TSGB: Target-Selective Gradient Backprop for Probing CNN Visual Saliency

Lin Cheng, Pengfei Fang, Yanjie Liang, Liao Zhang, Chunhua Shen, and Hanzi Wang, *Senior Member, IEEE*

***Abstract*— The explanation for deep neural networks has drawn extensive attention in the deep learning community over the past few years. In this work, we study the visual saliency, a.k.a. visual explanation, to interpret convolutional neural networks. Compared to iteration based saliency methods, single backward pass based saliency methods benefit from faster speed, and they are widely used in downstream visual tasks. Thus, we focus on single backward pass based methods. However, existing methods in this category struggle to successfully produce fine-grained saliency maps concentrating on specific target classes. That said, producing faithful saliency maps satisfying both target-selectiveness and fine-grainedness using a single backward pass is a challenging problem in the field. To mitigate this problem, we revisit the gradient flow inside the network, and find that the entangled semantics and original weights may disturb the propagation of target-relevant saliency. Inspired by those observations, we propose a novel visual saliency method, termed *Target-Selective Gradient Backprop* (TSGB), which leverages rectification operations to effectively emphasize target classes and further efficiently propagate the saliency to the image space, thereby generating *target-selective* and *fine-grained* saliency maps. The proposed TSGB consists of two components, namely, TSGB-Conv and TSGB-FC, which rectify the gradients for convolutional layers and fully-connected layers, respectively. Extensive qualitative and quantitative experiments on the ImageNet and Pascal VOC datasets show that the proposed method achieves more accurate and reliable results than the other competitive methods. Code is available at https://github.com/123fxdx/CNNvisualizationTSGB**

***Index Terms*— Model interpretability, explanation, saliency map, CNN visualization.**

I. Introduction

IN RECENT years, deep convolutional neural networks (CNNs) have revolutionized various computer vision tasks, including object classification [1], [2], semantic segmentation [3], [4], low-level image processing [5], etc. However, human's knowledge on how deep models make decisions is still limited, which affects the trustworthiness of such a "Black Box" in the deep learning community. Moreover, this trustworthiness issue limits the development of real-world applications, e.g., autonomous driving [6] and medical diagnoses [7].

To interpret the working mechanism of deep neural networks, some explanation methods [8]–[12] have been developed to help humans understand what we can trust and how we can improve the networks. This paper studies the visual saliency [13], [14], w.r.t. the target classes, to explain how CNNs make decisions for given input images. The visual saliency, a.k.a. visualization, or visual explanation, aims to highlight important features, which highly contribute to the network predictions. In addition, visual saliency is also a useful technique for some downstream tasks, e.g., weakly-supervised vision [15], [16], person re-identification [17]–[19], knowledge distillation [20], etc.

In general, iteration based methods and single backward pass based methods are two dominant groups of methods to probe the visual saliency. Iteration based methods can localize the important regions in images by conducting iterative feed-forwards or backwards [21]–[25]. Such an approach is time-consuming, and it may introduce adversarial noise to saliency maps [22]. In contrast, single backward pass based methods offer the advantage of being computationally efficient, without introducing adversarial noise. To benefit from these properties, this work focuses on the single backward pass based method to study the visual saliency.

Among single backward pass based methods, many works, e.g., GradBP [13], GuidedBP [26], and FullGrad [27], have been proposed to exploit the gradient to generate saliency maps, where dominant objects of input images are highlighted. However, such methods often fail to focus on the target class, leading to inferior results w.r.t. the target category of interest. As shown in Fig. 1, GuidedBP produces two similar saliency maps, which cannot focus on the target class. Assume an extreme situation: an explanatory result for a selected target turns out target-agnostic, which is meaningless for the explanatory work. Other works, e.g., EBP [28] and GradCAM [29], attempt to leverage the top-down relevance or the weighted activation maps to produce class-discriminative explanations. However, they fail to backpropagate the saliency to the input image space, thereby resulting in coarse saliency maps.

Manuscript received April 14, 2021; revised January 2, 2022; accepted February 23, 2022. Date of publication March 11, 2022; date of current version March 21, 2022. This work was supported by the National Natural Science Foundation of China under Grant U21A20514 and Grant 61872307. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christophoros Nikou. *(Corresponding author: Hanzi Wang.)*

Lin Cheng, Liao Zhang, and Hanzi Wang are with the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: cheng.charm.lin@hotmail.com; leochang@stu.xmu.edu.cn; hanzi.wang@xmu.edu.cn).

Pengfei Fang is with the College of Engineering and Computer Science, Australian National University, Canberra, ACT 2601, Australia (e-mail: pengfei.fang@anu.edu.au).

Yanjie Liang is with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: liangyj@pcl.ac.cn).

Chunhua Shen is with the State Key Laboratory of CAD and CG, Zhejiang University, Hangzhou 310027, China (e-mail: chunhua@me.com).

Digital Object Identifier 10.1109/TIP.2022.3157149

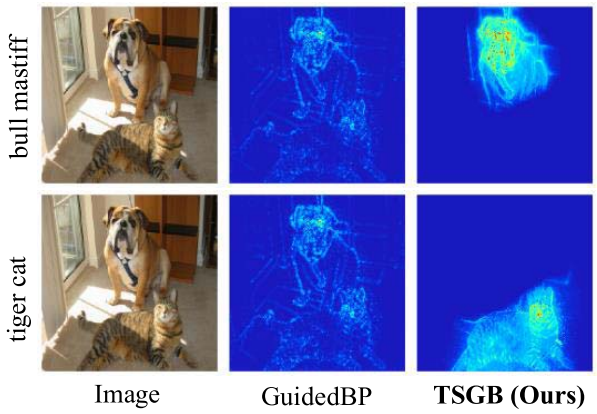


Fig. 1. Comparison of saliency maps w.r.t. target-selectiveness for the predictions of "bull mastiff" and "tiger cat". GuidedBP [13] produces two similar saliency maps, while the proposed TSGB produces the discriminative saliency maps.

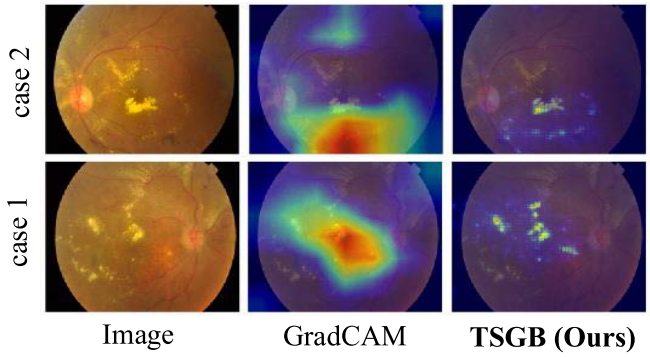


Fig. 2. Comparison of saliency maps w.r.t. fine-grainedness for the predictions of two cases of diabetic retinopathy. GradCAM [29] produces two coarse saliency maps, while the proposed TSGB can produce the fine-grained saliency maps.

Such coarse explanations are inadequate when fine-grained localization becomes a concern. For example, in the domain of medical image predictions, the fine-grained explanations are essential to discriminate the fine biological tissues [14], [30]. As shown in Fig. 2, GradCAM produces two coarse saliency maps, which cannot reveal the fine-grained patterns. Although many efforts have been made to study single backward pass based methods, developing an explanation approach that satisfies both class-level *target-selectiveness* and pixel-level *fine-grainedness* still needs further investigation.

In this paper, we attempt to address this challenging problem by taking a step back and rethinking the discipline of gradients inside neural networks. As noticed in the literature [31], [32], the network nodes in the intermediate layers may couple different semantic concepts. Interestingly, we find that even the final hidden layer before the output/prediction layer may contain entangled semantics. Such entangled nodes severely affect the target-selectiveness property when propagating the target attribution to the lower layers using gradients. On the other hand, we also observe that the saliency maps can be disturbed by the gradients using pre-trained parameters in convolutional layers. This impedes the attribution passing to the bottom to obtain fine-grained saliency maps.

Inspired by the above observations, we propose a novel visual saliency method, termed *Target-Selective Gradient Backprop* (TSGB), to generate target-specific and fine-grained saliency maps, which can explain how CNNs make decisions. The proposed TSGB consists of two modules, i.e., a target selection module for fully-connected (FC) layers and a fine-grained propagation module for convolutional (Conv.) layers. The target selection module exploits the contributions of sub-nodes to the target node, and emphasizes the negative connections by the ratio of positive contributions to negative contributions, which can disentangle the target class from the irrelevant classes and background in features. The fine-grained propagation module leverages the ratio of feature responses between two consecutive layers to propagate the visual saliency from the feature space to the image space.

