathbb{R}$ is a single-sample loss function defined in the linguist feature space. To make a class prediction, we independently train an inverse mapping $\phi^\dagger : \tilde{\mathcal{Y}} \to \mathcal{Y}$ from linguistic features to classes. Thus, we can make class predictions using $\phi^\dagger(\boldsymbol{\pi}(\boldsymbol{\psi}(\boldsymbol{X}_i), \boldsymbol{W}))$.

Feature extraction: Within our framework, there is a freedom to choose the feature extraction models $\boldsymbol{\psi}$ (see the left of Fig. 2). To illustrate this flexibility, we have used both pretrained ResNet101 [37] and VGG16 [38] for visual feature extraction. In both cases, we cut the network before the final dense layers.

A. Class Embedding

A central component in our approach is the introduction of a meaningful K-dimensional embedding space, $\tilde{\mathcal{Y}}$. We consider a mapping $\boldsymbol{\phi}$ from class $y_i \in \mathcal{Y}$ to an embedding vector $\tilde{\boldsymbol{y}}_i \in \tilde{\mathcal{Y}}$, where each component of $\tilde{\boldsymbol{y}}_i$ is a linguistic attribute. In practice, $\boldsymbol{\phi}$ would be a probabilistic embedding describing the conditional probability of $\mathbb{P}(\tilde{\boldsymbol{y}}_i \mid y_i)$; however, obtaining this is difficult. Instead, we start from a matrix $\mathbf{A} \in \mathbb{R}^{C \times K}$ provided by experts (in our case, this was conveniently provided by the zero-shot learning community; see Table I, for an example, [39]), which can be interpreted as $\mathbb{E}(\tilde{\boldsymbol{y}}_i \mid y_i)$. In our approach, we also need to learn the "inverse mapping," $\phi^\dagger(\tilde{\boldsymbol{y}}_i)$, giving $\mathbb{P}(y_i \mid \tilde{\boldsymbol{y}}_i)$. We learn this mapping using an MLP (MLP_L in our model, see the right of Fig. 2), where our inputs are noisy vectors $\tilde{\boldsymbol{y}}_i$ (i.e., the rows of matrix $\mathbf{A}$) and our targets are the classes y_i. This mapping is learned entirely without seeing the training images and is then frozen. Although this is a rather simple approach, it is extremely fast to learn and leads to good performance.

B. Explainable Neural Networks

Below we introduce the strategies regarding how the DNN $\boldsymbol{\pi}$ in our proposed multilevel architecture (middle of Fig. 2) can be explainable. Note that $\boldsymbol{\pi} : \boldsymbol{\psi}(\boldsymbol{X}_i) \to \boldsymbol{\phi}(y_i)$, where $\boldsymbol{X}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Since $\boldsymbol{\phi}(y_i)$ is not unique, we train $\boldsymbol{\pi}$ to learn the match between features and attributes in an unsupervised way (regarding $\tilde{\mathcal{Y}}$) using the training set $\mathcal{S}$ (i.e., image/label pairs), with the trained MLP_L and pre-trained $\boldsymbol{\psi}$.

For all $\boldsymbol{X}_j \in \mathcal{X}$ used for testing, let $\hat{y}_j = \phi^\dagger(\boldsymbol{\pi}(\boldsymbol{\psi}(\boldsymbol{X}_j)), \boldsymbol{W})$ be the predicted class label of the test image $\boldsymbol{X}_j$. Let $\pi_k(\boldsymbol{\psi}(\boldsymbol{X}_j))$ be the kth attribute of $\boldsymbol{\pi}(\boldsymbol{\psi}(\boldsymbol{X}_j))$, where $k \in \{1, 2, \ldots, K\}$. We define $\boldsymbol{u}_j \in \mathbb{R}^K$ as the importance of the attributes for the test image $\boldsymbol{X}_j$ and evaluate it by taking the gradient of the predicted class label $\hat{y}_j$ with respect to every attribute of $\boldsymbol{\pi}(\boldsymbol{\psi}(\boldsymbol{X}_j))$, i.e.,

$$\begin{aligned} \boldsymbol{u}_j &= \left(u_j^1, u_j^2, \ldots, u_j^K\right) \\ &= \left(\frac{\partial \hat{y}_j}{\partial \pi_1(\boldsymbol{\psi}(\boldsymbol{X}_j))}, \frac{\partial \hat{y}_j}{\partial \pi_2(\boldsymbol{\psi}(\boldsymbol{X}_j))}, \ldots, \frac{\partial \hat{y}_j}{\partial \pi_K(\boldsymbol{\psi}(\boldsymbol{X}_j))}\right). \end{aligned} \tag{2}$$

Then, the top K^* largest of $\{u_j^k\}_{k=1}^K$ will be selected as the *salient linguistic attributes* for image $\boldsymbol{X}_k$. In this sense, $\boldsymbol{\pi}$

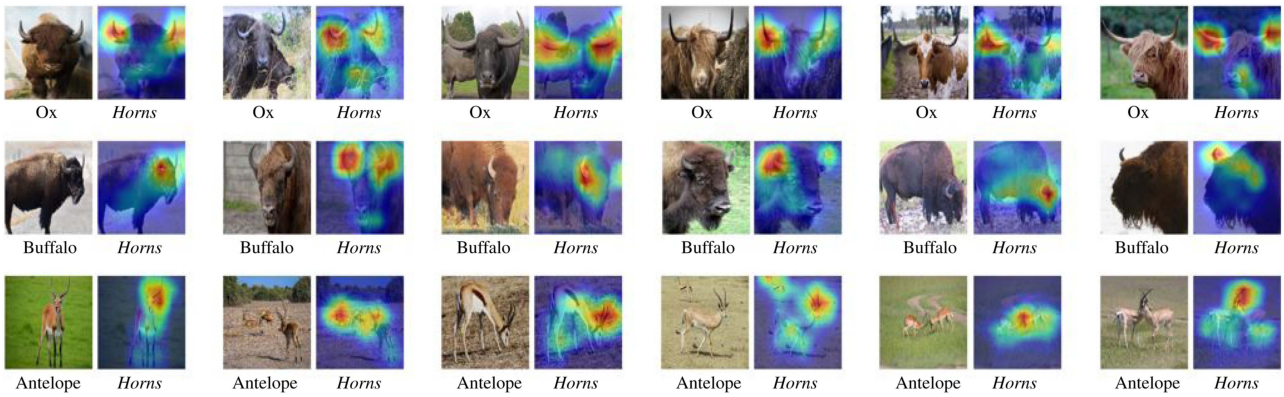

Fig. 5. Images from different classes in the AwA2 dataset with their obtained attribute-wise saliency maps by our approach. Examples from Ox, Buffalo, and Antelope classes show that the attribute *horns* is well captured by X-CNN.

