Synchronization-inspired Interpretable Neural Networks

Wei Han, Zhili Qin, Jiaming Liu, Christian Böhm, Junming Shao*

***Abstract*—Synchronization is a ubiquitous phenomenon in nature that enables the orderly presentation of information. In the human brain, for instance, functional modules such as the visual, motor, and language cortices form through neuronal synchronization. Inspired by biological brains and previous neuroscience studies, we propose an interpretable neural network incorporating a synchronization mechanism. The basic idea is to constrain each neuron, such as a convolution filter, to capture a single semantic pattern while synchronizing similar neurons to facilitate the formation of interpretable functional modules. Specifically, we regularize the activation map of a neuron to surround its focus position of the activated pattern in a sample. Moreover, neurons locally interact with each other, and similar ones are synchronized together during the training phase adaptively. Such local aggregation preserves the globally distributed representation nature of the neural network model, enabling a reasonably interpretable representation. To analyze the neuron interpretability comprehensively, we introduce a series of novel evaluation metrics from multiple aspects. Qualitative and quantitative experiments demonstrate that the proposed method outperforms many state-of-the-art algorithms in terms of interpretability. The resulting synchronized functional modules show module consistency across data and semantic specificity within modules.**

***Index Terms*—Interpretable Neural Networks, Synchronization, Active Interpretability, Interpretability Metrics.**

I. Introduction

Current deep neural network models (e.g. VGGNet, ResNet) are widely recognized as complex learning systems with an extremely large number of connections. Despite their excellent performance on various tasks, their internal knowledge and predictions are often difficult to interpret. The black-box property limits their further development and real-world applications, particularly in fields such as medicine and finance where results require careful consideration. Compared to biological brains, the success of deep neural network models is largely due to their large number of parameters and massive training data, while their network structures are relatively simple. The study of neural network interpretability has thus gained increasing attention in recent years. Instead of concentrating on the post-hoc interpretability of a trained neural network [1], [2], in this study, we introduce a new concept: **synchronization**, to improve the learning and interpretability of the artificial neural network.

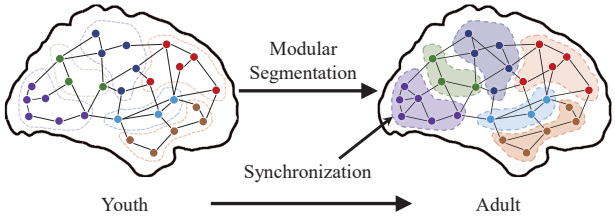


Fig. 1. Illustration of functional module formation in human brains via neuronal synchronization during aging. Here the motifs with different colors indicate distinct functional modules such as the visual cortex and motor cortex.

Synchronization is a fundamental phenomenon in which a group of events spontaneously come into co-occurrence with a common rhythm, despite differences in their individual rhythms. It is ubiquitous in human life, occurring in everything from metabolic processes in our cells to the highest cognitive tasks we perform as individuals within a group [3]. Emerging evidence of synchronization has been found in many neuroscience studies. For instance, cortical columns have been shown to contain neurons that respond to similar information [4], while groups of visual cortex neurons spontaneously synchronize together to perform similar functions in the human brain [5]. As synchronized neurons continually reinforce their connections, the brain network gradually segments into functional modules [6]. Figure 1 provides a simple illustration of functional module formation in human brains via neuronal synchronization during aging. The modular structure of the human brain provides many desirable properties, such as energy-efficient processing (i.e., a small-world network structure with a high clustering coefficient and low average path length) [7], robustness against damage by mutation or viral infection [8], and high interpretability (i.e., each module is often associated with a specific function) [9].

Synchronization promotes reasonable and sparse representation, enhancing the performance and interpretability of neural networks. Following the synchronization mechanism, neurons that respond to similar patterns interact with each other and adaptively aggregate into functional modules. The local aggregation of individual neurons does not significantly reduce the globally diverse response of all neurons to semantics. Such locally compact and globally distributed properties result

This work is supported by the Fundamental Research Funds for the Central Universities (ZYGX2019Z014), National Natural Science Foundation of China (61976044, 52079026), Fok Ying-Tong Education Foundation (161062), and Sichuan Science and Technology Program (2022YFG0260, 2020YFH0037).

W. Han, Z. Qin, J.-M. Liu, and J.-M. Shao are with the Department of Computer Science and Engineering, University of Electronic Science and Technology of China, China (E-mail: {weihan, qinzhili, liujiaming}@std.uestc.edu.cn, junmshao@uestc.edu.cn).

C. Böhm is with the Institute of Informatics, Ludwig Maximilian University of Munich, Germany (E-mail: boehm@ifi.lmu.de).

*Corresponding author: Junming Shao (junmshao@uestc.edu.cn).

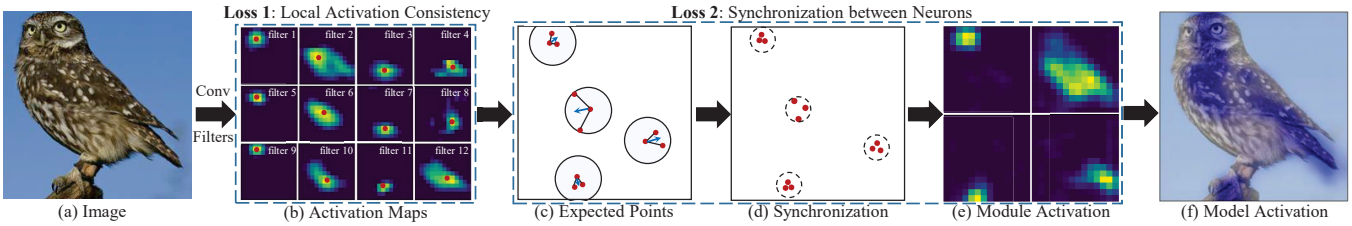


Fig. 2. Illustration of synchronization-inspired neural network. (a) A bird image. (b) The response of each neuron (i.e., the activation map of each convolutional filter in CNN) is constrained to a local continuous region with the loss of local activation consistency. (c) The activation points of filters in (b). During training, the activations of similar neurons (represented as neighboring activation points) are gradually synchronized together with synchronization loss. (d) After training, similar neurons are aggregated to achieve neuronal synchronization. (e) The activations of similar neurons form functional modules to capture the high-level semantics. (f) The visualization of functional module activation overlayed in the original bird, showing natural interpretability (i.e., each module activation corresponds to the bird's eye, abdomen, tail, and feet, respectively).

in a reasonable representation. The previous work [10] has demonstrated that the model with the best performance is of proper clustering coefficient and average path length, which is consistent among different data sets and neural network structures. Surprisingly, it is also similar to the real biological neural network. Additionally, the process of gathering scattered neurons into functional modules brings sparseness in model representation, allowing neurons to focus on meaningful semantics. Furthermore, compared to probing each neuron, investigating a small number of functional modules benefits the understanding of complex models. It alleviates gray-boxing of the interpretable model caused by excessive information and thus is a promising way.

In this study, we aim to generalize the concept of synchronization to construct an interpretable neural network. The key idea is to constrain each individual neuron to learn a simple semantic pattern, and group similar neurons together to form interpretable functional modules automatically. By constraining neuron activation consistency around a focus point, neurons will exhibit simple and clear functions. As illustrated in Fig. 2, for a given bird image, each convolutional filter (i.e., a neuron) activates a continual local image region to capture one semantic pattern of the bird only (e.g., bird's eye). Meanwhile, similar neurons are expected to synchronize together to form interpretable functional modules and thus allow capturing a higher semantic pattern. For instance, some convolutional filters activate neighboring image regions together to capture the bird's head in Fig. 2. Relying on the powerful concept of synchronization, the proposed interpretable neural network has many desirable properties, which are mainly summarized as follows.

- **Biological viewpoint.** Motivated by the advantages of biological systems, the inherited synchronization mechanism is first introduced to improve the interpretability of neural networks.
- **Automatic functional module formation**. Thanks to the synchronization-based clustering for automatic neuron aggregation, our proposed method allows forming interpretable functional modules to capture high-level semantic patterns.
- **High interpretability.** In contrast to the post-hoc interpretability of a trained neural network, the resulting interpretable functional module of our method lends itself to good interpretability. Extensive experiments have further demonstrated its superior interpretability compared with many state-of-the-art algorithms.

