Multi-scale Interpretation Model for Convolutional Neural Networks: Building Trust based on Hierarchical Interpretation

Xinrui Cui, Dan Wang, and Z. Jane Wang, *Fellow, IEEE*

***Abstract*—With the rapid development of deep learning models, their performances in various tasks are improved; while meanwhile their increasingly intricate architectures make them difficult to interpret. To tackle this challenge, model interpretability is essential and has been investigated in a wide range of applications. For end users, model interpretability can be used to build trust in the deployed machine learning models. For practitioners, interpretability plays a critical role in model explanation, model validation, and model improvement to develop a faithful model. In the paper, we propose a novel Multi-scale INTerpretation (MINT) model for convolutional neural networks using both the perturbation-based and the gradient-based interpretation approaches. It learns the class-discriminative interpretable knowledge from the multi-scale perturbation of feature information in different layers of deep networks. The proposed MINT model provides the coarse-scale and the fine-scale interpretations for the attention in the deep layer and specific features in the shallow layer, respectively. Experimental results show that the MINT model presents the class-discriminative interpretation of the network decision and explains the significance of the hierarchical network structure.**

***Index Terms*—Model interpretability, multi-scale interpretation, convolutional neural networks, model-agnostic.**

I. Introduction

In the era of big data, deep learning algorithms have achieved unprecedented breakthroughs in a wide range of computer vision tasks such as image classification [1], object detection [2], image captioning [3], semantic segmentation [4], visual question answering [5] and multimedia data generation. The increasing popularity of deep learning coincides with the surge in the capabilities of black-box models whose internal mechanisms are difficult to explain. For convolutional neural networks (CNNs), interpretability is sacrificed to achieve better performance through higher abstraction. For example, deep ResNets [1] are over 200-layer deep and highly non-linear. However, should we trust a prediction only based on high accuracy? In recent years, CNNs have been applied in critical areas such as medical care, criminal justice, and finance, the inability of a human to interpret these models could be problematic. When such models fail, they fail without warning or explanation, leaving end users wondering the reasons. Therefore, it becomes imperative for us to know what CNNs have learned and explain how the decisions are made. Nowadays, model interpretability has drawn increasing attention because of its potential applications, such as building trust for real-world users and offering the insight of the black-box model for practitioners [6], [7], [8].

An interpretation model should help understand the black-box model by information extraction and analysis from the model being explained. Model interpretability plays a critical role in explaining modeling results, enhancing trust, further designing more faithful models, and setting related regulations. For end users, they need to understand how a black-box model makes its decision. Then they can decide whether to trust the model or not. For researchers in machine learning, they can benefit from model interpretability by gaining deeper insight into the model. Model interpretability also motivates theories about mechanisms of black-box models and facilitates the debugging of unexpected errors when building a sophisticated machine learning or deep learning model. It is worth mentioning that model interpretability is becoming new regulations, e.g., the EU General Data Protection Regulation.

Model interpretability has long been a research topic in the machine learning field. Some machine learning algorithms, such as linear model, SVM, and decision tree, are interpretable. Typically, these classical rule-based methods [9] are highly interpretable. Decomposable approaches where each stage is hand-crafted are interpretable as each component assumes an intuitive explanation. However, with problems to be solved becoming more and more complex, these algorithms are less attractive, because the hand-designed features and linear models are not capable of representing such complex relations.

As a result, many complex models surge and provide superior performance, such as random forest and CNNs. However, the excellent performance of complex models comes at the cost of interpretability because these models offer little transparency regarding their intermediate mechanisms. For example, thousands of neurons in CNNs jointly implement a complex nonlinear mapping from the input to the output. Therefore, CNNs are intrinsically difficult to understand and interpret. Despite the good performance of CNNs, many applications in healthcare and medicine require justifications from the models as to why and how they come to the results. In fact, this is a major reason why many applications still use simple linear models or decision trees. For such situations, model interpretability of CNNs could be a remedy and make the general public trust them.

This work was supported in part by the Canadian Natural Sciences and Engineering Research Council (NSERC), the Four Year Doctoral Fellowship and the International Doctoral Fellowship at the University of British Columbia. (*Corresponding author: Dan Wang.*)

X. Cui, D. Wang, and Z. Wang are with the Department of Electrical and Computer Engineering, University of British Columbia, BC, Canada. e-mail: (xinruic@ece.ubc.ca; danw@ece.ubc.ca; zjanew@ece.ubc.ca.)

In order to interpret complex models, post-hoc interpretability is introduced, which refers to the extraction and analysis of information from an already learned model. It is beneficial in qualitatively understanding the nature of features and predictions. Moreover, the learned model usually does not have to sacrifice its performance in order to be interpretable.

Prevalent visualization approaches to post-hoc interpretability can be broadly classified as gradient-based methods and perturbation-based methods.

In gradient-based approaches, the gradient of the output with respect to the input is used to construct the heatmap. Such approaches involve a single forward and backward pass through deep networks to obtain heatmaps. Simonyan et al. [10] measured the relative importance of pixels in the input image by calculating the gradient of the output regarding the input. In the backward pass, the gradients of neurons with negative inputs were suppressed when propagating through ReLU layers. Guided Backpropagation [11] was proposed to build heatmaps by suppressing the flow of gradients through neurons wherein either of the inputs or incoming gradients were negative through ReLU. Class activation mapping (CAM) [12] and Grad-CAM [13] incorporated network activations into visualizations. Concretely, they visualized the linear combination of activations and class-specific gradients in a late layer. The above methods introduce different propagation heuristics for image saliency. However, gradient-based methods (which can interpret the class-discriminative property of networks) rely on the gradient of the output to compute the class-discriminative importance. Since the gradient of the output in the shallow layer is not as reliable as that in the deep layer, gradient-based methods mainly focus on the deep layer, such as CAM [12] and Grad-CAM [13]. Other gradient-based methods which can interpret the feature in the shallow layer provide explanations without the class-discriminative property [11], [14].

Perturbation-based methods involve perturbing the input image and observing the change of prediction. The underlying principle is that if the pixels which contribute maximally to the prediction are changed, the prediction will drop by the maximum amount. Zeiler et al. [14] occluded different patches of the input image in regular grids and monitored the prediction of CNN and the activation of feature map in the last layer that was maximally activated for the un-occluded input image. Fong et al. [15] perturbed the sensitivity heatmap and monitored the change of prediction to refine the heatmap to minimum pixels that can preserve the prediction score. Ribeiro et al. [16] occluded different super-pixels of the input image and proposed an interpretation model, Local Interpretable Model-agnostic Explanations (LIME), to approximate the predictions of black-box models. Compared with gradient-based methods, the image-level interpretable representation in perturbation-based methods is understandable to our human. For example, the interpretable representation in LIME [16] is obtained by the unsupervised segmentation algorithm which considers the color, boundary, and texture information. This process reflects some kind of human understanding way to input images but does not take into account that the network transforms the input information into the features of different layers in its internal mechanism. Meanwhile, because the interpretation results of perturbation-based methods are the important image regions for the predictions, their performance is limited by the methods used to obtain the interpretable representations, such as the unsupervised segmentation method in LIME [16]. The explanation provided by these methods only interprets the network decision from the input-level view. Therefore, it also restricts the performance of the perturbation-based interpretation methods.

To the best of our knowledge, most current interpretation methods only provide the single-scale visual explanation. They only present input image regions that are critical to the predictions but do not provide hierarchical visual interpretation. For human perception, the single-scale visual explanation is not enough. Generally, human focuses on both the entirety and the locality of a given object when recognizing it. Therefore, multi-scale visual explanation should be more beneficial to human understanding. On the other hand, CNNs consist of hierarchical structure of layers. The shallow layers extract fine-scale features such as the texture and boundary, while the deep layers focus on high-level semantic features such as coarse-scale attention. Therefore, it is desirable to build a multi-scale interpretation model for the hierarchical structure of CNNs.

To address the above challenges, we propose a Multi-scale INTerpretation (MINT) model for CNNs which exploit the benefits of both gradient-based and perturbation-based methods. The core idea of the perturbation-based methods is to perturb the input, see how the predictions change and figure out which part is more important to the prediction. This is a benefit in terms of interpretability because we can perturb the input by changing components that make sense to our human. In comparison, the gradient-based methods can extract the interpretable semantic information from the internal features. Therefore, it can give insight into the internal mechanisms of networks. In order to take advantage of both methods, the proposed method utilizes features and corresponding gradients to guide the segmentation for interpretable representation. And then the interpretation model perturbs the feature information in the shallow and deep layers to compute the class-discriminative importance of the multi-scale interpretable super-pixels. Finally, the proposed interpretation model provides the fine-scale and coarse-scale explanations for the shallow and deep layers, respectively. From this point of view, our model can be regarded as the hybrid interpretation model which utilizes both perturbation mechanism and the multi-scale features together with their gradient information.

The main contributions of our work are summarized as follows:

1) The MINT model can learn the class-discriminative interpretable knowledge from fine-scale and coarse-scale perturbations of feature information in the shallow and deep layers. Therefore, the MINT model exploits both perturbation-based and gradient-based interpretation mechanisms. No architectural change or re-training of the deep network is needed.
2) MINT provides the coarse-scale and fine-scale interpretations for the target-class prediction to explain the coarse-scale attention in the deep layer and the fine-scale feature in the shallow layer, respectively. Therefore, MINT can

provide the class-discriminative explanation to the given input, and also help us understand the meaning of the hierarchical features in networks.