The main contributions of this paper are summarized as follows:

- We study the influence of entangled semantics and original gradients on the backprop of visual saliency. Based on our findings, we propose a novel visual saliency method, i.e., TSGB, to explain CNNs' decisions. To our best knowledge, this is the first work to generate target-selective and fine-grained saliency maps in a single backward pass.
- We design a target selection module, i.e., TSGB-FC, for the backprop of FC layers. TSGB-FC adaptively enhances the negative connections inside the networks to make the visual saliency effectively focus on the target class.
- We devise a fine-grained propagation module, i.e., TSGB-Conv, for the backprop of Conv. layers and other advanced layers. TSGB-Conv exploits the information of feature maps rather than model parameters to efficiently produce high-resolution saliency maps.

Extensive experiments show the superiority of the proposed TSGB against the competitive methods in target-selectiveness, fine-grainedness, running speed, explanatory generalization, and faithfulness. Moreover, TSGB can be employed to diagnose the biases in the model and dataset. Furthermore, TSGB can be used to help human interpret the CNN model trained for medical diagnoses, and locate the critical biological structures.

The remainder of this paper is organized as follows: In Section II, some related works are described. In Section III, the factors disturbing the target-selectiveness and fine-grainedness during gradients backprop are analyzed. Based on the analysis, the proposed method, including the target selection module and the fine-grained propagation module, is presented in Section IV. In Section V, qualitative and quantitative experiments are conducted on various tasks to validate our method against the competitors. Conclusion and discussion are drawn in Section VI.

II. Related Work

A variety of saliency methods have been studied to interpret the decisions made by CNNs. Those methods can be categorized into two groups according to the number of processing of feedforward and backward, namely, single backward pass based methods as well as iteration based methods (i.e., multiple feedforward and backward pass based methods). We first

focus on discussing three kinds of single backward pass based methods in Section II-A. We then review several iteration based methods in Section II-B.

A. Single Backward Pass-Based Methods

1) Gradient Related Methods: GradBP [13] is one of the pioneering works for exploring visual saliency, which computes the gradient of the class score w.r.t. the input image to visualize the importance heatmap. Thereafter, GuidedBP [26] and Deconvolution [21] modify the backpropagated gradients, which makes the saliency maps sharper and clearer. Note that their explanation results fail to concentrate on the selected target [29], [33]. As the most recent work, FullGrad [27] improves the saliency maps by considering the multi-layer gradients aggregation.

2) Relevance Related Methods: Layer-wise Relevance Propagation [34] and Deep Taylor decomposition [35] explain the networks by decomposing the contribution of the target layer by layer. These methods pay attention to extensive existing objects, similar to [26]. Excitation Backprop (EBP) [28] uses the contrastive marginal winning probability to propagate the top-down attention. DeepLIFT [36] assigns the attribution by comparing the difference between the input and the reference data. CNN Fixation [8] measures the contributions between a pair of consecutive layers to uncover the pixel coordinates of saliency regions.

3) Activation Related Methods: CAM [12] and the generalized version GradCAM [29] utilize the gradient to weigh the feature maps to localize the important regions. This type of method is still the optimal one as noted in [37]. Guided GradCAM [29] ensembles GuidedBP and GradCAM, which actually needs more than one backward pass, and its target-selectiveness almost depends on GradCAM.

Despite that these single backward pass based methods are advanced in visual saliency, their explanatory results cannot satisfy the properties of target-selectiveness and fine-grainedness simultaneously.

B. Iteration-Based Methods

Another type of visual saliency method is based on iterations. Perturbation related methods, such as Occlusion [21], Meaningful Perturbation [22], RISE [23] and LIME [24], evaluate the output scores by occluding the input iteratively, which takes much running time and it is possible to introduce adversarial noise. Most recently, Score-CAM [38] masks the input according to the intermediate activation maps and repeatedly performs feedforwards N times (i.e., the number of activation maps) to obtain the importance scores. Optimization related methods, including Feedback [39] and FGVis [30], add the complex switch structure into the network and iteratively optimize the objective function to achieve the saliency maps. Some integration related methods, such as SmoothGrad [40], IntegratedGrad [41], and Integrated Grad-CAM [42], can be regarded as the ensembles over single propagation methods. We can also take advantages of these ensembles to improve our method. These iteration based methods are time-consuming and they do not achieve the optimal performance.

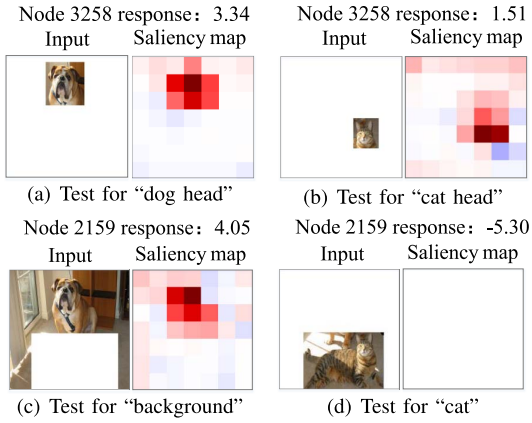


Fig. 3. Analysis of network nodes with entangled semantics. In each test, a feedforward is performed to obtain the node response. Then the saliency is backpropagated from the node to the top Conv. layer. Note that in (d), the backprop generates a blank map because of an inactivated state of the node, but we retain the negative response value for a better understanding.

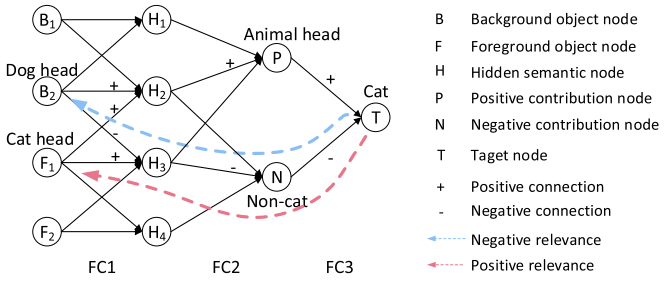
Compared to iteration based methods, single backward pass based methods run faster, and they are less likely to introduce the adversarial noise. Hence, we focus on investigating the single backward pass based visual saliency. Unlike all of these methods, our method rectifies the gradient backprop, which satisfies both the target-selectiveness and fine-graininess in a high-speed manner.

III. Analysis

During the procedure of generating the visual saliency, what exactly affects the selectiveness of the target in saliency maps? What disturbs the visual attribution when backpropagating saliency maps from the top layer to the low layer and what makes the visualized results rough rather than fine-grained? Driven by these two crucial questions, we attempt to explore the problem by revisiting the discipline of gradients inside the networks, as gradients indeed contain inherent and fundamental properties of the networks and they have been employed by many works [13], [26], [27], [29], [40], [41] for explanations.

A. Entangled Semantics in the FC Layer

In the following, we take the VGG16 model as an example. As shown in Fig. 4(a), one may intuitively think that the positive contribution nodes (with positive connections) to an output node "tiger cat" should encode the "cat" related semantic information, e.g., the cat head, the cat tail, etc. However, in practice, when testing on a positive contribution node (Fig. 3(a, b)), i.e., the *3258-th node* in the input of the FC3 layer, both "dog head" and "cat head" can activate the node. Meanwhile, the saliency regions with corresponding semantics are produced by the backprop from the node. Thus we naturally consider that positive contribution nodes encode entangled semantics, e.g., the "animal head", the range of which is even broader than that of the output node's semantics (Fig. 4(a)). When attributing the target class to the lower layers, passing gradients through these entangled nodes severely affects the target selection, as shown in Fig. 5(a) "Pool5".



(a) Conceptual diagram of node explanation

(b) Backprop from negative contribution nodes

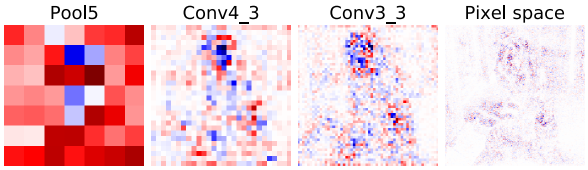
Fig. 4. An example of explanation for network nodes. The positive and negative relevance to the target are respectively marked in the red and blue colors[1]. An even number of negative connections make a positive contribution, depicted in the red dotted arrow.

