LeNo: Adversarial Robust Salient Object Detection Networks with Learnable Noise

He Wang[1,2], **Lin Wan**[1], **He Tang**[1]*

[1]School of Software Engineering, Huazhong University of Science and Technology
[2]School of Cyber Science and Engineering, Huazhong University of Science and Technology
{hew, linwan, hetang}@hust.edu.cn

Abstract

Pixel-wise prediction with deep neural network has become an effective paradigm for salient object detection (SOD) and achieved remarkable performance. However, very few SOD models are robust against adversarial attacks which are visually imperceptible for human visual attention. The previous work robust saliency (ROSA) shuffles the pre-segmented superpixels and then refines the coarse saliency map by the densely connected conditional random field (CRF). Different from ROSA that relies on various pre- and post-processings, this paper proposes a light-weight Learnable Noise (LeNo) to defend adversarial attacks for SOD models. LeNo preserves accuracy of SOD models on both adversarial and clean images, as well as inference speed. In general, LeNo consists of a simple shallow noise and noise estimation that embedded in the encoder and decoder of arbitrary SOD networks respectively. Inspired by the center prior of human visual attention mechanism, we initialize the shallow noise with a cross-shaped gaussian distribution for better defense against adversarial attacks. Instead of adding additional network components for post-processing, the proposed noise estimation modifies only one channel of the decoder. With the deeply-supervised noise-decoupled training on state-of-the-art RGB and RGB-D SOD networks, LeNo outperforms previous works not only on adversarial images but also on clean images, which contributes stronger robustness for SOD. Our code is available at https://github.com/ssecv/LeNo.

Introduction

The progresses of deep neural networks (DNNs) have significantly promoted the development of down-stream computer vision tasks such as image recognition (He et al. 2016), semantic segmentation (Chen et al. 2017), object detection (Ren et al. 2015) and salient object detection (Zhao et al. 2020). These data-driven models are usually trained with extensive input images and the annotations. Previous studies (Che et al. 2021) and (Li, Li, and Yu 2019) have shown that the saliency detection networks are fragile to adversarial attacks and the performance significantly decreases with even imperceptible perturbations. As shown in Fig. 1(a), the SOD network fails to detect the salient object of input image with adversarial perturbations, even if the salient object

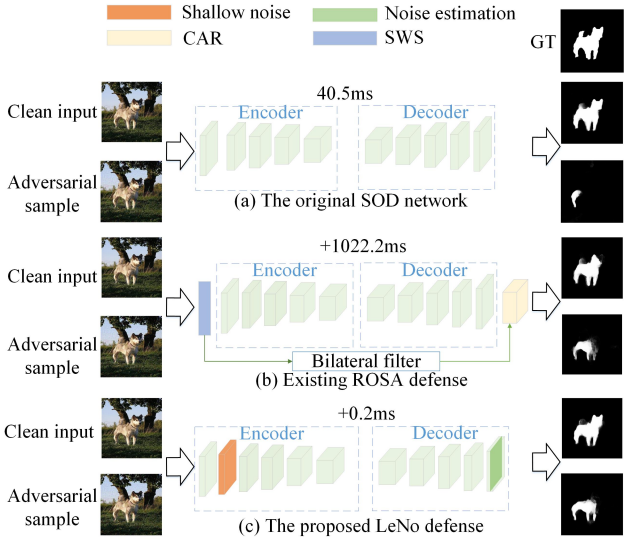

Figure 1: SOD adversarial defense comparison. (a) Original GateNet without any defense. (b) ROSA defense which adds three components to the front, middle and back of the network respectively. (c) The proposed LeNo defense, it embeds a lightweight learnable shallow noise and noise estimation and balances the performance of the clean image and adversarial image.

of the adversarial image is still obvious and can be easily distinguished from the background. As shown in Table 1, without modifying the ground truth annotations, under adversarial attacks, GateNet (Zhao et al. 2020) obtains only .2741 F_β on ECSSD dataset and BBSNet (Fan et al. 2020f) obtains only .3397 F_β on NJU2K dataset, the F_β drops by .6697 and .5677 respectively. This phenomenon indicates that state-of-the-art SOD models are easy to be fooled by adversarial attacks, even with the clean auxiliary depth map. The robustness of SOD models seriously concerns the security and perception gap between human and neural networks, because human visual system is hard to recognize the adversarial perturbations of the input image.

As shown in Fig. 1(b), recent study ROSA (Li, Li, and Yu 2019) has proposed a robust salient object detec-

*The corresponding author is He Tang.

tion framework that incorporates a segment-wise shielding (SWS) component in front of the backbone, a context-aware restoration (CAR) component after the decoder and a bilateral filter between input image and densely connected CRF. It performs well on 4 RGB SOD benchmark datasets with 3 SOD networks. The core idea of ROSA is to introduce another generic noise to destroy subtle curve-like pattern of adversarial perturbations. SWS component divides an image into superpixels by SLIC (Achanta et al. 2012) and shuffles the pixels within a superpixel randomly, it breaks adversarial pertubations by introducing random noise. The CAR component refines the saliency detection with a densely connected CRFasRNN (Zheng et al. 2015). However, the noise introduced by SWS is random and not learnable, making accuracy of the SOD model drops obviously on clean images, e.g., Fig. 1(b). Moreover, the ROSA is heavy and requires over 1 second additional time at inference stage. Thus it may retard some real-time SOD models.