3) We propose a novel multi-scale interpretable representation guided by the class-discriminative feature and its gradient information of CNN. Therefore the MINT model can provide insight into the hierarchical structure of networks. It is more reasonable for the network explanation than that only based on visual input.
4) To learn the MINT model, we design a multi-scale linear sparse interpretation optimization problem and propose an inter-scale constraint as a regularization to cooperate the coarse-scale and fine-scale information. We also introduce the MINT algorithm to solve the optimization problem.

II. Methodology

In this section, we present the proposed MINT model in detail. Its overall goal is to provide an interpretable class-discriminative explanation by exploiting the feature perturbation information in the shallow and deep layers.

A. Multi-scale Interpretable Representation

Before we present the MINT model, it is essential to clarify what interpretable representation is and why multi-scale analysis is needed. Here, an interpretable explanation needs to illustrate what a black-box model has learned for its prediction in a human-understandable way. Intuitively, a raw input of a black-box model, such as text and image, is understandable to a human. Therefore, a possible interpretable representation for image classification is the image region. An interpretable explanation can be obtained by projecting the prediction back into the raw input space, describing which part of the input is most influential to the prediction. A multi-scale analysis is needed for the following reasons. Firstly, human understanding involves the perception of locality and entirety. For example, a possible hierarchical interpretation for image classification contains both the main body of an object and its local important feature. The explanation that incorporates fine scale and coarse scale can be more beneficial to human understanding. Secondly, deep networks usually learn hierarchical features. For example, CNN maps an input image, via multiple layers, to a probability vector over different classes. Convolution and pooling operations enable features in deep layers larger receptive fields. In other words, shallow layers usually represent local, fine-scale features while deep layers show more holistic features. It means the multi-scale understanding exists not only in visual cognition of human but also in the hierarchy structure of networks. Accordingly, explanations that incorporate fine scale and coarse scale can be more beneficial to human understanding and the model interpretation.

In the MINT model, it needs to obtain the multi-scale interpretable representation. In previous related studies, they use different methods to obtain the interpretable representation. For example, Zeiler [14] uses image patches in regular grids as the interpretable representation and Ribeiro [16] utilizes an unsupervised segmentation to obtain super-pixels. However, they are not the multi-scale interpretable representation and have no connection with the internal feature in networks.

One strategy is to apply different segmentation algorithms to get large super-pixels and small super-pixels for the multi-scale interpretable representation. However, it still does not explore the meaning of the internal features. In CNNs, the deep layers have the higher-level semantic feature than that in shallow layers. Therefore, by combining the features and their gradients in deep layers, the MINT model can obtain a class-discriminative saliency map of the target class to guide the coarse-scale segmentation for the interpretable representation. This segmentation, as the coarse-scale interpretable representation for the deep layers, is more sensible than the segmentation which only considers the input pixel values. The MINT model utilizes not only the coarse-scale and fine-scale interpretable representations but also the multi-scale feature perturbation in shallow and deep layers. Subsequently, the MINT model provides multi-scale explanation for the features in shallow and deep layers.

Coarse-scale super-pixels. The coarse-scale explanation is designed to interpret the attention in deep layers. Accordingly, in the MINT model, the class-discriminative saliency map is obtained by the linear combination of features weighted by their gradients of target-class prediction in the deep layer. Subsequently, the class-discriminative saliency map is applied to the GrabCut segmentation algorithm [17] as prior knowledge of the target object. Consequently, the region highlighted in the feature of the deep layer can be segmented as a coarse super-pixel for the target class. Then, the background region is segmented into coarse super-pixels by an unsupervised segmentation algorithm. The coarse-scale super-pixel set contains the coarse super-pixels for target and background, which is denoted as

$$CS = \{cs_1, cs_2, \cdots, cs_i, \cdots, cs_p\} \tag{1}$$

where cs_i is the i-th coarse super-pixel and p is the total number of coarse-scale super-pixels.

Fine-scale super-pixels. For each coarse-scale super-pixel, we continue segmenting it into small super-pixels for the fine-scale explanation. The fine-scale super-pixels of the whole image are denoted as

$$FS = \{fs_1, fs_2, \cdots, fs_j, \cdots, fs_q\} \tag{2}$$

The total number of small super-pixels is q.

Perturbed samples. Here, $\mathbf{X}$ is denoted as the original representation of an instance. We draw the perturbed sample $\mathbf{Z}$ around $\mathbf{X}$ by selecting its small super-pixels uniformly at random. For fine-scale super-pixel of each perturbed sample, we use $\mathrm{s}^f \in \{0,1\}^{q\times 1}$ as a binary vector for the fine-scale perturbation state, where each entry represents the absence (0) or presence (1) of a particular fine-scale super-pixel. For each fine-scale super-pixel in the perturbed image, its pixel values remain original when the small super-pixel exists in the perturbed image, otherwise, its pixel values are grayed out. For coarse-scale super-pixels of each perturbed sample, $\mathrm{s}^c \in \mathbb{R}^{p\times 1}$ is introduced as the coarse-scale state vector where each entry

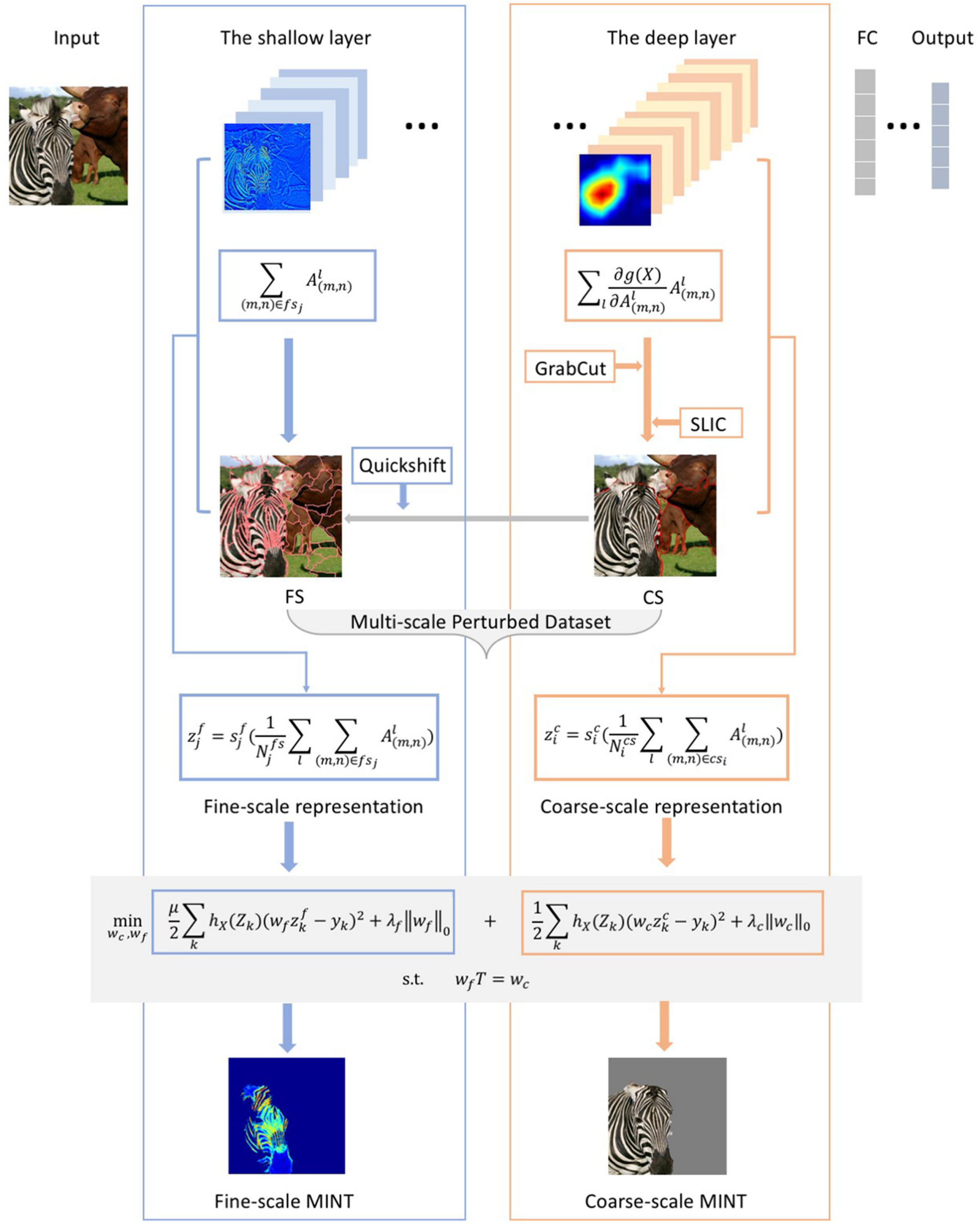
Input
The shallow layer
The deep layer
FC
Output
$\sum_{(m,n)\in fs_j} A^l_{(m,n)}$
$\sum_l \frac{\partial g(X)}{\partial A^l_{(m,n)}} A^l_{(m,n)}$
GrabCut
SLIC
Quickshift
FS
CS
Multi-scale Perturbed Dataset
$z^f_j = s^f_j(\frac{1}{N^{fs}_j}\sum_l \sum_{(m,n)\in fs_j} A^l_{(m,n)})$
$z^c_i = s^c_i(\frac{1}{N^{cs}_i}\sum_l \sum_{(m,n)\in cs_i} A^l_{(m,n)})$
Fine-scale representation
Coarse-scale representation
$\min_{w_c,w_f} \frac{\mu}{2}\sum_k h_X(Z_k)(w_f z^f_k - y_k)^2 + \lambda_f \|w_f\|_0 + \frac{1}{2}\sum_k h_X(Z_k)(w_c z^c_k - y_k)^2 + \lambda_c \|w_c\|_0$
s.t. $w_f T = w_c$
Fine-scale MINT
Coarse-scale MINT

Fig. 1. The process of the proposed MINT model.

represents the fraction of presented small super-pixels in a particular large super-pixel.

