CHIP: Channel-wise Disentangled Interpretation of Deep Convolutional Neural Networks

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

***Abstract*—With the widespread applications of deep convolutional neural networks (DCNNs), it becomes increasingly important for DCNNs not only to make accurate predictions but also to explain how they make their decisions. In this work, we propose a CHannel-wise disentangled InterPretation (CHIP) model to give the visual interpretation to the predictions of DCNNs. The proposed model distills the class-discriminative importance of channels in networks by utilizing the sparse regularization. Here, we first introduce the network perturbation technique to learn the model. The proposed model is capable to not only distill the global perspective knowledge from networks but also present the class-discriminative visual interpretation for specific predictions of networks. It is noteworthy that the proposed model is able to interpret different layers of networks without re-training. By combining the distilled interpretation knowledge in different layers, we further propose the Refined CHIP visual interpretation that is both high-resolution and class-discriminative. Experimental results on the standard dataset demonstrate that the proposed model provides promising visual interpretation for the predictions of networks in image classification task compared with existing visual interpretation methods. Besides, the proposed method outperforms related approaches in the application of ILSVRC 2015 weakly-supervised localization task.**

***Index Terms*—Model interpretability, class-discriminative visual interpretation, channel-wise disentanglement, sparsity.**

I. Introduction

In recent years, deep convolutional neural networks (DCNNs) have achieved impressive results in a number of computer vision tasks such as image classification [1], [2], object detection [3], image captioning [4], [5], and visual question answering [6]. Despite their success in these tasks, the complex internal working mechanism of DCNNs remains elusive and makes them highly non-transparent to human [7]. The lack of capability to make themselves understandable may result in a general mistrust of their results. The mistrust may hinder the application of DCNNs, especially for critical applications such as medical diagnosis and criminal justice where model reliability is critical [8]. The requirement in applications underlines the need for the interpretation of DCNNs.

Interpretability study has drawn increasing attention because of its potential advantages such as building trust for real-world users and offering insight into black-box models for practitioners [9], [10]. In this research field, visual interpretation is one of the prevalent interpretation methods. A number of visual interpretation studies have been proposed to interpret the decisions of DCNNs [11], [12], [13], [14]. A good visual interpretation should be a faithful representation of the network as well as be interpretable enough for users to understand. Specifically, a desired visual interpretation model should possess the following properties:

- The model should give class-discriminative visual interpretation, because the DCNN being explained is designed to distinguish different classes. The class-discriminative visual interpretation should highlight the important region of the input for a certain class prediction.
- The model should be able to present the fine-grained information that is important to the class prediction.
- The model is supposed to show interpretable knowledge of different layers and lend insight into the roles of different layers of DCNNs.
- Given an image, the model should give a specific explanation for the network decision. People can decide whether to trust that decision based on the instance interpretation.
- The model should give global perspective interpretation of networks. It implies that the general interpretation knowledge should be distilled from a large image dataset but rather a single image.

It should also be noted that there is a trade-off between the performance and the interpretability of DCNNs. Pursuing the interpretability of DCNNs may sacrifice their accuracies, which is undesirable [13]. In light of that, we focus on the post-hoc interpretability [14] which refers to the extraction and analysis of information from a trained network. In this situation, the network usually does not have to sacrifice its performance in order to be interpretable.

For visual interpretation, several approaches have been proposed to highlight a small portion of internal features that are critical to the decision by ascribing saliency [15], [16], [13], [14]. However, there are a few limitations to these methods. For example, Guided Backpropagation [15] cannot give class-discriminative visualization. Class Activation Mapping (CAM) [13] and Grad-CAM [14] are only constrained in the last convolutional layer to present the visual interpretation. Table I shows the properties of different visual interpretation methods.

In order to alleviate the above issues, we propose a CHannel-wise disentangled InterPretation (CHIP) model which can build trust in the network without sacrificing its accuracy. The main contributions of this work are summarized as follows:

- CHIP model can disentangle channels of different layers by learning the class-discriminative importance of features. Based on that, CHIP can give visual interpretation to network decisions without re-training.
- CHIP can distill the class-discriminative knowledge of the network from a large dataset and utilize the distilled

X. Cui, D. Wang, and Z. J. 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.)

TABLE I
PROPERTIES OF DIFFERENT VISUAL INTERPRETATION METHODS

Method \ Property	Instance interpretation		General interpretation	Class-discriminative interpretation	High-resolution interpretation	Multi-layer interpretation	Post-hoc interpretation
	Perturbation-based	Gradient-based					
Deconvolution[16]		✓			✓	✓	✓
Guided Backpropagation[15]		✓			✓		✓
CAM[13]		✓		✓			
Grad-CAM[14]		✓		✓			✓
Guided Grad-CAM[14]		✓		✓	✓		✓
LIME[17]	✓			✓			✓
DeepDraw[18]			✓	✓	✓		✓
Guided CHIP	✓	✓	✓	✓	✓	✓	✓
CHIP & Refined CHIP	✓		✓	✓	✓	✓	✓

knowledge to interpret the decision of the network for an instance image. It means our model can not only explore the general class-discriminative information without data bias but also give an interpretation for the prediction of a particular input image.

- In order to learn the CHIP model, we first introduce the network perturbation method by perturbing internal features. The underlying principle is that the class prediction would drop dramatically if the forward propagation of class-specific important channels are blocked.
- Our model utilizes the inherent sparse property of features in DCNNs as a regularization to distill class-discriminative knowledge. We also propose an algorithm to optimize the CHIP model.
- Through combining the class-discriminative visual interpretation in the shallow and the deep layers, we extend CHIP to Refined CHIP which can give a high-resolution class-discriminative explanation.
- The proposed interpretation model can be applied to tackle the weakly-supervised localization problem by utilizing the visual interpretation. Compared with previous related work, the proposed model achieves better performance in ILSVRC 2015 dataset.

The rest of the paper is structured as follows. Section II reviews the related work. The proposed CHIP model is presented in Section III. In Section IV, experiments are implemented to evaluate the efficacy of the proposed CHIP and Refined CHIP. The conclusion is drawn in Section V.

II. RELATED WORK

The most related studies to our work are visual interpretation of DCNNs and its application in the weakly-supervised localization. In this section, we describe the related work in comparison with our work.

Visual interpretation of DCNNs. Many studies have been conducted to understand DCNNs by visual interpretation. The visual interpretation of DCNNs is generally classified into two categories: one is instance interpretation which means the interpretation for a particular input; and the other is general interpretation which denotes the interpretation of the overall network. Both of them focus on providing visual interpretation to explore the internal working mechanism of DCNNs.

Methods in the first category can be further divided into gradient-based methods and perturbation-based ones. Gradient-based approaches give interpretation by using the gradient (or gradient variant) of the output or the internal unit. For instance, Guided Backpropagation [15] and Deconvolution [16] interpret the network by visualizing internal units. They design different backward pass ways to map neuron activation down to the input space, visualizing the input image pattern that is most discriminative to the neuron activation. Although they can give high-resolution visualization for a specific input image, the visualization is not class-discriminative. In contrast, CAM [13] and Grad-CAM [14] are able to give the class-discriminative interpretation. Specifically, they visualize the linear combination of activations and class-specific weights in the last convolutional layer. Although the obtained visualization is class-discriminative, they are not high-resolution which means they do not show the fine-grained details. In this situation, the visualization cannot clearly show what feature makes the network give its prediction despite the highlight of the related image region. To tackle the problem, Guided Grad-CAM is designed by combining Guided Backpropagation visualization. Another common drawback is that their visual interpretation is only effective for the last convolutional layer. In addition, CAM can only provide interpretation for a specific kind of DCNN architectures.

Perturbation-based methods mainly focus on the input image perturbation [16], [17], [11]. To be specific, they involve perturbing the input images and observing the change of predictions. The reason behind this is that the prediction would drop by the maximum amount when pixels contribute maximally to the prediction are modified. Although these methods can obtain the class-discriminative visualization, they do not explore the internal mechanism of networks and interpret the internal features. Meanwhile, their performance is largely limited by the size and shape of the occluded pieces in the perturbed images. For instance, image patches in regular grids are used for occlusion in [16] while super-pixels obtained by unsupervised segmentation are used in [17].

Methods in the other category focus on visualizing DCNNs in a global perspective by synthesizing the optimal image that can get the maximum activation of unit [18], [19], [20]. Although these methods are able to give class-discriminative interpretation, they are not designed for interpreting specific

input images. Therefore, from some point of view, they only provide high-level abstract interpretation and are not helpful to understand specific decisions of DCNNs.

There is a compromise between instance interpretation and general interpretation. Instance interpretation can explain network predictions for a given input. In comparison, the knowledge obtained by the general interpretation is more stable and representative without data bias. In order to balance the desired characteristics of two interpretation categories, the proposed interpretation model is designed to distill the class-discriminative knowledge of network from a large dataset and then utilize the knowledge to explain certain predictions. The dataset has about a thousand images for each class. Meanwhile, class loyalty measure is added for each image in the CHIP model. Both of them make the CHIP model learn the class-discriminative importance accurately. Further, Refined CHIP is designed to provide visual interpretation that is both class-discriminative and high-resolution.

