변조 기반의 시각적 해석과 객체 왜곡 방지 기술

김준호, 김성엽, 노용만

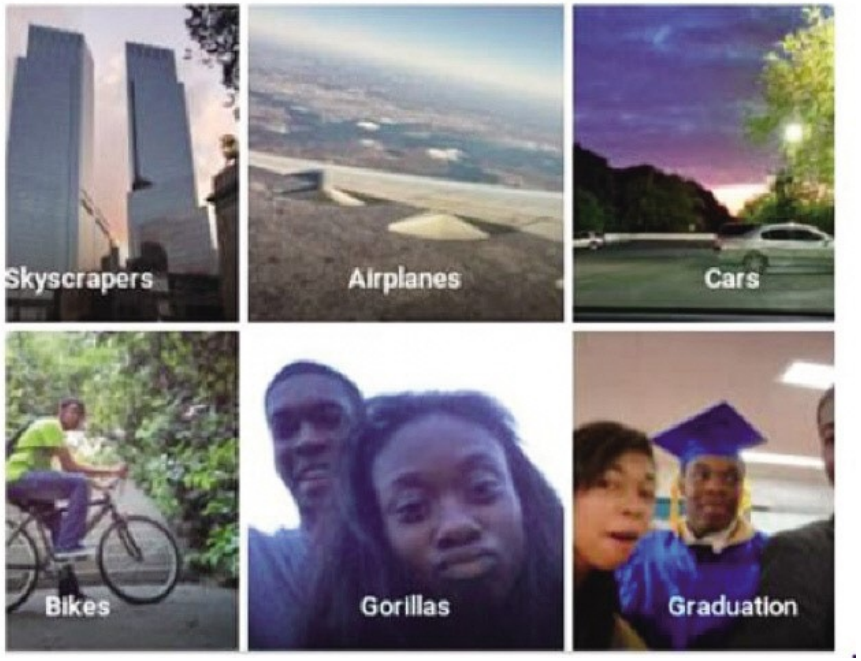
노용만 교수님 연구실

arkimjh@kaist.ac.kr

Image and Video Systems Lab.
School of Electrical Engineering
KAIST

Understanding Deep Neural Networks

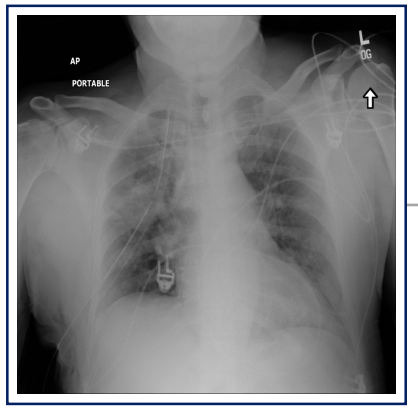
- Application of Deep Learning



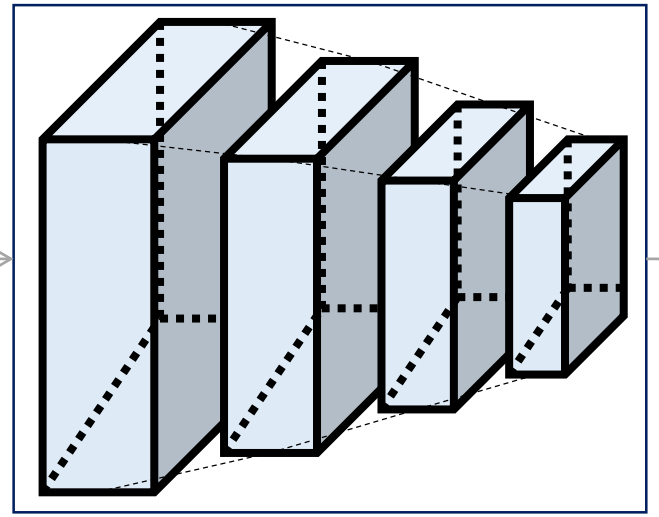
Explainable tools are crucial for high-impact / high-risk applications of Deep Learning

Understanding Deep Neural Networks

- Application of Deep Learning



Chest X-ray Image



CNN Architecture

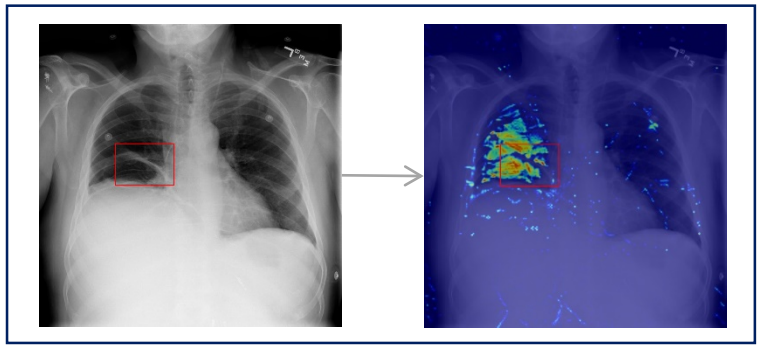
Atelectasis: 92.87%
Nodule: 2.43%
Effusion: 7.52%
Infiltration: 12.3%
⋮

Diagnosis Prediction

How to "explain" the network prediction to Atelectasis?