Coarse-scale representation for the deep layer. The MINT model perturbs the features in the deep layer to learn the coarse-scale explanation. In the deep layer, the coarse-scale representation is the average pooling of features weighted by the corresponding state value for each coarse-scale super-pixel. Consequently, the coarse-scale representation of the perturbed sample in the deep layer is denoted as

$$\mathbf{z}^c = \mathbf{s}^c \odot \left[a_1^{cs}\ a_2^{cs} \cdots a_i^{cs} \cdots a_p^{cs}\right]^T \tag{3}$$

where the a_i^{cs} is the average pooling feature of cs_i, and the operator $\odot$ denotes Hadamard product. The average pooling feature of cs_i is defined as

$$a_i^{cs} = \frac{1}{N_i^{cs}} \sum_l \sum_{(m,n)\in cs_i} \mathbf{A}^l_{(m,n)} \tag{4}$$

where the $\mathbf{A}^l_{(m,n)}$ is the feature value of l-th channel in coordinate (m,n), and N_i^{cs} is the pixel number of cs_i.

Fine-scale representation for the shallow layer. Similarly, in the MINT model, the features in the shallow layer are perturbed to learn the fine-scale interpretation of networks.

Therefore, the fine-scale representation of the shallow layer is defined as

$$\mathbf{z}^f = \mathbf{s}^f \odot \left[a_1^{fs}\ a_2^{fs} \cdots a_j^{fs} \cdots a_q^{fs}\right]^T \tag{5}$$

where the a_j^{fs} is the average pooling feature of fs_j. The average pooling feature of the j-th fine-scale super-pixel is defined as

$$a_j^{fs} = \frac{1}{N_j^{fs}} \sum_{l} \sum_{(m,n)\in fs_j} \mathbf{A}^l_{(m,n)} \tag{6}$$

where the N_j^{fs} is the pixel number of the j-th fine-scale super-pixel.

Multi-scale perturbed dataset. To learn our multi-scale interpretation model, we build a multi-scale perturbed dataset by sampling in locality. It is similar to the work [16], but utilizes multi-scale feature information to linearly approximate the network prediction. The model being explained is defined as $g : \mathbb{R}^d \rightarrow \mathbb{R}$. It is noteworthy that in image classification we explain each class prediction separately for multiple classes, thus $g(\mathbf{X})$ is the prediction of the target class. The perturbed sample set $\mathcal{D}_{img} = \{\mathbf{Z}_k\}$ is generated by sampling state vectors of super-pixels for the given instance $\mathbf{X}$. Given a perturbed sample $\mathbf{Z}$, we feed it into the model being explained and obtain $g(\mathbf{Z})$. Finally, we get the perturbed dataset $\mathcal{D}\{\mathbf{Z}_k, y_k, \mathbf{z}_k^c, \mathbf{s}_k^c, \mathbf{z}_k^f, \mathbf{s}_k^f\}$, where $y_k = g(\mathbf{Z}_k)$ is the prediction of a certain class for the k-th perturbed instance $\mathbf{Z}_k$.

B. MINT Model

In our approach, we adopt a local linear approximation perturbation approach which is similar with [16] to build our multi-scale interpretation model. The underlying principle of perturbation is that the image region which contributes maximally to the prediction, if occluded, would reduce the probability by the maximum amount. In the MINT model, we exploit the fine-scale and coarse-scale feature information in the shallow and deep layers to explain the internal mechanism of networks. The MINT model utilizes the multi-scale super-pixels guided by the features and gradients of networks as the perturbation units. The MINT model learns the behavior of the network by perturbing the multi-scale feature information in shallow and deep layers and observing how the prediction changes.

Now we will introduce the proposed MINT model. First, we define the coarse-scale and fine-scale linear interpretation functions as f_{coarse}, f_{fine}. Second, $\phi(f_{coarse}, f_{fine})$ is introduced as a constraint of coarse-scale and fine-scale explanations. Owing to the relationship between different scale explanation, the cooperation between coarse-scale and fine-scale interpretation would help build more reasonable multi-scale interpretation model. We also denote $\psi(f_{coarse})$ and $\psi(f_{fine})$ as the measures of complexity of the explanations f_{coarse}, f_{fine} because the interpretation model needs to be simple enough to be interpretable. The model being explained is defined as $g : \mathbb{R}^d \rightarrow \mathbb{R}$. $h_{\mathbf{X}}(\mathbf{Z})$ is denoted as a proximity measure between an perturbed instance $\mathbf{Z}$ and the original input $\mathbf{X}$. Finally, we use $\mathcal{L}(f_{coarse}, g, h_{\mathbf{X}})$ and $\mathcal{L}(f_{fine}, g, h_{\mathbf{X}})$ as measures of f_{coarse} and f_{fine} in approximating g in locality. To ensure both interpretability and local fidelity, the MINT model is obtained by

$$\begin{aligned} \arg \min_{f_{coarse}, f_{fine}} & \mathcal{L}(f_{coarse}, g, h_{\mathbf{X}}) + \psi(f_{coarse}) + \psi(f_{fine}) \\ & + \mu\mathcal{L}(f_{fine}, g, h_{\mathbf{X}}) + \phi(f_{coarse}, f_{fine}) \end{aligned} \tag{7}$$

where μ is regularization parameter.

Given the multi-scale perturbed dataset $\mathcal{D}$, we optimize Eq. (7) to get the multi-scale interpretation for the network decision. Specifically, we minimize $\mathcal{L}(f_{coarse}, g, h_{\mathbf{X}})$ and $\mathcal{L}(f_{fine}, g, h_{\mathbf{X}})$ with the perturbation dataset of the multi-scale feature information. In order to learn the local behavior of g as the input varies, we approximate $\mathcal{L}(f_{coarse}, g, h_{\mathbf{X}})$ and $\mathcal{L}(f_{fine}, g, h_{\mathbf{X}})$ by drawing perturbation samples, weighted by $h_{\mathbf{X}}$. To measure the difference between a perturbed instance $\mathbf{Z}$ and the original instance $\mathbf{X}$, $h_{\mathbf{X}}(\mathbf{Z})$ is defined as

$$h_{\mathbf{X}}(\mathbf{Z}) = \exp(-\frac{1}{\sigma^2}\|\mathbf{X} - \mathbf{Z}\|_F^2) \tag{8}$$

In terms of coarse-scale explanation, we adopt linear model to interpret the importance of each large super-pixel cs to the prediction of the model, which is written as

$$f_{coarse}(\mathbf{z}^c; \mathbf{w}_c) = \mathbf{w}_c\mathbf{z}^c \tag{9}$$

where $\mathbf{w}_c \in \mathbb{R}^{1\times p}$ indicates the weights of all coarse-scale super-pixels in the instance.

For fine-scale explanation, we also use linear model to calculate the importance of each small super-pixel fs to the prediction. For a perturbed instance $\mathbf{Z}$, the fine-scale explanation is

$$f_{fine}(\mathbf{z}^f; \mathbf{w}_f) = \mathbf{w}_f\mathbf{z}^f \tag{10}$$

where $\mathbf{w}_f \in \mathbb{R}^{1\times q}$ indicates the weight of all fine-scale super-pixels in the instance.

Therefore, the loss functions of our multi-scale interpretation model are denoted as

$$\mathcal{L}(f_{coarse}, g, h_{\mathbf{X}}) = \frac{1}{2}\sum_{\mathbf{Z}\in\mathcal{D}} h_{\mathbf{X}}(\mathbf{Z})[f(\mathbf{z}^c; \mathbf{w}_c) - g(\mathbf{Z})]^2 \tag{11}$$

$$\mathcal{L}(f_{fine}, g, h_{\mathbf{X}}) = \frac{1}{2}\sum_{\mathbf{Z}\in\mathcal{D}} h_{\mathbf{X}}(\mathbf{Z})[f((\mathbf{z}^f; \mathbf{w}_f) - g(\mathbf{Z})]^2 \tag{12}$$

Here, we use a linear constraint for coarse-scale and fine-scale explanations. It means the weight of each coarse-scale super pixel can be written as the linear combination of the weights of its corresponding fine-scale super pixels.

$$\mathbf{w}_f\mathbf{T} = \mathbf{w}_c \tag{13}$$

where $\mathbf{T} \in \mathbb{R}^{m\times p}$ is the multi-scale relationship matrix. The (i, j)-th element $t_{(i,j)}$ of $\mathbf{T}$ is defined

$$t_{(i,j)} = \left\{ \begin{array}{ll} 1 & i\text{-th small super-pixel} \in j\text{-th large super-pixel} \\ 0 & \text{Otherwise} \end{array} \right. \tag{14}$$

In our model, we use sparsity to measure the complexity of the interpretation model. Therefore, $\psi(f_{coarse})$ and $\psi(f_{fine})$ are denoted as

$$\psi(f_{coarse}) = \nu_1 \|\mathbf{w}_c\|_0 \tag{15}$$

$$\psi(f_{fine}) = \nu_2 \|\mathbf{w}_f\|_0 \tag{16}$$

which measure the number of non-zero entries in the multi-scale weights. ν_1 and ν_2 are regularization parameters.