II. Related Works

A. Neural Network Interpretability

The network interpretability benefits understanding the behavior of the neural network. Existing interpretability works [2] mainly focus on the post-hoc interpretability of a trained neural network. According to the type of explanations, they are divided into example-based, attribution-based, hidden-semantics-based, and rule-based categories. A local or global instance-based interpretation could be given via similar cases [11] or prototypes [12]. Key instances were proposed to be evaluated by measuring the parameter influence of removing them from the training set [13]. The goal of attribution-based interpretation methods is to figure out key features. Shapley value [14] is adapted from the game theory for a solution to the payoff assignment problem of attributions. Sensitivity analyses [15], [16] probe the feature importance by introducing disturbances. Visualization methods, such as Class Activation Map [17], [18], intuitively display attribution contributions. The hidden semantics-based method focuses on exploring semantic patterns in units of the network. Network dissection [19] takes the intersection between feature maps and annotations as the index, counting the selectivity of units to semantic parts. However, the correspondence between units and semantics is overlapping. Therefore, the semantics is proposed to be represented properly by the vector embedding based on the weights of units [20]. The rule-based explanation explains the behavior of complex models by simple rules. The common method is the surrogate model, which analogies a complex neural network model locally [21], [22] or globally [23], [24] as a simple model to achieve interpretability. In addition to image data, interpretable methods also have applications in other data modalities such as time series data [25], graph data [26], and heterogeneous data [27]. However, the post-hoc method is only capable to provide information in the existing model and cannot disentangle the units and semantics. In contrast, we focus on active interpretability models which provide a potential solution to the problem.

B. Active interpretability Models

Active interpretability models achieve model-intrinsic interpretability by regularizing model complexity. Involved in the training process, they are capable to improve semantic representation in the model. However, it is still an open question. A prototype layer [28], [29] is introduced to make predictions with case-based interpretability. By dividing task samples based on expert knowledge or category hierarchy, the transparent design of the model is realized in reinforcement learning [30], visual question answer [31] and image classification tasks [32]. Attribution-based active interpretability constraints are employed to improve interpretability on metrics [33], introduce attribution priors [34], or select feature subsets [35], [36]. As tree models are intuitive interpretable models, tree constraints [37], [38] are added to model training to achieve rule-based active interpretability models. Hidden-semantics-based active interpretation methods focus on regularizing neurons in neural networks, such as filters in convolutional neural networks. Assigning filters to specific classes, CSGCNN [39] and Decoup [40] alleviate filter-class entanglement during training. icCNN [41] computes the prior filter cluster structure before each epoch and encourages such cluster structure in the training. Introducing additional semantic categories as concepts, Concept Whiten [42] aligns model representations with those concepts. ICNN [43] clarifies filter representations in high conv-layers of CNNs by minimizing the mutual information between filter feature maps and Gaussian templates. In the study, we employ the synchronization mechanism to aggregate neurons with clear and simple functions into functional modules spontaneously, achieving reasonable interpretability of network representation.

C. Interpretability Metrics

Due to the lack of a specific definition of interpretability, there are various and messy interpretability metrics. Since each interpretability method gives explanations from different perspectives, the *readability* [44] of given explanations to human beings is a concern. *Plausibility* [44], [45] means how convincing the interpretation is to humans, while *faithfulness* [46], [47] measures how accurately it reflects the true reasoning process of the model. As the two notations are conflated in some works [48], [49], [50], an article [51] proposes to make a distinction between the two criteria for potential users. Another concept is *fidelity* [21], [22], which evaluates how well the explanation approximates the prediction of the black-box model. *Stability* [52], [53], [54] under perturbation of interpreted parts is also a side basis for interpretation assessment. As for active hidden-semantics-based interpretability, the main interpretability metric is the object-part interpretability of filters [19], [43], [39]. It is defined based on the IoU between activation and pixel-wise annotation. Based on the distance from landmarks, the standard deviation is used to determine *instability* [43]. The entropy is introduced to measure *inconsistency* [41], and the sliding curve is used to show the trade-off relationship between multiple metrics of the interpretable model. In this study, we integrate several filter interpretability metrics into the entropy framework and assess interpretability from multiple aspects.

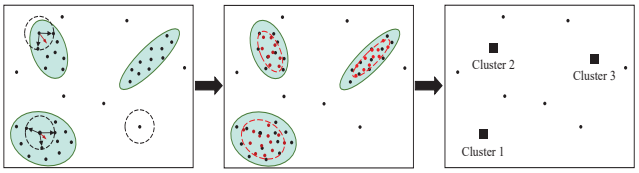


Fig. 3. Illustration of Synchronization-based clustering. The static data as phase oscillators consistently interact with others. Neighbors within the radius ϵ attract each other, resulting in multiple synchronized clusters. The synchronization mechanism ensures a globally distributed data structure while reducing local data complexity.

III. Methodology

In this section, we demonstrate a synchronization-inspired active interpretability model. We first introduce the basic concept of synchronization and its mathematical formulation as Preliminary in Section III-A. Then, the two parts of the proposed method are described in detail, including the local activation consistency constraint in Section III-B and the synchronization constraint in Section III-C. Finally, we present a series of quantitative metrics for active hidden-semantics-based interpretability in Section III-D.

A. Preliminary

To understand the synchronization concept and simulate its dynamics, many models have been proposed, e.g., the well-known Kuramoto model [55]. In this study, we extend the idea of synchronization-based clustering algorithms, such as SynC [56], to automatically group together similar neurons. The clustering by synchronization approach is to view each data object as a phase oscillator, with its feature vector representing its phase. By simulating the dynamic behaviors of these objects over time and their interaction with similar objects, the phase of an object gradually aligns with its neighbors, resulting in non-linear object movement driven by the local cluster structure. Ultimately, the objects in a cluster become synchronized, sharing the same feature vector (as illustrated in Fig. 3). Formally, let $x \in \mathbb{R}^d$ be an object in the data set D and $x_i(t)$ be its i-th dimension at time t respectively. The algorithm first identifies its ϵ-range neighbors $Nb_\epsilon(x(t))$ and then simulates the dynamics of each dimension x_i of the object x over time.

$$Nb_\epsilon(x) = \{y \in D | dist(y, x) \leq \epsilon\} \tag{1}$$

$$x_i(t+1) = x_i(t) + \frac{1}{|Nb_\epsilon(x(t))|} \cdot \sum_{y(t) \in Nb_\epsilon(x(t))} \sin\left(y_i(t) - x_i(t)\right) \tag{2}$$

Here, $|Nb_\epsilon(x(t))|$ is the neighbors number of the object $x(t)$. When applying synchronization to neural network models, some adaptations are necessary. Firstly, neurons have no specific semantic encoding as objects but only respond to specific samples. To achieve the aggregation of their functions, we

need to impose constraints on the neuron responses. Secondly, the parameters in the neural network constantly vary with iterations, just as the input samples do. In random iterations, the responses of all neurons may become neighbors to each other at some point, leading to an undesirable reduction in representational diversity of the neural network models [57]. To alleviate this problem in a dynamic environment, another force is needed to preserve diversity in the representation.

B. Neuronal Local Activation Consistency

To construct the synchronization-inspired interpretable neural network, the first key point is to ensure that the semantics of neurons are simple and clear. In a convolutional layer of a neural network, we take the feature map $A(f)$ of the outputs as the activation of the filter f. Since a semantic pattern usually corresponds to only one part of the data (e.g., the bird's eye is only one part of the image in Fig. 2), the feature map of each convolutional filter (i.e., each individual neuron) is thus supposed to be partially activated. In addition, another intuitive hypothesis on image data is that a semantic pattern always appears in a continual local area. For instance, in Fig. 2, the bird's eyes are located in the left-upper part of the image. If the filter is designed to capture the semantics of the bird's eye, the activation map of this filter should cover only the continual pixels of the eye. Therefore, we heuristically constrain the activation of a filter to concentrate in a single region, which is assumed to represent one object part. We refer to this constraint as **neuronal local activation consistency**.

Formally, given an image, let $A(f) = \{a_1(f), a_2(f), \cdots, a_M(f)\}$ be the activation map of a filter f at the convolution layer in a neural network, and let the corresponding activated location vector be $X(f) = \{x_1(f), x_2(f), \cdots, x_M(f)\}$, where $x_m(f)$ is the two-dimensional coordinate of m-th element on the activation map. The expectation of the activation position of a given activation map is calculated as $\bar{x}(f) = E_{m\sim q(a_m(f))}(X(f))$. The neuronal local activation consistency constraint is to regularize each activation point $x_m(f)$ to be close to $\bar{x}(f)$ to achieve a continuous local activation area. To this end, the activation consistency loss is defined as the cross-entropy-like distribution divergence between the actual activation probability $q(x_m(f))$ and the expected activation probability $p(x_m(f)|\bar{x}(f))$.

$$L_{lac} = -\frac{1}{N}\sum_f \sum_m p\big(x_m(f)|\bar{x}(f)\big)\log q\big(x_m(f)\big) \quad (3)$$

Here, N is the number of filters in the target convolution layer. $p\big(x_m(f)|\bar{x}(f)\big)$ is the activation probability of m-th position element on the condition of $\bar{x}(f)$. Intuitively, the closer $x_m(f)$ is to the expectation of the activation position $\bar{x}(f)$, the higher the activation probability. Thereby, we define the expected activation probability $p(x_m(f)|\bar{x}(f))$ using the softmax operation applied to the distance, as follows.

$$p\big(x_m(f)|\bar{x}(f)\big) = \frac{e^{-dist\left(x_m(f),\bar{x}(f)\right)}}{\sum_n e^{-dist\left(x_n(f),\bar{x}(f)\right)}} \quad (4)$$

where the Euclidean distance $dist(\cdot)$ is employed in this study. It is based on the softmax function which normalizes the value with the sum of activation values across every position n. Meanwhile, the actual activation probability $q(x_m(f))$ at each location m is calculated directly from the activation map $A(f)$.

$$q\big(x_m(f)\big) = \frac{e^{a_m(f)}}{\sum_n e^{a_n(f)}} \quad (5)$$

As activation positions are close to the activation position $\bar{x}(f)$, each neuron (i.e., a filter f) tends to capture one semantic pattern by activating a continual local area.