Weakly-supervised Localization. This task refers to localizing object only with the image-level class label.

One approach to the task is to utilize the visual interpretation distilled from the network to localize target object [13], [14]. From this point of view, visual interpretation model can be applied to the weakly-supervised localization task. In CAM, a network should be modified to a particular kind of architectures to learn the class activation map which can be used to generate an object bounding box for weakly-supervised localization. In the modified architectures, the convolutional layer is followed by the global average pooling layer and then the softmax layer. Compared with the original network, the modified network architecture may achieve inferior classification accuracy. Therefore, when addressing localization task based on class activation maps, the localization accuracy is also limited by the inferior accuracy of the modified network.

In Grad-CAM, the gradient of the output to feature maps are used as the weights of feature maps to obtain class activation map. Compared with CAM, Grad-CAM does not need to modify the network architecture and can be applied to different networks. It should be noted that the class-discriminative weights in Grad-CAM are only obtained from a single image. Such distilled knowledge from a single image are different from the class-discriminative knowledge which is learned from a large dataset in the training process of networks. In contrast, the proposed interpretation model distills the class-discriminative knowledge from a large dataset. Therefore, our visual interpretation is more reliable and informative, which provides more insight into the network and also enhances the localization accuracy.

III. Methodology

In this section, we present the proposed CHIP model. Its objective is to provide insight into neural networks based on the distilled class-discriminative knowledge. With this model, we can give visual interpretation for the decisions of networks and apply it to the weakly-supervised object localization. In section III-A, we describe the disentangled channels based on perturbed networks. In section III-B, we present the CHIP model to distill important channels for different classes. We also describe the way to obtain class-discriminative visual interpretation and the application of our model to the weakly-supervised localization. Section III-C describes the proposed CHIP algorithm.

A. Disentangled Channels based on Perturbed Networks

Class-discriminative importance of channels. In order to disentangle the roles of channels, our interpretation model is designed to learn the importance of channels for different classes. Through turning off partial channels each time, we can learn the class-discriminative importance of channels by analyzing the variation of predictions. The underlying principle is that the class prediction would drop dramatically if the forward propagation of important channels are blocked.

Specifically, for the l-th layer, the class-discriminative importance matrix is denoted as $\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2 \cdots \mathbf{w}_c \cdots \mathbf{w}_C]^T$, where C is the number of image category. For the c-th class, the class-discriminative importance is $\mathbf{w}_c = [w_c^1\ w_c^2 \cdots w_c^k \cdots w_c^K]$, where w_c^k represents the importance of the k-th channel and K is the number of channels in the l-th layer.

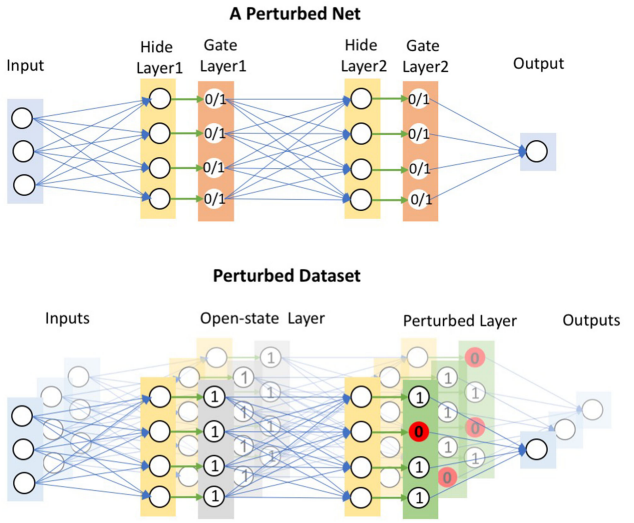

Fig. 1. Illustration of a perturbed network and the perturbed dataset.

Channel-wise perturbed networks. Inspired by the perturbation-based method [17], we perturb a pre-trained network $f(\cdot)$ by channel gates to learn the class-discriminative importance of channels. As shown in Fig. 1(top), each layer is associated with a gate layer in which each channel gate controls the state of the corresponding channel in the former layer. Here, in the l-th layer, we use a binary vector $\mathbf{d} = [d_1\ d_2 \cdots d_k \cdots d_K]^T$ to denote the channel gate. The k-th channel is turned off if d_k is zero. The perturbed network is generated by adding the channel gate layer after each original layer. We denote the original features in the l-th layer as $\mathbf{A} \in \mathcal{R}^{u,v,K}$, where u and v are the width and height of the channel. For the k-th channel in the l-th layer, the output of the channel gate layer is

$$\hat{\mathbf{A}}_k = d_k \mathbf{A}_k \tag{1}$$

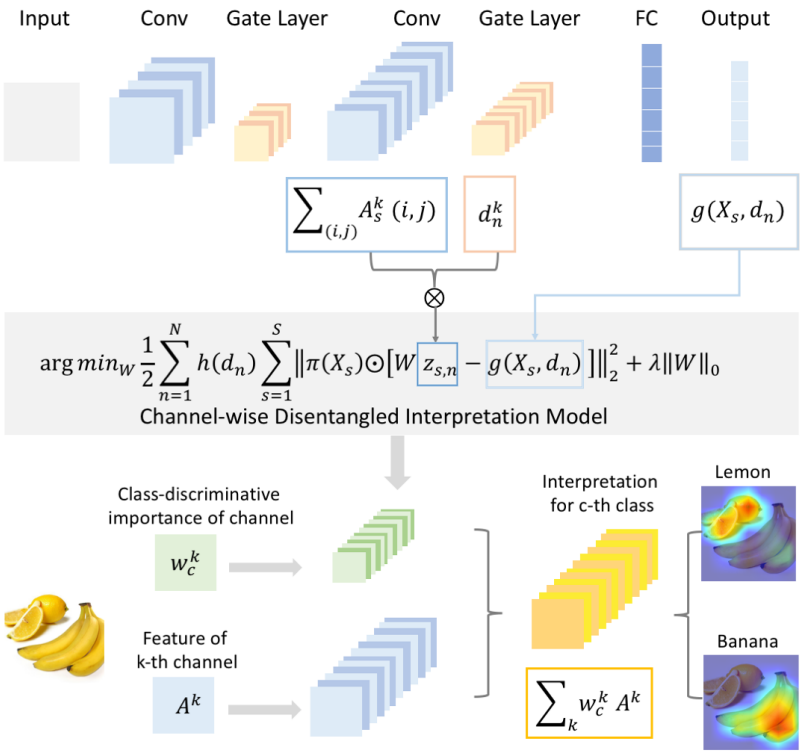

Fig. 2. The workflow of CHIP model. In the first stage, we learn the CHIP model based on the perturbed dataset which contains around a hundred million perturbed networks for the overall 1000 classes in ILSVRC 2015 dataset. After the optimization, we obtain the distilled class-discriminative importance of channels for different classes. In the second stage, given a specific image and the class of interest (banana and lemon), the class-discriminative visual interpretation is obtained for the target class by utilizing the distilled knowledge.

For the l-th layer, the global average pooling of the channel gate layer is $\mathbf{z} = [z_1\ z_2 \cdots z_k \cdots z_K]^T$. For the k-th channel, the global average pooling is

$$z_k = \frac{1}{uv} \sum_{i,j} \hat{\mathbf{A}}_k(i,j) \tag{2}$$

In the perturbed network, we add control gate layers behind each layer without changing the original weights of the pre-trained network. We can obtain different perturbed networks by changing the values in control gate layers.

Perturbed dataset. In order to learn the class-discriminative importance of channels, we need to generate the perturbed dataset, as shown in the bottom image of Fig. 1. Specifically, to learn $\mathbf{W}$ for the l-th layer, the perturbed dataset is obtained by the following three steps: The first step is to generate the perturbed networks. For the chosen l-th layer, we sample the channel gate values by using $\mathcal{D}_{gate} = \{\mathbf{d}_n\}_{n=1}^N$. And for other layers, we freeze the channel gate to be open. Secondly, we feed each image from image dataset $\mathcal{D}_{img} = \{\mathbf{X}_s\}_{s=1}^S$ into each perturbed network and get the results. For example, the results of n-th perturbed network for s-th image is denoted as

$$g(\mathbf{X}_s, \mathbf{d}_n) = [g^1(\mathbf{X}_s, \mathbf{d}_n) \cdots g^c(\mathbf{X}_s, \mathbf{d}_n) \cdots g^C(\mathbf{X}_s, \mathbf{d}_n)]^T \tag{3}$$

where $g^c(\mathbf{X}_s, \mathbf{d}_n)$ is the prediction for the c-th class. Meanwhile, its global average pooling in the l-th channel gate layer is recorded as $\mathbf{z}_{s,n}$. Finally, we get the perturbed dataset $\mathcal{D}\{\mathbf{X}_s, \mathbf{d}_n, \mathbf{z}_{s,n}, f(\mathbf{X}_s), g(\mathbf{X}_s, \mathbf{d}_n)\}$, where the results of the original network are denoted as

$$f(\mathbf{X}_s) = [f^1(\mathbf{X}_s) \cdots f^c(\mathbf{X}_s) \cdots f^C(\mathbf{X}_s)]^T \tag{4}$$

B. CHIP Model

Given the perturbed dataset $\mathcal{D}$, we formulate the CHIP model as

$$\arg\min_{\mathbf{W}} \frac{1}{2} \sum_{n=1}^{N} h(\mathbf{d}_n) \sum_{s=1}^{S} \|\pi(\mathbf{X}_s) \odot [\mathbf{W}\mathbf{z}_{s,n} - g(\mathbf{X}_s, \mathbf{d}_n)]\|_2^2 + \lambda \|\mathbf{W}\|_0 \tag{5}$$

where λ is the regularization parameter, and $\odot$ denotes Hadamard product.