→ Find the region where network is looking at. (Visual Explanation)

Visual Explanation

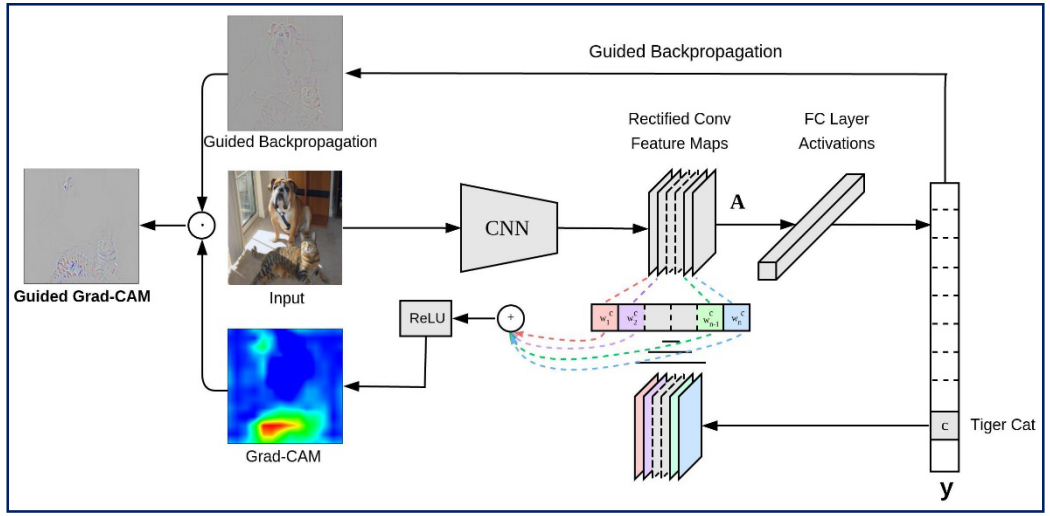


Attribution Map for Atelectasis

Visual Explanation

- Visualizing Attribution Maps

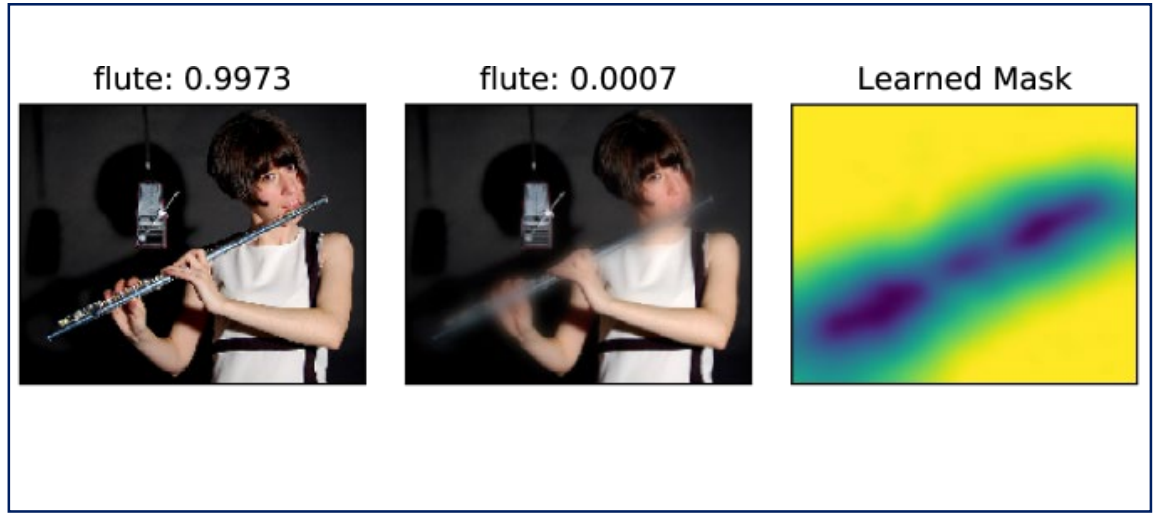
Gradient-based Methods



- Pros: Simple and Fast.

- Cons: Need tractable internal components (e.g., gradient, activation), Network architecture dependency.

Perturbation-based Methods



- Pros: Model-agnostic property, Interpretability for the black-box models.

- Cons: difficult to optimize.

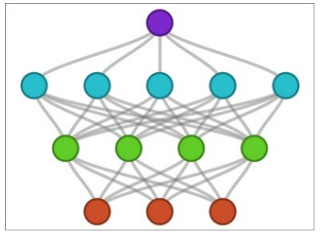
[Selvaraju et al. 2017; Fong et al. 2017]

Perturbation-based Methods

- Perturbed Mask Optimization

Goal: Optimizing a mask that changes network response for the target object to the maximum.

Input $\boldsymbol{x}$
Target c: Flute



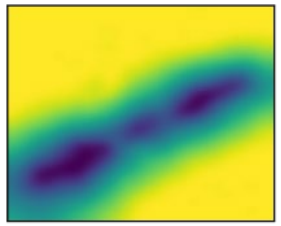
Pretrained Network f

Step 1: Perturb input image with a mask

$$\boldsymbol{x}_{pt} = \boldsymbol{m} \otimes \boldsymbol{x} + (\boldsymbol{1} - \boldsymbol{m}) \otimes \boldsymbol{x}_r$$

Step 2: Find optimal mask that change target prediction to maximum.

$$\boldsymbol{m}^* = \underset{\boldsymbol{m}}{\operatorname{argmin}}\, f(\boldsymbol{x}_{pt}; c) + R(\boldsymbol{x}_{pt})$$



