Interpreting Bottom-up Decision-Making of CNNs via Hierarchical Inference

Dan Wang[2], Xinrui Cui[1†], Xun Chen[1], Rabab Ward[2], *Fellow, IEEE* and Z. Jane Wang[2], *Fellow, IEEE*

***Abstract*—With the great success of convolutional neural networks (CNNs), interpretation of their internal network mechanism has been increasingly critical, while the network decision-making logic is still an open issue. In the bottom-up hierarchical logic of neuroscience, the decision-making process can be deduced from a series of sub-decision-making processes from low to high levels. Inspired by this, we propose the Concept-harmonized HierArchical INference (CHAIN) interpretation scheme. In CHAIN, a network decision-making process from shallow to deep layers is interpreted by the hierarchical backward inference based on visual concepts from high to low semantic levels. Firstly, we learned a general hierarchical visual-concept representation in CNN layered feature space by concept harmonizing model on a large concept dataset. Secondly, for interpreting a specific network decision-making process, we conduct the concept-harmonized hierarchical inference backward from the highest to the lowest semantic level. Specifically, the network learning for a target concept at a deeper layer is disassembled into that for concepts at shallower layers. Finally, a specific network decision-making process is explained as a form of concept-harmonized hierarchical inference, which is intuitively comparable to the bottom-up hierarchical visual recognition way. Quantitative and qualitative experiments demonstrate the effectiveness of the proposed CHAIN at both instance and class levels.**

***Index Terms*—Interpretation model, bottom-up hierarchical visual recognition, decision-making process.**

I. Introduction

Convolutional neural networks (CNNs) have been successfully applied to various tasks, such as image classification [1], object detection [2], and image processing [3]. However, the internal mechanism of networks is still unclear [4] and has attracted increasing research attention. Among them, visual interpretation is a popular direction. Previous visual interpretation methods mainly focus on visualizing internal layers and explaining their patterns [5].

In recent works [6], [7], [8], researchers explain CNNs by utilizing interpretable visual concepts in human perception. In [6], authors learn the concept representation in an internal CNN layer by aligning visual concepts with network units. However, existing concept-based interpretation works only focus on explaining features in an internal layer by concepts, not the network decision-making process from shallow to deep layers. Interpretation of the network decision-making process remains an open issue.

Neuroscientists report that hierarchical decisions are formed in the visual cortex [9]. To explore human visual recognition, neuroscientists have proposed different mechanisms, including bottom-up hierarchy, integrative hierarchies, and parallel processing [10].

In the network decision-making process, features from low to high semantic levels are learned from shallow to deep layers in CNNs [5]. Therefore, we intend to explore the network decision-making process in a way similar to the bottom-up hierarchy in neuroscience. The bottom-up hierarchy is viewed as a visual recognition processing in which information is processed sequentially with increasing complexities [11], [12], [13]. At the early vision stage, basic features of an image, such as color, texture, are determined. Then, basic features are combined into recognizable object groups at the middle vision stage. High-level vision is the stage that understands the scene. Inspired by the bottom-up hierarchy, we attempt to explore how the CNN decision-making process can be understood similarly. To answer it, we need to address the two following issues:

- For human visual recognition, our logic is built based on hierarchical visual concepts. The first issue is how to explain feature representations in CNNs based on hierarchical visual concepts.
- In human bottom-up logic, a decision can be hierarchically deduced from sub-decisions [9]. Therefore, the second issue is how to interpret a specific network decision-making process as the form of a hierarchical inference.

For the first issue, the concept harmonizing model is applied to interpret network features from shallow to deep layers by utilizing visual concepts from low to high semantic levels. To address the second issue, we propose the Concept-harmonized HierArchical INference (CHAIN) interpretation to explain the network decision-making process. Given a specific network decision-making process being interpreted, CHAIN can infer a concept at a deeper layer backward to concepts at a shallower layer, as shown in Fig. 1.

The main contributions are as follows:

- We propose the CHAIN interpretation to explain the network decision-making process for a specific instance. It can interpret the bottom-up decision-making process of CNNs via hierarchical inference of semantic concepts from low to high levels.
- To interpret the CNN logic inference from shallow to deep layers, CHAIN models a deeper concept as a sparse combination of shallow concepts.

† Corresponding author: Xinrui Cui. e-mail: (xinruic@ece.ubc.ca).

1. Department of Electronic Engineering and Information Science, University of Science and Techonology of China.

2. Department of Electrical and Computer Engineering, University of British Columbia, BC, Canada.

e-mail: (danw@ece.ubc.ca; xinruic@ece.ubc.ca; xunchen@ustc.edu.cn; rababw@ece.ubc.ca; zjanew@ece.ubc.ca)

Contact xinruic@ece.ubc.ca for further questions about this work.

Fig. 1. Illustration of the CHAIN interpretation. CHAIN provides both instance-level and class-level interpretations to explain the network decision-making process. Specifically, a network decision-making process can be interpreted by the hierarchical inference of visual concepts from low to high semantic levels.

- We can also use CHAIN to give qualitative and quantitative interpretations for intra-class and inter-class network decisions at the class level.

II. Related Work

In this section, we review three main directions of network interpretation:

Input-based network interpretation. This direction explains CNNs by learning critical input regions to a particular network prediction. Some methods utilize the perturbation mechanism in which critical input regions are obtained by perturbing different input regions and observing the variation of the prediction [14], [15], [16], [17]. Zeiler et al. [14] slide a regular grid as the perturbation patch over an image to find an important square region for the prediction. LIME [15] uses super-pixels as perturbation patches and calculates the importance of each super-pixel via local linear regression. Wagner et al. [18] optimize a perturbed version of an input image to highlight the evidence in the image for a specific prediction. These methods map from the output space to the input space. However, such visual interpretations are only in the input space and do not explain the internal network mechanism.

Feature-based network interpretation. Another interpretation direction is based on network features. Some approaches use visualization to understand network features. Guided Backpropagation [19] and Deconvolution [14] visualize the image pattern which can obtain the largest activation of a particular network unit. Olah et al. [5] give a reliable tool to visualize network units with high quality in different layers. However, these methods only provide a general interpretation, which means they cannot explain a specific network decision for a given input. In comparison, CAM [20], and Grad-CAM [21] provide the class-discriminative interpretation of a specific network decision by visualizing the linear combination of features. Nevertheless, these approaches are not suited for fine-grained interpretation. Also, their interpretations are not as effective for shallow layers as those for the last layer. To tackle the challenge, CHIP distills class-discriminative network units from shallow to deep layers to interpret network features [22].

Concept-based network interpretation. Researchers propose the concept-based interpretation to interpret network features as semantic concepts [6]. In [7], network features are interpreted as visual concepts by evaluating the overlap between a concept annotation and the saliency region of a network feature. However, it is a one-to-one alignment and does not consider the one-to-many situation. In comparison, Zhang et al. [25] develop a method to learn an explanatory graph that automatically disentangles object parts from each unit without any part annotations. Zhou et al. [8] propose to represent the target image prediction as a linear combination of different concepts. In [8], the prediction and candidate concepts need to be represented utilizing network units from the same layer. However, studies show that different semantic-level concepts match network units at different layers. Therefore, for concept representations in the last convolutional layer, low semantic-level concepts may not be as active as high semantic-level concepts. Meanwhile, these methods do not explain the layer structure for the network decision-making process.

Here, we propose the CHAIN interpretation to explain the network decision-making process. It interprets hierarchical

Ability \ Method	CAM	Grad-CAM	TCAV	Network Dissection	FLOWIN	Our CHAIN
Interpret a specific instance	✓	✓	✓		✓	✓
Interpret internal units with semantic concepts			✓	✓		✓
Interpret a inference relationship between two layers					✓	✓
Interpret a CNN bottom-up decision-making process as a hierarchical concept-Inference in Neuroscience						✓

Fig. 2. Scope of interpretation for different methods (CAM [20], GradCAM [21], TCAV [6], Network Dissection [23] and FLOWIN [24], and CHAIN).

network learning as a decision-making process of visual concepts from shallow to deep layers. Fig. 2 shows the scope of interpretation for different methods. Existing methods interpret networks from different angles. In comparison, the proposed CHAIN aims to provide a more comprehensive interpretation.