The first term in our interpretation model is the loss function, which is denoted as

$$\mathcal{L} = \frac{1}{2} \sum_{n=1}^{N} h(\mathbf{d}_n) \sum_{s=1}^{S} \|\pi(\mathbf{X}_s) \odot [\mathbf{W}\mathbf{z}_{s,n} - g(\mathbf{X}_s, \mathbf{d}_n)]\|_2^2 \tag{6}$$

In the loss function, $h(\mathbf{d}_n)$ is denoted as the proximity measure between a binary channel gate vector $\mathbf{d}$ and the all-one vector $\mathbf{1}$. Specifically, it is defined as

$$h(\mathbf{d}_n) = \exp(-\frac{1}{\sigma^2} \|\mathbf{d}_n - \mathbf{1}\|_2^2) \tag{7}$$

We also denote $\pi(\mathbf{X}_s)$ as the loyalty measure of image $\mathbf{X}_s$ to different categories. Specifically, it is defined as

$$\pi(\mathbf{X}_s) = [\sqrt{f^1(\mathbf{X}_s)} \cdots \sqrt{f^c(\mathbf{X}_s)} \cdots \sqrt{f^C(\mathbf{X}_s)}]^T \tag{8}$$

where $f^c(\mathbf{X}_s)$ denotes the network prediction score of c-th class.

In our model, the second term is the regularization term. We use sparsity to measure the complexity of the interpretation model because of the inherent sparse property of features in DCNNs. Since the interpretation model needs to be simple enough to be interpretable, we employ the sparse regularization term as

$$\Phi(\mathbf{W}) = \|\mathbf{W}\|_0 \tag{9}$$

which measures the number of non-zero entries in the weight. Fig. 2 illustrates the workflow of the proposed CHIP model.

The proposed CHIP model learns the class-discriminative importance of channels for different classes. The distilled knowledge is capable to be further applied to visual interpretation and weakly-supervised object localization.

Class-discriminative Visualization. In the proposed interpretation model, we obtain the optimal class-discriminative importance of channels for different classes. By combining class-discriminative importance and corresponding feature maps, we can get the visual interpretation for a specific network decision of an input image. Specifically, for a certain layer, the visual interpretation for the c-th class prediction of a particular instance is denoted as

$$\tilde{\mathbf{A}}^c = \sum_k w_c^k \mathbf{A}_k \tag{10}$$

where w_c^k represents the optimal importance of the k-th channel to the c-th class, and $\mathbf{A}_k$ denotes the feature in the k-th channel for the given image.

The CHIP model can give visual interpretation for the decision of the network. Further, we design Refined CHIP to present high-resolution visual interpretation which can show the detailed features distilled from networks. It is commonly known that features in the shallow layer are of higher resolution than that in the deep layer. Conversely, the semantic representation in the deep layer is at higher level than that in the shallow layer. Likewise, CHIP interpretation for shallow layers and deep layers also possess similar properties. We can combine the distilled interpretation in different layers to obtain the Refined CHIP visual interpretation. Specifically, Refined CHIP result is obtained by the point-wise multiplication of CHIP visual interpretation in the first and the last convolutional layers. Therefore, our Refined CHIP interpretation is not only high-semantic but also high-resolution. Here, the Refined CHIP interpretation utilizes the distilled interpretation knowledge in different layers, which also reveals the roles of different layers. In comparison, Guided Grad-CAM combines Guided Backpropagation visualization with Grad-CAM to give a high-resolution class-discriminative interpretation. To compare with Guided Grad-CAM in the experiment, we also combine the Guided Backpropagation and CHIP visualization inspired by [15].

Weakly-supervised Localization. It is well known that deeper layers in DCNNs capture higher-level semantic information which can be regarded as the object saliency information for localization. Therefore, the intuition here is that our model distills the class-discriminative knowledge in the deep layer and transfers this knowledge from the pre-trained classification network to the localization task. Here, the visual interpretation can be regarded as a saliency map to localize the object. For this task, the bounding box can be obtained based on the saliency map distilled from the last convolutional layer. Because the visual interpretation obtained by our model is learned without the ground truth bounding box annotation, our localization method is also a weakly-supervised approach.

C. CHIP Algorithm

The proposed interpretation model is used to distill the important channels of a pre-trained network for different image categories. Here, we propose a channel-wise disentangled interpretation algorithm to optimize our interpretation model.

Here, the optimization of CHIP model in Eq. (5) can be divided into solving the optimization problem for each class separately. Due to the discrete and nonconvex nature of L_0 norm, the problem is NP hard which means that it is too complex to solve. To get an approximate solution to the problem, we replace L_0 norm with L_1 norm, since L_1 norm is naturally the best convex approximation of L_0 norm. For the c-th class, the optimization problem of interpretation model turns into

$$\begin{aligned}\arg\min_{\mathbf{w}_c} \frac{1}{2}\sum_{n=1}^{N} h(\mathbf{d}_n) \sum_{s=1}^{S} f^c(\mathbf{X}_s)[\mathbf{w}_c \mathbf{z}_{s,n} - g^c(\mathbf{X}_s, \mathbf{d}_n)]^2 \\ + \lambda \|\mathbf{w}_c\|_1\end{aligned} \tag{11}$$

The optimization problem in Eq. (11) is convex and can be solved. Here, we design a channel-wise disentangled interpretation algorithm by adopting the alternating iteration rule to learn $\mathbf{w}_c$.

The optimization problem can be converted into the equivalent formulation

$$\begin{aligned}\arg\min_{\mathbf{w}_c, \mathbf{m}_c} \frac{1}{2}\sum_{n=1}^{N} h(\mathbf{d}_n) \sum_{s=1}^{S} f^c(\mathbf{X}_s)[\mathbf{w}_c \mathbf{z}_{s,n} - g^c(\mathbf{X}_s, \mathbf{d}_n)]^2 \\ + \lambda \|\mathbf{m}_c\|_1 \\ \text{subject to} \quad \mathbf{w}_c = \mathbf{m}_c\end{aligned} \tag{12}$$

The augmented Lagrangian for the above problem is

$$\begin{aligned}\arg\min_{\mathbf{w}_c, \mathbf{m}_c, \mathbf{q}} \frac{1}{2}\sum_{n=1}^{N} h(\mathbf{d}_n) \sum_{s=1}^{S} f^c(\mathbf{X}_s)[\mathbf{w}_c \mathbf{z}_{s,n} - g^c(\mathbf{X}_s, \mathbf{d}_n)]^2 \\ + \lambda \|\mathbf{m}_c\|_1 + \frac{\rho}{2}\|\mathbf{w}_c - \mathbf{m}_c - \mathbf{q}\|_2^2\end{aligned} \tag{13}$$

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

$$\begin{aligned}\mathbf{w}_c^{i+1} \leftarrow (\sum_{s,n} f^c(\mathbf{X}_s) h(\mathbf{d}_n) g^c(\mathbf{X}_s, \mathbf{d}_n) {\mathbf{z}_{s,n}}^T + \rho \mathbf{m}_c^i \\ + \rho \mathbf{q}^i)(\sum_{s,n} f^c(\mathbf{X}_s) h(\mathbf{d}_n) \mathbf{z}_{s,n} {\mathbf{z}_{s,n}}^T + \rho \mathbf{I})^{-1}\end{aligned} \tag{14}$$

In order to calculate $\mathbf{m}_c$, the solution is

$$\mathbf{m}_c^{i+1} \leftarrow soft(\mathbf{w}_c^{i+1} - \mathbf{q}^i, \frac{\lambda}{\rho}) \tag{15}$$

