Poster Location:
West Building Exhibit
Halls ABC 388
WED-PM-388

ADVERSARIAL COUNTERFACTUAL VISUAL EXPLANATIONS

Normandy University, ENSICAEN, UNICAEN, CNRS, GREYC, France

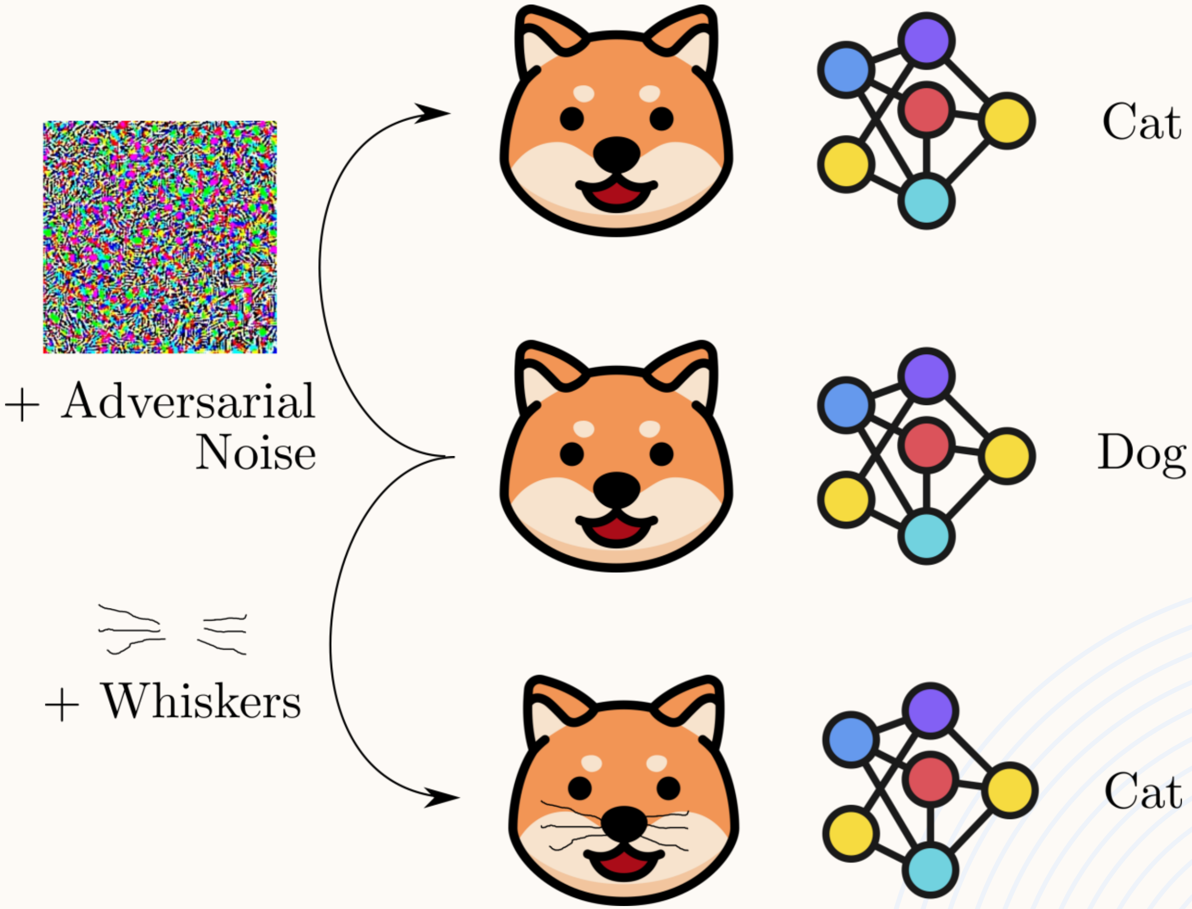
Guillaume Jeanneret Loïc Simon Frédéric Jurie

INTRODUCTION

- In this work, we study **counterfactual explanations.**
- Counterfactual explanations search to answer "**what does X have to change to alter the prediction from Y to Y'**?"
- The changes consist in human-understandable modifications.

INTRODUCTION

- Adversarial attacks seeks to **fool** the classifier as well.
- Attacks towards traditional classifier are **imperceptible**
- Counterfactual attempt to find **perceptual changes**.
- This property is **undesirable** as the explanations cannot be studied.
- Nonetheless, attacks towards robust models generate **semantic changes**.



INTRODUCTION

- Objectives:
    1. Flip the classifier decision
    2. Sparse and Proximal Changes
    3. Perceptually Realistic Modifications

CONTRIBUTIONS

How may we use **adversarial attacks** to generate **semantic changes** for any model, **regardless of its robustification**?

