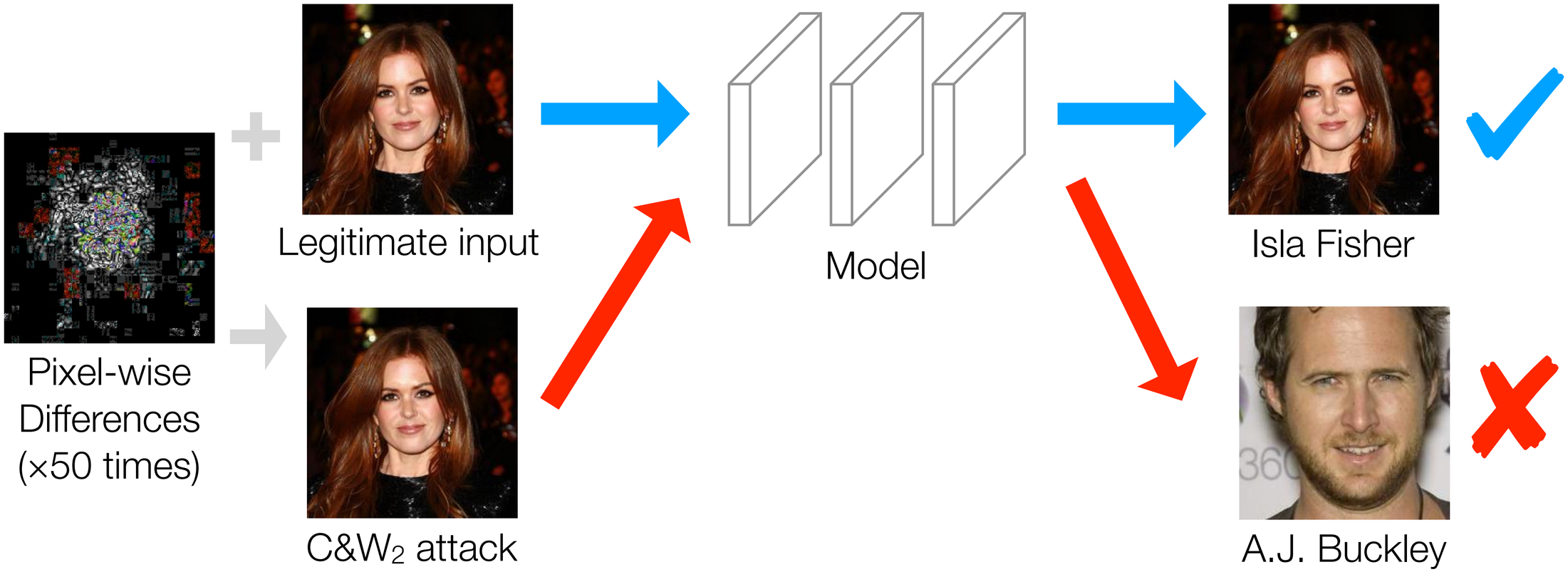
Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples

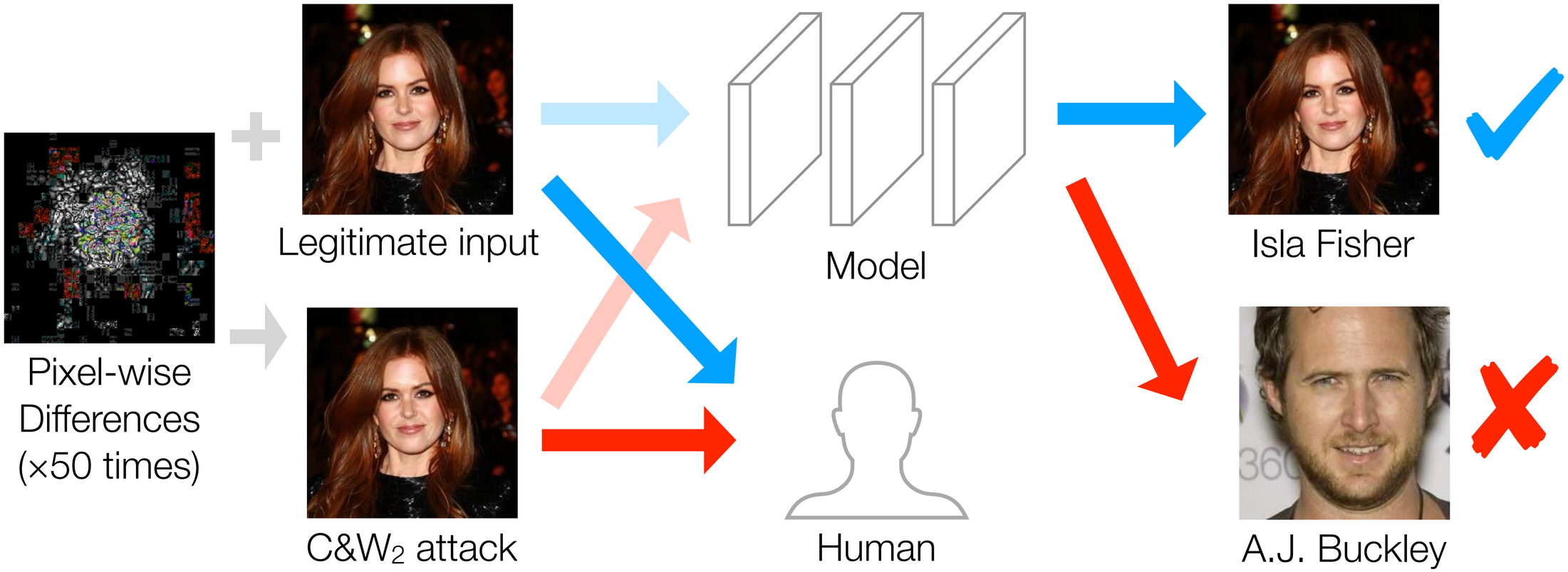
Guanhong Tao, Shiqing Ma, Yingqi Liu, Xiangyu Zhang