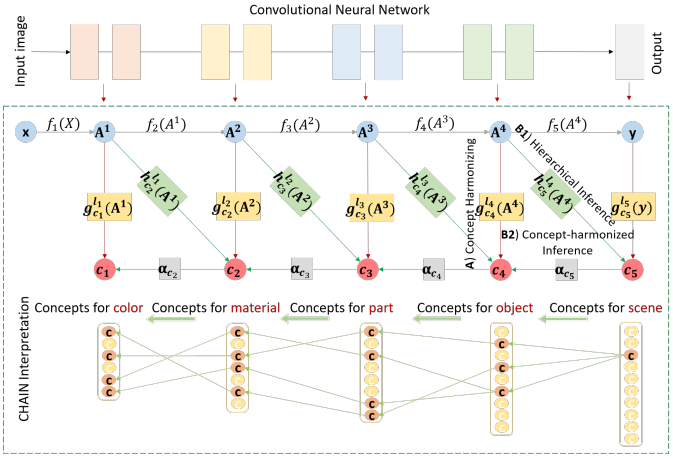


Fig. 3. The workflow of CHAIN interpretation. A) Firstly, the concept harmonizing model learns global feature representations (functions $G = \{g^l_{c'}\}_{l,c'}$, yellow rectangles) for concepts (c', red dots) from high to low semantic levels (the overall network $F_L : \mathbf{X} \to \mathbf{y}$, an internal layer $f_l : \mathbf{A}^{l-1} \to \mathbf{A}^l$, blue dots). For a given input, CHAIN explains its network decision-making process as a hierarchical concept inference. From a deeper layer to a shallower one, the inference process has two steps: B1) In the hierarchical inference model, concept c in a deeper layer (the $(l+1)$-th layer) can be decomposed into critical features $\mathbf{A}^l$ in a shallower layer (the l-th layer) by optimizing the function $h^l_c(\mathbf{A}^l)$ (green rectangles). B2) Accordingly, in the concept-harmonized inference model, we can represent the deeper-layer concept c as a sparse representation of shallower-layer concepts weighted by $\boldsymbol{\alpha}_c$ (gray rectangles). (Bottom: CHAIN interpretation) Based on the inference process, we firstly deduce the target scene concept backward into object ones. Each critical object concept is continuously deduced into part ones and so on. Therefore, a network decision-making process can be interpreted by presenting its concept inference from scene to color levels.

III. Methodology

Our work interprets networks in a hierarchical inference way analogous to bottom-up hierarchy in Neuroscience.

To accomplish it, in Section III-A, we first build a concept harmonizing model to explore the relationships between CNN units and semantic concepts hierarchically, as shown in Phase A of Fig. 3. This model is trained on the Broden dataset [7] and works as the pre-stage of our CHAIN interpretation framework. Based on the Broden dataset, the concept harmonizing model can learn global and reliable concept representations in the feature space. In this way, we build a link between the semantic-stratified structure for visual concepts and the layer-stratified structure in networks.

Subsequently, in section III-B, we propose the concept-harmonized hierarchical inference (CHAIN) to interpret a specific network decision-making process, as shown in Phase B of Fig. 3. We first deduce the target scene concept in the output layer backward into object ones in the shallower layer. Each critical object concept is continuously deduced into part concepts and so on. At the instance level, a network decision-making process can be interpreted as the concept-harmonized hierarchical inference from scene to color semantic levels.

Specifically, from a deeper layer to a shallower layer, the inference process has two steps. In the hierarchical inference model (section III-B1), the target concept in a deeper layer is decomposed into critical network features in a shallower layer, as shown in Step B1 of Fig. 3. In the concept-harmonized inference model (section III-B2), we represent the deeper-layer concept as a sparse representation of shallower-layer concepts, as shown in Step B2 of Fig. 3.

For the interpretation of network decisions for images from the same class, we build the class-level CHAIN interpretation by selecting common concepts among instance-level CHAIN interpretations for different images in the same category. In CHAIN, a network decision-making process is interpreted as the concept-harmonized hierarchical inference from high to low semantic levels.

A. Concept Harmonizing in Feature Space

In recent works [23], [6], researchers explain the meaning of CNN features (units) as visual semantic concepts. By quantifying the alignment of unit-concept pairs, they explore the expression ability of CNN features in different layers for visual concepts in different semantic levels. Experiments demonstrate that for a network trained on a specific image dataset, the latent representation of hidden units for each concept is on a regular basis rather than a random basis [23], [6]. Experiments on different CNNs also show that CNN features learn visual concepts with higher semantic complexity with deepening CNN layers. It means that the concept representation in feature space has a regular hierarchical pattern by using a semantic concept dataset [23], [6]. Therefore, it is reasonable to set the correspondence between CNN layers and concepts in the concept harmonizing model.

In this section, we learn the concept representation in feature space. Visual concepts from low to high semantic levels are

aligned with network units from shallow to deep layers, as shown in Phase A of Fig. 3.

Visual concepts. Based on the Broden dataset [7], we utilize visual concepts in five semantic levels, i.e., color, material, part, object, and scene from low to high levels. Network features in five layers are selected to align with concepts from low to high semantic levels. The selected five-level visual concepts are all labeled at the pixel level except scene-level ones.

Training dataset for concept harmonizing model. For the c'-th concept in the l-th layer, we design a concept harmonizing function $g^l_{c'}(\cdot)$ mapping from feature space $\mathcal{A}^l$ to concept-score space $\mathcal{Y}_{c'}$. For an image $\mathbf{X}$, $y_{c'}(\mathbf{X})$ is the binary concept label where each value represents the absence (0) or presence (1) of the c'-th concept. During the training of $g^l_{c'}(\cdot)$, an image pixel $\mathbf{X}_{(i,j)}$ containing the target concept is the positive sample and its concept-score is $y_{c'}(\mathbf{X})_{(i,j)} = 1$. Other image pixels are negative samples. Training samples of a target concept are fed into the network to obtain features in the corresponding layer. The concept harmonizing model $g^l_{c'}(\cdot)$ is optimized via

$$\arg\min_{g^l_{c'}} \frac{1}{N_{(i,j)}} \sum_{(i,j)}^{N_{(i,j)}} \left[g^l_{c'}(F_l(\mathbf{X})_{(i,j)}) - y_{c'}(\mathbf{X})_{(i,j)} \right]^2 \quad (1)$$

where the network function F_l maps from image space $\mathcal{X}$ to feature space $\mathcal{A}^l$.

Similar to [6], we utilize a linear model as the concept harmonizing model to align units with visual concepts. Compared with [6], we further interpret CNNs based on the semantic-stratified visual concepts obtained by the concept harmonizing model.

Harmonizing weights of units for a visual concept. The concept harmonizing function is defined as

$$g^l_{c'}(F_l(\mathbf{X})_{(i,j)}) = \mathbf{v}^l_{c'} F_l(\mathbf{X})_{(i,j)} \quad (2)$$

where $F_l(\mathbf{X})_{(i,j)} \in \mathbb{R}^{I\times 1}$ is feature in the l-th layer and I is the number of units. Units in the l-th layer are harmonized with the c'-th concept by the harmonizing weight $\mathbf{v}^l_{c'} = [v^{l,1}_{c'}\ v^{l,2}_{c'} \cdots v^{l,i}_{c'} \cdots v^{l,I}_{c'}]$. $v^{l,i}_{c'}$ is the harmonizing weight of the i-th unit.

Similarly, we calculate concept harmonizing functions for concepts in different layers. We can learn optimal concept harmonizing functions $G = \{g^l_{c'}\}_{l,c'}$ for visual concepts at different semantic levels.

B. Concept-harmonized Hierarchical Inference (CHAIN)

On top of the optimal concept harmonizing functions, we propose the CHAIN framework to interpret the network decision-making process of an input image.

For an input image, CHAIN explains its network decision-making process as a concept-harmonized hierarchical inference from shallow to deep layers. As shown in Step B1 and B2 of Fig. 3, CHAIN includes two steps, the hierarchical inference model and the concept-harmonized inference model.

Firstly, a concept in a deeper layer is decomposed into critical features in a shallower layer. Secondly, we interpret these critical shallower-layer features by concepts harmonized in the shallower layer. Accordingly, we can represent a concept in a deeper layer as a sparse representation of concepts in a shallower layer.