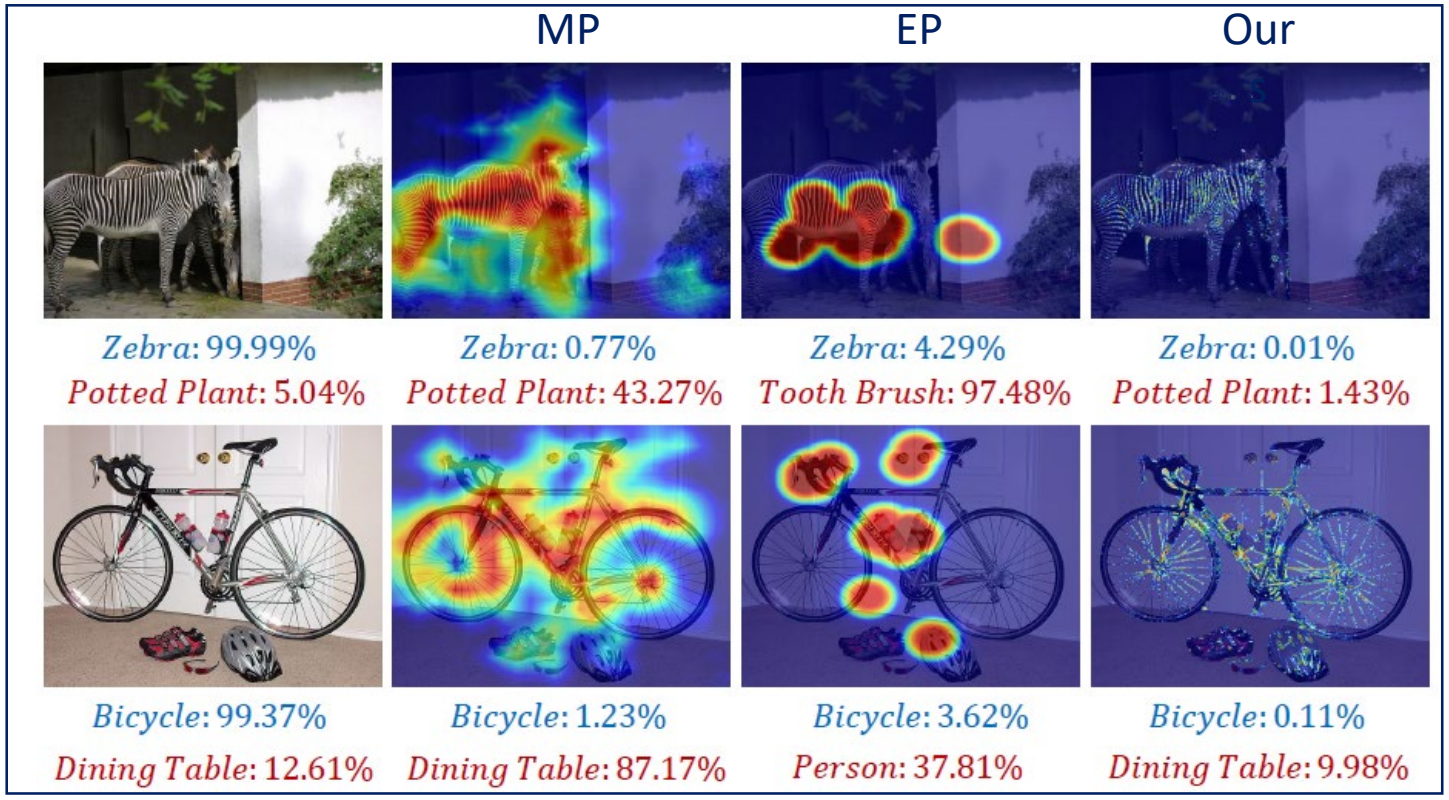
: Learned Mask

Flute: 0.9973 → 0.0007

[Fong et al. 2017]

Class Distortion Problem

- Unexpected Changes of Prediction

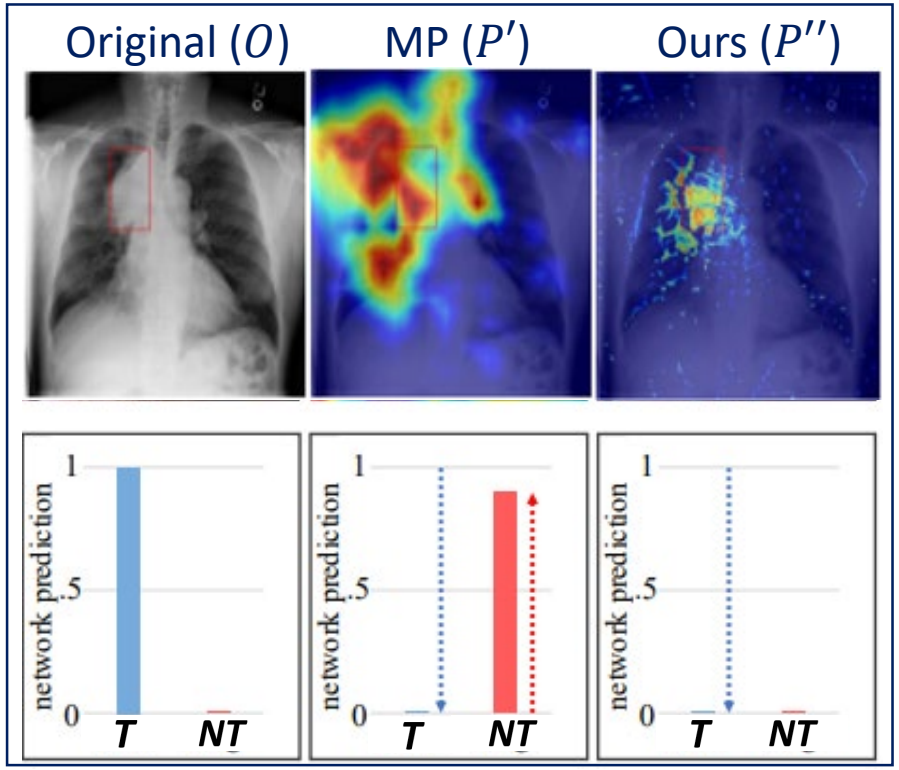


Difficult to understand the relationship between the masked image and the outputs of the network prediction.

