ROSA: Robust Salient Object Detection Against Adversarial Attacks

Haofeng Li, Guanbin Li, *Member, IEEE*, and Yizhou Yu, *Fellow, IEEE*

***Abstract*—Recently, salient object detection has witnessed remarkable improvement owing to the deep convolutional neural networks which can harvest powerful features for images. In particular, the state-of-the-art salient object detection methods enjoy high accuracy and efficiency from fully convolutional network (FCN)-based frameworks which are trained from end to end and predict pixel-wise labels. However, such framework suffers from adversarial attacks which confuse neural networks via adding quasi-imperceptible noises to input images without changing the ground truth annotated by human subjects. To our knowledge, this paper is the first one that mounts successful adversarial attacks on salient object detection models and verifies that adversarial samples are effective on a wide range of existing methods. Furthermore, this paper proposes a novel end-to-end trainable framework to enhance the robustness for arbitrary FCN-based salient object detection models against adversarial attacks. The proposed framework adopts a novel idea that first introduces some new generic noise to destroy adversarial perturbations, and then learns to predict saliency maps for input images with the introduced noise. Specifically, our proposed method consists of a segment-wise shielding component, which preserves boundaries and destroys delicate adversarial noise patterns and a context-aware restoration component, which refines saliency maps through global contrast modeling. The experimental results suggest that our proposed framework improves the performance significantly for state-of-the-art models on a series of datasets.**

***Index Terms*—Adversarial attack, deep neural network, salient object detection.**

I. Introduction

SALIENT object detection aims at locating and segmenting objects, which are most visually distinctive to human subjects, in an image or a video frame. Designing a salient object detection model for simulating this process not only improves our understanding of the inner mechanism of human vision and psychology, but also benefits many applications in the field of computer vision and graphics. For example, salient object detection has been widely studied and applied to robotics [1]; context-aware image editing [2]; object segmentation [3], [4]; and person reidentification [5]. Since salient object detection algorithms are usually adopted during the initialization or preprocessing stage of a system, efficiency and robustness are of considerable importance. Imagine, if the performance of the preprocessing stage is seriously affected by corrupt input images, succeeding stages might produce unpromising results, which could be a catastrophe to the entire system.

For the last several years, significant successes have been achieved in the computer vision community, as training deep convolutional neural networks (CNNs) on large-scale datasets becomes feasible. A deep CNN is composed of stacked convolution filters with learnable parameters. Since those filters harvest information naturally from local neighborhoods in the input image and their parameters are adaptively determined by a training set, deep CNNs demonstrate a high fitting capacity superior to traditional methods using handcrafted features. These days, deep learning has been widely employed in image classification, semantic segmentation, object localization, as well as salient object detection.

Deep learning-based salient object detection models can be roughly divided into two groups. One group adopts segment-wise labeling while the other group predicts pixel-level results. Segment-wise labeling methods first divide an image into regions. Pixels in the same region most probably share similar saliency values. CNN features for each region are then extracted to evaluate its saliency. In contrast, pixel-wise methods usually embrace fully convolutional network (FCN) architectures, which take an entire image as input and yield a dense saliency map directly. Such methods not only demonstrate higher efficiency but also achieve state-of-the-art accuracy in virtue of their end-to-end trainable property.

However, those FCN-driven approaches have weaknesses that might degrade their performance in practice. First, being end-to-end trainable allows gradients propagated easily from supervision target to the input image, which puts the salient object detection models at the risk of adversarial attacks. Adversarial attacks generate adversarial samples that do not change the ground truth assigned by human subjects but increase the prediction error of neural models by making visually imperceptible changes to the image, as shown in Fig. 1. Second, dense labeling models do not explicitly model contrast among different image parts but implicitly estimate saliency in a single FCN. Once input images are polluted by adversarial noise, low-level features and high-level features that cannot

Manuscript received December 5, 2018; revised April 17, 2019; accepted April 26, 2019. Date of publication May 17, 2019; date of current version October 26, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61702565 and Grant U1811463, and in part by the Science and Technology Planning Project of Guangdong Province under Grant 2017B010116001. This paper was recommended by Associate Editor H. Lu. *(Corresponding author: Guanbin Li.)*

H. Li is with the Department of Computer Science, University of Hong Kong, Hong Kong (e-mail: lhaof@foxmail.com).

G. Li is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: liguanbin@mail.sysu.edu.cn).

Y. Yu is with the Department of Computer Science, University of Hong Kong, Hong Kong, and also with AI Lab, Deepwise Healthcare, Beijing 100080, China (e-mail: yizhouy@acm.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2019.2914099

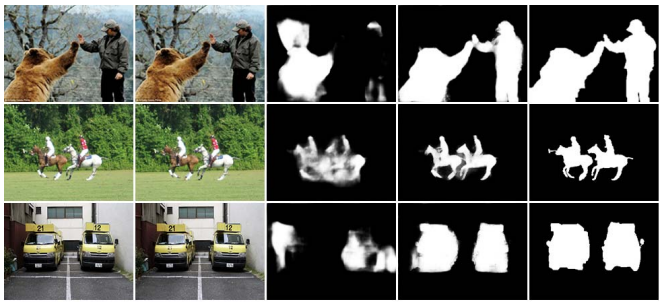


Fig. 1. Effectiveness of our proposed method. The leftmost column shows the original images while the second column from the left displays the corresponding adversarial samples. The L_∞-norm of the adversarial perturbations is set as 25 pixel values. Column (a) is the saliency maps of adversarial samples, and is predicted by DSS [6]. Column (b) is the saliency maps of adversarial samples, and is predicted by our proposed method with DSS as the backbone network stream. The rightmost column is the ground truth saliency maps. As can be seen in the above, the adversarial samples are almost visually the same as their original images. Besides, DSS incorporated with our proposed method yields saliency maps of higher quality, in comparison to the original DSS.

correct themselves will be affected as well. Third, current largest training datasets of salient object detection contain only several thousands of images, in comparison to some image classification benchmark with millions of samples [7]. At the same time, the salient object categories included are very limited. Thus, to some extent, existing models are fitting bias within the data, for example, detecting objects frequently appearing in training set rather than locating the most distinctive ones. Those approaches might rely on capturing too much high-level semantics and could be sensitive to low-level perturbations, such as adversarial noises.

Segment-wise labeling approaches enjoy higher robustness as they model contrast explicitly and determine saliency score depending on multiple regions, such as the considered segment and its context. For different segments, gradients calculated by the same target might conflict when propagating on the input image, since different regions could share the same local or global context. Nevertheless, it is inefficient to adopt sparse labeling methods in practice, due to evaluating hundreds of segments.

To enhance the robustness and maintain efficiency for existing dense labeling methods, this paper proposes a novel framework robust saliency (ROSA) that can take any FCN as a backbone. We first observe that adversarial noise itself is fragile as it is computed accurately by backward propagation. Adversarial noise forms some subtle curve-like pattern that may play an important role. Destructing such patterns could reduce the attack effects. And then we notice that CNNs are less sensitive to some generic noise than the adversarial noises, since adversarial samples are aimed at neural models. We also consider *a priori* that nearby pixels with similar low-level features have similar saliency values. Thus, we come up with a novel framework that first destroys adversarial perturbations by introducing some new generic noise, and then learns to adaptively predict saliency maps against the new introduced noise. To destroy adversarial noises, we develop a segment-wise shielding component placed before the backbone network. Segment-wise shielding component divides an image into small parts according to low-level similarity and shuffles pixels in each part randomly. It introduces another generic noise to destroy the structural pattern in adversarial samples and, therefore, alleviates the attack effect. To refine results affected by the newly introduced noise, we conceive another component known as context-aware restoration placed after the backbone network. The restoration component adjusts the saliency score at some position according to similarities among raw pixel values of the position and its context. The overall system with a backbone network is fine-tuned end-to-end in the training stage.

Our proposed framework demonstrates several strengths in the following. First, the ROSA framework is not so susceptible to adversarial attacks. Since the shielding component has no learnable parameters, it does not support backward propagating gradients onto the input image to generate adversarial samples. Even when adversarial samples are found, their adversarial noise can still be destroyed by the shielding component during the testing stage. Second, the shielding component shuffles pixels in the same segment and thus does boundary less harm. Moreover, ROSA adopts an FCN-based model as its backbone and the restoration component is implemented by convolutional operator that supports parallel computing. Both designs help maintain acceptable efficiency for the entire system.

In short, our contributions have three-folds.

1) We, for the first time, launch adversarial attacks on the state-of-the-art salient object detection models successfully.
2) We propose a novel salient object detection framework that first introduces some new noise to resist adversarial perturbations, and then adaptively predicts saliency maps for inputs with the new introduced noises. The proposed framework is instantiated by an arbitrary FCN backbone and two strongly coupled and complementary components.
3) Experimental results verify that the implemented adversarial attacks are effective for a wide range of existing salient object detection models. Moreover, extensive experiments demonstrate that the our proposed framework is resistant to adversarial samples, and more robust than existing defense baselines.