In order to handle the aforementioned limitations, this paper proposes a learnable noise against adversarial attacks of SOD networks. The learnable noise consists of a shallow noise and a noise estimation. Different from SWS that introduces noise directly to the input image, the shallow noise inserts a noisy layer between stem and stage 1 of the backbone. It introduces noise in feature-level, thereby the noise is learnable and able to balance learning clean images with learning adversarial images. Inspired from the image denoising method (Guo et al. 2019), we propose a lightweight noise estimation component to refine the feature of adversarial images. Our shallow noise and noise estimation are embedded in the encoder and decoder respectively, allowing parallel computation. Furthermore, the noise estimation only affects one channel of the decoder. Consequently, our defense method introduces much less extra time and performs better than ROSA, see Fig. 1(b) and (c).

Our main contributions can be summarized as three-folds:

1. We successfully launch adversarial attacks on state-of-the-art RGB and RGB-D SOD models. Experimental results verify that a wide range of existing SOD models are sensitive to adversarial perturbations.
2. We propose a simple but efficient learnable noise (LeNo) which hardly modifies the original SOD network structure. It consists of a plug-and-play shallow noise and noise estimation. It is parallel computing and hardly influences the inference speed.
3. With the deeply-supervised noise-decoupled training scheme, the proposed defense method promotes adversarial robustness of extensive RGB and RGB-D SOD networks. The experimental results show that our proposed defense method outperforms previous works not only on adversarial images but also on clean images.

Related Works

Salient Object Detection

An impressive mechanism of human vision system is the internal process that quickly scans the global image to obtain region of interest. In the field of computer vision, this task is referred to as Salient Object Detection. It plays a key role in a range of real-world applications, such as medical image segmentation (Fan et al. 2020b; Wu et al. 2021a), camouflaged object detection (Fan et al. 2020a), etc. Although significant progress has been made in the past several years (Zhao et al. 2019; Li et al. 2018; Qin et al. 2019), there is still room for improvement when faced with challenging factors, such as complicated backgrounds or varying lighting conditions in the scenes. One way to overcome such challenges is to employ depth maps, which provide complementary spatial information and have become easier to be captured due to the ready availability of depth sensors. Recently, RGB-D based salient object detection has gained increasing attention, and various methods have been developed (Fan et al. 2020c; Chen et al. 2021). Early RGB-D based salient object detection models tended to extract handcrafted features and then fused the RGB image and depth map. Despite the effectiveness of traditional methods using handcrafted features, their low-level features tend to limit generalization ability, and they lack the high-level reasoning required for complex scenes. To address these limitations, several deep learning based RGB-D salient object detection methods (Fan et al. 2020d) have been developed, with improved performance.

Adversarial Attacks

Existing adversarial attacks consist of several groups, one-step gradient-based methods; iterative methods (Dong et al. 2018); optimization-based methods (Xiao et al. 2018); and generative networks (Zhao, Dua, and Singh 2017; Poursaeed et al. 2018) based methods.

FGSM Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014) is an efficient single-step adversarial attack method. Given vectorized input x and corresponding target label y, FGSM alters each element of x along the direction of its gradient w.r.t the inference loss $\partial L/\partial x$. The generation of adversarial example $\hat{x}$ (i.e., perturbed input) can be described as:

$$\hat{x} = x + \epsilon \cdot sgn\left(\nabla_x \mathcal{L}(g(x;\theta), y)\right), \tag{1}$$

where ϵ is the perturbation constraint that determines the attack strength. $g(x;\theta)$ computes the output of DNN paramterized by θ. $sgn\left(\cdot\right)$ is the sign function.

PGD Projected Gradient Descent (PGD) (Madry et al. 2017) is a multi-step variant of FGSM, which is one of the strongest L_∞ adversarial example generation algorithm. With $\hat{x}_{t=1} = x$ as the initialization, the iterative update of perturbed data $\hat{x}$ in iteration t can be expressed as:

$$\hat{x}_t = \Pi_{P_\epsilon(x)}\left(\hat{x}_{t-1} + a \cdot sgn\left(\nabla_x \mathcal{L}\left(g\left(\hat{x}_{t-1};\theta\right), y\right)\right)\right), \tag{2}$$

where $P_\epsilon(x)$ is the projection space which is bounded by $x \pm \epsilon$, and a is the step size. (Madry et al. 2017) also proposes that PGD is a universal adversary among all the first-order adversaries.

ROSA ROSA (Li, Li, and Yu 2019) is an iterative gradient-based pipeline. It is the first adversarial attack on the state-of-the-art salient object detection models. They try to make the predictions of all pixels in x go wrong. In each

iteration t, supposing that adversarial sample $\hat{x}$ from previous time step or initialization is prepared, the adversarial sample is updated as:

$$\hat{x}_0 = x, \hat{x}_{t+1} = \hat{x}_t + p_t, \tag{3}$$

$$p'_t = \sum_{i \in S_t} \left[\nabla_{\hat{x}_t} g_{i,1-y_i}(\hat{x}_t; \theta) - \nabla_{\hat{x}_t} g_{i,y_i}(\hat{x}_t; \theta) \right]. \tag{4}$$

Here, p_t denotes the adversarial perturbations computed at t-th step, it is obtained by normalization as $\alpha \cdot p'_t / \|p'_t\|_\infty$ where α is a fixed step length, i denotes one pixel in x, S_t denotes the set of pixels that g can still classify correctly and y_i denotes two categories: salient and nonsalient.

Defenses against Adversarial Attacks

Many researchers resort to randomization schemes (Liu et al. 2018; Dhillon et al. 2018) for mitigating the effects of adversarial perturbations in the input/feature domain. The intuition behind this type of defense is that DNNs are always robust to random perturbations. A randomization based defense attempts to randomize the adversarial effects into random effects, which are not a concern for most DNNs. Some of them also add noise to the network as we do, but their noise is random and not learnable, like ROSA.