C. Interpretable Functional Module Formation with Neuronal Synchronization

The synchronization mechanism is a powerful and intrinsic concept in the human brain that automatically groups similar neurons together. Inspired by neuroscience studies, we aim to introduce this mechanism into artificial neural networks. In neuronal synchronization of neural networks, neurons that process similar functions are aggregated to form functional modules that represent high-level semantics. In the study, we define the function of a neuron as the expected semantic response when processing information from all samples. Thus, the semantic response in a single sample can be seen as a sampling of the neuron function, and the aggregation of neuron functions on the global level is achieved by synchronizing in the continuous sampling. Compared to the constraint on predefined filter clusters [41], which is based on statistics from the entire training set, the synchronization mechanism adaptively segments functional modules.

Building upon the neuronal local activation consistency constraint, each neuron is now capable of capturing a single semantic pattern. To reduce the computational cost, we adopt the neuron's response position $\bar{x}$ on the sample to represent its response. In each iteration, neurons with similar functions are expected to have similar response positions. Similar to synchronization-based clustering, where similar objects are grouped together by aligning with their ϵ-neighborhoods, we propose the **synchronization loss** to encourage neurons that are neighbors to have a higher mutual activation probability in one sampling. We denote the co-activated probability between neurons within the ϵ-neighborhood as $p(f_j|f_i)$, and the mutual activation probability as $q(f_j|f_i)$.

$$L_{sync} = -\sum_{f_i}\sum_{f_j\in Nb_\epsilon(f_i)} p(f_j|f_i)\log q(f_j|f_i) \quad (6)$$

Formally, we first define the neuronal co-activation probability $p(f_j|f_i)$ by the uniform distribution of the ϵ-neighborhood regarding the filter f_i during forward information processing. Here, the ϵ-neighborhood is a hard division based on the distance between the activation position of filters f_i and f_j. When the filters f_i and f_j are co-activated at a nearby position, we assume that they process similar information and strengthen their relationship (i.e., the mutual activation probability $q(f_j|f_i)$). The closer the neuronal activation positions are, the more likely they are to synchronize with each other.

$$p(f_j|f_i) = \frac{1}{|Nb_\epsilon(f_i)|} \quad (7)$$

$$q(f_j|f_i) = \frac{e^{-dist\left(\bar{x}(f_i),\bar{x}(f_j)\right)}}{\sum_{f_j} e^{-dist\left(\bar{x}(f_i),\bar{x}(f_j)\right)}} \tag{8}$$

To understand the synchronization constraint, we conduct a further derivation of the synchronization loss. It could be decomposed into two parts, including the pull on ϵ-neighbor elements and the push on overall elements. This approach ensures the aggregation of local information while avoiding the collapse of global representations.

$$\begin{aligned} L_{sync} &= -\sum_{f_i}\sum_{f_j\in Nb_\epsilon} p(f_j|f_i)\log q(f_j|f_i) \\ &= -\sum_{f_i}\sum_{f_j\in Nb_\epsilon(f_i)} \frac{1}{|Nb_\epsilon(f_i)|}\log\frac{e^{-dist\left(\bar{x}(f_i),\bar{x}(f_j)\right)}}{\sum_{f_j} e^{-dist\left(\bar{x}(f_i),\bar{x}(f_j)\right)}} \\ &= \frac{1}{|Nb_\epsilon(f_i)|}\sum_{f_i}\Big(\sum_{f_j\in Nb_\epsilon(f_i)} dist\big(\bar{x}(f_i),\bar{x}(f_j) \\ &\qquad + \log\sum_{f_j} e^{-dist\left(\bar{x}(f_i),\bar{x}(f_j)\right)}\Big) \end{aligned} \tag{9}$$

Finally, the proposed method optimizes the interpretable neural network model by jointly considering the cross entropy loss of the classification task L_{ce}, the local activation consistency loss, and the synchronization loss, with balance coefficients λ_1 and λ_2, respectively.

$$L = L_{ce} + \lambda_1 L_{lac} + \lambda_2 L_{sync} \tag{10}$$

D. Various Metrics of Interpretability

In previous works, the mean object-part interpretability of neurons [19], [43], [39] is adopted to assess hidden-semantics-based interpretability, which is based on the intersection-over-union score (IoU). The IoU score provides a similarity measurement between part mask annotations and the receptive field of convolutional filters. Specifically, the IoU score between filter f and the k-th part on the image I is defined as $IoU^I_{f,k} = |S^I_f \cap S^I_k|/|S^I_f \cup S^I_k|$, where S^I_f and S^I_k denote the part region of filter f and the ground-truth mask of the k-th part, respectively. Based on such a definition, the mean object-part interpretability gives the probability of a filter associating with a specific semantic part. In accordance with previous work [43], we assign filter f to the k-th part in image I if $IoU^I_{f,k} > \phi$, where ϕ is the assigning threshold. The object-part interpretability of filters is calculated by reporting the highest probability among all parts across all images.

$$Interp(f) = \max_k \text{mean}_I \mathbb{I}\left(IoU^I_{f,k} > \phi\right), \tag{11}$$

where $\mathbb{I}(condition)$ is an indicator function, which is 1 if the condition satisfies and 0 otherwise. Please note that the filter interpretability in the previous work [43] is calculated with the filters that are assigned to specific classes in multi-category classification. However, we choose not to split filters in the study to obtain a comprehensive result. For example, if 512 filters are assigned to 200 classes in the CUB200 data set, some classes may be associated with many filters, while other classes may not have any.

In addition to the direct interpretability metric, we introduce entropy-based metrics to assess interpretability from various perspectives. First, the stability of the interpretability is considered. It measures whether the filter associates with specific patterns stably. The entropy $H(\cdot)$ means the degree of chaos, and the entropy in the *Stability* metric would be low if a filter is consistently associated with one semantic part among all images. Second, we investigate whether a filter only focuses on one part of one image. The *Purity* metric calculates the entropy of the responded parts of the filter. Although this interpretability measure may be high, it would be meaningless if all filters only focus on one or a few semantic parts. Therefore, we take the entropy of activated parts on an image as the *Diversity* metric to demonstrate the coverage of network activations on global semantic parts.

$$Stability(f) = H_k\left(\text{mean}_I 1(IoU^I_{f,k} > \phi)\right) \tag{12}$$

$$Purity(f) = \text{mean}_I\left(H_k\left(1(IoU^I_{f,k} > \phi)\right)\right) \tag{13}$$

$$Diversity(I) = H_k\left(\max_f 1(IoU^I_{f,k} > \phi)\right) \tag{14}$$

The values fed into the entropy function are normalized via dividing by their summed value, while the calculated entropy values are normalized to [0, 1] via dividing by the biggest value, namely, the entropy of the uniform distribution.

In terms of the data set with only landmark annotations, we take the negative distance between filter focus and landmarks as the similarity measure $Sim^I_{f,k}$, instead of $IoU^I_{f,k}$ in the metric calculation. However, this similarity based on the distance between the filter focus and landmarks is equivalent to identifying coverage upon a circle that takes the landmark as the center and ϵ as the radius. The shape of semantic parts is always not circular, and the positions of some semantic objects may overlap. Metrics based on this similarity measure can easily fall into the extreme case of highly overlapping or exclusive. Therefore, inspired by the previous work [41], we introduce a sliding assigning threshold and compare the changes of metric-pair curves to perform an assessment in the datasets with landmark annotations.