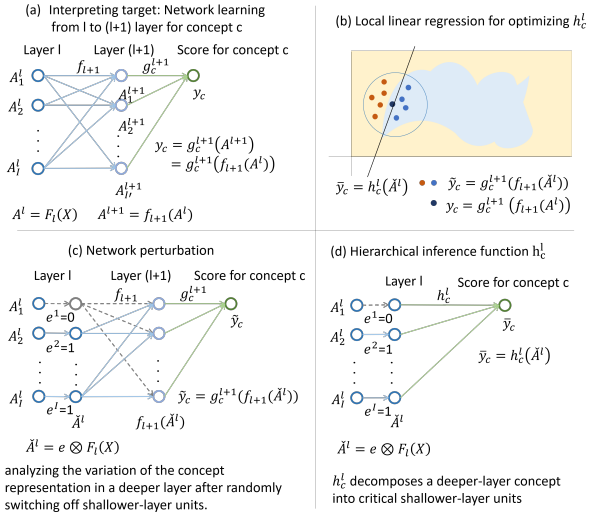


Fig. 4. Illustration of the hierarchical inference model.

*1) **Hierarchical Inference**:* The hierarchical inference function is designed to approximate the network inference process from a shallow layer to a deep layer for the target concept. It is optimized by a local linear regression problem based on the network perturbation, as shown in Fig. 4.

The network inference function from a shallow layer to a deep layer for the target concept, as shown in Fig. 4(a), is highly nonlinear. The idea of local linear regression is that if the network inference function has sufficient smoothness, then the model will look linear in local regions, as shown in Fig. 4(b). It is based on locally fitting a line rather than a constant. The network function F_l can be expressed as $F_l(\mathbf{X}) = f_l(f_{l-1}(\cdots(f_1(\mathbf{X}))\cdots))$, where f_l is the function in the l-th layer.

Inspired by the network perturbation [22], the hierarchical inference model is learned by analyzing the variation of the concept representation in a deeper layer after randomly switching off shallower-layer units, as shown in Fig. 4(c). The unit-wise perturbed vector is denoted as a binary vector $\mathbf{e}$. The l-th layer is perturbed by $\mathbf{e} = [e^1\ e^2 \cdots e^i \cdots e^I]$ in which each item controls the state of the corresponding unit in the l-th layer. The i-th unit is turned off if e^i is zero.

By selecting elements of $\mathbf{e}$ as 1 uniformly at random, we collect samples in local region to approximate the network inference function for a specific instance in the local linear regression method, as shown in Fig. 4(b). The underlying principle is that the concept in a deeper layer would drop dramatically if its important shallower-layer units are blocked. For the c-th concept in the $(l+1)$-th layer, its hierarchical inference function $h^l_c(\cdot)$ maps from feature space $\mathcal{A}^l$ in the l-th layer to concept-harmonizing space $\mathcal{G}^{l+1}_c$ for the c-th concept in the $(l+1)$-th layer, as shown in Fig. 4(d).

To infer the c-th concept in the $(l+1)$-th layer from features in the l-th layer for a given image $\mathbf{X}$, we formulate the

hierarchical inference model as

$$\arg\min_{h_c^l} \mathbb{E}_{\mathbf{e}} \left[\mathcal{L}(h_c^l(\breve{\mathbf{A}}^l), g_c^{l+1}(f_{l+1}(\breve{\mathbf{A}}^l))) \right] + \psi(h_c^l) \tag{3}$$

where a unit-wise perturbed feature in l-th layer is denoted as $\breve{\mathbf{A}}^l = \mathbf{e} \otimes F_l(\mathbf{X})$ and $\otimes$ is a broadcasted outer product. During the optimization of Eq. (3), $\mathbf{e}$ is generated by selecting its elements as 1 uniformly at random. We obtain $\{\mathbf{e}_n\}_{n=1}^N$ by generating N perturbed vector samples.

Inference weights of shallower-layer units to a concept in a deeper layer. For the c-th concept in the $(l+1)$-th layer, its hierarchical inference function is $h_c^l(\breve{\mathbf{A}}^l) = \mathbf{w}_c^l * \breve{\mathbf{A}}^l$, where $*$ is a 1×1 convolutional operation by the filter $\mathbf{w}_c^l$. The inference weight vector is $\mathbf{w}_c^l = [w_c^{l,1}\ w_c^{l,2} \cdots w_c^{l,i} \cdots w_c^{l,I}]$, where $w_c^{l,i}$ denotes the importance of the i-th unit in the l-th layer to the c-th concept in the $(l+1)$-th layer. In the hierarchical inference function, feature $\breve{\mathbf{A}}^l$ in the l-th layer is resized to the spatial dimension of feature $f_{l+1}(\breve{\mathbf{A}}^l)$ in the $(l+1)$-th layer.

In Eq. (3), the first term is the loss function

$$\begin{aligned} &\mathcal{L}(h_c^l(\breve{\mathbf{A}}^l), g_c^{l+1}(f_{l+1}(\breve{\mathbf{A}}^l))) \\ =&\frac{1}{2} r(\mathbf{e}) \| h_c^l(\mathbf{e} \otimes F_l(\mathbf{X})) - g_c^{l+1}(f_{l+1}(\mathbf{e} \otimes F_l(\mathbf{X})))\|_F^2 \end{aligned} \tag{4}$$

where $r(\mathbf{e})$ is the proximity measure between a binary vector $\mathbf{e}$ and the all-one vector $\mathbf{1}$. Specifically, it is

$$r(\mathbf{e}) = \exp(-\frac{1}{\sigma^2}\|\mathbf{e} - \mathbf{1}\|_2^2) \tag{5}$$

The second term of Eq. (3) is the sparse regularization term $\psi(h_c^l) = \lambda \|\mathbf{w}_c^l\|_1$. To make the hierarchical inference model be simple enough to be interpretable, the sparsity of inference weights measures the complexity of the interpretation model. λ is the regularization parameter.

To optimize Eq. (3), we design a hierarchical inference algorithm by adopting the alternating iteration rule, which is described in section I of the supplementary material.

*2) **Concept-harmonized Inference**:* A deeper-layer concept can be represented as a sparse combination of concepts in a shallower layer.

To infer the c-th concept in the $(l+1)$-th layer from concepts in the l-th layer for a given image $\mathbf{X}$, the concept-harmonized inference model is optimized by

$$\arg\min_{\alpha_c} \quad \|\mathbf{V}^l \boldsymbol{\alpha}_c - (\mathbf{w}_c^l)^T\|_2^2 + \|\boldsymbol{\alpha}_c\|_0 \tag{6}$$

where $\mathbf{V}^l = [\mathbf{v}_1^l; \mathbf{v}_2^l; \cdots; \mathbf{v}_{c'}^l; \cdots; \mathbf{v}_{C'}^l]^T$ is the harmonizing weights of a concept set in the l-th layer.

Contribution weights of concepts in a shallower layer to a concept in a deeper layer. The contribution weight vector of concepts in the l-th layer to the c-th concept in the $(l+1)$-th layer is defined as $\boldsymbol{\alpha}_c = [\alpha_c^1\ \alpha_c^2 \cdots \alpha_c^{c'} \cdots \alpha_c^{C'}]^T$. Specifically, $\alpha_c^{c'}$ is the contribution of the c'-th concept in the l-th layer to the c-th concept in the $(l+1)$-th layer.

Based on the contribution weights, the higher level concept is inferred from lower level concepts. Here, we use Orthogonal Matching Pursuit (OMP) algorithm [26] to obtain $\boldsymbol{\alpha}_c$. The selected inference concept set is denoted as $\mathcal{C}_c^{l+1}$ which contains critical concepts in the l-th layer for the concept c in the $(l+1)$-th layer. The stop condition in OMP is $\| \sum_{c' \in \mathcal{C}_c^{l+1}} \mathbf{V}_{c'}^l \boldsymbol{\alpha}_c^{c'} - (\mathbf{w}_c^l)^T\|_2^2 \leq 0.1 \|\mathbf{w}_c^l\|_2^2$.