TABLE I
EXCERPT OF THE ATTRIBUTE-CLASS MATRIX **A** FOR THE AWA2 DATASET

Classes \ Attributes	*Gray*	*Patches*	*spots*	*Lean*	*Tail*	*Strong*	*Muscle*	···
Antelope	12.34	16.11	9.19	39.99	40.59	33.56	26.14	···
Grizzly bear	3.75	1.25	0	0	9.38	78.48	48.89	···
Killer whale	1.25	68.49	32.69	22.68	41.67	63.35	10.45	···
Beaver	7.5	0	7.5	8.75	86.56	32.81	24.38	···
Dalmatian	0	37.08	100	63.68	53.75	34.93	23.75	···

Attribute values are in [0, 100] and standardized before use.

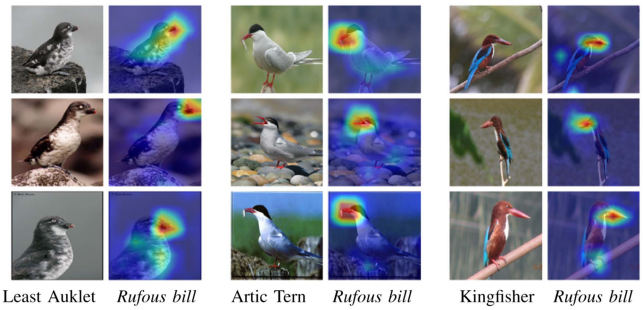

Fig. 6. Images from different classes in the CUB dataset with their obtained attribute-wise saliency maps by our approach. Examples from Least Auklet, Artic Tern, and Kingfisher classes show that the attribute *rufous bill* is well captured by X-CNN.

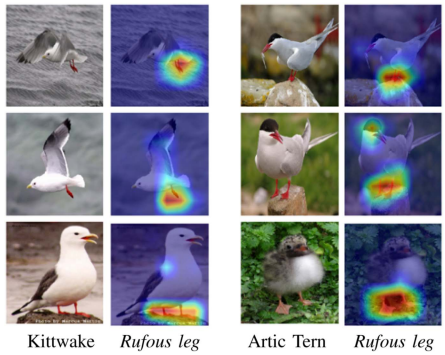

Fig. 7. Images from different classes in the CUB dataset with their obtained attribute-wise saliency maps by our approach. Examples from Kittiwake and Artic Tern classes show that the attribute *rufous leg* is well captured by X-CNN.

therefore can be explained by these salient linguistic terms. As examples, two of the most common neural networks—MLP and CNN—are adopted for $\boldsymbol{\pi}$ as follows.

1) Explainable MLP (X-MLP): Here, $\boldsymbol{\pi}$ represents an MLP, say $\boldsymbol{\pi}_{\mathrm{MLP}}$ consisting of a few dense layers. $\boldsymbol{\pi}_{\mathrm{MLP}}$ can then be explained by the obtained salient linguistic terms in a single level manner. We also call $\boldsymbol{\pi}_{\mathrm{MLP}}$ X-MLP.

2) Explainable CNN (X-CNN): Here, $\boldsymbol{\pi}$ represents a CNN, say $\boldsymbol{\pi}_{\mathrm{CNN}}$ consisting of convolutional layers with K channels followed by a global average pooling layer. $\boldsymbol{\pi}_{\mathrm{CNN}}$ can then be explained by the obtained salient linguistic terms. Moreover, we can also find out where these salient attributes are related in the given test image $\boldsymbol{X}_j$ by exploiting the spatial information preservation property of the CNN structure. For the ith attribute, $i \in \{1, 2, \ldots, K\}$, its attribute-wise saliency map (i.e., heat map mask) say $\bar{\boldsymbol{X}}_j^i$ can be obtained by the output of the last CNN layer after being upsampled to the same size of $\boldsymbol{X}_j$; in other words, the salient part of $\boldsymbol{X}_j$ is corresponding to the ith salient attribute $\bar{\boldsymbol{X}}_j^i$.

In contrast to $\boldsymbol{\pi}_{\mathrm{MLP}}$, $\boldsymbol{\pi}_{\mathrm{CNN}}$ provides both the salient linguistic terms and the corresponding attribute-wise saliency maps in a multilevel manner with no extra cost. For ease of reference, we call $\boldsymbol{\pi}_{\mathrm{CNN}}$ X-CNN. The procedures regarding training and test of our approach are summarized in Algorithm 1 in the supplementary material.

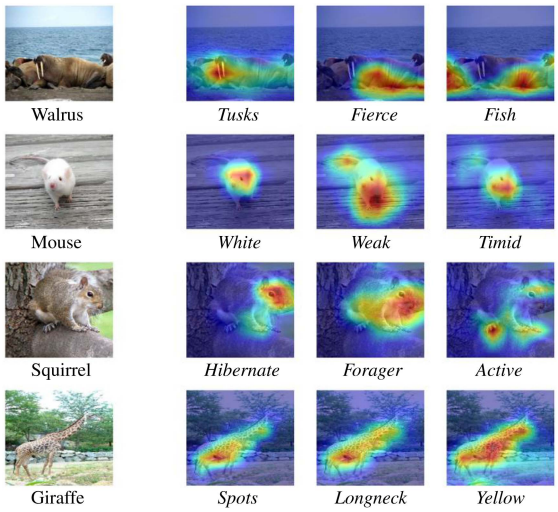

Fig. 8. Explainability of the proposed approach for correct class prediction. (Left) Randomly selected test images in the AwA2 dataset. (Right) Top three most salient attributes helping the neural network make the correct classification, and the corresponding attribute-wise saliency maps.

