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

Lin Cheng, Pengfei Fang, Yanjie Liang, Liao Zhang, Chunhua Shen, 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 are widely used in downstream visual tasks. Thus our work focuses on single backward pass approaches. 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 framework, termed *Target-Selective Gradient* (TSG) backprop, which leverages rectification operations to effectively emphasize target classes and further efficiently propagate the saliency to the input space, thereby generating *target-selective* and *fine-grained* saliency maps. The proposed TSG consists of two components, namely, TSG-Conv and TSG-FC, which rectify the gradients for convolutional layers and fully-connected layers, respectively. Thorough qualitative and quantitative experiments on ImageNet and Pascal VOC show that the proposed framework achieves more accurate and reliable results than other competitive methods.**

***Index Terms*—Model interpretability, explanation, visual saliency, 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], and low-level image processing [5], etc. However, people'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 will limit the development of real-world applications, e.g., autonomous driving [6] and medical diagnoses [7].

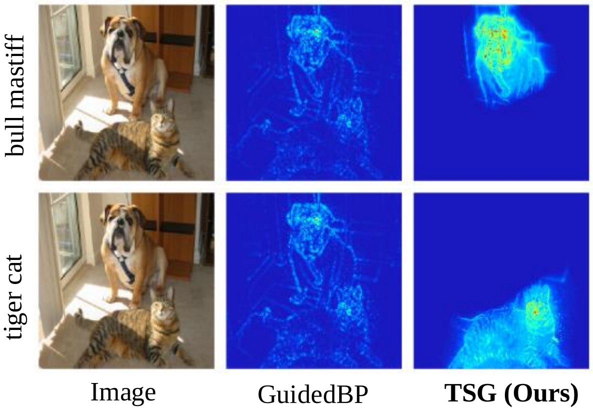


Fig. 1. Comparison of saliency maps w.r.t. target-selectiveness for the predictions of “bull mastiff” and “tiger cat”. GuidedBP produces two similar saliency maps, while the proposed TSG produces discriminative saliency maps.

To interpret the working mechanism of deep neural networks, many 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 the 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 visual task [15], [16], person re-identification task [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 through conducting iterative feedforward or backward [21]–[24]. Such a protocol is time-consuming and may bring 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 natures, 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 [25], and FullGrad [26], have been proposed to exploit the gradient to generate saliency maps, where the 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

This work is supported by the National Natural Science Foundation of China (Grant No. 61872307).

L. Cheng, Y. Liang, L. Zhang, and H. Wang are with Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, P.R. China (e-mail: cheng.charm.lin@hotmail.com, yanjieliang@stu.xmu.edu.cn, leochang@stu.xmu.edu.cn, hanzi.wang@xmu.edu.cn).

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

C. Shen is with Monash University, Australia (e-mail: chunhua@icloud.com).

*Corresponding author

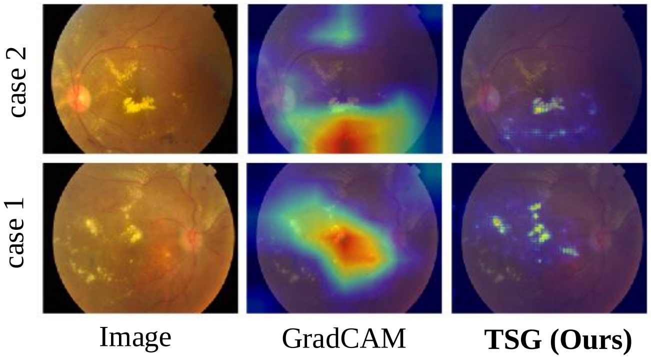


Fig. 2. Comparison of saliency maps w.r.t. fine-grainedness for the predictions of two cases of diabetic retinopathy. GradCAM produces two coarse saliency maps, while the proposed method TSG can produce fine-grained saliency maps.

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 explanation work. Other works, e.g., EBP [27] and GradCAM [28], attempt to leverage the top-down relevance or weighed activations to produce class-discriminative explanations. However, they fail to backprop the saliency to the input image space, thereby resulting in coarse 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], [29]. As shown in Fig. 2, GradCAM produces two coarse saliency maps, which cannot reveal the fine-grained patterns. Although much effort has 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 [30], [31], 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 further observe that the saliency maps are also 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.

Drawing the inspiration from the analysis, we propose a novel visual saliency framework, termed *Target-Selective Gradient* (TSG) backprop, to generate target-specific and fine-grained saliency maps, which explain how CNNs make decisions. The proposed TSG 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 explores the contributions of sub-nodes to the target node, and emphasizes the negative connections by the ratio of the positive contributions to the 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 to the image space.

The main contributions of this paper are summarized as follows:

- We explore the influence of entangled semantics and original gradients on the backprop of visual saliency. Based on our findings, we propose a novel visual saliency framework, i.e., TSG, to explain CNNs' decisions. To our best knowledge, this is the first work to simultaneously generate target-selective and fine-grained saliency maps in a single backward pass.
- We design a target selection module, i.e., TSG-FC, for the backprop of FC layers. TSG-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., TSG-Conv, for the backprop of Conv layers and other advanced layers. TSG-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 TSG against the competitive methods in target-selectiveness, fine-grainedness, running speed, explanatory generalization, and faithfulness. Moreover, TSG can be employed to diagnose the biases in the model and dataset. Furthermore, TSG 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 effects 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 pioneer 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 [25] 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 [28], [32]. The most recent work, FullGrad [26] improves the saliency maps by considering the multi-layer gradients aggregation.