The optimization problem of the MINT model turns into

$$\begin{aligned} \arg\min_{\mathbf{w}_f,\mathbf{w}_c} \quad & \frac{1}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_c\mathbf{z}_k^c - y_k)^2 + \lambda_c\|\mathbf{w}_c\|_0 \\ & + \frac{\mu}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_f\mathbf{z}_k^f - y_k)^2 + \lambda_f\|\mathbf{w}_f\|_0 \\ \text{subject to} \quad & \mathbf{w}_f\mathbf{T} = \mathbf{w}_c \end{aligned} \tag{17}$$

The detailed process of the MINT model is shown in Figure 1.

C. MINT Algorithm

The proposed interpretation model is used to interpret the prediction of a black-box model being explained. It incorporates both coarse-scale and fine-scale explanations. Here, we propose a MINT algorithm to optimize our interpretation model.

The augmented Lagrangian for the optimization problem of MINT model in (17) is

$$\begin{aligned} \arg\min_{\mathbf{w}_f,\mathbf{w}_c,\mathbf{p}} \quad & \frac{1}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_c\mathbf{z}_k^c - y_k)^2 + \lambda_c\|\mathbf{w}_c\|_0 \\ & + \frac{\mu}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_f\mathbf{z}_k^f - y_k)^2 + \lambda_f\|\mathbf{w}_f\|_0 \\ & + \mathbf{p}^T(\mathbf{w}_f\mathbf{T} - \mathbf{w}_c) + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c\|_2^2 \end{aligned} \tag{18}$$

The equation can be rewritten as

$$\begin{aligned} \arg\min_{\mathbf{w}_f,\mathbf{w}_c,\mathbf{g}} \quad & \frac{1}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_c\mathbf{z}_k^c - y_k)^2 + \lambda_c\|\mathbf{w}_c\|_0 \\ & + \frac{\mu}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_f\mathbf{z}_k^f - y_k)^2 \\ & + \lambda_f\|\mathbf{w}_f\|_0 + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2 \end{aligned} \tag{19}$$

where

$$\mathbf{g} \equiv -\frac{1}{\lambda}\mathbf{p} \tag{20}$$

Due to the discrete and nonconvex nature of L_0 norm, the problem is NP hard which means that it is combinatorial and too complex to solve. To get an approximate solution to the problem, several sparse algorithms have been proposed. One popular strategy is to replace L_0 norm with L_1 norm, since L_1 norm is naturally the best convex approximation of L_0 norm. Then, the augmented Lagrangian for the optimization problem of MINT model in Eq. (19) can be converted to

$$\begin{aligned} \arg\min_{\mathbf{w}_f,\mathbf{w}_c,\mathbf{g}} \quad & \frac{1}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_c\mathbf{z}_k^c - y_k)^2 + \lambda_c\|\mathbf{w}_c\|_1 \\ & + \frac{\mu}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_f\mathbf{z}_k^f - y_k)^2 + \lambda_f\|\mathbf{w}_f\|_1 \\ & + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2 \end{aligned} \tag{21}$$

The optimization problem in Eq. (21) is convex and can be solved. Here, we design MINT algorithm by adopting the alternating iteration rule to learn $\mathbf{w}_f$ and $\mathbf{w}_c$. When learning $\mathbf{w}_c$, $\mathbf{w}_f$ is regarded as fixed value, and vice versa.

Given that the optimization is considered over the variable $\mathbf{w}_c$, the optimization function can be reduced to

$$\begin{aligned} \arg\min_{\mathbf{w}_c} \quad & \frac{1}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_c\mathbf{z}_k^c - y_k)^2 + \lambda_c\|\mathbf{w}_c\|_1 \\ & + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2 \end{aligned} \tag{22}$$

According to the objective function in Eq. (22), it can be converted into the equivalent formulation

$$\begin{aligned} \arg\min_{\mathbf{w}_c,\mathbf{m}_c} \quad & \frac{1}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_c\mathbf{z}_k^c - y_k)^2 + \lambda_c\|\mathbf{m}_c\|_1 \\ & + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2 \\ \text{subject to} \quad & \mathbf{w}_c = \mathbf{m}_c \end{aligned} \tag{23}$$

The augmented Lagrangian for the above problem is

$$\begin{aligned} \arg\min_{\mathbf{w}_c,\mathbf{m}_c,\mathbf{p}_c} \quad & \frac{1}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_c\mathbf{z}_k^c - y_k)^2 + \lambda_c\|\mathbf{m}_c\|_1 \\ & + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2 + \mathbf{p}_c^T(\mathbf{w}_c - \mathbf{m}_c) \\ & + \frac{\rho_c}{2}\|\mathbf{w}_c - \mathbf{m}_c\|_2^2 \end{aligned} \tag{24}$$

The equation can be rewritten as

$$\begin{aligned} \arg\min_{\mathbf{w}_c,\mathbf{m}_c,\mathbf{g}_c} \quad & \frac{1}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_c\mathbf{z}_k^c - y_k)^2 + \lambda_c\|\mathbf{m}_c\|_1 \\ & + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2 + \frac{\rho_c}{2}\|\mathbf{w}_c - \mathbf{m}_c - \mathbf{g}_c\|_2^2 \end{aligned} \tag{25}$$

where

$$\mathbf{g}_c \equiv -\frac{1}{\rho_c}\mathbf{p}_c \tag{26}$$

Through a careful choice of the new variable, the initial problem is converted into a much simple problem. Given that the optimization is considered over the variable $\mathbf{w}_c$, the optimization function can be reduced to

$$\begin{aligned}\mathbf{w}_c \leftarrow \arg\min_{\mathbf{w}_c} & \frac{1}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_c\mathbf{z}_k^c - y_k)^2 \\ & + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2 \\ & + \frac{\rho_c}{2}\|\mathbf{w}_c - \mathbf{m}_c - \mathbf{g}_c\|_2^2\end{aligned} \tag{27}$$

The solution is

$$\begin{aligned}\mathbf{w}_c^{s+1} \leftarrow & (\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k) y_k (\mathbf{z}_k^c)^T + \lambda\mathbf{w}_f\mathbf{T} - \lambda\mathbf{g} + \rho_c\mathbf{m}_c^s \\ & + \rho_c\mathbf{g}_c^s)(\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)\mathbf{z}_k^c(\mathbf{z}_k^c)^T + \lambda\mathbf{I} + \rho_c\mathbf{I})^{-1}\end{aligned} \tag{28}$$

In order to calculate $\mathbf{m}_c$, the optimization problem to be solved is

$$\mathbf{m}_c \leftarrow \arg\min_{\mathbf{m}_c} \lambda_c\|\mathbf{m}_c\|_1 + \frac{\rho_c}{2}\|\mathbf{w}_c - \mathbf{m}_c - \mathbf{g}_c\|_2^2 \tag{29}$$

The solution is

$$\mathbf{m}_c^{s+1} \leftarrow soft(\mathbf{w}_c^{s+1} - \mathbf{g}_c^s, \frac{\lambda_c}{\rho_c}) \tag{30}$$

Lagrange multipliers of $\mathbf{g}_c$ update to

$$\mathbf{g}_c^{s+1} \leftarrow \mathbf{g}_c^s - (\mathbf{w}_c^{s+1} - \mathbf{m}_c^{s+1}) \tag{31}$$

Similarly, given that the optimization is considered over the variable $\mathbf{w}_f$, the optimization function can be reduced to

$$\begin{aligned}\arg\min_{\mathbf{w}_f} & \frac{\mu}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_f\mathbf{z}_k^f - y_k)^2 + \lambda_f\|\mathbf{w}_f\|_1 \\ & + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2\end{aligned} \tag{32}$$

The objective function in Eq. (32) is rewritten as

$$\begin{aligned}\arg\min_{\mathbf{w}_f,\mathbf{m}_f} \quad & \frac{\mu}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_f\mathbf{z}_k^f - y_k)^2 + \lambda_f\|\mathbf{m}_f\|_1 \\ & + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2\end{aligned} \tag{33}$$

subject to $\quad \mathbf{w}_f = \mathbf{m}_f$

The augmented Lagrangian for the above problem becomes

$$\begin{aligned}\arg\min_{\mathbf{w}_f,\mathbf{m}_f,\mathbf{p}_f} & \frac{\mu}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_f\mathbf{z}_k^f - y_k)^2 + \lambda_f\|\mathbf{m}_f\|_1 \\ & + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2 + \mathbf{p}_f^T(\mathbf{w}_f - \mathbf{m}_f) \\ & + \frac{\rho_f}{2}\|\mathbf{w}_f - \mathbf{m}_f\|_2^2\end{aligned} \tag{34}$$