PURDUE UNIVERSITY®

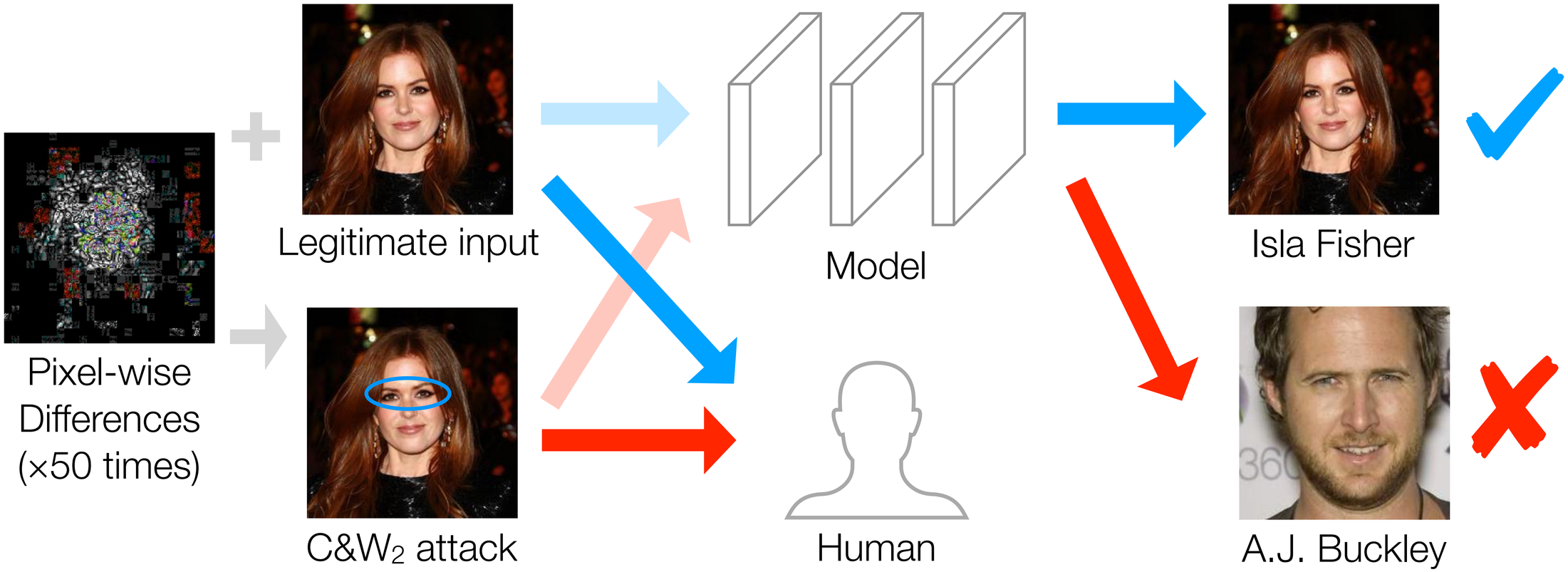
Understanding Adversarial Samples



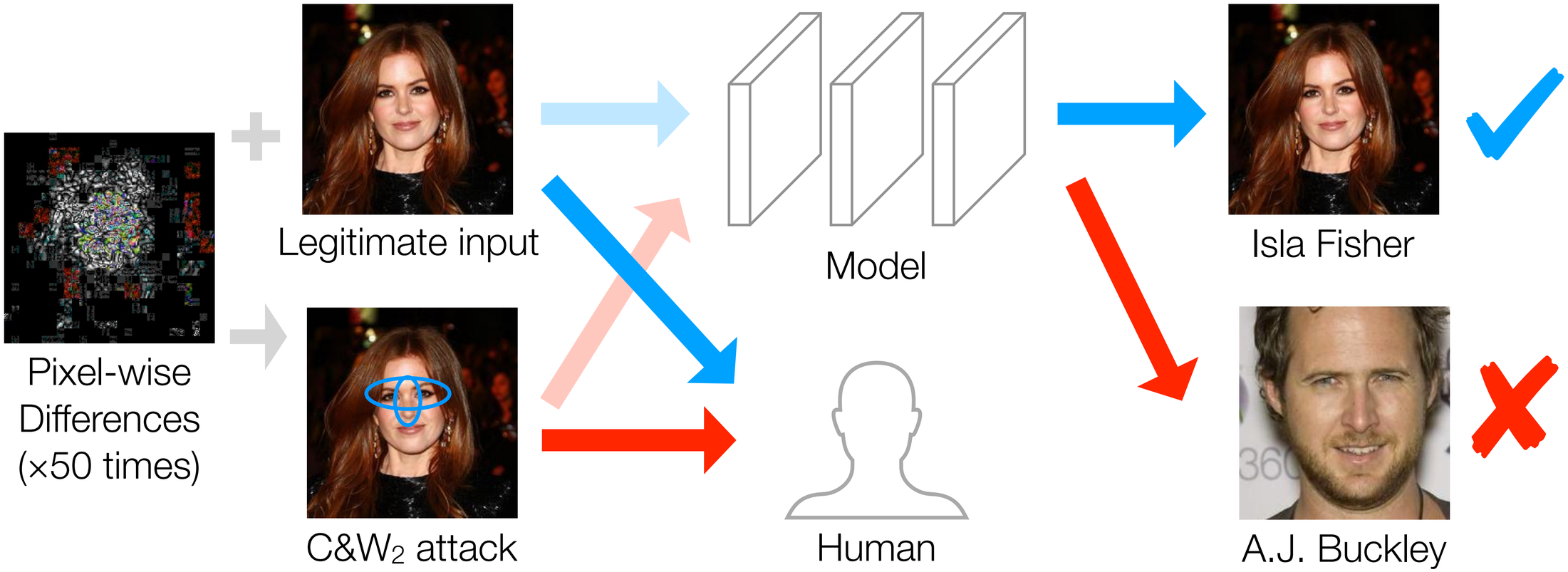
Understanding Adversarial Samples



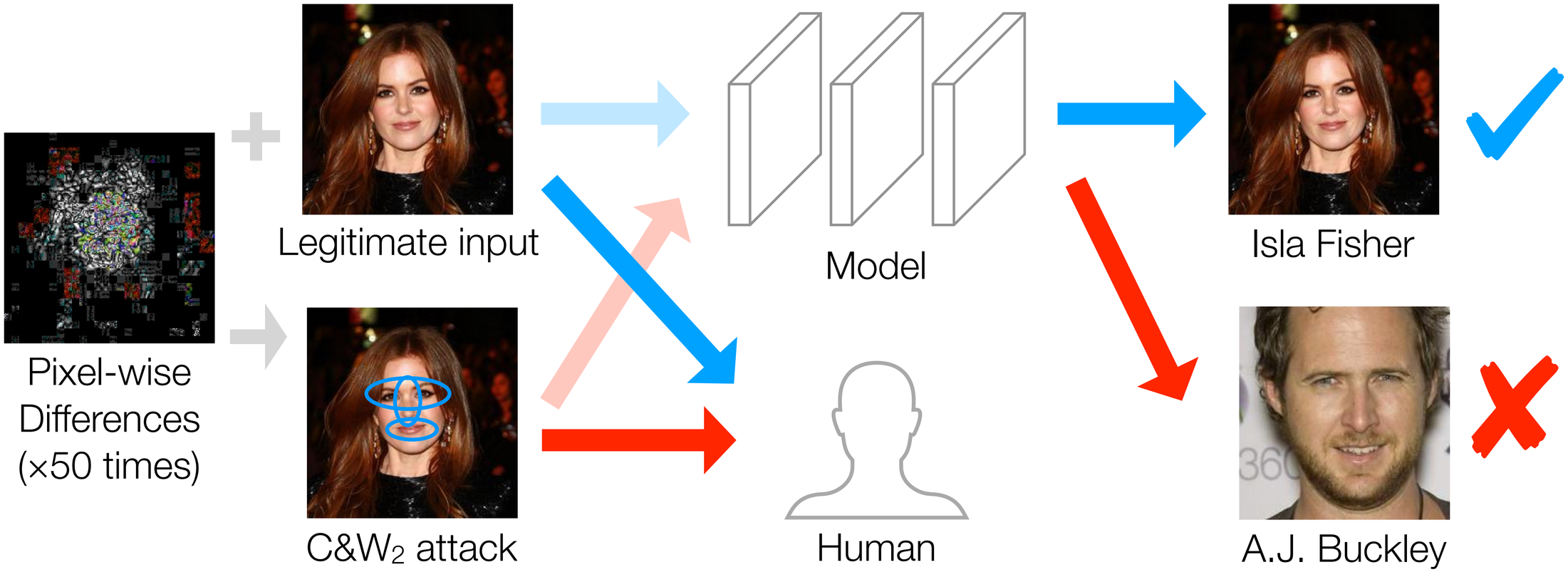
Understanding Adversarial Samples



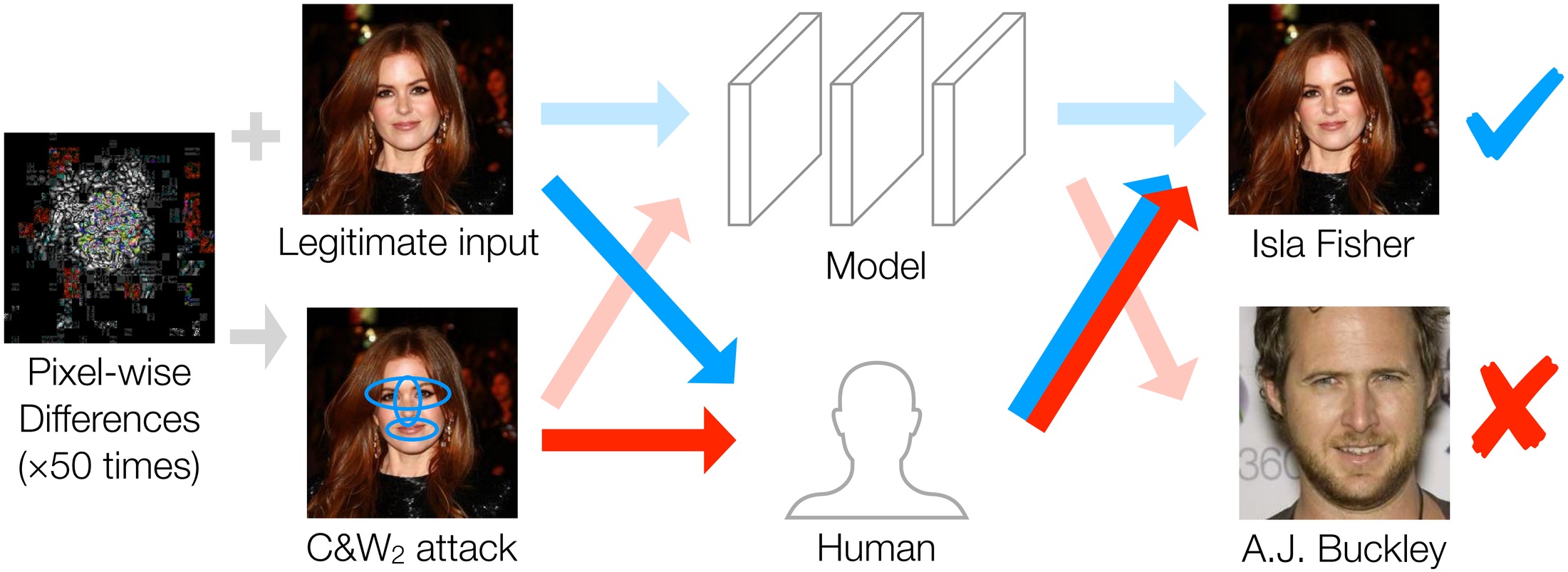
