Robust Adversarial Perturbation on Deep Proposal-based Models

Yuezun Li[1]
yli52@albany.edu
Daniel Tian[2]
dtian19@berkshireschool.org
Mingching-Chang[1]
mchang2@albany.edu
Xiao Bian[3]
xiao.bian@ge.com
Siwei Lyu[1]
slyu@albany.edu

[1] University at Albany, State University of New York, USA
[2] Berkshire School, Massachusetts, USA
[3] GE Global Research Center, Niskayuna, New York, USA

Abstract

Adversarial noises are useful tools to probe the weakness of deep learning based computer vision algorithms. In this paper, we describe a robust adversarial perturbation (R-AP) method to attack deep proposal-based object detectors and instance segmentation algorithms. Our method focuses on attacking the common component in these algorithms, namely Region Proposal Network (RPN), to universally degrade their performance in a black-box fashion. To do so, we design a loss function that combines a label loss and a novel shape loss, and optimize it with respect to image using a gradient based iterative algorithm. Evaluations are performed on the MS COCO 2014 dataset for the adversarial attacking of 6 state-of-the-art object detectors and 2 instance segmentation algorithms. Experimental results demonstrate the efficacy of the proposed method.

1 Introduction

Deep learning based algorithms achieve superior performance in many problems in computer vision, including image classification, object detection and segmentation. However, it has been recently shown that algorithms based on Convolutional Neural Network (CNN) are vulnerable to *adversarial perturbations*, which are intentionally crafted noises that are imperceptible to human observer, but can lead to large errors in the deep network models when added to images. To date, most existing adversarial perturbations are designed to attack CNN image classifiers, *e.g.*, [4, 6, 10, 14, 15, 16, 19, 23].

Recently, attention has been shifted to finding effective adversarial perturbation to CNN-based object detectors [13, 22]. Compared to image classification, finding effective perturbations for object detectors is more challenging, as the perturbation should affect not just the class label, but also the location and size of each object within the image. Existing methods [13, 22] mostly design specific loss functions based on the final prediction to disturb object

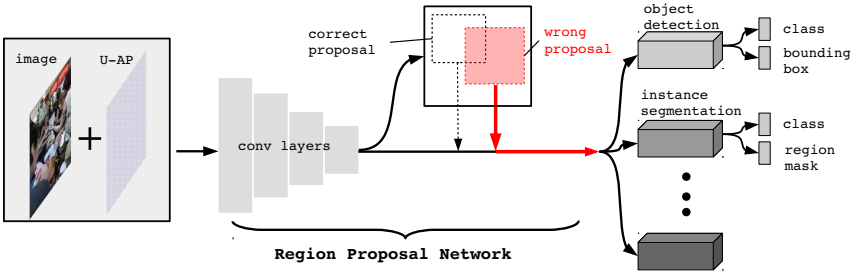

Figure 1: *Overview of the Robust Adversarial Perturbation (R-AP) method. Our method attacks Region Proposal Network (RPN) [17] in deep proposal-based object detectors and instance segmentation algorithms.*

class labels. As such, these methods are model dependent, which require detailed knowledge of the network architectures.

In this work, we develop a *Robust Adversarial Perturbation* (R-AP) method to universally attack deep proposal-based models that are fundamental to majority of object detectors and instance segmentation algorithms. Our method is based on the fact that a majority of recent object detectors and instance segmentation algorithms, *e.g.*, [7, 8, 11, 17] use a Region Proposal Network (RPN) [17] to extract object-like regions, known as *proposals*, from an image and then process the proposals further to obtain object class labels and bounding boxes in object detection, and the instance class labels and region masks in instance segmentation. If a RPN is successfully attacked by an adversarial perturbation, such that no correct proposals generated, the subsequent process in the object detection or instance segmentation pipeline will be affected. Figure 1 overviews the proposed R-AP method. The investigation of adversarial perturbation on deep proposal-based models can lead to further understanding of the vulnerabilities of these widely applied methods. The efforts can also aid improving the reliability and safety of the derived technologies, including computer vision guided autonomous cars and visual analytics.