Algorithm 1: Pseudocode of the CHIP Algorithm

Input: the perturbed dataset $\mathcal{D}\{\mathbf{X}_s, \mathbf{d}_n, \mathbf{z}_{s,n}, f(\mathbf{X}_s), g(\mathbf{X}_s, \mathbf{d}_n)\}$;
the original network feature $\mathbf{A}$ for the instance being explained;
Output: optimal $\mathbf{W}^*$;
the class-discriminative interpretation obtained by CHIP;
Initialization: set $c = 0$, $i = 0$, $\mathbf{m}_c^0$, $\mathbf{q}^0$, $\lambda > 0$, $\rho \geq 0$;
repeat
  The optimization problem for class-discriminative importance $\mathbf{w}_c$ of a certain layer for the c-th class;
  $\arg\min_{\mathbf{w}_c} \frac{1}{2} \sum_{n=1}^{N} h(\mathbf{d}_n) \sum_{s=1}^{S} f^c(\mathbf{X}_s)[\mathbf{w}_c \mathbf{z}_{s,n} - g^c(\mathbf{X}_s, \mathbf{d}_n)]^2 + \lambda \|\mathbf{w}_c\|_1$;
  repeat
    1: $\mathbf{w}_c^{i+1} \leftarrow (\sum_{s,n} f^c(\mathbf{X}_s) h(\mathbf{d}_n) g^c(\mathbf{X}_s, \mathbf{d}_n) \mathbf{z}_{s,n}{}^T + \rho \mathbf{m}_c^i + \rho \mathbf{q}^i)(\sum_{s,n} f^c(\mathbf{X}_s) h(\mathbf{d}_n) \mathbf{z}_{s,n} \mathbf{z}_{s,n}{}^T + \rho \mathbf{I})^{-1}$;
    2: $\mathbf{m}_c^{i+1} \leftarrow soft(\mathbf{w}_c^{i+1} - \mathbf{q}^i, \frac{\lambda}{\rho})$;
    3: update lagrange multipliers: $\mathbf{q}^{i+1} \leftarrow \mathbf{q}^i - (\mathbf{w}_c^{i+1} - \mathbf{m}_c^{i+1})$;
    4: **update iteration:** $i \leftarrow i + 1$;
  until *stopping criterion is satisfied*;
  update iteration: $c \leftarrow c + 1$;
until *class-discriminative importance of a certain layer for all categories are optimized*;
return $\mathbf{W}^*$;
CHIP: for a given instance, the visual interpretation of a certain layer for the target class is obtained by $\tilde{\mathbf{A}}^c = \sum_k w_c^k \mathbf{A}_k$;

Lagrange multipliers update to

$$\mathbf{q}^{i+1} \leftarrow \mathbf{q}^i - (\mathbf{w}_c^{i+1} - \mathbf{m}_c^{i+1}) \tag{16}$$

The pseudocode of CHIP algorithm is shown in Algorithm 1.

IV. Experiments

In this section, we conduct a series of experiments to evaluate the performance of CHIP model and the Refined CHIP interpretation. Before the analysis of experiment results, we describe the experiment settings of our interpretation model in section IV-A. Then, we conduct experiments in four aspects. In section IV-B, we give visual interpretation for images in different classes based on the proposed CHIP and Refined CHIP. In section IV-C, the performance of CHIP model on weakly-supervised localization task is evaluated. In section IV-D, we show the visual interpretation in different layers of the neural network. In section IV-E, the class-discriminative importance of channels for different classes are compared.

A. Experiment Settings

Image Dataset. Here, we use the ILSVRC 2015 [21] training dataset to learn the CHIP model. The ILSVRC 2015 training dataset includes 1281167 images of total 1000 classes. For each class, the number of images ranges from 732 to 1300.

Perturbed Dataset. To be consistent with earlier work, we use the off-the-shelf VGG16 model [22] from the Caffe [23] Model Zoo as the network to be interpreted. In VGG16 model, there are 16 layers with learnable weights. The number of channels is not uniform in different layers. For example, the first convolutional layer contains 64 channels while the last convolutional layer has 512 channels.

By adding gate layers on top of each original layer in VGG16, we can build perturbed networks for the learning of our interpretation model. Each control gate in a gate layer controls the state of the corresponding channel in the original layer of VGG16. Specifically, for a control gate, zero and one represent the close and open of the corresponding channel in the previous layer, respectively.

In order to learn the class-discriminative importance of channels in a certain layer for 1000 classes, we generate $100 \times Image\,Number\,in\,Dataset$ perturbed networks, totally around a hundred million perturbed networks for the overall 1000 classes. In each perturbed network, we only perturb the layer of interest, while keeping other gate layers in open states. We get the perturbed dataset by feeding images into perturbed networks, as described in Section III-A. After the optimization of CHIP model, we can get the class-discriminative importance of channels in the layer of interest for each class.

B. Visual Interpretation for Different Classes

In this section, we evaluate the performance of the visual interpretation of CHIP and Refined CHIP for different classes. Here, we learn class-discriminative importance of channels in the first and the last convolutional layers for 1000 classes. For evaluation, we choose images from ILSVRC 2015 validation dataset as the test images.

Given the optimized results of our CHIP model, we can present class-discriminative visual interpretation for the predictions of the network. Previous work has demonstrated that the semantic representation in the deep convolutional layer is more class-specific than that in the shallow convolutional layer. In light of that, CHIP model focuses on the class-discriminative visual interpretation in the last convolutional layer. Meanwhile, features in the shallow layer are of higher resolution than that in the deep layer. Therefore, the distilled interpretation in the first and the last convolutional layers can be combined to obtain the Refined CHIP interpretation which is both high-semantic and high-resolution. Here, we show the CHIP and Refined CHIP results in comparison with the LIME [17],

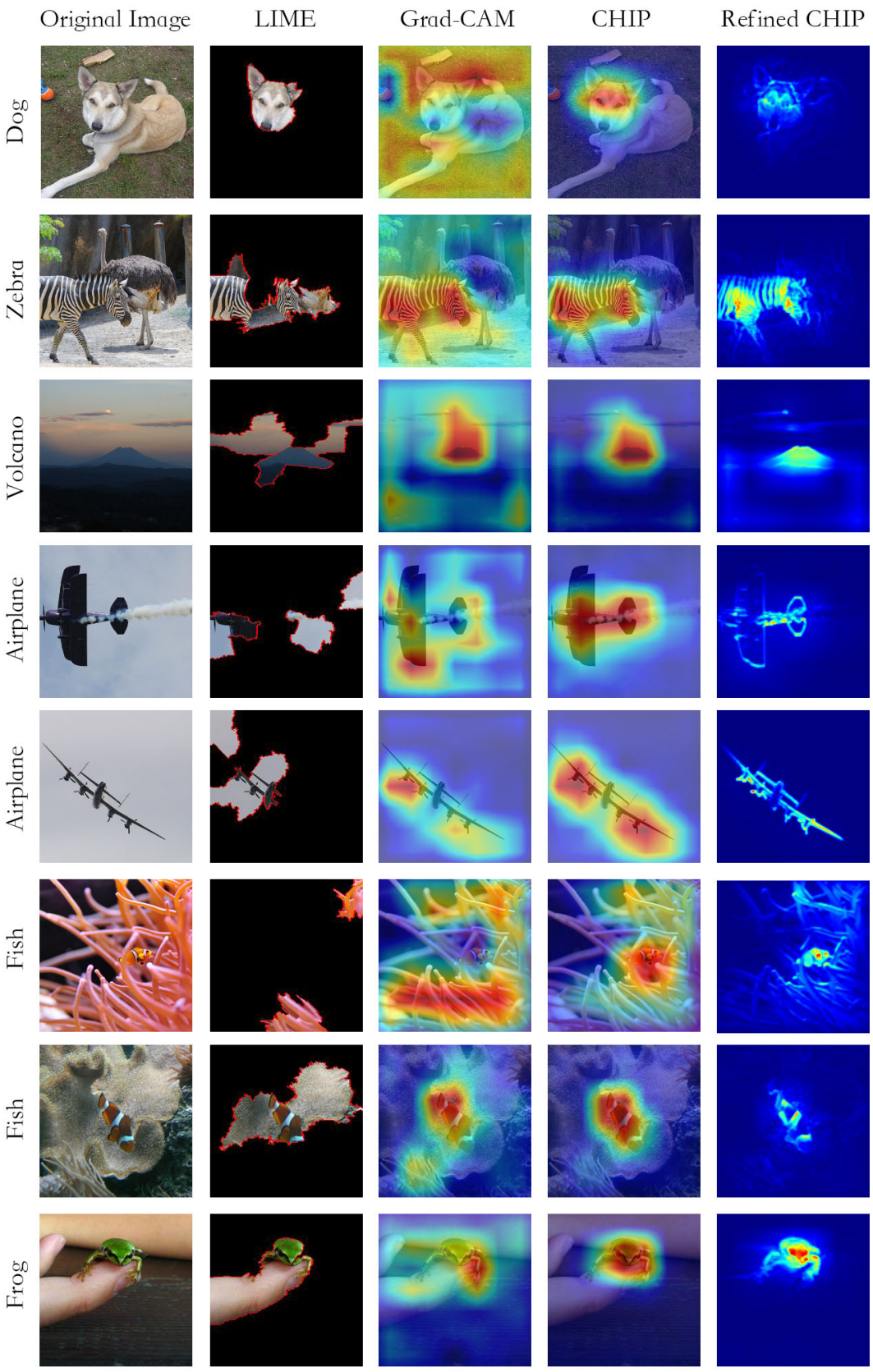

Fig. 3. Comparison of visual interpretation for simple one-object images.

