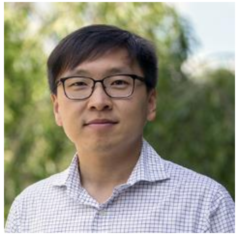
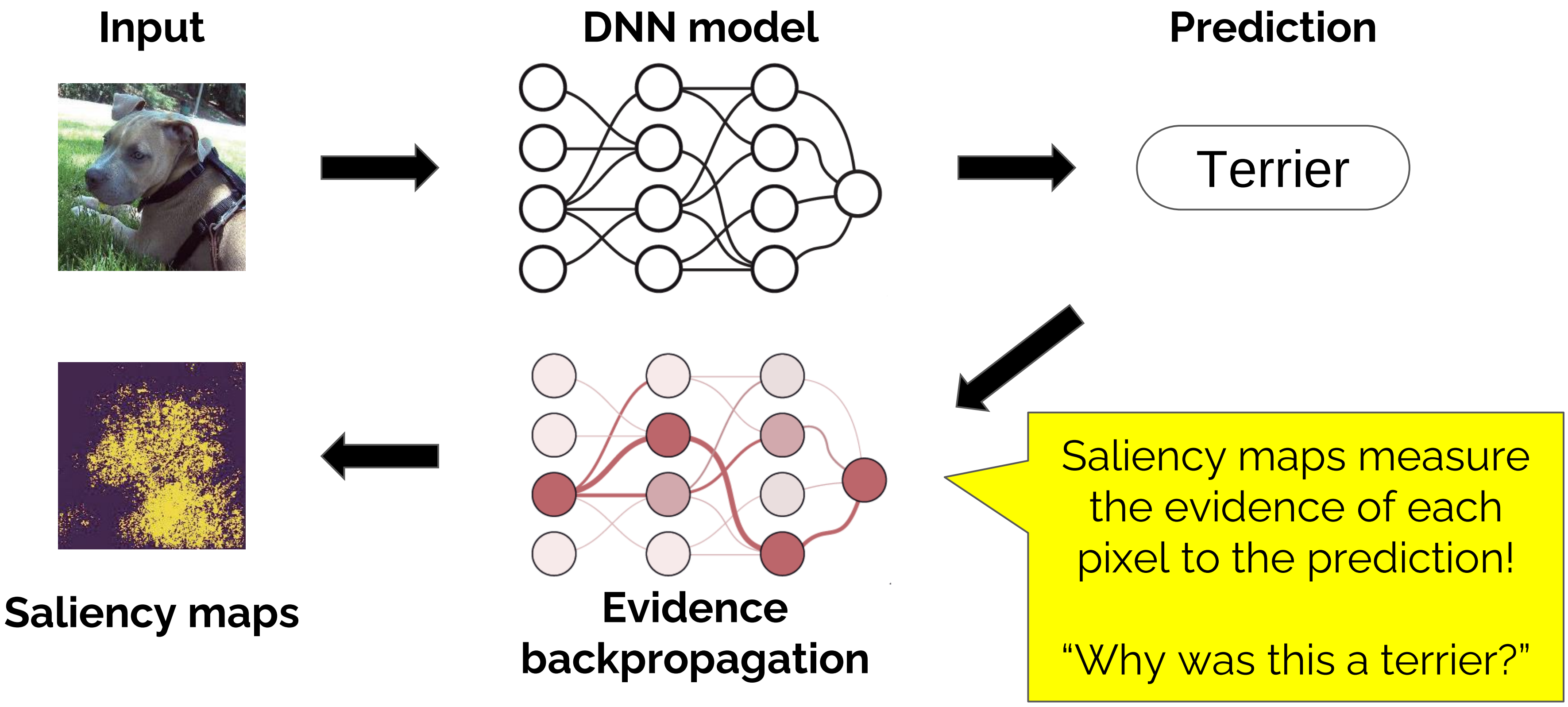
DANCE: Enhancing saliency maps using decoys

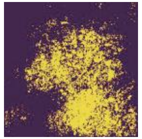
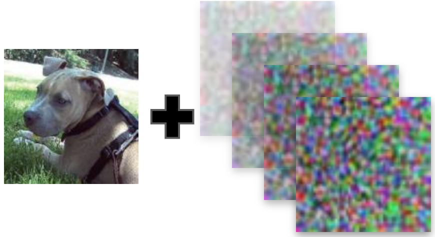
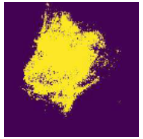
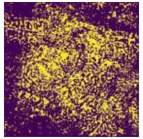
Yang Lu*, Wenbo Guo*, Xinyu Xing, William Stafford Noble

