Exploring the Vulnerability of Single Shot Module in Object Detectors via Imperceptible Background Patches

BMVC CARDIFF 2019

Yuezun Li [1] Xiao Bian [2] Ming-Ching Chang [1] Siwei Lyu [1]

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

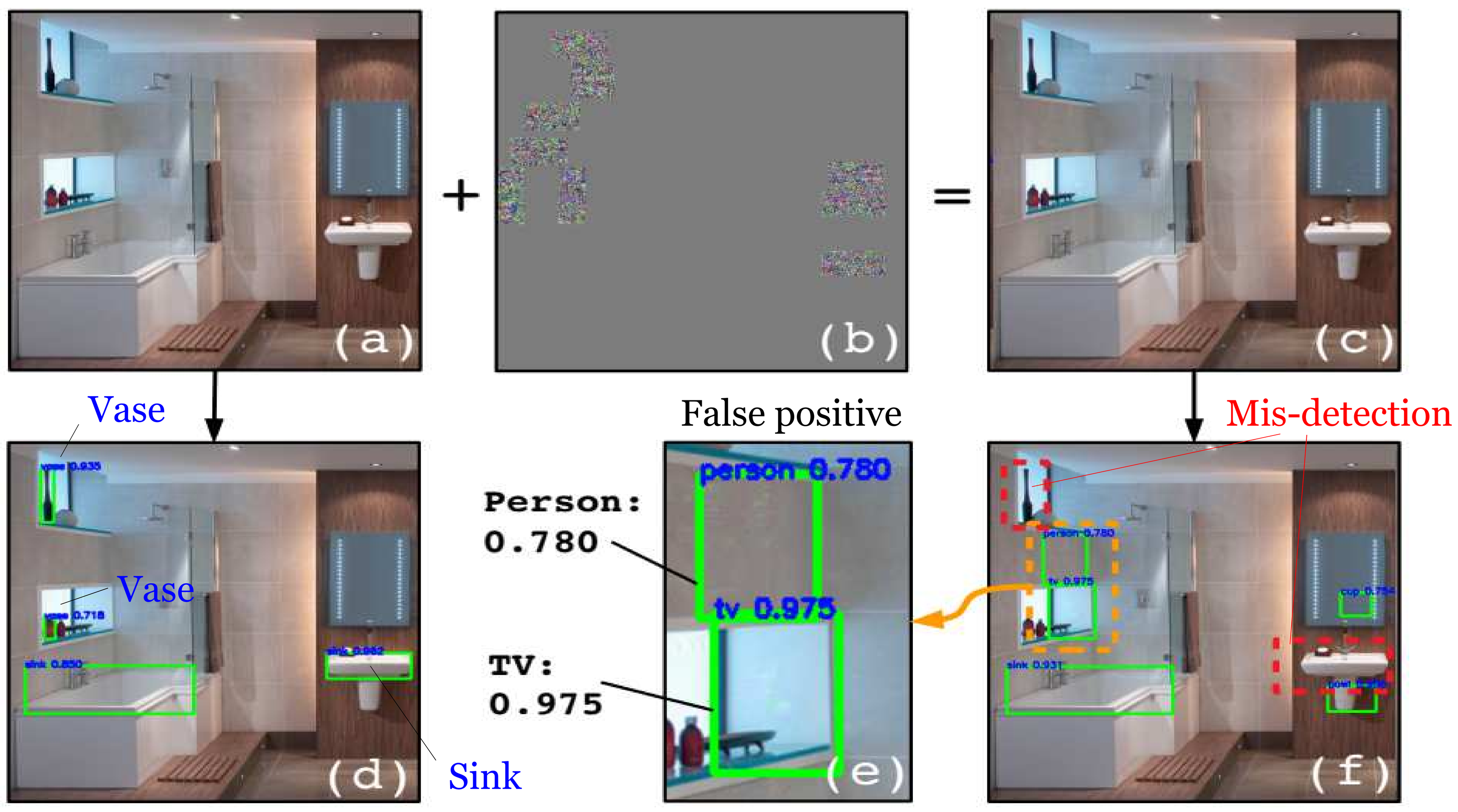
Abstract

We explore vulnerability of the **Single Shot Module (SSM)** commonly used in recent object detectors, by adding small perturbations to patches in the background outside object of interest to attack the object detection task.