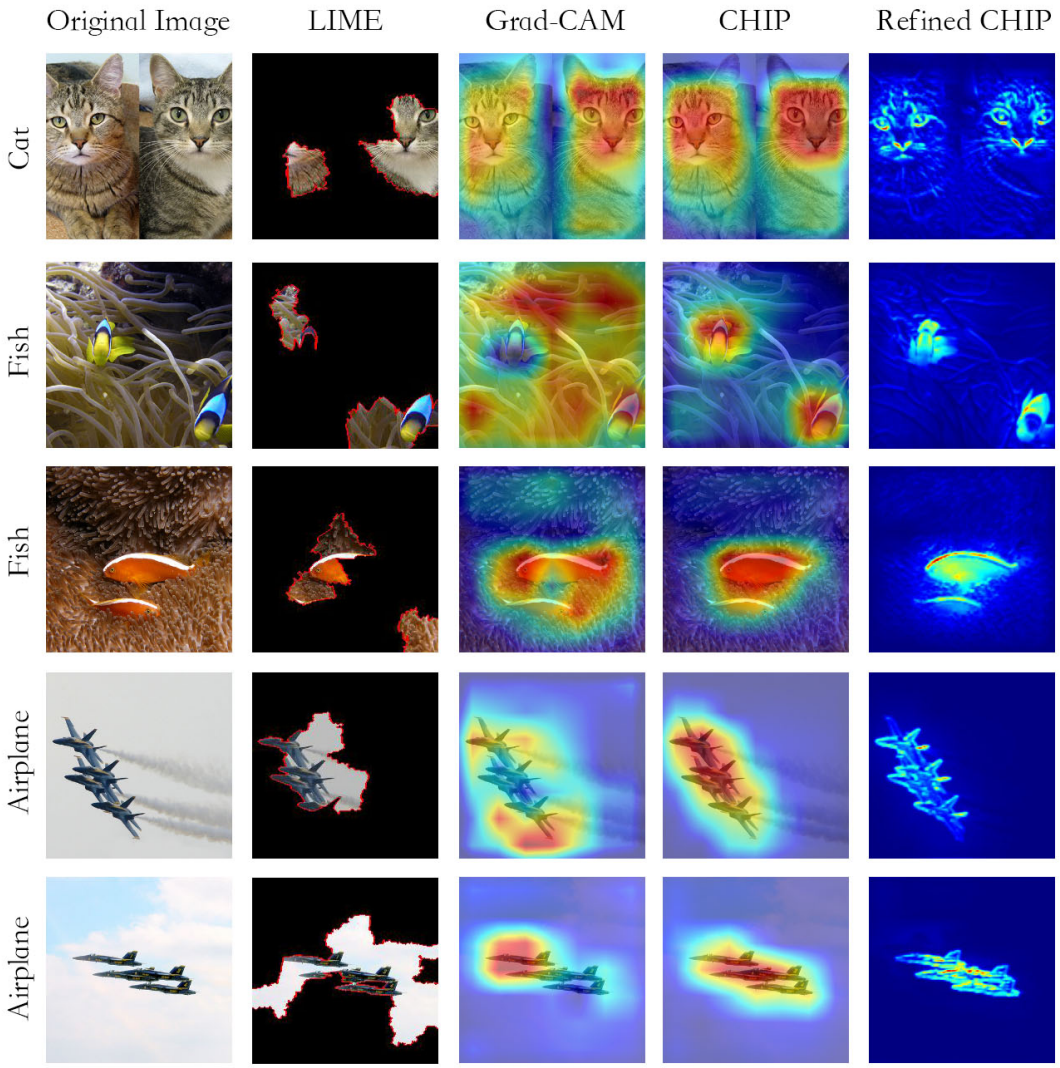

Fig. 4. Comparison of visual interpretation for complex multi-object images.

Grad-CAM [14], Guided Backpropagation [15] and Guided Grad-CAM [14]. LIME and Grad-CAM can only present coarse class-discriminative visual interpretation. LIME explains the prediction by selecting image region that is important to the target class. In contrast, Grad-CAM provides visual interpretation for the prediction by utilizing the internal features. Guided Backpropagation and Guided Grad-CAM can provide the high-resolution visual interpretation. However, Guided Backpropagation is not a class-discriminative interpretation method.

Fig. 3 shows the comparison of visual interpretation of different methods for simple images each of which only contains a single target of the selected class. For instance, the first original image in Fig. 3 contains a dog, and the visual interpretation results of the different methods are provided for the dog class. For LIME, the class-discriminative importance is calculated for each super-pixel. In the experiment, LIME shows the top 5 important super-pixels for the target class. For Grad-CAM and CHIP, the visual interpretation is designed at pixel level, which can also be regarded as the saliency maps for the target class. Similar with previous work, the visual interpretation of Grad-CAM and CHIP is displayed by superimposing saliency map on the original image. Because the results of Grad-CAM and CHIP are not high-resolution, this displaying way can make it easier to evaluate whether the important pixels in visual interpretation belong to the object region of target class. In contrast, this displaying way is not necessary for Refined CHIP because it is high-resolution interpretation containing fine-scale features. Therefore, we directly show the interpretation results of Refined CHIP.

From Fig. 3, we can observe that LIME, in most cases, is able to capture partial target object as the visual interpretation, except for the sixth original image where a fish is hiding near coral. The flaw of LIME is that its visual interpretation is limited by image segmentation result. LIME also cannot explore inside the network and understand the internal mechanism, which restricts its performance. Its results usually contain image region that does not belong to the target object, such as the

fourth LIME explanation where the sky is included as the critical image region for the airplane.

Compared with LIME and Grad-CAM, CHIP is more reasonable which always captures the target object. For instance, in the explanation of the sixth original image, the compared methods both miss the fish object region and incorrectly highlight the background except for the proposed CHIP and Refined CHIP. It illustrates that CHIP and Refined CHIP have better class-discriminative characteristics. On top of CHIP, Refined CHIP presents fine-grained details for the target object, which further illustrates the inherent characteristics of features in different layers. For example, for the second original image in Fig. 3, Refined CHIP depicts the zebra-stripes in the visual interpretation for zebra. In contrast, other compared methods can only provide coarse visual interpretation, such as the super-pixel in LIME and the saliency region in Grad-CAM.

Fig. 4 further shows the visual interpretation in complex images each of which contains multiple objects of the target class. In this situation, a good visual interpretation is expected to highlight all objects of target category, which is more difficult. In this figure, the compared methods neglect some target objects in some cases. As shown in the second row of Fig. 4, LIME and Grad-CAM both miss one fish. In contrast, CHIP provides more comprehensive visual interpretation which captures all objects of the target category. Refined CHIP further shows more detailed visual interpretation for the target category, such as the explanation for the fourth original image where the contour and texture of airplanes are presented.

Fig. 5 shows the visual interpretation for an image where there are multiple target categories. The target categories are lemon and banana in the original image. The visual interpretation for banana and lemon are respectively in the top and bottom half of Fig. 5. In this figure, we also show the performance of Refined CHIP compared with Guided Backpropagation, Guided Grad-CAM, and Guided CHIP.

Fig. 5 indicates that CHIP performs more reasonable than Grad-CAM. As shown in Fig. 5(h), Grad-CAM does not provide a reasonable explanation for lemon, because it also highlights partial region of banana. Fig. 5(g) and (m) indicate Guided Backpropagation provides high-resolution but not class-discriminative visualization, because its visualization results for lemon and banana are similar. The visual interpretation of Guided Grad-CAM and Guided CHIP involve rich details by combining Guided Backpropagation. However, because of the inferior visual interpretation of Grad-CAM, Guided Grad-CAM is also less reasonable than Guided CHIP. As shown in Fig. 5(k) and (l), Guided Grad-CAM for lemon highlights banana region, while Guided CHIP only highlights lemon region. As shown in Fig. 5(d) and (j), Refined CHIP presents visual interpretation that is both high-resolution and class-discriminative without using Guided Backpropagation. Except for the correctly highlighted object region, the fine-grained information in Refined CHIP further identifies the important fine-scale features in the network for the predicted class.

C. Weakly-supervised Localization

In weakly-supervised object localization task, competing methods should localize the target object without the ground-truth localization annotation. In this experiment, competing approaches are evaluated in the ILSVRC 2015 weakly-supervised localization challenge where they are required to provide object bounding boxes together with classification predictions. Specifically, given an input image, the original classification network gives its class prediction. And different competing methods are used to learn the saliency map from the interpretation of the classification network for the predicted class. The obtained saliency map is binarized with the optimal threshold of the maximal intensity. The corresponding bounding box is obtained around the largest partition in the binarized saliency map. For each competing method, a grid search is implemented to select the optimal threshold for the best localization performance. Finally, the compared methods can get the bounding boxes as the localization results by utilizing the interpretation results. Here, quantitative and qualitative analyses are provided to evaluate the results of competing methods.

In the quantitative comparison, we evaluate the localization error using the Intersection over Union (IoU) metric. The IoU metric is defined as follows:

$$\mathbf{IoU} = \frac{\textit{Area of Overlap}}{\textit{Area of Union}} \tag{17}$$

where the numerator refers to the area of overlap between the predicted bounding box and the ground-truth bounding box for the target category. And the denominator denotes the area of the union which includes both the predicted bounding box and the ground-truth bounding box.