II. Related Work

In this section, we brief several groups of previous work related to our proposed approach, salient object detection, adversarial attacks, and defenses against adversarial attacks.

A. Salient Object Detection

Algorithms for detecting salient objects can be separated into two categories. One category is the conventional methods that do not use neural networks but resort to prior knowledge and handcrafted features [2], [8]–[17]. Ranking saliency [16]

is a saliency detection algorithm based on graph-based manifold ranking, which ranks the relevances of images elements to foreground or background seeds. Another category driven by deep CNNs can be categorized as two groups: 1) sparse labeling and 2) dense labeling. Sparse labeling methods [18], [19] appeared in early years. Li and Yu [18], [20] trained a binary classifier to estimate visual saliency for each superpixel with multiscale learned CNN features. Wang *et al.* [21] developed a local DNN estimating coarse saliency for object proposals and a global DNN evaluating weights to combine different proposals. Zhao *et al.* [22] employed a deep CNN to predict visual saliency for single superpixel with local context and global context. Qin *et al.* [23] introduced a single-layer cellular automata (SCA) which can exploit the intrinsic relevance of similar image regions to detect salient objects, based on extracted deep features. Since these methods take a region as a unit of computation and contain two separate steps of feature extraction and salient value inference, they are generally inefficient and require a large amount of space for feature storage. Inspired by the successful application of FCNs in pixel-level semantic segmentation; recently, dense labeling approaches have established the new state-of-the-art in salient object detection [24]–[33]. Li and Yu [34] modeled visual saliency by combining a fully convolutional stream with a segment-wise spatial pooling stream. Wang *et al.* [35] employed FCNs to refine coarse saliency maps based on prior knowledge in a recurrent way. Hou *et al.* [6] adapted holistically nested edge detector (HED) [36] architecture by introducing short connections to the skip-layer structures.

B. Adversarial Attack

Existing adversarial attacks consist of several groups, one-step gradient-based methods [37]; iterative methods [38]–[44]; optimization-based methods [45], [46]; and generative networks [47], [48]-based methods. The fast gradient sign method (FGSM) [37] computes one-step gradient to maximize the loss $L(\cdot)$ between the model output and the ground truth, within some L_∞-norm bound ϵ. FGSM generates adversarial sample as

$$x^* = x + \epsilon \cdot \text{sign}(\bigtriangledown_x L(f(x;\theta), y)) \tag{1}$$

where x^*, x, and y are the adversarial sample, original image, and ground truth, respectively. $f(\cdot;\theta)$ denotes some neural model with parameters θ. Iterative approaches [41], [44] conduct FGSM multiple times with a small step length α as

$$x^*_{t+1} = \text{clip}\big(x^*_t + \alpha \cdot \text{sign}(\bigtriangledown_x L(f(x;\theta), y)), \epsilon\big) \tag{2}$$

where x^*_t denotes an adversarial sample obtained at t-th time step. x^*_0 is initialized as x. $\text{clip}(x, \epsilon)$ keeps each element x_i of x within the range of $[x_i - \epsilon, x_i + \epsilon]$. Szegedy *et al.* [49] solved a box-constrained optimization with L-BFGS to find an adversarial sample. Dong *et al.* [38] proposed an iterative algorithm that integrates a momentum term into the iterative process to boost adversarial attacks. Cisse *et al.* [40] proposed an approach called "Houdini" to attack structured prediction problems (including human pose estimation and speech recognition) whose final performance measure is a combinatorial nondecomposable quantity. Dai *et al.* [50] proposed to fool a family of graph neural networks by modifying the combinatorial structure of data. They developed a reinforcement learning-based methods, variants of genetic algorithms, and gradient-based methods to attack graph neural networks. Adversarial attack on salient object detection remains a gap before this paper.

C. Defense Against Adversarial Attacks

Some defense methods are proposed to protect attacked target neural networks from potential adversarial samples [51]–[58]. Metzen *et al.* [52] augmented the attacked target network by small subnetworks, which take the output feature maps at some layers as inputs, and predict a probability of the input containing adversarial noise. SafetyNet [53] equips a CNN classifier with an RBF-SVM to detect adversarial samples with discrete codes calculated from the final RELU outputs. Images transformations, including bit quantization, vectorization [59], JPEG compression, and total variance minimization may remove or destroy adversarial perturbations [54], before feeding an input image into the target network. Xie *et al.* [55] proposed a simple method that randomly resizes an input image and pads it with zeros, to destroy the effect of adversarial attacks. Liao *et al.* [57] developed a neural network-based denoiser that is trained with a loss function based on some high-level features of the attacked target classifier. Many existing defense baselines struggle to remove potential adversarial noises from input images, which is different from our proposed idea that adaptively predicts saliency maps for inputs with some new introduced generic noise to resist adversarial attacks.

III. Method

This section first describes how we launch an adversarial attack on the state-of-the-art visual saliency models, and then detail how our proposed robust salient object detection framework works.

A. Adversarial Samples for Salient Object Detection

Adversarial attack aims at synthesizing some perturbed input that fools neural models without changing its ground truth label. In this section, we introduce the pipeline to yield an adversarial sample for a given salient object detection model f, and it can be directly used to attack other detection models as most of the existing visual saliency models have similar FCN-based network architectures and are usually initialized by the same pretrained image classification model [60], [61]. Adversarial samples can be divided into two categories. *Targeted* adversarial samples make the attacked models produce specific results as predicted saliency maps while *nontargeted* ones maximize the mean absolute error (MAE) and/or minimize F_β-measure. In this paper, only nontargeted attacks are concerned and targeted samples may be investigated in the future.

Inspired by [62], we implement an iterative gradient-based pipeline to synthesize adversarial samples. To generate those samples, it requires a neural network pretrained on salient

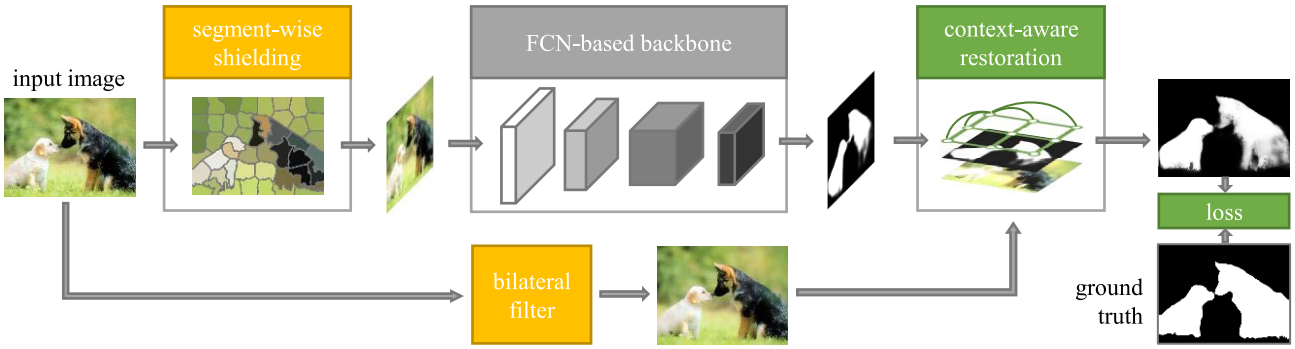


Fig. 2. ROSA, robust salient object detection framework. To defend against adversarial samples, the proposed method exploits a novel idea that first introduces some novel generic noise to destroy adversarial perturbations, and then learns to predict saliency maps for images with the introduced generic noise. Such idea is different from those resorting to image smoothing or transformations before feeding input images into target networks. As shown in the above, the segment-wise shielding component introduces some generic noise to perturb the adversarial noises. Then, an FCN-based backbone takes the noisy image as input, and yields a coarse saliency map. Afterward, a context-aware restoration component utilizes a graph model to refine the coarse saliency map, with a smoothed image providing pairwise pixel-level similarity. The smoothed image is obtained by applying bilateral filter on the input image. Lastly, a pixel-wise binary cross-entropy loss function is calculated between the refined saliency map and the ground truth. The FCN-based backbone network and the context-aware restoration component are end-to-end trained to adapt to the input images with the introduced noise. The proposed method can theoretically incorporate arbitrary FCN-based backbone network.

object detection, some natural images and their corresponding saliency maps densely labeled at pixel level. Let $f(\cdot, \theta)$ be the pretrained model with parameters θ. x, x^*, and y denote a natural image, its corresponding adversarial sample, and ground truth, respectively. Before synthesizing the adversarial sample, x is subtracted by mean pixel values. After the generation, x^* is enlarged to the range of [0, 255] and rounded to RGB image. Each element y_i of y belongs to {0, 1}, with 0 denoting nonsalient and 1 denoting salient. To ensure the adversarial perturbation unnoticeable, parameter ϵ is set as upper bound of L_∞-norm such that $||x - x^*|| \leq \epsilon$. The maximum number of iterations T limits the overall running time cost. Once T iterations are finished or the L_∞-norm bound is reached, the generation stops and returns adversarial sample obtained at the current time step.