IV. Experiments

A. Experimental Setup

1) Data sets: In this study, we select four real-world datasets with part mask/landmark annotations, including the PASCAL VOC Part dataset [58], the CUB200-2011 dataset [59], the ILSVRC 2013 DET Animal-Part dataset [60], and the BORDEN dataset [19]. Since the BORDEN data set does not have multi-classification task labels, we train the model in Imagenet-100 [61] and verify its interpretability on the BORDEN dataset. Following the protocol [43], there are 4 or 5 coarse semantic parts of six animal categories annotated in the PASCAL VOC Part dataset. The objects whose size is bigger than 50×50 are cropped. Meanwhile, the BRODEN dataset contains over 60,000 images with pixel-level and image-level annotations for 1,197 fine-grained concepts. The Imagenet-100 is a subset of the ILSVRC dataset [62]. We use 30 animal categories in the ILSVRC DET dataset for training and testing. The ground-truth positions of the head,

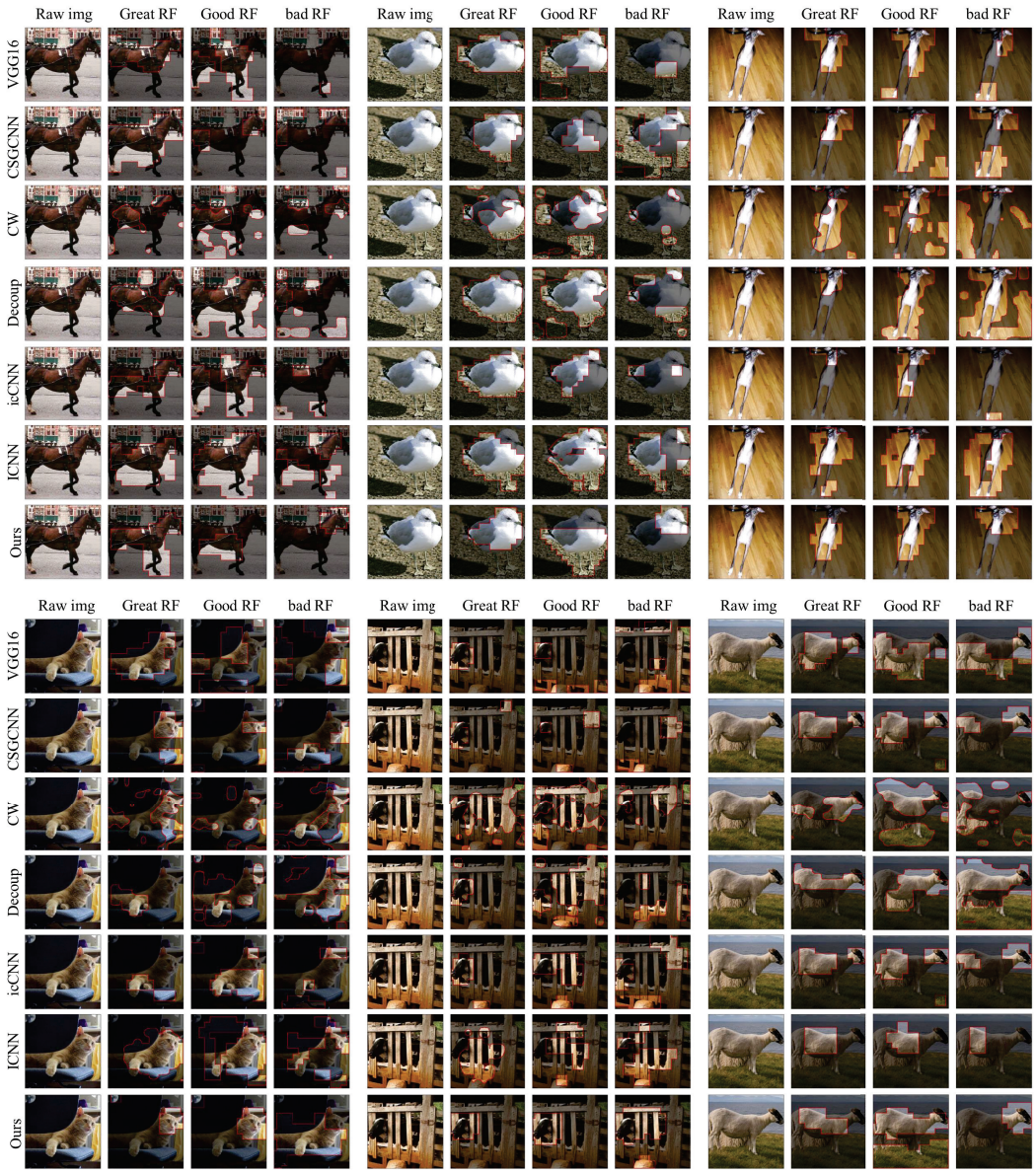


Fig. 4. Illustration of filter activation map of different methods. Here we select the activation maps at 95, 70, and 30 percent quantiles of IoU to show as the great, good, and bad activation maps, respectively. The proposed method enables a single neuron to respond to a single region, which is specific for a semantic part.

the back, and the tail of birds in the CUB200 dataset are employed as landmarks, while the annotations of the head and frontal legs in the ILSVRC DET Animal-Part dataset are given as landmarks. However, the annotation in the ILSVRC DET Animal-Part dataset is sparse, we adopt all annotated training and testing data for the quantitative interpretability evaluation. The input images of all datasets are 224×224. The training and testing splits are based on given split files of downloaded data sets.

2) Comparison Methods: The proposed method aims to achieve a hidden semantics-based interpretability of neural networks. We compare it against state-of-the-art methods, including the following baselines: (1) **Baseline** serve as the backbone for all methods, including two classic architectures: VGG16 [63] and ResNet18 [64]; (2) **CSGCNN** [39] trains interpretable convolutional neural networks by differentiating class-specific filters; (3) **CW** [42] whiten the representation based on additional concept samples; (4) **Decoup** [40] employs hard activation routing to show the information processing of neural networks and supervises learning of interpretable information based on class labels. (5) **icCNN** [41] encourages pre-defined filter clusters in the kernel space to regularize filters to represent semantic parts consistently and diversely; (6) **ICNN** [43] minimizes the mutual information between filter feature maps and Gaussian templates to clarify filter representations in high conv-layers of CNNs.

3) Evaluation Metrics: Besides the filter interpretability (*Interp.*) measure, a series of metrics are also adopted to assess the interpretability from various aspects, consisting of *Stability*, *Purity* and *Diversity*. Meanwhile, we employ

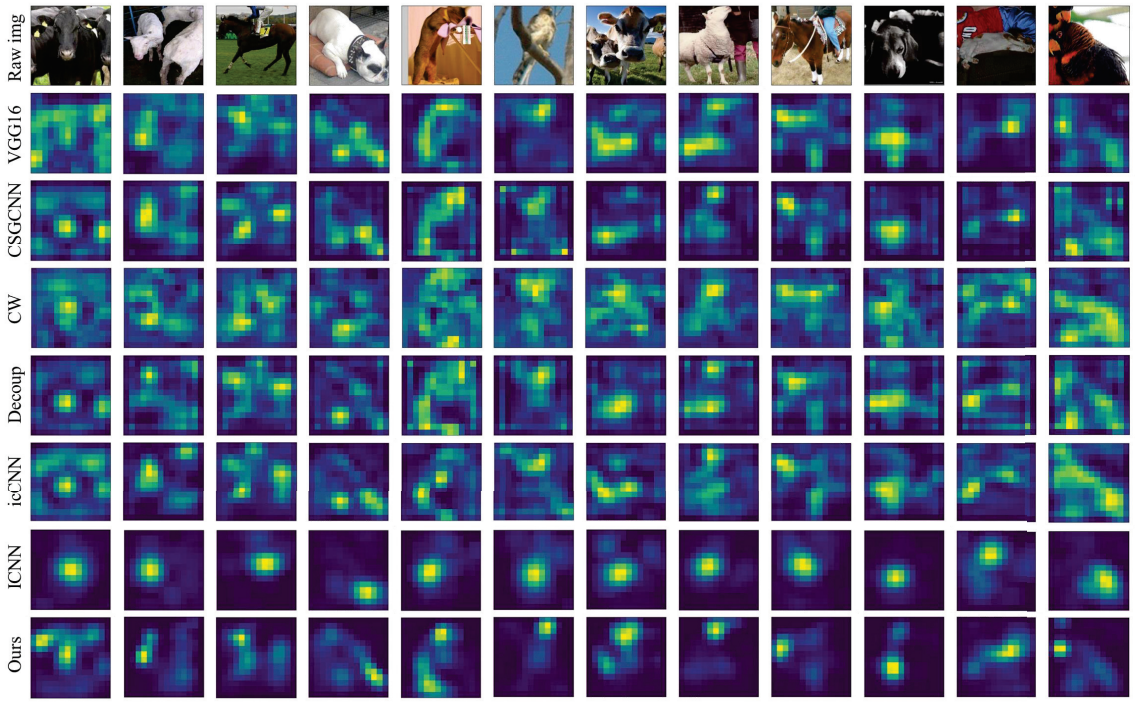


Fig. 5. Average feature maps of all filters among different models. Due to the concentration of individual neuron responses and the local aggregation of neuron functions, the global average activation map of the proposed neural network exhibits pure, accurate, and diverse properties.

classification accuracy as the fundamental metric to illustrate the effect of different interpretable constraints on the model performance.