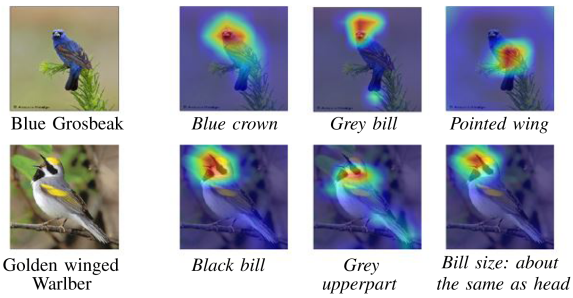

Fig. 9. Explainability of the proposed approach for correct class prediction. (Left) Randomly selected test images in the CUB dataset. (Right) Top three most salient attributes helping the neural network make the correct classification, and the corresponding attribute-wise saliency maps.

IV. Data

Animals with Attributes (AwA1) is a well-known dataset used in zero-shot learning [39]. Due to the absence of raw images and copyright issues, an alternative version of it named AwA2 was introduced in [32]. It is a medium-scale coarse-grained dataset with 37 322 images from 50 classes collected from public web sources, including 85 attributes per class available. Table I (see the supplementary material for its full version) presents a size of 5×7 excerpt of AwA2, exemplifying the nature between the attributes and different classes. The other benchmark dataset used in this work is caltech-UCSD birds (CUB)-200-2011 [40]. CUB is a fine-grained dataset containing around 11 800 images of 200 different bird classes, including 312 attributes per class. These linguistic attributes will be exploited to create self-explainable DNNs under our proposed methodology.

A. Generation of the Training Dataset $\mathcal{T}$ for $\mathrm{MLP_L}$

Recall that we have the attribute-class matrix $\mathbf{A} \in \mathbb{R}^{C \times K}$, where C and K represent the number of classes and the number of attributes per class, respectively. Table I in the supplementary material shows the full matrix $\mathbf{A}$ for the AwA2 dataset. The original matrix $\mathbf{A}$ can directly form a dataset, i.e.,

$$\mathcal{T}_0 = \left\{ (\tilde{\boldsymbol{y}}_k, y_k) \mid \tilde{\boldsymbol{y}}_k \in \tilde{\mathcal{Y}}, y_k \in \mathcal{Y}, k = 1, 2, \ldots, C \right\} \quad (3)$$

but this is too small to train $\mathrm{MLP_L}$; note that $\tilde{\boldsymbol{y}}_k = (\tilde{y}_k^1, \ldots, \tilde{y}_k^K)$ is the kth row of $\mathbf{A}$ for class k, and $\tilde{y}_k^i$ is the ith attribute of $\tilde{\boldsymbol{y}}_k$.

We augment $\mathcal{T}_0$ by upsampling each sample $(\tilde{\boldsymbol{y}}_k, y_k) \in \mathcal{T}_0$ to $\hat{C}$ number of samples by randomly manipulating $\hat{K}$ number of attributes of $\tilde{\boldsymbol{y}}_k$ among the total K, with the aim of perturbing the original samples. The values of the selected attributes for manipulation can be conducted randomly. In our setting, we use two values $\beta_0 > 0$ and $\beta_1 < 0$, and change the positive values of the selected attributes to β_1, otherwise, to β_0. Note that this setting is an arbitrary choice (here, we use $\beta_0 = 1.5$ and $\beta_1 = -0.5$), which can be replaced by other ways appropriate. We finally generate a training dataset $\mathcal{T}$ with $C\hat{C}$ ($\hat{C} > 0$) number of samples. In our experiments, $\hat{C}$ is set to 100. The data generation process is summarized in Algorithm 2 in the supplementary material.

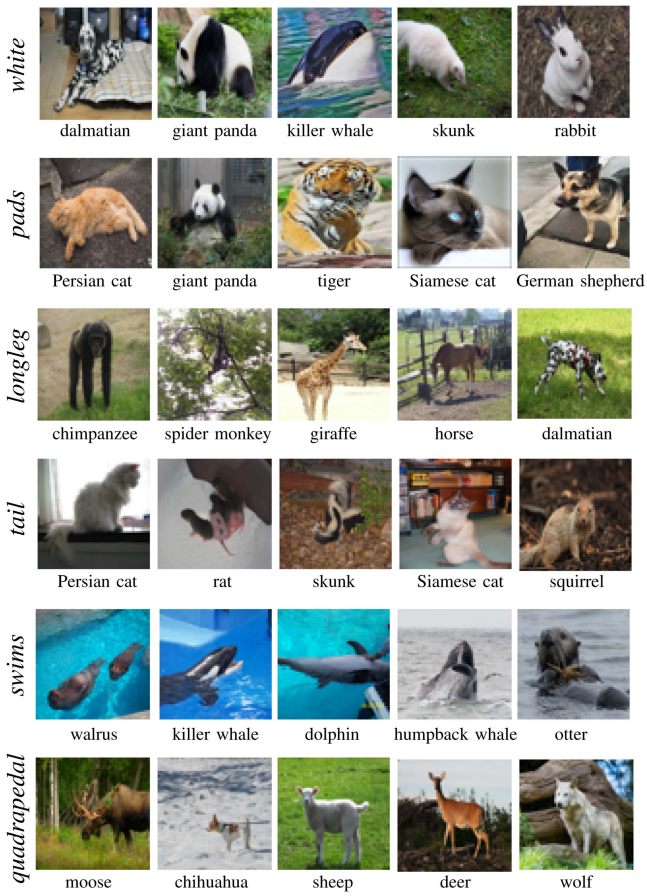

Fig. 10. Top five classes that maximally activate the given individual attributes on the left. For better illustration, one representative image from each class is also shown together with the class name.

V. Experiments

All the experiments were implemented on a personal laptop with the following specifications: i7-8750H CPU; GeForce

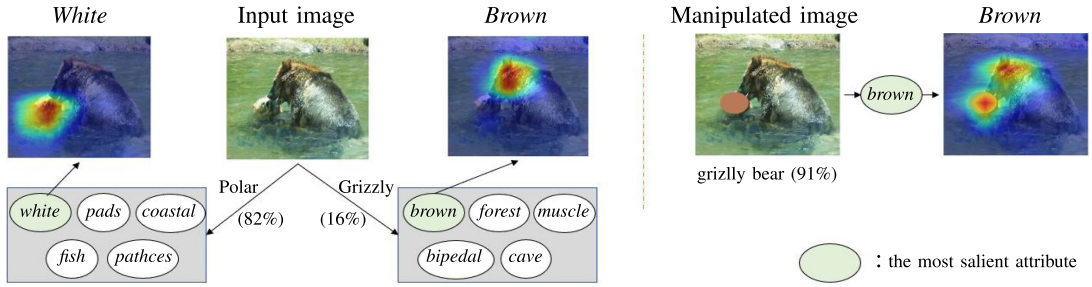

Fig. 11. Explainability of the proposed approach for incorrect prediction. (Left) Grizzly bear, which is misclassified as polar bear. Top five salient attributes are shown in ellipses for the highest probable class (i.e., polar bear) and second class (i.e., grizzly bear). The most salient attributes (i.e., white and brown) and their saliency maps provide insights of the prediction. (Right) Manipulated grizzly bear image (obtained by replacing the area related to the attribute "white" by a brown patch), which is then correctly classified by our approach with high confidence (91%).