Similarly, the concept-harmonized inference for other critical concepts in the $(l+1)$-th layer are also optimized. In the process, we denote $\mathcal{S}^l = \bigcup_{c \in \mathcal{S}^{l+1}} \mathcal{C}_c^{l+1}$ as the concept set being inferred in the l-th layer. Subsequently, the optimization of the concept-harmonized inference model is conducted to infer concepts in the l-th layer.

Finally, a network decision-making process is interpreted as the hierarchical inference of concepts. For a network decision, CHAIN interpretation learns its optimal inference concept set $\mathcal{C} = \{\mathcal{C}_c^{l+1}\}_{l,c}$. Meanwhile, we can quantitatively analyze the CHAIN interpretation for a network decision. The pseudocode of the CHAIN interpretation is shown in Algorithm 1.

For the inference of part-level concepts, we only care about the most important material-level concept. It is also applied to the inference from the material-level concept to the color-level concept. Therefore, in the concept-harmonized inference model Eq. (6), the sparsity of $\boldsymbol{\alpha}_c$ is set as 1 to select the most critical shallower-layer concept for the target deeper-layer concept. This also can be achieved by utilizing the derivation of concept directional-derivative, described in section II of the supplementary material.

IV. Experiments

Experiments are conducted to analyze qualitatively and quantitatively the performance of the proposed interpretation model. Experimental setting and quantitative analysis metrics are described in section IV-A. In section IV-B, we provide the instance-level CHAIN interpretation. In section IV-C, the CHAIN interpretation is applied to explain network predictions at the class level. In section IV-D, we study the influence of a shallower-layer concept on a deeper-layer concept in CHAIN interpretation by manipulating the shallower-layer concept in the feature space. In section IV-E, networks can be further understood by the intra-class and inter-class inference distance of concepts in CHAIN.

A. Experimental setting

*1) **ResNets on the Places365 scene classification dataset**:* CHAIN interpretation is applied to explain ResNet-18 and ResNet-50 [1] that are pre-trained on the Places365 scene classification dataset [27]. In the concept harmonizing model, we use five layers (i.e., the output, layer4, layer3, layer2, and layer1) to align with five semantic level concepts (scene, object, part, material, and color).

*2) **Concept harmonizing dataset**:* In the concept harmonizing model, Broden dataset is utilized as the concept dataset [7]. Broden dataset contains a hierarchical level of labeled visual-concept samples. Concept annotations are in pixel level except for the scene annotation which is in image level. The five semantic level concepts are from the ADE20K [28], Pascal-Part [29], and OpenSurfaces datasets [30].

*3) **Influence score**:* The CHAIN interpretation distills critical concepts in a shallower layer for a concept in a deeper layer. Therefore, we manipulate network features by enhancing or weakening a shallower-layer concept to examine its influence on a deeper-layer concept.

Algorithm 1: Pseudocode of the CHAIN Interpretation

Input: a network $F_L = f_L(f_{L-1}(\cdots(f_1(\mathbf{X}))\cdots))$ being interpreted; an image $\mathbf{X}$ being interpreted; the concept harmonizing functions $\{g_c^{l+1}\}_{l,c}$ for visual concepts;
Output: for a network decision, the optimal inference concept set $\mathcal{C} = \{\mathcal{C}_c^{l+1}\}_{l,c}$;
Initialization: the being-inferred concept set $\mathcal{S}^L$ in the L-th layer and $\mathcal{S}^l = \emptyset (l < L)$; the inference concept set $\mathcal{C}_c^{l+1} = \emptyset$ for the concept c in the $(l+1)$-th layer; $l = L-1$;
repeat
  for each concept $c \in \mathcal{S}^{l+1}$ **do**
    $h_c^l \leftarrow \arg\min_{h_c^l} \mathbb{E}_{\mathbf{e}}\left[\mathcal{L}(h_c^l(\breve{\mathbf{A}}^l), g_c^{l+1}(f_{l+1}(\breve{\mathbf{A}}^l)))\right] + \psi(h_c^l)$. (in Eq. (3));
    $\boldsymbol{\alpha}_c \leftarrow \arg\min_{\alpha_c} \|\mathbf{V}^l\boldsymbol{\alpha}_c - (\mathbf{w}_c^l)^T\|_2^2 + \|\boldsymbol{\alpha}_c\|_0$. (in Eq. (6));
    $\mathcal{S}^l = \mathcal{S}^l \bigcup \mathcal{C}_c^{l+1}$
    return the optimal inference concept set $\mathcal{C}_c^{l+1}$ for the concept c in the $(l+1)$-th layer
  end for
  return the being-inferred concept set $\mathcal{S}^l$ in the l-th layer
  update iteration: $l \leftarrow l-1$
until *the CHAIN interpretation is optimized from the deepest to shallowest layer*;
return the optimal inference concept set $\mathcal{C} = \{\mathcal{C}_c^{l+1}\}_{l,c}$ for a network decision being interpreted.

Fig. 5. Instance-level CHAIN interpretation for images from four classes (i.e. Farm, Orchard, House, and Pasture).

The influence of enhancing concept c' in the l-th layer on concept c in the $(l+1)$-th layer for pixel (i,j) is

$$\begin{aligned}&\triangle(concept_c^{l+1}; concept_{c'}^{l})^{+}\\ &\stackrel{\text{def}}{=} g_c^{l+1}(f_{l+1}(\mathbf{A}^l + \mathbf{v}_{c'}^l * \boldsymbol{\delta})_{(i,j)}) - g_c^{l+1}(f_{l+1}(\mathbf{A}^l)_{(i,j)})\end{aligned} \tag{7}$$

where $\mathbf{A}^l \in \mathbb{R}^{W\times H\times I}$ is network features and I is the number of units. For the c'-th concept in the l-th layer, its harmonizing weight is $\mathbf{v}_{c'}^l = [v_{c'}^{l,1}\ v_{c'}^{l,2}\cdots v_{c'}^{l,i}\cdots v_{c'}^{l,I}]$, which can also be regarded as a concept direction in the feature space of the l-th layer. Unit influence rate is denoted as $\boldsymbol{\delta} \in \mathbb{R}^{W\times H\times I}$.

The influence of weakening concept c' in the l-th layer on concept c in the $(l+1)$-th layer for pixel (i,j) is

$$\begin{aligned}&\triangle(concept_c^{l+1}; concept_{c'}^{l})^{-}\\ &\stackrel{\text{def}}{=} g_c^{l+1}(f_{l+1}(\mathbf{A}^l - \mathbf{v}_{c'}^l * \boldsymbol{\delta})_{(i,j)}) - g_c^{l+1}(f_{l+1}(\mathbf{A}^l)_{(i,j)})\end{aligned} \tag{8}$$

For a deeper-layer $concept_c^{l+1}$, its important shallower-layer $concept_{c'}^{l}$ is assumed to have positive value for $\triangle(concept_c^{l+1}; concept_{c'}^{l})^{+}$, and negative value for $\triangle(concept_c^{l+1}; concept_{c'}^{l})^{-}$. In contrast, its unimportant shallower-layer $concept_{c'}^{l}$ is assumed to have non-positive value for $\triangle(concept_c^{l+1}; concept_{c'}^{l})^{+}$, and non-negative value for $\triangle(concept_c^{l+1}; concept_{c'}^{l})^{-}$. In section IV-D, we evaluate CHAIN interpretation by the influence score.

*4) **Inference distance**:* The hierarchical inference model obtains critical features in a shallower layer for a concept in a deeper layer. Therefore, we define the inference distance to quantitatively analyze the inference of concepts on intra-class and inter-class network decisions in section IV-E.

Inference distance of concept c for image set s is defined as the average Euclidean distance of inference weights for images in image set s, which is denoted as

$$\begin{aligned} &dist(concept_c, set_s) \\ &\stackrel{\text{def}}{=} \frac{1}{N_i} \sum_i \| \mathbf{w}_{set_{s,i}}^{concept_c} - \bar{\mathbf{w}}_{set_s}^{concept_c} \|_2 \end{aligned} \tag{9}$$

where the center of inference distance $\bar{\mathbf{w}}_{set_s}^{concept_c}$ is

$$\bar{\mathbf{w}}_{set_s}^{concept_c} \stackrel{\text{def}}{=} \frac{1}{N_i} \sum_i \mathbf{w}_{set_{s,i}}^{concept_c} \tag{10}$$

Here, $\mathbf{w}_{set_{s,i}}^{concept_c}$ is the inference weight vector of shallower-layer units to the c-th concept in the deeper layer for the i-th image in image set s. N_i is the number of images in set s.