4) Experimental Settings: All models in this study employ the baseline model as the backbone which is pre-trained with ImageNet images, and the base task is the multi-category classification. The interpretable layer is set to be the top convolutional layer, and the classification part is a fully connected layer. The quantitative interpretability metrics are calculated based on the activation of all filters in the interpretable layer. Since the classic method ICNN[1] does not offer the code of interpretability evaluation, we adopt the corresponding code of CSGCNN[2] to evaluate the filter object-part interpretability. As comparison methods do not provide a well-trained model and the results of metrics introduced in the paper, we re-run the comparison methods with official codes on GitHub. Besides the implementation of the proposed method[3], the results of algorithms presented in the paper are based on the same training setting. However, the data setting in the ICNN code limits its multi-category experiment on CUB200 and ImageNet-100 datasets. Meanwhile, the official code is based on the SimpleNN framework in MATLAB, which does not support ResNet. As a result, it cannot be directly transformed into the backbone of ResNet. For a fair comparison, the extra mask operation of ICNN for activation maps is not adopted in interpretability evaluation, and the hard routing layer of Decoup is only applied to the target layer. Meanwhile, the sparse constraint in Decoup will greatly impair interpretability, and therefore, we set its balance coefficient to 0. Apart from ICNN being run on Matlab 2018b, all other experiments are implemented by Pytorch 1.5.0 and run on an RTX 2080Ti card under Ubuntu 18.04.

5) Implementation Details: In the training, the SGD optimizer with an initial learning rate of 0.01 is adopted. It is accompanied by a $0.1\times$ learning rate decay at $\{50, 75, 90\}$ during 100 epochs. The batch size is set as 32 for VGGNet and 64 for ResNet, and the weight decay is $5e^{-4}$. The ϵ for neighbor identification is assigned as 0.03. The λ_1 for neural local activation consistency loss is 0.2, while λ_2 for synchronization loss is 0.1. These two parameters were halved in ResNet, with λ_1 set to 0.1 and λ_2 set to 0.05. Another halving occurs in the multi-classification task of ImageNet-100 for both architectures. In the mask-based evaluation, we fix the assigning threshold ϕ of 0.2 for animal datasets and 0.04 for the BORDEN dataset. In the landmark-based evaluation, we vary the assigning thresholds to obtain the metric-pair curves.

B. Qualitative Evaluation Results

1) Interpretability of Individual Neuronal Activation: To gain insight into filter interpretability, we provide an intuitive demonstration of individual activations by visualizing the receptive field of filters. The computation of the filter receptive field follows the method [19], which scales up the valid feature map to the image resolution. We select filters in various conditions according to a unified standard to showcase the interpretability of different methods on the same image. Specifically, the activation maps at the 95th, 70th, and 30th

[1]https://github.com/zqs1022/interpretableCNN

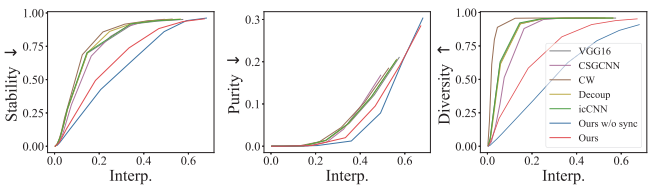
[2]https://github.com/hyliang96/CSGCNN

[3]Our codes with interpretability evaluation are provided in the supplementary material and will be released to GitHub after the paper is accepted.

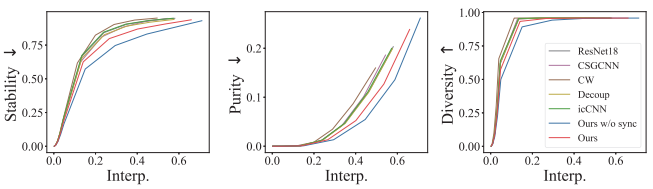
TABLE I
QUANTITATIVE RESULTS OF DIFFERENT ALGORITHMS UPON VARIOUS INTERPRETABILITY METRICS ON THE VOC PART DATASET.

	Method	VGG16				ResNet18			
		Interp.↑	Stability↓	Purity↓	Diversity↑	Interp.↑	Stability↓	Purity↓	Diversity↑
VOC Part	Baseline	0.4613	0.5526	0.0614	**0.8845**	0.4388	0.5371	0.0667	**0.8905**
	CSGCNN	0.5055	0.5202	0.0698	0.8539	0.3311	0.5294	0.0452	0.8733
	Decoup	0.4571	0.5242	0.0577	0.8626	0.3704	0.5121	0.0474	0.8880
	CW	0.4218	0.4743	**0.0411**	0.6542	0.3114	0.5378	**0.0289**	0.8730
	IcCNN	0.4505	0.5492	0.0556	0.8790	0.4464	0.4886	0.0723	0.8612
	ICNN*	0.4991	0.4970	0.0581	0.3409	-	-	-	-
	Ours w/o sync	**0.6825**	0.4593	0.0810	0.6950	**0.7282**	0.4431	0.0965	0.5128
	Ours	0.6418	**0.4346**	0.0511	0.8203	0.6598	**0.4290**	0.0652	0.7452
BORDEN	Baseline	0.2171	0.0018	0.1773	0.7915	0.4210	0.0017	0.3742	0.7643
	CSGCNN	0.2432	0.0015	0.1829	0.8530	0.3958	0.0018	0.3268	**0.8235**
	Decoup	0.2321	0.0019	0.1751	0.7284	0.4150	0.0018	0.3671	0.7769
	CW	0.2497	0.0054	0.2025	**0.9130**	0.3646	**0.0013**	**0.3105**	0.1979
	IcCNN	0.2304	0.0128	**0.1590**	0.8569	0.4013	0.0016	0.3337	0.8149
	ICNN*	-	-	-	-	-	-	-	-
	Ours w/o sync	**0.3323**	0.0015	0.2258	0.6512	0.4441	0.0021	0.3902	0.7567
	Ours	0.3156	**0.0012**	0.1697	0.7529	**0.4621**	0.0017	0.3417	0.7762

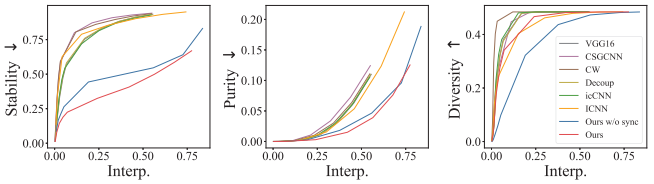
* The official code is based on the SimpleNN structure in MATLAB, which does not support ResNet. Meanwhile, its data setting focuses on single classification tasks and does not support large-scale multi-classification experiments, such as CUB200 and ImageNet-100.



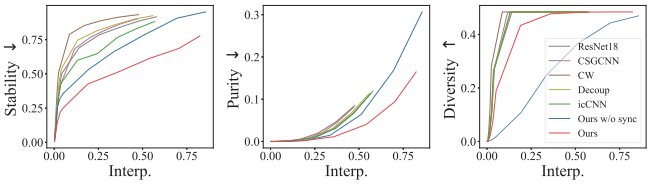
(a) VGG16 as backbone on CUB200-2011 Dataset



(b) ResNet18 as backbone on CUB200-2011 Dataset



(c) VGG16 as backbone on ILSVRC Part Dataset



(d) ResNet18 as backbone on ILSVRC Part Dataset

Fig. 6. Interpretability metric-pair curves in datasets with landmark-based annotations.

percentiles of the IoU score are chosen as representative of great, good, and bad activations, respectively.

Fig. 4 depicts filter receptive fields in top conv-layers of an interpretable neural network trained for multi-category classification. In the baseline model, the filter cover multiple semantic parts and produce multiple activation areas. Meanwhile, CSGCNN and Decoup only clarify the class-filter relationship without constraining features, resulting in scattered filter responses. This unclear situation is even more obvious on the filter activation map of CW, which may be that the inappropriate selection of the concept category will make the features whiten in an undesirable direction. The filters in icCNN are encouraged to be different in different filter groups. The content of activation maps in ICNN is more concentrated than the one in other algorithms, since it regularizes the model with Gaussian templates. Moreover, the proposed method considers both the effects within and between filters. As shown, the response of a single filter focuses on a specific semantic object, and multiple filters show distributed representations.