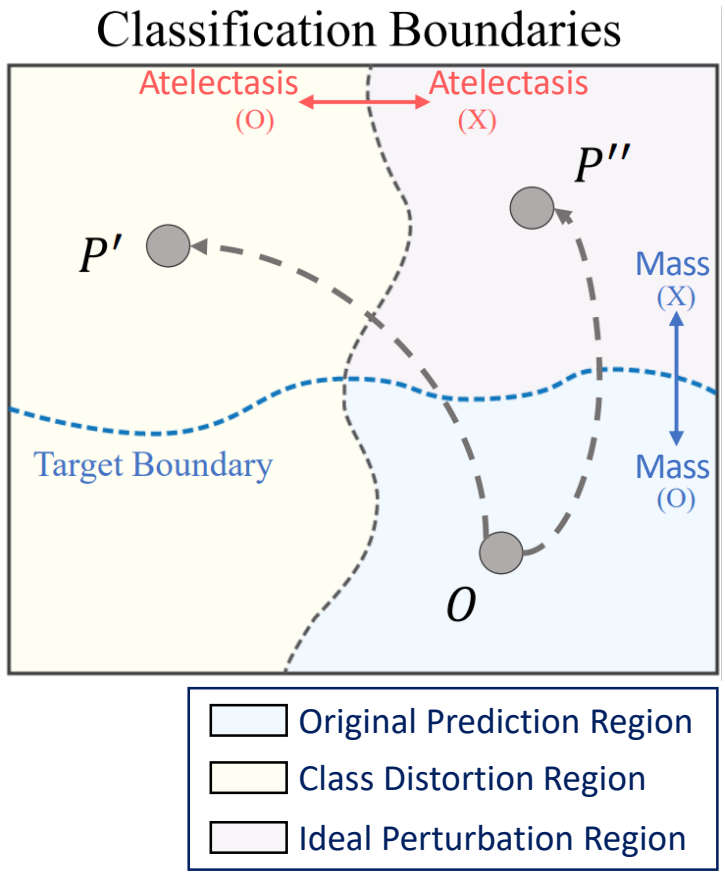
[MP: Fong et al. 2017; EP: Fong et al. 2019]

Class Distortion Problem

- Analysis on Classification Boundaries



T: Atelectasis / ***NT***: Mass

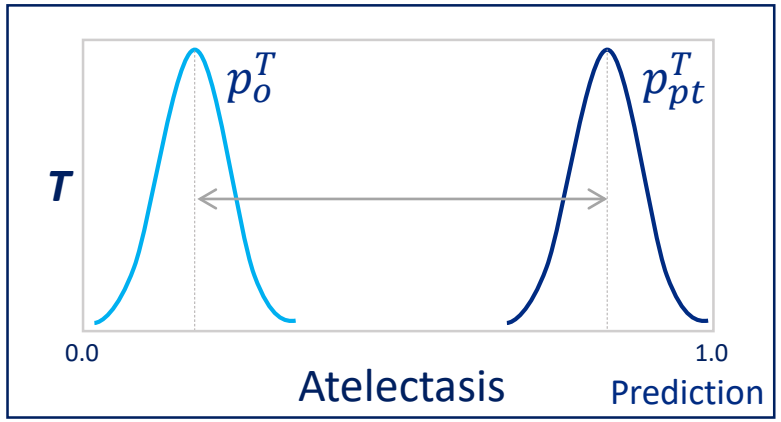


Only considering target class prediction of perturbation, may cause Class Distortion for the non-target classes that can be unintentionally changed.

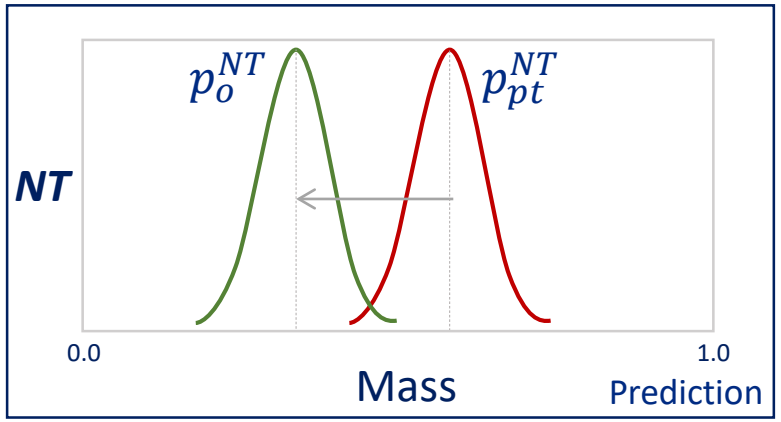
[MP: Fong et al. 2017]

Our Approach: Robust Perturbation

- Robust Mask Perturbation



p_o^T: ***T*** Prediction for $\boldsymbol{x}$ / p_{pt}^T: ***T*** Prediction for $\boldsymbol{x_{pt}}$



p_o^{NT}: ***NT*** Prediction for $\boldsymbol{x}$ / p_{pt}^{NT}: ***NT*** Prediction for $\boldsymbol{x_{pt}}$

Find optimal mask $\boldsymbol{m}^*$:

$$\mathcal{L}_M = -\mathcal{D}_T\left(\boldsymbol{P_o^T} \,\middle\|\, \boldsymbol{P_{pt}^T}\right) + \mathbb{E}_{\bar{c}}\left[\mathcal{D}_{NT}\left(\boldsymbol{P_o^{NT}} \,\middle\|\, \boldsymbol{P_{pt}^{NT}}\right)\right]$$