Inference distance between concept c for image set s and concept c' for image set s' is obtained by

$$\begin{aligned} &dist(concept_c, set_s; concept_{c'}, set_{s'}) \\ &\stackrel{\text{def}}{=} \| \bar{\mathbf{w}}_{set_s}^{concept_c} - \bar{\mathbf{w}}_{set_{s'}}^{concept_{c'}} \|_2 \end{aligned} \tag{11}$$

B. Instance-level CHAIN interpretation

For illustration, we use CHAIN to explain network predictions of four randomly selected images from different classes. ResNet-18 accurately predicts their scene classes. More results can be found in section III of the supplementary material.

Fig. 5 shows the instance-level CHAIN interpretation for images from four classes (i.e., farm, orchard, house, and pasture). For a specific network decision, CHAIN shows the hierarchical concept inference to explain the network decision-making process. For example, in the bottom right of Fig. 5, CHAIN interpretation shows that the pasture scene prediction is inferred from the horse, grass, and fence concepts. Moreover, the object-level horse concept is inferred from the part-level torso concept that is deduced from the material-level skin concept. Finally, the horse concept can be hierarchically inferred from the orange concept.

Meanwhile, CHAIN provides visualization for concepts at each semantic level. The visualization of concepts in the CHAIN interpretation can localize the corresponding image regions, which validates the effectiveness of concept representations and interprets the network feature learning for visual concepts.

To demonstrate that CHAIN is applicable to different classification tasks, we also conduct CHAIN interpretation experiments on object classification based on the PASCAL VOC dataset. The results are shown in section IV of the supplementary material.

Meanwhile, in section V of the supplementary material, we qualitatively compare the proposed CHAIN interpretation method with GradCAM [21], and FLOWIN [24].

C. Class-level CHAIN interpretation

In this experiment, we give the intra-class and inter-class CHAIN interpretation. Specifically, in section IV-C1, house images with different surroundings are selected for the intra-class analysis. In section IV-C2, three scene classes are chosen for the inter-class CHAIN interpretation for different networks.

*1) **Intra-class CHAIN interpretation**:* For the house scene class, the class-level interpretation is analyzed by house images with three typical surroundings (i.e., curb, hedge, and swimming pool). We randomly choose twenty house images for each surrounding as the corresponding image set to explain their ResNet-18 predictions.

Fig. 6 displays the CHAIN interpretation for the house class. The top row shows house image examples with the three surroundings. Sunburst charts for images with three surroundings are in the middle. The innermost circle of a sunburst chart shows object-level concepts that are crucial to the house-scene prediction. The outer circle shows the lower-level concepts that are important to concepts in the inner circle. The proportion of each visual concept in the innermost circle indicates its contribution to the network scene prediction. Similarly, the contribution of a part concept to its object concept is shown by its proportion in the outer circle. In each sunburst chart, concepts at each level are sorted in descending order of the contribution to their corresponding high-level concept. In the three sunburst charts, the object-level house concept has the most significant contribution to the house scene prediction. Meanwhile, hedge and swimming pool concepts own the second-largest contribution to the network prediction of house image sets with hedge and swimming pool, respectively. The curb concept also contributes a lot to the house scene prediction of the house image set with the curb. Therefore, CHAIN can explain the intra-class network predictions by presenting common and unique concepts for images within a class.

At the bottom of Fig. 6, we present the CHAIN interpretation diagram for the house class in which the network output is inferred from the scene to the color level. The fraction enclosed by the purple dashed line denotes the house-related concepts shared by the three types of images. Fractions in yellow, green, and red dashed rectangles are the concepts for different surroundings. The diagram is consistent with the observation of the three sunburst charts. For example, the hedge concept is inferred from the plant, then to the foliage, and finally to the green color. Meanwhile, the hedge concept is shared in the interpretation of house image sets with hedge and swimming pool. At the instance level, it is observed that many images in the image set with a swimming pool involve a hedge. Therefore, the CHAIN interpretation diagram is reasonable.

*2) **Inter-class CHAIN interpretation for different networks**:* This section analyzes inter-class CHAIN interpretation of three scene classes (Living room, Farm, and Pasture) for ResNet18 and ResNet50. We use all images in the Places365 validation dataset for each scene class to analyze the inter-class CHAIN interpretation.

Results are shown in Fig. 7 for ResNet18 (top) and ResNet50 (bottom). In each bar chart, concepts with top contribution

Fig. 6. Class-level CHAIN interpretation for the house class. (Top) House images with three types of surroundings, i.e. curb, hedge, and swimming pool. We randomly show sixteen images from each image set for illustration. (Middle) Sunburst charts of images with different surroundings. The innermost ring shows concepts that are crucial to the house-class prediction. The expansion of a concept section to its outer ring shows the lower-level concepts that are important to the concept itself. (Bottom) CHAIN interpretation diagram for the house-class images. The fraction enclosed by the purple dashed line denotes the house-related concepts shared by the three types of images. Fractions in yellow, green, and red dashed rectangles are the concepts for different surroundings.

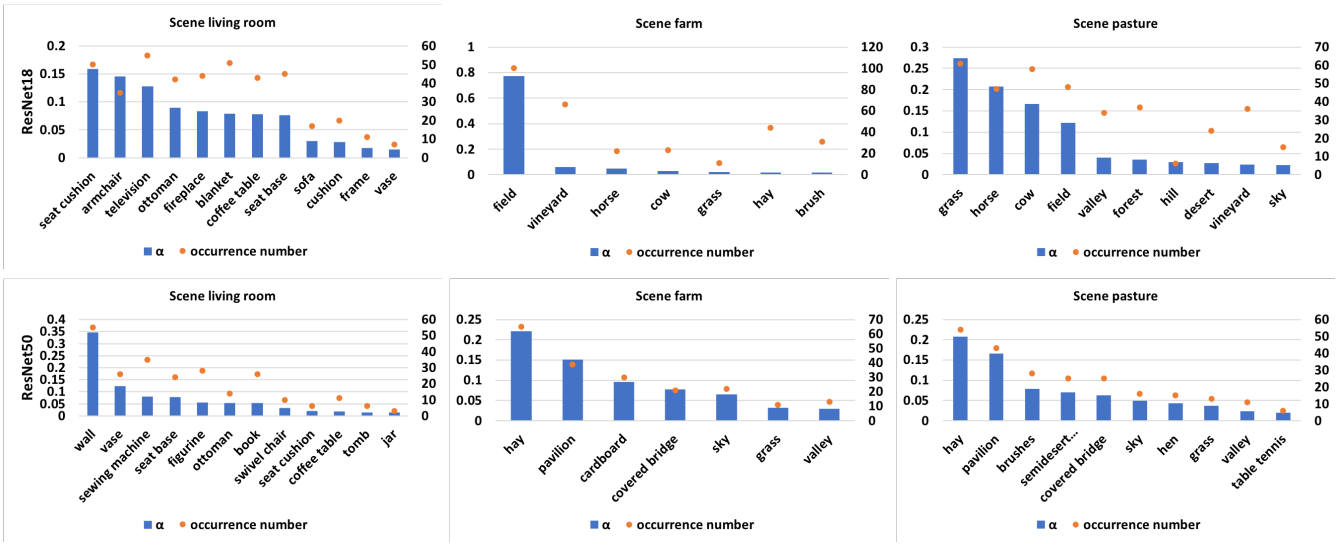


Fig. 7. Comparison of CHAIN interpretation in different networks. Contribution weights α and occurrence numbers of object-level concepts to scene-level concepts are compared.