In each iteration t, supposing that adversarial sample x_t^* from previous time step or initialization is prepared, we update the adversarial sample as

$$x_0^* = x, x_{t+1}^* = x_t^* + p_t \tag{3}$$

where p_t denotes the adversarial perturbation computed at t-th step. We formulate the goal making the predictions of all pixels in x go wrong as $\forall i, \text{argmax}_c\{f_{i,c}(x_t^* + p_t; \theta)\} \neq y_i$. Here, i denotes one of all n pixels in x and c denotes two categories: salient and nonsalient. To determine p_t, gradient descent algorithm is applied as

$$p_t' = \sum_{i \in S_t} \left[\nabla_{x_t^*} f_{i,1-y_i}(x_t^*; \theta) - \nabla_{x_t^*} f_{i,y_i}(x_t^*; \theta)\right] \tag{4}$$

where S_t denotes the set of pixels that f still can classify correctly. Then, p_t is obtained by normalization as $\alpha \cdot p_t'/||p_t'||_\infty$, where α is a fixed step length. The pseudocode of the entire generation pipeline is shown in Algorithm 1.

Algorithm 1 Adversarial Sample Generation

Require: natural image x;
corresponding saliency annotation y;
pre-trained visual saliency model $f(\cdot; \theta)$;
pixels set $S = \{1, 2, \ldots, n\}$ of x;
maximum number of iterations T;
step length α; upper bound ϵ of L_∞ norm;
$x_0^* \leftarrow x, p \leftarrow 0, t \leftarrow 0, e \leftarrow 0, S_0 = S$;
while $t < T$ and $e \leq \epsilon$ and $|S_t| > 0$ **do**
  calculate p_t' by Equation 4;
  $p_t \leftarrow \alpha \cdot p_t'/||p_t'||_\infty$;
  $p \leftarrow p + p_t$;
  calculate x_{t+1}^* by Equation 3;
  $e \leftarrow ||x_{t+1}^* - x||_\infty$;
  $t \leftarrow t + 1$;
  $S_t \leftarrow \{i | argmax_c\{f_{i,c}(x_t^*; \theta)\} = y_i\}$;
end while
$x^* \leftarrow x_t^* + \bar{x}$;
$x^* \leftarrow round(x^*)$;
return x^*;

B. Robust Salient Object Detection Framework

In this section, we propose a novel salient object detection framework ROSA that demonstrates the high robustness against adversarial attacks. As shown in Fig. 2, the ROSA framework consists of a segment-wise shielding component, an FCN-based backbone network, and a context-aware restoration component. Virtually, the backbone can be chosen as an arbitrary FCN-based visual saliency model that takes an entire image as input and yields a densely labeled saliency map. The FCN backbone enjoys high efficiency and accuracy but displays sensitivity on adversarial samples. The shielding component and the restoration component play an important role in improving the robustness of the proposed framework.

A segment-wise shielding component destroys potential adversarial noise patterns in an input image before sending it to the backbone, by introducing some "shuffling" noise that is easier to resist. The observation behind is that adversarial noises are some delicate perturbations deliberately synthesized

for CNNs, while CNN is not sensitive to and can adapt to some other noise. To alleviate harms caused by the new noise, the shielding component first divides the input image into nonoverlapping regions, namely superpixels. We follow the region decomposition method developed by [63]. Specifically, k cluster centers in the joint space of color and pixel position are initialized by sampling pixels at regular grid steps. Then, we assign each pixel to the cluster center with minimum distance and update each cluster center as the mean vector of pixels belonging to the cluster, in an iterative way. The iteration ends when L2-norm error between new location and previous location of each cluster center converges.

After the region decomposition, we permute all pixels within the same superpixel randomly. Such shuffling operation strongly destroy the adversarial perturbation while it limits the introduced noise within each single superpixel. Thus, object boundaries that those superpixels are adhere to are not spoiled and the noisy saliency map output by the backbone network has a chance to be restored. Some may suggest an option that smooths each superpixel by averaging pixels inside. Recall what we argue in Section I, existing FCN models overfit too much high-level semantics in visual saliency data. The random permutation makes capturing high-level semantics more difficult and enforces neural networks to harvest low-level contrast among regions. It also plays a role in augmenting dataset and reducing the overfitting issue.

A context-aware restoration component exploits low-level similarity between each pixel and its context to refine the saliency scores provided by the backbone network. As adversarial perturbations aim at parameterized convolution filters, the restoration component adopts a complete graph model instead of CNN architecture. We measure similarity among pixels in low-level color space and spatial position, since previous high-level convolutional features have been polluted. The restoration component adjusts saliency maps by minimizing some energy function as

$$E(y^*) = \sum_i E_u(y_i^*, y_i) + \sum_{i<j} E_p\left(y_i^*, y_j^*\right) \tag{5}$$

where y denotes the coarse saliency map and y^* denotes the resulted saliency map. The first unary energy term measures the cost (inverse likelihood) of assigning i with y_i^*. The second term pairwise energy measures the cost of assigning i and j with y_i^* and y_j^* at the same time. It encourages similar nearby pixels to be labeled the same. The pairwise energy is defined as (6) where p denotes the pixel position and x_i' denotes the pixel color. x' is a smoothed image output by the bilateral filter that takes the adversarial sample x^* as input, as shown in Fig. 2. ω_1 and ω_2 are tuned by training. θ_α, θ_β, and θ_γ are chosen as 160, 3, and 3, respectively. μ is a learnable label compatibility function that penalizes assigning i and j with different labels

$$E_p\left(y_i^*, y_j^*\right) = \mu\left(y_i^*, y_j^*\right)\left\{\omega_1 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|x_i' - x_j'|^2}{2\theta_\beta^2}\right) + \omega_2 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right)\right\}. \tag{6}$$

We realize the component following some previous work [64], [65] that solves (5) as densely connected conditional random field with a recurrent neural network. The neural network is implemented with and enjoys efficiency from 1×1 convolutional layers. Since the restoration component makes use of global context to refine results, it is more difficult to change the prediction by adversarial noises of some limited perturbation strength. In order to influence the prediction results of pixels at certain specific locations, intricate changes involving a larger range of feature vectors may be required, which in turn results in larger pixel value perturbations.

C. Training Scheme

The following explains how we train the entire framework in an end-to-end scheme. In the beginning, the FCN-based backbone of ROSA framework is initialized as some pretrained visual saliency model while the parameters of context-aware restoration component are set up according to [65]. Then, the parameters of the backbone network and the restoration component are fine-tuned together. As the segment-wise shielding component contains no learnable parameters, gradients are not passed backward through that component. To maintain generalization ability against different kinds of adversarial perturbation, our training set does not include adversarial samples but only natural images. These training samples are fed into the segment-wise shielding component. As shown in Fig. 2, a pixel-wise cross-entropy loss function is computed between the ground truth saliency map and the output of the context-aware restoration component. SGD algorithm is used to train the proposed method. The learning rate of the context-aware restoration component is set as 10^{-10} while that of other parts is selected as 10^{-13}. The momentum and weight decay are set as 0.9 and 0.0005, respectively. For each backbone FCN in this paper, fine-tuning with our proposed framework takes no more than 5 epochs. We adopt early-stopping strategy and terminate the training if the performance on validation set is not improved after 2 consecutive epochs. If the proposed method adopts DSS as its backbone, a forward pass on an image costs about 0.8 s.

IV. Experiment

In this section, we conduct three groups of experiments. First, we launch adversarial attacks on existing visual saliency models and investigate how they are affected. Then, we integrate our proposed framework with current models to present how the proposed framework enhances the robustness for those models. Lastly, we verify the effectiveness of each component in the ROSA framework.

A. Dataset

In this paper, we conduct experiments on MSRA-B dataset [66], HKU-IS dataset [18], DUT-OMRON dataset [9], and ECSSD dataset [67]. The MSRA-B dataset contains a train set of 2500 images, a validation set of 500 images, and a test set of 2000 images. The HKU-IS dataset includes 2500 images, 500 images, and 1447 images in train set, validation set, and test set, respectively. We follow the released data

split in the MSRA-B and HKU-IS dataset. For DUT-OMRON dataset, we randomly separate all the 5168 images into a train set of 2500 images, a validation set of 500 images, and a test set of 2168 images. For ECSSD dataset, all 1000 images are taken as testing samples.