Previous works in feature denoising (Liao et al. 2018) attempt to alleviate the effects of adversarial perturbations on high-level features learned by DNNs. Unlike them, we defend against adversarial perturbations by affecting low-level features, making a lightweight defense framework.

L2P (Jeddi et al. 2020) is an end-to-end feature perturbation learning approach for image classification. Inspired by it, we adopt alternating back-propagation to effectively update parameters. The difference is that we propose a new cross-shaped noise with fewer parameters and without adversarial training.

Methodology

Adversarial Attacks on Salient Object Detection Models

We perform a total of three attacks, namely FGSM, PGD and ROSA. Their step sizes are 0.3, 0.04 and 0.1, respectively. The max interations of PGD and ROSA are chosen as 10 and 30. Our bound is set to 20 just like ROSA. For RGB-D salient object detection, we generate the corresponding adversarial dataset for every dataset used in our experiment with pretrained BBSNet. These adv-datasets are used to attack other RGB-D models. In this process, we include white-box attack for BBSNet and black-box attack for CoNet and MobileSal. For RGB salient object detection, same as before, we choose GateNet as the victim model. PiCANet and PFSNet are tested on adv-datasets generated by GateNet.

Learnable Noise against Adversarial Attacks

SOD is a pixel-wise dense prediction task. The receptive field of shallow layer is smaller than deep layer and could be sensitive to low-level perturbations, such as adversarial noise. As shown in Fig. 2, adversarial images cause a great suppression of the shallow activation maps of the network, the few remaining activation leads to the final huge deviation in deep layers. In this paper, our goal is to rectify the perturbated activations as much as possible. Besides, high-level features are aggregated from low-level features in DNN, so it is more effective to defend after the stem than at deeper stage. This is the core motivation for proposing the shallow noise.

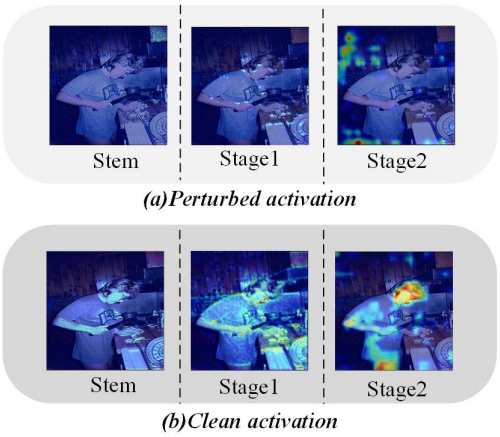

Figure 2: Activation maps when PGD perturbed image and clean image are passed through backbone.

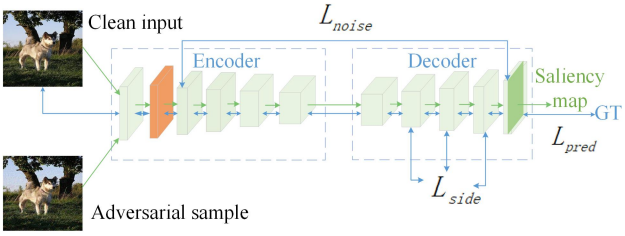

Figure 3: The overall architecture of LeNo. The shallow noise is inserted into the backbone and noise estimation is part of the last stage of decoder. The blue arrows are the data flow during training. The output of shallow noise is used to calculate loss of noise estimation, i,e. L_{noise}. The green arrows are the data flow when LeNo detects salient objects.

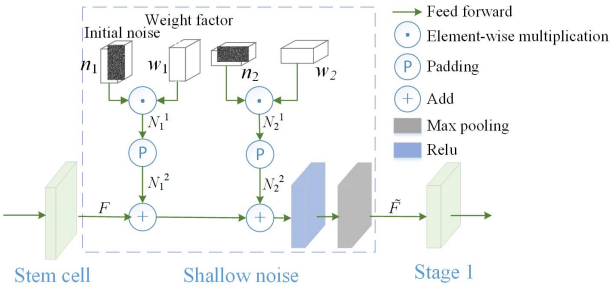

Figure 4: Illustration of shallow noise.

Shallow Noise. We attempt to rectify the perturbed acti-

vation by adding noise to the shallow stage, as shown in Fig. 3 and Fig. 4. First, we initialize two noises, which are shaped as $\mathcal{R}^{C\times H\times (W/2)}$ and $\mathcal{R}^{C\times (H/2)\times W}$. Here, $\mathcal{R}^{C\times H\times W}$ is the shape of input of shallow noise. Both of them are sampled from gaussian distribution and multiplied with a weight factor independently. The weight factor has the same shape with its corresponding noise and each element of it is initialized as 0.25. The whole process can be mathematically described as:

$$N_1^1 = n_1 * w_1, n_1 \sim N(0,1), \tag{5}$$

$$N_2^1 = n_2 * w_2, n_2 \sim N(0,1). \tag{6}$$

Then, we pad them with zero to the shape of $\mathcal{R}^{C\times H\times W}$ as:

$$N_1^2 = padding(N_1^1), \tag{7}$$

$$N_2^2 = padding(N_2^1). \tag{8}$$

At this point, the noise is generated and can be added directly to the input. The output of shallow noise can be mathematically represented as:

$$\tilde{F} = maxpool(RELU(F + N_1^2 + N_2^2)), \tag{9}$$

where F is the original noise-free feature which is appended with learnable noise later.

In each channel of N_1^2 and N_2^2, noise is not global. Some area has noise which is added later while others don't. As shown in Fig. 5, inspired by the center prior of human visual attention mechanism, we propose a cross-shaped noise which is stronger at the center area and noise-free in the corner.