Saliency maps: the most popular interpretability method for deep neural network (DNN) models



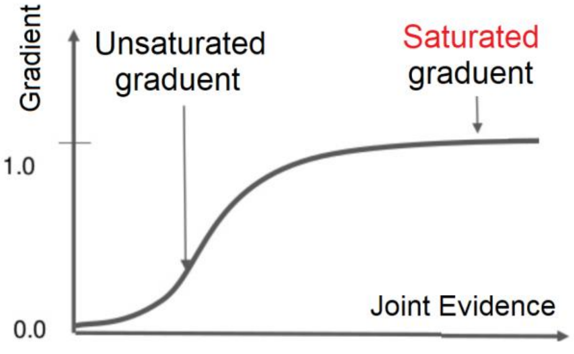
Existing saliency maps can be summarized into a “variations-and-aggregation” paradigm

	Input	**Input variations**	**Saliency aggregation**	**Output**
Vanilla Gradient *(Simonyan et al. 2013)*			NA	
Smooth Gradient *(Smilkov et al. 2017)*			Average	
Integrated Gradient *(Sundararajan et al. 2017)*			Weighted Sum	

Existing saliency maps suffer from following limitations

Gradient saturation

(Shrikumar et al. 2017)



- Strong **joint** evidences together with others.
- Diminishing **marginal** evidence alone.

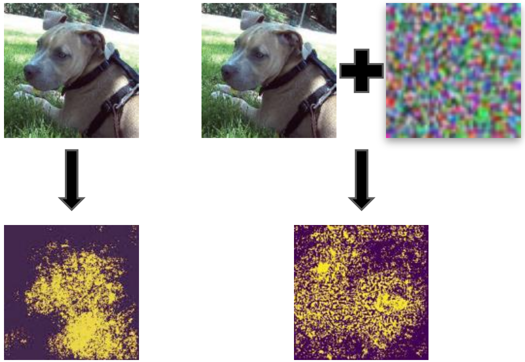
Isolated importance

(Singla et al. 2019)

- The gradient is calculated by **fixing** other features.
- **Smoothness** in input doesn't hold in saliency maps.