The proposed R-AP method attacks a RPN based on the optimization of two loss functions: *(i)* the *label loss* and *(ii)* the *shape loss*, each of which targets a specific aspect of RPN. First, inspired by recent methods [13, 22] that attacks CNN-based object detectors, we design the label loss to disturb the label prediction (which indicates whether a proposal represents an object or not). Second, we also design a shape loss, which attacks the shape regression step in RPN, so that even if an object is correctly identified, the bounding box cannot be accurately refined to the object shape. Note that our R-AP method can be combined with existing adversarial perturbation method such as [22] to jointly attack corresponding network, since R-AP specifically focuses on attacking RPN, which is the intermediate stage of network compared to others which target the entire network.

Experimental validations are performed on the MS COCO 2014 [12], the current largest dataset used for training and evaluating mainstream object detectors and instance segmentation algorithms. Our experimental results demonstrate that the proposed R-AP attack can significantly degrade the performance of several state-of-the-art object detectors and instance segmentation algorithms, with minimal perceptible artifacts to human observers.

Our contributions are summarized in the following:

- To the best of our knowledge, this is the first work to thoroughly investigate the effects of adversarial perturbation on RPN, which universally affects the performance of deep

proposal-based object detectors and instance segmentation algorithms.

- In contrast to previous attack paradigms that only disturb object class label prediction, our method not only disturbs the proposal label prediction in RPN, but also distracts the shape regression, which can explicitly degrade the bounding box prediction in proposal generation.

2 Related Work

Deep proposal-based models follow a common paradigm of two steps — proposal generation and proposal refinement. A majority of recent object detectors and instance segmentation algorithms are deep proposal-based models. For object detection [2, 17], a Region Proposal Network (RPN) generates object proposals, which are refined in subsequent network modules for the exact bounding boxes and class labels. The state-of-the-art instance segmentation [8, 11] can be viewed as an extended version of object detection, which also use RPN to generate object proposals and refine them to semantic mask of objects.

Region Proposal Network (RPN) is a CNN-based model for object proposal generation. A RPN starts with a (manually specified) fixed size of multi-scale anchor boxes for each cell in feature map. At training phase, each anchor box is matched to ground truth. If the overlap between an anchor box and a ground truth is greater than threshold, this anchor box will be marked as positive example, otherwise negative example. Moreover, the shape offset between positive anchor box and the matched ground truth is recored for bounding box shape regression. At testing phase, the label and offset predictions of all anchor boxes are generated within a single forward. Compared to the selective search method [21] used in RCNN [5], RPN is much more efficient and accurate, such that it is widely used in current deep models to provide proposals.

Adversarial perturbation is an intentionally crafted noise that aims to perturb deep learning based models with minimal perception distortion to the image. Many methods [4, 6, 10, 14, 15, 16, 19, 23] have been proposed to fool image classifiers. Szegedy *et al*. [19] first described this intriguing property and formulated adversarial perturbation generation as an optimization problem. Goodfellow *et al*. [6] proposed an optimal max-norm constrained perturbations, referred as "fast gradient sign method" (FGSM), to improve the running efficiency. Kurakin *et al*. [10] proposed a "basic iterative method" which generates perturbation iteratively using FGSM. Papernot *et al*. [16] constructed an adversarial salieny map to indicate the desired places that can be affected efficiently. The DeepFool of Moosavi *et al*. [14] further improves the effectiveness of adversarial perturbation. Moosavi *et al*. [15] discovered the existence of image agnostic adversarial perturbations for image classifier.

Recently, adversary attack on object detectors has attracted many attentions. Lu *et al*. [13] attempted to generate adversarial perturbations on "stop" sign and "face" to mislead corresponding detectors. Xie *et al*. [22] proposed a dense adversarial generation method to iteratively incorrect predictions of object detectors. However, these methods are task-specific, which designs loss functions based on the final predictions. They do not address the adversarial perturbation universally. In contrast, we focus on attacking RPN, a common component of deep proposal-based models, to universally degrade their performance without knowing the details of their architecture.