Understanding Adversarial Samples



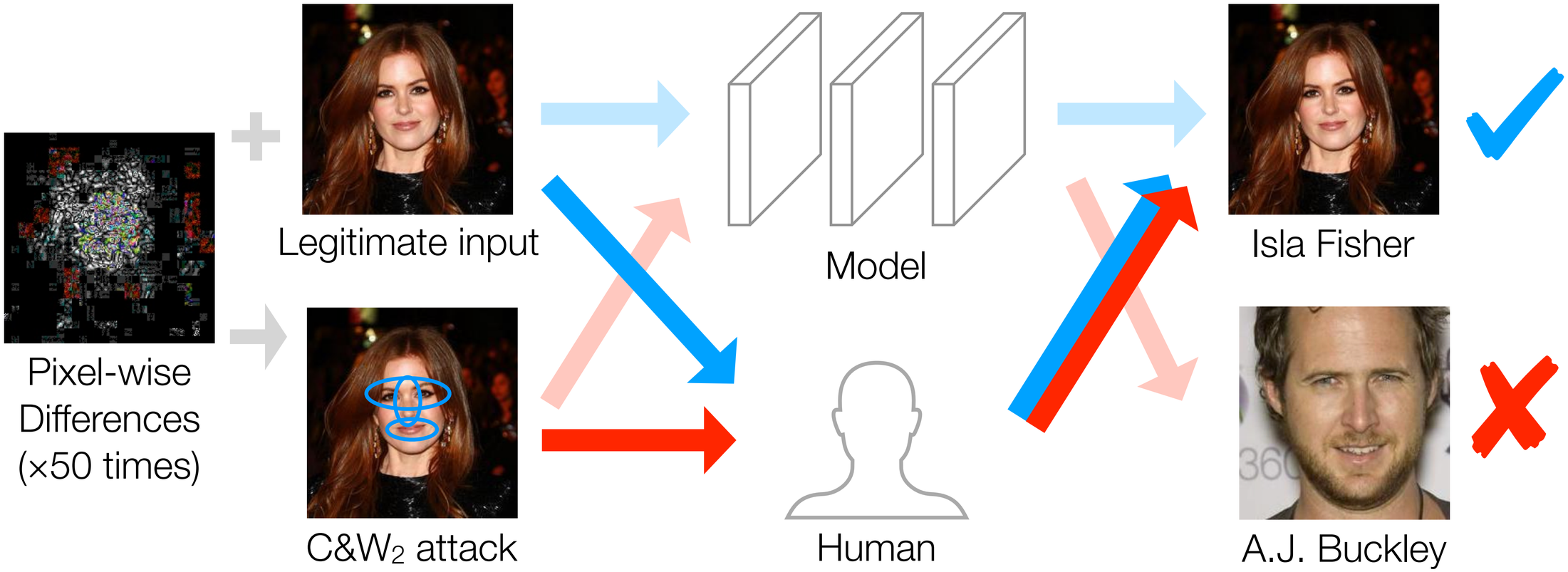
Understanding Adversarial Samples



Understanding Adversarial Samples



Understanding Adversarial Samples



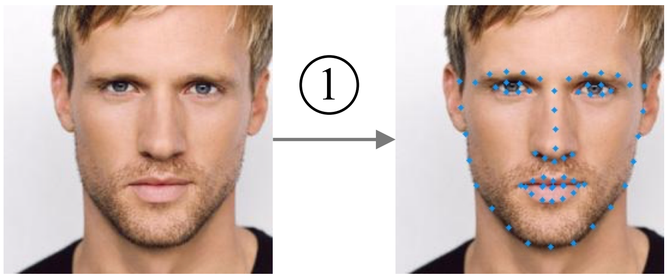
- Idea: is the classification result of a model mainly based on human perceptible attributes?