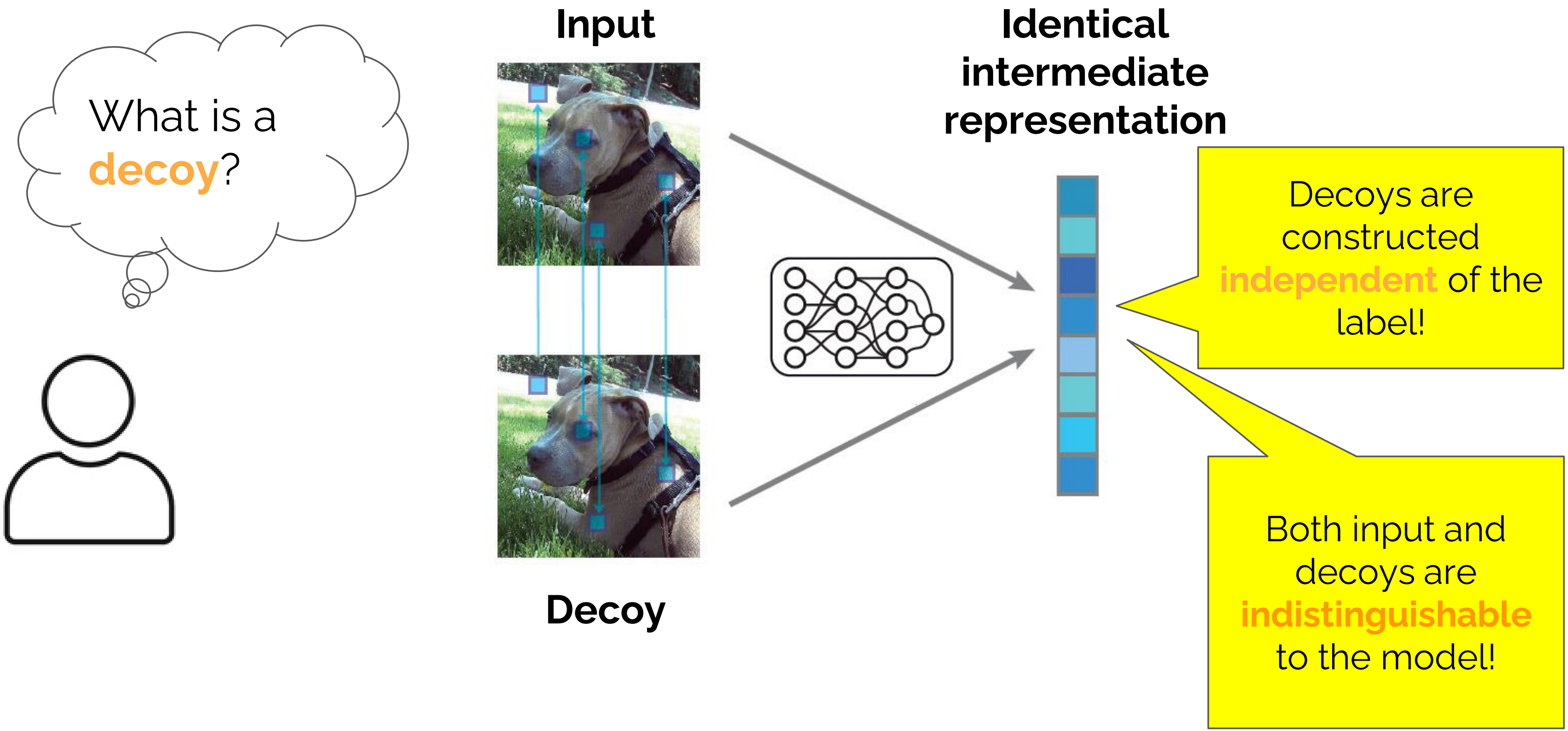
Sensitive to perturbation

(Ghorbani et al. 2017)

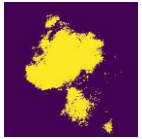


- Even **imperceivable** noises can **drastically** change the saliency maps.

We propose DANCE: decoy-enhanced saliency maps



DANCE can also be decomposed into the “variations-and-aggregation” paradigm

	Input	**Decoys**	**Saliency aggregation**	**Output**
DANCE			Pixel-wise range of decoy saliencies	

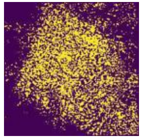
Highlights:

- ❑ Decoys cannot be **out-of-distribution** by design.
- ❑ Decoy can be **constructed efficiently** by optimization.
- ❑ **Theoretical soundness** in mitigating aforementioned limitations

Two different metrics are used to quantitatively evaluate the performance of DANCE

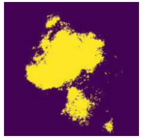
Fidelity

(Dabkowski & Gal, 2017)



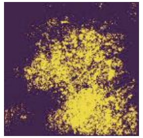
The saliency map is **less coherent** to the prediction

The saliency map is **more coherent** to the prediction



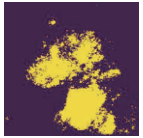
Sensitivity

(Alvarez-Melis & Jaakkola. 2018)