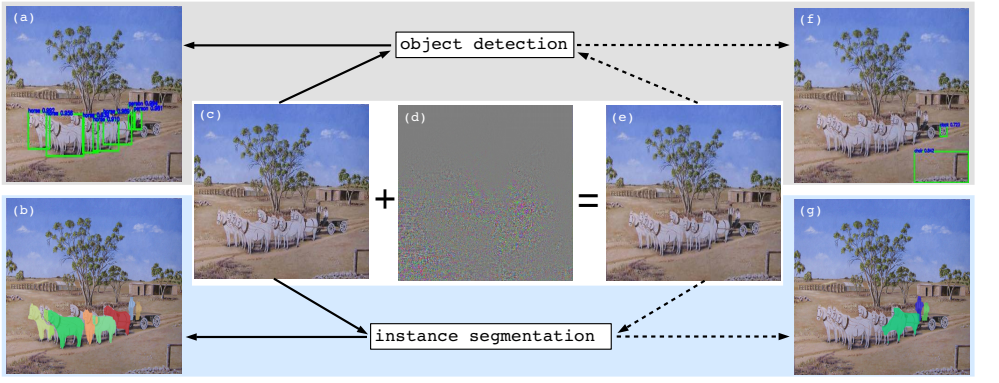

Figure 2: *Illustration of performing R-AP on the object detector Faster-RCNN [17] and the instance segmentation algorithm FCIS [17]. (c) is the original image. (d) is the noise generated from R-AP, amplified by a factor of* 10 *for better visibility. (e) is the perturbed image by adding (c) and (d). (a,f) are the Faster-RCNN object detection of (c,e), and (b,g) are the FCIS instance segmentation of (c,e), respectively.*

3 Method

The proposed method attacks deep proposal-based object detectors and instance segmentation algorithms by adding minimal adversarial noises to the input image that can effectively disturb the predictions of Region Proposal Network (RPN). Given an input image and a pre-trained RPN, we design a specific objective function, a combination of two terms — the label loss and the shape loss, to calculate the adversarial perturbation. In particular, we optimize this objective function with respect to image using an iterative gradient based method.

Note all mainstream deep proposal-based object detectors and instance segmentation algorithms rely on a few standard RPNs to provide proposals for subsequent processes. Once the RPN is disturbed, the performance of these deep models is naturally degraded. As such, our R-AP method is suitable in nature for black-box attack to these models, *i.e.*, without the need to know their implementation details. Inspired by [17], we generate adversarial perturbations for different RPN and combine them together to improve the robustness of black-box attack. Figure 2 illustrates an example of R-AP attack on object detection and instance segmentation.

Section 3.1 describes the notations and the general paradigm of label loss for generating adversarial perturbations. Then we introduce our new shape loss that can explicitly disturb proposal shape regression. Section 3.2 presents the details of iterative adversarial perturbation generation scheme.

3.1 Notations and Problem Formulation

Denote $\mathcal{I}$ as the input image that contains n ground truth bounding boxes for objects $\{\bar{b}_i = (\bar{x}_i, \bar{y}_i, \bar{w}_i, \bar{h}_i)\}_{i=1}^{n}$, where $\bar{x}_i, \bar{y}_i, \bar{w}_i, \bar{h}_i$ are the x- and y-coordinate of the box center point, the width and height of bounding box $\bar{b}_i$, respectively. Let $\mathcal{F}_\theta$ denote a Region Proposal Network (RPN) with model parameters θ. Let $\mathcal{F}_\theta(\mathcal{I}) = \{(s_j, b_j)\}_{j=1}^{m}$ denotes the set of m generated proposals with input image $\mathcal{I}$, where s_j denotes the confidence score (probability after sigmoid function) of j-th proposal and b_j is the bounding box of j-th proposal. Let

$b_j = (x_j, y_j, w_j, h_j)$, where x_j, y_j, w_j, h_j are the x- and y-coordinate of the box center point, the width and height of bounding box b_j, respectively.

Our goal is to seek an minimal adversarial perturbation added to image $\mathcal{I}$ to fail a RPN . The adversarial perturbation generation can be casted as an optimization problem of specific designed loss. In our method, we design the loss as the summation of *(i)* the *label loss* L_{label}, which is a general paradigm used in previous methods to disturb label prediction, and *(ii)* the *shape loss* L_{shape}, which is our newly proposed term to explicitly disturb bounding box shape regression. As *Peak Signal-to-Noise Ratio* (PSNR) is an approximation of human perception of image quality, we employ it to evaluate the distortion of adversarial perturbation. Less perturbation results in higher PSNR. Throughout this work, we assume the model parameters θ of PRN are fixed, and the R-AP algorithm generates a perturbed image $\mathcal{I}$ by optimizing the following loss as

$$\min_{\mathcal{I}} \quad L_{label}(\mathcal{I}; \mathcal{F}_\theta) + L_{shape}(\mathcal{I}; \mathcal{F}_\theta), \ \text{s.t. PSNR}(\mathcal{I}) \geq \varepsilon, \tag{1}$$

where $\text{PSNR}(\mathcal{I})$ denotes the PSNR of luminance channel in image $\mathcal{I}$, ε is the lower bound of PSNR. We describe the label loss L_{label} and shape loss L_{shape} in sequel.