GTX 1060 GPU; and 16-GB RAM. The proposed methodology is trained and tested on the coarse-grained and fine-grained benchmark datasets. Training of $\mathrm{MLP_L}$ takes around 15 min. Training of X-MLP and X-CNN takes around 30 min and 80 min, respectively. The pretrained feature extraction models (i.e., ψ) ResNet101 and VGG16 are downloaded from Keras' website.[1] The implementation setup and results in the XAI regime are given follows.

A. Implementation Setup

1) For the feature extraction model ψ, pretrained ResNet101 and VGG16 are, respectively, used for datasets AwA2 and CUB. The sizes of the extracted features for each image in datasets AwA2 and CUB are, respectively, $8 \times 8 \times 2048$ and $8 \times 8 \times 512$.
2) The language model $\mathrm{MLP_L}$ is a few layers wide MLP (here, three layers are used). To train it, two training sets, $\mathcal{T}$, with size of 5000 and 20 000, respectively, for datasets AwA2 and CUB are formed. $\mathrm{MLP_L}$, including the order of the attributes, is frozen after the training completes.
3) $\boldsymbol{\pi}_{\mathrm{MLP}}$ is a few layers wide MLP (here, four layers are used) taking 2048 and 512 features extracted by ψ and outputting 85 and 312 attributes for datasets AwA2 and CUB, respectively. $\boldsymbol{\pi}_{\mathrm{CNN}}$ for simplicity is set to one single convolutional layer with the size of $8 \times 8 \times 85$ and $8 \times 8 \times 312$ for datasets AwA2 and CUB, respectively. A $30/70$ split of the data was formed for training/test.
4) For comparison, *fine-tuned* ResNet101 and VGG16 are obtained by directly using the training set of image/label pairs of AwA2 and CUB, respectively.

The Adam optimizer with a learning rate of 0.001 and batch size of 32 is used in all experiments. The number of epochs is set to 100 and early stopping is applied (with patience set to 10 based on the validation loss). We stress that our main goal is to make DNNs self-explainable rather than accuracy driven. It is expected that the prediction performance reported could be improved for example with hyperparameter fine tuning and/or wiser selection of the feature extraction model ψ.

[1]Pretrained ResNet101 and VGG16: https://keras.io/api/applications/

B. Classification Performance of $\mathrm{MLP_L}$

Although our main focus in this article is XAI rather than the classification accuracy of $\mathrm{MLP_L}$, it is worth evaluating the classification accuracy performance of $\mathrm{MLP_L}$ under the training dataset $\mathcal{T}$. To do so and also investigate the impact of the parameters $\hat{C}$ and $\hat{K}$, we first generate different training dataset $\mathcal{T}$ using different upsampling rate $\hat{C}$ and different value of $\hat{K}$ (i.e., the number of attributes whose values are manipulated) ranging from 0 to 65 (i.e., up to two thirds of the total 85 attributes per class in the AwA2 dataset). Afterwards, to evaluate the accuracy performance of $\mathrm{MLP_L}$ using the generated datasets $\mathcal{T}$ regarding different $\hat{C}$ and $\hat{K}$, we further split each dataset $\mathcal{T}$ into two parts with the ratio of 70/30 for training and test, respectively.

Fig. 3 shows the classification accuracy of $\mathrm{MLP_L}$ corresponding to different $\hat{K}$ and $\hat{C}$ values. It is seen that the accuracy decreases when $\hat{K}$ becomes larger for each $\hat{C}$, which is reasonable, since the larger the $\hat{K}$, the higher the perturbation of the ground-truth attribute-class samples. The results also show that larger $\hat{C}$ (which leads to a bigger dataset $\mathcal{T}$) results in more robust models against more noisy samples, e.g., see the yellow line in Fig. 3 for the case of $\hat{C} = 100$. In contrast, when $\hat{C}$ is small, there is a significant classification performance drop as shown in Fig. 3; e.g., the classification accuracy drops from over 90% (for $\hat{C} = 100$) to 65% (for $\hat{C} = 10$) when $\hat{K} = 32$.

Finally, for completeness, we investigate the case of no perturbation, i.e., the case of $\hat{C} = 0$. In this case, Table I in the supplementary material, $\mathcal{T}_0$, is directly used for the $\mathrm{MLP_L}$ training. We generate test sets following the same way of generating $\mathcal{T}$ by manipulating the samples in $\mathcal{T}_0$ once with different $\hat{K}$ values. As shown in Fig. 3 for $\hat{C} = 0$, the model's performance drops significantly when $\hat{K}$ becomes larger, indicating the limited performance of the model trained just by using the original set $\mathcal{T}_0$. In our XAI experiments, we pick $\hat{C} = 100$ and $\hat{K} = 8$ and remark that there is a freedom of choice for these hyperparameters to obtain better results.

C. Classification Performance of the Proposed Architecture

Table II shows that the proposed multilevel XAI architecture in Fig. 2 can achieve surprisingly good performance (i.e., over 90% and $\sim 50\%$ accuracy for the 50-class and 200-class datasets AwA2 and CUB, respectively) in classification accuracy against

TABLE II
CLASSIFICATION ACCURACY

Data	Model	Test Accuracy	Explainability
AwA2	ResNet101	95.8 ± 1.3	N/A
	X-MLP	90.5 ± 0.8	Unilevel
	X-CNN	90.1 ± 1.1	Multilevel
CUB	VGG16	57.2 ± 1.4	N/A
	X-MLP	54.9 ± 1.5	Unilevel
	X-CNN	44.6 ± 1.1	Multilevel

X-MLP/X-CNN can achieve comparable performance against the fine-tuned ResNet101 and VGG16, which, however, lack linguistic and visual explainability that X-MLP/X-CNN delivers.

the fine-tuned neural networks (i.e., ResNet101 and VGG16 trained directly on the labeled data, which lack explainability) on hold-out test set even though this is not the main aim of this work. The neural networks' performance in accuracy highly depends on the quality of the data acquired. However, most of the data researchers work on, if not all, could be biased, insufficient and/or sensitive. Creating architectures that can explain themselves and simultaneously reach high prediction performance—just like the one introduced in this work—is arguably the long-term pursuit in machine learning.