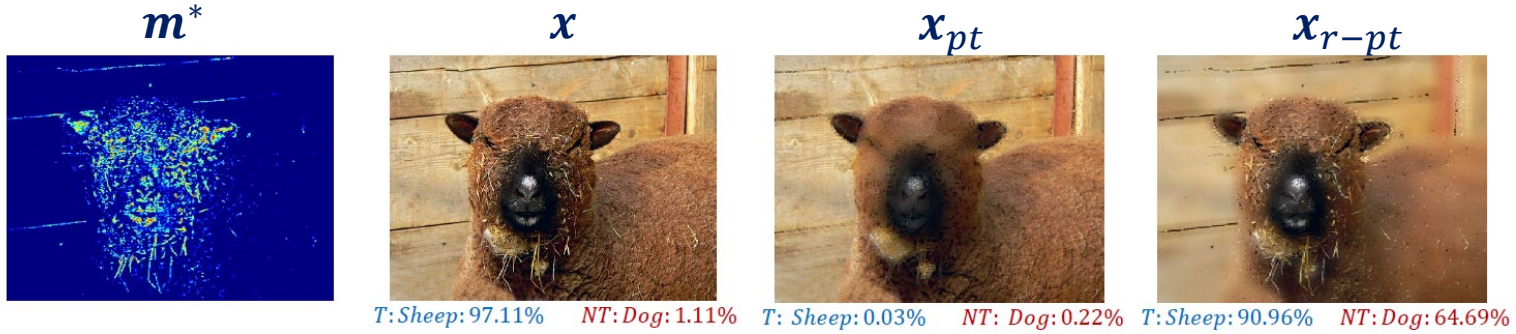
$$\boldsymbol{m}^* = \underset{\boldsymbol{m}}{\operatorname{argmin}}\, \mathcal{L}_M + |\mathbf{1} - \boldsymbol{m}|_1$$

$\mathcal{L}_M$ leads to generate an optimal that changes the target prediction only, while maintaining the non-target predictions.

Our Approach: Robust Perturbation

- Reversed Mask Perturbation

Question: “Can the network revert to original prediction of the target class with the reversely applied mask?” – Another Distortion induced.



$\boldsymbol{x}$: Original Image / $\boldsymbol{m}^*$: Optimal Mask

$$\boldsymbol{x}_{pt} = \boldsymbol{m} \otimes \boldsymbol{x} + (\mathbf{1} - \boldsymbol{m}) \otimes \boldsymbol{x}_r$$

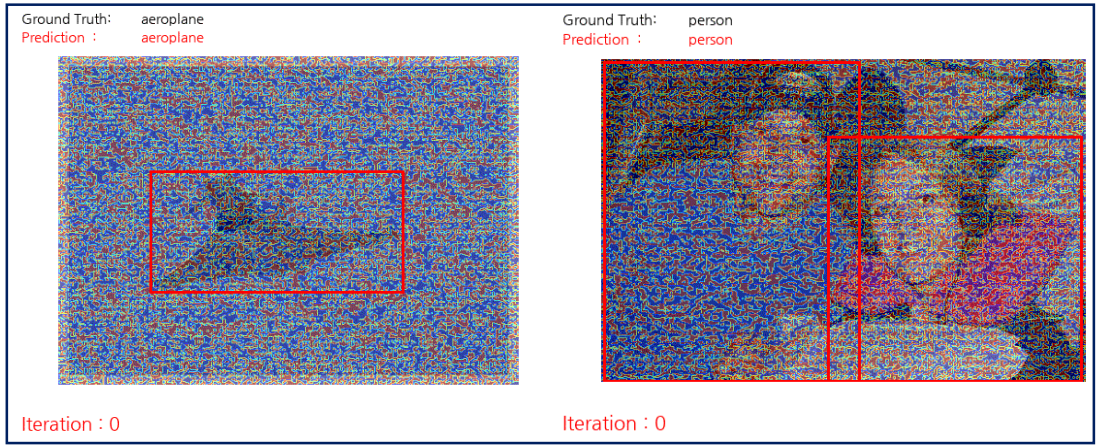
$$\boldsymbol{x}_{r-pt} = (\mathbf{1} - \boldsymbol{m}) \otimes \boldsymbol{x} + \boldsymbol{m} \otimes \boldsymbol{x}_r$$

Cross-checking mask optimization:

$$\mathcal{L}_{RM} = \mathcal{D}_T\left(p_o^T \,\middle\|\, p_{r-pt}^T\right) + \mathbb{E}_{\bar{c}}\left[f\left(\boldsymbol{x}_{r-pt}; \bar{c}\right)\right]$$