Label Loss. The label loss L_{label} is designed to disrupt the label prediction of proposals, which is in analogy to existing adversarial perturbation methods [13, 22] for object detectors. Denote $z_j \in \{0, 1\}$ as the indicator of j-th proposal, where $z_j = 1$ means that j-th proposal is positive, otherwise negative. We let $z_j = 1$ if (1) the bounding box intersect-over-union (IoU) of j-th proposal with an arbitrary ground truth object is greater than a preset threshold μ_1; (2) the confidence score of j-th proposal is greater than another preset threshold μ_2, otherwise we set $z_j = 0$. The above rule can be formulated as $z_j = 1$, if $\exists\, i$, $\text{IoU}(\bar{b}_i, b_j) > \mu_1$ and $s_j > \mu_2$, and 0 otherwise.

Note that RPN initially generates a large amount of proposals, and in R-AP, we only focus on disturbing positive proposals as they are the key to the subsequent algorithms. The label loss L_{label} is given by

$$L_{label}(\mathcal{I}; \mathcal{F}_\theta) = \textstyle\sum_{j=1}^{m} z_j \log(s_j). \tag{2}$$

In other words, minimizing this loss is equivalent to decreasing confidence score of positive proposals.

Algorithm 1 *Adversarial Perturbation Generation*

Require: RPN model $\mathcal{F}_\theta$; input image $\mathcal{I}$; maximal iteration number T.

$\mathcal{I}_0 = \mathcal{I}, t = 0$;
while $t < T$ and $\sum_{j=1}^{m} z_j \neq 0$ **do**
  $\hat{p}_t = \nabla_{\mathcal{I}_t}(L_{label} + L_{shape})$;
  $p_t = \frac{\lambda}{||\hat{p}_t||_2} \cdot \hat{p}_t$; ▷ λ *is a fixed scale parameter*
  $\mathcal{I}_{t+1} = \text{clip}(\mathcal{I}_t - p_t)$;
  if $\text{PSNR}(\mathcal{I}_t) < \varepsilon$ **then**
    break
  $t = t + 1$;
$p = \mathcal{I}_t - \mathcal{I}_0$;

Ensure: adversarial perturbation p

Shape Loss. Shape regression is an important step to refine the bounding box of object detections or proposals. Specifically, in RPN, shape regression is used to adjust the anchor

Table 1: *Performance of R-AP on 6 state-of-the-art object detectors at mAP* 0.5 *and* 0.7. *Lower value denotes better attacking performance.*

	FR-v16	**FR-mn**	**FR-rn50**	**FR-rn101**	**FR-rn152**	**RFCN [2]**
origin	59.2/47.3	47.1/32.6	59.5/49.4	63.5/53.6	64.8/54.5	60.1/50.0
random	58.7/46.5	46.5/32.6	59.6/48.9	63.2/53.2	64.6/54.4	59.9/49.6
v16 (p_1)	**5.1/3.1**	34.8/22.2	47.9/36.8	52.7/42.4	55.5/45.0	54.5/43.8
mn (p_2)	56.8/44.4	**11.0/6.1**	56.7/45.2	60.6/50.2	62.3/51.4	57.5/46.6
rn50 (p_3)	53.8/41.2	39.5/25.7	**10.5/6.6**	52.8/42.2	55.9/44.7	53.7/42.6
rn101 (p_4)	54.8/42.6	41.0/27.4	50.0/39.2	**16.8/11.0**	56.0/45.3	52.0/40.4
rn152 (p_5)	55.0/41.9	41.8/27.4	49.8/38.3	53.6/42.2	**17.3/10.6**	54.4/42.9
$\mathbf{P} = \alpha \cdot \sum_{i=1}^{5} p_i$	37.5/25.6	26.4/16.5	31.3/21.3	37.9/27.2	41.4/30.1	**47.0/35.9**