The equation is reformulated as

$$\begin{aligned}\arg\min_{\mathbf{w}_f,\mathbf{m}_f,\mathbf{g}_f} & \frac{\mu}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_f\mathbf{z}_k^f - y_k)^2 + \lambda_f\|\mathbf{m}_f\|_1 \\ & + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2 \\ & + \frac{\rho_f}{2}\|\mathbf{w}_f - \mathbf{m}_f - \mathbf{g}_f\|_2^2\end{aligned} \tag{35}$$

where

$$\mathbf{g}_f \equiv -\frac{1}{\rho_f}\mathbf{p}_f \tag{36}$$

The above problem is converted to a simple problem by a prudent choice of the new variable. Given that the optimization is considered over the variable $\mathbf{w}_f$, the optimization function becomes

$$\begin{aligned}\mathbf{w}_f \leftarrow \arg\min_{\mathbf{w}_f} & \frac{\mu}{2}\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)(\mathbf{w}_f\mathbf{z}_k^f - y_k)^2 \\ & + \frac{\lambda}{2}\|\mathbf{w}_f\mathbf{T} - \mathbf{w}_c - \mathbf{g}\|_2^2 \\ & + \frac{\rho_f}{2}\|\mathbf{w}_f - \mathbf{m}_f - \mathbf{g}_f\|_2^2\end{aligned} \tag{37}$$

The solution is

$$\begin{aligned}\mathbf{w}_f^{r+1} \leftarrow & (\mu\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k) y_k (\mathbf{z}_k^f)^T + \lambda\mathbf{w}_c\mathbf{T}^T + \lambda\mathbf{g}\mathbf{T}^T + \rho_f\mathbf{m}_f^r \\ & + \rho_f\mathbf{g}_f^r)(\mu\sum_{k=1}^{K} h_{\mathbf{X}}(\mathbf{Z}_k)\mathbf{z}_k^f(\mathbf{z}_k^f)^T + \lambda\mathbf{T}\mathbf{T}^T + \rho_f\mathbf{I})^{-1}\end{aligned} \tag{38}$$

The optimization problem of $\mathbf{m}_f$ is

$$\mathbf{m}_f \leftarrow \arg\min_{\mathbf{m}_f} \lambda_f\|\mathbf{m}_f\|_1 + \frac{\rho_f}{2}\|\mathbf{w}_f - \mathbf{m}_f - \mathbf{g}_f\|_2^2 \tag{39}$$

The solution is

$$\mathbf{m}_f^{r+1} \leftarrow soft(\mathbf{w}_f^{r+1} - \mathbf{g}_f^r, \frac{\lambda_f}{\rho_f}) \tag{40}$$

Lagrange multipliers of $\mathbf{g}_f$ update to

$$\mathbf{g}_f^{r+1} \leftarrow \mathbf{g}_f^r - (\mathbf{w}_f^{r+1} - \mathbf{m}_f^{r+1}) \tag{41}$$

Finally, lagrange multipliers of $\mathbf{g}$ in the optimization problem in Eq. (21) update to

$$\mathbf{g}^{t+1} \leftarrow \mathbf{g}^t - (\mathbf{w}_f^{t+1}\mathbf{T} - \mathbf{w}_c^{t+1}) \tag{42}$$

Through several rounds of alternating optimization of $\mathbf{w}_c$ and $\mathbf{w}_f$, we finally get the optimal results $\mathbf{w}_c^*$ and $\mathbf{w}_f^*$. The pseudocode of MINT algorithm is shown in Algorithm 1.

Given the optimized MINT model, we can present both coarse-scale and fine-scale explanations to the prediction of the model being explained. As for the coarse-scale visual interpretation, we generate corresponding visual explanation by choosing the most critical coarse-scale super-pixels for the deep layer. Specifically, for a particular coarse-scale super-pixel, its class-discriminative importance for the network prediction is calculated by combining its optimal coarse-scale weight with corresponding average pooling feature value in the deep layer. Accordingly, the class-discriminative important vector of coarse-scale super-pixels is denoted as

$$\hat{\mathbf{w}}_c^* = \mathbf{w}_c^* \odot \left[a_1^{cs} \; a_2^{cs} \cdots a_i^{cs} \cdots a_p^{cs}\right] \tag{43}$$

where the a_i^{cs} is the average pooling feature of cs_i, and the operator $\odot$ denotes Hadamard product. Therefore, the coarse-scale visual interpretation in the deep layer is the selected coarse-scale super-pixel whose class-discriminative importance value ranks the first.

Algorithm 1: Pseudocode of the MINT Algorithm

```
Input: the perturbed dataset $\mathcal{D}\{\mathbf{Z}_k, y_k, \mathbf{z}_k^c, \mathbf{s}_k^c, \mathbf{z}_k^f, \mathbf{s}_k^f\}$, the optimization objective of MINT model
Output: optimal $\mathbf{w}_c^*$, $\mathbf{w}_f^*$
Initialization: set $t = 0$, $s = 0$, $r = 0$, $\rho_c \geq 0$, $\rho_f \geq 0$, $\mathbf{m}_c^0$, $\mathbf{g}_c^0$, $\mathbf{m}_f^0$, $\mathbf{g}_f^0$, $\mathbf{g}^0$, $\mathbf{w}_c^{parm}$, $\mathbf{w}_f^{parm}$
repeat
    The optimization problem for $\mathbf{w}_c$
    Input initialization: $\mathbf{m}_c^0$, $\mathbf{g}_c^0$, $\mathbf{w}_f^{parm}$,$\mathbf{g}^0$
    repeat
        $\mathbf{w}_c^{s+1} \leftarrow (\sum_{k=1}^K h_{\mathbf{X}}(\mathbf{Z}_k) y_k (\mathbf{z}_k^c)^T + \lambda(\mathbf{w}_f \mathbf{T} - \mathbf{g}) + \rho_c \mathbf{m}_c^s + \rho_c \mathbf{g}_c^s)(\sum_{k=1}^K h_{\mathbf{X}}(\mathbf{Z}_k) \mathbf{z}_k^c (\mathbf{z}_k^c)^T + \lambda \mathbf{I} + \rho_c \mathbf{I})^{-1}$
        $\mathbf{m}_c^{s+1} \leftarrow soft(\mathbf{w}_c^{s+1} - \mathbf{g}_c^s, \frac{\lambda_c}{\rho_c})$
        Update Lagrange multipliers:
              $\mathbf{g}_c^{s+1} \leftarrow \mathbf{g}_c^s - (\mathbf{w}_c^{s+1} - \mathbf{m}_c^{s+1})$
        Update iteration: $s \leftarrow s + 1$
    until some stopping criterion is satisfied;
    Update initialization:
        $\mathbf{m}_c^0 \leftarrow \mathbf{m}_c^*$
        $\mathbf{g}_c^0 \leftarrow \mathbf{g}_c^*$
        $\mathbf{w}_c^{parm} \leftarrow \mathbf{w}_c^*$
    The optimization problem for $\mathbf{w}_f$
    Input initialization: $\mathbf{m}_f^0$, $\mathbf{g}_f^0$, $\mathbf{w}_c^{parm}$,$\mathbf{g}^0$
    repeat
        $\mathbf{w}_f^{r+1} \leftarrow (\mu \sum_{k=1}^K h_{\mathbf{X}}(\mathbf{Z}_k) y_k (\mathbf{z}_k^f)^T + \lambda(\mathbf{w}_c + \mathbf{g})\mathbf{T}^T + \rho_f \mathbf{m}_f^r + \rho_f \mathbf{g}_f^r)(\mu \sum_{k=1}^K h_{\mathbf{X}}(\mathbf{Z}_k) \mathbf{z}_k^f (\mathbf{z}_k^f)^T + \lambda \mathbf{T}\mathbf{T}^T + \rho_f \mathbf{I})^{-1}$
        $\mathbf{m}_f^{r+1} \leftarrow soft(\mathbf{w}_f^{r+1} - \mathbf{g}_f^r, \frac{\lambda_f}{\rho_f})$
        Update Lagrange multipliers:
            $\mathbf{g}_f^{r+1} \leftarrow \mathbf{g}_f^r - (\mathbf{w}_f^{r+1} - \mathbf{m}_f^{r+1})$
        Update iteration: $r \leftarrow r + 1$
    until some stopping criterion is satisfied;
    Update initialization:
        $\mathbf{m}_f^0 \leftarrow \mathbf{m}_f^*$
        $\mathbf{g}_f^0 \leftarrow \mathbf{g}_f^*$
        $\mathbf{w}_f^{parm} \leftarrow \mathbf{w}_f^*$
    Update Lagrange multipliers:
        $\mathbf{g}^{t+1} \leftarrow \mathbf{g}^t - (\mathbf{w}_f^{parm} \mathbf{T} - \mathbf{w}_c^{parm})$
    Update initialization:
        $\mathbf{g}^0 \leftarrow \mathbf{g}^{t+1}$
    Update iteration: $t \leftarrow t + 1$
until some stopping criterion is satisfied;
```

In comparison, for the fine-scale visual interpretation, the corresponding visual explanation is provided by presenting the specific feature in the most critical fine-scale super-pixels for the shallow layer. The class-discriminative importance of a particular fine-scale super-pixel for the network prediction is computed by combining its optimal fine-scale weight with corresponding average pooling feature value in the shallow layer. Consequently, the class-discriminative important vector of fine-scale super-pixels is denoted as

$$\hat{\mathbf{w}}_f^* = \mathbf{w}_f^* \odot \left[a_1^{fs} \ a_2^{fs} \cdots a_j^{fs} \cdots a_q^{fs} \right] \tag{44}$$

where the a_j^{fs} is the average pooling feature of fs_j. Because of the sparsity of the fine-scale weight, the class-discriminative importance vector of fine-scale super-pixels is also sparse. Given the optimal class-discriminative importance of each fine-scale super-pixel, a threshold for importance is selected empirically. If the importance of a fine-scale super-pixel is higher than it, the feature of this fine-scale super-pixel will be chosen for the visual interpretation of the shallow layer. Therefore, for the certain prediction of the model being explained, coarse-scale and fine-scale interpretation are presented for the deep and shallow layer, respectively.