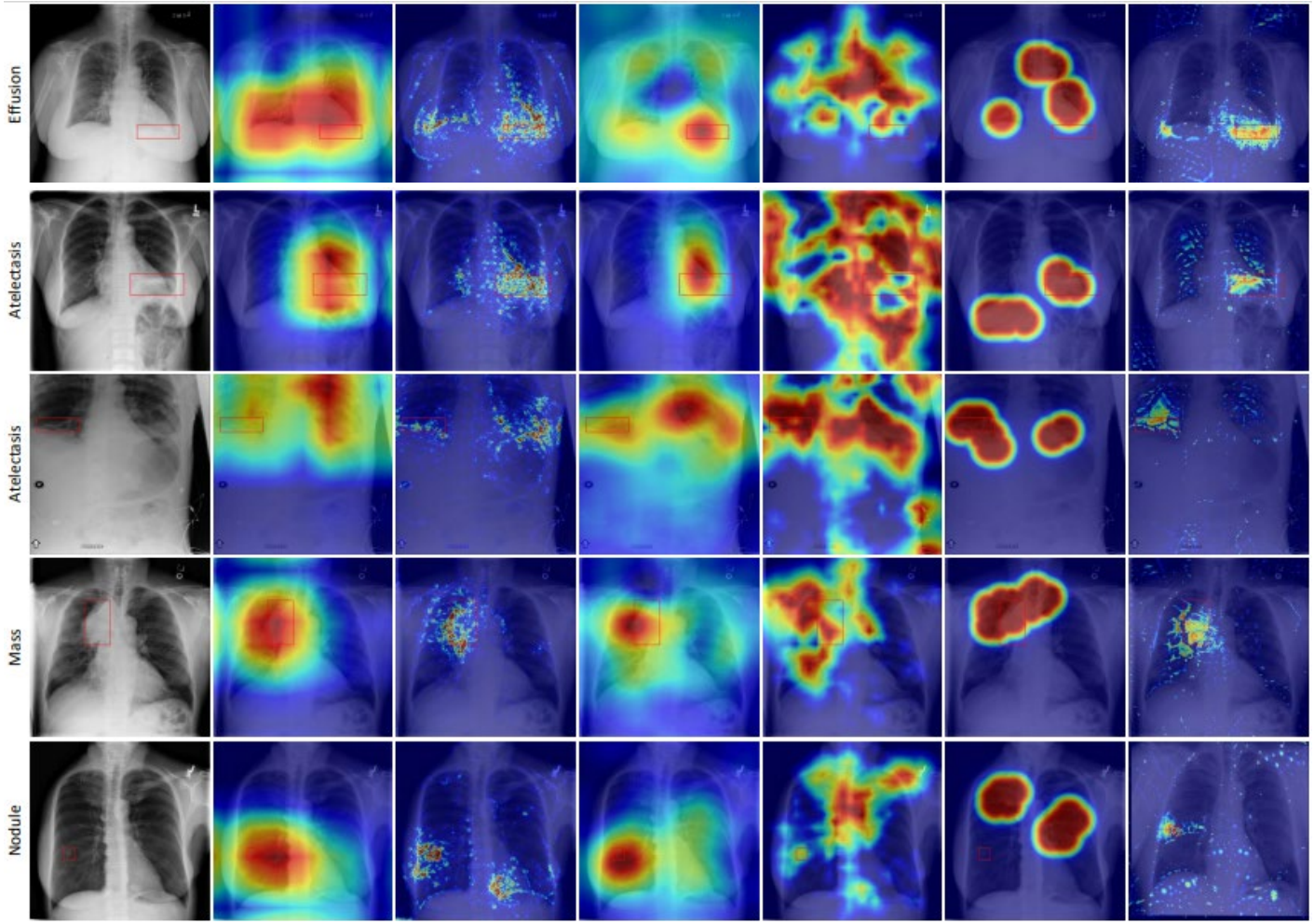
$$\boldsymbol{m}^* = \underset{\boldsymbol{m}}{\operatorname{argmin}}\, \mathcal{L}_M + \mathcal{L}_{RM} + |\mathbf{1} - \boldsymbol{m}|_1$$

Example of Robust Perturbation



Experiments

- Qualitative Results: Medical Image



- Experimental Setting

- Chest X-ray 14 Dataset

14 Thoracic Disease Classes

112,120 X-ray Images

- Baseline Visualization Methods

- Gradient-based Methods

Grad-CAM [Selvaraju et al. 2017]

Gradient Backpropagation [Simonyan et al. 2014]

- Perturbation-based Methods

RISE [Petsiuk et al. 2018]

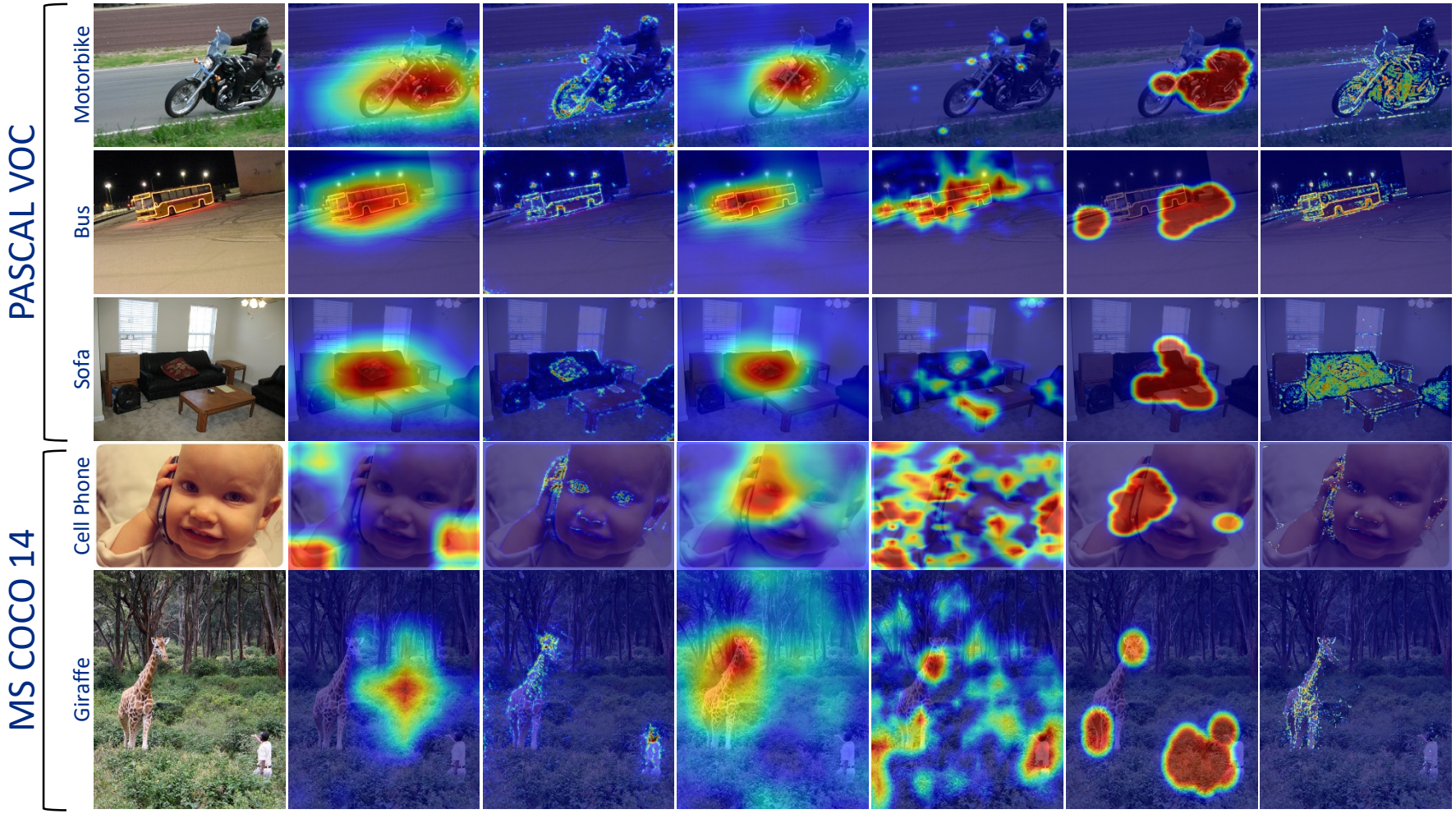
Meaningful Perturbation [Fong et al. 2017]