boxes to the ground truth bounding boxes of the object by minimizing the offset between them. Therefore, we design a specific shape loss to explicitly disturb the bounding box shape regression. Let $\Delta x_j, \Delta y_j, \Delta w_j, \Delta h_j$ be the predicted x- and y- coordinate center location offsets, width and height offset of bounding box b_j, respectively. To explicitly disturb the shape regression, we define a new loss function L_{shape} as

$$L_{shape}(\mathcal{I}; \mathcal{F}_\theta) = \sum_{j=1}^{m} z_j((\Delta x_j - \tau_x)^2 + (\Delta y_j - \tau_y)^2 + (\Delta w_j - \tau_w)^2 + (\Delta h_j - \tau_h)^2), \quad (3)$$

where $\tau_x, \tau_y, \tau_w, \tau_h$ are large offsets defined to substitute the real offset between anchor boxes and matched ground truth bounding boxes. We are only concerned about the predicted offset of positive proposals, as it is inappropriate to consider the bounding boxes of negative proposals . By minimizing Eq. (3), the R-AP method forces predicted offset $\Delta x_j, \Delta y_j, \Delta w_j, \Delta h_j$ approaching $\tau_x, \tau_y, \tau_w, \tau_h$ respectively, such that the shape of bounding box b_j will be incorrect.

3.2 The Robust Adversarial Perturbation (R-AP) Algorithm

To generate the proposed R-AP, we optimize Eq. (1) using an iterative gradient descent scheme, as mentioned in [22]. Let t denote iteration number. We calculate the gradient of $L_{label} + L_{shape}$ with respect to image $\mathcal{I}$ at t as $\hat{p}_t$. We normalize $p_t = \frac{\lambda}{\|\hat{p}_t\|_2} \cdot \hat{p}_t$ to keep perturbation minimal perceptive and stability of each iteration, where λ is a fixed scale parameter, $\|\cdot\|_2$ is $L2$ norm metric. Then image $\mathcal{I}_{t+1}$ is updated by $\mathcal{I}_t - p_t$ and we clip the pixel value back to $[0, 255]$ at the end of each iteration. The process is repeated until (1) the maximum iteration number T is reached, or (2) positive proposals cease to exist, *i.e.*, $\sum_{j=1}^{m} z_j = 0$, or (3) *Peak Signal-to-Noise Ratio* (PSNR) is less than a threshold ε. The algorithm of adversarial perturbation generation is listed in Algorithm 1.

Note that R-AP is not mutually exclusive to other adversarial perturbation method such as [22] for object detectors. For instance, we can combine R-AP with method in [22] to generate more effective adversarial perturbations, since our loss function is based on a different stage of networks compared to other state-of-the-arts algorithms.

4 Experimental Results

In this section, we report the experimental evaluation of R-AP on several state-of-the-art object detectors and instance segmentation algorithms. Section 4.1 describes the dataset

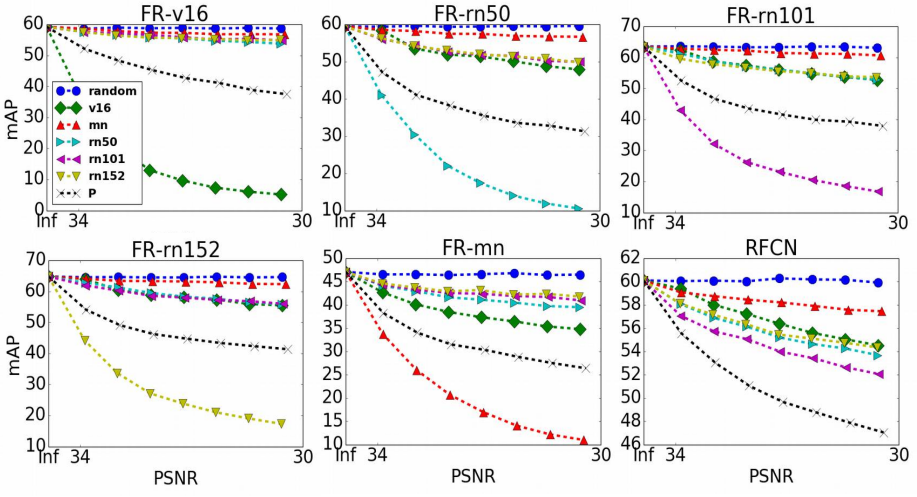

Figure 3: *Illustration of R-AP performance under different PSNR value at mAP 0.5 on 6 object detectors.*

and evaluation metric. Section 4.2 describes the R-AP settings in all experiments. Section 4.3 and section 4.4 presents the R-AP attack experiments on object detectors and instance segmentation algorithms respectively.