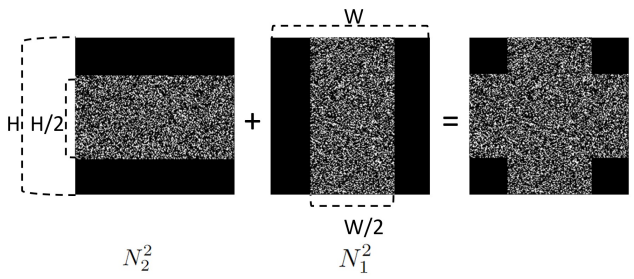

Figure 5: Generation of cross-shaped noise.

Noise Estimation. We replace a channel of the last stage of decoder to estimate the noise and refine the feature, minimizing changes to the network while improving the predictive performance. At the last stage of the decoder, we select the first channel for this task. We calculate the average map of output of the shallow noise along channel dimension as noise ground truth to supervise the learning process. It can be mathematically expressed as:

$$N_{gt} = Avg(\tilde{F}), \tag{10}$$

where $\tilde{F}$ is the output of shallow noise.

Deeply-supervised Noise-decoupled Training Scheme

As shown in Fig. 3, the network is trained with only clean images, but it performs well on both clean and adversarial images. The loss function can be mathematically discribed as:

$$L_{noise} = BCE(N_{est}, N_{gt}) + \lambda MSE(N_{est}, N_{gt}), \tag{11}$$

$$L_{side} = \sum_{i=1}^{3} BCE(S_i, G), \tag{12}$$

$$L_{pred} = BCE(S, G), \tag{13}$$

$$L = L_{noise} + L_{side} + L_{pred}, \tag{14}$$

where N_{est} denotes noise estimation result in decoder, N_{gt} denotes noise GT which is the average map of the output of shallow noise on channels. S_i denotes one of the sideouts induced from decoder. S is the output of the SOD network. L is the summation of the three losses, we set λ to 0.1 in our experiments.

Broadly speaking, our training process is divided into two phases. In the first phase, we decouple the learning process to effectively update network and noise parameters. In the second phase, we freeze parameters of shallow noise and previous layers to get noise GT, preparing for noise estimation. Specifically, our detailed training procedure is described as follows:

1. Initialize the network parameters θ_n^0 with the pretrained weight.
2. Randomly sampling shallow noise from a Gaussian distribution, with weights θ_w^0 initialized as 0.25 for all elements.
3. Based on θ_n^0, θ_w^0 and the sum of L_{side} and L_{pred}, utilize SGD to train the network and alternately update these two parameters, obtaining θ_n^1 and θ_w^1.
4. Freeze θ_w^1 and layers before shallow noise.
5. Based on θ_n^1, θ_w^1 and L, utilize SGD to train the network and update network parameters, obtaining θ_n^2.

Experimental Results

Datasets

In this paper, we focus on both RGB-D and RGB based salient object detection. For RGB-D based salient object detection, we experiment on NJU2K (Ju et al. 2014a), NLPR (Peng et al. 2014), LFSD (Li et al. 2014a), SIP (Fan et al. 2020e), NJUD (Ju et al. 2014b), STEREO (Niu et al. 2012) and DUTS-D (Piao et al. 2019). We train on the training set of NJU2K and NLPR, their test set and other public datasets are all for testing. For RGB based salient object detection, we experiment on ECSSD (Shi et al. 2015), HKU-IS (Li and Yu 2015), DUTS (Wang et al. 2017), DUT-OMRON (Yang et al. 2013), PASCAL-S (Li et al. 2014b) and SOD (Martin et al. 2001). We train on DUTS-TR, the remaining datasets are used for testing. We follow the released data split.

	$F_\beta \uparrow$		MAE$\downarrow$		$F_\beta \uparrow$		MAE$\downarrow$		$F_\beta \uparrow$		MAE$\downarrow$	
	clean	adver	clean	adver	clean	adver	clean	adver	clean	adver	clean	adver
	NJU2K				NLPR				SIP			
BBS	.9074	.3397	.0419	.3783	.8991	.1976	.0286	.4492	.8780	.2718	.0589	.3755
CoNet	.8946	.7952	.0455	.0885	.8928	.7736	.0280	.0626	.8693	.7981	.0607	.0883
MobileSal	.9356	.7711	.0312	.1174	.8902	.7081	.0336	.0956	.8640	.7149	.0703	.1236
	ECSSD				HKU-IS				DUTS-TE			
GateNet	.9438	.2741	.0332	.4591	.9432	.3127	.0301	.4293	.8920	.2077	.0382	.4547
PiCANet	.8770	.6846	.0582	.1273	.8710	.7275	.0524	.0995	.7984	.5916	.0540	.1119
PFSNet	.9461	.5938	.0305	.2075	.9434	.6941	.0285	.1625	.8913	.5186	.0366	.2134
EDNet	.9476	.4351	.0265	.2847	.9431	.5357	.0245	.2412	.8913	.3708	.0341	.2851

Table 1: RGB and RGB-D SOD models under adversarial attacks.

Evaluation Metrics

We select MAE and F_β-measure as evaluation metrics. MAE measures pixel-level difference between the saliency map S and ground truth G as

$$\mathrm{MAE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S_{i,j} - G_{i,j}|, \tag{15}$$

where W and H denote the width and height of the saliency map, respectively. Precision is the ratio of ground truth salient pixels in the predicted salient area while recall is the ratio of predicted salient pixels in the ground truth salient area. F_β-measure is defined as

$$F_\beta = \frac{\left(1+\beta^2\right) \times Precision \times Recall}{\beta \times Precision + Recall}, \tag{16}$$

where β^2 is set as 0.3 to emphasize the precision.