On the other hand, a negative contribution node, i.e., the *2159-th node* in the input of the FC3 layer, is further tested, as illustrated in Fig. 3(c, d). We observe that the node's response value is negative for a cat being the input, whereas it is positive for a dog and background being the input. Thus, this node may encode the "non-cat" information. This suggests that the negative contribution is also important to help the network make a right decision. Furthermore, we surprisingly find that the negative contribution nodes in the FC3 layer contain the class information, as shown in Fig. 4. Specifically, using all final negative contribution nodes can result in a class-discriminative saliency map (Fig. 4(b)), which concentrates on the target and suppresses the background. The reasonable explanation is the transformation of gathering the connections with negative signs (i.e., even number of negatives make a positive). For example, the "cat head" is negatively relevant to the "non-cat" and the "non-cat" is negatively relevant to the "cat", leading that the "cat head" is positively relevant to the "cat" (Fig. 4(a)).

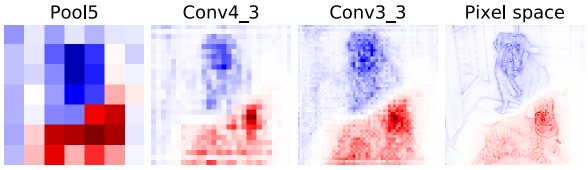
B. Backprop Noise in the Conv. Layer

As illustrated in Fig. 5(a), the gradient backprop generates a lot of noise, losing the target concentration. A similar result is also observed in [29], [40]. One possible reason can be explained as follows. The gradient can be regarded as an approximation to the importance score assigned to per feature. Conventionally, model parameters in the Conv. layers are trained for the feedforward to extract features. Here, in the procedure of the gradient backprop, directly using the original parameters to compute the saliency in convolutions (i.e., deconvolution operations) may introduce biases. This is more severe than the situation in the FC layers, because of dozens of local perceptions inside the convolutions. Moreover, the biases are accumulated layer by layer, leading to increasing noise

[1]This color setting can better distinguish the preserved negative values from positive values in the analysis, which differs from that in the experiment.



(a) Saliency maps of the original gradient backprop



(b) Saliency maps of TSGB

Fig. 5. Comparison of backprops using the original gradient and the proposed TSGB. Both propagations are from the target output node "tiger cat" down to low layers. We input the same image in Fig. 4(b) in this test.

along with the gradient backprop, which prevents achieving a fine-grained explanation.

IV. Methodology

Based on the above analysis, we propose a novel CNN visual saliency method, i.e., target-selective gradient backprop (TSGB), composed of a target selection module and a fine-grained propagation module, as shown in Fig. 6. For a pre-trained CNN model, the FC layers usually encode high-level semantic features related to the target classes, while the Conv. layers encode local features related to the object details. Given this prior knowledge, we design two modules of TSGB separately for the FC layers and Conv. layers. We will detail these two modules in the following.

A. Target Selection Module

According to the analysis of "entangled semantics", we propose a target selection module for the FC layers to select the target and suppress the target-irrelevant background.

Let g_i^l denote the target-selective gradient (TSG) of the i-th node in the l-th layer, and g_j^{l+1} denote the propagated gradient of the j-th node in the $(l+1)$-th layer. Additionally, the normal gradient $\tilde{g}_i^l = \sum_j w_{ij} g_j^{l+1}$ is given for reference. Firstly, given an input image and a pre-trained CNN, we perform a forward propagation and obtain the output scores before the softmax function. We set the initial gradient of the target node c in the output layer to 1, (i.e., $g_{j=c}^{l+1} = 1$), and set the rest nodes' initial gradients to 0, (i.e., $g_{j\neq c}^{l+1} = 0$). Then, we compute the TSG layer by layer in a top-down manner. In the final FC layer, the TSG of the lower layer g_i^l is calculated by enhancing the negative connection:

$$g_i^l = \sum_j (w_{ij}^+ + E_j(x^l, w) w_{ij}^-) g_j^{l+1}, \tag{1}$$

where w_{ij} is the connection weight, and $w_{ij}^+ = \mathrm{ReLU}(w_{ij})$, $w_{ij}^- = w_{ij} - w_{ij}^+$. Let x_i^l denote the feature of the i-th node in the l-th layer. The enhancement factor $E_j(x^l, w)$ is obtained

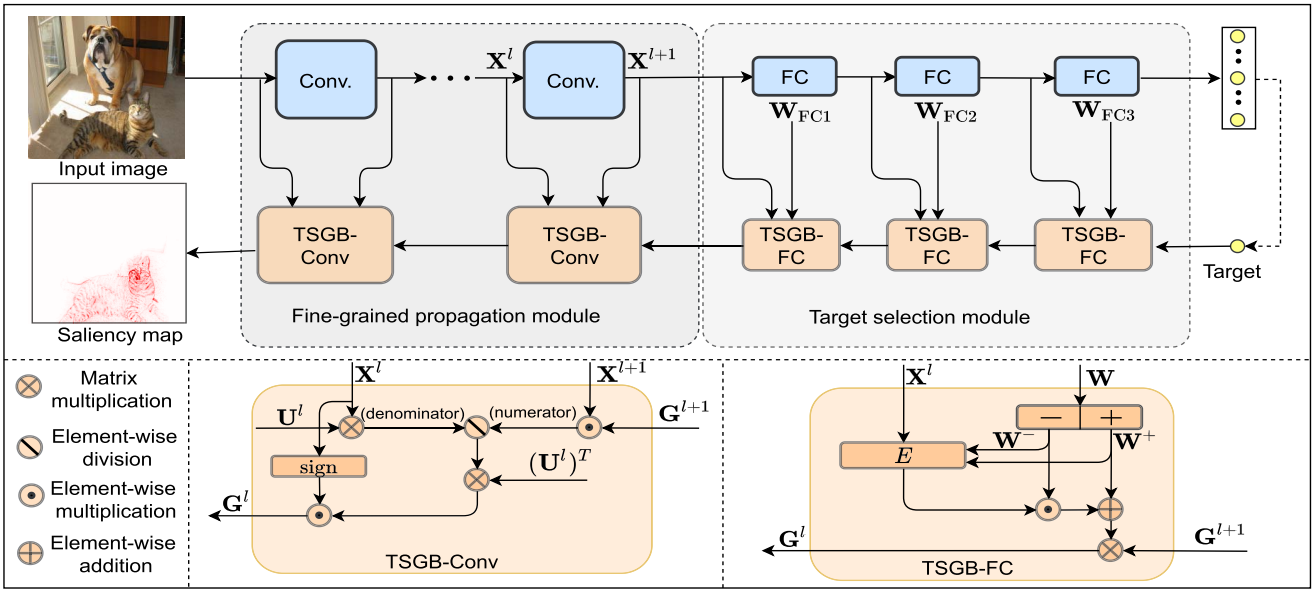


Fig. 6. The pipeline of the proposed target-selective gradient backprop (TSGB). Here, we use the VGG network as an example.

by the ratio of positive contributions to negative contributions:

$$E_j(x^l, w) = \alpha \frac{\sum_i x_i^l w_{ij}^+}{\sum_i |x_i^l w_{ij}^-|}. \tag{2}$$

When the positive contribution is larger, or the negative contribution is smaller, the relative entangled strength $\sum_i x_i^l w_{ij}^+ / \sum_i |x_i^l w_{ij}^-|$ will be larger, thereby leading to a larger ratio. α is a positive scale coefficient, which adjusts the enhancement ratio. It can be deduced that the ratio $\sum_i x_i^l w_{ic}^+ / \sum_i |x_i^l w_{ic}^-|$ for the target c is always larger than 1 if the output of the target c is positive, as $(\sum_i x_i^l w_{ic}^+ - \sum_i |x_i^l w_{ic}^-|) > 0$. However, if the ratio is much larger than 1, it may result in too strong suppression for the foreground objects. Thus, we use the scale coefficient to slightly adjust the enhancement ratio.

Note that we only rectify the gradients in the final FC layer, and calculate the gradients with the original weights, i.e., $E = 1$, for the other FC layers if there exist, such as in the VGG net. This is because that the other FC layers' information is integrated into the final layer's input, which is included in Eq. (2). Different from EBP [28] which only uses positive weights, we make use of both positive and negative weights for the other FC layers, as both of them are necessary for the whole module to select the target and suppress the background. For example, we use the proposed module to produce the saliency map of "Pool5" layer, which is the input layer of FC layer. As shown in Fig. 7, we can observe that the target is gradually disentangled from the irrelevant classes and background when the enhancement factor increases.