D. Explainability Performance of Our Multilevel XAI Method

1) Explainability for Correct Prediction: For a given image Least Auklet from the CUB dataset (see Fig. 1), both the fine-tuned DNN (VGG16 in our case) and our proposed method can easily classify it correctly. However, the fine-tuned DNN gives no explanation on why it reaches a decision by itself. *Post hoc* XAI methods (such as [16], [17], [29]) could be employed to see whether the classified object as a whole in the given image is the main part that the fine-tuned DNN focuses on (i.e., left of Fig. 1), but this level of explanation is rather limited and is an incomplete reflection of human-like explanations as discussed throughout this article. In contrast, the attribute-wise level of explanation that the proposed multilevel XAI model delivers (i.e., right of Fig. 1) is much richer, wider, deeper, and self-explainable thanks to the linguistic attributes. In detail, some of the most salient attributes that affect the prediction are presented as striped belly, rufous bill, and gray breast. Their corresponding saliency maps convincingly highlight the correct part of the image for the mentioned individual attributes. This type of explanation is desirable and is an important indicator of the match between image features and class-wise attributes that are learnt in an unsupervised way by the proposed architecture (i.e., unsupervised in the sense that the training images have not been labelled by linguistic attributes and/or salient regions have not been given).

Fig. 4 demonstrates the power of the proposed model in explainability with more challenging images. Linguistic self-explainable attributes of stripes and group are outputted as salient for zebra, while quadrupedal and horns are outputted for antelope by X-MLP and X-CNN. Attribute-wise saliency maps for horn and group outputted by X-CNN show the human-like explanation power of our approach. Some attributes can be well captured by DNNs with examples shown in Figs. 1, 4, and 14. To further demonstrate this property, we present as follows more images from a variety of classes, showing that the presented attributes are indeed learnt rather than special to the given images or classes.

The attribute-wise saliency maps, for example, in Fig. 5, show that the attribute *horns* is clearly learnt for the Ox, Antelope, and Buffalo classes; for more results, see Figs. 6 and 7. Further examples of the correct class prediction with the multilevel explainability by our approach are shown in Figs. 8 and 9. For abstract attributes, their saliency maps could give us an idea of what part of the image activates that specific attribute, e.g., active or weak. Moreover, the attribute values given by experts in Table I for the predicted classes indicate whether experts think these attributes are helpful to discriminate one class from the others. After checking, we can see these salient attributes obtained by our approach for the predicted classes are indeed meaningful.

2) Explainability Regarding Attribute-Class Prediction: Although the class-wise attributes are learnt in an unsupervised way by our approach, the previous experiments have shown the power of our model in activating meaningful attributes. To further demonstrate its ability in attribute prediction, we present the attribute-class prediction averaged over the test samples for each class; see Table II in the supplementary material. The predicted attribute values in this table are expected to be close to Table I (i.e., the ground-truth) in the supplementary material. Moreover, Fig. 10 shows the top five classes that maximally activate the given individual attributes (i.e., white, pads, etc.), together with one representative image from each class. It shows that the maximally activated attributes are indeed meaningful and highly relevant to the classes that activate them. More results are given in the supplementary material.

3) Explainabilility for Incorrect Prediction: Reaching 100% prediction accuracy is not the case for any method given a nontrivial task. Therefore, investigating the reason behind wrong predictions is equally important. The left of Fig. 11 shows a grizzly bear, which is misclassified as polar bear, and to understand the reason behind, the top five salient attributes obtained by the proposed approach for both grizzly bear and polar bear classes are presented. These attributes are indeed the ones that differentiate these two classes (*cf.,* the full attribute-class matrix in Table I in the supplementary material). The attribute-wise saliency maps of the most salient attributes (i.e., white and brown) for both classes provide further insights regarding why this misclassification occurred. After occluding the part considered as "white" (by our approach) with a brown patch, the manipulated image is then correctly classified by our approach as grizzly bear with a high confidence (see the right of Fig. 11); furthermore, the saliency map for the most salient attribute, "brown," now indeed highlights both the head of the bear and the brown patch, showing that our approach clearly learns what brown is and considers it as a strong indicator of grizzly bear class.

Further examples of incorrect class prediction with the multilevel explainability by our approach are shown in Figs. 12 and 13. Grizzly bear classified as polar bear and whale classified as dolphin are some of the most frequent misclassification cases

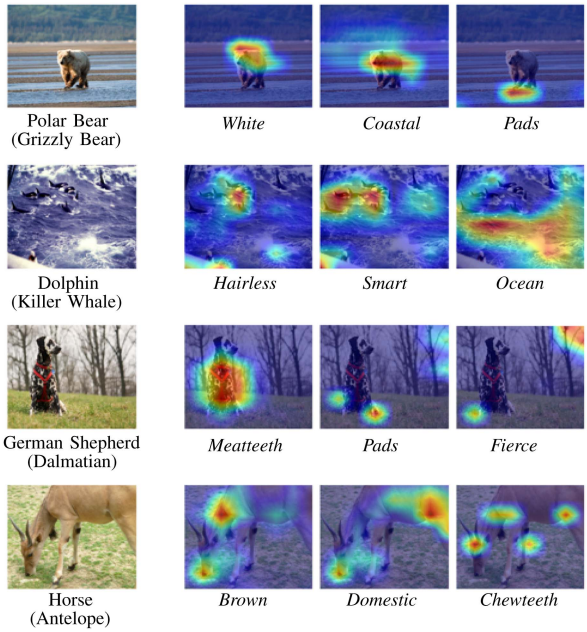

Fig. 12. Explainability of the proposed approach for the incorrect class prediction. (Left) Randomly selected test images in dataset AwA2; e.g., Polar Bear (Grizzly Bear) means the Grizzly Bear class is incorrectly predicted to be Polar Bear. (Right) Top three most salient attributes helping the neural network make the incorrect classification, and the corresponding attribute-wise saliency maps.