B. Evaluation

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

$$\text{MAE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S_{i,j} - G_{i,j}| \tag{7}$$

where W and H denote the width and height of the saliency map, respectively. To compute F_β-measure, we binarize each saliency map with an image-dependent threshold proposed by Achanta *et al.* [63]. The threshold T is calculated as

$$T = \frac{2}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} S_{i,j} \tag{8}$$

where W and H denote the width and height of the saliency map S. Pixels with saliency value larger than T form the predicted salient region. 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 [63]

$$F_\beta = \frac{\left(1 + \beta^2\right) \times \text{Precision} \times \text{Recall}}{\beta \times \text{Precision} + \text{Recall}} \tag{9}$$

where β^2 is set as 0.3 to emphasize the precision. To draw PR curves, a list of equally spaced thresholds are sampled. For each threshold value, each predicted saliency map in the benchmark is quantized into a binary mask. Precision and recall are calculated with each binary mask and its ground truth annotation. The precision and the recall corresponding to each threshold are computed by, respectively, taking average of precision and recall for all binary masks. Then, we obtain a list of (precision, recall) pairs and plot it as a PR curve.

C. Effectiveness of Adversarial Attack

We demonstrate the performance of eight state-of-the-art visual saliency models: 1) DSS [6]; 2) DCL [34]; 3) RFCN [35]; 4) Amulet [28]; 5) UCF [29]; 6) MC [22]; 7) LEGS [21]; and 8) MDF [18] on natural images and adversarial samples which are synthesized with a pretrained DSS model. For efficiency, the above neural models are trained on the train set of MSRA-B and tested on the test set of HKU-IS. The upper bound of L_∞-norm ϵ is chosen as 20. Qualitative results can be found in Fig. 5, where each sample consists of two rows. The upper are natural image and its predicted saliency maps while the lower are adversarial samples and their corresponding results. The second column from the left are the ground truth saliency maps denoted as GT, DSS, DCL, RFCN, Amulet, and UCF. Predicted saliency maps on adversarial samples change significantly, compared with that on natural images. For MC, LEGS, and MDF, predictions on adversarial samples and that on original images are visually approximate.

TABLE I
F_β-MEASURE AND MAE ON NATURAL AND ADVERSARIAL EXAMPLES

Model	F_β measure		MAE	
	Original	Adversarial	Original	Adversarial
DSS	87.13%	51.08%	0.0491	0.2251
DCL	86.82%	56.84%	0.0583	0.1921
RFCN	87.68%	75.39%	0.0544	0.0976
Amulet	84.94%	75.63%	0.0511	0.0906
UCF	81.64%	75.03%	0.0734	0.1107
LEGS	76.39%	75.43%	0.1192	0.1231
MC	75.31%	74.65%	0.0982	0.0999
MDF	82.05%	80.99%	0.0946	0.0999

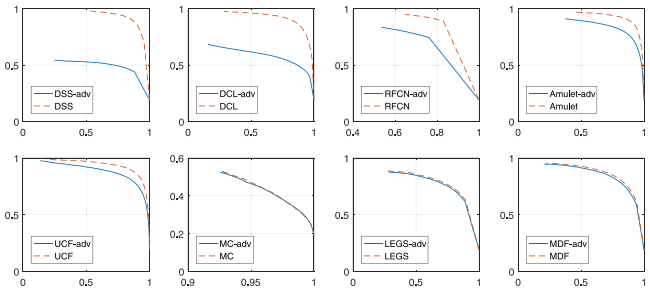


Fig. 3. Effectiveness of adversarial attack on PR curves. As shown in the above figures, the PR curve of DSS on adversarial samples drops the most seriously. The PR curves of DCL, RFCN, Amulet, and UCF also degrade to some extent, which suggests that adversarial samples yielded by some FCN network are transferable to attack other FCN variants. For MC, LEGS, and MDF, their PR curves tested on natural images and adversarial samples are relatively close to each other, which indicates that the sparse labeling-based methods are insensitive to adversarial noises.

As shown in Table I, F_β-measure of DSS and DCL drop 30%–36% when exposed to the adversarial samples. The adversarial attack reduces F_β-measure of RFCN, Amulet, and UCF by 6%–12% while it only lowers F_β of MC, LEGS, and MDF by 0.7%–1.1%. As shown in Table I, MAE of DSS and DCL are increased by 0.176 and 0.1338, respectively, on the adversarial samples while that of RFCN, Amulet, and UCF are raised by around 0.04. MAE of MC, LEGS, and MDF change less than 0.01. These results indicate that DSS suffers most from the adversarial attack for the adversarial samples are synthesized using a DSS model. DCL, RFCN, Amulet, and UCF are affected to different extent, which may depend on the similarity between their architectures and the pretrained model used to launch attacks.

Fig. 3 demonstrates the comparison of PR curves with respect to these above-mentioned salient object detection models tested on natural images and adversarial samples. N-adv denotes some neural network N tested on adversarial samples. The results tested on adversarial samples are plotted using blue solid curves while those tested on original images are draw with orange dashed curves. As shown in Fig. 3, the PR curves of DSS and DCL are significantly higher than DSS-adv and DCL-adv. It indicates that DSS and DCL suffer the most from adversarial samples. Because the adversarial samples are

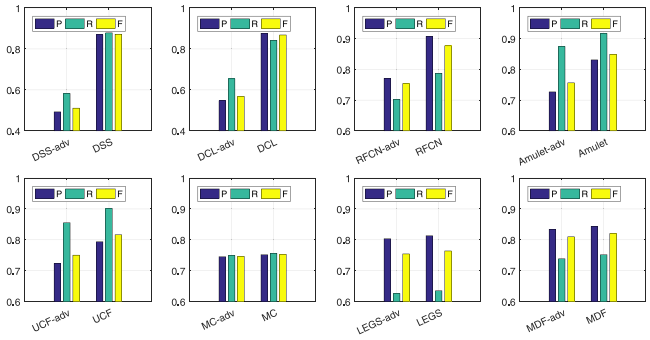


Fig. 4. Effectiveness of adversarial attack on precision, recall, and F_β-measure. N-adv denotes the result of neural network N tested on adversarial samples. As shown in the above bar diagrams, the precision, recall, and F_β of DSS-adv and DCL-adv decline by a wide margin, in comparison to DSS and DCL, respectively. For RFCN, Amulet, and UCF, their precision, recall, and F_β tested on adversarial samples also decrease to some degree. For sparse labeling methods (MC, LEGS, and MDF), their performances against adversarial attacks is almost unchanged.

synthesized using exactly the same DSS model. DCL also severely affected by the adversarial attacks, possibly because it has similar network structure with DSS. The PR curves of RFCN-adv, Amulet-adv, and UCF-adv also decline by a considerable margin, respectively, in comparison to RFCN, Amulet, and UCF. That is to say, the adversarial samples generated by some FCN-based model are transferable to degrade other FCN-based methods to different extent. Even attackers are unaware of the target neural network, they still have chances to launch successful attacks using some arbitrary FCN model. It reveals that existing visual saliency models based on dense labeling are threatened by adversarial attacks. For MC, LEGS, and MDF, their PR curves on natural images and adversarial samples almost completely overlap. It suggests that sparse labeling methods are quite robust against existing gradient-based attacks, since gradients propagated from different segments to the same image position very possibly conflict with each other.

Fig. 4 are the bar diagrams of existing visual saliency methods tested on natural images and adversarial samples. P, R, and F denote precision, recall, and F_β-measure, respectively, in the color of blue, green, and yellow. The results shown in Fig. 4 draws similar conclusions with Fig. 3. Precision, recall, and F_β of DSS and DCL decrease the most seriously against adversarial attacks. For RFCN, Amulet, and UCF, their precision, recall, and F_β are also harmed by adversarial samples to some degree. Segment-based models such MC, LEGS, and MDF are more robust and present negligible degeneration.

D. Robustness of the Robust Salient Object Detection Framework

To demonstrate the robustness of ROSA, we present extensive experiments on four datasets (HKU-IS, ECSSD, DUT-OMRON, and MSRA-B), with three state-of-the-art saliency models (DSS, DCL, and RFCN) as baselines. All models in the section are trained on a dataset that includes training sets of HKU-IS, DUT-OMRON, and MSRAB. Adversarial samples are synthesized with a DSS model pretrained on the above-mentioned dataset. The L_∞-norm upper bound ϵ of the adversarial noise is chosen as 25. We also compare our proposed method with several existing defending algorithms, which are developed for robust image classification and can be transferred to other tasks. Smooth denotes a spatial smooth filter in [68]. JPEG [54], [69] denotes applying JPEG compression on input images before feeding them into target networks. The quality of the compressed image is set to 75 according to [54]. Quant [70] denotes bit reduction that quantizes 8-bit RGB images into pixel values with less bits. We reduce images to 3 bits following [54]. TVM [54] denotes total variation minimization that aims at reducing difference between adjacent pixels. TVM is implemented using [71]. Tables II–IV are the numeric results with a baseline model as DSS, DCL, and RFCN, respectively. D-adv denotes experiments on the adversarial samples of dataset D. $N+M$ denotes the neural network N equipped with the defense method M.