Table II shows the localization errors of competing methods in the ILSVRC-15 validation dataset. Here, the proposed method is compared with Grad-CAM [14], c-MWP[24], and Backpropagation methods [12]. To evaluate the pre-trained VGG-16 neural network in the context of image classification, both top-1 and top-5 classification error in validation dataset are reported in Table II. Because all competing methods are based on the same pre-trained VGG-16 network, their classification accuracies are the same. Table II shows both top-1 and top-5 localization errors in validation dataset. Table II indicates that the localization error of our model is lower than other methods. It also indicates that our model provides better class-discriminative saliency maps than other visual interpretation models in this task.

Fig. 6 shows the qualitative comparison of Grad-CAM and the proposed approach. Because Grad-CAM achieves better object localization results than c-MWP and Backpropagation, here we only show the results of Grad-CAM for comparison. It can be seen that the object localization of CHIP outperforms that of Grad-CAM. As shown in Fig. 6, the predicted bounding boxes of Grad-CAM often cover the whole image region, such as the Grad-CAM results for images in the top two rows. The reason behind it is the saliency map of Grad-CAM sometimes highlights the background region rather than the target object, which results in the inaccurate localization. In comparison, the predicted bounding boxes of the proposed method have a higher overlap ratio with the ground-truth annotations, resulting in a lower localization error of CHIP model.

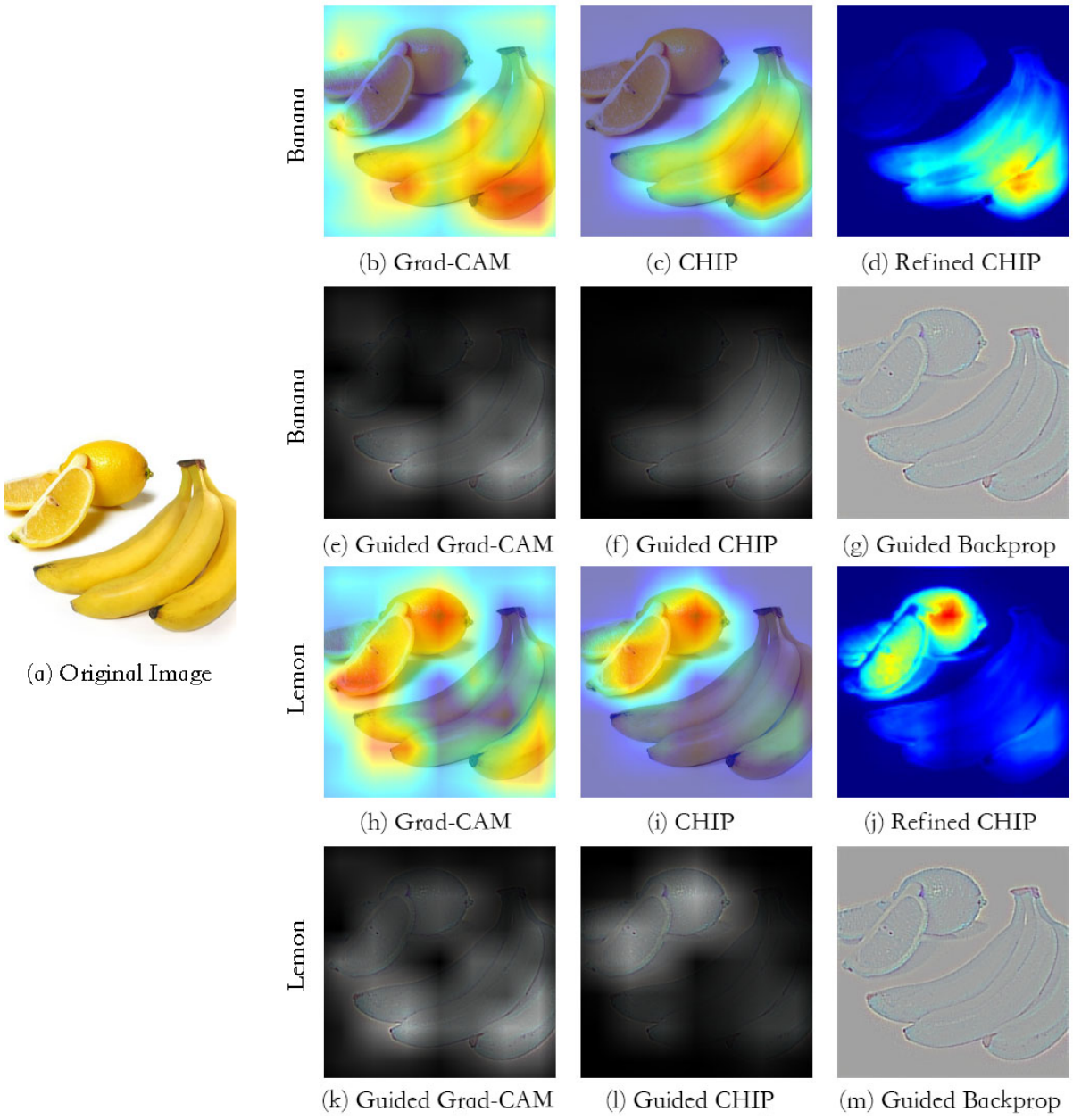

Fig. 5. Comparison of visual interpretation for the multi-category image.

TABLE II
LOCALIZATION AND CLASSIFICATION ERRORS ON ILSVRC-15 VALIDATION SET

Localization method	Top-1 loc error	Top-5 loc error	Classification network	Top-1 cls error	Top-5 cls error
Backpropagation [12]	61.11	51.43	VGG16	30.12	10.85
c-MWP [24]	70.92	63.01			
Grad-CAM [14]	56.47	46.35			
CHIP	**51.45**	**40.16**			

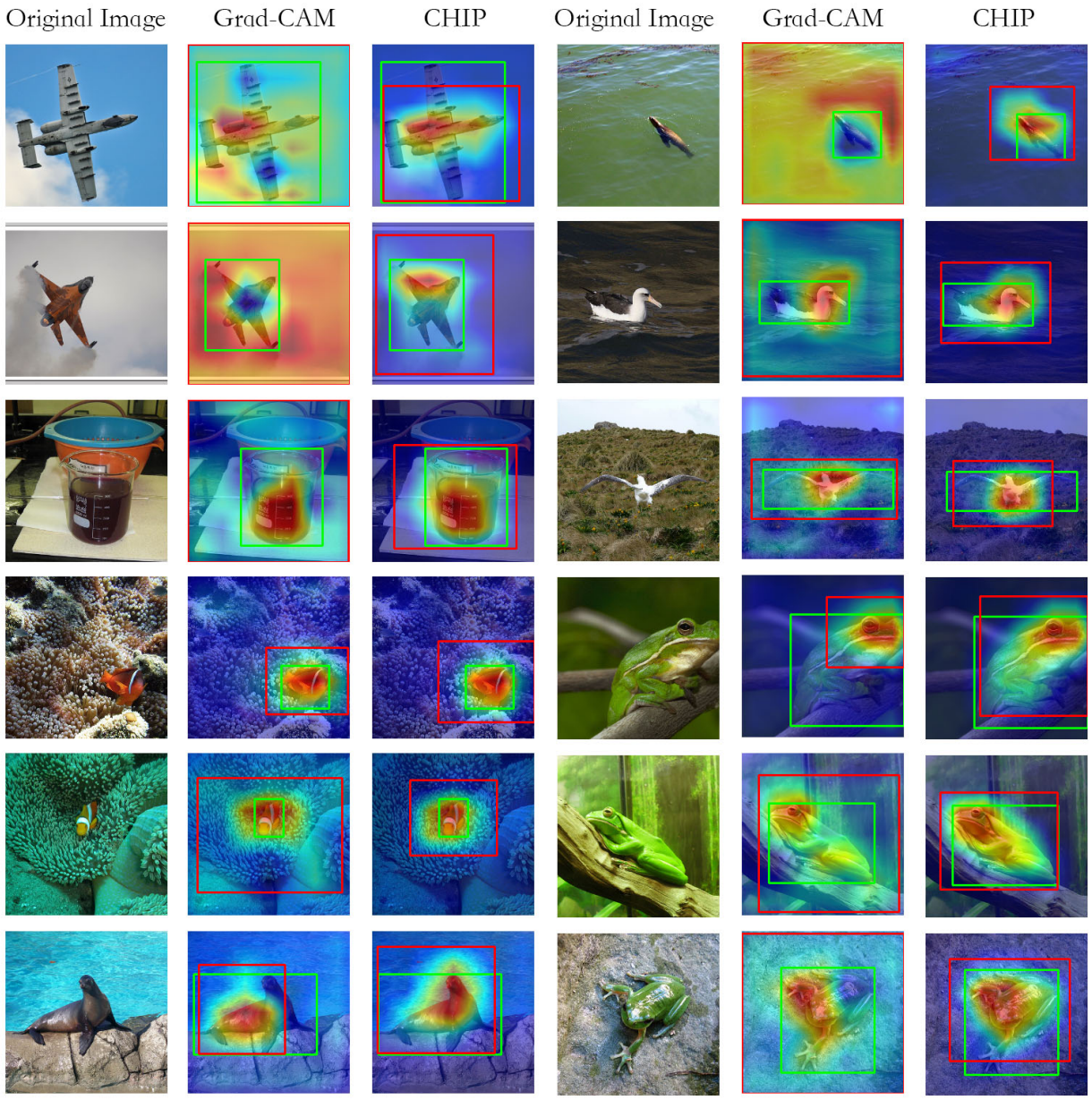

Fig. 6. Visual examples depicting the object localization performance of Grad-CAM and CHIP. The green boxes are ground-truth annotations for images and red boxes are predicted bounding boxes for the competing methods.