2) Interpretability of Functional Module: To further explore the global effects of functional modules in neural networks, we visualize the overall activation of the high-level convolution layer in image processing. Figure 5 shows the

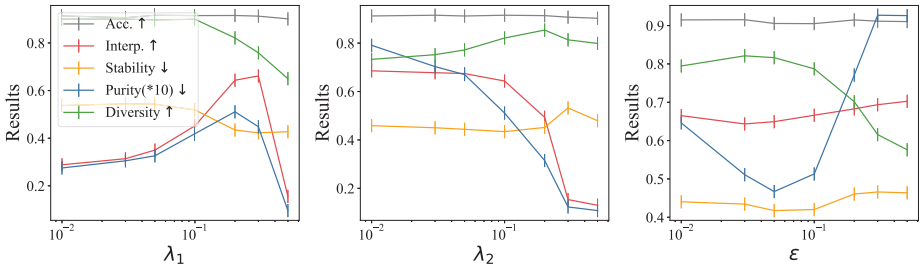


Fig. 7. Parameter sensitivity analysis on VOC Part dataset.

TABLE II
THE MULTI-CATEGORY CLASSIFICATION ACCURACY OF INTERPRETABLE NEURAL NETWORKS ON REAL-WORLD DATASETS.

Methods	VGG16			
	VOC Part	CUB-200	ILSVRC Part	ImageNet-100
Baseline	90.99%	81.39%	90.71%	84.12%
CSGCNN	90.66%	77.20%	86.65%	80.70%
Decoup	91.43%	80.69%	91.04%	**84.64%**
CW*	90.72%	79.58%	**91.73%**	83.86%
icCNN	90.44%	80.36%	90.84%	84.30%
ICNN	**91.97%**	-	86.30%	-
Ours w/o sync	91.10%	**82.68%**	90.81%	83.96%
Ours	91.53%	82.48%	91.10%	84.40%
Methods	**ResNet18**			
	VOC Part	CUB-200	ILSVRC Part	ImageNet-100
Baseline	88.26%	73.90%	88.72%	81.18%
CSGCNN	87.11%	71.82%	88.76%	80.68%
Decoup	88.50%	74.01%	87.17%	81.86%
CW*	**90.77%**	73.42%	**90.48%**	81.28%
icCNN	88.63%	73.80%	88.82%	81.44%
ICNN	-	-	-	-
Ours w/o sync	88.26%	74.04%	88.89%	**82.14%**
Ours	88.53%	**74.34%**	89.09%	82.04%

* The CW method involves extra concept data.

mean feature maps across all filters for different models. Since the VGG16, CSGCNN, Decoup, and icCNN lack direct regularization on the response within filters, their average activation maps appear scattered. Unfortunately, CW fails to show an advantage in feature interpretability. In contrast, both ICNN and our model are of such constraints, and it is conceivable that the average activation maps of the two methods are purer than the others. However, the filters of ICNN collapsed to the same activation of a single or few semantic parts, which may be sufficient for the classification task but insufficient for global semantic representation. In contrast, our proposed method achieves both clear activation and distributed representation. While the activation consistency constraint ensures the purity of a single filter to semantic objects, the

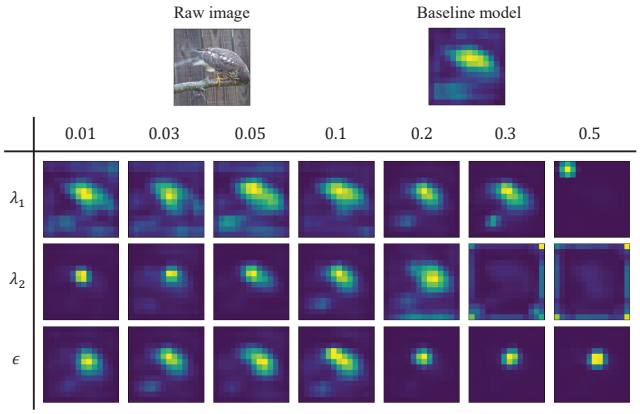


Fig. 8. Average filter activation maps under different hyper-parameter settings.

functional modules formed by neuronal synchronization show a distributed response to global semantic objects.

C. Quantitative Evaluation Results

1) Mask-based Metric Results: The quantitative metrics from various perspectives give a comprehensive assessment of hidden-semantics-based active interpretability, listed in Table I. *Interp.* is the main goal of interpretable models. Under rigorous task and evaluation settings, the existing methods have made incremental progress. Compared with state-of-the-art methods, our method achieves significant improvement on the metric and hits the top in both architectures. However, ICNN and ours with only the activation consistency constraint present a low diversity on semantic parts. The proposed method with the synchronization mechanism preserves diverse interpretability. Meanwhile, the improvement of *Stability* and *Purity* metrics means that the proposed interpretable model is capable of consistently associating filters with corresponding specific semantic patterns.

2) Mark-based Metric Curves: We employ the metric pair curve to assess interpretable neural networks on landmark-annotated datasets. Numerical comparison under the same interpretability scale is valid. Figure 6 illustrates the metric-pair curves of different CNNs on different datasets, with *Interp.* as the base axis. In both datasets and architectures, the interpretability of the proposed method achieves superior stability and purity, as well as comparable diversity. Moreover, our model has a broader interpretability scale and exhibits

TABLE III
MODULE CONSISTENCY. THE LESS IS BETTER.

Methods	VOC Part	CUB-200	ILSVRC Part	ImageNet-100
VGG16	0.4617	0.4855	0.4847	0.4858
icCNN	0.4481	0.4827	0.4653	0.4711
Ours w/o sync	**0.1501**	0.4925	**0.1507**	0.4872
Ours	0.3267	**0.4306**	0.2663	**0.3571**

TABLE IV
ALIGNMENT BETWEEN MODULES AND SEMANTIC PARTS.

Methods	NMI	ARI	AVI
VGG16	0.1432	0.1152	0.1267
icCNN	0.1599	0.1277	0.1532
Ours w/o sync	0.2502	0.3126	0.2285
Ours	**0.3538**	**0.3905**	**0.3542**

better interpretability than state-of-the-art interpretable models. Our model with only the activation consistency constraint has a greater advantage on the CUB200 dataset. This may be because the CUB200 dataset focuses only on birds, and the patterns across the entire dataset are not complicated.

3) Classification Accuracy: Table II summarizes the classification accuracy, demonstrating the trade-off effect of different interpretable constraints on multi-category classification tasks. While the proposed method performs well on model interpretability, it also provides comparable or even superior results compared to other methods. The CSGCNN method assigns a filter to a specific category. As the number of categories increases, the number of filters per category will decrease. This damages the capacity and representation ability of the model, leading to a decline in performance, as observed in the CUB200, ILSVRC Part, and ImageNet-100 datasets. Other algorithms do not cause great degradation of classification performance. Even the CW algorithm utilizes extra concept data, boosting its classification accuracy.

D. Sensitivity Analysis

The synchronization-inspired interpretable neural network involves two constraints and several hyper-parameters. To investigate the effect of each component, we conduct a parameter sensitivity analysis on the balance coefficients λ_1 and λ_2, as well as the neighbor range ϵ. Specifically, on the VOC Part dataset with VGG16 as the backbone, we vary one parameter within the range of $\{0.01, 0.03, 0.05, 0.1, 0.2, 0.3, 0.5\}$ while keeping the other parameters at their optimal value.

Figure 7 summarizes the quantitative results of the sensitivity experiments, while Figure 8 shows the qualitative effect of hyper-parameters on global filter activation. The accuracy remains robust to changes in hyper-parameters, with only a slight decrease under extreme settings. Increasing the coefficient λ_1 of the activation consistency constraint enables filters to focus on semantic objects rather than noisy areas, thereby enhancing *Interp.* and *Stability*. However, excessive consistency constraints can cause filters to completely collapse, damaging model performance on the metrics. Strengthening the synchronization constraint, i.e., increasing λ_2, promotes filters with similar functions to mutually excite. This allows the filters in the modules to focus on specific semantic parts in a stable and pure manner, while the filters between the modules exhibit diversity. However, extremely high values can cause filters to associate with meaningless spaces. The neighboring range ϵ affects the synchronization range between filters. When its value is too small, a filter cannot be synchronized with other filters, reducing the effect of synchronization loss. In contrast, when its value is too large, most filters are grouped into a cluster, deepening the collapse of filter functions and damaging the model performance on various indicators.