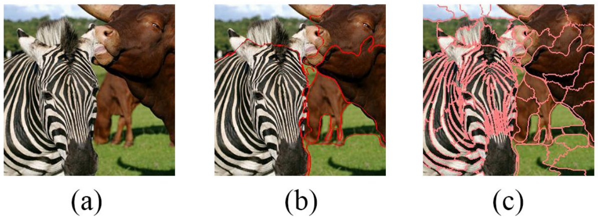

Fig. 2. An example of the original image (a), its coarse-scale segmentation (b), and fine-scale segmentation (c).

III. Experiments

In this section, we conduct experiments to evaluate the performance of our MINT model. For image classifiers, our interpretation model provides multi-scale interpretation of a specific class prediction. In the experiments, we explain top-class predictions of certain pre-trained neural networks including VGG [18] and Inception-V3 [19]. All these networks achieve high image classification accuracy in ImageNet dataset. These networks have strong learning ability to obtain a complex mapping from the input to the output. Interpretation models are expected to present what these networks have learned from different layers. As representative methods in the model interpretability field, Grad-CAM [13] and LIME [16] are compared with our interpretation model. Grad-CAM is a gradient-based interpretation model which is designed to explain deep networks, and LIME is a perturbed-based interpretation model.

A. Experiment Settings

1) Multi-scale Super-pixels Generation: In the experiment, we need to generate perturbed image samples for a given input image. In order to get the perturbed samples, the MINT model needs to generate multi-scale super-pixels. For the coarse-scale super-pixel, since features in the deep layer have high-level semantic meanings in the networks, the class-discriminative saliency map of the target class can be obtained by combining the features and their gradients in the deep layer. GrabCut [17] utilizes the class-discriminative saliency map as prior knowledge of the target object. As a result, the region which is highlighted for the target class in the feature of the deep layer is divided as a coarse-scale super-pixel from the background. Subsequently, the background region is segmented into coarse super-pixels by Simple linear iterative clustering (SLIC) [20] which is an unsupervised segmentation algorithm. For the fine-scale explanation, Quickshift [21], an unsupervised segmentation algorithm, is applied to segment each coarse-scale super-pixel into small super-pixels. Figure 2 shows an example of the original image, its coarse-scale and fine-scale segmentation results.

2) Perturbed Image Dataset Generation: We sample 5000 perturbed image instances by selecting fine-scale super-pixels uniformly at random. For each fine-scale super-pixel in the perturbed image, its pixel values remain original when it exists in the perturbed image. Otherwise, its pixel values are grayed out, which means the values are set to zero. Each perturbed image instance is fed into networks to get its class prediction. Figure 3 shows examples of the perturbed images and their "zebra" class prediction values. The probability of "zebra" class for the original image is 0.7843. While for the perturbed images, their prediction values for "zebra" class vary due to different occlusion parts.

The proposed interpretation model is qualitatively compared with Grad-CAM and LIME. Quickshift is used as the super-pixel segmentation algorithm in LIME. The parameter settings in quickshift are the same as in LIME and the proposed interpretation method. It is worth mentioning that quantitative criterion for interpretation models remains an open problem, mainly because currently there lacks of error metrics for the explanation results.

B. The Interpretation for the Correct Prediction of Networks

1) The interpretation for the correct prediction of a single class: Figures 4 and 5 show visual explanations for predictions of a single class obtained by Inception-V3 and VGG, respectively. The first column lists the original images. From the second column to the last one are the results of Grad-CAM, LIME, coarse-scale explanation of MINT, and fine-scale explanation of MINT, respectively. The networks being explained correctly predict all images as their ground-truth classes. In this experiment, images from the validation dataset of ImageNet 2015 are chosen as the instances to be explained.

As shown in Figures 4 and 5, the proposed MINT model presents explanations at multiple scales for certain class prediction.

For the coarse-scale explanation, the MINT model extracts important regions for the network prediction. From the experiment results, in most cases, coarse-scale explanation learned by the features in the deep layer coincides with the region of the target class. Coarse-scale explanation of MINT shows that the network has the coarse-scale attention to focus on the target object.

For the fine-scale explanation, the MINT model provides the important fine-scale feature in shallow layers for the network decision. For example, the fine-scale explantation of "zebra" presents the texture of zebra-stripe. Likewise, the fine-scale explanation of "albatross" shows the beak. Fine MINT shows that the network learns the important characteristics of the target class from the shallow layer.

The explanation of MINT that combines fine scale with coarse scale can be more beneficial to human understanding. Meanwhile, the MINT model interprets the mechanism of the hierarchical structure in which shallow layers usually represent local, fine-scale features while deep layers show more holistic features. Therefore, MINT can not only provide the multi-scale interpretation to the given instance but also assist us to understand the significance of the internal features in the network.

For the gradient-based interpretation, Grad-CAM generates the class-discriminative heatmap by utilizing the gradient of output with respect to features in the last convolutional layer. However, its explanation is not high-resolution and it cannot provide sensible interpretation for the shallow layer. For the perturbation-based interpretation, LIME is used to identify critical input image regions that contribute to the target-class prediction. Nevertheless, it cannot interpret the internal features of networks. In contrast, the MINT model can be considered as the hybrid interpretation model exploiting both perturbation mechanism and the features with their gradients. Consequently, the MINT model generates the fine-scale and coarse-scale explanations for the shallow and deep layers, respectively. The coarse-scale explanations shown in Figures 4 and 5, present almost complete class object while Grad-CAM and LIME results only show partial object. Meanwhile, Grad-CAM and LIME cannot provide a fine-scale explanation of the network

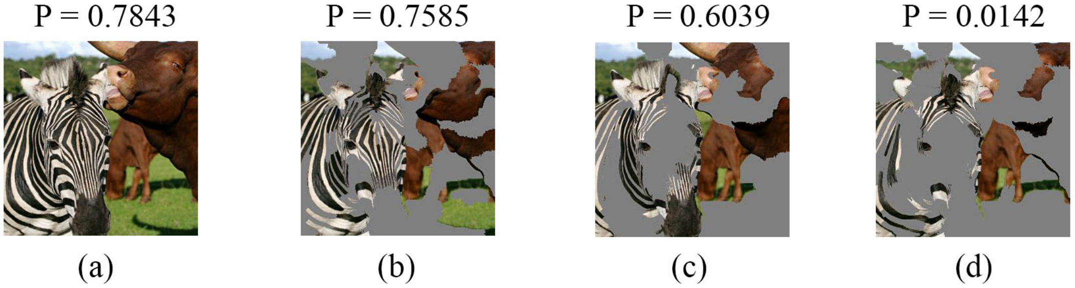

Fig. 3. An example of the original image (a) and the perturbed image samples (b-d) with corresponding "zebra" class prediction values.

decision for the shallow layer. Figures 4 and 5 show MINT model obtains more reasonable interpretation by cooperating the fine-scale and coarse-scale explanations.

2) The interpretation for correct predictions of multiple classes: Figure 6 shows the explanations for predictions of multiple classes obtained by Inception-V3. In this experiment, images in which multiple objects exist are chosen as the instances to test the performance of the compared interpretation models. For Figure 6(a), the network predicts "lion" and "Egyptian cat" as its top 2 classes accurately. For Figure 6(b), "ram" and "border collie" are correctly predicted as the top 2 classes.

As shown in the fourth and fifth columns of Figure 6, the coarse-scale and fine-scale explanations of MINT model present intuitively correct visual explanations for both class predictions. The coarse-scale explanation shows the important image region of particular class prediction, which interprets the class-discriminative attention captured by the deep layer. The fine-scale explanation further visualizes critical features learned from the shallow layer.

In comparison, the explanation of LIME (Figure 6(a-2)) is not reasonable when compared with the proposed model (Figure 6(a-3) and (a-4)). Meanwhile, LIME fails to provide an explanation for the network features. The explanation of Grad-CAM is not high-resolution and cannot interpret the features in shallow layers. To sum up, the results of our explanation model are more sensible in human understanding and the network structure than that of Grad-CAM and LIME.