2) Relevance Related Methods: Layer-wise Relevance Propagation [33] and Deep Taylor decomposition [34] explain the networks by decomposing the contribution of the target layer by layer. These methods pay attention to extensive existing objects, similar to [25]. Excitation Backprop (EBP) [27] proposes a contrastive marginal winning probability to propagate the top-down attention. DeepLIFT [35] assigns the attribution by comparing the difference between the input and the reference. CNN Fixation [8] measures the contribution strength between a pair of consecutive layers to uncover the pixel coordinates of saliency regions.

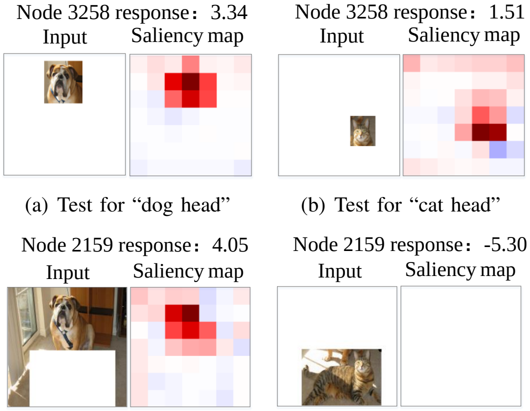
3) Activation Related Methods: CAM [12] and the generalized version GradCAM [28] utilize the gradient weighting the feature maps to localize the important regions. This type of methods are still the optimal noted in [36]. Guided GradCAM [28] ensembles GuidedBP and GradCAM, which actually needs more than 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 methods are 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, that take much running time and are possible to introduce adversarial noise. Most recently, Score-CAM [37] masks the input according to the intermediate activation maps and repeatedly performs N times (the number of activation maps) feedforwards to obtain the importance scores. Optimization related methods, including Feedback [38] and FGVis [29], add the complex switch structure into the network and iteratively optimize the object function to achieve the saliency maps. Some integration related methods, such as SmoothGrad [39], IntegratedGrad [40], and Integrated Grad-CAM [41], can be regarded as packages over single propagation methods. We can also take advantages of these packaging tools to increment our method effects. These iteration based methods are time-consuming and do not achieve the optimal performance.

Compared to iteration based methods, single backward pass based methods run faster, meanwhile without introducing 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.



(a) Test for "dog head"

(b) Test for "cat head"

(c) Test for "background"

(d) Test for "cat"

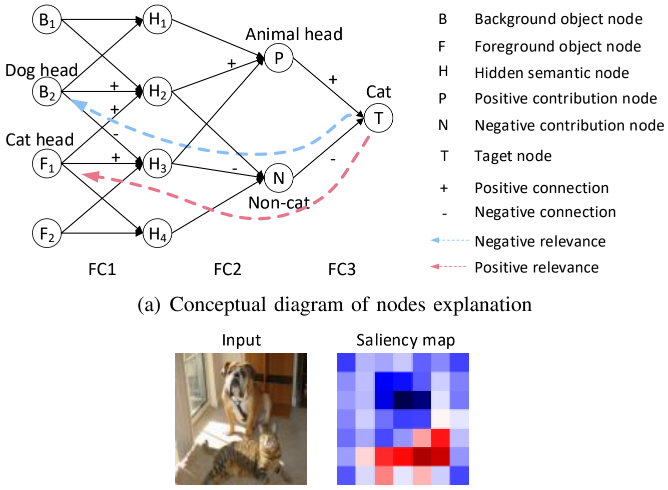
Fig. 3. Analysis of network nodes with entangled semantics. In each test, a feedforward is performed to obtain the node responses. Then the saliency is backproped from the node to the top Conv layer. Note that in (d), the backprop generates a blank map because of an inactivated state of the node, but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

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 and which makes the visualized results rough rather than fine-grained? Driven by these two essential 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 have been employed by many works [13], [25], [26], [28], [39], [40] for explanations.

1) Entangled Semantics in FC Layer: In the following, we take VGG16 model as an example. 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., cat head, 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 FC3 layer, both "dog head" and "cat head" can activate the node. Meanwhile, the saliency regions with corresponding semantics are produced by backprop from the node. Thus we naturally consider that positive contribution nodes encode entangled semantics, e.g., "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".

On the other hand, a negative contribution node, i.e., the *2159-th node* in the input of 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 as the input, whereas positive for a dog and background as 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,



(a) Conceptual diagram of nodes explanation

(b) Backprop from negative contribution nodes

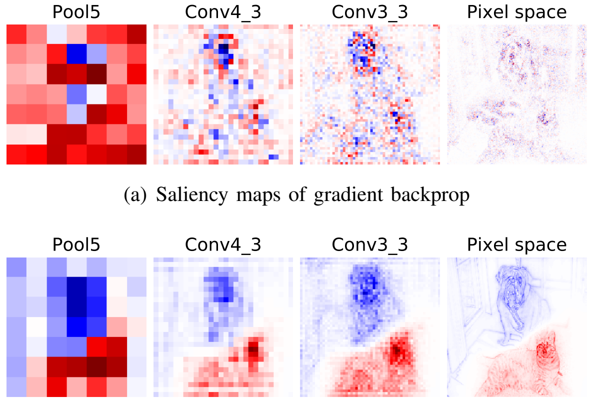
Fig. 4. An example of explanation for network nodes. The positive and negative relevance to the target are respectively marked in red and blue color[1]. Even number of negative connections make a positive contribution, depicted in the red dotted arrow.

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, "cat head" is negatively relevant to "non-cat" and "non-cat" is negatively relevant to "cat", leading that "cat head" is positively relevant to "cat" (Fig. 4(a)).