Sensitivity of The Models to Adversarial Attacks

We demonstrate the performance of three RGB-D based visual saliency models: 1) BBSNet (Fan et al. 2020f); 2) CoNet (Ji et al. 2020); 3) MobileSal (Wu et al. 2021b) on natural images and PGD adversarial samples which are synthesized with a pretrained BBSNet. For RGB based visual saliency models, we demonstrate the performance of: 1) GateNet (Zhao et al. 2020); 2) PiCANet (Liu, Han, and Yang 2018); 3) PFSNet (Ma, Xia, and Li 2021) 4) EDNet (Wu et al. 2022) on natural images and PGD adversarial samples which are synthesized with a pretrained GateNet.

As shown in Table 1, F_β-measure of BBSNet and GateNet drop .56–.7 when exposed to the adversarial samples. And F_β-measure of EDNet drops .41–.52 in adversarial environment. The adversarial attack reduces F_β-measure of PFSNet by .24–.37 while it lowers F_β of CoNet, MobileSal and PiCANet by .07–.2. As shown in Table 1, MAE of BBSNet and GateNet are increased by .3578 and .4183, respectively, on the adversarial samples. MAE of EDNet are raised by around .251 while that of PFSNet are raised by around .1626. MAE of CoNet, MobileSal and PiCANet change around .05. These results indicate that BBSNet and GateNet suffer most from the adversarial attack since the adversarial samples are synthesized with their own architecture. CoNet, MobileSal, PiCANet, PFSNet and EDNet are affected to different extent, which may depend on the similarity between their architectures and the model used to launch attacks. Overall, adversarial attack has a non-negligible impact on all experimental models.

Comparison with The State-of-the-art

To demonstrate the performance of our method, we present experiments with six state-of-the-art saliency models. Our base models include BBSNet, CoNet, MobileSal, GateNet, PiCANet and PFSNet which cover RGB-D based and RGB based salient object detection. We compare our method with existing defending algorithm ROSA (Li, Li, and Yu 2019) and L2P (Jeddi et al. 2020). ROSA is the most related work which is the first and up to date adversarial defense designed for SOD without adversarial training. We do not experiment on the base models used in ROSA to obtain up to date results. We treat ROSA as a defense method rather than a specific network and conduct experiments on a wide range of models. In all of the above settings, our approach is universally applicable.

As Table 2 and Tabel 3 show, our method achieves better F_β and MAE in most environments. For clean images, LeNo perfroms at most .0328 F_β higher and .0276 MAE lower than second place. For adversarial images, our proposed method is at most .0358 F_β higher and .0199 MAE lower than second place. Here, dataset-adv denotes experimenting on the adversarial samples of corresponding dataset. As Table 7 shows, our method has almost no effect on inference time compared with ROSA. Qualitative results are shown in Figure 6, all the input images are adversarial samples. The visual comparison demonstrates that our proposed LeNo strengthens the robustness of SOD network, and the final saliency maps are closer to GT than others.

Ablation Studies

Contribution of each component. As shown in Table 4, after adding shallow noise to the original model, the new network's performance decreases on clean images slightly, but increases on adversarial images greatly. In the worst case, it dropped by .0523 in F_β and increased by .0231 in MAE on clean images, but at this point there is a .2724F_β and .3016 MAE improvement in performance against adversarial samples. After continuing to add noise estimation, the performance of the model is improved on clean images while it is

	ECSSD		HKU-IS		DUTS-TE		DUT-O		PASCAL		SOD	
	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$
GateNet	.9438	.0332	.9432	.0301	.8920	.0382	.8190	.0545	.8750	.0653	.8714	.0990
GN+ROSA	.9292	.0425	.9275	.0377	.8592	.0486	.7969	.0614	**.8592**	**.0767**	**.8572**	**.1090**
GN+L2P	.8355	.1097	.8244	.0986	.7063	.1185	.6850	.1190	.7661	.1436	.7402	.1774
GN+LeNo	**.9360**	**.0399**	**.9279**	**.0372**	**.8612**	**.0471**	**.7975**	**.0606**	.8514	.0776	.8462	.1116
PiCANet	.9278	.0455	.9214	.0419	.8455	.0525	.7762	.0661	.8490	.0785	.8426	.1082
PN+ROSA	.9022	.0536	.9073	.0514	.8216	.0609	.7659	.0737	.8340	.0905	.8272	.1172
PN+L2P	.9177	.0490	.9157	.0472	**.8338**	**.0575**	.7741	.0727	**.8380**	**.0868**	.8306	.1163
PN+LeNo	**.9264**	**.0482**	**.9183**	**.0464**	.8316	.0597	**.7768**	**.0704**	.8347	.0891	**.8337**	**.1145**
PFSNet	.9461	.0305	.9434	.0285	.8913	.0366	.8197	.0553	.8740	.0637	.8815	.0890
PFSN+ROSA	.8778	.0698	.8943	.0491	.8193	.0592	.7725	**.0615**	.8101	.0982	.8122	.1386
PFSN+L2P	.9222	.0525	.9031	.0496	.8242	.0620	.7751	.0683	.8194	.0988	.8190	.1365
PFSN+LeNo	**.9263**	**.0391**	**.9249**	**.0387**	**.8525**	**.0521**	**.7929**	.0656	**.8373**	**.0840**	**.8518**	**.1089**

Table 2: Comparison with the state-of-the-art for RGB based models on clean images