(1) METHODOLOGY

We propose ACE, a method based on Adversarial Attacks to generate Counterfactual Explanations, regardless of its robustness towards them.

(2) PERFORMANCE

ACE performs competitively with respect to the prior State-of-the-Art, surpassing it in multiple measurement metrics in various datasets.

(3) EXTENSION

We point out some defects of current evaluation protocols and amend them. As well as, extend them toward a general setup.

(4) ACTIONABILITY

By analyzing the counterfactual explanations generated by ACE, we were able to produce actionable changes in order to flip the prediction of the classifier.

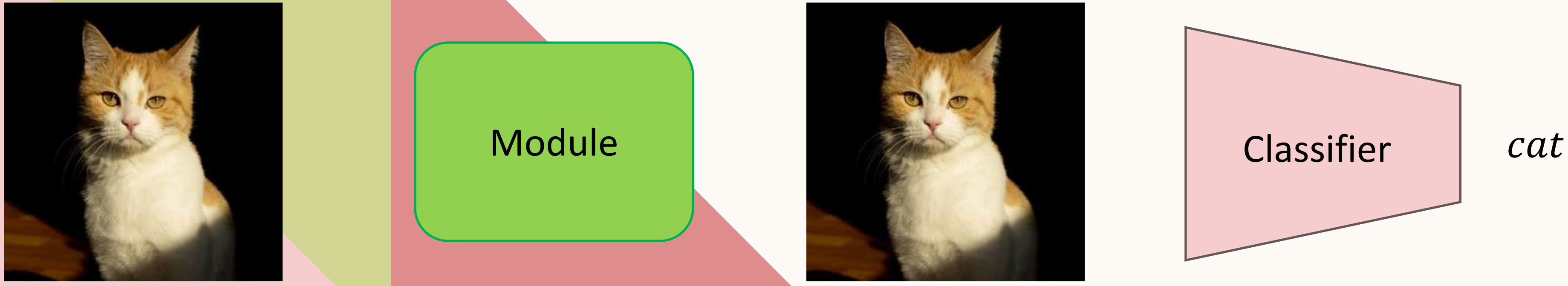
METHODOLOGY

Adversarial Counterfactual Visual Explanations

WHAT PROPERTIES WE WANT TO FIND

- We search to add a module to robustify the input model.

- This module should:
  - Filter high frequency signals without changing the original prediction.
  - The final product must lie in the image manifold.



METHODOLOGY

- Diffusion Models achieve this property!

- Core Idea: Use adversarial attacks, but instead of attacking directly the classifier, we attack the image before its filtering process.

$$E = \underset{x'}{\operatorname{argmax}}\left(L_{class}(x',y') + d(x',x)\right) \text{ transform to } E = \underset{x'}{\operatorname{argmax}}\left(\boxed{L_{class}(F(x'),y')} + d(x',x)\right)$$