TABLE I
INFLUENCE OF OBJECT-LEVEL CONCEPTS ON THE SCENE-LEVEL HOUSE CONCEPT

Influence of object concepts in CHAIN on the scene house								
	house	roof	shutter	curb	chimney	hedge	palm	swimming pool
$\triangle^+$	0.6209	0.4898	0.3390	0.1882	0.2991	0.2666	0.1667	0.0590
$\triangle^-$	-0.6209	-0.4898	-0.3390	-0.1882	-0.2991	-0.2666	-0.1667	-0.0590
Influence of unrelated object concepts on the scene house								
	ice	basket	airplane	book	dog	boat	bicycle	bed
$\triangle^+$	-0.2502	-0.2231	-0.1953	-0.1763	-0.1653	-0.1562	-0.1520	-0.1460
$\triangle^-$	0.2502	0.2231	0.1953	0.1763	0.1653	0.1562	0.1520	0.1460

TABLE II
INFLUENCE OF PART-LEVEL CONCEPTS ON THE OBJECT-LEVEL HOUSE CONCEPT

Influence of part concepts in CHAIN on the object house				
	dormer	roof	porch	bush
$\triangle^+$	5.2879	4.9468	3.3671	3.7148
$\triangle^-$	-7.6448	-6.4966	-4.8573	-4.9578
Influence of unrelated part concepts on the object house				
	taillight	crosswalk	bumper	sash
$\triangle^+$	-1.9990	-1.8485	-1.7969	-1.6713
$\triangle^-$	0.4857	0.7345	0.4081	0.5885

weights α are shown for comparison. Meanwhile, the occurrence numbers of concepts are presented as yellow dots. Contribution weights α and occurrence numbers in ResNet50 are more harmonized compared with those in ResNet18.

In Fig. 7, CHAIN provides reasonable inferences for different scene classes and networks. Besides, shallow-layer concepts that are important to the specific deep-layer one vary in different networks. As shown in the first column of Fig. 7, the important shallow-layer concepts for the living room are different between ResNet18 and ResNet50. They share a portion of important concepts for the scene living room, such as seat cushions, ottoman, coffee table, and vases. We observe similar results in CHAIN interpretation for other scenes. It is reasonable because the decision-making process of different networks may not be identical. By comparing CHAIN interpretation in different networks, we have a better understanding of the internal mechanisms of networks.

D. Influence of a shallower-layer concept on a deeper-layer concept

In this experiment, we analyze the influence of a shallower-layer concept on a deeper-layer one in CHAIN interpretation by manipulating concepts in the feature space, as described in section IV-A3. We enhance or weaken a shallower-layer concept in feature space to examine its influence on a deeper-layer concept. House images with three types of surroundings are used to evaluate CHAIN interpretation on ResNet-18.

*1) **Influence of a shallower-layer concept on a deeper-layer concept at the class level**:* Tab. I shows the influence of object-level concepts in a shallower layer on the scene-level house concept in a deeper layer. The top is the influence of object-level concepts in CHAIN, while the bottom is the influence of unrelated object-level concepts randomly selected from the complementary set of object-level concepts in CHAIN interpretation. Tab. I reveals that the manipulation of object-

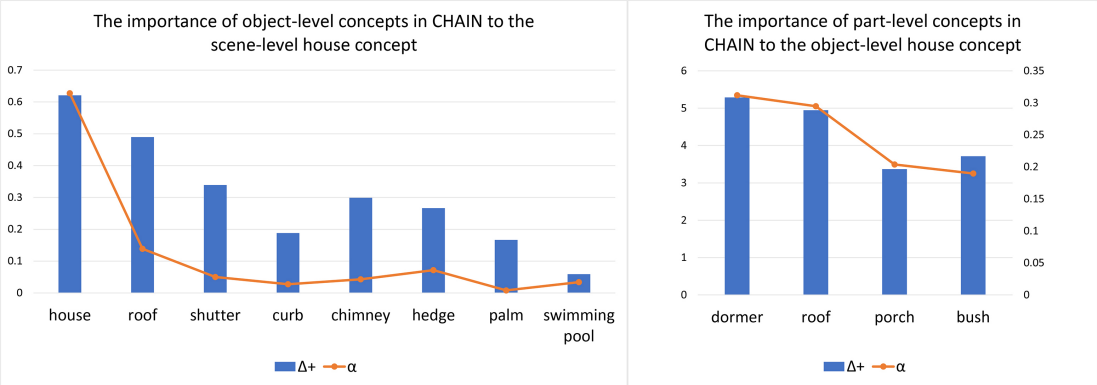


Fig. 8. The importance of shallower-layer concepts to a deeper-layer concept.

TABLE III
INFLUENCE OF OBJECT-LEVEL CONCEPTS ON THE SCENE-LEVEL HOUSE CONCEPT FOR INTRA-CLASS IMAGES

Influence of object concepts in CHAIN on the scene house								
	house	roof	shutter	curb	chimney	hedge	palm	swimming pool
	house images with curb							
$\triangle^+$	0.6960	0.5822	0.4080	**0.2784**	**0.3601**	0.2489	0.1321	-0.0118
$\triangle^-$	-0.6960	-0.5822	-0.4080	**-0.2784**	**-0.3601**	-0.2489	-0.1321	0.0118
	house images with hedge							
$\triangle^+$	0.6725	0.5091	0.3466	0.1844	0.3102	**0.3643**	0.1784	0.0138
$\triangle^-$	-0.6725	-0.5091	-0.3466	-0.1844	-0.3102	**-0.3643**	-0.1784	-0.0138
	house images with swimming pool							
$\triangle^+$	0.4943	0.3780	0.2624	0.1018	0.2269	0.1867	**0.1895**	**0.1749**
$\triangle^-$	-0.4943	-0.3780	-0.2624	-0.1018	-0.2269	-0.1867	**-0.1895**	**-0.1749**

Influence of unrelated object concepts on the scene house								
	ice	basket	airplane	book	dog	boat	bicycle	bed
	house images with curb							
$\triangle^+$	-0.3078	-0.2789	-0.2164	-0.1791	-0.1856	-0.1930	-0.1740	-0.1587
$\triangle^-$	0.3078	0.2789	0.2164	0.1791	0.1856	0.1930	0.1740	0.1587
	house images with hedge							
$\triangle^+$	-0.2584	-0.2493	-0.2145	-0.1995	-0.1856	-0.2176	-0.1616	-0.1761
$\triangle^-$	0.2584	0.2493	0.2145	0.1995	0.1856	0.2176	0.1616	0.1761
	house images with swimming pool							
$\triangle^+$	-0.1843	-0.1411	-0.1551	-0.1504	-0.1248	-0.0579	-0.1205	-0.1033
$\triangle^-$	0.1843	0.1411	0.1551	0.1504	0.1248	0.0579	0.1205	0.1033

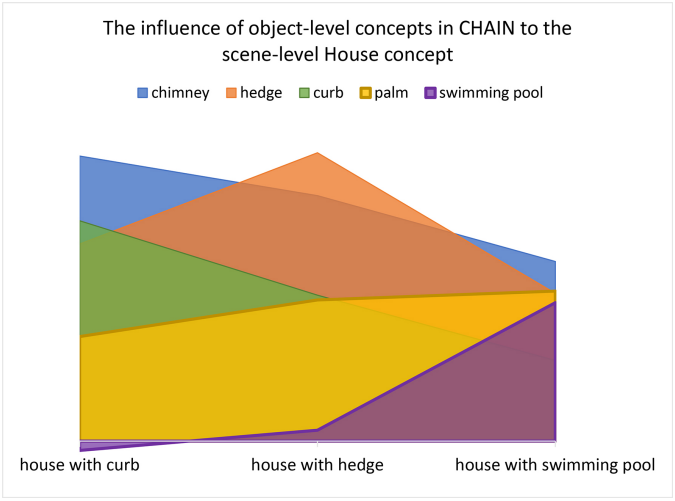
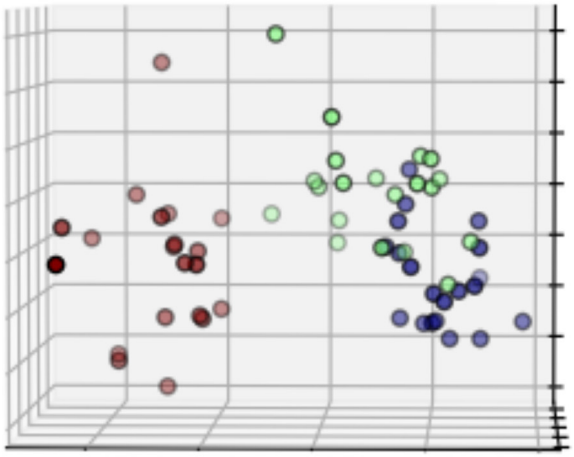


Fig. 9. The influence of enhancing object-level concepts on the scene-level house concept for intra-class images.

level concepts in CHAIN has a positive correlation with the scene-level house concept. Concretely, enhancing object concepts in CHAIN leads to an increase in the scene-level house concept. In comparison, the manipulation of unrelated concepts is inversely correlated with the scene-level house concept. It validates the efficacy of object-level concepts in CHAIN to the scene-level house concept. It is noted that the scale of enhancing and weakening influences are the same owning to linear network mapping functions (fully connected layer) from layer4 (object-level layer) to the output (scene-level layer).