		PGD ATTACK				ROSA ATTACK				FGSM ATTACK			
Dataset	Metric	GateN.	+ROSA	+L2P	+LeNo	GateN.	+ROSA	+L2P	+LeNo	GateN.	+ROSA	+L2P	+LeNo
HKU-adv	$F_\beta \uparrow$	.3127	.5544	.6515	**.6538**	.7998	.9210	.8171	**.9262**	.7720	.8314	.7762	**.8379**
	MAE$\downarrow$	.4293	.2532	.1857	**.1842**	.1118	.0420	.1009	**.0394**	.0967	.0769	.1142	**.0739**
DUTS-adv	$F_\beta \uparrow$	.2077	.3757	.4934	**.5019**	.7909	.8603	.7230	**.8606**	.6503	.7095	.6534	**.7284**
	MAE$\downarrow$	.4547	.3184	.2450	**.2285**	.0909	.0501	.1194	**.0479**	.1123	.1016	.1306	**.0936**
DUT-O-adv	$F_\beta \uparrow$	.1848	.2930	.4055	**.4170**	.6825	.7935	.6801	**.7981**	.5808	.6273	.6393	**.6442**
	MAE$\downarrow$	.4882	.3647	.2760	**.2725**	.1002	.0594	.1207	**.0578**	.1249	.1215	.1294	**.1160**
PASC.-adv	$F_\beta \uparrow$	.2915	.4645	**.5997**	.5697	.7480	**.8505**	.7635	.8498	.6798	.7361	.7268	**.7514**
	MAE$\downarrow$	.4696	.3300	**.1761**	.2509	.1608	.0855	.1451	**.0821**	.1603	.1375	.1629	**.1293**
Dataset	Metric	PiCAN.	+ROSA	+L2P	+LeNo	PiCAN.	+ROSA	+L2P	+LeNo	PiCAN.	+ROSA	+L2P	+LeNo
HKU-adv	$F_\beta \uparrow$	.8081	.8324	.8343	**.8468**	.9172	.9031	.9129	**.9133**	.8498	.8604	.8635	**.8712**
	MAE$\downarrow$	.0892	.0819	.0811	**.0782**	.0443	.0522	.0497	**.0488**	.0670	.0668	.0638	**.0635**
DUTS-adv	$F_\beta \uparrow$	.6690	.6971	.7055	**.7188**	.8415	.8172	.8277	**.8279**	.7431	.7501	.7594	**.7629**
	MAE$\downarrow$	.1177	.1108	.1091	**.1051**	.0539	.0629	.0615	**.0612**	.0858	.0843	.0829	**.0827**
DUT-O-adv	$F_\beta \uparrow$	.6167	.6416	.6460	**.6658**	.7771	.7632	.7725	**.7739**	.6879	.7003	.7045	**.7102**
	MAE$\downarrow$	.1287	.1269	.1238	**.1195**	.0659	.0748	**.0706**	.0731	.0933	.0956	**.0937**	.0946
PASC.adv	$F_\beta \uparrow$	.7159	.7419	.7416	**.7495**	.8420	.8279	**.8326**	.8301	.7757	.7761	.7865	**.7866**
	MAE$\downarrow$	.1569	.1476	.1440	**.1416**	.0836	.0954	**.0902**	.0924	.1201	.1204	**.1147**	.1172
Dataset	Metric	PFSN.	+ROSA	+L2P	+LeNo	PFSN.	+ROSA	+L2P	+LeNo	PFSN.	+ROSA	+L2P	+LeNo
HKU-adv	$F_\beta \uparrow$	.6941	.7716	.8151	**.8291**	.9363	.8826	.8962	**.9230**	.8447	.8209	.8282	**.8640**
	MAE$\downarrow$	.1625	.0997	.0871	**.0809**	.0344	.0560	.0530	**.0398**	.0681	.0777	.0827	**.0628**
DUTS-adv	$F_\beta \uparrow$	.5186	.6168	.6927	**.6992**	.8886	.8053	.8149	**.8480**	.7347	.6969	.7205	**.7541**
	MAE$\downarrow$	.2134	.1292	.1203	**.1146**	.0395	.0652	.0644	**.0544**	.0863	.0933	.0943	**.0852**
DUT-O-adv	$F_\beta \uparrow$	.4571	.5879	.6383	**.6511**	.8205	.7652	.7665	**.7951**	.6700	.6685	.6987	**.7036**
	MAE$\downarrow$	.2417	.1318	.1295	**.1284**	.0515	**.0639**	.0697	.0655	.0985	.0909	**.0887**	.0941
PASC.-adv	$F_\beta \uparrow$	.5998	.6564	**.7221**	.7143	.8611	.7944	.8162	**.8369**	.7512	.7253	.7483	**.7612**
	MAE$\downarrow$	.2247	.1779	**.1543**	.1611	.0717	.1077	.1019	**.0878**	.1259	.1404	.1423	**.1290**

Table 3: Comparison with the state-of-the-art for RGB based models on adversarial images

	DUTS-D				NLPR				SIP			
	$F_\beta \uparrow$		MAE$\downarrow$		$F_\beta \uparrow$		MAE$\downarrow$		$F_\beta \uparrow$		MAE$\downarrow$	
	clean	adver	clean	adver	clean	adver	clean	adver	clean	adver	clean	adver
BBSNet	.8117	.2473	.0828	.5398	.8991	.1976	.0286	.4492	.8780	.2718	.0589	.3755
+Shallow noi.	.7594	.5197	.1059	.2382	.8535	.7580	.0413	.0767	.8599	.7395	.0689	.1184
+Noi. estimation	.7852	.5125	.0977	.2413	.8744	.7574	.0385	.0819	.8628	.7317	.0681	.1245

Table 4: The ablation study of different components on the validations set of DUTS-D, NLPR and SIP