C. The Interpretation for the Incorrect Prediction of Networks

In addition to the explanation for correct class prediction, the explanation for incorrect class prediction is also crucial for model interpretation. In this experiment, the compared interpretation models are applied to explain the incorrect decision of Inception-V3. Images which contain multiple objects are tested to analyze different interpretation models.

Figure 7(a) is the original confusing image which includes a guitar. Among the labels of ImageNet dataset, "acoustic guitar" and "electric guitar" are both included. From appearance, two kinds of guitars only have small difference. Therefore, it makes networks difficult to give the correct prediction. In the experiment, Inception-V3 predicts "brown bear", "acoustic guitar" as the top 2 classes. The prediction of "electric guitar" is also within the top 5 predictions. In this scenario, our interpretation model is expected to give explanations for these predictions.

"Brown bear" and "acoustic guitar" do exist in the image. When the network gives the right prediction for "brown bear" and "acoustic guitar", the proposed MINT model is expected to give explanations why we should trust the predictions. As shown in Figure 7, the explanations for correct predictions "brown bear" (as shown in the first row) and "acoustic guitar" (as shown in the second row) are all natural to human.

However, "electric guitar" does not exist in the image. When the network gives the wrong prediction for "electric guitar", the MINT model is expected to give explanation why the prediction fails. As shown in Figure 8(a) and (b), the similarity between "electric guitar" and "acoustic guitar" is that they both have fretboard; their difference is the body where sound hole only exists in "acoustic guitar". For the wrong prediction "electric guitar", the MINT model shows reasonable fine-scale interpretation for the shallow layer: the fretboard (as shown in Figure 8(e)). In MINT model, none of the coarse-scale image regions is chosen because their importance weights are all minimal. In contrast, some fine-scale features learned from the shallow layer are chosen because of their relatively high importance weights in fine MINT. The reason is that the network learns the feature of electric guitar (such as fretboard) from fine-scale image regions. But no coarse-scale image region matches the feature of electric guitar. This indicates that even for the wrong prediction, the deep network being explained does not act unreasonably.

In comparison, Grad-CAM results (Fig. 7(a-5) and Fig. 8c) and LIME results (Fig. 7(a-6) and Fig. 8d) for "acoustic guitar" and "electric guitar" are similar, which both contain fretboard and guitar sound hole region. Therefore Grad-CAM and LIME may fail to give reasonable explanations for the wrong class prediction of networks.

IV. Conclusion

In the paper, we propose the MINT model, to provide multi-scale visual explanations for the predictions of CNN deep networks without requiring the architectural modification or retraining of deep networks. The MINT model provides coarse-scale interpretation for the attention in the deep layer and fine-scale interpretation for specific features in the shallow layer. Consequently, the MINT model can provide class-discriminative interpretation of the network decision, and also explain the significance of the hierarchical structure of networks.

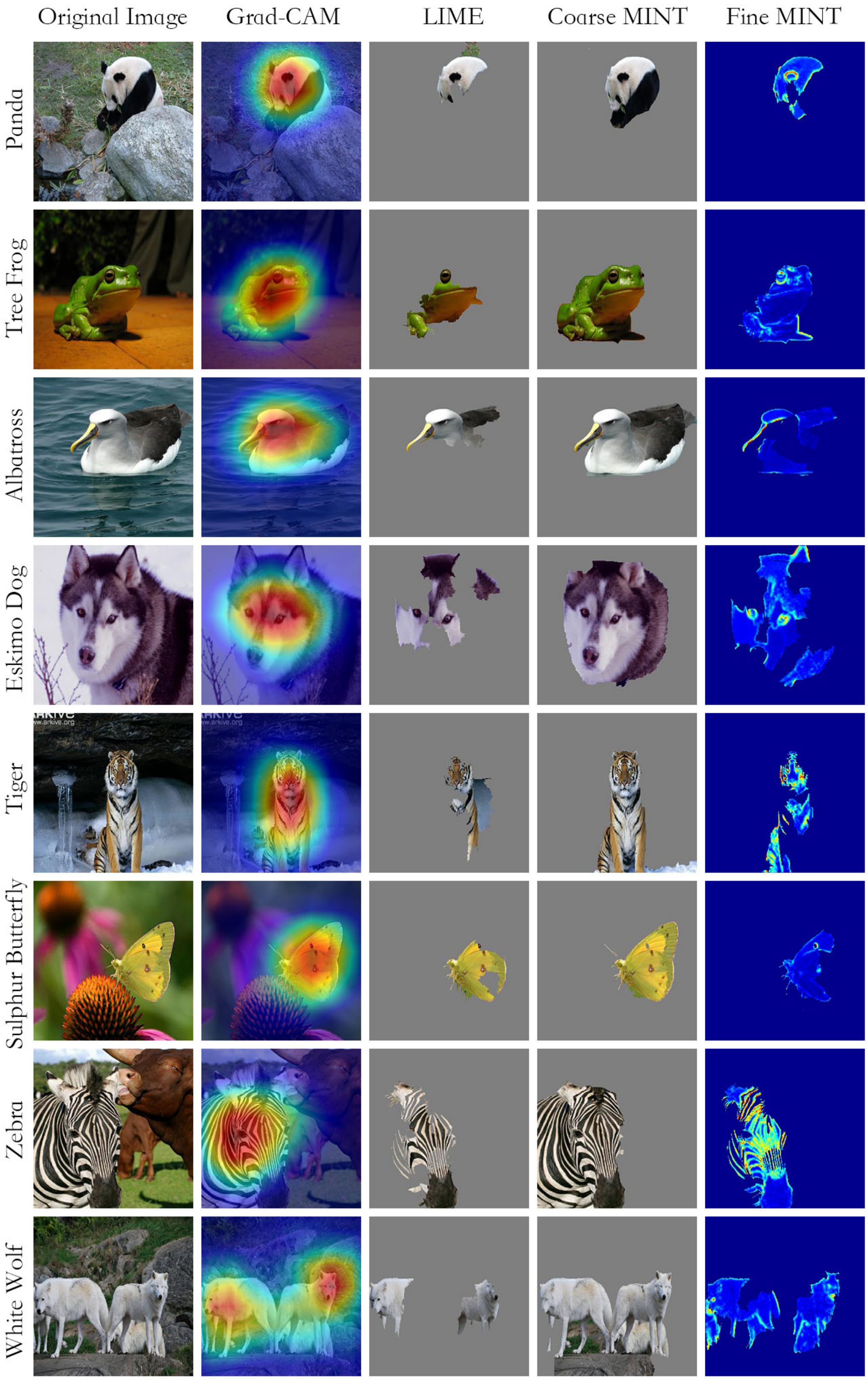

Fig. 4. Visual explanations for the top 1 class predictions of Inception-V3.