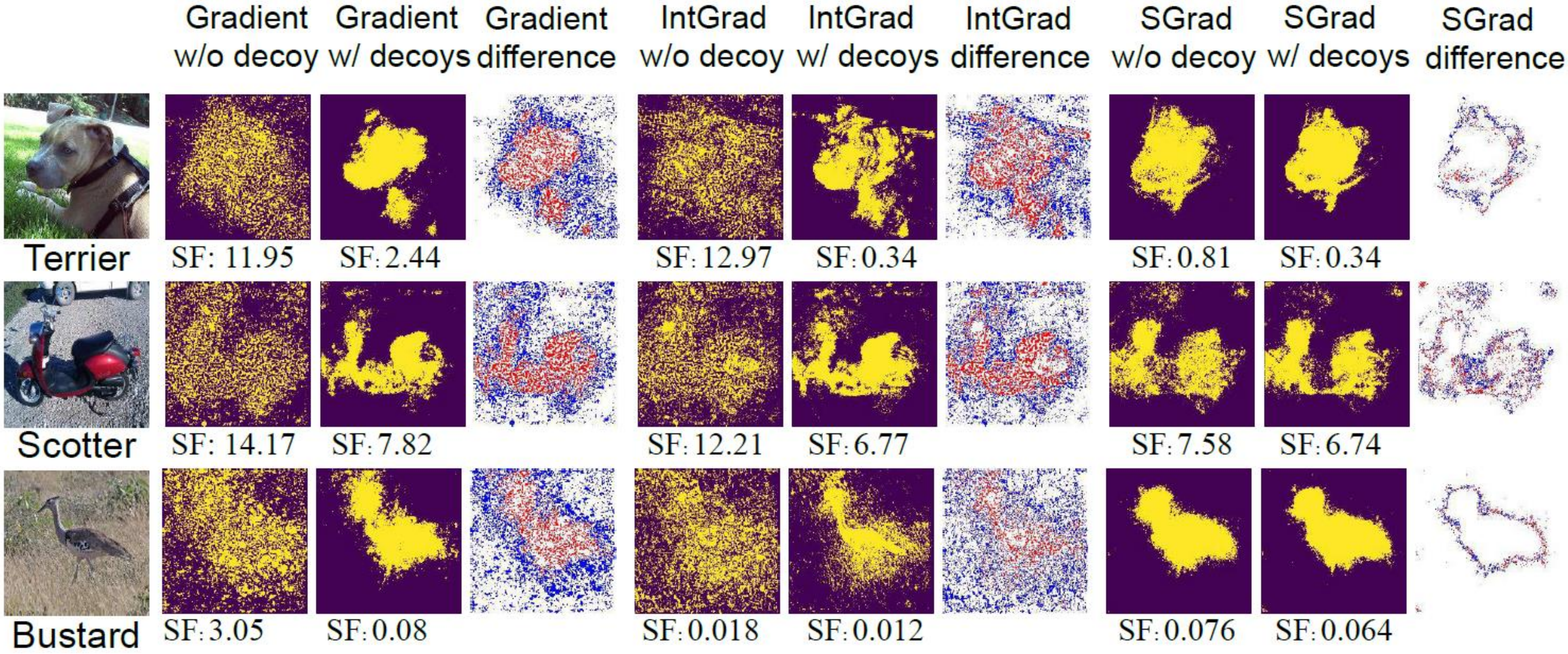


The saliency map is **less robust** to adversarial attack

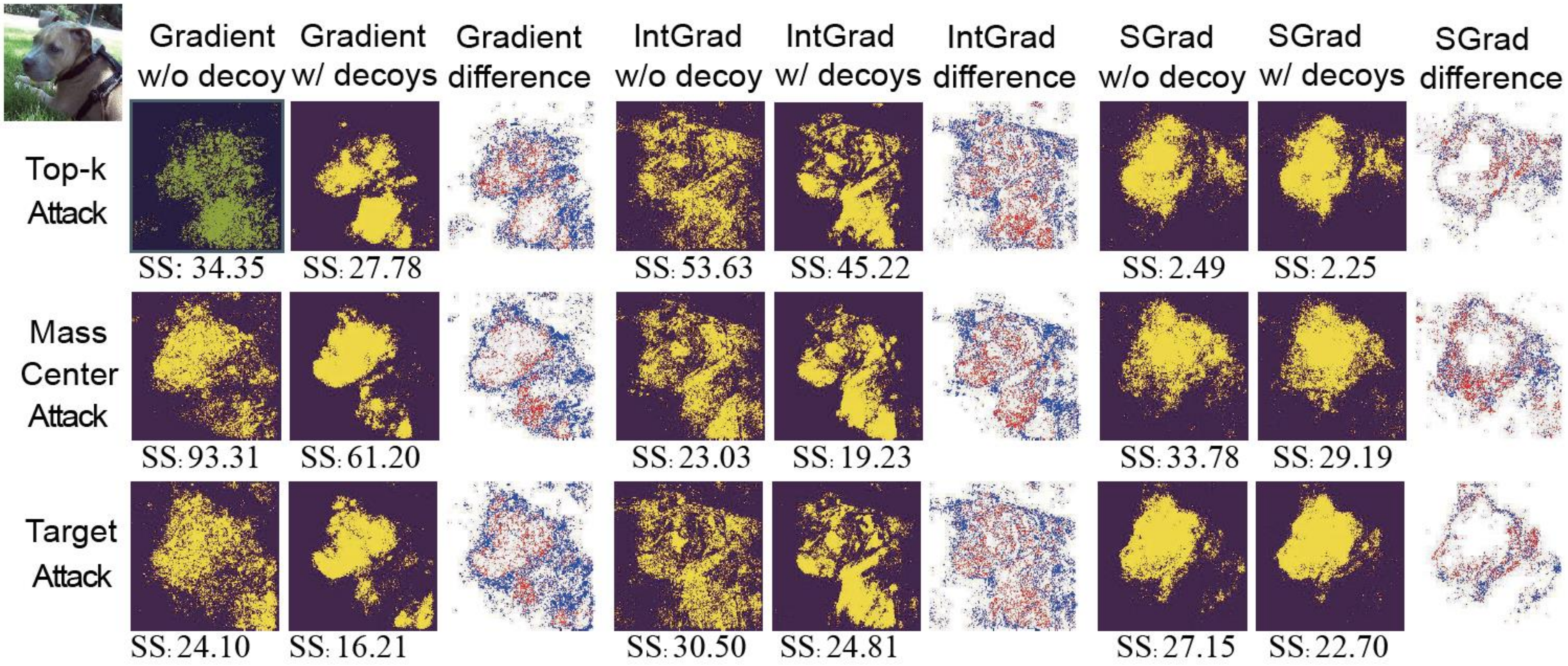
The saliency map is **more robust** to adversarial attack