As Table II shows, DSS+ours outperforms DSS* by 71.9%–86.2% with respect to F_β-measure on adversarial samples. DSS* displays seriously degraded F_β lower than 1.0% because the adversarial samples are synthesized with the same DSS model. For fair comparison, we also attack DSS+ours with samples produced by its own DSS backbone, which is denoted as DSS+ours*. DSS+ours* still significantly surpasses DSS* by 71.86%–86.05% F_β. The difference of the performance between DSS+ours* and DSS+ours is quite small and less than 0.78% F_β. Note that on natural images, DSS+ours* and DSS+ours have exactly the same numerical results, since their difference lies in using DSS models of different weights to synthesize adversarial samples. For simplicity, the cell corresponding to DSS+ours tested on original images leave a blank. Compared with existing defense baselines, DSS+ours exceeds the second best DSS+TVM by 6.08% F_β and 0.0337 MAE on the adversarial samples of HKU-IS dataset. On ECSSD-adv dataset, our proposed method outperforms TVM by 2.92% F_β and 0.0327 MAE. DSS+ours surpasses the second best DSS+TVM by 6.05% F_β and 0.0206 MAE on DUT-OMRON-adv dataset. On the adversarial samples of MSRA-B, the proposed framework also obtains higher F_β and smaller MAE than other defense methods. As for natural images, the performances of different models are close to each other, because of no threats caused by adversarial noises. Existing defense approaches act as small variations on input images and result in slight degeneration on performance. In most cases, the proposed framework achieves the best F_β-measure and MAE on clean input images, which suggests that our method improve the backbone model on both adversarial samples and natural images.

In the case of DCL shown in Table III, our proposed methods presents the highest F_β and the smallest MAE on both original images and adversarial samples of all four benchmarks. For example, DCL+ours outperforms DCL+TVM by 3.35% F_β and DCL+Quant by 0.0112 MAE on HKU-IS-adv. F_β and MAE of DCL+ours are superior to those of DCL by 1.65% and 0.0033 on HKU-IS dataset. On MSRA-B-adv dataset, our proposed defense framework surpasses the second best TVM by 2.24% and Smooth by 0.013. On the

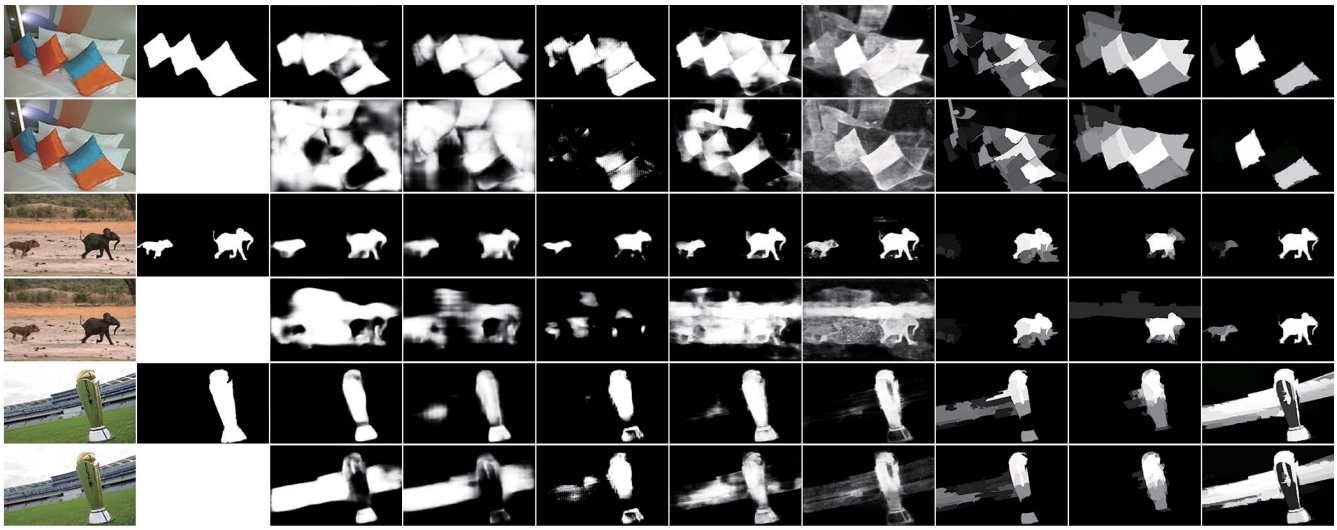


Fig. 5. Effectiveness of adversarial attack. The leftmost column is the input images in which the upper one is a natural image and the lower one is the corresponding adversarial samples. The second column from the left is the ground truth saliency maps in which the lower position leave vacant because natural images and their adversarial samples share the same ground truth. As shown in the above examples, saliency maps predicted by FCN-based salient object detection models, including DSS, DCL, RFCN, Amulet, and UCF are deteriorated by input images with adversarial perturbations. Segment-based models, such as MC, LEGS, and MDF, produce more consistent results between natural images and adversarial samples.

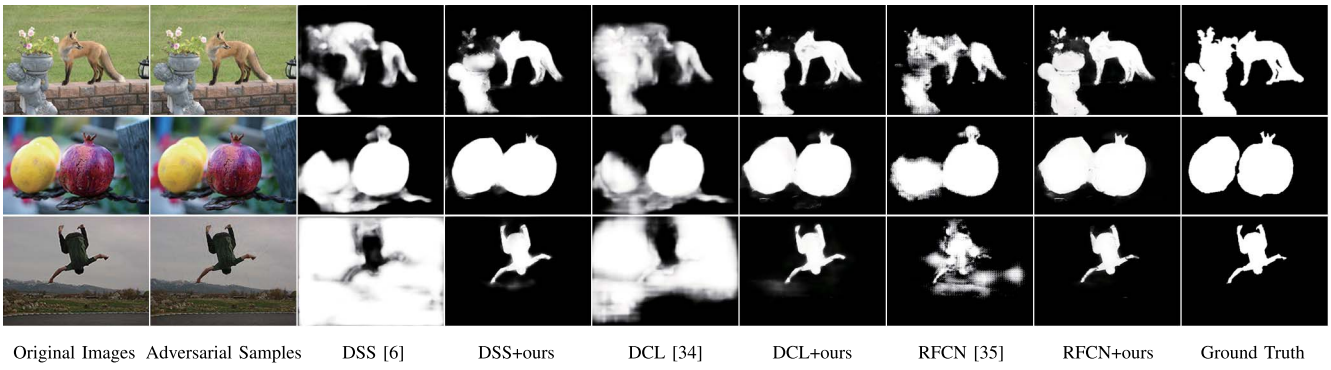


Fig. 6. Robustness of ROSA. The leftmost column is the original natural images. The second column from the left is the corresponding adversarial samples. The rightmost column is the ground truth. N+ours denotes some neural model N incorporated with our proposed method ROSA. All salient object detection models in the above are tested on adversarial samples. As shown in the above examples, our proposed framework enhances the prediction accuracy of three backbone network DSS, DCL, and RFCN.

natural images of MSRA-B, DCL+ours also obtains better F_β and MAE than the second best DCL+Smooth and DCL by 1.33% and 0.0042, respectively. In the case of RFCN shown in Table IV, the proposed defense framework achieves the best F_β and MAE on the adversarial samples of all four benchmarks. For example, RFCN+ours outperforms the second best RFCN+TVM by 4.09% F_β and RFCN+Quant by 0.0318 MAE on ECSSD-adv dataset. RFCN+ours surpasses the second best RFCN+TVM by 4.54% and RFCN+Quant by 0.0341 MAE on MSRA-B-adv dataset. As for natural images, our proposed method RFCN+ours achieves competitive or better results than RFCN.