Tab. II shows the influence of part-level concepts on the object-level house concept. The top is the influence of part-level concepts in CHAIN, while the bottom is the influence of randomly selected unrelated part-level concepts. Tab. II validates that the manipulation of part-level concepts in CHAIN has a positive correlation with the object-level house concept. In comparison, unrelated part-level concepts are negatively correlated with the object-level house concept. The observation is in accordance with that in Tab. I.

To further validate CHAIN, we compare the contribution



Inference distance

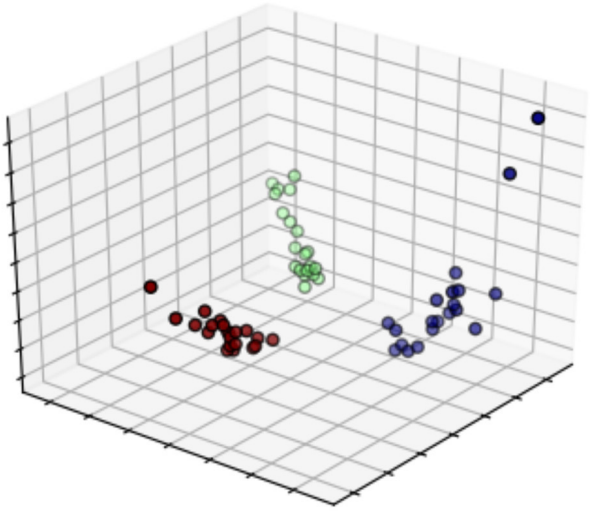
	Scene concept: House Image set: house with curb	Scene concept: House Image set: house with hedge	Scene concept: House Image set: house with swimming pool
Scene concept: House Image set: house with curb	0.0661	0.0723	**0.1344**
Scene concept: House Image set: house with hedge	0.0723	0.0804	**0.1115**
Scene concept: House Image set: house with swimming pool	**0.1344**	**0.1115**	0.0937

Scene: House Set: house with curb

Scene: House Set: house with hedge

Scene: House Set: house with swimming pool

Fig. 10. Intra-class inference distance of a scene-level concept. The left chart plots the inference weights of the house concept (scene level) for three image sets (houses with curb, hedge, and swimming pool) in the 3D-PCA space. The right table shows their inference distances.



Inference distance

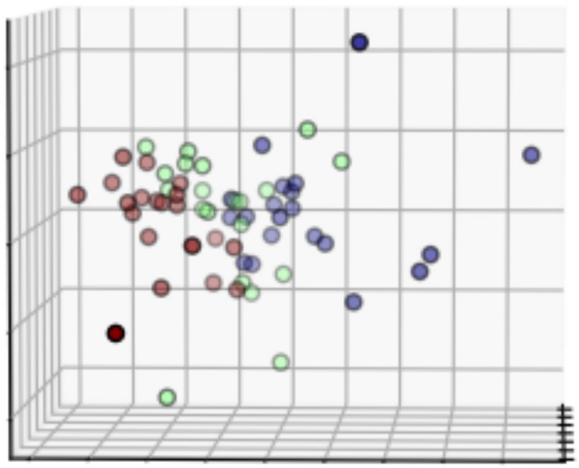
	Object concept: Curb Image set: house with curb	Object concept: Hedge Image set: house with hedge	Object concept: Swimming pool Image set: house with swimming pool
Object concept: Curb Image set: house with curb	0.2484	0.3932	0.4496
Object concept: Hedge Image set: house with hedge	0.3932	0.2065	0.4596
Object concept: Swimming pool Image set: house with swimming pool	0.4496	0.4596	0.1701

Object: Curb Set: house with curb

Object: Hedge Set: house with hedge

Object: Swimming pool Set: house with swimming pool

Fig. 11. Intra-class inference distance of object-level concepts. The left chart plots the inference weights of three object concepts (curb, hedge, and swimming pool) for their corresponding image sets (houses with curb, hedge, and swimming pool) in the 3D-PCA space. The right table shows their inference distances.



Inference distance

	Object concept: House Image set: house with curb	Object concept: House Image set: house with hedge	Object concept: House Image set: house with swimming pool
Object concept: House Image set: house with curb	0.2138	0.1127	0.1512
Object concept: House Image set: house with hedge	0.1127	0.2007	0.1268
Object concept: House Image set: house with swimming pool	0.1512	0.1268	0.2303

Object: House Set: house with curb

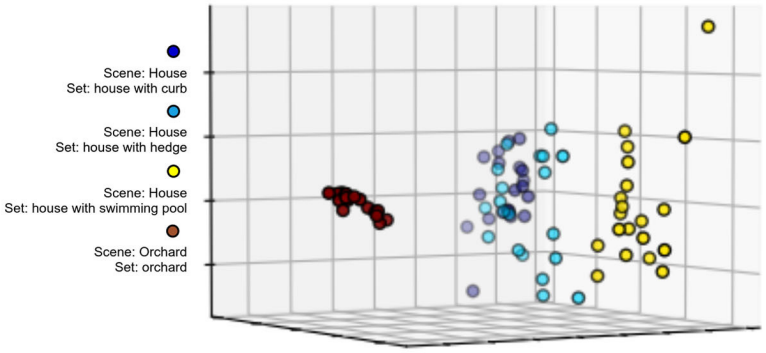
Object: House Set: house with hedge

Object: House Set: house with swimming pool

Fig. 12. Intra-class inference distance of an object-level concept. The left chart plots the inference weights of the house concept (object level) for three image sets (houses with curb, hedge, and swimming pool) in the 3D-PCA space. The right table shows their inference distances.

Inference distance

	Scene concept: House Image set: house with curb	Scene concept: House Image set: house with hedge	Scene concept: House Image set: house with swimming pool	Scene concept: Orchard Image set: orchard
Scene concept: House Image set: house with curb	0.0661	0.0723	0.1344	**0.2502**
Scene concept: House Image set: house with hedge	0.0723	0.0804	0.1115	**0.2366**
Scene concept: House Image set: house with swimming pool	0.1344	0.1115	0.0937	**0.2695**
Scene concept: Orchard Image set: orchard	**0.2502**	**0.2366**	**0.2695**	0.0822



	Object concept: House Image set: house with curb	Object concept: House Image set: house with hedge	Object concept: House Image set: house with swimming pool	Object concept: Tree Image set: orchard
Object concept: House Image set: house with curb	0.2138	0.1127	0.1512	**0.3948**
Object concept: House Image set: house with hedge	0.1127	0.2007	0.1268	**0.3621**
Object concept: House Image set: house with swimming pool	0.1512	0.1268	0.2303	**0.3928**
Object concept: Tree Image set: orchard	**0.3948**	**0.3621**	**0.3928**	0.3399

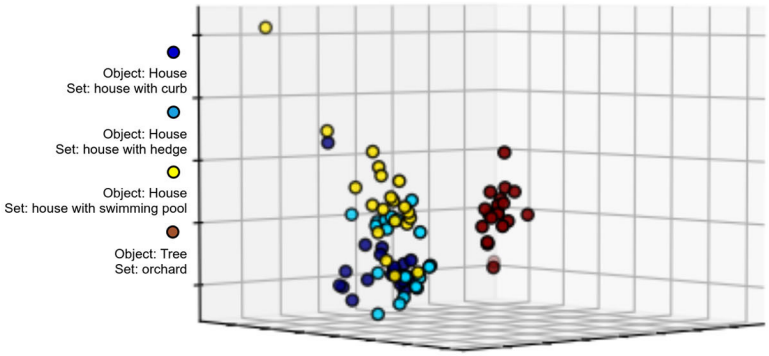


Fig. 13. Inter-class inference distance of concepts for the orchard and the house classes. The top row shows samples for two classes. The right chart (in the middle) plots the inference weights of the scene concepts for four image sets in the 3D-PCA space. The left table (in the middle) shows their inference distances of scene-level concepts. Similarly, the bottom row shows the inference distances of object-level concepts.

weight α and the influence $\triangle^+$ of a shallower-layer concept to a deeper-layer concept in Fig. 8. It shows a shallower-layer concept with a higher contribution weight generally comes with a more considerable enhancing influence. It also validates that the contribution weight α and the influence $\triangle^+$ reflect comparable results on measuring the importance of a shallower-layer concept to a deeper-layer concept. In CHAIN interpretation, a deeper-layer concept is inferred from a sparse combination of shallow-layer concepts weighted by the contribution weights α. In contrast, influence $\triangle^+$ is computed independently for each shallower-layer concept. Therefore, contribution weight α is more reasonable.