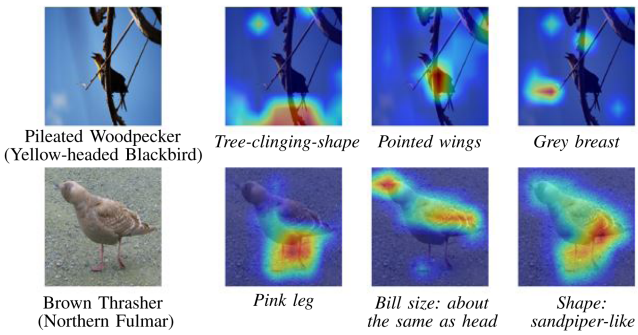

Fig. 13. Explainability of the proposed approach for incorrect class prediction. (Left) Randomly selected test images in dataset CUB. () Top three most salient attributes helping the neural network make the incorrect classification, and the corresponding attribute-wise saliency maps.

detected. After checking the attribute values in Table I given by the experts for the predicted classes and the ground-truth classes, we can see these salient attributes obtained by our approach are indeed consistent with the ones given by experts for the predicted classes; see also more discussion in Section VI for the challenges in, e.g., the linguistic alignment and the nature of explainability.

4) Sensitivity Between Attributes and Features: To further investigate the effectiveness of the linguistic attributes in our method in explainability, we test a zebra image and its artificial conversion to a horse using CycleGan [41], i.e., the attribute *stripes* is removed from the zebra; see Fig. 14. Again, all three models (i.e., fine-tuned ResNet101, X-MLP, and X-CNN) classified the zebra image as zebra and the artificially generated horse as horse. At this point, the fine-tuned ResNet101 has no explanation ability to show what changed in the original image

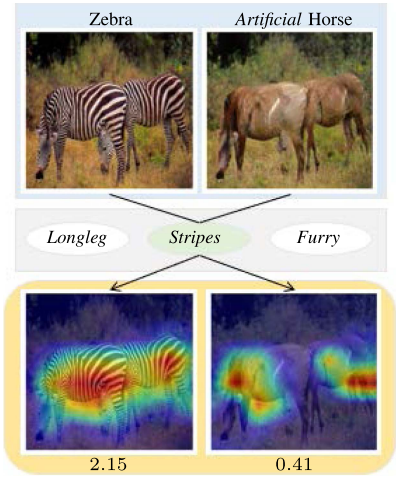

Fig. 14. Effectiveness of linguistic attributes in our approach. "Stripes" is one of the salient attributes for zebra with value of 2.15 and its saliency map reasonably highlights the body of zebra. For the artificial horse image, the attribute "stripes" is of value 0.41 and its saliency map is meaningfully unrelated.

that forces it to output "horse." In contrast, our model clearly shows that "stripes" is one of the salient attributes for the original zebra image with the attribute value of 2.15 and it drops to 0.41 for the artificially generated horse image. To validate the reason behind this visually, the attribute-wise saliency maps, generated by our approach with *no extra cost*, indicate that the X-CNN model focuses on the body of the zebra where the "stripes" lie, whereas arbitrary parts of the artificially generated horse image are highlighted when asked to show where the stripes are; see the bottom of Fig. 14.

5) Class Embeddings With Shuffled Attribute Values: Previous experiments are conducted using the prior information (i.e., the attribute-class matrix shown in Table I) provided by experts. An interesting and natural question is: what will the results be if the attribute-class matrix takes different values? In other words, to what extent, will the prior information in the attribute-class matrix be helpful in interpretability? To investigate this, we first shuffled all the columns of the attribute-class matrices for both datasets. In an extreme case, say the values of "ground" and "water" for the tiger class may be switched, which apparently would cause a dramatic information loss against the one provided by experts. The shuffled datasets are then used to train the model $\mathrm{MLP_L}$. Surprisingly, we found that it converged as fast as using the original data; moreover, the newly trained models X-MLP and X-CNN also reached an accuracy close to the ones obtained by using the original data. Fig. 15 shows the interpretability results provided by our approach in this attributes shuffling scenario. Obviously the linguistic attributes become meaningless; moreover, we also observe that the salient regions no longer appear to be associated with the object being recognized and defy an easy human explanation in contrast to what was observed in Fig. 1. This finding by our approach shows that those predetermined attribute lists are crucial to explainability. It also suggests that purely relying on the accuracy of DNNs (which might be trained on data with unknown flaws) could

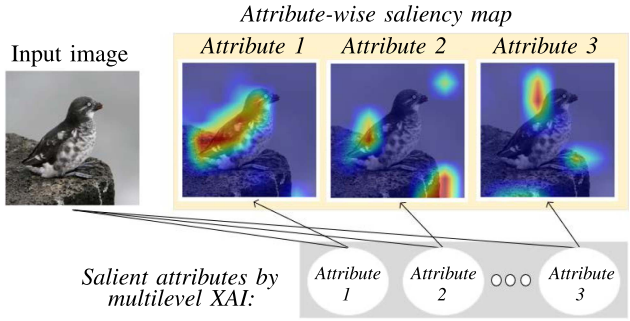

Fig. 15. Explainability of the proposed approach in the scenario of attributes shuffling (*cf.*, Fig. 1). Model accuracy reaches the same level as when we use the true linguistic attribute values, whereas the attribute-wise saliency maps are meaningless to humans, illustrating the importance of prior knowledge of the attribute values.

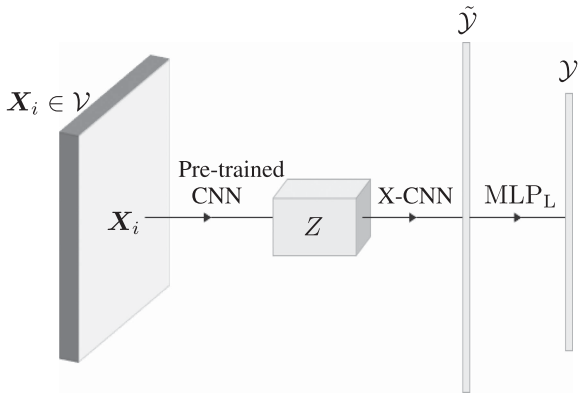

Fig. 16. Simplified version of the proposed multilevel XAI architecture. An image $\boldsymbol{X}_i$ with a class label in $\mathcal{Y}$ is drawn from the dataset $\mathcal{V} \subseteq \mathcal{X}$. It is then passed to a pretrained CNN. After passing through the last CNN layer, Z, it is passed to X-CNN, which produces a linguistic representation in $\tilde{\mathcal{Y}}$. Finally, it is passed to the $\mathrm{MLP_L}$ with a softmax output, giving a predicted label in $\mathcal{Y}$ for $\boldsymbol{X}_i$.

be perilous and the corresponding interpretability is essential. Further discussion is in Section VI.