B. Fine-Grained Propagation Module

In this subsection, we further propose a fine-grained propagation module for the Conv. layers to efficiently propagate the saliency to the input image space.

In the Conv. layers, there exists the local perception over each space location, which is different from the FC layers.

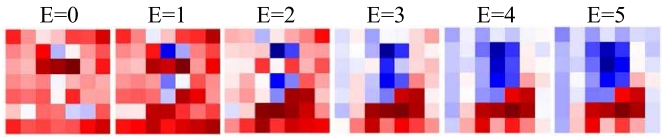


Fig. 7. The influence of the enhancement factor E. The saliency maps are from the "Pool5" layer for the target "tiger cat" as in Fig. 5.

Considering this difference, we implement the backprop of Conv. layers differently. The TSG of the lower layer g_i^l in the Conv. layers is devised as

$$g_i^l = \text{sign}(x_i^l) \sum_j \frac{x_j^{l+1} u_{ij}}{\sum_i |x_i^l u_{ij}|} g_j^{l+1}, \tag{3}$$

where $u_{ij} = 1$ if x_i^l is inside the receptive field of x_j^{l+1}, and 0 otherwise. The denominator is actually the convolution operation with the kernel, each of whose elements is 1. $\text{sign}(\cdot)$ is the sign function. We leverage the information of feature maps rather than model parameters to propagate the saliency map to the pixel space. This is because that feature maps are dependent on the input instance, while model parameters are input-agnostic. Feature maps are more accurate for assigning the importance score per feature for a specific instance during propagation. As Eq. (3) shows, given an identical input feature, a larger output response indicates the stronger relevance of the input feature to the output feature, leading to a larger TSG. Note that although no model parameters are explicitly included in the equation, the TSG is related to model parameters as well. Actually, the computation of the feature x_j^{l+1} is determined by model parameters, which are implicitly contained in Eq. (3).

Eq. (3) can be rewritten in a tensor form. Let $\mathbf{X}^l \in \mathbb{R}^{M \times H_l \times W_l}$ and $\mathbf{X}^{l+1} \in \mathbb{R}^{N \times H_{l+1} \times W_{l+1}}$ denote the feature maps in the l-th and $(l+1)$-th layer, respectively. M and N are the channel numbers of $\mathbf{X}^l$ and $\mathbf{X}^{l+1}$, respectively. $\mathbf{U}^l \in \mathbb{R}^{M \times N \times K_h \times K_w}$ is a set of defined Conv. kernels with the spatial size of $K_h \times K_w$ in the l-th layer ($\mathbf{U}^l$ has the same

dimension as the original weight). $\mathbf{G}^l$ and $\mathbf{G}^{l+1}$ are the TSG maps in the l-th and $(l+1)$-th layers, respectively. The m-th map $\mathbf{G}^l_m \in \mathbb{R}^{H_l \times W_l}$ is formulated as

$$\mathbf{G}^l_m = \frac{\mathbf{X}^{l+1} \odot \mathbf{G}^{l+1}}{|\mathbf{X}^l| * \mathbf{U}^l} * (\mathbf{U}^l_m)^T \odot \ \text{sign}(\mathbf{X}^l_m), \tag{4}$$

where $\odot$, $*$, and $|\cdot|$ denote the element-wise multiplication, convolution operation, and element-wise absolute value operation, respectively. In our formulation, all elements in $\mathbf{U}$ are ones.

Let $\mathbf{u} \in \mathbb{R}^{K_h \times K_w}$ denote a single channel of Conv. kernel in $\mathbf{U}$. To speed up the computation, we further obtain the following derivation from Eq. (4):

$$\begin{aligned}\mathbf{G}^l_m &= \Big(\sum_{n=1}^{N} \frac{\mathbf{X}_n{}^{l+1} \odot \mathbf{G}_n{}^{l+1}}{|\mathbf{X}^l| * \mathbf{U}^l}\Big) * (\mathbf{u}^l)^T \odot \ \text{sign}(\mathbf{X}^l_m) \\ &= \frac{\sum_{n=1}^{N} \mathbf{X}_n{}^{l+1} \odot \mathbf{G}_n{}^{l+1}}{\sum_{m=1}^{M} |\mathbf{X}^l_m| * \mathbf{u}^l} * (\mathbf{u}^l)^T \odot \ \text{sign}(\mathbf{X}^l_m). \end{aligned} \tag{5}$$

Note that in Eq. (4), each channel in $\mathbf{U}^l_m$ is equal, leading to obtaining the first line in Eq. (5). Similarly, considering the first line in Eq. (5), each channel in $\mathbf{U}^l$ is equal leading to each channel in the result of $|\mathbf{X}^l| * \mathbf{U}^l$ equal, and thereby obtaining the second line. By this transformation, the convolution operation is turned from multiple channels to one channel. Here, we further analyze the computation complexity of the equation.

1) Computation Complexity: For convenience, we ignore the difference between the multiplication and addition operations, as well as the difference of space scale between input and output layers. The computation complexity of Eq. (4), depends on the term $|\mathbf{X}^l| * \mathbf{U}^l$, and thereby the computation complexity is $O(M \times N \times H \times W \times K \times K)$. On the other hand, the computation complexity of the second line in Eq. (5) depends on the term $\sum_{m=1}^{M} |\mathbf{X}^l_m| * \mathbf{u}^l$, with the computation complexity being $O(M \times H \times W \times K \times K)$. Thus, the transformation of Eq. (5) reduces the computation cost N times for $\mathbf{G}^l_m$, and $M \times N$ times for $\mathbf{G}^l$.

2) Other Layers: We formulate the backprop of Normalization layer, including the Batch Normalization layer and the Local Response Normalization layer, as

$$g^l_i = \frac{x^{l+1}_j}{x^l_i} g^{l+1}_j. \tag{6}$$

Eq. (6) is also utilized for the backprop of a type of Average Pooling layer whose input features contain negative values, such as in the case of DenseNet. Otherwise, we directly use the original gradient operations for the other layers in CNNs, including ReLU, Max Pooling, Adaptive Pooling, Skip Connection, Concat layer, and common Average Pooling layer, etc.

As shown in Fig. 5(b), this fine-grained propagation module can effectively deliver the TSG to the image space to generate high-resolution saliency maps, meanwhile keeping the target concentration. Note that the TSG can be propagated to any layer inside the network to analyze the attributions of channels of interest according to different demands of semantic levels and spatial scales.

V. Experiments

In this section, we first qualitatively validate the proposed TSGB via visual comparisons. Then in quantitative experiments, we evaluate the proposed TSGB with weakly-supervised localization tasks on the ImageNet dataset [43] and the Pascal VOC dataset [44]. Furthermore, we evaluate the faithfulness of the explanations with pixel perturbation [23] and sanity check [45]. Finally, we perform the bias diagnosis, the medical image test and the ablation study.

We compare our method with several other competitors including GradBP [13], GradCAM [29], DeepLIFT [36], [46], EBP [28], FullGrad [27] and Fixation [8]. These competitors are the state-of-the-art saliency methods in the single backward pass type, which is consistent with the type of our method. For our method, we set the scale coefficient to 0.5∼1.3 for the negative enhancement in Eq. (2). More details can be found in Section V-F. We follow the processing in [27] to obtain final saliency maps by first multiplying the produced target-selective gradients to feature maps, and then summing all the elements along the channel dimension.

A. Visual Comparison

1) Comparison on Different Samples: We employ TSGB to generate saliency maps from different targets on different samples in comparison with the other competitors. As shown in Fig. 8, GradBP and DeepLIFT generate noisy maps, which highlight most foreground objects, even including some target-irrelevant objects. FullGrad focuses on the most dominant objects rather than the target. Fixation only generates almost the same saliency maps w.r.t. different targets of each image. One reasonable explanation for Fixation is that the backprop in the FC layers neglects the negative connections, leading to the lack of the target-selectiveness. GradCAM and EBP can produce class-discriminative maps. However, their results still contain irrelevant backgrounds, especially on the borders of images, such as in the cases of "goldfish", "cabinet", "cheetah" and "zebra". It is also worth mentioning that the generated saliency maps from FullGrad, GradCAM, and EBP are coarse. In contrast, TSGB can produce target selective and fine-grained maps with clear targets' boundaries and fewer irrelevant backgrounds. Furthermore, the explanatory results of TSGB are more human interpretable, when compared to its competitors.