2) Backprop Noise in 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 [28], [39]. 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 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., deconvolutions operation) may introduce biases. This is more severe than the situation in FC layers, because of dozens of local perceptions inside the (de)convolutions. Moreover, the biases are accumulated layer by layer, leading to increasing noise 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 framework, i.e., target-selective gradient (TSG) backprop, composed of a target selection module and a fine-grained propagation module, as shown in Fig. 6. For a pre-trained CNN model, FC layers usually encode high-level semantic features related to the target class, while Conv layers encode local features related to the object details. Given this prior knowledge, we design the two modules of TSG backprop separately for FC layers and Conv layers. We will detail these two modules in the following.



(a) Saliency maps of gradient backprop

(b) Saliency maps of TSG

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

A. Target Selection Module

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

Let g_i^l denote the 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 $\mathrm{softmax}$ 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 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, as

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

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 by the ratio of positive contributions to negative contributions, as

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

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, resulting in too strong suppression for the foreground

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

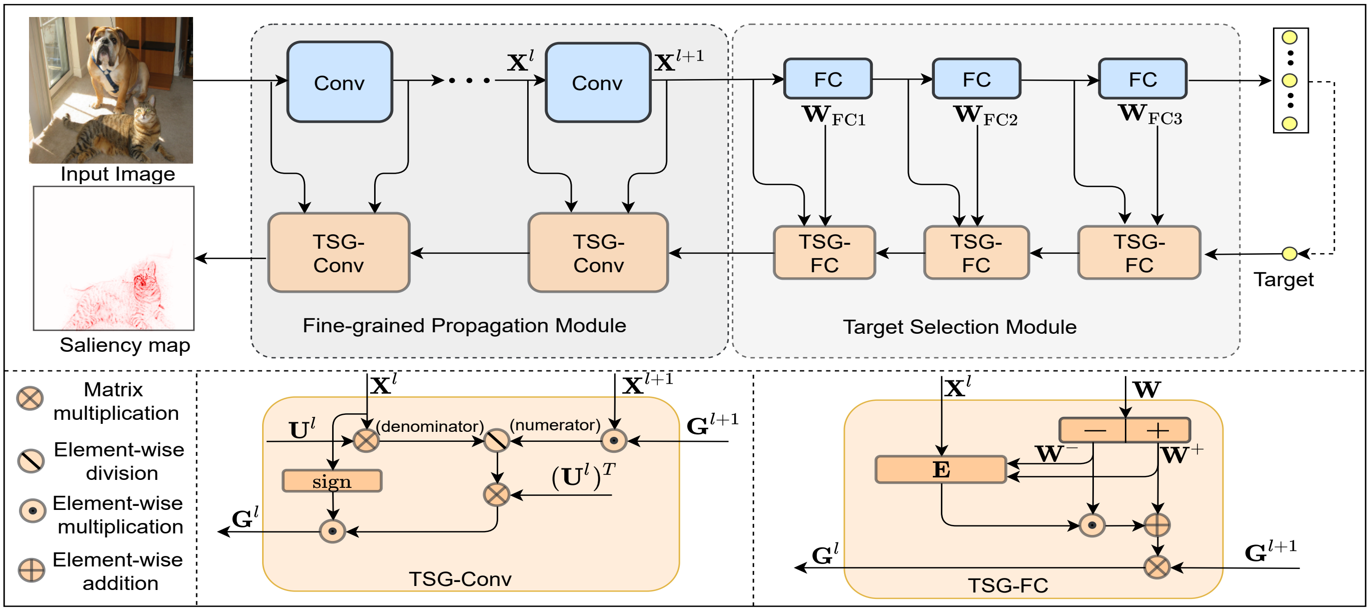


Fig. 6. The pipeline of proposed target-selective gradient (TSG) backprop framework. Here, we use VGG network as an example backbone network.

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 original weights, i.e., $E = 1$, for other FC layers if there exist, such as in VGG net. This is because other FC layers' information is integrated into the final layer's input which is included in Eq. (2). Different from EBP [27] which only uses positive weights, we make use of both positive and negative weights for 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. 5(b), we can observe that the target is effectively selected (red regions) and the irrelevant background is suppressed (blue regions).

B. Fine-Grained Propagation Module

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

In Conv layers, there exists the local perception over each space location, which is different from FC layers. Considering this difference, we implement the backprop of Conv layers differently. The TSG of the lower layer g_i^l in 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 a signum 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. $\mathbf{U}^l \in \mathbb{R}^{M \times N \times K_h \times K_w}$ is a set of Conv kernels with the 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 $\mathbf{G}_m^l \in \mathbb{R}^{H_l \times W_l}$ is formulated as

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

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 kernels in $\mathbf{U}$. To speed up the computation, we further obtain the following derivation from Eq. (4):