Architecture of Aml

Architecture of AmI

Input

Architecture of AmI

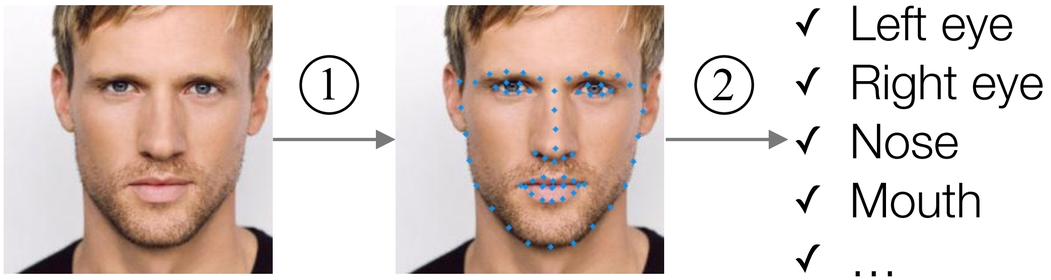


Input

①

Landmark generation

Architecture of AmI



① ②

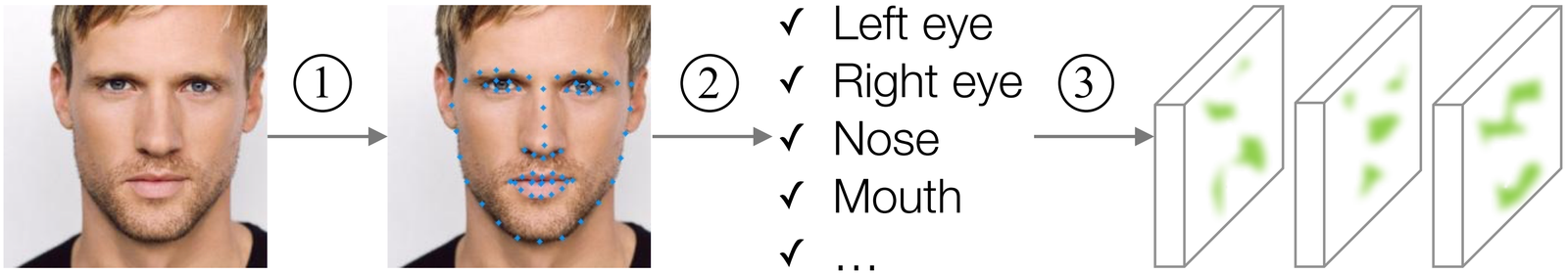
- ✓ Left eye
- ✓ Right eye
- ✓ Nose
- ✓ Mouth
- ✓ …