Extremal Perturbation [Fong et al. 2019]

Robust Perturbation (Ours)

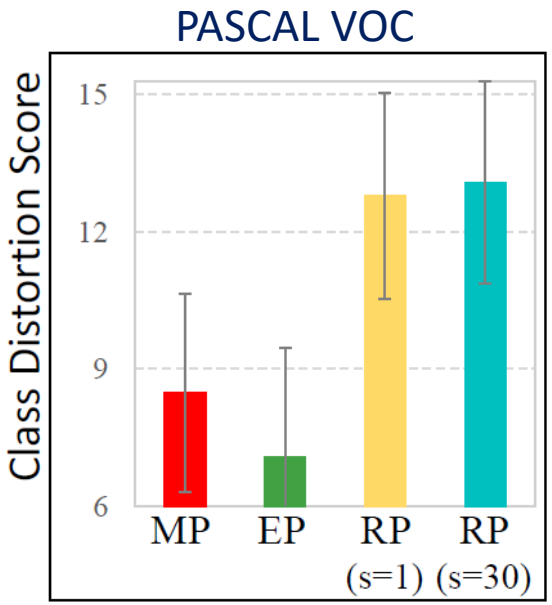
Experiments

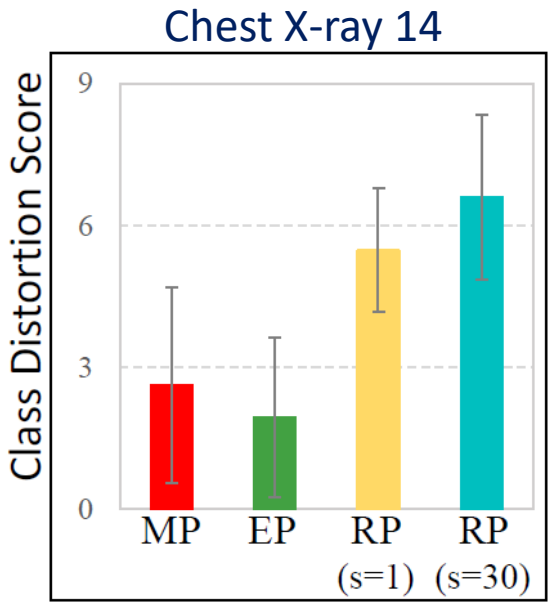
- Qualitative Results: Generic Recognition Datasets