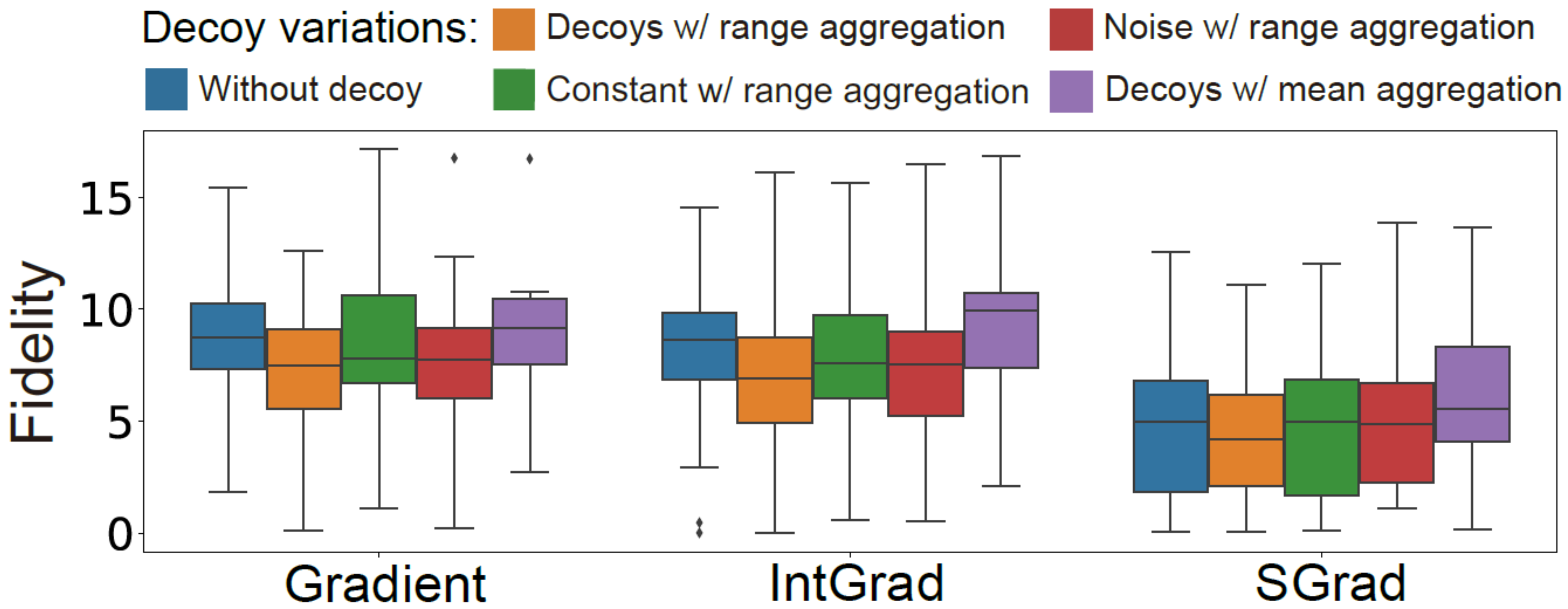
DANCE achieves more coherent saliency maps both qualitatively and quantitatively



DANCE is more robust to various types of adversarial attacks both qualitatively and quantitatively



A control study demonstrates the necessity of both decoy and range-based aggregation steps in DANCE



Conclusions

- Empirically, DANCE performs qualitatively and quantitatively better than existing methods.
- Theoretically, DANCE mitigates three limitations commonly suffered by existing methods: gradient saturation, isolated importance, and sensitivity to perturbation.
- We have demonstrated the wide applicability in a variety of domains.
- Code availability: https://bitbucket.org/noblelab/dance