Input

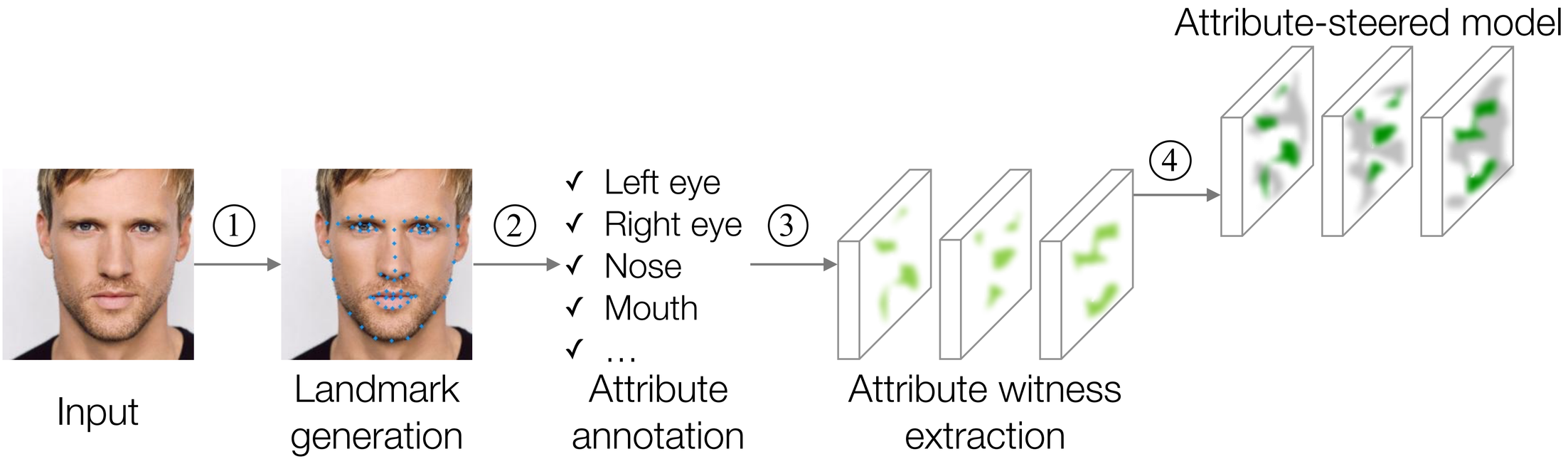
Landmark generation

Attribute annotation

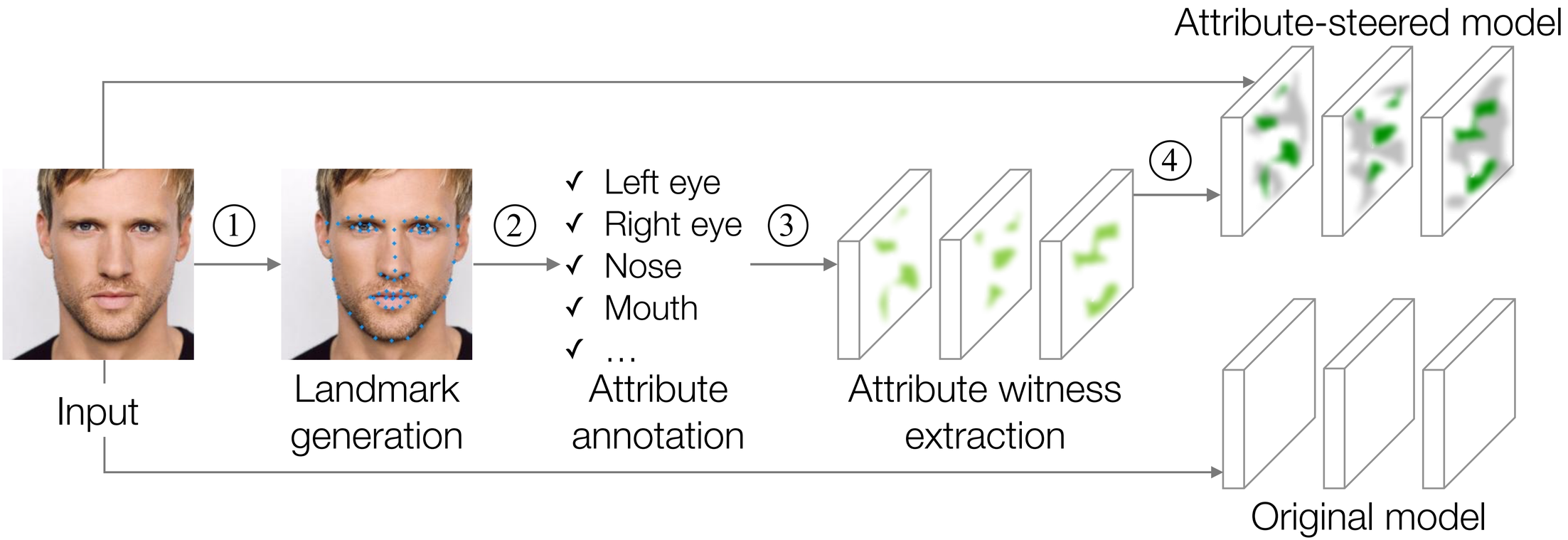
Architecture of AmI



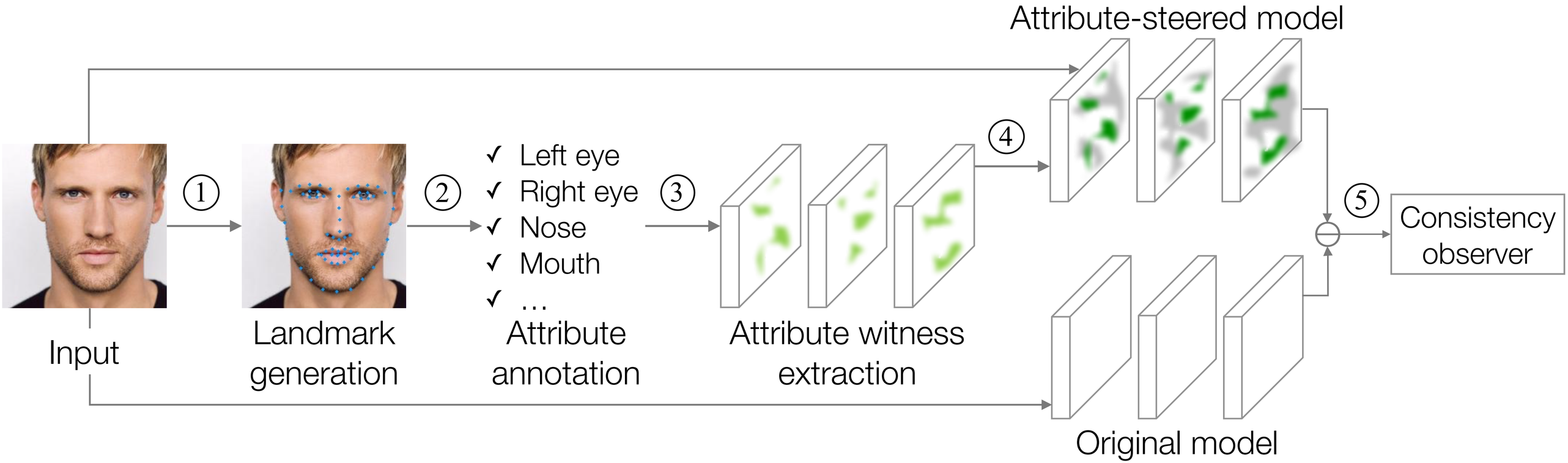
Architecture of AmI