2) Comparison on Different Models: To verify the generalization of the proposed TSGB, we further conduct the experiments across various CNN models, along with the competitors. From Fig. 9, we can find that over several cases EBP can produce saliency maps with fewer background areas than GradBP, DeepLIFT and FullGrad, while it totally fails on DenseNet121 and MobileNetV2. The main reason is that the features in DenseNet121 and MobileNetV2 contain negative values, which affects the robustness of EBP. GradCAM is valid for these tested models, while its results cannot discriminate the borders of targets precisely. Moreover, GradCAM also fails like EBP if the gradients are backpropagated to the low layer [29]. Unlike these competitors, TSGB shows its advantage of being target-selective, fine-grained, and robust

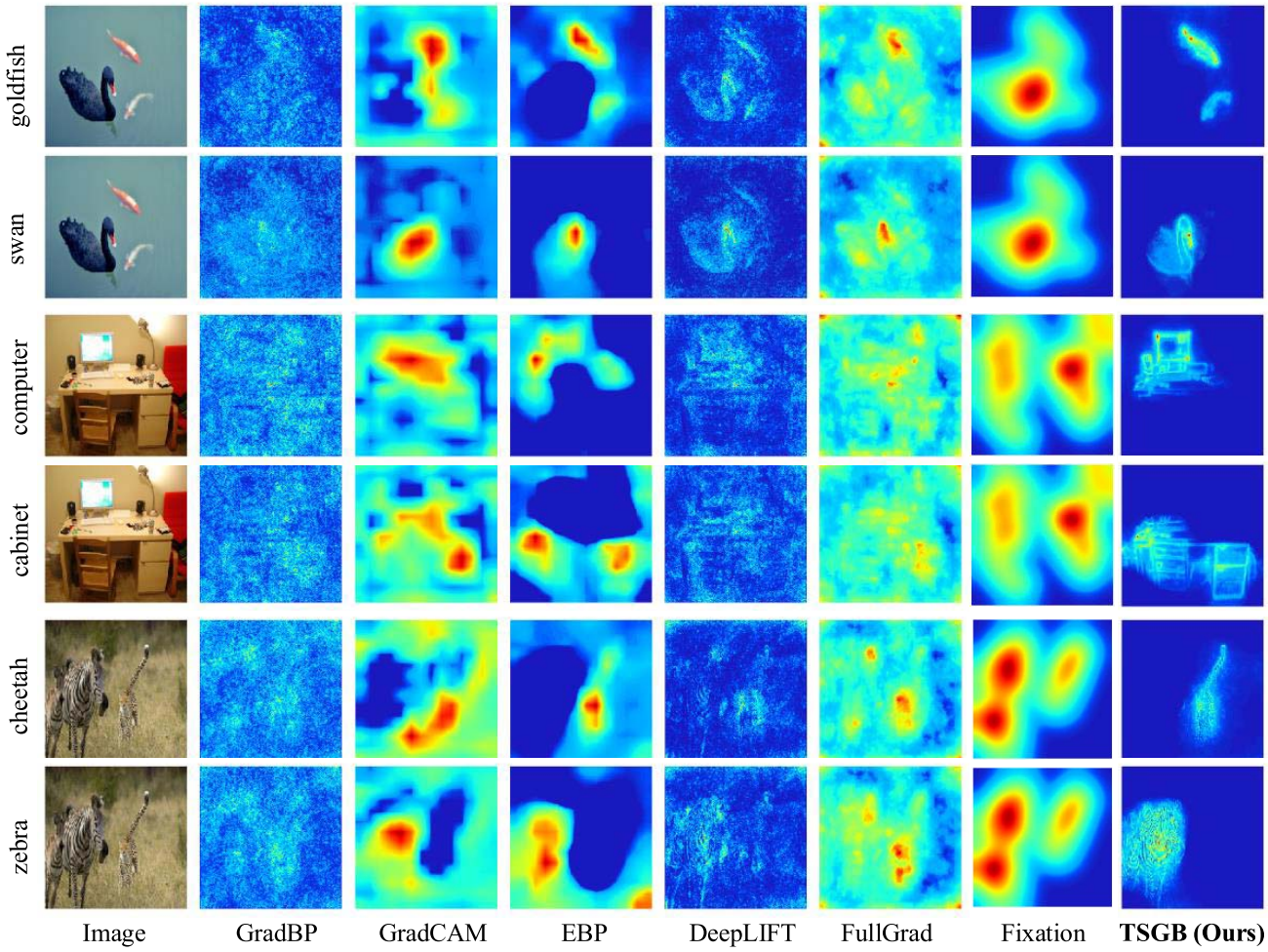


Fig. 8. Comparison of different methods on different samples. The saliency maps are generated from different targets, annotated on the left side, in each sample. The deep blue color represents the background, and all other colors represent varying degrees of the target evidence. The negative values are truncated for better contrast.

for extensive models, even for the models containing negative-value features (i.e., DenseNet121, MobileNet, etc.). In addition, we find that saliency maps generated on ResNet50 and VGG16 are better than the other models. The target saliency on MobileNetV2 is relatively blurry when compared to the other networks. One possible reason is that the computation-efficient model cannot learn good features as discriminative as other conventional models. Since the official code of Fixation does not support the models of ResNet50, ResNeXt, DenseNet and MobileNet, we omit the evaluation of Fixation on these models.

B. Weakly-Supervised Localization

1) Object Localization: A satisfactory saliency method is expected to generate target-relevant saliency maps, where the areas with high intensity indicate the positions of targets. Following [28], [29], [48], we evaluate the visual saliency methods with the weakly-supervised object localization task on the ImageNet dataset using the VGG16 and ResNet50 models, which are pre-trained with the classification labels.

On the ImageNet 2012 validation (val) set, we first predict categories, and then use saliency methods to generate the saliency maps. The top-5 localization (LOC) error is evaluated under the protocol of the ILSVRC challenge [43].

TABLE I

LOCALIZATION ON THE IMAGENET VAL SET (LOWER IS BETTER). ERROR RATES OF GRADBP, GRADCAM, AND EBP FOR VGG16 ARE TAKEN FROM [29]. DEEPLIFT REFERS TO THE "CAPTUM" PACKAGE IN PYTORCH1.4 [47]. FIXATION REFERS TO THE OFFICIAL CODE IN [8]

Method	VGG16		ResNet50	
	Top5 LOC error (%)	FPS	Top5 LOC error (%)	FPS
GradBP [13]	51.46	25.64	55.44	**34.19**
GradCAM [29]	46.41	32.26	40.73	29.63
DeepLIFT [36]	55.32	7.12	53.11	17.78
EBP [28]	63.04	23.26	44.44	26.67
FullGrad [27]	47.82	12.05	46.35	9.93
Fixation [8]	58.38	0.39	-	-
TSGB (Ours)	**43.46**	**43.48**	**40.49**	31.25

After achieving the saliency maps, we search the best performing thresholds for different methods and binarize the saliency maps with the selected thresholds to obtain the bounding boxes. Besides the binarization, we do not append any other post-processing techniques to our method. As shown