4.1 Dataset

The performance of the R-AP is evaluated on MS COCO 2014 dataset [12], which is a large scale dataset containing 80 object categories for multiple tasks, including object detection and instance segmentation. Experiments are conducted on a subset (random 3000 images) of the MS COCO 2014 validation set and evaluated using "mean average precision" (mAP) metric [3] at intersection-over-union (IoU) threshold 0.5 and 0.7.

4.2 R-AP Settings

We generate adversarial perturbations for five different RPN architectures: vgg16 (**v16**) [18], mobilenet (**mn**) [9], resnet50 (**rn50**), resnet101 (**rn101**) and resnet152 (**rn152**) [7], where the adversarial perturbations are denoted p_1, p_2, p_3, p_4, p_5, respectively. These RPN architectures are extracted from Faster-RCNN (**FR**) object detectors [17] implemented by [1]. We also generate Gaussian noise (**random**) as a perturbation baseline in comparison to demonstrate the effectiveness of R-AP. Inspired by [22], we accumulate these perturbations as $\mathbf{P} = \alpha \cdot \sum_{i=1}^{5} p_i$, where α is a scale parameter to control the distortion.

The following parameter values are used throughout the paper: overlap threshold $\mu_1 = 0.1$, confidence score threshold $\mu_2 = 0.4$, offset $\tau_x = \tau_y = \tau_w = \tau_h = 10^5$, scale parameters $\lambda = 30, \alpha = 0.5$, and maximum iteration number $T = 210$. In general, typical values of PSNR in lossy image compression is between 30 and 50 dB [20]. Therefore, we set $\varepsilon = 30$ dB as the lower bound of PSNR. In our experiments, all perturbations are terminated at maximum iteration number T with PSNR $> \varepsilon$.

4.3 Object Detection

We study six state-of-the-art object detectors, including five Faster-RCNN based methods **FR-v16**, **FR-mn**, **FR-rn50**, **FR-rn101**, **FR-rn152** with various base networks, and the Region Fully Convolutional Network **RFCN** [2]. We denote FR-v16 as vgg16 based Faster-RCNN in short, and the others are named in the same way accordingly throughout the paper.

Table 1 illustrates the performance of R-AP generated from different RPN on the six object detectors, where we report mAP metric at 0.5 and 0.7. The 2^{nd} row of Table 1 (random) shows the added Gaussian noise as perturbation for a baseline comparison, and the result shows that the performance degradation is $< 1\%$. In contrast, the R-AP generated from *v16, mn, rn50, rn101, rn152* can each reduce the performance of the object detectors by a larger amount. Since we extract *v16, mn, rn50, rn101, rn152* based RPN from *FR-v16, FR-mn, FR-rn50, FR-rn101, FR-rn152* detectors respectively, the R-AP works as white-box attack for their corresponding RPNs. Thus, the degradation for the respective object detector (highlighted in bold) is significantly larger. For example, the detection performance of *FR-v16* is degraded greatly from 59.2% to 5.1% at mAP 0.5, and from 47.3% to 3.1% at mAP 0.7.

In comparison, the *RFCN* works as a black-box detector in our experiment. Observe that in the *RFCN* column of Table 1, the R-AP generated from *v16, mn, rn50, rn101, rn152* based RPN can effectively reduce the detection performance. In particular, the R-AP based on *rn101* can reduce the performance from 60.1% to 52.0% at mAP 0.5, and from 50.0 to 40.4% at mAP 0.7. The last row (*P* for RFCN) shows the scaled accumulation of 5 perturbations $(p1, p2, p3, p4, p5)$, which essentially represents the combination of effects learned from multiple networks that can notably degrade the performance of *RFCN* by 13.1% at mAP 0.5 and 14.1% at mAP 0.7.