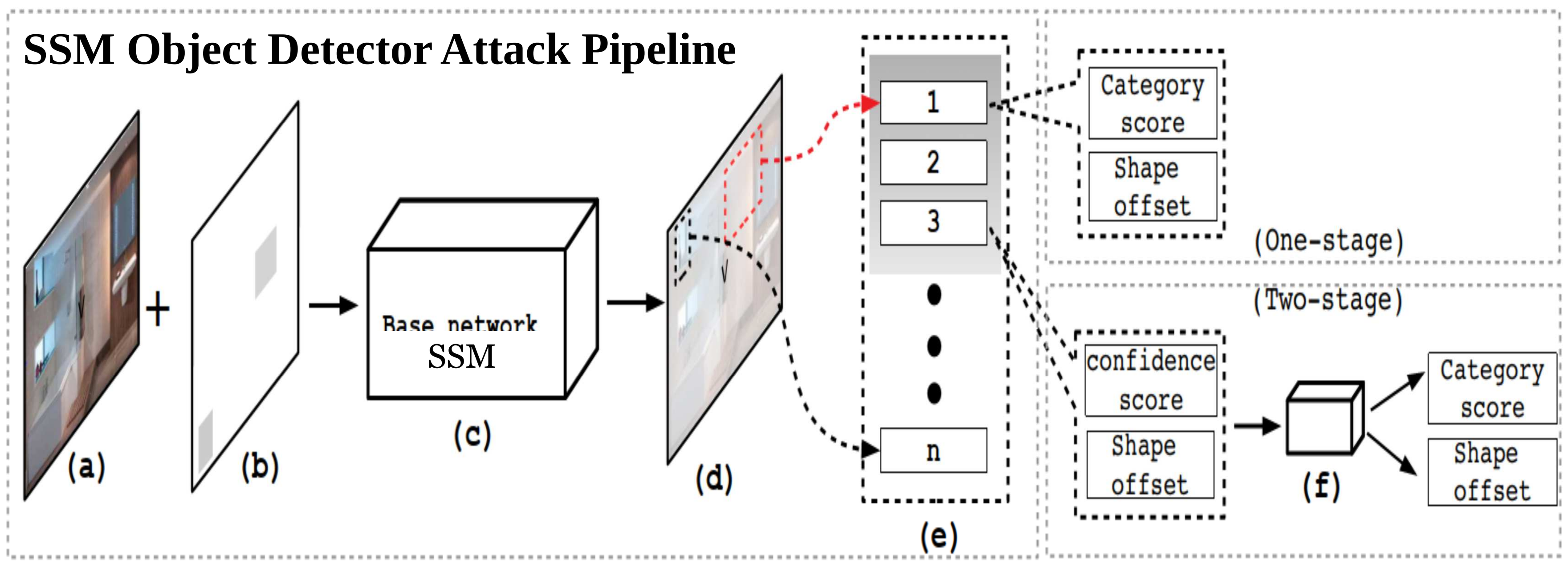
Backgrounds

- Single Shot Module (SSM) refers to (1) the Region Proposal Network (RPN) commonly used in two-stage object detectors, or (2) the single-stage object detector itself.
- Adversarial perturbations are intentionally designed noises that are imperceptible to human observers yet can seriously harm the performance of a deep neural network.

Overview



Visual illustration of the SSM background patch attack on object detectors.



(a) Input image.
(b) Background patches generated from our method.
(c) SSM base-network (RPN of two-stage detectors or the single-stage detector).
(d) Output of SSM. *Our attack can effectively disrupt the top ranked results by decreasing true positives (black) and increasing false positives (red).*
(e) Top ranked object proposals (for two-stage detectors) or detection results (for single-stage detectors).
(f) Sub-network of two-stage object detectors for class labels prediction and shape refinements.

This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number 73042CS.

Methods

Problem Formulation

The proposed SSM adversarial attack is to search for suitable background patches in terms of ***geometry*** *(location, size and shape)* and ***pixel changes*** to be altered.

Given input image $\mathcal{I}$, we formulate this optimization as the minimization of three loss terms:

(1) True Positive Class (**TPC**) loss: L_{tpc}
(2) True Positive Shape (**TPS**) loss: L_{shape}
(3) False Positive Class (**FPC**) loss: L_{fpc}

$$\min_{\mathcal{I}\odot\mathcal{Q}} \left\{L_{tpc}(\mathcal{I}\odot\mathcal{Q};\mathcal{F})+L_{shape}(\mathcal{I}\odot\mathcal{Q};\mathcal{F})+L_{fpc}(\mathcal{I}\odot\mathcal{Q};\mathcal{F})\right\},$$
$$\text{s.t. } \mathrm{PSNR}(\mathcal{I}\odot\mathcal{Q}) \geq \varepsilon,$$

Threshold : 35 dB

Location and shape of the background patches $\mathcal{Q}$ and the included pixel value in image $\mathcal{I}$