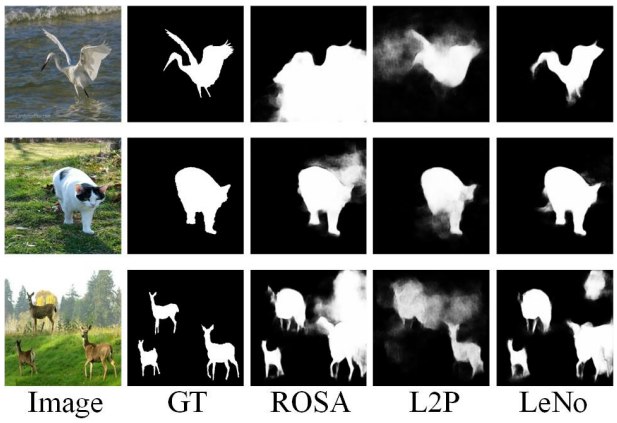

Figure 6: Qualitative comparison with the state-of-the-art. The leftmost column is the adversarial images. The second column is the ground truth. The third, the fourth and the last column denote GateNet prediction incorporated with ROSA, L2P and LeNo defense respectively.

basically maintained on adversarial images.

Noise initialization. There are three different noise initialization settings. We test 1) uniform distribution with zero mean and unit variance, 2) gaussian distribution with zero mean and variance 0.734 and 3) constant setting of .5. Such a choice of parameters ensures their average perturbation intensity is equal. As Table 5 shows, Gaussian distribution is the optimal choice in most cases.

Metric	$F_\beta \uparrow$	MAE↓	$F_\beta \uparrow$	MAE↓	$F_\beta \uparrow$	MAE↓
DS	NJU2K-adv		SIP-adv		DUTS-D-adv	
Unif.	.7775	**.0927**	.6910	.1257	.4734	.1899
Cons.	.7652	.0979	.6768	.1303	.4922	.1906
Gaus.	**.7876**	.0947	**.7252**	**.1187**	**.5282**	**.1786**
DS	DUTS-TE-adv		DUT-O-adv		PASCAL-adv	
Unif.	.6773	.1078	.6017	.1255	.7076	.1491
Cons.	.6754	.1074	.5981	.1240	.7039	.1507
Gaus.	**.6950**	**.1030**	**.6162**	**.1212**	**.7201**	**.1448**

Table 5: Effect of different noise initial distribution

Number of noise layer. As Table 6 shows, whether on the clean images or the adversarial images, single noise layer is always the best. Superabundant noise may hamper the final prediction. Here, +1 denotes adding noise layer after every BottleNeck in stem and stage1. And +2 denotes adding noise layer after every BottleNeck in stem, stage1 and stage2.

Noise area. Finally, we investigate how the noise area affects the performance. We report the results in FGSM adversarial environment. Here, network-full denotes that all elements of the feature map are added with effective noise, which is adopted by (Jeddi et al. 2020). Network-center denotes only the elements in the center area are added with effective noise, which is an intuitive consideration for Salient Object Detection. Here, the central area is a square with the sides halved. Salient objects usually appear at the center part

Method	Metric	NJ.-adv	SIP-adv	NJU2K	SIP
+1	$F_\beta \uparrow$	.7677	**.7521**	.7956	.7838
	MAE↓	.1104	.1187	.1002	.1078
+2	$F_\beta \uparrow$	.7521	.7261	.7743	.7519
	MAE↓	.1176	.1348	.1124	.1255
LeNo	$F_\beta \uparrow$	**.7956**	.7345	**.8207**	**.7852**
	MAE↓	**.0929**	**.1170**	**.0815**	**.0988**
Method	Metric	EC.-adv	HKU-adv	ECSSD	HKU
+1	$F_\beta \uparrow$	.8310	.8299	.8687	.8678
	MAE↓	.1062	.0935	.0922	.0826
+2	$F_\beta \uparrow$	.7993	.8113	.7993	.8113
	MAE↓	.1200	.1015	.1200	.1015
LeNo	$F_\beta \uparrow$	**.8775**	**.8712**	**.9264**	**.9183**
	MAE↓	**.0732**	**.0635**	**.0482**	**.0472**

Table 6: Effect of number of noise layer

Model	Inference Time (ms)		
	w/o defense	ROSA	LeNo
BBS	28.9	+616.7	**+0.5**
CoNet	26.6	+341.2	**+0.2**
GateNet	40.5	+1022.2	**+0.2**
PiCANet	166.6	+737.2	**+0.1**
PFSNet	36.4	+550.3	**+0.3**
MobileSal	16.8	+270.8	**+0.1**

Table 7: Comparison with the state-of-the-art on speed, each inference time is averaged by 10 times computing.

of the image, so we make the central area the focus of our defense. Adding stronger noise at the center of the feature map makes the defense more targeted. In order to balance the noise strength and computational complexity, we propose the cross-shaped noise as Fig. 5 shows. Experiments show that cross-shaped noise is the optimal.

Inference time. The proposed LeNo defense can be computed in parallel since it is embedded in the encoder and decoder of the network. The extra time of ROSA is introduced by the summation of superpixel shuffle (Achanta et al. 2012), CRFRNN (Zheng et al. 2015) and bilateral filter. As Table 7 shows, LeNo is significantly faster than ROSA on all base models.

Conclusion

In recent years, salient object detection with deep neural network has achieved good performance. However, it is sensitive to adversarial attacks, which rises a huge security challenge. In this paper, a lightweight Learnable Noise (LeNo) is proposed against adversarial attacks for SOD models. Different from the previous work ROSA which relies on various pre- and post-processings, the shallow noise with a cross-shaped Gaussian distribution and noise estimation for refining the feature are introduced to replace SWS and CAR components to speed up inference. Extensive experiments show that LeNo gets better result in RGB and RGB-D SOD benchmarks with much lower inference latency.

Acknowledgments

This work was supported by the National Natural Science Foundation of China Grant 61902139.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11): 2274–2282.