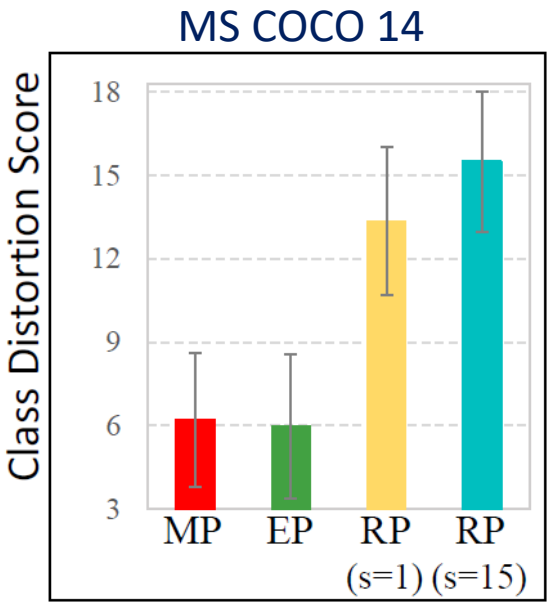


Experiments

- Class Distortion Score







Measuring the occurrence of Class Distortion

Perturbation Confidence

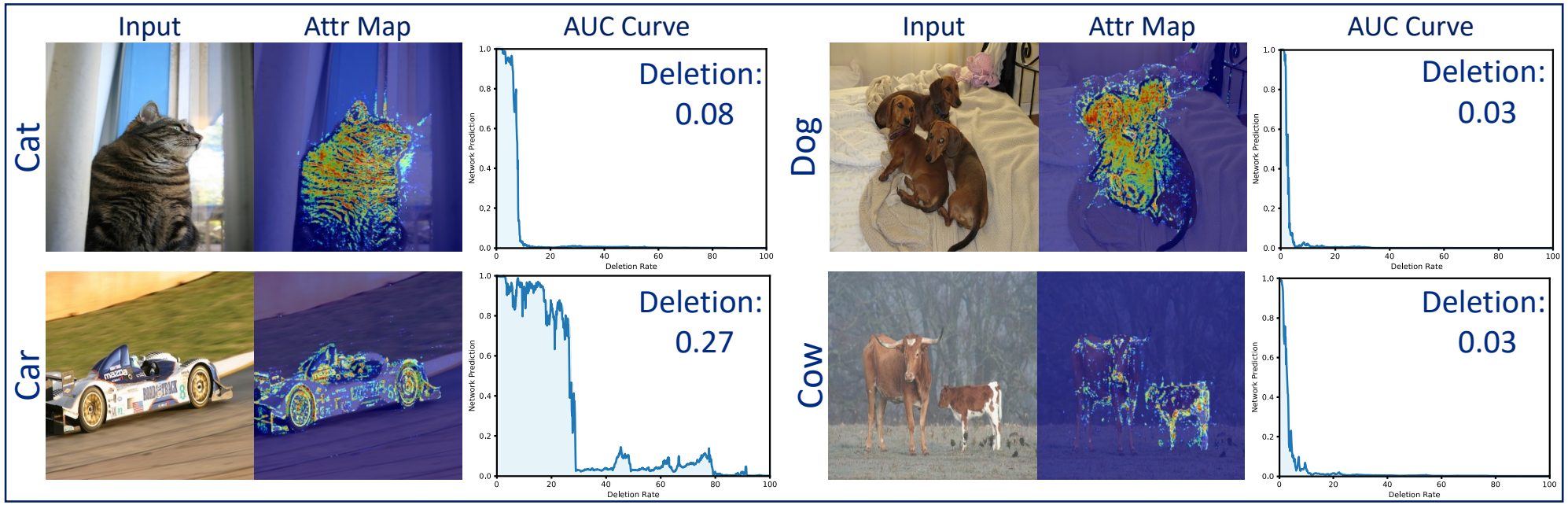
$$CDS = \Delta_{\mathrm{T}}\log-\mathrm{odds} - \mathbb{E}_{\bar{c}}\{\Delta_{\mathrm{NT}}\log-\mathrm{odds}\}$$

Penalization for Class Distortion

Experiments

- Deletion Game

Validating the influence of the highlighted pixels on the network decision.



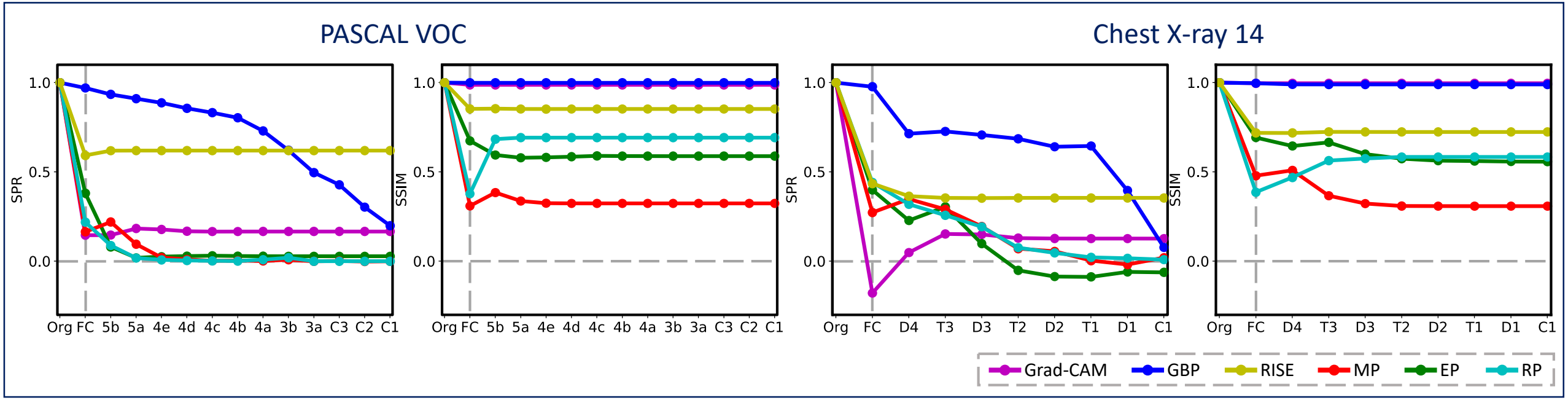
Model	VOC 07	Chest X-ray	COCO 14
EP	0.4329	0.1511	0.5621
RISE	0.3216	0.1502	0.4664
MP	0.2806	0.1389	0.4945
Ours	**0.2068**	**0.1248**	**0.4299**

AUC for Deletion Game
- Lower AUC implies the attribution map in a better quality

[Petsiuk et al. 2018]

Experiments

- Sanity Checks: Model Parameters Randomization



Generated attribution map should be sensitive to changes of the network parameters if the visual methods can properly provide visual explanations for the given network

[Adebayo et al. 2018]

Thank You

김준호, 김성엽, 노용만

노용만 교수님 연구실

arkimjh@kaist.ac.kr

Image and Video Systems Lab.
School of Electrical Engineering
KAIST