Fig. 7 demonstrates the PR curves of existing defense baselines and the proposed method with three backbone networks, DSS, DCL, and RFCN, respectively. Solid curves denote the proposed defense framework while dashed ones denote existing defense algorithms. Note that in the leftmost subfigure of Fig. 7, DSS+ROSA* denotes attacking DSS+ROSA via adversarial samples that are generated using the backbone of DSS+ROSA itself. As shown in Fig. 7, the proposed defense framework displays better PR curves than the second best TVM with DSS as backbone. On the case of DCL and RFCN, the PR curves of our proposed method are also higher than the second best Quant by a considerable margin. The existing defense algorithms perform significantly worse with DSS than DCL and RFCN, since the adversarial samples are synthesized using a DSS model. In short, our proposed framework not only significantly enhances the robustness of backbone against adversarial attacks but also demonstrates comparable or better performance on natural images. Fig. 6 presents some qualitative comparisons on the robustness of ROSA.

E. Ablation Study

This section verifies the effectiveness of each part in the proposed ROSA framework. We integrate each component of

TABLE II
ROBUSTNESS OF ROSA WITH DSS

Dataset	Metric	DSS*	DSS+ours*	DSS+ours	DSS+Smooth	DSS+JPEG	DSS+Quant	DSS+TVM
HKU-IS-adv	F_β	0.74%	83.18%	83.52%	54.93%	12.61%	31.80%	77.44%
	MAE	0.7495	0.0654	0.0644	0.1831	0.4638	0.3224	0.0981
HKU-IS	F_β	87.50%		88.48%	85.92%	87.28%	87.07%	84.39%
	MAE	0.0436		0.0341	0.0528	0.0446	0.0479	0.0670
ECSSD-adv	F_β	0.90%	81.48%	82.26%	55.09%	12.91%	34.84%	79.34%
	MAE	0.7763	0.0940	0.0915	0.2202	0.4960	0.3428	0.1242
ECSSD	F_β	87.69%		87.33%	86.56%	87.59%	87.23%	85.62%
	MAE	0.0608		0.0475	0.0712	0.0618	0.0689	0.0891
DUT-OMRON-adv	F_β	0.41%	72.27%	72.31%	35.41%	22.62%	26.83%	66.26%
	MAE	0.8038	0.0687	0.0690	0.2265	0.3409	0.2885	0.0896
DUT-OMRON	F_β	76.20%		79.63%	74.96%	76.19%	75.32%	73.70%
	MAE	0.0547		0.0469	0.0591	0.0547	0.0581	0.0675
MSRA-B-adv	F_β	0.51%	86.56%	86.71%	49.79%	35.48%	46.35%	84.11%
	MAE	0.8021	0.0549	0.0548	0.2460	0.3376	0.2608	0.0812
MSRA-B	F_β	89.59%		89.76%	89.11%	89.58%	89.06%	88.63%
	MAE	0.0440		0.0366	0.0484	0.0440	0.0489	0.0567

TABLE III
ROBUSTNESS OF ROSA WITH DCL

Dataset	Metric	DCL	DCL+ours	DCL+Smooth	DCL+JPEG	DCL+Quant	DCL+TVM
HKU-IS-adv	F_β	77.84%	84.38%	79.89%	79.36%	79.53%	81.03%
	MAE	0.0866	0.0695	0.0843	0.0810	0.0807	0.0911
HKU-IS	F_β	85.22%	86.87%	84.67%	85.08%	84.72%	83.59%
	MAE	0.0540	0.0507	0.0671	0.0550	0.0579	0.0797
ECSSD-adv	F_β	79.22%	83.76%	80.50%	80.85%	80.58%	82.28%
	MAE	0.1070	0.0960	0.1081	0.1008	0.1047	0.1229
ECSSD	F_β	85.82%	86.25%	85.02%	85.63%	85.41%	84.82%
	MAE	0.0698	0.0677	0.0867	0.0712	0.0783	0.1093
DUT-OMRON-adv	F_β	60.46%	69.68%	60.34%	64.46%	62.54%	67.95%
	MAE	0.1091	0.0794	0.1092	0.0934	0.0995	0.0912
DUT-OMRON	F_β	70.30%	72.23%	67.92%	70.30%	69.10%	70.66%
	MAE	0.0723	0.0683	0.0891	0.0723	0.0769	0.0834
MSRA-B-adv	F_β	79.87%	87.82%	83.62%	83.53%	82.68%	85.58%
	MAE	0.0960	0.0645	0.0775	0.0783	0.0819	0.0837
MSRA-B	F_β	87.80%	89.29%	87.96%	87.76%	87.02%	87.19%
	MAE	0.0563	0.0521	0.0595	0.0563	0.0609	0.0743

TABLE IV
ROBUSTNESS OF ROSA WITH RFCN

Dataset	Metric	RFCN	RFCN+ours	RFCN+Smooth	RFCN+JPEG	RFCN+Quant	RFCN+TVM
HKU-IS-adv	F_β	76.58%	85.76%	76.38%	78.36%	79.83%	78.72%
	MAE	0.0985	0.0599	0.1036	0.0916	0.0862	0.1068
HKU-IS	F_β	86.94%	87.18%	85.12%	86.24%	85.65%	82.69%
	MAE	0.0533	0.0535	0.0668	0.0563	0.0612	0.0893
ECSSD-adv	F_β	77.24%	84.89%	78.27%	79.69%	80.61%	80.80%
	MAE	0.1290	0.0836	0.1319	0.1190	0.1154	0.1382
ECSSD	F_β	87.22%	85.76%	85.73%	86.77%	86.04%	84.05%
	MAE	0.0735	0.0802	0.0911	0.0770	0.0869	0.1207
DUT-OMRON-adv	F_β	56.09%	72.57%	60.60%	63.79%	62.30%	66.21%
	MAE	0.1257	0.0691	0.1132	0.1002	0.1048	0.1029
DUT-OMRON	F_β	72.18%	74.37%	71.28%	72.15%	70.02%	70.14%
	MAE	0.0728	0.0638	0.0786	0.0728	0.0802	0.0918
MSRA-B-adv	F_β	76.32%	89.68%	79.96%	83.46%	83.36%	85.14%
	MAE	0.1210	0.0507	0.1053	0.0852	0.0848	0.0912
MSRA-B	F_β	89.45%	90.53%	88.43%	89.45%	88.25%	87.47%
	MAE	0.0518	0.0469	0.0605	0.0519	0.0604	0.0769

ROSA with DSS, DCL, and RFCN, respectively. For simplicity, the above models are trained on MSRA-B train set and tested on HKU-IS test set. SWS denotes segment-wise shielding component and CAR denotes context-aware restoration component. To validate the effect of CAR/SWS, we compare *+SWS/*+CAR with *+ROSA. According to Table V, even though *+CAR exceed *+ROSA by 5.22%, 3.5%, and 2.75% F_β on natural images, *+ROSA outperforms *+CAR by

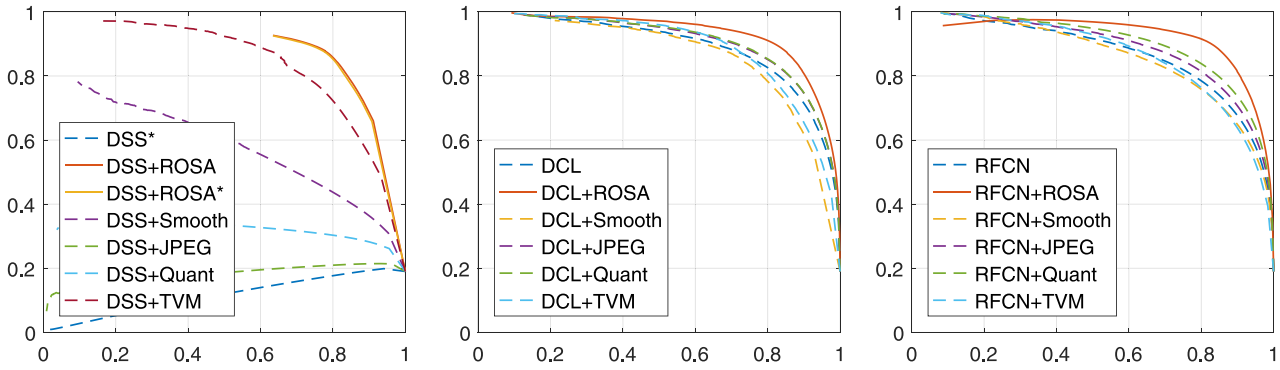


Fig. 7. Quantitative analysis on the robustness of ROSA in terms of PR curve. It shows the PR curves of existing defense methods and our proposed framework with DSS, DCL, and RFCN, respectively. Solid curves denote our proposed method while dashed curves denote other defense algorithms. DSS+ROSA* denotes DSS+ROSA tested on the adversarial samples synthesized by the DSS backbone of DSS+ROSA itself. As shown in the above figures, our proposed method achieves higher PR curves than other defense baselines with DSS, DCL, and RFCN.

