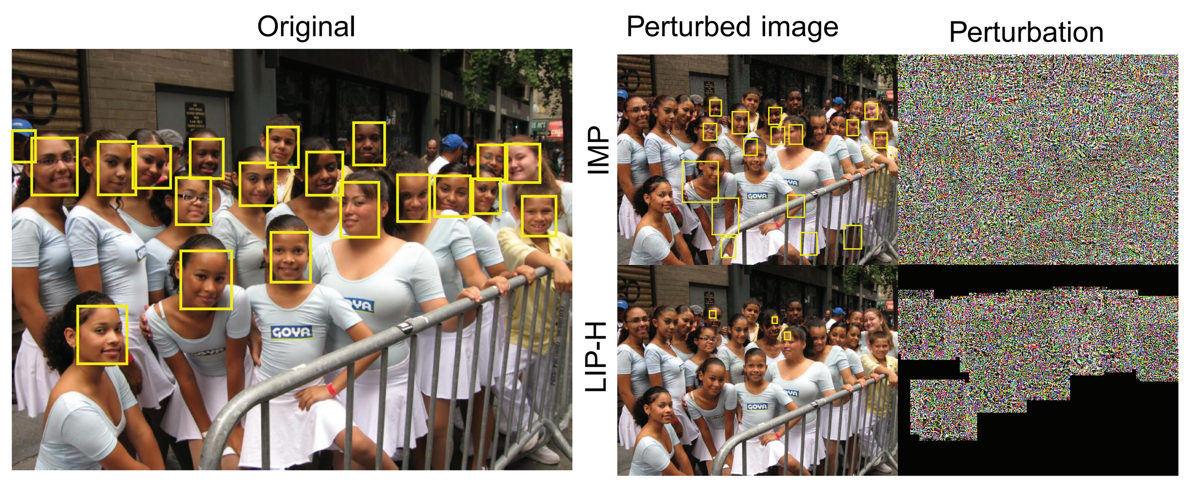
Using LIP to Gloss Over Single-Stage Face Detection Networks

Siqi Yang, Arnold Wiliem, Shaokang Chen and Brian C. Lovell
The University of Queensland, Australia

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

ECCV 2018 European Conference on Computer Vision

Can we attack a face detector?



Adversarial Perturbations:

- Imperceptible perturbations that change the neural network output significantly
- Fast Gradient Sign Method (FGSM) [1]:

$$X^{adv} = X + \alpha \cdot sign(\nabla_x \ell(f_\theta(X), y^{true}))$$

- Prior works are in image classification [1], semantic segmentation [2,3] and object detection [3]
- The attack in object detection is more difficult: Need to ensure all region proposals associated with the object/instance are successfully attacked

We are the first to study adversarial attack in single-stage face detection:

- Single-stage detector: Performs object classification and localization simultaneously, e.g. YOLO and SSD. This work uses the face detector, HR [4]

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. ICLR, 2015.
[2] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. In ICCV, 2017.
[3] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. Adversarial examples for semantic segmentation and object detection. In ICCV, 2017.
[4] P. Hu and D. Ramanan. Finding tiny faces. In CVPR, 2017.
[5] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.
[6] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In NIPS, 2016.
[7] V. Jain and E. G. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. UMass Amherst Technical Report, 2010.

Acknowledgements

This work has been funded by Sullivan Nicolaides Pathology, Australia, and the Australian Research Council (ARC) Linkage Projects Grant LP160101797. Arnold Wiliem is funded by the Advance Queensland Early-Career Research Fellowship.

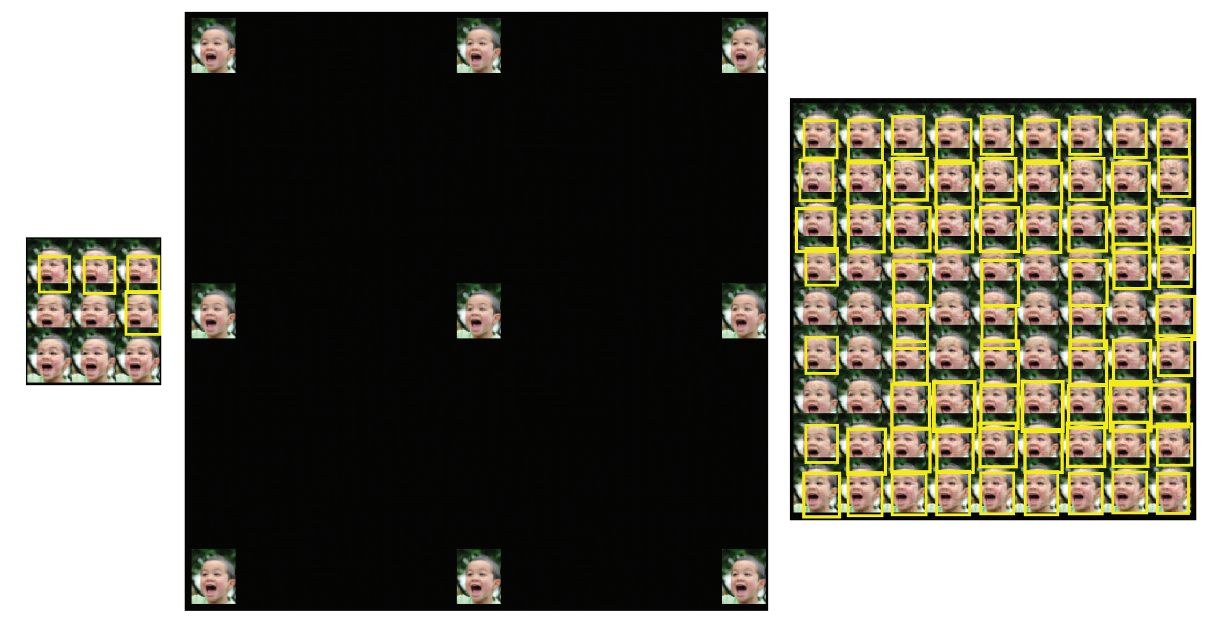
Instance Perturbation Interference (IPI) Problem