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

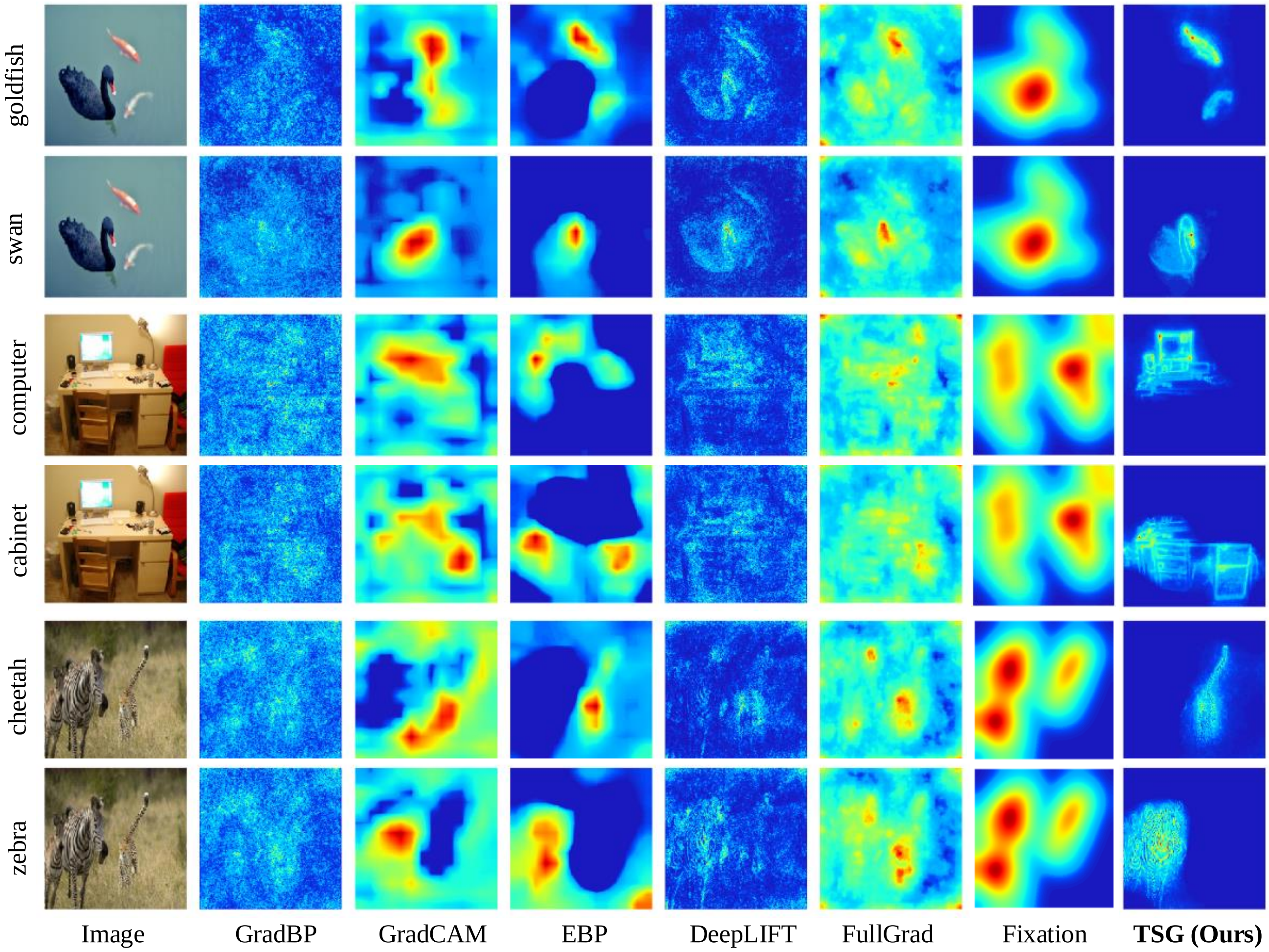


Fig. 7. Comparison of different methods on different samples. The saliency maps are generated from different targets, annotated on the left side, in each sample. The deep blue color represents the background and all other colors represent varying degrees of the target evidence. The negative values are truncated for better contrast.

Note that in Eq. (4), each channel in $\mathbf{U}_m^l$ 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.

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

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

$$g_i^l = \frac{x_j^{l+1}}{x_i^l} g_j^{l+1}. \tag{6}$$

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 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 input space to generate high-resolution saliency maps, meanwhile keeping the target concentration. Nevertheless, the TSG can be propagated to any layer inside the network to analyse 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 TSG via visual comparisons. Then in quantitative experiments, we evaluate the proposed TSG with weakly-supervised localization tasks on the ImageNet dataset [42] and Pascal VOC dataset [43]. Furthermore, we evaluate the faithfulness of the explanations with pixel perturbation [23] and sanity check [44]. Finally, we perform the bias diagnosis, extra tests on medical images, and the ablation study.

We compare our method with GradBP [13], GradCAM [28], DeepLIFT [35], [45], EBP [27], FullGrad [26] 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$\sim$1.3 for the negative enhancement in Eq. (2). More details can be found in Section V-F. We follow the processing in [26] 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 TSG to generate saliency maps from different targets on different samples in comparison with other competitors. As shown in Fig. 7, GradBP and DeepLIFT generate noisy maps, which highlight most of 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 for one image w.r.t. to different targets. One reasonable explanation for Fixation is that the backprop in 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, TSG can produce target selective and fine-grained maps with clear targets’ boundaries and fewer irrelevant backgrounds. Furthermore, visual explanatory results of TSG are more human interpretable, as compared to its competitors.

2) Comparison on Different Models: To verify the generalization of the proposed TSG, we further conduct the experiments across various CNN models, along with the competitors. From Fig. 8, we can find that over several cases EBP can produce saliency maps with fewer background areas than GradBP, DeepLIFT and FullGrad, while 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 presented models, while its results cannot discriminate the borders of targets precisely. Moreover, GradCAM also fails as EBP if the gradients are backproped to the low layer [28]. Unlike these competitors, TSG shows its advantage of being target-selective, fine-grained, and robust for extensive models, even for the models containing negative-value features (i.e., DenseNet121, MobileNet, etc). In addition, we find that saliency maps generated from ResNet50 and VGG16 are better than other backbone models. The target saliency from MobileNetV2 is relatively blurry when compared to 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.