The performance of R-AP under different PSNR value on six object detectors are shown in Figure 3. Observe that Gaussian noise (random) only produces small effects as the PSNR decreased. In contrast, the mAP curves of *v16, mn, rn50, rn101, rn152* in respective detector plots drop significantly compared to others. The black curve in each plot is the performance of accumulated perturbation *P*. We can see in pure black-box object detector *RFCN*, the accumulated perturbation curve drops notably and achieves the best results.

We illustrate the visual performance of accumulated P on several object detectors in the first four rows of Figure 5. Due to the degradation of RPN after R-AP attack, the person in *FR-rn50* (b) is not detected. For the case of *FR-mn* (d), despite the target is correctly identified, the bounding box is disturbed to an undesired shape.

4.4 Instance Segmentation

We evaluate the proposed R-AP on attacking two of the state-of-the-art instance segmentation methods in a black-box setting — **FCIS** [11] and Mask-RCNN (**MR**) [8]. Table 2 summarizes the performance degradation after applying R-AP at mAP 0.5 and 0.7. The 2^{nd} row of the table (random) shows a baseline by adding simple Gaussian noise as the perturbation, which is known to be ineffective in attacking, *i.e.* with only $< 1\%$ performance decreased. In contrast, R-AP based on *v16, mn, rn50, rn101, rn152* each leads to large degradation of the performance. In particular, R-AP based on *rn101* degrades the performance of *FCIS* by 10.2% at mAP 0.5 and 9.8% at mAP 0.7, and reduces *MR* performance by 9.3% at mAP 0.5 and 7.8% at mAP 0.7. Notably, the accumulated perturbation *P* degrades the performance of both methods — a decrease of 15.1%/13.2% on *FCIS* and 16.7%/13.8% on *MR*.

Figure 4 shows the performance evaluation of the R-AP black-box attack under different

Table 2: *Performance of black-box attack on 2 state-of-the-art instance segmentation algorithms at mAP* 0.5 *and* 0.7. *Lower value denotes better attacking performance.*

	FCIS [11]	**MR** [8]
origin	61.0/45.8	60.3/45.6
random	60.4/45.6	59.4/45.0
v16 (p_1)	54.7/40.0	51.3/38.1
mn (p_2)	57.6/42.4	55.5/41.9
rn50 (p_3)	52.8/38.2	52.0/38.1
rn101 (p_4)	50.8/36.1	51.0/37.8
rn152 (p_5)	53.4/39.1	51.9/38.8
$\mathbf{P} = \alpha \cdot \sum_{i=1}^{5} p_i$	**45.9/32.6**	**43.6/31.8**

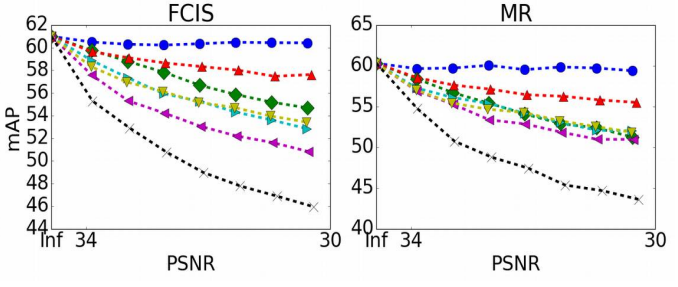

Figure 4: *Illustration of R-AP performance under different PSNR value at mAP 0.5 on 2 instance segmentation algorithms.*

PSNR on the two instance segmentation methods. The blue curve corresponding to Gaussian noise is mostly flat, which shows the inefficacy of attack. In contrast, R-AP is effective for both instance segmentation methods. Notably, the black curve corresponding to the accumulated perturbation P achieves the largest degredation among all.

Visual illustration of the accumulated P attack for instance segmentation is shown in the last two rows in Figure 5. Observe that the object instances are poorly segmented after the R-AP perturbation.