IMage based Perturbation (IMP):

- Following the FGSM, the perturbations are generated and applied w.r.t. the entire image

Existence of the IPI problem:

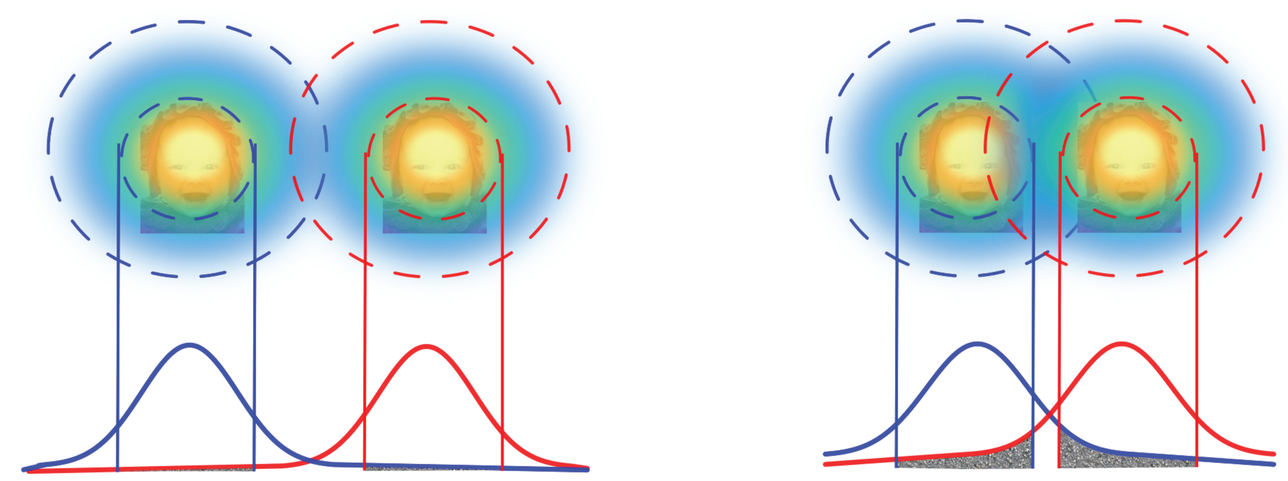
Number of Faces	Distance	Attack Success Rate (%)
1	40	100
9	40	51.5
	160	56
	240	63.9
64	40	18.3



(a) (b) (c)

- The attack success rate drops when the number of faces increases
- With the same number of faces, the attack success rate can be increased as the distances among faces increase

Proposed Method: LIP



Perturbations overlap with the neighboring face ERF, which might disrupt the attack

Explanations of the IPI problem:

- Our adversarial perturbation is a 2D Gaussian distribution:

$$\nabla_X L(f_\theta(X,t_c),-1) = \frac{\partial L(f_\theta(X,t_c),-1)}{\partial f_\theta(X,t_c)} \boxed{\frac{\partial f_\theta(X,t_c)}{\partial X}}$$

In CNNs, the distribution of impact within the Theorectical Receptive Field (TRF) follows a 2D Gaussian distribution [6]:

$$\frac{\partial f_\theta(X,t_c)}{\partial X}$$

- The Effective Receptive Field (ERF) is a fraction of TRF, where pixels have significant impact to the neuron decision [6]

Localized Instance Perturbation (LIP):

Aim: eliminating the interfering perturbation

- Perturbation cropping according to the instance ERF:

$$R_{m_i} = C_{e_i} \cdot \nabla_X L_{m_i} \quad \text{, where } C_{e_i}(w,h) = \begin{cases} 1, (w,h) \in e_i \\ 0, otherwise \end{cases}$$

- Individual instance perturbation (processing each instance separately): $R = \sum_{i=1}^{N} C_{e_i} \cdot \nabla_X L_{m_i}$