D. Visual Interpretation in Different Layers

In this section, we conduct experiments to analyze the visual interpretation in different layers. Here, we select five pooling layers from VGG16 as the target layers. For each pooling layer, the class-discriminative importance of channels for different classes are learned by the CHIP model. Visual interpretation is then obtained based on the learned importance of channels and the corresponding features. Here, we choose images from ILSVRC 2015 validation dataset as test images.

Fig. 7 shows the visual interpretation for simple images. Here, "simple" means the object of target class is single and can be easily distinguished from background scenario. In Fig. 7, we select images from three classes, cat, dog, and bird, to show the visualization. Original images are in the leftmost column, and corresponding visualization is aligned in the right side. Visual interpretation in the pool1 layer highlights fine-grained object details of target class, such as edges and textures shown in the visualization for cat. In contrast, the visualization in deeper layers capture higher-level semantic features of objects, such as eyes and ears of dog shown in the visualization of the pool4 layer. This inference is consistent with the existing work about visualization [25], [26]. Here, it should be noted that the visualization in different layers highlights the image region that is specific to the class of interest, which qualitatively demonstrates the effectiveness of the class-discriminative property of CHIP model.

Moreover, Fig. 8 compares our visual interpretation with Grad-CAM in different layers for zebra class in complex

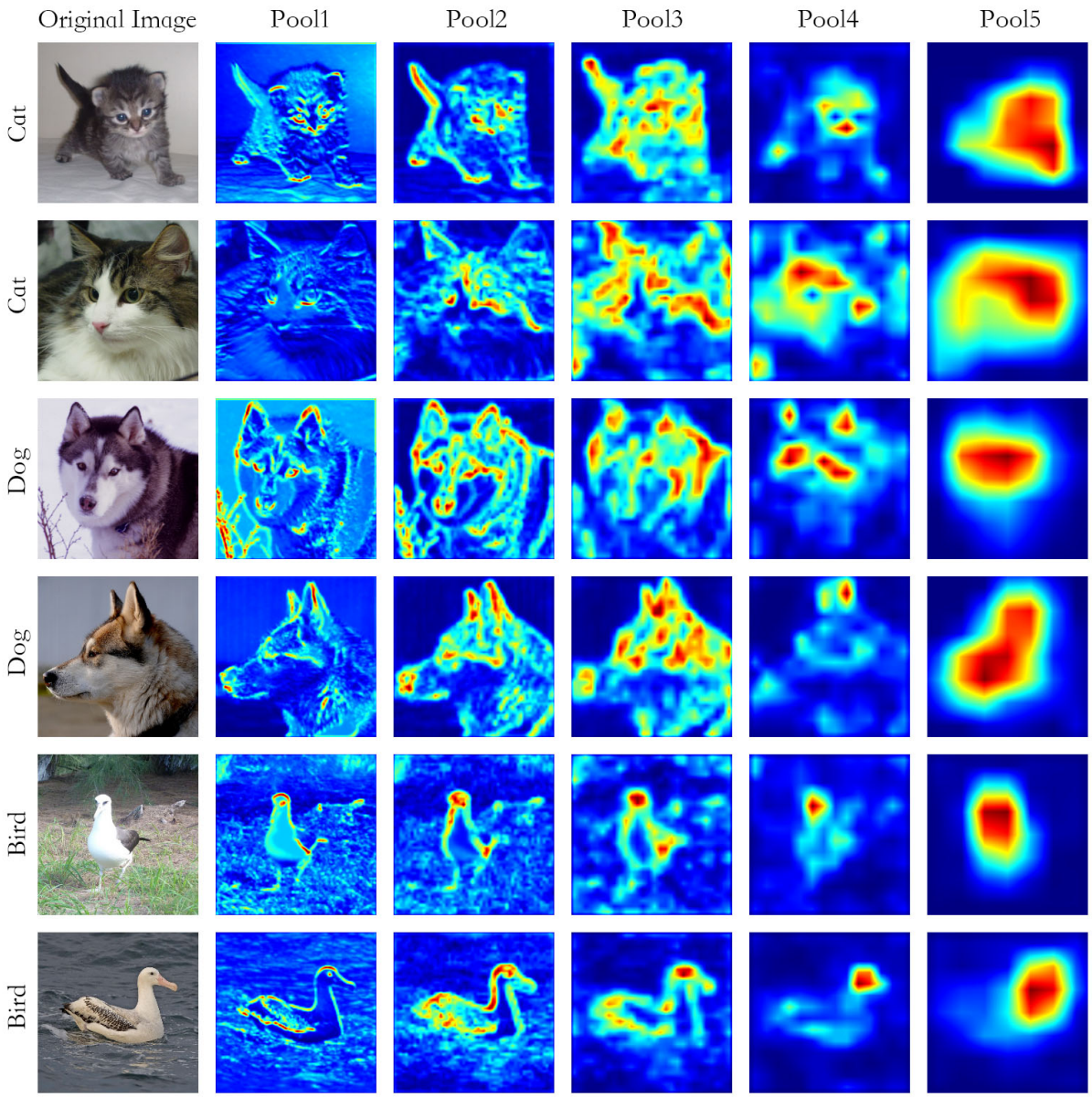

Fig. 7. Visual interpretation in different layers of VGG16 model for simple images.

images. Here, "complex" means there are multiple objects of different classes in the complex background. For example, the top original image has two zebras and some bisons in the grassland.

In Grad-CAM, the visualization is obtained by the weighted sum of features in the last convolutional layer. Grad-CAM is only applied to the last convolutional layer in [14]. Meanwhile, the channel weights in Grad-CAM are obtained from an individual image. Fig. 8 indicates that the visualization of Grad-CAM in former layers is not reasonable, which always concentrate on the background. For example, for the top image in Fig. 8, Grad-CAM highlights the background but rather zebra object from pool1 layer to pool4 layer.

In contrast, because the proposed model is based on the network perturbation, our model can give visual interpretation in different layers. The class-discriminative importance of the CHIP model is obtained from an image dataset. Therefore, CHIP captures more reliable class-discriminative interpretation in different layers. In Fig. 8, the CHIP model can always highlight the zebra region in these layers. For the pool5 layer, Grad-CAM presents more reasonable results than its former pooling layers. However, for the middle image in Fig. 8, our approach achieves better visual interpretation than the compared method in the pool5 layer.

Fig. 7 and Fig. 8 both illustrate that the visualization in the deeper layer captures better class-discriminative representation than the former layers, which means the saliency of the target object is more obvious in deeper layers.