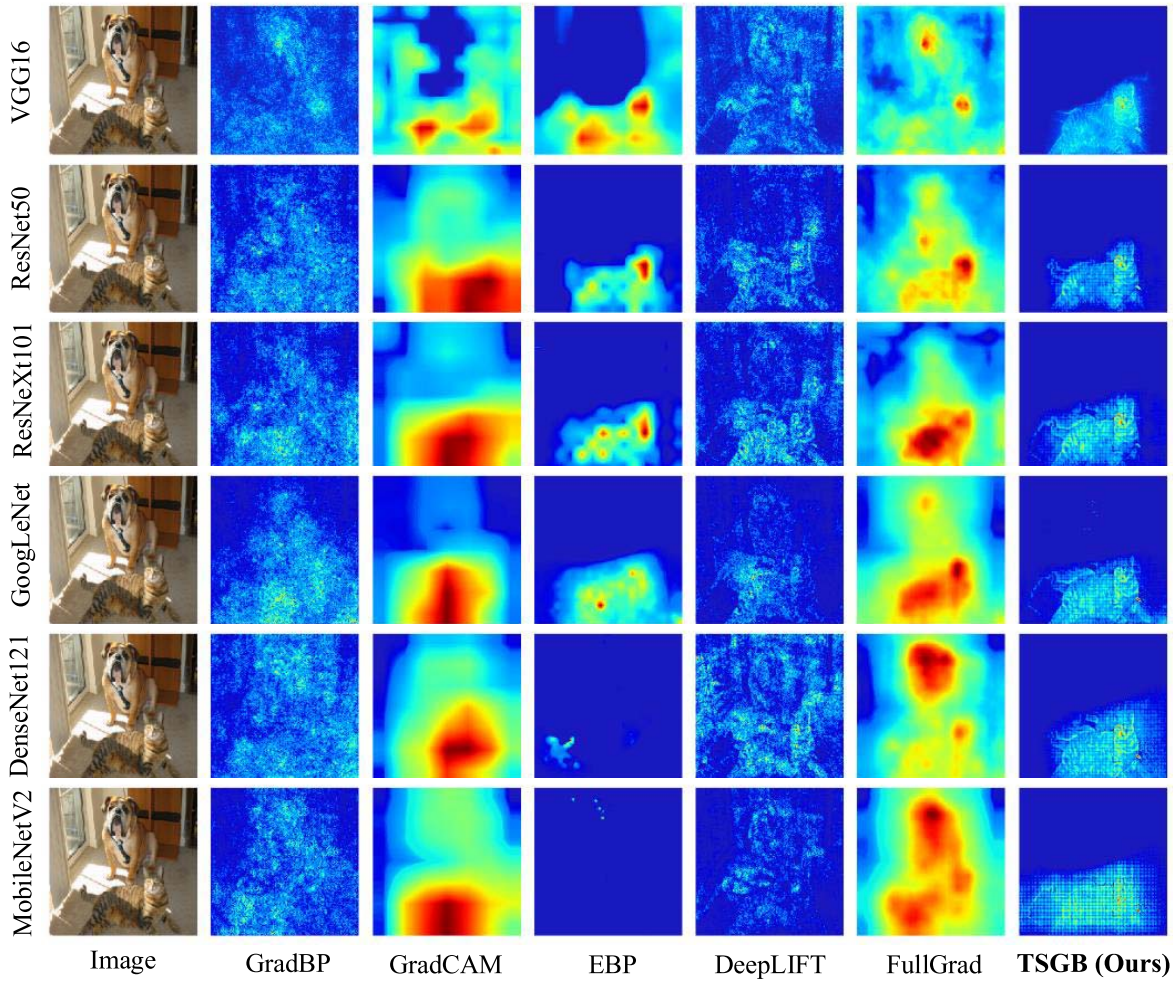


Fig. 9. Comparison of different methods on different models. The models' names are annotated on the left side. The saliency maps are generated from the same target, i.e., "tiger cat."

in Table I, TSGB outperforms the other methods in localization errors both explained models. For example, the results of TSGB are 43.46/40.49, as compared to 46.41/40.73 of the second-best method, i.e. GradCAM, on VGG16/ResNet50. Note that GradCAM additionally applies the post-processing technique, i.e., searching for the largest connected component after binarization. Compared to ResNet, the VGG model has more FC layers, which are possible to involve the stronger entanglement as stated in Section III. In this situation, using TSGB to disentangle the semantics in the FC layers will boost the performance of VGG. TSGB outperforms FullGrad by 4.36%/5.86% on VGG16/ResNet50. FullGrad aggregates all Conv-layer saliency maps to improve the performance but it consumes much more computation memory. We find that most methods achieve lower error rates on ResNet50 than VGG16. This is likely owing to the higher classification capacity of ResNet50, leading to better localization performances.

Moreover, we test the average running speed on a GeForce GTX 1080 Ti GPU. The proposed TSGB achieves the highest speed at 43 frames per second (FPS), which is 6 times faster than the DeepLIFT (7 FPS). Note that Fixation does not support GPU computation in its backprop, resulting in the slow running speed.

2) Point Localization: Considering that the explanatory results intend to focus on the most discriminative regions of targets, we use another popular evaluation metric, Pointing Game [28], to measure the explanatory results. This metric is defined as the ratio of hits, where a hit is counted if the maximum point of the saliency map is inside the target region. As shown in Table II, our method achieves the superior performance over the other methods on the Pascal VOC2007 test set, which can be attributed to the target-selectiveness of TSGB. GradBP and Fixation achieve much lower accuracy. This is probably because that the point localization of GradBP is easily interfered by the noise, and Fixation cannot focus on the target class, which is consistent with the visual comparison experiments (see Fig. 8).

C. Faithfulness Check

1) Pixel Perturbation: In order to evaluate the faithfulness of explanatory results at pixel level, we use the deletion metric [23] to test TSGB. The intuition behind this metric is that if the saliency region is responsible for the model prediction, the prediction probability will descend when erasing the corresponding region. This protocol is to measure the decline in prediction probability of classification when

TABLE II

POINTING GAME ON THE VOC2007 TEST SET (HIGHER IS BETTER). THE RESULTS OF GRADBP, GRADCAM, AND EBP ARE TAKEN FROM [28]

	VGG16		ResNet50	
Method	Mean accuracy (%)	FPS	Mean accuracy (%)	FPS
GradBP [13]	76.00	18.18	65.80	**16.95**
GradCAM [29]	86.60	**18.52**	90.60	16.25
DeepLIFT [36]	79.05	7.69	82.72	3.12
EBP [28]	80.00	10.41	89.20	6.31
FullGrad [27]	84.16	8.56	88.99	3.14
Fixation [8]	74.52	0.44	-	-
TSGB (Ours)	**89.33**	18.18	**90.68**	11.82

TABLE III

PIXEL PERTURBATION ON THE VOC2012 VAL SET (LOWER IS BETTER). LOWER DELETION SCORE MEANS HIGHER FAITHFULNESS OF SALIENCY METHODS

Method	Deletion score	FPS
GradBP [13]	0.1932	10.10
GradCAM [29]	0.1680	10.10
DeepLIFT [36]	0.1621	1.42
EBP [28]	0.5028	5.41
FullGrad [27]	0.1667	4.78
TSGB (Ours)	**0.1564**	**14.29**

iteratively perturbing the important pixels according to the rank of saliency values generated by a saliency method. The steeper the decline (i.e., the lower deletion score) is, the more reliable the saliency method is. As shown in Table III, TSGB achieves the lowest score, which suggests TSGB is the most faithful to the model predictions and capable of capturing the fine-grained details corresponding to the targets.

2) Sanity Check: As suggested by [45], we conduct the sanity check for the proposed TSGB to validate whether the explanatory results are sensitive to the model parameters or not. If the explanatory results are similar before and after the model parameters are randomized, the corresponding saliency method is more risky in trustworthiness. We evaluate the similarity with Spearman rank correlation before and after the randomization of model parameters for our TSGB and the other comparative methods, including GuidedBP [26] for reference. As illustrated in Fig. 10, TSGB and GradCAM are sensitive to the change of the parameter values while GuidedBP is much independent of model parameters.

D. Diagnosing Bias and Failure Cases

We adopt TSGB to diagnose the biases in the VGG16 network pre-trained on ImageNet. As shown in Fig. 11, the man is recognized as "basketball" (top left), and the station is recognized as "train" (bottom left), which makes it difficult to catch the failure clues by only knowing the prediction possibilities. Fortunately, with the help of target-selective and fine-grained saliency maps generated by TSGB, one can easily understand the reason why the model makes such decisions. For example, the "basketball" class is predicted by seeing the sports suit, and the "train" class is predicted by seeing the rail. One reasonable explanation of model biases is that co-occurring objects, e.g., sports suit and basketball, rail and train, exist in the training dataset. For the right two cases in Fig. 11, our method fails to produce the target-specific visualized maps,

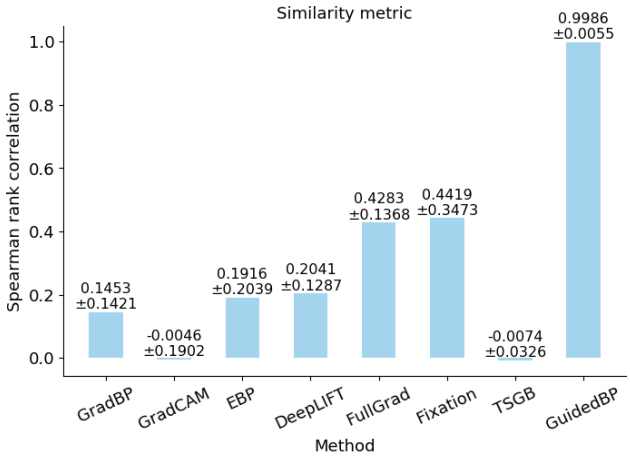


Fig. 10. Sanity check with similarity metric for model randomization. Spearman rank correlation is taken as the similarity metric. The values above the bar are the means and standard deviations of similarities between the original explanations and the randomized explanations on ImageNet. Lower similarity denotes better faithfulness of explanations.