ACE ALGORITHM

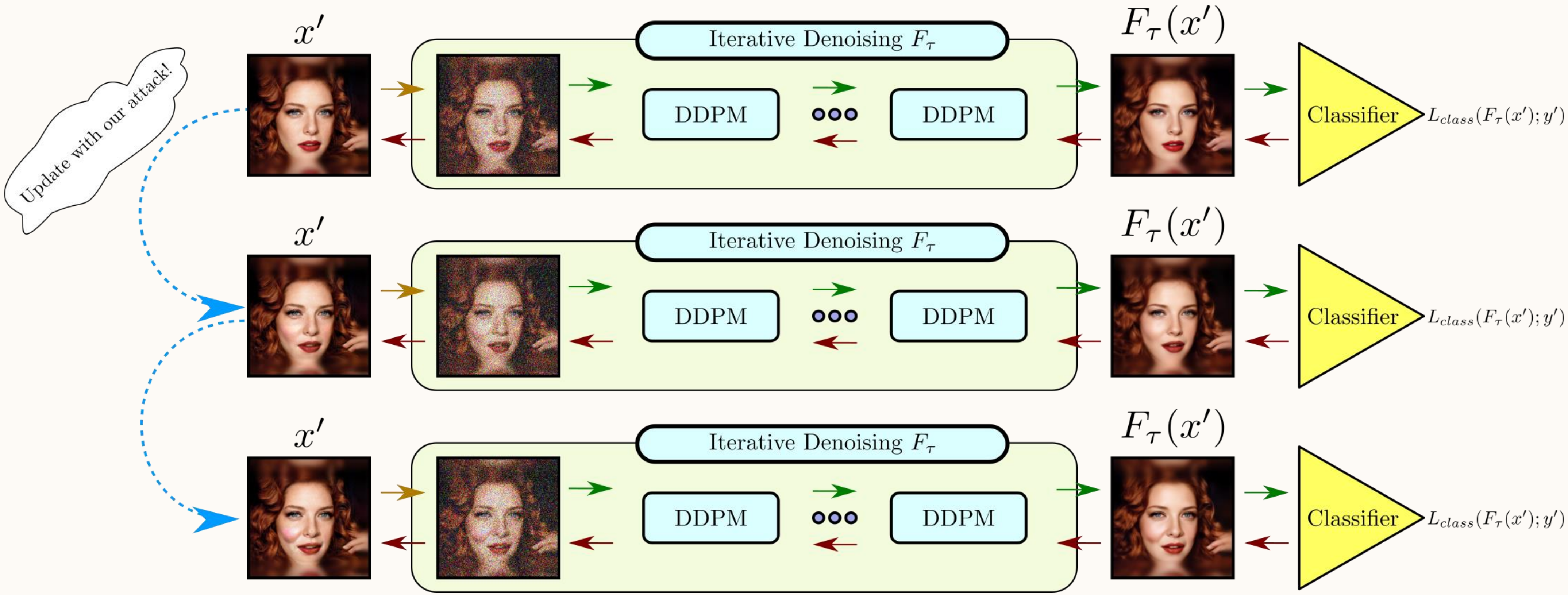
ADD NOISE

DDPM FILTERING

PREDICTION

COMPUTE LOSS

UPDATE IMAGE

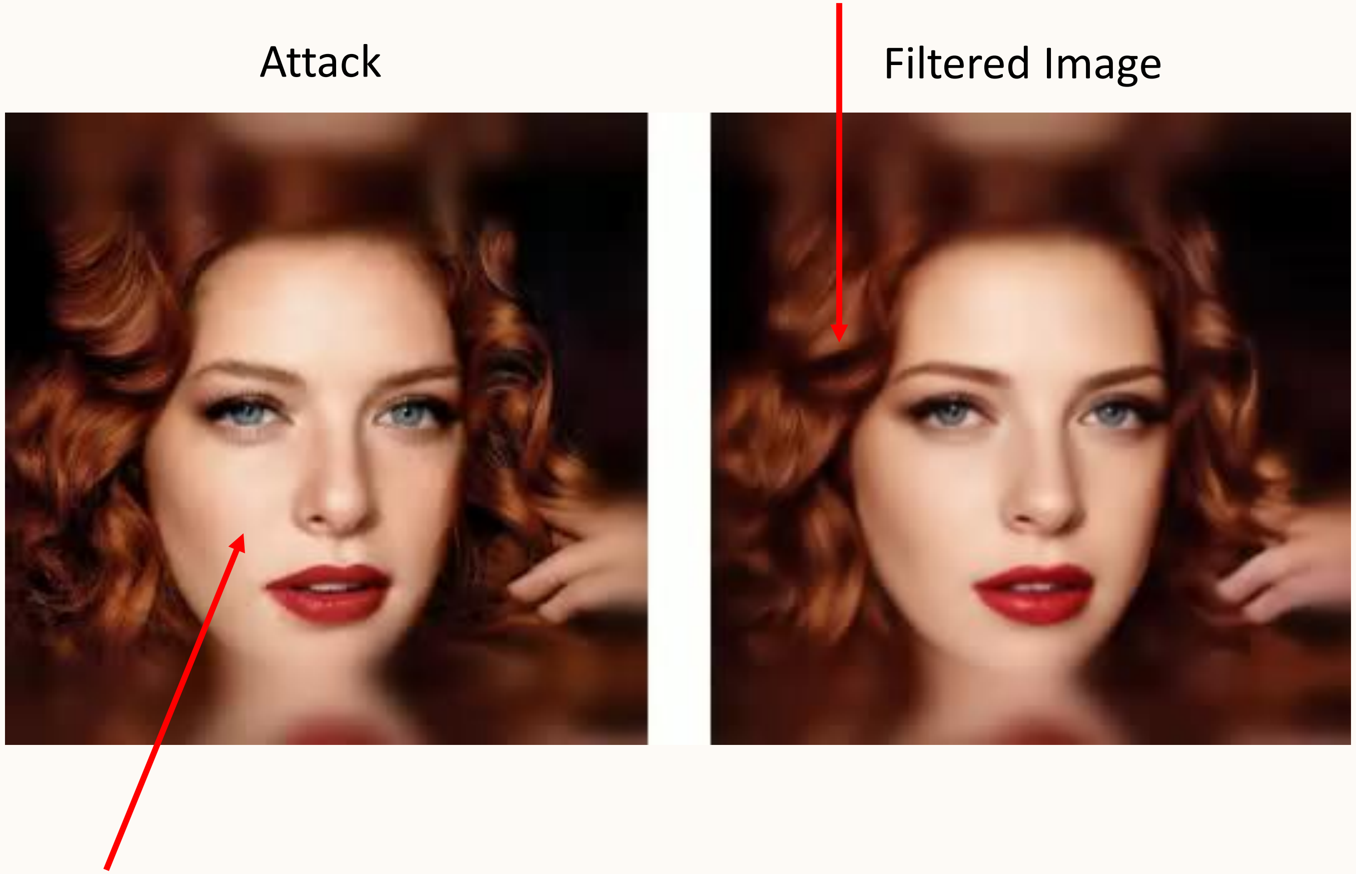

Update with our attack!
x'
Iterative Denoising F_τ
DDPM
DDPM
$F_\tau(x')$
Classifier
$L_{class}(F_\tau(x'); y')$
x'
Iterative Denoising F_τ
DDPM
DDPM
$F_\tau(x')$
Classifier
$L_{class}(F_\tau(x'); y')$
x'
Iterative Denoising F_τ
DDPM
DDPM
$F_\tau(x')$
Classifier
$L_{class}(F_\tau(x'); y')$

EVOLUTION OF THE ATTACK

Filtering modifies all!