6) Information of Linguistic Attributes: We now study the information of linguistic attributes, which will give us some guidance about the relationship between each class and its attributes.

For a set of classes $\mathcal{Y}$ (in our case, the 50 species in the dataset AwA2), let Y be a random variable describing our classes, and define the probability of a class having class label $y \in \mathcal{Y}$ as $\mathbb{P}[Y = y]$. Then, the entropy (or uncertainty) in the class label is

$$H_Y = -\sum_{y \in \mathcal{Y}} \mathbb{P}[Y = y] \log \left(\mathbb{P}[Y = y] \right). \tag{4}$$

To compute the mutual information for a linguistic attribute, $\tilde{y}^k$, we need to know $\mathbb{P}[Y \mid \tilde{y}^k]$, which can be estimated from a proper image set $\mathcal{V} \subseteq \mathcal{X}$. If we choose an image $\boldsymbol{X}_i \in \mathcal{V}$, we can feed it in our network (i.e., see Fig. 16) to find its attribute value, $\tilde{y}_i^k, 1 \leq k \leq K$. To simplify the calculation, let us set a threshold on the attribute value so that if it is above the threshold, we treat $\tilde{y}_i^k = 1$, otherwise $\tilde{y}_i^k = 0$. For each $y \in \mathcal{Y}$ and each $\alpha \in \{0, 1\}$, we can estimate various probabilities, e.g.,

$$\mathbb{P}\left[Y = y, \tilde{y}^k = \alpha\right] \approx \frac{\sum_{\boldsymbol{X}_i \in \mathcal{V}} [\![\boldsymbol{X}_i \text{ in class } y]\!] \, [\![\tilde{y}_i^k = \alpha]\!]}{|\mathcal{V}|}$$

$$\mathbb{P}\left[\tilde{y}^k = \alpha\right] \approx \frac{\sum_{\boldsymbol{X}_i \in \mathcal{V}} [\![\tilde{y}_i^k = \alpha]\!]}{|\mathcal{V}|} \tag{5}$$

where $[\![\text{predicate}]\!]$ is an indicator function (equal to 1 if the predicate is true and 0 otherwise). We can then compute

$$\mathbb{P}\left[Y = y \mid \tilde{y}^k = \alpha\right] = \frac{\mathbb{P}\left[Y = y, \tilde{y}^k = \alpha\right]}{\mathbb{P}\left[\tilde{y}^k = \alpha\right]}. \tag{6}$$

The conditional entropy of the classes given the (binarized) linguistic attribute is given by

$$H_{Y|\tilde{y}^k} = -\sum_{\tilde{y}^k \in \{0,1\}} \sum_{y \in \mathcal{Y}} \mathbb{P}\left[Y = y, \tilde{y}^k = \alpha\right] \log \left(\mathbb{P}\left[Y = y \mid \tilde{y}^k = \alpha\right] \right). \tag{7}$$

The mutual information on the classes due to the linguistic attributes is given by

$$I\left(Y; \tilde{y}^k\right) = H_Y - H_{Y|\tilde{y}^k}. \tag{8}$$

Calculating the mutual information by attributes $\tilde{y}^k$ predicted by our approach and (8) gives us a measure of how important each attribute is in differentiating the classes; see Table III in the supplementary material. Some of them seem useless by themselves as seen in Table III in the supplementary material; however, they could be important in combination with others.

An alternative way to calculate the mutual information is by following the same steps above but obtaining $\tilde{y}^k$ values from the attribute-class matrix given in Table I instead of dataset $\mathcal{V}$. These alternative mutual information values represent the importance of each attribute for human experts who created Table I. The calculated alternative mutual information is also presented in Table III in the supplementary material.

We now have two mutual information values per attribute, i.e., one is by using the dataset $\mathcal{V}$ (here, we use the test set) and the other is by using the prior knowledge given in Table I; see Table III in the supplementary material. Comparison between these two values for each attribute is helpful to see the discrepancy between what attributes people think are important and the ones that our trained DNNs learn. To give an example, "small" reduces the uncertainty by 0.98 bits when using the table values given by experts. However, it takes the value of 0.13 when using images, which suggests that "small" is not one of the best attributes for the trained DNN to differentiate the given classes, although people think it is. Further discussion is in the following section.

VI. Discussion and Limitations

To the best of our knowledge, the proposed approach is totally novel to XAI. It also raises a number of questions that are not commonly addressed in this context. Many of these open research questions we believe are important to the further development of XAI. Some of these are outlined as follows.