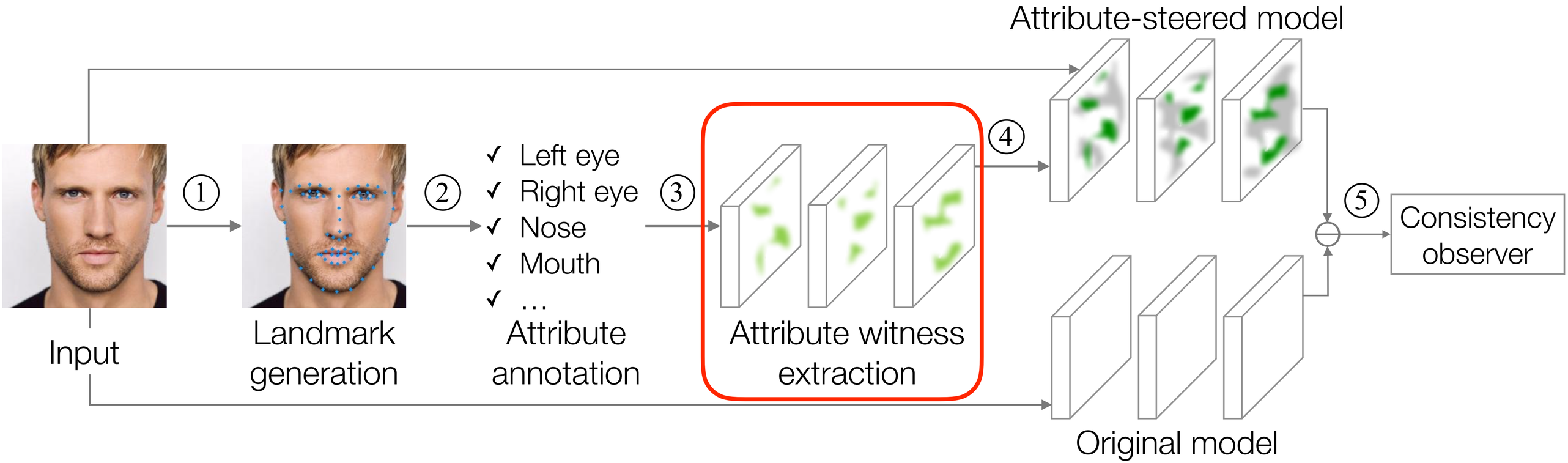
Architecture of AmI



Architecture of AmI



Architecture of AmI



Challenges

- Are there correspondences between attributes and neurons?
- If yes, how to extract corresponding neurons?