*2) **Influence of object-level concepts on the scene-level house concept for intra-class images**:* The influence of object-level concepts on the scene-level house concept for intra-class images is shown in Tab. III. Bold numbers indicate the largest influences on scene-level house concept among three house image subsets. Fig. 9 shows the influence of enhancing object-level concepts on the scene-level house concept. In Fig. 9, concepts of surroundings (curb, hedge, and swimming pool) impact the most on house images with corresponding

surroundings.

In Tab. III, it is also observed that the swimming pool has a slightly negative influence on house images with a curb. This observation is reasonable because the swimming pool barely appears when we look at a house on the street. However, in Tab. III, the curb has a weak positive influence on house images with a swimming pool. The reason is that the edge of a swimming pool paved by marble tiles looks similar to a curb. This observation indicates that CHAIN interpretations for intra-class images coincide with visual perception.

E. Inference distance of concepts in CHAIN interpretation

In this section, we study the inference process of concepts for intra-class and inter-class network decisions based on the inference distance defined in section IV-A4. Here, we apply CHAIN to explain the ResNet-18 decision making-process.

1) ***Inference distance of concepts for the intra-class interpretation****:* In the experiment, house images with three types of surroundings are selected for evaluation.

Fig. 10 depicts the intra-class CHAIN interpretation for the house class at the scene level. The left plot shows the inference weights of the scene-level house concept for three image sets (houses with curb, hedge, and swimming pool) in the 3D-PCA space. The right table shows their inference distances.

In the table of Fig. 10, bold numbers (0.1344 and 0.1115) are larger than others. For the left plot, red points can be easily separated from other points. In comparison, green and blue points have some overlap. Therefore, at the scene level, the inference of the house concept for the swimming pool set is different from those for the curb and hedge sets. In visual perception, house surroundings in the swimming pool set also have a huge difference from those for the curb and hedge sets.

Fig. 11 presents the intra-class CHAIN interpretation for the house class on object level. The left chart plots the inference weights of three object concepts (curb, hedge, and swimming pool) in their corresponding image sets (houses with curb, hedge, and swimming pool) in the 3D-PCA space. The right table shows their inference distances.

In the left plot of Fig. 11, color points can be clustered into three groups separately. In the table, diagonal entries are smaller than the others. Therefore, at the object level of the CHAIN interpretation, the inference of the three object concepts can be easily distinguished between each other.

Fig. 12 shows the intra-class CHAIN interpretation for the house class on object level. The left chart plots the inference weights of the house concept (object level) for three image sets (houses with curb, hedge, and swimming pool) in the 3D-PCA space. The right table shows their inference distances.

The left plot in Fig. 12 shows that color points mix with each other, which is also testified by the right table. It means at the object level, inferences of the house concept for three image sets are similar. For the three image sets, the house object is similar despite the difference in their surroundings. In summary, CHAIN interpretation learns the similarity and variance within a class, which is consistent with our visual understanding.

2) ***Inference distance of concepts for the inter-class interpretation****:* In this experiment, orchard and house are utilized for the study of the inter-class inference distance of concepts in CHAIN interpretation.

Fig. 13 shows the inter-class CHAIN interpretation for orchard and house classes. The top row shows samples from four image sets (orchard, houses with curb, hedge, and swimming pool) for two classes. The right chart (in the middle) plots the inference weights of scene concepts for four image sets in the 3D-PCA space. The left table (in the middle) shows their inference distances. Similarly, the bottom row shows the analysis for the object-level concepts.

In the left two tables of Fig. 13, the bold entries are higher than the others. In the right two plots, data in red color can be easily separated from other data. At the scene level, the CHAIN interpretation of the scene-level orchard concept is different from that of the scene-level house concept. Likewise, there is a large discrepancy between the interpretation of the object-level house concept and the tree concept. Therefore, the inter-class difference between orchard and house is larger than the intra-class difference, which is also aligned with the visual perception of images. Therefore, CHAIN interpretation can be used for the inter-class investigation.

V. Conclusion

In this paper, the CHAIN interpretation is proposed to explain the network decision-making process. Specifically, the CHAIN interpretation hierarchically reasons a network decision to be visual concepts from the high to the low level. Hierarchical visual concepts also help explain the layer structure of the network. Except for the instance-level interpretation, the CHAIN interpretation can also provide inference at the class level. Experiment results demonstrate that the proposed CHAIN model can provide reasonable interpretations at both intra-class and inter-class levels.

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)*, 2016, pp. 770–778.

[2] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 821–830.

[3] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, 2017.

[4] J. Townsend, T. Chaton, and J. M. Monteiro, "Extracting relational explanations from deep neural networks: A survey from a neural-symbolic perspective," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2019.

[5] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017.

[6] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018, pp. 2668–2677.

[7] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3319–3327.

[8] B. Zhou, Y. Sun, D. Bau, and A. Torralba, "Interpretable basis decomposition for visual explanation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 119–134.

[9] J. A. Lorteije, A. Zylberberg, B. G. Ouellette, C. I. De Zeeuw, M. Sigman, and P. R. Roelfsema, "The formation of hierarchical decisions in the visual cortex," *Neuron*, vol. 87, no. 6, pp. 1344–1356, 2015.

[10] B. Epshtein, I. Lifshitz, and S. Ullman, "Image interpretation by a single bottom-up top-down cycle," *Proceedings of the National Academy of Sciences*, vol. 105, no. 38, pp. 14 298–14 303, 2008.

[11] D. Marr, "Early processing of visual information," *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 275, no. 942, pp. 483–519, 1976.

[12] M. J. Riddoch and G. W. Humphreys, *Object recognition*. Psychology Press Hove, 2001.

[13] G. W. Humphreys, C. J. Price, and M. J. Riddoch, "From objects to names: A cognitive neuroscience approach," *Psychological research*, vol. 62, no. 2-3, pp. 118–130, 1999.

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

[15] 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.

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

[17] X. Cui, D. Wang, and Z. J. Wang, "Multi-scale interpretation model for convolutional neural networks: Building trust based on hierarchical interpretation," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2263–2276, 2019.

[18] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, and S. Behnke, "Interpretable and fine-grained visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9097–9107.

[19] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," in *Workshop in International Conference on Learning Representations*, 2015.

[20] 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)*, 2016, pp. 2921–2929.

[21] 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)*, 2017, pp. 618–626.

[22] X. Cui, D. Wang, and Z. J. Wang, "Chip: Channel-wise disentangled interpretation of deep convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2019.

[23] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2131–2145, 2019.

[24] X. Cui, D. Wang, and Z. J. Wang, "Feature-flow interpretation of deep convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1847–1861, 2020.

[25] Q. Zhang, X. Wang, R. Cao, Y. N. Wu, F. Shi, and S. Zhu, "Extracting an explanatory graph to interpret a cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[26] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4680–4688, 2011.

[27] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

[28] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5122–5130.

[29] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1971–1978.

[30] S. Bell, K. Bala, and N. Snavely, "Intrinsic images in the wild," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 1–12, 2014.