TABLE I
LOCALIZATION ON IMAGENET VAL SET (LOWER IS BETTER). ERROR RATES OF GRADBP, GRADCAM, AND EBP FOR VGG16 ARE TAKEN FROM [28]. DEEPLIFT REFERS TO THE “CAPTUM” PACKAGE IN PYTORCH1.4 [46]. FIXATION REFERS TO THE OFFICIAL CODE IN [8].

	VGG16		ResNet50	
Method	Top5 LOC error (%)	FPS	Top5 LOC error (%)	FPS
GradBP [13]	51.46	25.64	55.44	**34.19**
GradCAM [28]	46.41	32.26	40.73	29.63
DeepLIFT [35]	55.32	7.12	53.11	17.78
EBP [27]	63.04	23.26	44.44	26.67
FullGrad [26]	47.82	12.05	46.35	9.93
Fixation [8]	58.38	0.39	-	-
TSG (Ours)	**43.46**	**43.48**	**40.49**	31.25

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 [27], [28], [47], we evaluate the visual saliency methods with weakly-supervised object localization task on ImageNet dataset on 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 [42].

After achieving the saliency maps, we search the best performing thresholds for different methods and binarize the saliency maps at the selected thresholds of the maximum value to obtain the bounding boxes. Besides the binarization with the threshold, we do not append any other post-processing techniques to our method. As shown in Table I, TSG outperforms other methods in localization errors on the both explained models. For example, the results of TSG read as 43.46/40.49, as compared to 46.41/40.73 of the second best method, i.e. GradCam, on VGG16/ResNet50. Compared to ResNet, VGG model has more FC layers, which are possible to involve the stronger entanglement as stated in Section III. In this situation, using TSG to disentangle the semantics in FC layers will boost the performance of VGG. Note that GradCAM additionally applies the post-processing technique, i.e., searching for the largest connected component after binarization. TSG also presents better performance gains over FullGrad by 4.36%/5.86% on VGG16/ResNet50, where FullGrad aggregates all Conv-layer saliency maps to improve the performance but 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 clas-

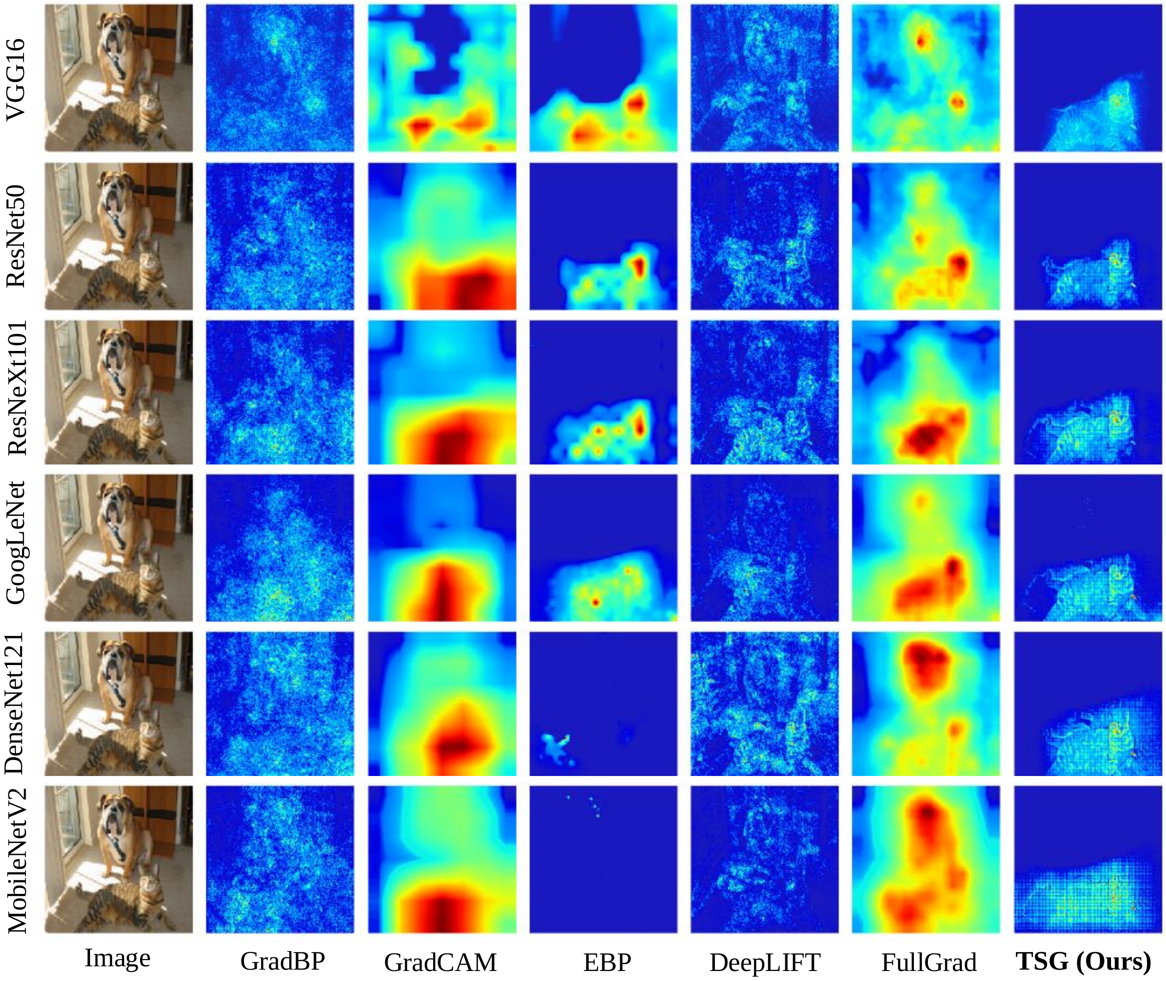


Fig. 8. Comparison of different methods on different models. The models' names are annotated on the left side. The saliency maps are generated from the same target, i.e., "tiger cat".

sification capacity of ResNet50, leading to better localization performances.