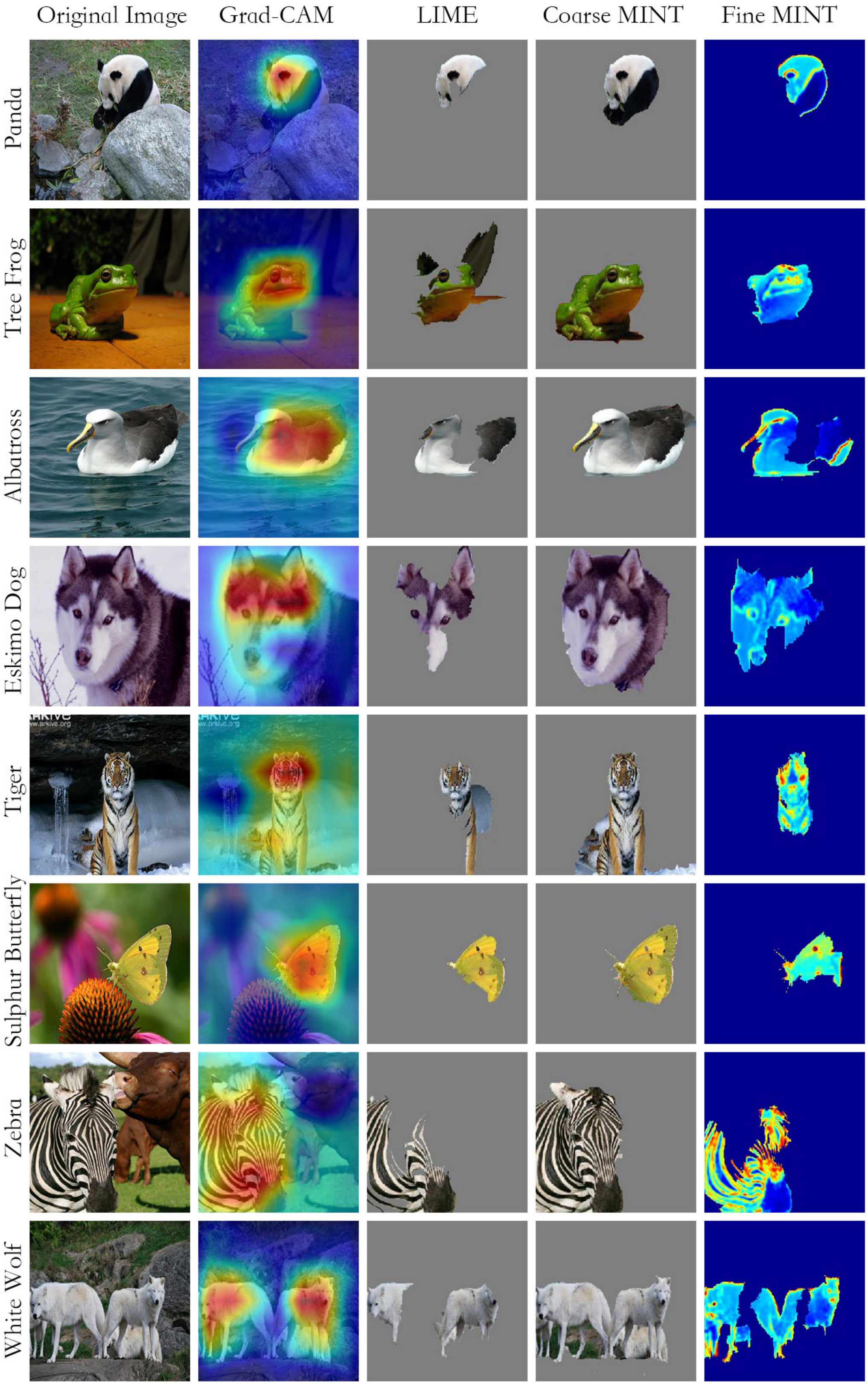

Fig. 5. Visual explanations for the top 1 class predictions of VGG.

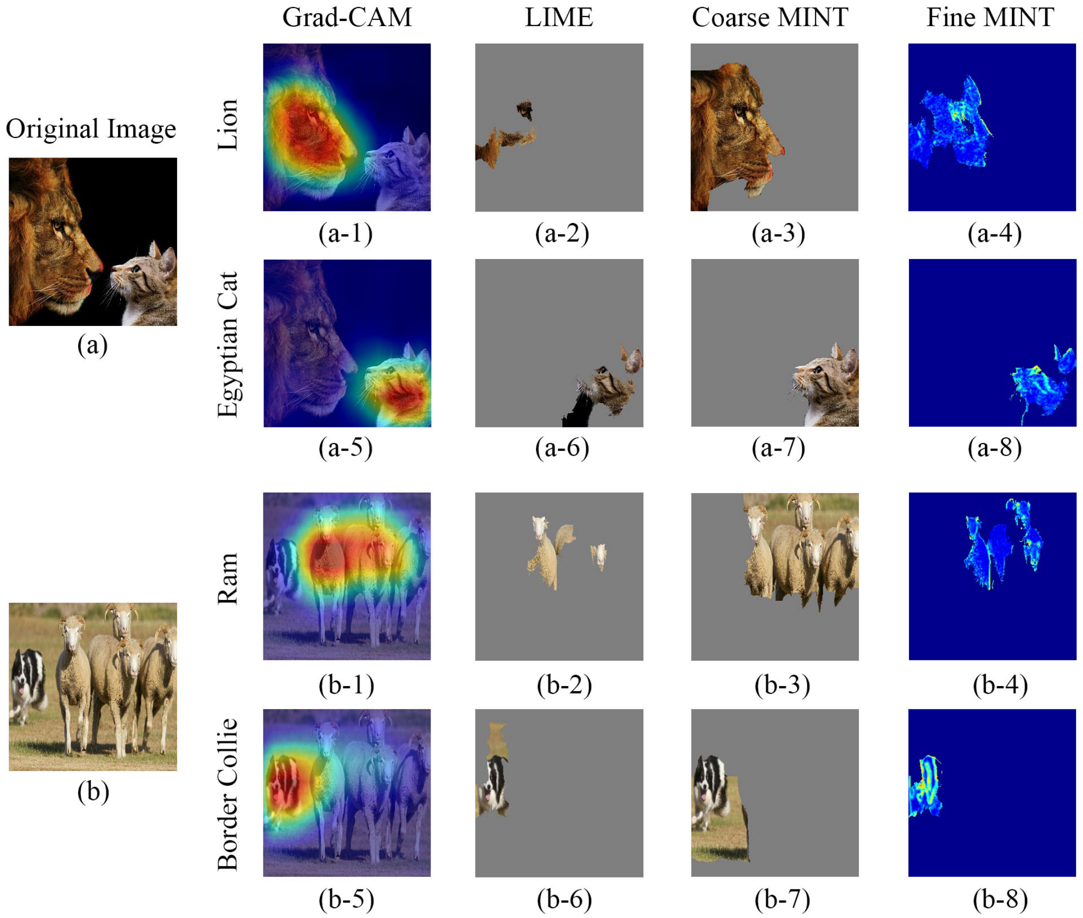

Fig. 6. Visual explanations for the top 2 class predictions of images (a, b). From the second column to the fifth one are results of Grad-CAM, LIME, coarse-scale explanation and fine-scale explanation of the MINT model, respectively.

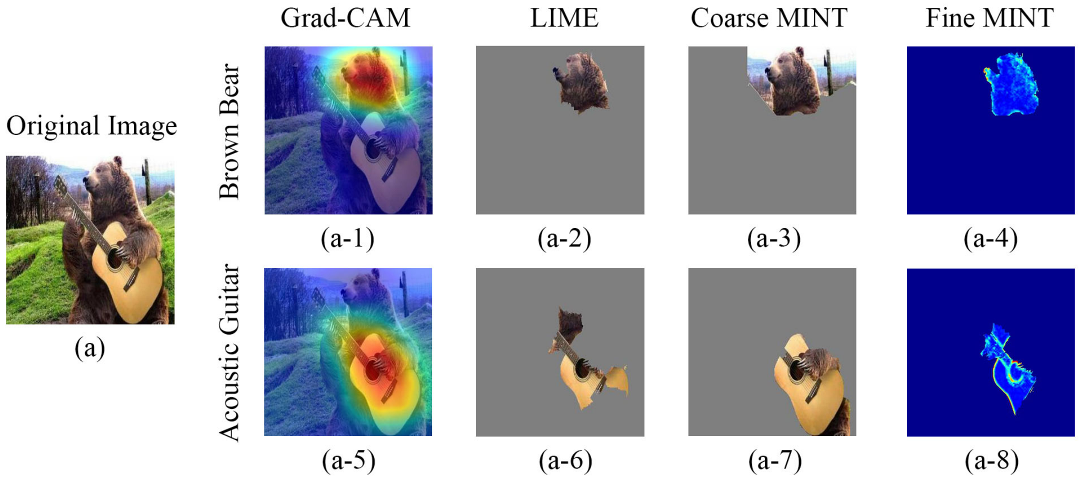

Fig. 7. Visual explanations for top class predictions of an image (a). The selected class predictions of (a) are "brown bear" (first row) and "acoustic guitar" (second row). From the second column to the fifth one are results of Grad-CAM, LIME, coarse-scale explanation and fine-scale explanation of the MINT model, respectively.

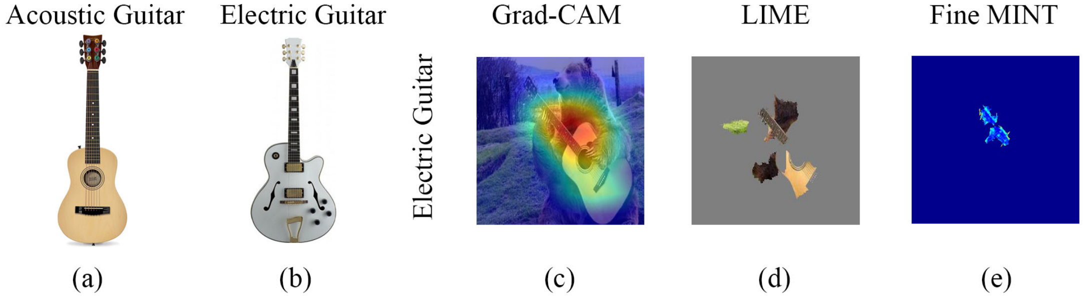

Fig. 8. Visual explanations for the "electric guitar" class prediction of the image in Fig. 7(a). (a) is acoustic guitar and (b) is electric guitar. (c) is the Grad-CAM result; (d) is the LIME result; (e) is the fine-scale explanation of the MINT model.

Experimental results demonstrate that the visual explanations of the proposed model are intuitively reasonable in image classification and help build trust of the model being explained.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[3] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4565–4574.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[5] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A deep learning approach to visual question answering," *International Journal of Computer Vision*, vol. 125, no. 1-3, pp. 110–135, 2017.

[6] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644–655, Sep. 1998.

[7] X. Cui, D. Wang, and Z. J. Wang, "Chip: Channel-wise disentangled interpretation of deep convolutional neural networks," *arXiv preprint arXiv:1902.02497*, 2019.

[8] N. Doulamis and A. Doulamis, "Evaluation of relevance feedback schemes in content-based in retrieval systems," *Signal Processing: Image Communication*, vol. 21, no. 4, pp. 334 – 357, 2006.

[9] P. Jackson, "Introduction to expert systems," 1986.

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

[11] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 2921–2929.

[13] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[14] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[15] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," *arXiv preprint arXiv:1704.03296*, 2017.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[17] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut": Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

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

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[20] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[21] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *European Conference on Computer Vision*. Springer, 2008, pp. 705–718.