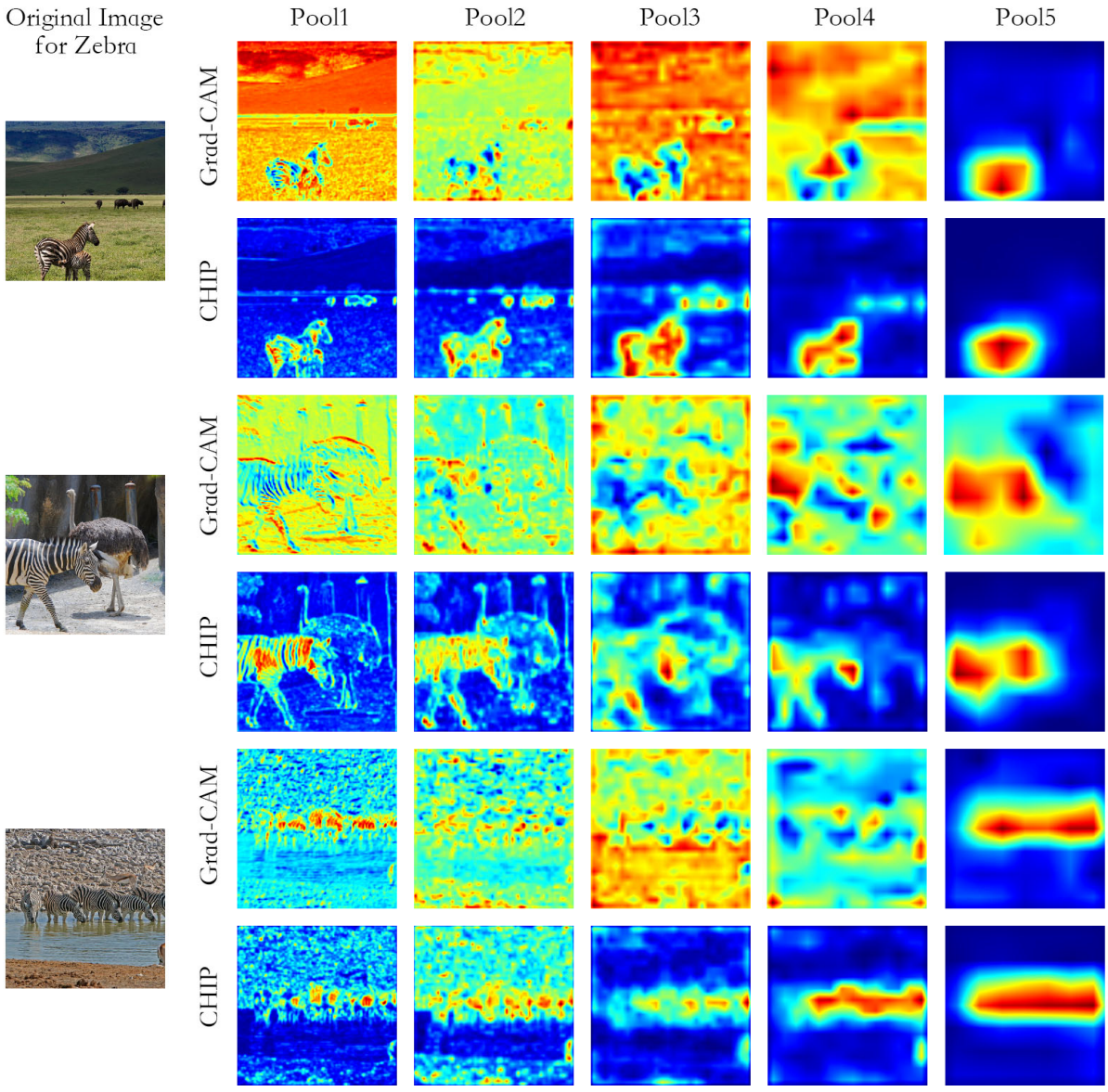

Fig. 8. Visual interpretation in different layers of VGG16 model for complex images.

E. Class-discriminative Importance of Channels for Different Classes

In the section, we compare the class-discriminative importance of channels between different classes. Here, we select three classes as examples: bullfrog, tree frog, and zebra. The former two classes belong to the different species of frogs. While the third class barely shares similarity with the former two classes. Because deeper layer can capture higher-level semantic representation, we select the last convolutional layer to study the property of the class-discriminative importance.

Fig. 9 shows the comparison of the class-discriminative importance of channels for different classes, where the vertical axis and horizontal axis represent the channel index and the class respectively. The figure illustrates the sparsity of the class-discriminative importance of channels for different classes. Only a small subset of channels are important for each class. Fig. 9 also shows that similar classes (bullfrog and tree frog) have some common important channels, while different classes (bullfrog and zebra) seldom have overlapped important channels.

To further illustrate this observation, we plot two Venn diagrams comparing the number of overlapped important channels among the three classes. In Fig. 10, the left Venn diagram shows the number of overlapped channels within the top 10 importance channels. The right one shows the number of overlapped channels in the important channels whose importance are larger than a thousandth of the highest one for each class. In the right Venn diagram, the number of important channels for each class is not uniform. Fig. 10 illustrates that the number of overlapped important channels between similar classes (bullfrog and tree frog) is larger than that between the dissimilar classes (bullfrog and zebra).

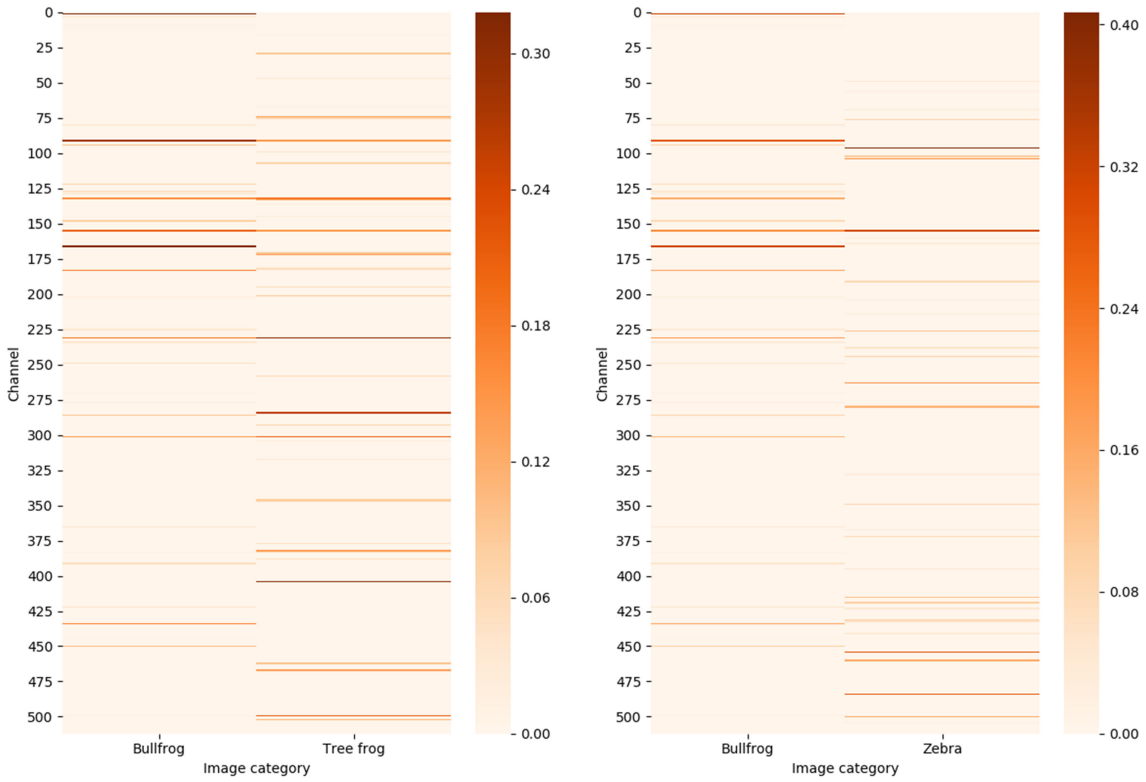

Fig. 9. Comparison of the class-discriminative importance of channels for different classes.

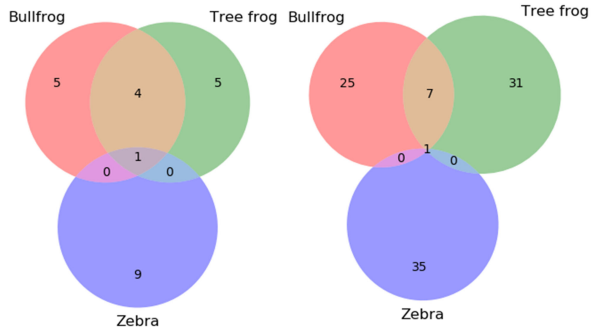

Fig. 10. The Venn diagrams of the overlapped channels between different classes.

V. Conclusion

In the work, we proposed a novel CHIP model, which can provide visual interpretation for the predictions of networks without requiring the retraining of networks. Further, we combine the visual interpretation in the first and the last convolutional layers to obtain Refined CHIP visual interpretation that is class-discriminative and high-resolution. Through experiment evaluation, we have demonstrated that the proposed interpretation model can provide more reasonable visual interpretation compared with previous methods. The proposed method can also outperform other visual interpretation methods in the application of weakly-supervised object localization in ILSVRC 2015 benchmark.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2980–2988.

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

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and VQA," *CoRR*, vol. abs/1707.07998, 2017. [Online]. Available: http://arxiv.org/abs/1707.07998

[6] 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, pp. 110–135, Dec 2017.

[7] Z. C. Lipton, "The mythos of model interpretability," *CoRR*, vol. abs/1606.03490, 2016. [Online]. Available: http://arxiv.org/abs/1606.03490

[8] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "Mdnet: A semantically and visually interpretable medical image diagnosis network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3549–3557.

[9] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 1885–1894.

[10] H. Lakkaraju, E. Kamar, R. Caruana, and E. Horvitz, "Identifying unknown unknowns in the open world: Representations and policies for guided exploration," in *AAAI*, 2017.

[11] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 3449–3457.

[12] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *CoRR*, vol. abs/1312.6034, 2013. [Online]. Available: http://arxiv.org/abs/1312.6034

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

[14] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 618–626.

[15] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," *CoRR*, vol. abs/1412.6806, 2014. [Online]. Available: http://arxiv.org/abs/1412.6806

[16] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[17] 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.

[18] A. Qygard, "Deepdraw," https://github.com/auduno/deepdraw, 2015.

[19] D. Wei, B. Zhou, A. Torralba, and W. T. Freeman, "Understanding intra-class knowledge inside CNN," *CoRR*, vol. abs/1507.02379, 2015. [Online]. Available: http://arxiv.org/abs/1507.02379

[20] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017, https://distill.pub/2017/feature-visualization.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec 2015.

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

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 675–678.

[24] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, Oct 2018.

[25] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *CoRR*, vol. abs/1310.1531, 2013. [Online]. Available: http://arxiv.org/abs/1310.1531

[26] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2014, pp. 512–519.