TABLE V
ABLATION STUDY OF ROSA

Model	F_β measure		MAE	
	Original	Adversarial	Original	Adversarial
DSS+SWS	82.89%	79.09%	0.0660	0.0802
DSS+CAR	90.96%	53.86%	0.0428	0.2169
DSS+ours	85.78%	82.18%	0.0541	0.0683
DCL+SWS	78.19%	74.20%	0.0852	0.1015
DCL+CAR	88.36%	60.17%	0.0500	0.1754
DCL+ours	84.86%	81.46%	0.0579	0.0708
RFCN+SWS	77.07%	74.45%	0.0916	0.1004
RFCN+CAR	88.80%	77.98%	0.0549	0.0904
RFCN+ours	86.05%	83.83%	0.0603	0.0688

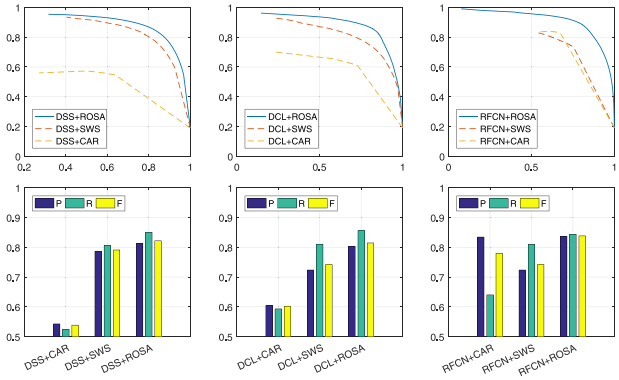


Fig. 8. Ablation study of ROSA. The above compares the entire proposed framework with its separated components, segment-wise shielding component denoted as SWS, and context-aware restoration component denoted as CAR. The upper row are PR curves with DSS, DCL, and RFCN while the lower one are bar diagrams of precision, recall, and F_β. As shown in the above figures, the performance of the entire proposed framework is superior to those of its internal components. It suggests that the internal components in our proposed method acts as different roles to complement and enhance each other significantly.

28.32%, 21.29%, and 5.85% against adversarial attacks, which indicates the effectiveness of SWS. MAE results in Table V also draw a similar conclusion. Even though MAE of *+CAR are 0.0113, 0.0079, and 0.0054 less than that of *+ROSA on natural images, *+ROSA lowers MAE by 0.1486, 0.1046, and 0.0216 on adversarial samples by a larger margin. As for components *+SWS, *+ROSA surpasses *+SWS by 2.89%, 6.67%, and 8.98% F_β on original samples and 3.09%, 7.26%, and 9.38% F_β on adversarial samples. Besides, *+ROSA reduces MAE by 0.0119, 0.0273, and 0.0313 on natural images and 0.0119, 0.0307, and 0.0316 on adversarial samples in comparison to *+SWS, which authenticates the effectiveness of CAR. We claim that SWS and CAR are two strongly coupled components. For example, DCL obtains 56.84% F_β against adversarial samples as shown in Table I while DCL+CAR achieves 60.17% F_β. CAR improves DCL by 3.33% F_β. However, DCL+ours (SWS+CAR) outperforms DCL+SWS by 7.26% F_β more than 3.33%. That is to say, with SWS component, the effectiveness of CAR is more significant. The cases of DSS and RFCN draw the same conclusion. Thus, SWS and CAR are not a separate process but two complementary steps of one core idea, adaptively predicting saliency for inputs with the new introduced noise to resist adversarial attacks. In short, CAR component can refine saliency maps predicted by models with SWS component. These two components are complementary and contribute to the robustness of our proposed method.

Fig. 8 compares the entire proposed framework with its internal components with DSS, DCL, and RFCN, on PR curves and bar diagrams of precision–recall–F_β. The upper row are PR curves while the lower one are bar diagrams. Among these PR curves, the blue solid ones denote the entire proposed method while the orange/yellow dashed ones denote the internal components SWS and CAR. In the bar diagrams, P, R, and F denote precision, recall, and F_β in the color of blue, green, and yellow, respectively. In Fig. 8, the entire proposed framework displays higher PR curves than its internal components with three different backbone networks. Besides, the precision, recall, and F_β of the entire proposed method also surpass those of *+SWS and *+CAR. In detail, *+CAR perform the worst with DSS and DCL, while *+SWS is close to *+CAR with RFCN. It indicates that CAR component almost cannot resist adversarial attacks without SWS component. Note that *+ROSA achieves the best and outperform *+SWS with different backbones. It suggests that CAR component can further improve *+SWS by a significant margin.

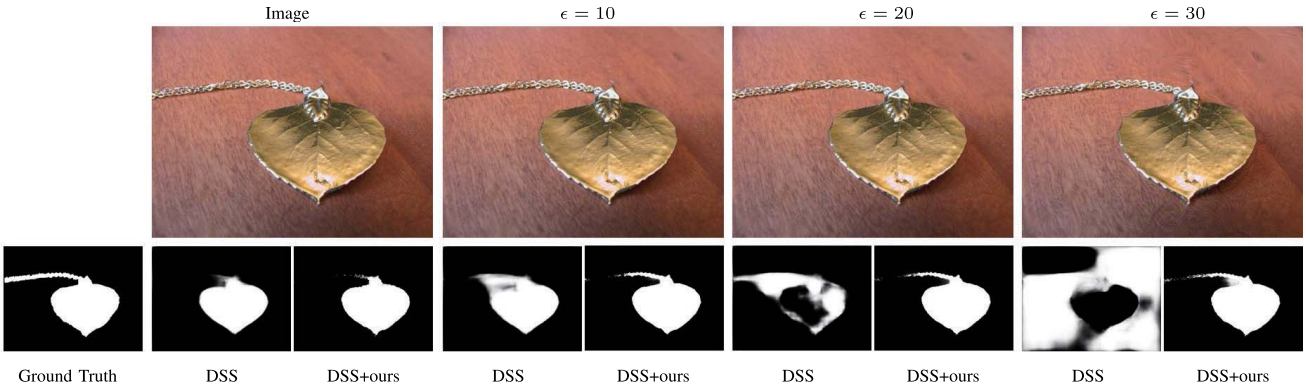
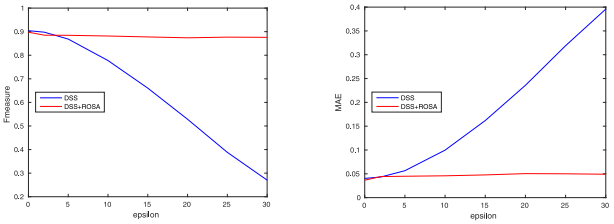


Fig. 9. Qualitative comparison among adversarial attacks of different strengths. ϵ denotes the L_∞-norm upper bound of adversarial perturbations. As shown in the above, larger ϵ achieves stronger attack and results in worse performance of a pretrained DSS model. But the curve-like patterns of adversarial noises are more easily observed. No matter what ϵ is set in the range of [0, 30], the proposed method denoted as DSS+ours presents stable and fine saliency maps.



(a) (b)

Fig. 10. Investigation of attack strength. Epsilon denotes the L_∞-norm upper bound of adversarial perturbations. F-measure denotes the evaluation metric, F_β-measure. (a) F-measure–Epsilon curve. (b) MAE–Epsilon curve.

F. Investigation of Attack Strength

In this section, we investigate how a hyper-parameter, the upper bound of L_∞-norm (denoted as ϵ), affects the strength of the proposed adversarial attack. We also study how our proposed defense method performs against adversarial attacks of different strengths. For efficiency, 500 images are randomly selected from the test set of MSRA-B dataset. These 500 images are named MSRA-B500. A list of ϵ is sampled in the range of [0, 30]. For each ϵ, a set of adversarial samples is synthesized for the entire MSRA-B500. These adversarial samples are computed using some pretrained DSS model as target network. Each set of adversarial samples is tested by another trained DSS model, and our proposed method with DSS, respectively. F_β-measure and MAE are calculated for each set of adversarial samples. We plot these results as F-measure–Epsilon curves and MAE–Epsilon curves in Fig. 10. As Fig. 10 shows, the blue curve denotes the performance of DSS while the red one represents the proposed method (denoted as DSS+ROSA). As ϵ increases, the F_β of DSS drops dramatically. It suggests that the strength of an adversarial attack grows with the increase of its L_∞-norm upper bound. Notice that as ϵ rises, the performance of our proposed method only degrades slightly and then becomes stable. It indicates that the proposed defense framework is robust to adversarial samples of different strengths. Fig. 9 demonstrates a qualitative comparison among adversarial attacks of different strengths. Setting ϵ as 30 achieves the strongest attack and DSS incorrectly predicts the reverse of the ground truth as salient regions. However, the adversarial noise is perceptible and curve-like patterns can be observed in the top right of the adversarial sample. For $\epsilon = 20$, the adversarial perturbations are hard to spot and the DSS model is still seriously affected. Thus, we suggest that choosing ϵ around 20 helps launch a strong and quasi-imperceptible adversarial attack on salient object detection models.