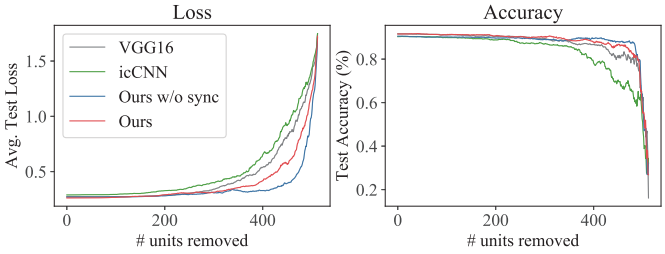


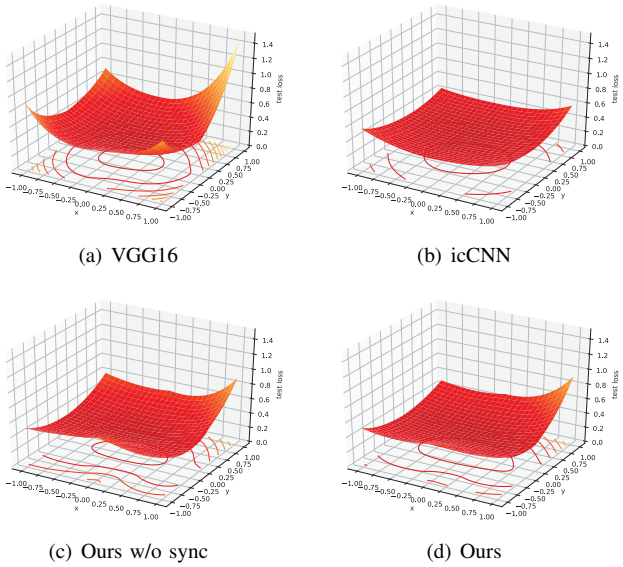
Fig. 9. Model redundancy investigation by pruning.

E. Functional Module Analysis

The proposed method aims to co-activate similar neurons to form functional modules by neuronal synchronization. This raises several interesting questions: 1) Do the gathered neurons consistently belong to the same module? 2) Does the functional module enhance model interpretability? 3) Is the information processing redundant in the model with modules? 4) Does the modular structure enhance the model robustness against the parameter disturbance? In this section, we conduct functional module analysis on the VOC Part dataset to investigate these questions. Additionally, since icCNN groups neurons with pre-defined clusters, we further employ this method for comparison.

1) Module Consistency: We define the connectivity of filters by *cosine* similarity and spectral clustering. Specifically, the normalized feature map $A(f_i)$ as the filter activation of the image is adopted to represent filter behaviors. The original high-dimensional feature map is used here, therefore, the cosine similarity is selected instead of Euclidean distance. The filter similarity on the image is calculated by the dot product, i.e., $sim_{i,j} = A(f_i) \cdot A(f_j)$. Based on the similarity matrix, spectral clustering is conducted, with the number of clusters set as the pre-defined number of semantic parts in the dataset. The connectivity of filters on one image is the upper triangular matrix of the adjacency matrix given by the clustering. Finally, module consistency is defined as the average standard deviation of filter connectivity across different images.

The module consistency of different models is listed in Table III. The smaller the index value, the better the module



(a) VGG16 (b) icCNN

(c) Ours w/o sync (d) Ours

Fig. 10. Disturbance sensitivity of models on VOC Part dataset.

consistency. icCNN only slightly outperforms the baseline model, indicating that the icCNN method cannot guarantee consistency in the module of gathered filters among different instances. Meanwhile, when only the activation consistency constraint is adopted, our model shows the best module consistency. This is because most filters collapsed into a module. In contrast, the synchronization constraint retains the global structure, making a trade-off on the module consistency of the proposed model.

2) Model Interpretability with Modules: To investigate whether functional modules in the high-level convolution layer benefit model interpretability, we compare the relationship between filters and modules with the relationship between filters and semantic parts. The IoU-based index mentioned in Section III-D provides a way to assign filters to semantic parts. Meanwhile, the previous sub-section describes how to cluster filters into modules. Given these two relationships, we introduce the metrics including NMI [65] (Normalized Mutual Information), ARI [66] (Adjusted Rand Index), and AVI [67] (Adjusted Variation of Information) to measure their alignments. Table IV illustrates the alignment between modules and semantic parts, showing how functional modules help filter interpretability. In the table, the icCNN method did not demonstrate significant gains in the relationship between modules and semantic parts. The indicator values of the proposed method are greater than those of other models, especially ours without the synchronization constraint. It means that the proposed synchronization loss imposes a clear correspondence of modules to semantic parts, enabling a white-box model through a small number of functional modules.

3) Model Redundancy with Modules: We encourage the mutual excitation of neurons activated by approximate patterns to form modules expressing similar functions. Therefore, the emergence of functional modules should increase the function redundancy of neurons in the neural network model. To explicitly investigate this phenomenon, we introduce the neural network pruning method based on Taylor expansion [68]. This experiment explores function redundancy in neural networks by sequentially removing the redundant filter that has the least impact on results. Figure 9 shows that our method exhibits higher redundancy to prevent performance degradation. However, since neurons have collapsed, the model with only the activation consistency constraint is of more redundant units. Surprisingly, icCNN is difficult to resist filter pruning, possibly because it encouraged differences between different groups to a greater extent.

4) Model Robustness with Modules: Since neurons in the same module are redundant, can the redundancy of the module benefit model robustness? We employ the technology of loss landscape [69] to investigate the sensitivity of model performance to disturbances. The parameters in the high-level convolution layer are changed in two orthogonal directions. Based on sampling, the loss surface of the disturbed model upon test data is drawn in Fig. 10. As shown, models with modular constraints have lower loss values and smoother loss landscapes than the baseline model. Namely, icCNN and the proposed model with the synchronization loss perform more stably against parameter disturbance than ours without the synchronization loss and the baseline model, respectively.

V. Challenges

While active interpretation methods in neural networks provide real-time and credible advantages, how to develop an effective constraint is still the primary challenge in this field. Different from existing methods that require additional semantic annotations, class information, or prior clustering information, the introduced synchronization mechanism facilitates adaptive interactions between neurons, imposing effective constraints toward the ordered presentation of information. Meanwhile, balancing model performance and interpretability is a crucial challenge when imposing interpretability constraints, as these constraints reduce the complexity of model representation. The synchronization method allows for local aggregation of representation while maintaining global distribution, preserving the model capacity. Eventually, the lack of a reasonable and comprehensive evaluation system remains a significant obstacle to the development of active interpretability methods. The assessment of active interpretability toward optimized neurons is mainly based on filter object-part interpretability, whereas post-hoc interpretability offers a comprehensive hexagonal capability map [70]. To comprehensively evaluate the interpretability of constrained neurons, we propose a set of metrics that examine the neuron interpretability from multiple perspectives.

VI. Limitations

Although our proposed method has demonstrated promise in reducing representation complexity and enhancing network interpretability by simplifying and aggregating neurons in the spatial domain, it does have certain limitations. Specifically, our method shares a common limitation with existing feature-based actively interpretable neural network methods, namely, it is applicable to high-level convolutional layers. This is due to

the enrichment of semantic information in neural networks as layer depth increases, making it difficult to constrain neuron representation of low-level semantics which is in the wide distribution in the spatial domain at lower layers. In the next work, we will focus on addressing this limitation by implementing a model-based constraint. Moreover, although our method performs well under the same parameter settings across multiple datasets, optimal performance still requires parameter tuning for different neural network architectures. Automated parameter tuning on small datasets may help alleviate this issue. Despite these limitations, our proposed method significantly enhances neural network interpretability and provides novel insights and ideas. We hope our findings will inspire further research.

VII. Future Research Directions

In future research, we aim to overcome the existing limitations. Namely, we plan to give an explanation for every element in actively interpretable neural networks. It enables a fully transparent design of neural networks. Therefore, how to extend the scope of active interpretability constraints is the main research direction in the future. Additionally, as there is no prior standard for semantics, the semantics of constrained neurons cannot be directly described. In future research, promoting the human-friendly interpretation of actively interpretable models without additional semantic supervision is a valuable research direction. Finally, for powerful pre-training models, incorporating active interpretability constraints in unsupervised comparative training presents an interesting research direction.

VIII. Conclusion

In this paper, we propose a biological-inspired interpretable neural network, by introducing the synchronization mechanism to construct interpretable functional modules. To this end, we constrain each neuron to capture one semantic pattern with local activation consistency loss. Afterward, the synchronization loss is proposed so that neurons responding to the same pattern are aggregated together to form functional modules, preserving globally distributed representation. A series of evaluation metrics from different aspects are introduced for comprehensive interpretability assessment. Qualitative and quantitative experiments have demonstrated that the proposed method provides superior interpretability compared with many state-of-the-art algorithms. We hope this study will enlighten people to consider interpretability from a new biological perspective.

References

[1] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[2] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, 2021.

[3] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou, "Synchronization in complex networks," *Physics reports*, vol. 469, no. 3, pp. 93–153, 2008.

[4] J. C. Horton and D. L. Adams, "The cortical column: a structure without a function," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1456, pp. 837–862, 2005.