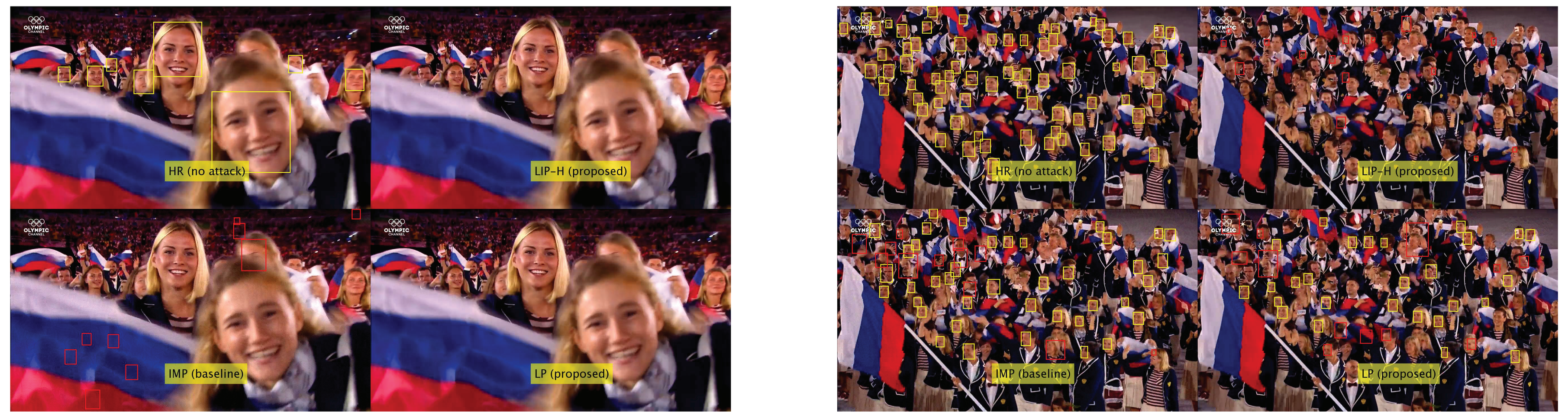
15seconds-Summary

Questions: Why existing adversarial perturbation methods are not effective when there are multiple objects/instances?

Contributions:

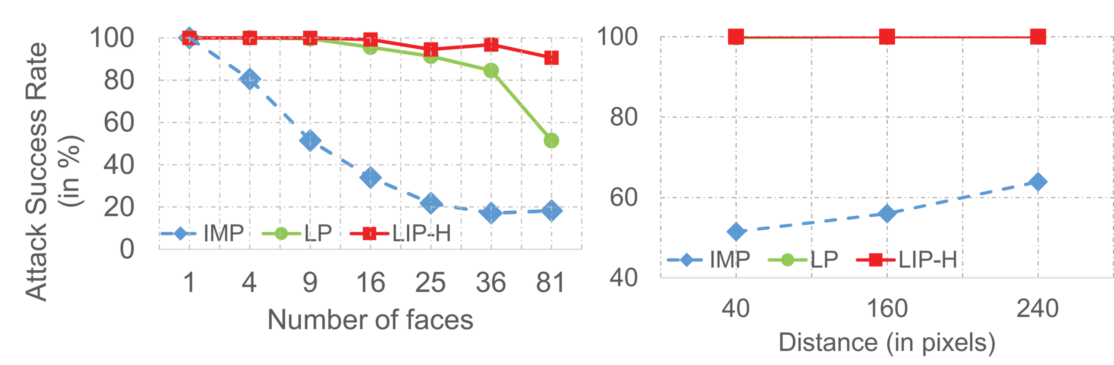
1. **IPI Problem:** The interfering perturbations disrupt the adversarial perturbations generated for the neighboring objects/instances
2. **Explanations:** Perturbations overlap with the neighboring object Effective Receptive Field
3. **Method:** We propose the Localized Instance Perturbation (LIP) that confines the perturbation inside the Effective Receptive Field of a target.

Results



The detection results by the HR are shown in original and perturbed images. (Yellow: true positives; Red: false positives)

Evaluation on Synthetic Images:



Evaluation on Face Detection Datasets:

Perturbations	Sets	None	I-FGSM IMP	I-FGSM LP	I-FGSM LIP-A	I-FGSM LIP-H
Detection Rate (%)	Easy	92.4	46.2	30.1	28.2	**26.5**
	Medium	90.7	50.7	34.7	32.2	**31.1**
	Hard	77.3	45.9	29.3	**23.6**	26.6
Attack Success Rate (%)	Easy	-	50.0	67.4	69.5	**71.3**
	Medium	-	44.1	61.7	64.5	**65.7**
	Hard	-	40.6	62.1	**69.5**	65.6

Evaluation on Object Detection Datasets:

Perturbations	IMP	LP
Average Recall	7.9	**2.2**
Average Precision	6.9	**1.9**