Attack

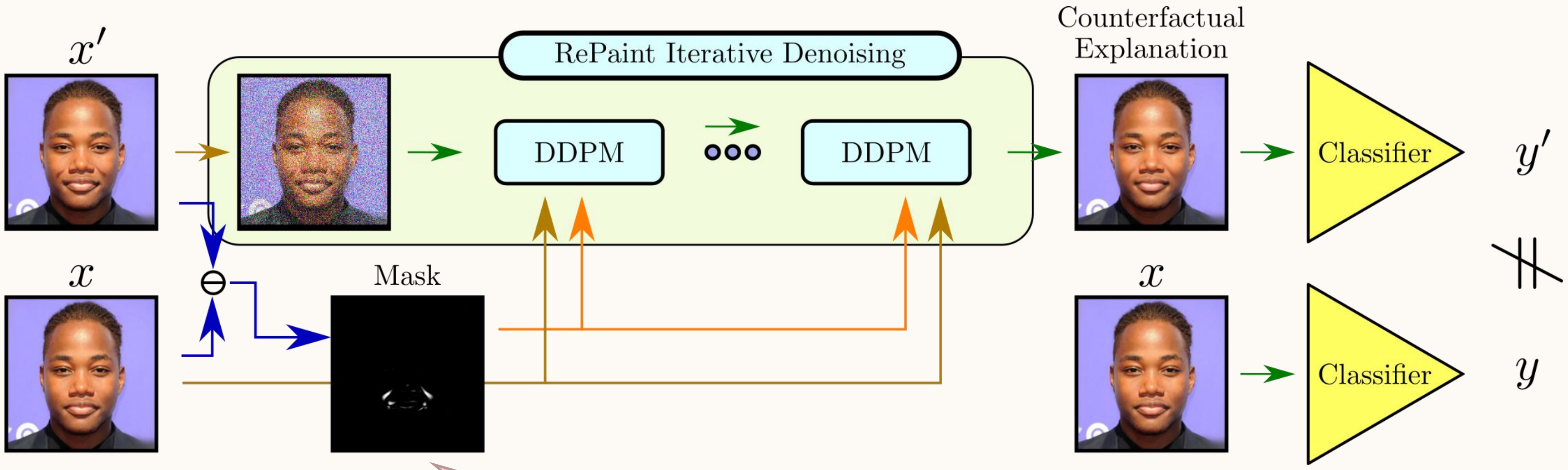
Filtered Image



Only changes this

BRINGING THE EXPLANATIONS CLOSER

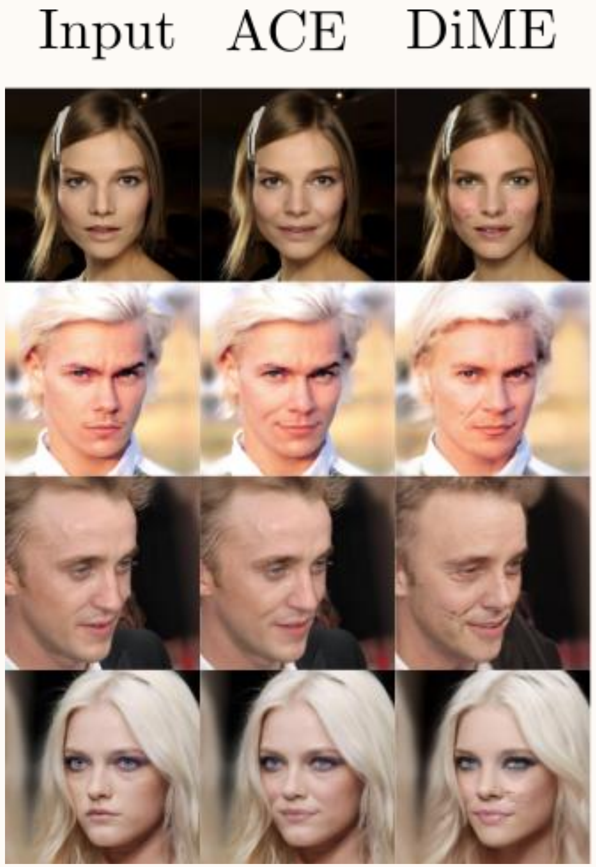
- Adapt Inpainting strategies
- Requiring use input
- We can use the difference between the pre-explanation as the guiding mask!