Moreover, we test the average running speed with 224 × 224 images on a GeForce GTX 1080 Ti GPU. The proposed TSG 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 very slow running.

2) Point Localization: Considering that the explanatory results intend to focus on the most discriminative regions of targets, we use another popular metric, Pointing Game [27], to measure 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 superior performance over other methods on Pascal VOC2007 test set, which can be attributed to the target-selectiveness of TSG. GradBP and Fixation have 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. 7).

TABLE II
POINTING GAME ON VOC2007 TEST SET (HIGHER IS BETTER). THE RESULTS OF GRADBP, GRADCAM, AND EBP ARE COPIED FROM [27].

	VGG16		ResNet50	
Method	Mean accuracy (%)	FPS	Mean accuracy (%)	FPS
GradBP [13]	76.00	18.18	65.80	**16.95**
GradCAM [28]	86.60	**18.52**	90.60	16.25
DeepLIFT [35]	79.05	7.69	82.72	3.12
EBP [27]	80.00	10.41	89.20	6.31
FullGrad [26]	84.16	8.56	88.99	3.14
Fixation [8]	74.52	0.44	-	-
TSG (Ours)	**89.33**	18.18	**90.68**	11.82

C. Faithfulness Check

1) Pixel Perturbation: In order to evaluate the faithfulness of explanatory results in pixel level, we use the deletion metric [23] to test TSG. 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 iteratively perturbing the important pixels according to the rank of saliency values generated by saliency method. The

TABLE III
PIXEL PERTURBATION ON VOC2012 VAL SET (LOWER IS BETTER). LOWER DELETION SCORE SUGGESTS HIGHER FAITHFULNESS OF SALIENCY METHODS.

Method	Deletion score	FPS
GradBP [13]	0.1932	10.10
GradCAM [28]	0.1680	10.10
DeepLIFT [35]	0.1621	1.42
EBP [27]	0.5028	5.41
FullGrad [26]	0.1667	4.78
TSG (Ours)	**0.1564**	**14.29**

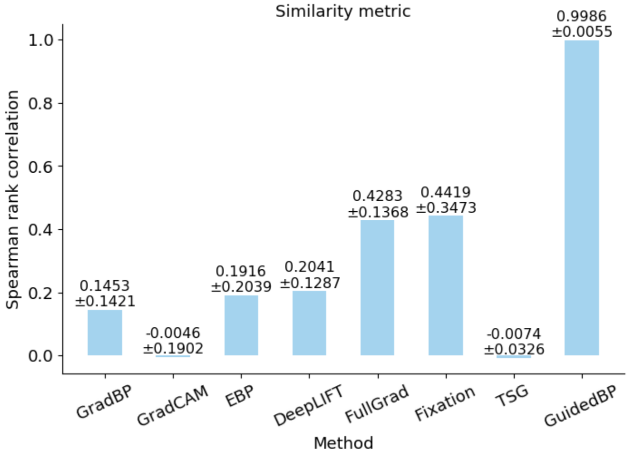


Fig. 9. Sanity check with similarity metric for model randomization. Spearman rank correlation is taken as the similarity metric. The values above the bar are means and standard deviations of similarities between original explanations and randomized explanations on ImageNet. Lower similarity denotes better faithfulness of explanations.

steeper the decline (i.e., the lower deletion score) is, the more reliable the saliency method is. As shown in Table III, TSG achieves the lowest score, that suggests TSG is faithful to the model predictions and capable of capturing the fine-grained details corresponding to the targets.

2) Sanity Check: As suggested by [44], we conduct sanity check for the proposed TSG to validate whether the explanatory results are sensitive to the model parameters or not. If the explanatory results are much similar before and after the model parameters are randomized, the saliency method is more risky to be trusted. We evaluate the similarity with Spearman rank correlation before and after the randomization of model parameters for our TSG and other comparative methods, including GuidedBP [25] for reference. As illustrated in Fig. 9, TSG and GradCAM are sensitive to the change of the parameters, while GuidedBP is much independent on model parameters.

D. Diagnosing Bias and Failure Cases

We adopt TSG to diagnose the biases in the VGG16 network pre-trained on ImageNet. As shown in Fig. 10, the man is recognized as "basketball" (top left), and the station is recognized as "train" (bottom left), which makes it difficult to interpret the failure clues by only knowing the prediction possibilities. Fortunately, with the help of target-selective and fine-grained saliency maps generated by TSG, one can breezily 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. 10, our method fails to produce the target-specific visualized maps, 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 TSG is faithful to the model.

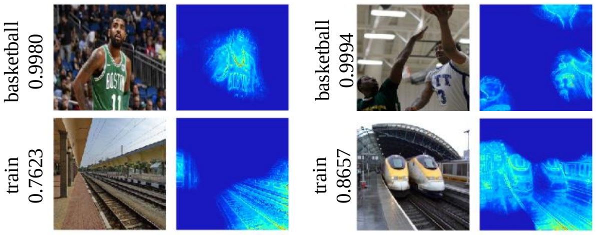


Fig. 10. Diagnosing bias and failure cases. The texts denote the predicted targets and possibilities. TSG can help diagnose the biases in the model and dataset without suppressing the useful information even in the background.