[5] S. Zeki and S. Shipp, "The functional logic of cortical connections," *Nature*, vol. 335, no. 6188, p. 311, 1988.

[6] G. L. Baum, R. Ciric, D. R. Roalf, R. F. Betzel, T. M. Moore, R. T. Shinohara, A. E. Kahn, S. N. Vandekar, P. E. Rupert, M. Quarmley *et al.*, "Modular segregation of structural brain networks supports the development of executive function in youth," *Current Biology*, vol. 27, no. 11, pp. 1561–1572, 2017.

[7] K. Hilger, M. Fukushima, O. Sporns, and C. J. Fiebach, "Temporal stability of functional brain modules associated with human intelligence," *Human brain mapping*, vol. 41, no. 2, pp. 362–372, 2020.

[8] H. Aerts, W. Fias, K. Caeyenberghs, and D. Marinazzo, "Brain networks under attack: robustness properties and the impact of lesions," *Brain*, vol. 139, no. 12, pp. 3063–3083, 2016.

[9] M. Loukas, C. Pennell, C. Groat, R. S. Tubbs, and A. A. Cohen-Gadol, "Korbinian brodmann (1868–1918) and his contributions to mapping the cerebral cortex," *Neurosurgery*, vol. 68, no. 1, pp. 6–11, 2011.

[10] J. You, J. Leskovec, K. He, and S. Xie, "Graph structure of neural networks," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 881–10 891.

[11] R. Caruana, H. Kangarloo, J. D. Dionisio, U. Sinha, and D. Johnson, "Case-based explanation of non-case-based learning methods." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 1999, p. 212.

[12] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," *The Annals of Applied Statistics*, pp. 2403–2424, 2011.

[13] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1885–1894.

[14] L. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, pp. 307–317, 1953.

[15] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3429–3437.

[16] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[19] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6541–6549.

[20] R. Fong and A. Vedaldi, "Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8730–8738.

[21] 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*, 2016, pp. 1135–1144.

[22] G. Plumb, D. Molitor, and A. Talwalkar, "Model agnostic supervised local explanations," *arXiv preprint arXiv:1807.02910*, 2018.

[23] R. Setiono and H. Liu, "Understanding neural networks via rule extraction," in *IJCAI*, vol. 1. Citeseer, 1995, pp. 480–485.

[24] R. Chen, H. Chen, J. Ren, G. Huang, and Q. Zhang, "Explaining neural networks semantically and quantitatively," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9187–9196.

[25] B.-J. Hou and Z.-H. Zhou, "Learning with interpretable structure from gated rnn," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2267–2279, 2020.

[26] I. Spinelli, S. Scardapane, and A. Uncini, "A meta-learning approach for training explainable graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[27] D. Liu, X. Gao, C. Peng, N. Wang, and J. Li, "Heterogeneous face interpretable disentangled representation for joint face recognition and synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[28] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[29] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," *Advances in neural information processing systems*, vol. 32, 2019.

[30] J. Andreas, D. Klein, and S. Levine, "Modular multitask reinforcement learning with policy sketches," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 166–175.

[31] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4942–4950.

[32] W. Han, C. Zheng, R. Zhang, J. Guo, Q. Yang, and J. Shao, "Modular neural network via exploring category hierarchy," *Information Sciences*, vol. 569, pp. 496–507, 2021.

[33] G. Plumb, M. Al-Shedivat, Á. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar, "Regularizing black-box models for improved interpretability," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[34] E. Weinberger, J. Janizek, and S.-I. Lee, "Learning deep attribution priors based on prior knowledge," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 034–14 045, 2020.

[35] M. Wojtas and K. Chen, "Feature importance ranking for deep learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5105–5114, 2020.

[36] D. Vlahek and D. Mongus, "An efficient iterative approach to explainable feature learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[37] M. Wu, S. Parbhoo, M. Hughes, R. Kindle, L. Celi, M. Zazzi, V. Roth, and F. Doshi-Velez, "Regional tree regularization for interpretability in deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6413–6421.

[38] M. Wu, M. C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez, "Beyond sparsity: Tree regularization of deep models for interpretability," in *Association for the Advancement of Artificial Intelligence*, 2018.

[39] H. Liang, Z. Ouyang, Y. Zeng, H. Su, Z. He, S.-T. Xia, J. Zhu, and B. Zhang, "Training interpretable convolutional neural networks by differentiating class-specific filters," in *European Conference on Computer Vision*. Springer, 2020, pp. 622–638.

[40] Y. Li, R. Ji, S. Lin, B. Zhang, C. Yan, Y. Wu, F. Huang, and L. Shao, "Interpretable neural network decoupling," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 653–669.

[41] W. Shen, Z. Wei, S. Huang, B. Zhang, J. Fan, P. Zhao, and Q. Zhang, "Interpretable compositional convolutional neural networks," *arXiv preprint arXiv:2107.04474*, 2021.

[42] Z. Chen, Y. Bei, and C. Rudin, "Concept whitening for interpretable image recognition," *Nature Machine Intelligence*, vol. 2, no. 12, pp. 772–782, 2020.

[43] Q. Zhang, Y. Nian Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8827–8836.

[44] B. Herman, "The promise and peril of human evaluation for model interpretability," *arXiv preprint arXiv:1711.07414*, 2017.

[45] E. M. Kenny and M. T. Keane, "On generating plausible counterfactual and semi-factual explanations for deep learning," *AAAI-21*, pp. 11 575–11 585, 2021.

[46] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 7786–7795.

[47] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[48] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[49] N. Poerner, B. Roth, and H. Schütze, "Evaluating neural network explanation methods using hybrid documents and morphological agreement," *arXiv preprint arXiv:1801.06422*, 2018.

[50] J. Wu and R. J. Mooney, "Faithful multimodal explanation for visual question answering," *arXiv preprint arXiv:1809.02805*, 2018.

[51] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?" *arXiv preprint arXiv:2004.03685*, 2020.

[52] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3681–3688.

[53] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, "On the (in) fidelity and sensitivity of explanations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[54] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," *arXiv preprint arXiv:1908.04626*, 2019.

[55] Y. Kuramoto, "Self-entrainment of a population of coupled non-linear oscillators," in *International symposium on mathematical problems in theoretical physics*. Springer, 1975, pp. 420–422.

[56] C. Böhm, C. Plant, J. Shao, and Q. Yang, "Clustering by synchronization," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 583–592.

[57] D. Liu, S. Wang, J. Ren, K. Wang, S. Yin, and Q. Zhang, "Trap of feature diversity in the learning of mlps," *arXiv preprint arXiv:2112.00980*, 2021.

[58] 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.

[59] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[60] Q. Zhang, R. Cao, Y. N. Wu, and S.-C. Zhu, "Growing interpretable part graphs on convnets via multi-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[61] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[62] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

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

[65] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research*, vol. 3, no. Dec, pp. 583–617, 2002.

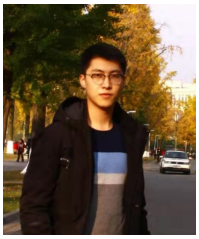
[66] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[67] X. V. Nguyen, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" in *ICML*, 2009.

[68] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," in *International Conference on Learning Representations*, 2016.

[69] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 6391–6401.

[70] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, "Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond," *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1–11, 2023.



Wei Han received his B.S. degree in Electronic and Electric Engineering in 2017, from the University of Electronic Science and Technology of China. He is currently a Ph.D. student at the University of Electronic Science and Technology of China. His main research interest is interpretable machine learning, transfer learning and adversarial examples.

Zhili Qin received his B.S. degree in Computer Science in 2017, from the Hefei University of Technology. He is currently a Ph.D. student at the University of Electronic Science and Technology of China. His main research interest is multi-label learning, few-shot learning and zero-shot learning.

Jiaming Liu received his B.S. degree in 2017 from the University of Electronic Science and Technology of China. He is currently a Ph.D. student at the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His main research interest is data stream classification, clustering and semi-supervised learning.

Christian Böhm received the Ph.D. degree, in 1998 and the habilitation degree, in 2001. He is currently a professor of computer science at Ludwig-Maximilians-Universität München, Munich, Germany. His research interests include database systems and data mining, particularly index structures for similarity search and clustering algorithms. He has received several research awards at top-tier data mining conferences.

Junming Shao received his Ph.D. degree with the highest honor ("Summa Cum Laude") at the University of Munich, Germany, in 2011. He became the Alexander von Humboldt Fellow in 2012. Currently, he is a professor of Computer Science at the University of Electronic Science and Technology of China. His research interests include data mining and neuroimaging. He not only published papers on top-level data mining conferences like KDD, ICDM, and SDM (three of those papers have won the Best Paper Awards), but also published data mining-related interdisciplinary work in leading journals including Brain, Neurobiology of Aging, and Water Research.