Challenges

- Are there correspondences between attributes and neurons?
- If yes, how to extract corresponding neurons?
- **Propose: Bi-directional reasoning**

Challenges

- Are there correspondences between attributes and neurons?
- If yes, how to extract corresponding neurons?
- **Propose: Bi-directional reasoning**
  - Forward: attribute changes —> neuron activation changes

Challenges

- Are there correspondences between attributes and neurons?
- If yes, how to extract corresponding neurons?
- **Propose: Bi-directional reasoning**
  - Forward: attribute changes —> neuron activation changes
  - Backward: neuron activation changes —> attribute changes

Challenges

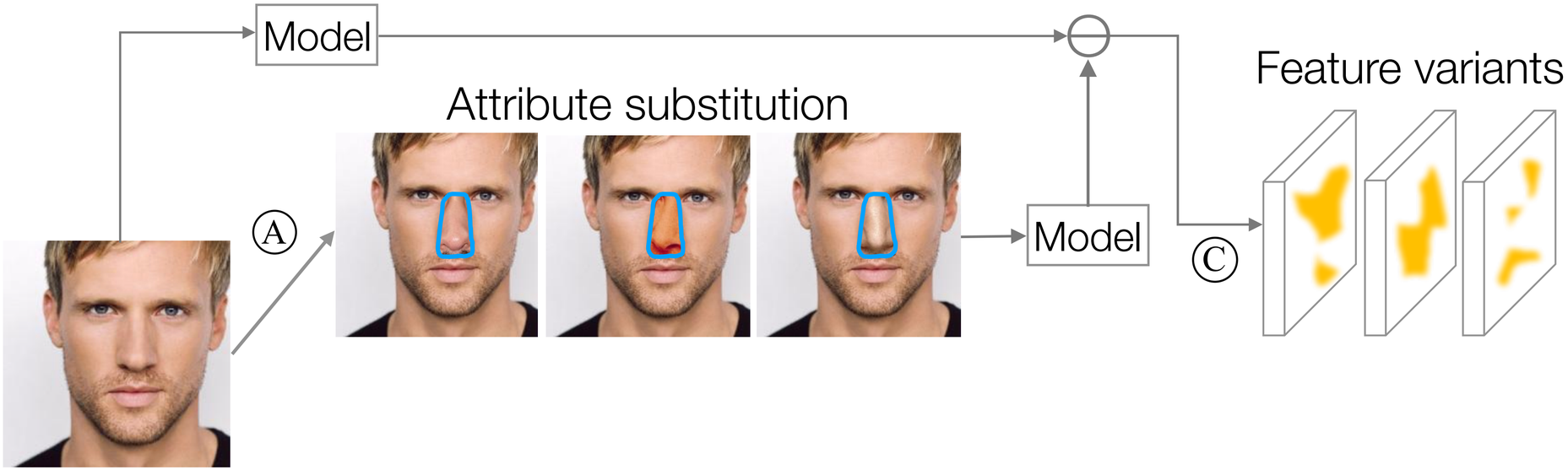
- Are there correspondences between attributes and neurons?
- If yes, how to extract corresponding neurons?
- **Propose: Bi-directional reasoning**
  - Forward: attribute changes —> neuron activation changes
  - Backward: neuron activation changes —> attribute changes
  - Backward: no attribute changes —> no neuron activation changes