E. Explanation for Medical Images

To test the generalization of TSG on the different type of images, we use TSG to explain the deep neural model trained on the medical image dataset, i.e., Kaggle Diabetic Retinopathy. 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 Kaggle Diabetic Retinopathy dataset and ImageNet dataset. The explained model is ResNet152 trained on Kaggle Diabetic Retinopathy dataset with image-level labels. It took around more than 100 epochs to train this model to reach 97% accuracy for classification. As shown in Fig. 11, TSG obtains more reliable explanatory results than other comparative methods. Benefited from the target-selectiveness, TSG can focus on the disease-relevant regions. More importantly, with the property of fine-grainedness, TSG can 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 with the proposed fine-grained propagation module via ablation study on ImageNet localization task. We choose the vanilla gradient backprop as our baseline. As shown in Table IV, when replacing the vanilla gradient in FC layers with the target selection module, the LOC error is 4.39% lower. When replacing the vanilla gradient in Conv layers with the fine-grained propagation module, the LOC error is 1.29% lower. Although the fine-grained propagation module looks less useful, when we further append the fine-grained propagation module to the target selection module, the LOC

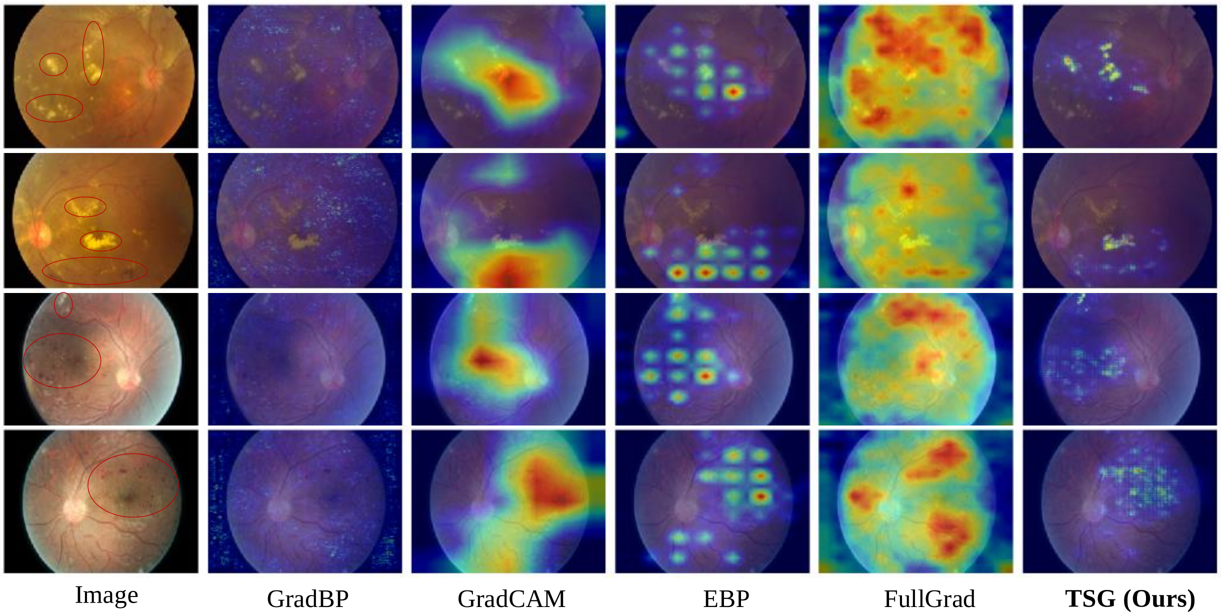


Fig. 11. Explanation for medical images. We compare the proposed TSG with GradBP [13], GradCAM [28], EBP [27], and FullGrad [26]. The red ovals denote the lesions of retinas.

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

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

error continues to decrease, i.e., 5.14% lower than the "TSG-FC+Grad". This shows that both of the proposed modules are necessary to TSG and provide complementary benefits to TSG. Those two modules are tied tightly in one framework, achieving a superior performance than a single module.

2) Fine-Grained Propagation Module vs. Edge Detector: Benefited from the fine-grained propagation module, the saliency maps can highlight clear details relevant to the targets, such as examples in Fig. 7. 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 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 pattern and color pattern, besides the edges pattern, such as "orange" and "dark glasses" in Fig. 12. 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 "dark glasses" and a bar in "soccer ball" in Fig. 12. In addition, the fine-grained propagation module can propagate the saliency maps to different semantic levels in different spatial scales, and can analyze the attributions of different features channels.

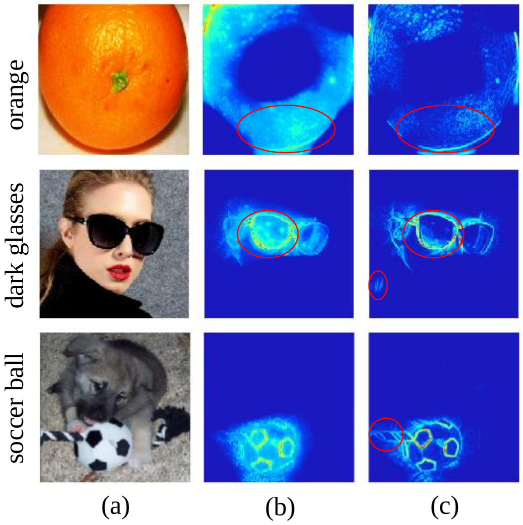


Fig. 12. Fine-grained propagation module vs. edge detector. (a) Input image. (b) Saliency maps from TSG. (c) Replacement of the fine-grained propagation module in TSG with the edge detector.