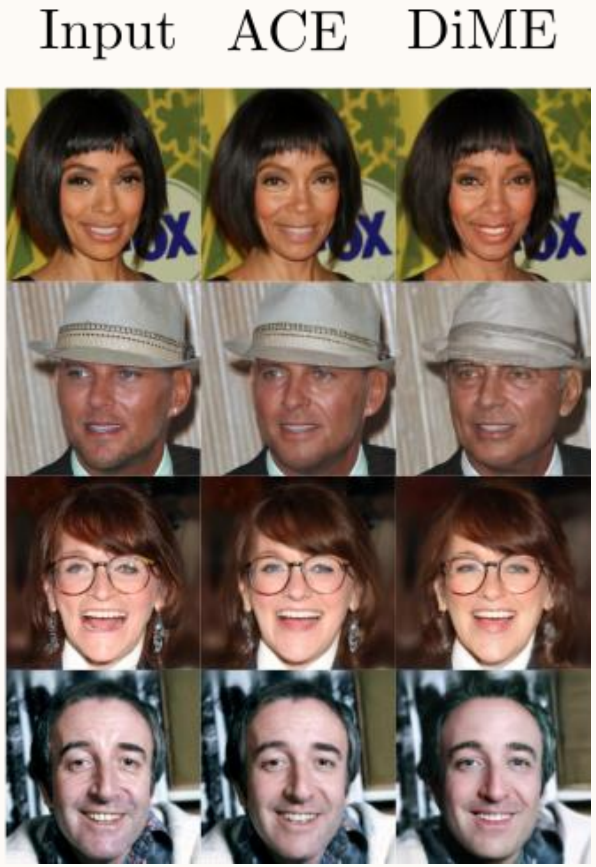


This mask tells use where to add the changes

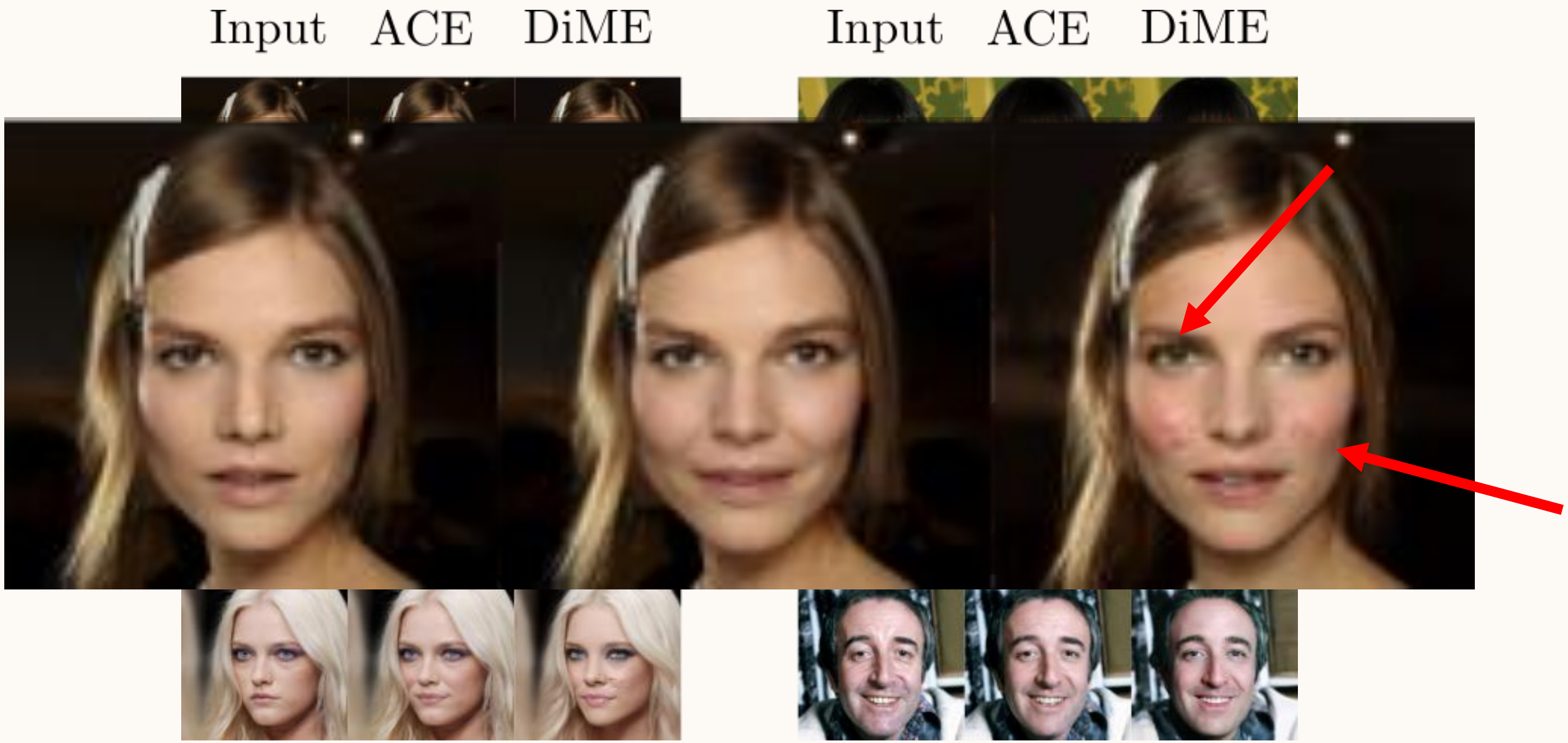
RESULTS

ACE IN CONTRAST TO PREVIOUS SOTA

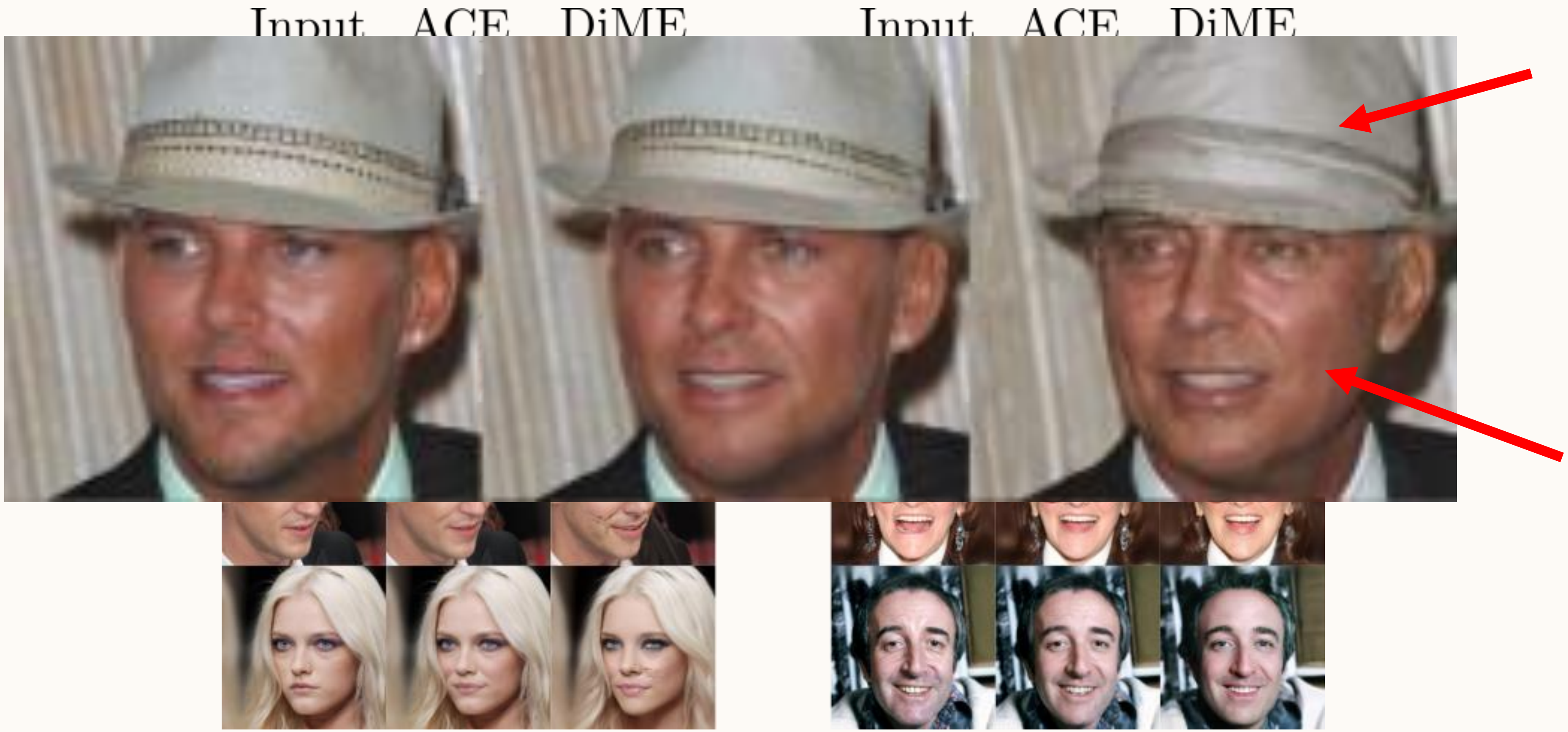




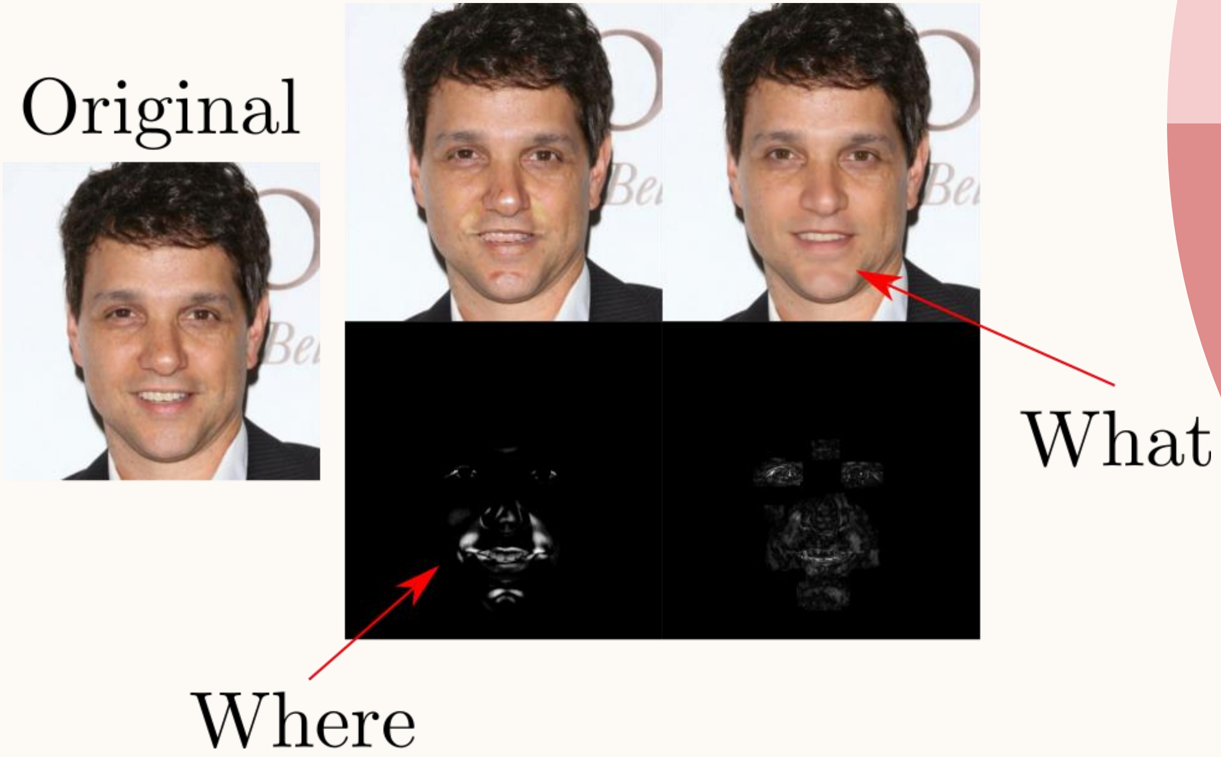
ACE IN CONTRAST TO PREVIOUS SOTA



ACE IN CONTRAST TO PREVIOUS SOTA

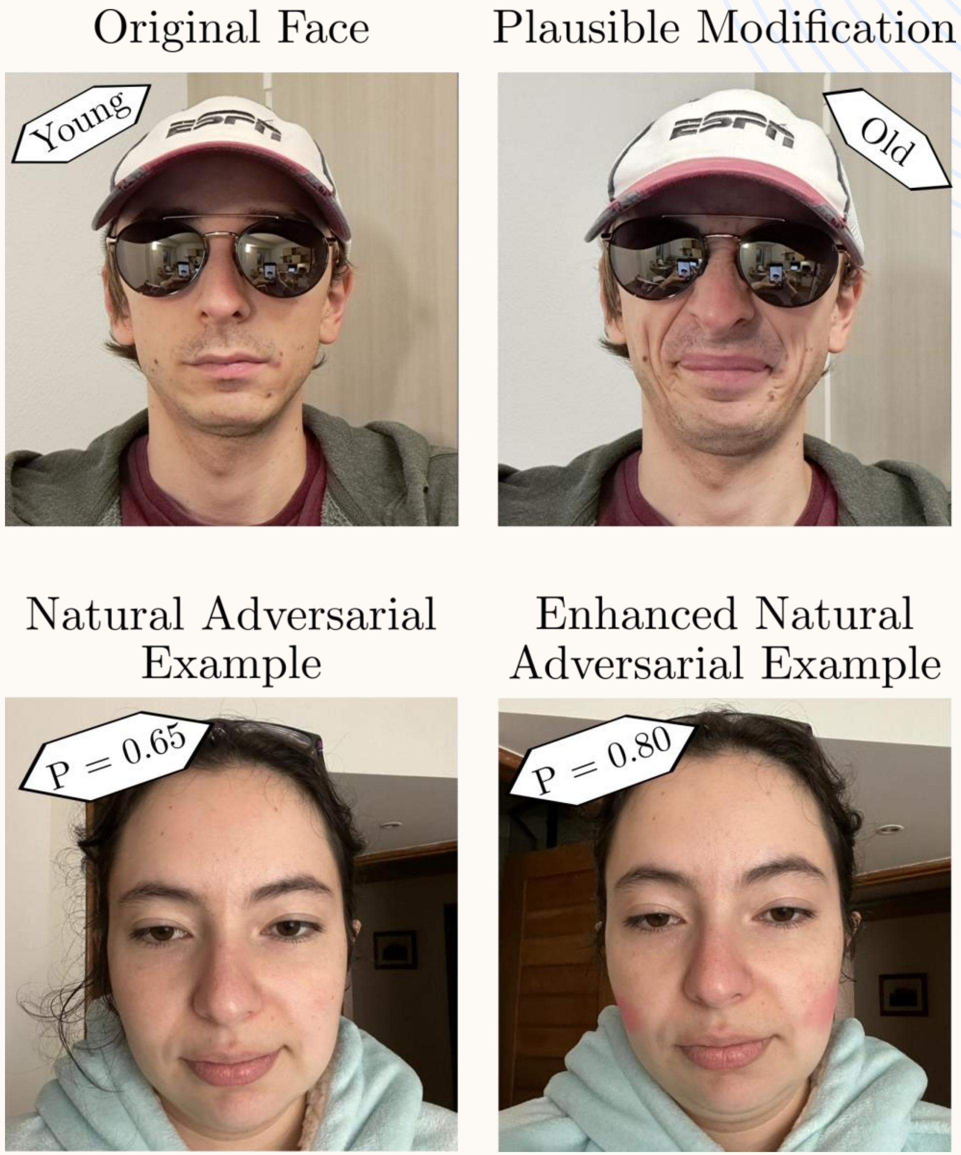


ACE GIVES ADDITIONAL CUES!



FINDING A CLASSIFIER INDUCTIVE BIAS

By studying our counterfactual explanations, we were able to decipher some of the biases learned by the classifiers.


Original Face
Plausible Modification
Young
Old
Natural Adversarial Example
Enhanced Natural Adversarial Example
P = 0.65
P = 0.80

PERFORMANCE AND ADDITIONAL METRICS

- We extended current evaluation metric to assess general counterfactual explanations in images.
- ACE outperforms previous SOTA in multiple metrics.
- Please refer to our paper for more details.

CONCLUSION

- We proposed Adversarial Counterfactual Explanation (ACE) to generate counterfactual explanations via adversarial attacks.
- ACE provides an additional mask to localize the changes.
- With our proposed method, we were able to discover one of the most recurrent biases for classifiers. With our findings, we were able to make actionable changes in real life to change the output of the classifier.

THANK YOU VERY MUCH!

If you have any question, please visit us at West Building Exhibit Halls ABC 388.

Webpage

Github