5 Conclusion

In this paper, we propose a robust adversarial perturbation (R-AP) method to attack deep proposal-based object detectors and instance segmentation algorithms. To the best of our knowledge, this work is the first to investigate the universal adversarial attack of the deep proposal-based models. Our method focuses on attacking Region Proposal Network (RPN), a component used in most deep proposal-based models, to universally affect the performance of their respective tasks. We describe a new loss function to not only disturb label prediction but also degrade shape regression. Evaluations on the MS COCO 2014 dataset shows that the R-AP can effectively attack several state-of-the-art object detectors and instance segmentation algorithms.

Future work includes conducting further experiments on additional deep proposal-based models, including the part-based human pose detection. This work also opens up new opportunities on how to effectively improve the robustness of RPN-based networks.

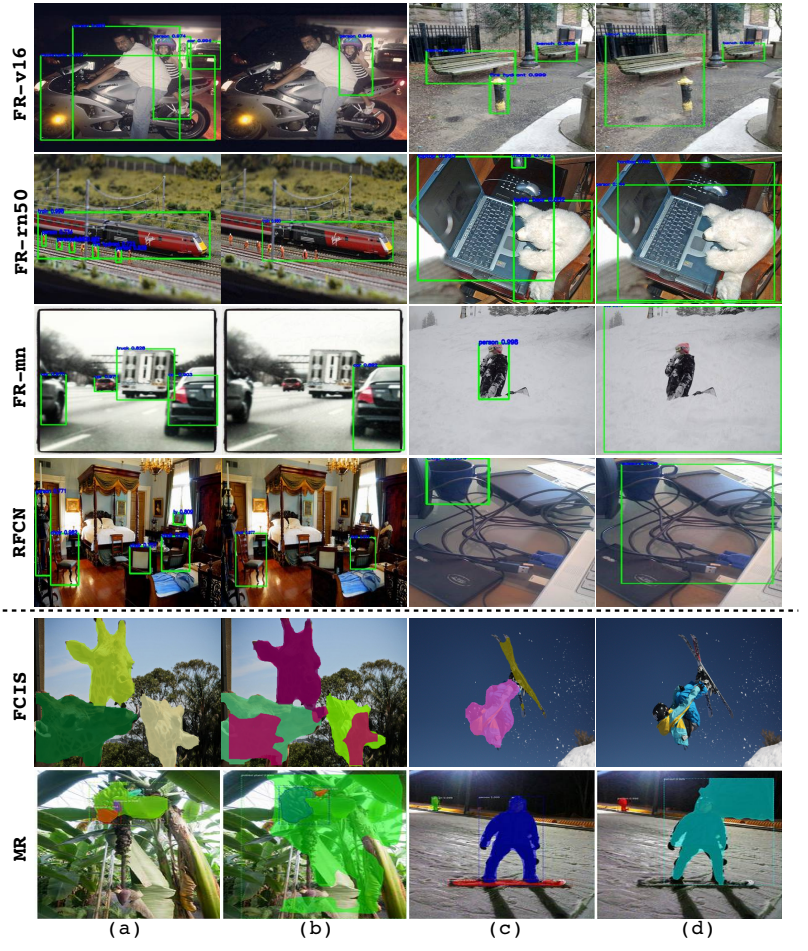

Figure 5: *Visual results of the R-AP attack on several mainstream object detectors (first 4 rows) and instance segmentation algorithms (last 2 rows). Columns (a,c) are the original results, and (b,d) are the R-AP attacked results.*

References

[1] Xinlei Chen and Abhinav Gupta. An implementation of Faster RCNN with study for region sampling. *arXiv 1702.02138*, 2017.

[2] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*. 2016.

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 2010.

[4] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. In *CVPR*, 2018.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[9] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv 1704.04861*, 2017.

[10] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR*, 2017.

[11] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, 2017.

[12] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[13] Jiajun Lu, Hussein Sibai, and Evan Fabry. Adversarial examples that fool detectors. *arXiv 1712.02494*, 2017.

[14] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

[15] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017.

[16] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *EuroS&P*, 2016.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *TPAMI*, 2017.

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 2014.

[19] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv 1312.6199*, 2013.

[20] Prabhakar Telagarapu, V Jagan Naveen, A Lakshmi Prasanthi, and G Vijaya Santhi. Image compression using DCT and wavelet transformations. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 2011.

[21] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *IJCV*, 2013.

[22] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, 2017.

[23] Xiaohui Zeng, Chenxi Liu, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. *arXiv 1711.07183*, 2017.