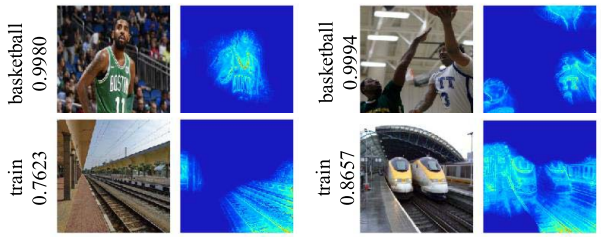


Fig. 11. Diagnosing bias and failure cases. The texts denote the predicted targets and possibilities. TSGB can help diagnose the biases in the model and dataset without suppressing the useful information, even in the background.

where some relevant backgrounds are not suppressed. This is because that these backgrounds are involved in the model predictions of the target classes. This also suggests that TSGB is faithful to the model.

E. Explanation for Medical Images

To test the generalization of TSGB on the different types of images, we use TSGB to explain the deep neural model trained on the medical image dataset, i.e., the Kaggle Diabetic Retinopathy dataset. The images in this dataset contain various texture features and color features, which are non-object-like features. Thus there is a big domain gap between the Kaggle Diabetic Retinopathy dataset and the ImageNet dataset. The explained model is ResNet152 trained on the Kaggle Diabetic Retinopathy dataset with image-level labels. It took around more than 100 epochs to train this model to achieve 97% accuracy for classification. As shown in Fig. 12, TSGB obtains more reliable explanatory results than the other competing methods. Benefiting from the target-selectiveness, TSGB can focus on the disease-relevant regions. More importantly, with the property of fine-grainedness, TSGB can effectively highlight the detailed patterns in the medical images.

F. Ablation Study

1) Target Selection Module vs. Fine-Grained Propagation Module: We compare the proposed target selection module

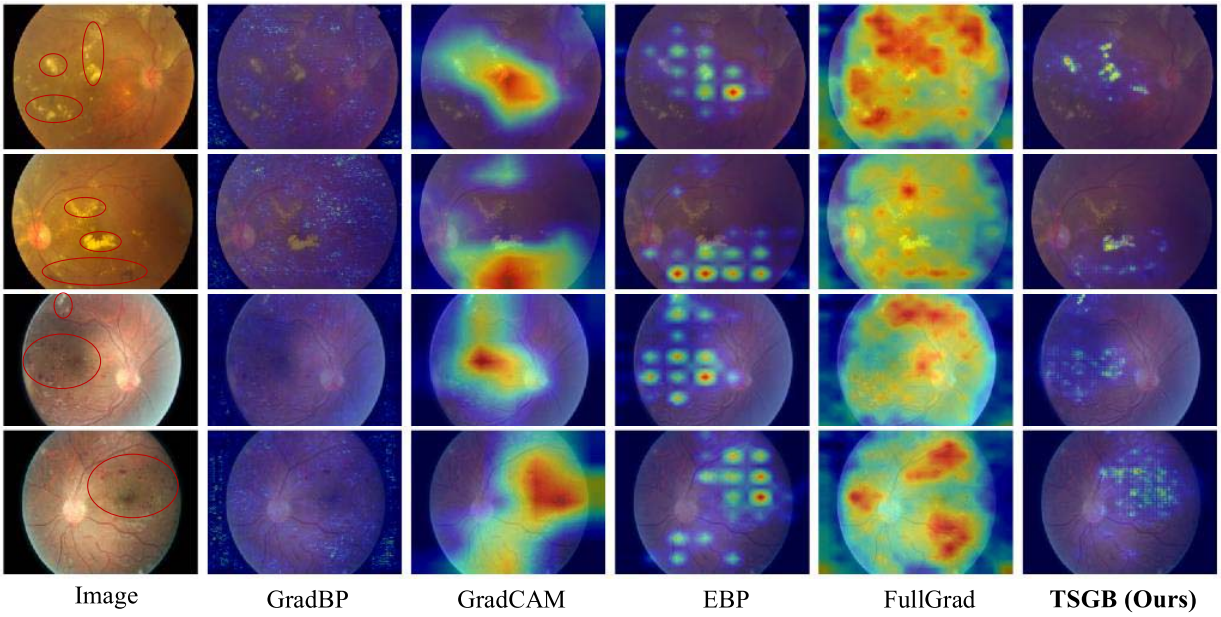


Fig. 12. Explanation for several medical images. We compare the proposed TSGB with GradBP [13], GradCAM [29], EBP [28], and FullGrad [27]. The red ovals denote the lesions of retinas.

TABLE IV

ABLATION STUDY FOR TSGB (LOWER IS BETTER). "GRAD" DENOTES THE VANILLA GRADIENT BACKPROP. "TSGB-FC" DENOTES THE TARGET SELECTION MODULE."TSGB-CONV" DENOTES THE FINE-GRAINED PROPAGATION MODULE

Methods	TSGB-FC	TSGB-Conv	LOC error (%)
Grad			52.99
TSGB-FC+Grad	✓		48.60
Grad+TSGB-Conv		✓	51.70
TSGB (Ours)	✓	✓	**43.46**

with the proposed fine-grained propagation module via ablation study on the ImageNet localization task. We choose the vanilla gradient backprop as our baseline. As shown in Table IV, when replacing the vanilla gradient backprop in the FC layers with the target selection module, the LOC error achieved by TSGB-FC+Grad becomes 4.39% lower. When replacing the vanilla gradient backprop in the Conv. layers with the fine-grained propagation module, the LOC error achieved by Grad+TSGB-Conv becomes 1.29% lower. Although the fine-grained propagation module seems less useful, when we further append the fine-grained propagation module to the target selection module, the LOC error continues to decrease, i.e., 5.14% lower than that obtained by TSGB-FC+Grad. This shows that both of the proposed modules are necessary to TSGB and provide complementary benefits to TSGB. Those two modules are tied tightly in one framework, achieving better performance than only using a single module.

2) Fine-Grained Propagation Module vs. Edge Detector: Benefiting from the fine-grained propagation module, the saliency maps can highlight clear details relevant to the targets, such as the examples in Fig. 8. When we propagate the saliency maps to the pixel level, the visualizations appear more edges-like patterns, since the low layers in CNN are inclined to extract edge features and other low-level features from the

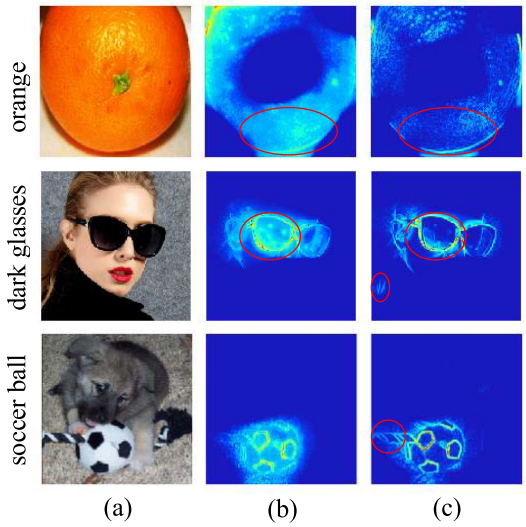


Fig. 13. Influence of the fine-grained propagation module and the edge detector. (a) Input images. (b) Saliency maps generated by TSGB. (c) Saliency maps generated by replacing of the fine-grained propagation module in TSGB with the edge detector.

images. Nevertheless, the fine-grained propagation module is not equivalent to an edge detector. We replace the fine-grained propagation module with the edge detector and evaluate the new setting with the pixel perturbation experiment, which turns out 11.06% worse. Compared with the edge detector, the fine-grained propagation module can also highlight the texture and color patterns, besides the edge patterns, such as the "orange" and the "dark glasses" in Fig. 13. Moreover, the fine-grained propagation module can refine the details in coarse saliency maps in a top-down manner, such that it further suppresses the irrelevant object parts, such as the hair tail in the "dark glasses" and the bar in the "soccer ball" in Fig. 13. In addition, the

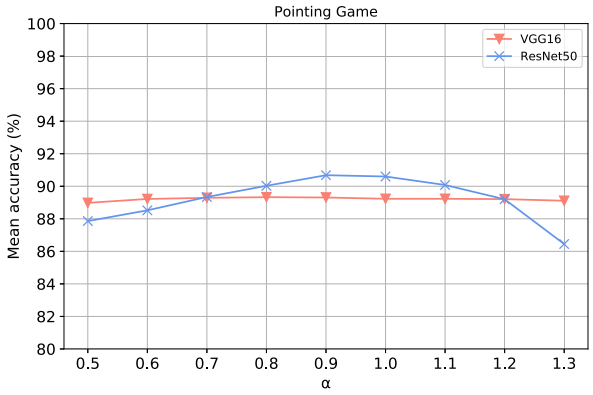


Fig. 14. Influence of different values of the scale coefficient α on the performance of TSGB.

fine-grained propagation module can propagate the saliency maps to different semantic levels at different spatial scales, and it can analyze the attributions of different features channels.