Attribute Witness Extraction

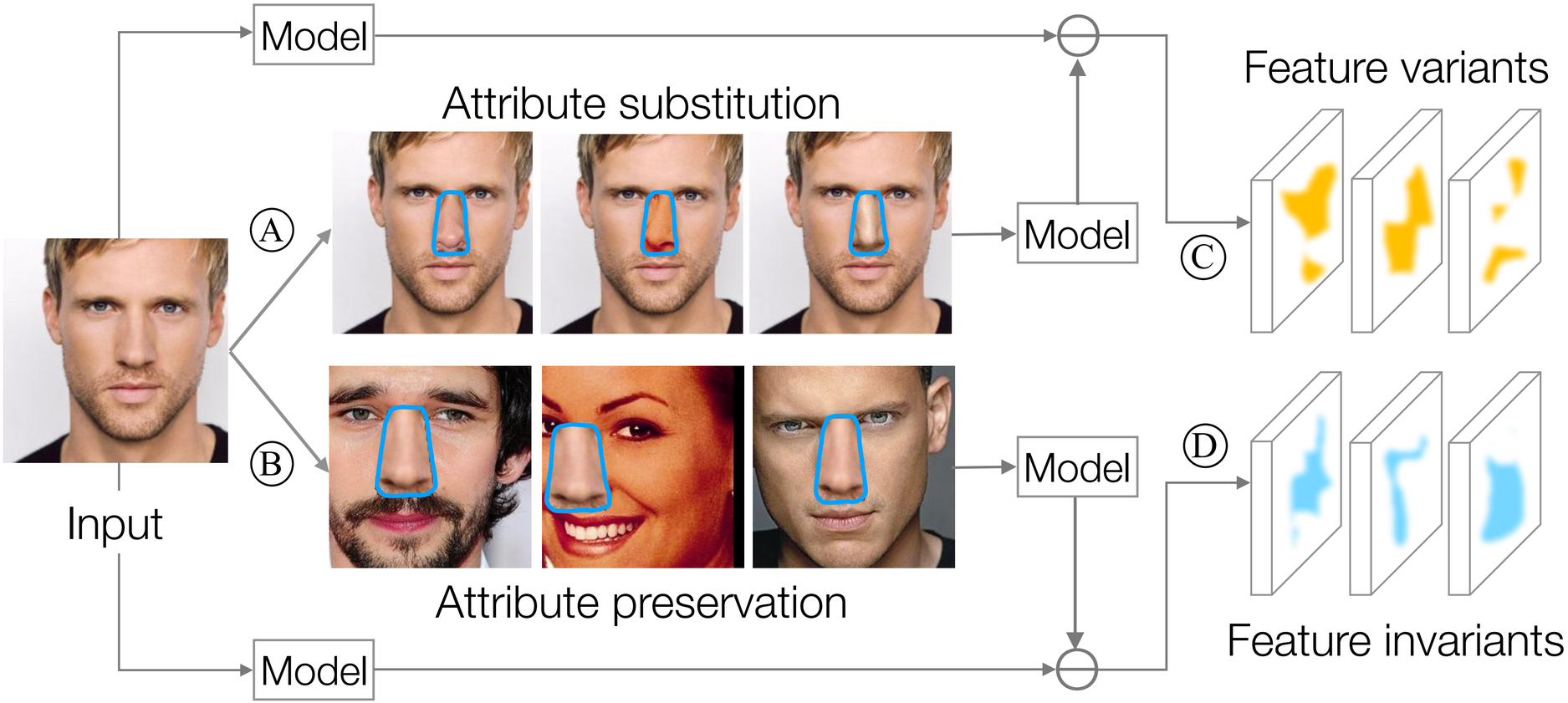
Attribute Witness Extraction

Input

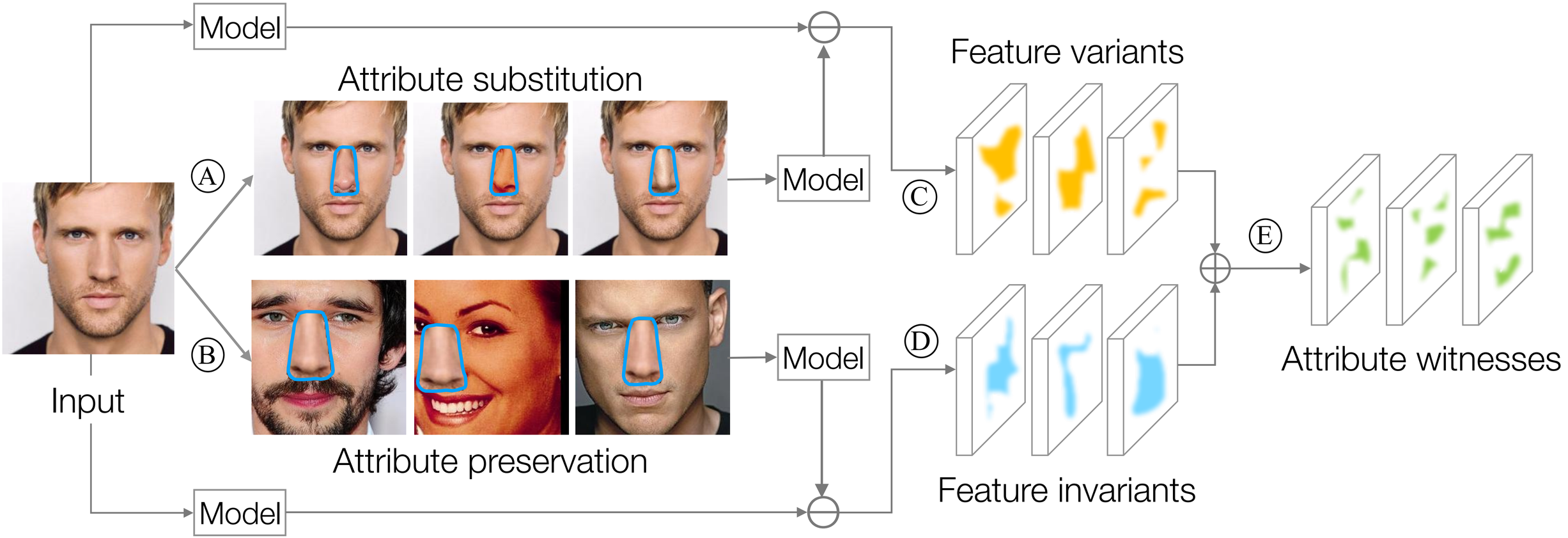
Attribute Witness Extraction



Attribute Witness Extraction



Attribute Witness Extraction



Experimental Results

Experimental Results

- Attribute witnesses

Experimental Results

- Attribute witnesses
  - The number of witnesses extracted is **smaller than 20**, although there are **64-4096** neurons in each layer

Experimental Results

- Attribute witnesses
  - The number of witnesses extracted is **smaller than 20**, although there are **64-4096** neurons in each layer
- Adversary detection

Experimental Results

- Attribute witnesses
  - The number of witnesses extracted is **smaller than 20**, although there are **64-4096** neurons in each layer
- Adversary detection
  - Achieve **94%** detection accuracy for **7** different kinds of attacks with **9.91%** false positives on benign inputs

Experimental Results

- Attribute witnesses
  - The number of witnesses extracted is **smaller than 20**, although there are **64-4096** neurons in each layer
- Adversary detection
  - Achieve **94%** detection accuracy for **7** different kinds of attacks with **9.91%** false positives on benign inputs
  - A state-of-the-art technique ***Feature Squeezing*** **(NDSS '18)** can only achieve **55%** accuracy with **23.3%** false positives for face recognition systems

Thank you!

Please visit our poster #99

05:00-07:00 PM @ Room 210 & 230 AB