V. Conclusion

In this paper, we, for the first time, achieve successful attacks on state-of-the-art visual saliency methods. We experimentally confirm that existing FCN-based models are sensitive to adversarial perturbation. In addition, this paper proposed a novel salient object detection framework that first brings some new generic noise to input images, and then adaptively detects salient objects for the inputs with the new noise. The proposed framework is instantiated by an arbitrary FCN-based backbone network, a segment-wise shielding component and a context-aware restoration component. Experimental results suggest that these two components are strongly coupled and significantly complement each other. Besides, extensive comparisons show that the entire framework can effectively strengthen the robustness of FCN-based saliency models, superior to the existing defense baselines. We believe that developing an accurate, fast, and robust model will be a new trend in salient object detection.

References

[1] Y. Yu, G. K. I. Mann, and R. G. Gosine, "An object-based visual attention model for robotic applications," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 5, pp. 1398–1412, Oct. 2010.

[2] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[3] Y. Wei *et al.*, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2017.

[4] C. Wang, Y. Guo, J. Zhu, L. Wang, and W. Wang, "Video object co-segmentation via subspace clustering and quadratic pseudo-Boolean optimization in an MRF framework," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 903–916, Jun. 2014. [Online]. Available: https://doi.org/10.1109/TMM.2014.2306393

[5] S. Bi, G. Li, and Y. Yu, "Person re-identification using multiple experts with random subspaces," *J. Image Graph.*, vol. 2, no. 2, pp. 151–157, 2014.

[6] Q. Hou *et al.*, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.

[7] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 248–255.

[8] M.-M. Cheng *et al.*, "Efficient salient region detection with soft image abstraction," in *Proc. Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, 2013, pp. 1529–1536.

[9] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 3166–3173.

[10] H. Jiang *et al.*, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 2083–2090.

[11] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[12] X. Huang and Y.-J. Zhang, "300-FPS salient object detection via minimum directional contrast," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4243–4254, Sep. 2017.

[13] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. Int. Conf. Comput. Vis.*, Las Vegas, NV, USA, 2016, pp. 2334–2342.

[14] J. Zhang *et al.*, "Minimum barrier salient object detection at 80 FPS," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1404–1412.

[15] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 2814–2821.

[16] L. Zhang, C. Yang, H. Lu, X. Ruan, and M.-H. Yang, "Ranking saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1892–1904, Sep. 2017.

[17] L. Zhang, J. Ai, B. Jiang, H. Lu, and X. Li, "Saliency detection via absorbing Markov chain with learnt transition probability," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 987–998, Feb. 2018.

[18] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 5455–5463.

[19] J. Kim and V. Pavlovic, "A shape-based approach for salient object detection using deep learning," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 455–470.

[20] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[21] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. Int. Conf. Comput. Vis.*, Boston, MA, USA, 2015, pp. 3183–3192.

[22] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1265–1274.

[23] Y. Qin, M. Feng, H. Lu, and G. W. Cottrell, "Hierarchical cellular automata for visual saliency," *Int. J. Comput. Vis.*, vol. 126, no. 7, pp. 751–770, 2018.

[24] G. Li and Y. Yu, "Contrast-oriented deep neural networks for salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6038–6051, Dec. 2018.

[25] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 2386–2395.

[26] G. Li, Y. Xie, and L. Lin, "Weakly supervised salient object detection using image labels," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 7024–7031.

[27] Z. Luo *et al.*, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, no. 6, 2017, p. 7.

[28] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

[29] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 212–221.

[30] T. Wang *et al.*, "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3127–3135.

[31] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 234–250.

[32] S. Chen, B. Wang, X. Tan, and X. Hu, "Embedding attention and residual network for accurate salient object detection," *IEEE Trans. Cybern.*, to be published.

[33] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[34] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 478–487.

[35] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 825–841.

[36] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1395–1403.

[37] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[38] Y. Dong *et al.*, "Boosting adversarial attacks with momentum," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

[39] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.

[40] M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured visual and speech recognition models with adversarial examples," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds. Red Hook, NY, USA: Curran Assoc., 2017, pp. 6977–6987. [Online]. Available: http://papers.nips.cc/paper/7273-houdini-fooling-deep-structured-visual-and-speech-recognition-models-with-adversarial-examples.pdf

[41] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[42] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[43] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 86–94.

[44] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–23.

[45] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 427–436.

[46] C. Xiao *et al.*, "Spatially transformed adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–30.

[47] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4422–4431.

[48] Z. Zhao, D. Dua, and S. Singh, "Generating natural adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[49] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–9.

[50] H. Dai *et al.*, "Adversarial attack on graph structured data," in *Proc. 35th Int. Conf. Mach. Learn.*, Jul. 2018, pp. 1115–1124.

[51] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Security Privacy (SP)*, San Jose, CA, USA, 2016, pp. 582–597.

[52] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[53] J. Lu, T. Issaranon, and D. A. Forsyth, "SafetyNet: Detecting and rejecting adversarial examples robustly," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 446–454.

[54] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in *Proc. Int. Conf. Learn. Represent.*, 2018. [Online]. Available: https://openreview.net/forum?id=SyJ7ClWCb

[55] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. L. Yuille, "Mitigating adversarial effects through randomization," in *Proc. Int. Conf. Learn. Represent.*, 2018.

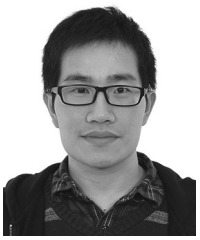
[56] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "PixelDefend: Leveraging generative models to understand and defend against adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[57] F. Liao *et al.*, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1778–1787.
[58] N. Akhtar, J. Liu, and A. Mian, "Defense against universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.
[59] C. Wang, J. Zhu, Y. Guo, and W. Wang, "Video vectorization via tetrahedral remeshing," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1833–1844, Apr. 2017. [Online]. Available: https://doi.org/10.1109/TIP.2017.2666742
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.
[62] C. Xie *et al.*, "Adversarial examples for semantic segmentation and object detection," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1378–1387.
[63] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 1597–1604.
[64] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFS with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.
[65] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1529–1537.
[66] T. Liu *et al.*, "Learning to detect a salient object," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[67] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended CSSD," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 717–729, Apr. 2016.
[68] X. Li and F. Li, "Adversarial examples detection in deep networks with convolutional filter statistics," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5775–5783.
[69] N. Das *et al.*, "SHIELD: Fast, practical defense and vaccination for deep learning using JPEG compression," in *Proc. Knowl. Disc. Data Min.*, 2018, pp. 196–204.
[70] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proc. Netw. Distrib. Syst. Security Symp.*, 2018.
[71] P. Getreuer, "Rudin–Osher–Fatemi total variation denoising using split Bregman," *Image Process. Line*, vol. 2, pp. 74–95, May 2012.

Haofeng Li received the B.S. degree from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, in 2015. He is currently pursuing the Ph.D. degree with the Department of Computer Science, University of Hong Kong, Hong Kong.

His current research interests include computer vision, image processing, and deep learning.

Mr. Li was a recipient of the Hong Kong Ph.D. Fellowship.



Guanbin Li (M'15) received the Ph.D. degree from the University of Hong Kong, Hong Kong, in 2016.

He is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He has authorized and co-authorized over 20 papers in top-tier academic journals and conferences. His current research interests include computer vision, image processing, and deep learning.

Dr. Li was a recipient of the Hong Kong Ph.D. Fellowship. He serves as an Area Chair for the conference of VISAPP. He has been serving as a Reviewer for numerous academic journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON COMPUTERS, CVPR, AAAI, and IJCAI.

Yizhou Yu (M'10–SM'12–F'19) received the Ph.D. degree from the University of California at Berkeley, Berkeley, CA, USA, in 2000.

He is a Professor with the University of Hong Kong, Hong Kong, and the Chief Scientist with Deepwise Healthcare, Beijing, China. He was a Faculty Member with the University of Illinois at Urbana–Champaign, Champaign, IL, USA, for 12 years. His current research interests include computer vision, deep learning, biomedical data analysis, computational visual media, and geometric computing.

Prof. Yu was a recipient of the 2002 U.S. National Science Foundation CAREER Award, the 2007 NNSF China Overseas Distinguished Young Investigator Award, and the ACCV 2018 Best Application Paper Award. He has served on the editorial board of *IET Computer Vision*, *Visual Computer*, and the IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS. He has also served on the Program Committee of many leading international conferences, including SIGGRAPH, SIGGRAPH Asia, and International Conference on Computer Vision.