3) Influence of Scale Coefficient: To analyze the influence of choosing different values of the scale coefficient α in Eq. (2), we test the proposed TSGB with Pointing Game, as mentioned in "Point localization" in Section V-B. We record the experimental results corresponding to varied $\alpha \in [0.5 : 0.1 : 1.3]$ both on the VGG16 and ResNet50 models. As shown in Fig. 14, we can find that each model has one peak mean accuracy as α varies from 0.5 to 1.3, where VGG16 obtains the best result at $\alpha = 0.8$ and ResNet50 obtains the best result at $\alpha = 0.9$. Furthermore, the results on VGG16 are less sensitive to the scale coefficient α than those on ResNet50. Specially, the accuracy on the VGG16 model fluctuates within a very small extent, i.e., 0.35, in the whole range of α. On both models, there is less fluctuation in the mean accuracy for $\alpha \in [0.6, 1.2]$.

VI. Conclusion and Discussion

To probe the CNN visual saliency, we propose a novel saliency backprop method, i.e., target-selective gradient backprop (TSGB), which consists of a target selection module and a fine-grained propagation module. The target selection module adaptively enhances the negative connections to disentangle the target class from the irrelevant classes and background. The fine-grained propagation module leverages the information of feature maps to propagate the visual saliency and produces high-resolution saliency maps. Qualitative experiments show that TSGB can more discriminately explain different targets and generate clearer saliency maps than the competitive methods. Moreover, TSGB can be used for most of the CNN models. Quantitative experiments reveal that TSGB achieves superior localization performance, and stronger reliability over the competitive methods. Furthermore, we also validate that TSGB is faithful to the explained models.

Note that this explanatory work is mainly based on the visual aspect, as it is difficult to establish a set of rigorous mathematical explanations. We leave the theoretical study for the future research.

References

[1] L. Song *et al.*, "A deep multi-modal CNN for multi-instance multi-label image classification," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 6025–6038, Dec. 2018.

[2] Y. Ding *et al.*, "AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2826–2836, 2021.

[3] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic segmentation with context encoding and multi-path decoding," *IEEE Trans. Image Process.*, vol. 29, pp. 3520–3533, 2020.

[4] L. Jing, Y. Chen, and Y. Tian, "Coarse-to-fine semantic segmentation from image-level labels," *IEEE Trans. Image Process.*, vol. 29, pp. 225–236, 2020.

[5] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[6] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[7] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "MDNet: A semantically and visually interpretable medical image diagnosis network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3549–3557.

[8] K. R. Mopuri, U. Garg, and R. V. Babu, "CNN fixations: An unraveling approach to visualize the discriminative image regions," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2116–2125, May 2019.

[9] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 9273–9282.

[10] S. Wickramanayake, W. Hsu, and M. Lee, "FLEX: Faithful linguistic explanations for neural net based model decisions," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 2539–2546.

[11] J. Sun, X. Cao, H. Liang, W. Huang, Z. Chen, and Z. Li, "New interpretations of normalization methods in deep learning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5875–5882.

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[13] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Int. Conf. Learn. Represent.*, 2014.

[14] Z. Fang, K. Kuang, Y. Lin, F. Wu, and Y.-F. Yao, "Concept-based explanation for fine-grained images and its application in infectious keratitis classification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 700–708.

[15] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6488–6496.

[16] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5267–5276.

[17] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive Siamese networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5735–5744.

[18] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1389–1398.

[19] P. Fang, J. Zhou, S. Roy, L. Petersson, and M. Harandi, "Bilinear attention networks for person retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8030–8039.

[20] P. Dhar, R. V. Singh, K.-C. Peng, Z. Wu, and R. Chellappa, "Learning without memorizing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5138–5146.

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

[22] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3449–3457.

[23] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 151–165.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
[25] Y. Wang, H. Su, B. Zhang, and X. Hu, "Learning reliable visual saliency for model explanations," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1796–1807, Jul. 2020.
[26] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. Int. Conf. Learn. Represent.*, 2015.
[27] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," in *Proc. Conf. Neural Inf. Process. Syst.*, 2019, pp. 4126–4135.
[28] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, 2016.
[29] 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," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2019.
[30] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9097–9107.
[31] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, and S. Zhu, "Interpreting CNN knowledge via an explanatory graph," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 4454–4463.
[32] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8827–8836.
[33] A. Mahendran and A. Vedaldi, "Salient deconvolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 120–135.
[34] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.
[35] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017.
[36] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.
[37] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim, "Evaluating weakly supervised object localization methods right," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3130–3139.
[38] H. Wang *et al.*, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 111–119.
[39] C. Cao *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2956–2964.
[40] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," Jun. 2017, *arXiv:1706.03825*.
[41] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.
[42] S. Sattarzadeh, M. Sudhakar, K. N. Plataniotis, J. Jang, Y. Jeong, and H. Kim, "Integrated grad-CAM: Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1775–1779.
[43] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[44] M. Everingham, S. M. A. Eslami, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.
[45] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 9525–9536.
[46] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2018.
[47] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.
[48] C. Cao, Y. Huang, Y. Yang, L. Wang, Z. Wang, and T. Tan, "Feedback convolutional neural network for visual localization and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1627–1640, Jul. 2019.

Lin Cheng received the M.S. degree in mechanical engineering from Hangzhou Dianzi University, Hangzhou, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Informatics, Xiamen University, Xiamen, China. His current research interests include computer vision, machine learning, and deep learning theory. He serves as a Reviewer for CVPR and ICCV.

Pengfei Fang received the B.E. degree in automation from Hangzhou Dianzi University in 2014 and the M.E. degree in mechatronics from the Australian National University (ANU) in 2017. He is currently pursuing the joint Ph.D. degree with ANU and the Data61-CSIRO. He is also a Visiting Scholar with Westlake University. His research interests include computer vision, machine learning, and cooperative control.

Yanjie Liang received the Ph.D. degree from the School of Informatics, Xiamen University, Xiamen, China, in 2021. He is currently a Postdoctoral Researcher with the Peng Cheng Laboratory, China. He has published several papers in IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *Pattern Recognition*, ACM MM, ICME, and ACCV. His current research interests include computer vision, machine learning, and visual tracking.

Liao Zhang received the B.E. degree in the Internet of Things from Fuzhou University, Fuzhou, China, in 2016, and the M.S. degree in computer science from Xiamen University, Xiamen, China, in 2020. His research interests include computer vision and machine learning.

Chunhua Shen is a Professor of computer science with Zhejiang University, China.

Hanzi Wang (Senior Member, IEEE) received the Ph.D. degree in computer vision from Monash University, Australia. He was an Adjunct Professor from 2010 to 2012 and a Senior Research Fellow from 2008 to 2010 at The University of Adelaide, Australia; an Assistant Research Scientist from 2007 to 2008 and a Postdoctoral Fellow from 2006 to 2007 at Johns Hopkins University; and a Research Fellow at Monash University from 2004 to 2006. He is currently a Distinguished Professor of Minjiang Scholars in Fujian and the Founding Director of the Center for Pattern Analysis and Machine Intelligence (CPAMI), Xiamen University (XMU). His research interests are concentrated on computer vision and pattern recognition. He was awarded the Douglas Lampard Electrical Engineering Research Prize and Medal for his Ph.D. thesis at Monash University. He has served on the Program Committee (PC) of ICCV, ECCV, CVPR, IJCAI, SIGGRAPH, and AAAI; and he has served on the reviewer panel for more than 40 journals and conferences. He was an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2010 to 2015 and a Guest Editor of *Pattern Recognition Letters* in September 2009.