1) *Linguistic alignment:* We train the X-CNN network to find some visual features that separate the set of animals with high scores for a linguistic feature from those with low scores. We hope in doing so that the feature we learn captures some salient aspect of the linguistic feature. The extent to which we succeed we call the "linguistic alignment." Because we are learning this correlation in an unsupervised manner, we are not guaranteed that the visual feature aligns correctly with the linguistic label. This is likely to improve if we were to use more classes. On a small database of animals dominated by undulates and fish, it would be easy to confuse the linguistic term "hooves" with "legs." Such confusion is much less likely if the dataset contained other animals, such as dogs or primates.
   The linguistic alignment is complicated as a DNN is likely to assess circumstantial or contextual evidence for the existence of an attribute. As hooves are highly correlated with legs, it would not be surprising if the legs were considered highly salient to the presence of hooves in an image. This appears to be common in many of the attributes, where saliency maps appear to take in a much larger area than the feature described by the linguistic attribute. At some level, this clearly makes sense. A hoof-like object that is not attached to a leg is unlikely to be a hoof. Similarly, where a hoof is occluded, there may be enough context to infer that the animal is very likely to be hooved. However, this contextual evidence reduces the linguistic alignment. This is a feature of explainability rather than a fault of our approach. However, an important line of research in XAI is to separate circumstantial and contextual evidence from direct evidence.
2) *The nature of explainability:* Explanations for classifications are not unique. This was shown when we randomized the elements in the matrix, $\mathbf{A}$, between classes and linguistic attributes. In doing so, it seems highly unlikely that any attribute has a simple linguistic description. Yet, we can train our model with these random attributes and still obtain classification levels of around 90% in AwA2. This seems at first sight counter-intuitive, although given that we have 85 continuous features they have the potential to carry sufficient information to separate the classes with high accuracy. What separates, at least, some of the true linguistic attributes from other attributes is the information content of the linguistic attributes; that is the linguistic attributes that have a high mutual information in regard to the classes. However, some of these attributes are hard for a DNN to learn.
3) *What attributes DNNs learn:* We have shown examples of attributes such as "rufous bill" that appear to be well captured by the networks. However, through using an independent test set we find that some of the linguistic attributes appear not to have been learned by the network. An example of this is "big," clearly an attribute that would be useful for humans to distinguish an elephant from a squirrel. Because of the nature of the training set, where objects tend to be resized to fill most of the image, this turns out to carry little information about the classes. However, the lack of success of these attributes is very informative regarding how DNNs perform a discriminative task. All the linguistic attributes used in the data are chosen by experts because linguistically they carry considerable mutual information about the classes. The failure of DNNs to exploit some of these terms obtained by our approach conveys important insights that directly address the issue of what information a neural network is actually learning—a core concern of XAI.
4) *Tangible versus abstract attributes:* The linguistic attributes used in this article (that we inherited from the zero-shot learning community) interestingly incorporate both tangible and abstract attributes. For the tangible attributes such as "horn," we would expect the corresponding saliency map to highlight horn (although as we have argued it may highlight areas that are important contextual clues to the presence of a horn). The more abstract attributes such as "domestic" or "fast" are less easily attributed to a particular area of the image. They may however be highly informative, for example, in differentiating between cat and lion. When these attributes are informative, then it is clearly important to understand whereabouts in the image these attribute are inferred. Again our approach goes some way toward addressing this issue.
5) *Atomic and compound attributes:* In our approach, we have treated all attributes as atomic. Consequently, pointy, fluffy, and large ears would all be treated as separate linguistic attributes. However, each attribute would correspond to the same area of the image, and in many cases, it seems more natural to treat attributes as compound entities. We have not attempted to do this, but if we wish to scale up our approach to larger datasets, this seems to us to be an important area of future research.

VII. Conclusion

High-performance DNNs are highly desirable when they can reason about their decisions. We presented a new XAI methodology—multilevel XAI—with self-explainable models delivering human-like multilevel explanations alongside the class probabilities. Explaining why a certain prediction is made using linguistic terms and attribute-wise saliency maps without requiring per-image ground-truth explanations in the training phase makes the proposed technique efficient and inexpensive. The results in explainability demonstrated by the match between image features and class embeddings greatly empower the explainability of DNNs while preserving their prediction ability at a reasonable level. Given the importance of XAI and the power of the newly introduced approach, we believe this could spark new avenues in XAI and shed light on developing and applying AI in more sensible ways.

References

[1] 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.

[2] H. Hagras, "Toward human-understandable, explainable AI," *Computer*, vol. 51, no. 9, pp. 28–36, 2018.

[3] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[4] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[5] O. Loyola-Gonzalez, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154096–154113, 2019.

[6] W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowl.-Based Syst.*, vol. 263, 2023, Art. no. 110273.

[7] W. D. Heaven, "Hundreds of AI tools have been built to catch covid. None of them helped," MIT Technology Review, 2021. [Online]. Available: https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

[8] M. Roberts et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Mach. Intell.*, vol. 3, no. 3, pp. 199–217, 2021.

[9] P. McCausland, "Self-driving uber car that hit and killed woman did not recognize that pedestrians jaywalk," 2019. [Online]. Available: https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281

[10] T. Olavsrud, "7 famous analytics and AI disasters," 2022. [Online]. Available: https://www.cio.com/article/190888/5-famous-analytics-and-ai-disasters.html

[11] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[12] S. Tan, R. Caruana, G. Hooker, and Y. Lou, "Distill-and-compare: Auditing black-box models using transparent model distillation," in *Proc. AAAI/ACM Conf. Artif. Intell., Ethics, Soc.*, 2018, pp. 303–310.

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

[14] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable AI methods–A brief overview," in *Proc. xxAI-Beyond Explainable AI*, Vienna, Austria, 2022, pp. 13–38.

[15] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in *Proc. Brit. Mach. Vis. Conf.*, Durham, UK: BMVA Press, Sep. 3–6, 2018, Art. no. 151.

[16] 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.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2018, Art. no. 187.

[18] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[19] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. Int. Conf. Learn. Representations (Workshop Track)*, San Diego, CA, USA, May 2015.

[20] 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.

[21] A. Sarkar, D. Vijaykeerthy, A. Sarkar, and V. N Balasubramanian, "A framework for learning ante-hoc explainable models via concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10286–10295.

[22] P. W. Koh et al., "Concept bottleneck models," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5338–5348.

[23] R. Goebel et al., "Explainable AI: The new 42?," in *Proc. Int. Cross-Domain Conf. Mach. Learn. Knowl. Extraction*, Springer, 2018, pp. 295–303.

[24] D. H. Park et al., "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8779–8788.

[25] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 549–565.

[26] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[27] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," *Ann. Appl. Statist.*, vol. 9, no. 3, pp. 1350–1371, 2015.

[28] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Mach. Learn.*, vol. 102, no. 3, pp. 349–391, 2016.

[29] R. Ramprasaath et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[30] W. Wang, C. Han, T. Zhou, and D. Liu, "Visual recognition with deep nearest centroids," 2022, *arXiv:2209.07383*.

[31] T. Zhou, W. Wang, E. Konukoglu, and L. Van Gool, "Rethinking semantic segmentation: A prototype view," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2582–2593.

[32] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Sep. 2019.

[33] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1425–1438, Jul. 2016.

[34] A. Frome et al., "DeViSE: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[35] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1410–1418.

[36] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2152–2161.

[37] 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.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[39] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[40] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. 2010-001, 2011.

[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.