3) Influence of Scale Coefficient: To analyze the influence of choosing different scale coefficients α in Eq. (2), we test the proposed TSG with Pointing Game, as mentioned in "Point localization" in Section V-B. We record the experimental results corresponding to $\alpha \in [0.5 : 0.1 : 1.3]$ both on VGG16 and ResNet50 models. As depicted in Fig. 13, we can find that each model has one peak mean accuracy as α varies during the interval, where VGG16 achieves the best result at $\alpha = 0.8$ and ResNet50 achieves the best result at $\alpha = 0.9$. Furthermore, the results on VGG16 are less sensitive to the scale coefficient α than that on ResNet50. Specially, the accuracy on VGG16

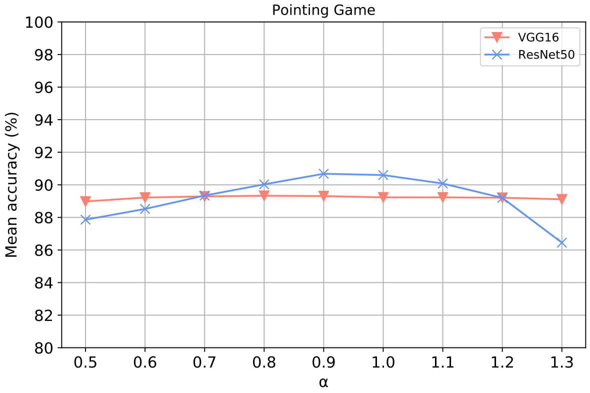


Fig. 13. Influence of different scale coefficients α on the performance of TSG.

model fluctuates within a very small range, i.e., 0.35, during the whole interval of α. On both models, there is a smaller 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 framework, i.e., target-selective gradient (TSG) backprop, which is comprised 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 TSG can more discriminately explain different targets and generate clearer saliency maps than the competitive methods. Moreover, TSG can be used for most of the CNN models. Quantitative experiments reveal that TSG gains superior localization performance, and stronger reliability over the competitive methods. Furthermore, we also verify that TSG 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 future research.

References

[1] L. Song, J. Liu, B. Qian, M. Sun, K. Yang, M. Sun, and S. Abbas, "A deep multi-modal CNN for multi-instance multi-label image classification," *IEEE Transaction on Image Processing*, vol. 27, no. 12, pp. 6025–6038, 2018.

[2] Y. Ding, Z. Ma, S. Wen, J. Xie, D. Chang, Z. Si, M. Wu, and H. Ling, "AP-CNN: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification," *IEEE Transaction on Image Processing*, 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 Transaction on Image Processing*, vol. 29, pp. 3520–3533, 2020.

[4] L. Jing, Y. Chen, and Y. Tian, "Coarse-to-fine semantic segmentation from image-level labels," *IEEE Transaction on Image Processing*, 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 Transaction on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[6] J. Kim and J. F. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proceedings of IEEE International Conference on Computer Vision*, 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 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 Transaction on Image Processing*, vol. 28, no. 5, pp. 2116–2125, 2019.

[9] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Proceedings of Neural Information Processing Systems*, 2019, pp. 9273–9282.

[10] S. Wickramanayake, W. Hsu, and M. Lee, "FLEX: faithful linguistic explanations for neural net based model decisions," in *Proceedings of AAAI Conference on Artificial Intelligence*, 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 *Proceedings of AAAI Conference on Artificial Intelligence*, 2020, pp. 5875–5882.

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

[13] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proceedings of International Conference on Learning Representations*, 2014.

[14] Z. Fang, K. Kuang, Y. Lin, F. Wu, and Y. Yao, "Concept-based explanation for fine-grained images and its application in infectious keratitis classification," in *Proceedings of ACM International Conference on Multimedia*, 2020, pp. 700–708.

[15] Y. Wei, J. Feng, X. Liang, M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5267–5276.

[17] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1389–1398.

[19] P. Fang, J. Zhou, S. K. Roy, L. Petersson, and M. Harandi, "Bilinear attention networks for person retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8030–8039.

[20] P. Dhar, R. V. Singh, K. Peng, Z. Wu, and R. Chellappa, "Learning without memorizing," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5138–5146.

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

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

[23] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *Proceedings of British Machine Vision Conference*, 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 *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

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

[26] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," in *Proceedings of Conference on Neural Information Processing Systems*, 2019, pp. 4126–4135.

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

[28] 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," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

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

[30] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, and S. Zhu, "Interpreting CNN knowledge via an explanatory graph," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2018, pp. 4454–4463.

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

[32] A. Mahendran and A. Vedaldi, "Salient deconvolutional networks," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 120–135.

[33] 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, p. e0130140, 2015.

[34] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, p. 211–222, 2017.

[35] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of International Conference on Machine Learning*, 2017, pp. 3145–3153.

[36] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim, "Evaluating weakly supervised object localization methods right," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3130–3139.

[37] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 111–119.

[38] C. Cao, X. Liu, Y. Yi, Y. Yu, and T. S. Huang, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 2956–2964.

[39] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *CoRR*, vol. abs/1706.03825, 2017.

[40] M. Sundararajan, A. Taly, , and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of International Conference on Machine Learning*, 2017, pp. 3319–3328.

[41] 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," *CoRR*, vol. abs/2102.07805, 2021.

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

[43] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2014.

[44] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proceedings of Neural Information Processing Systems*, 2018, pp. 9525–9536.

[45] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *Proceedings of International Conference on Learning Representations*, 2018.

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, and *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of Neural Information Processing Systems*, 2019, pp. 8024–8035.

[47] C. Cao, Y. Huang, Y. Yang, L. Wang, Z. Wang, and T. Tan, "Feedback convolutional neural network for visual localization and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1627–1640, 2019.