Che, Z.; Borji, A.; Zhai, G.; Ling, S.; Li, J.; Tian, Y.; Guo, G.; and Le Callet, P. 2021. Adversarial attack against deep saliency models powered by non-redundant priors. *IEEE Transactions on Image Processing*, 30: 1973–1988.

Chen, H.; Li, Y.; Deng, Y.; and Lin, G. 2021. CNN-Based RGB-D Salient Object Detection: Learn, Select, and Fuse. *International Journal of Computer Vision*, 129(7): 2076–2096.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.

Dhillon, G. S.; Azizzadenesheli, K.; Lipton, Z. C.; Bernstein, J.; Kossaifi, J.; Khanna, A.; and Anandkumar, A. 2018. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020a. Camouflaged Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020b. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, 263–273. Springer.

Fan, D.-P.; Lin, Z.; Zhang, Z.; Zhu, M.; and Cheng, M.-M. 2020c. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems*, 32(5): 2075–2089.

Fan, D.-P.; Lin, Z.; Zhang, Z.; Zhu, M.; and Cheng, M.-M. 2020d. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems*, 32(5): 2075–2089.

Fan, D.-P.; Lin, Z.; Zhang, Z.; Zhu, M.; and Cheng, M.-M. 2020e. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems*, 32(5): 2075–2089.

Fan, D.-P.; Zhai, Y.; Borji, A.; Yang, J.; and Shao, L. 2020f. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In *European conference on computer vision*, 275–292. Springer.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guo, S.; Yan, Z.; Zhang, K.; Zuo, W.; and Zhang, L. 2019. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1712–1722.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jeddi, A.; Shafiee, M. J.; Karg, M.; Scharfenberger, C.; and Wong, A. 2020. Learn2perturb: an end-to-end feature perturbation learning to improve adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1241–1250.

Ji, W.; Li, J.; Zhang, M.; Piao, Y.; and Lu, H. 2020. Accurate RGB-D salient object detection via collaborative learning. In *European Conference on Computer Vision*, 52–69. Springer.

Ju, R.; Ge, L.; Geng, W.; Ren, T.; and Wu, G. 2014a. Depth saliency based on anisotropic center-surround difference. In *2014 IEEE international conference on image processing (ICIP)*, 1115–1119. IEEE.

Ju, R.; Ge, L.; Geng, W.; Ren, T.; and Wu, G. 2014b. Depth saliency based on anisotropic center-surround difference. In *2014 IEEE international conference on image processing (ICIP)*, 1115–1119. IEEE.

Li, G.; and Yu, Y. 2015. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5455–5463.

Li, H.; Li, G.; and Yu, Y. 2019. ROSA: Robust salient object detection against adversarial attacks. *IEEE transactions on cybernetics*, 50(11): 4835–4847.

Li, N.; Ye, J.; Ji, Y.; Ling, H.; and Yu, J. 2014a. Saliency detection on light field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2806–2813.

Li, X.; Yang, F.; Cheng, H.; Liu, W.; and Shen, D. 2018. Contour knowledge transfer for salient object detection. In *Proceedings of the european conference on computer vision (ECCV)*, 355–370.

Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; and Yuille, A. L. 2014b. The secrets of salient object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 280–287.

Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Hu, X.; and Zhu, J. 2018. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1778–1787.

Liu, N.; Han, J.; and Yang, M.-H. 2018. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3089–3098.

Liu, X.; Cheng, M.; Zhang, H.; and Hsieh, C.-J. 2018. Towards robust neural networks via random self-ensemble. In

Proceedings of the European Conference on Computer Vision (ECCV), 369–385.

Ma, M.; Xia, C.; and Li, J. 2021. Pyramidal feature shrinking for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2311–2318.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, 416–423.

Niu, Y.; Geng, Y.; Li, X.; and Liu, F. 2012. Leveraging stereopsis for saliency analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 454–461. IEEE.

Peng, H.; Li, B.; Xiong, W.; Hu, W.; and Ji, R. 2014. RGBD salient object detection: A benchmark and algorithms. In *European conference on computer vision*, 92–109. Springer.

Piao, Y.; Ji, W.; Li, J.; Zhang, M.; and Lu, H. 2019. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7254–7263.

Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S. 2018. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4422–4431.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7479–7489.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Shi, J.; Yan, Q.; Xu, L.; and Jia, J. 2015. Hierarchical image saliency detection on extended CSSD. *IEEE transactions on pattern analysis and machine intelligence*, 38(4): 717–729.

Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 136–145.

Wu, Y.-H.; Gao, S.-H.; Mei, J.; Xu, J.; Fan, D.-P.; Zhang, R.-G.; and Cheng, M.-M. 2021a. Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. *IEEE Transactions on Image Processing*, 30: 3113–3126.

Wu, Y.-H.; Liu, Y.; Xu, J.; Bian, J.-W.; Gu, Y.-C.; and Cheng, M.-M. 2021b. MobileSal: Extremely efficient RGB-D salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wu, Y.-H.; Liu, Y.; Zhang, L.; Cheng, M.-M.; and Ren, B. 2022. EDN: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing*, 31: 3125–3136.

Xiao, C.; Zhu, J.-Y.; Li, B.; He, W.; Liu, M.; and Song, D. 2018. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*.

Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3166–3173.

Zhao, J.-X.; Liu, J.-J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019. EGNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8779–8788.

Zhao, X.; Pang, Y.; Zhang, L.; Lu, H.; and Zhang, L. 2020. Suppress and balance: A simple gated network for salient object detection. In *European conference on computer vision*, 35–51. Springer.

Zhao, Z.; Dua, D.; and Singh, S. 2017. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*.

Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; and Torr, P. H. 2015. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, 1529–1537.