SSM

Background Patches Generation

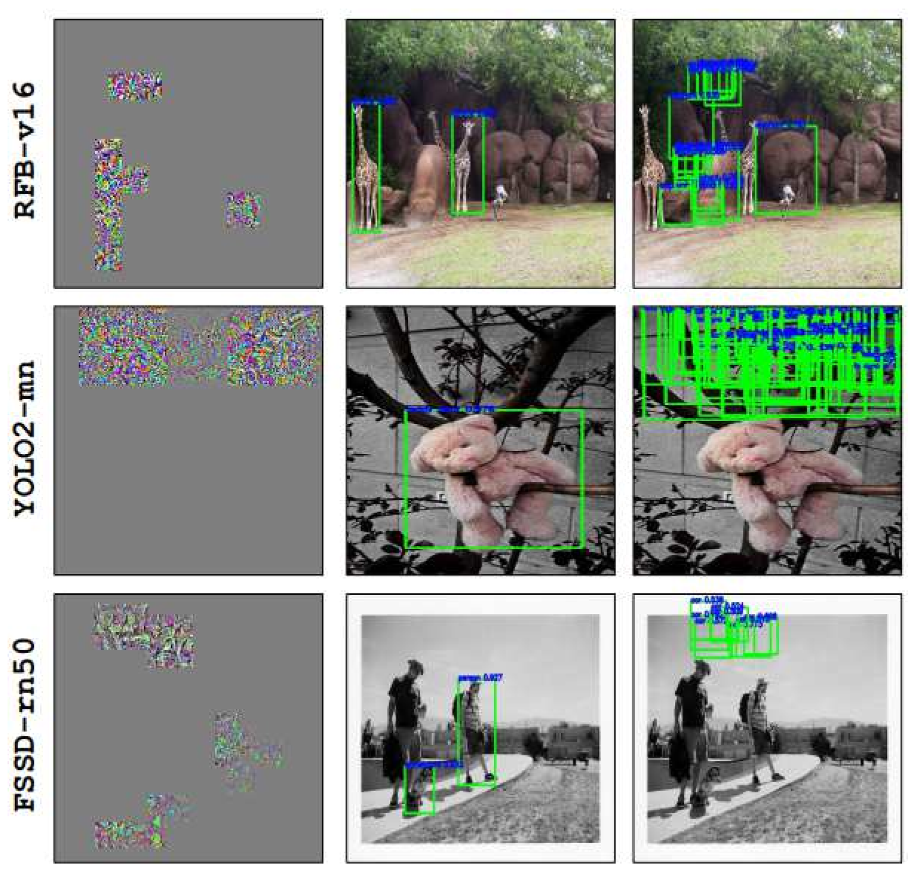
Algorithm 1 *Background Patch Generation*

Require: SSM model $\mathcal{F}$; input image $\mathcal{I}$; maximal iteration T
$\mathcal{I}_0 = \mathcal{I}, t = 0$
while $t < T$ and $\sum_{j=1}^{M} z_j \neq 0$ **do**
  $\mathcal{G}_t = \nabla_{\mathcal{I}_t}\left[L_{tpc}(\mathcal{I}_t;\mathcal{F})+L_{shape}(\mathcal{I}_t;\mathcal{F})+L_{fpc}(\mathcal{I}_t;\mathcal{F})\right]$
  if $t = 0$ **then**
    $\mathcal{Q}_0 \leftarrow$ initial background patches
  else
    $\mathcal{Q}_t \leftarrow$ expanded background patches
  $\mathcal{P}_t = \mathcal{G}_t \odot \mathcal{Q}_t$
  $\hat{\mathcal{P}}_t = \frac{\lambda}{||\mathcal{P}_t||_2} \cdot \mathcal{P}_t$
  $\mathcal{I}_{t+1} = \mathrm{clip}(\mathcal{I}_t - \hat{\mathcal{P}}_t)$
  if $\mathrm{PSNR}(\mathcal{I}_{t+1} \odot \mathcal{Q}_t) < \varepsilon$ **then**
    break
  $t = t + 1$
Ensure: Adversarial perturbed image $\mathcal{I}_t$

Prefer: (1) distance between background patch and objects greater than a threshold, (2) patch with largest sum of gradient intensities, (3) no overlap between selected patches.

Expanding in one of the 4 possible directions (left, right, top, down) and the expanding direction is determined by whose gradient intensity increases the most for the patch.

Results



** FR: Faster-RCNN*

*Experiment on **5** two-stage object detectors with **5** different Region Proposal Networks (RPNs) and **8** single-stage object detectors at mAP at th = 0.5 / 0.7 (1st / 2nd value in table).*
** v16:vgg16, mn: MobileNet, rn50:ResNet50, rn101:ResNet101, rn152: ResNet152.*

	No Noise	Random	TPC+TPS+FPC
FR-v16	62.4/48.7	62.5/48.9	**41.9/32.7**
FR-mn	46.1/32.9	46.4/32.9	**26.6/19.3**
FR-rn50	64.7/52.7	64.7/52.2	**39.8/33.4**
FR-rn101	66.0/56.0	65.8/55.7	**36.2/31.2**
FR-rn152	70.0/60.0	69.1/58.9	**36.8/31.7**
SSD-rn50	46.6/37.2	47.2/37.1	**27.9/20.9**
SSD-v16	48.3/37.0	47.8/37.1	**24.5/17.4**
RFB-rn50	48.9/40.3	48.7/41.2	**26.1/20.5**
RFB-v16	48.3/37.9	46.5/37.3	**26.0/19.4**
YOLO2-mn	46.6/30.4	45.4/29.9	**22.3/15.3**
YOLO3-mn	49.0/36.0	49.6/36.5	**33.3/21.8**
FSSD-rn50	51.2/41.5	51.4/42.2	**28.8/20.8**
FSSD-v16	54.0/44.2	53.9/43.5	**33.5/24.1**

Conclusion

- The proposed **SSM background-patch attack** can effectively harm mainstream deep object detection networks *by only altering imperceptible pixels in the background* that results in significantly decreased true positives and increased false positives.
- Experiments on mainstream object detectors expose such vulnerability.