

# Explainable AI for diffraction images

...

Barbara Klaudel

Supervisors: David Pennicard, Shah Nawaz, Vahid Rahmani

# Introduction

1. Motivation for explainable artificial intelligence (XAI).
2. The challenge.
3. Selected XAI methods.
4. Results.

# Deep Learning Black-Box Pitfalls

BBC

## Tay: Microsoft issues apology over racist chatbot fiasco

Microsoft has apologised for creating an artificially intelligent chatbot that quickly turned into a holocaust-denying racist.

Mar 25, 2016



Slate

## Facebook Apologizes After its AI Mislabels Video of Black Men as “Primates”

Within Facebook there has also been controversy regarding racial issues. A few years ago, CEO Mark Zuckerberg called on employees at the ...  
18 hours ago



The New York Times

## Apple Card Investigated After Gender Discrimination Complaints

Apple Card Investigated After Gender Discrimination Complaints ... A prominent software developer said on Twitter that the credit card was “sexist ...  
Nov 10, 2019



MIT Technology Review

## AI is sending people to jail—and getting it wrong

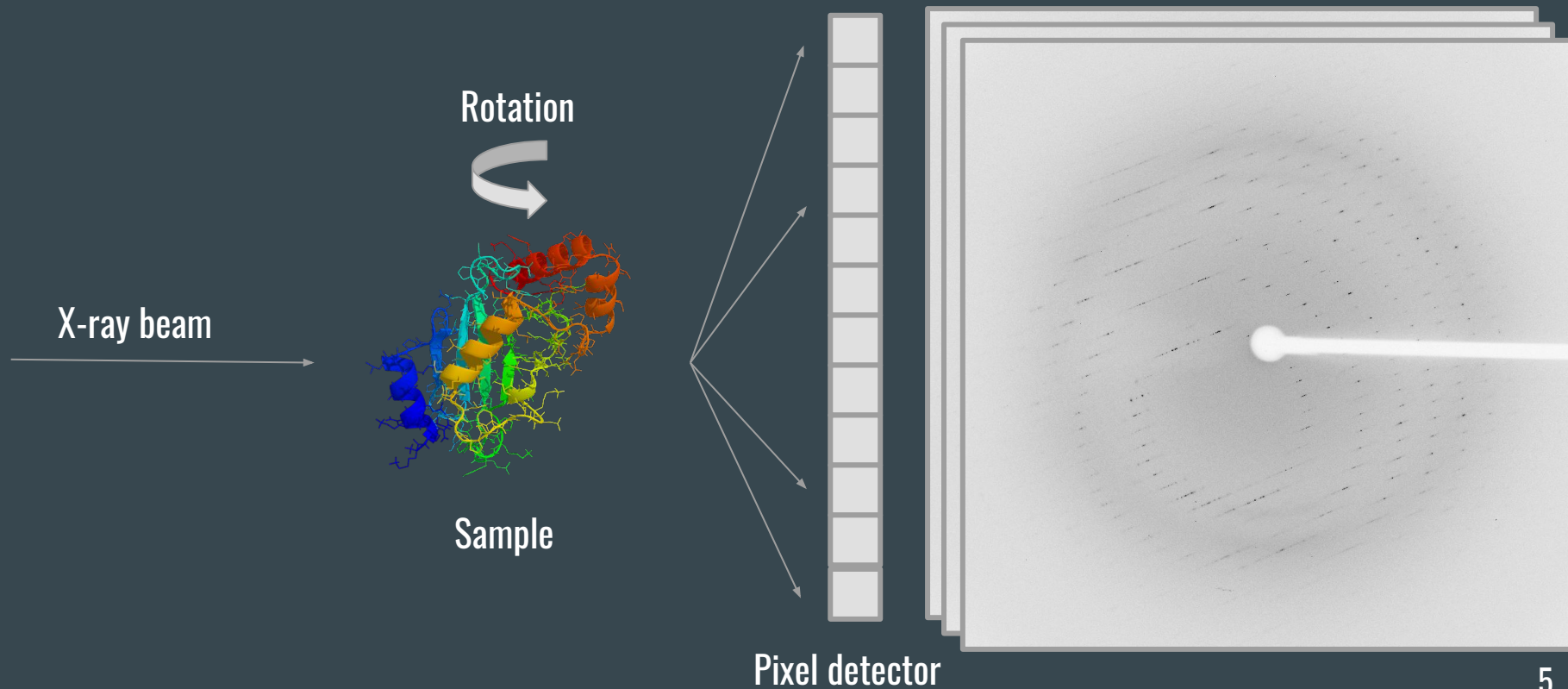
As a result, the algorithm could amplify and perpetuate embedded biases and generate even more bias-tainted data to feed a vicious cycle.  
Jan 21, 2019



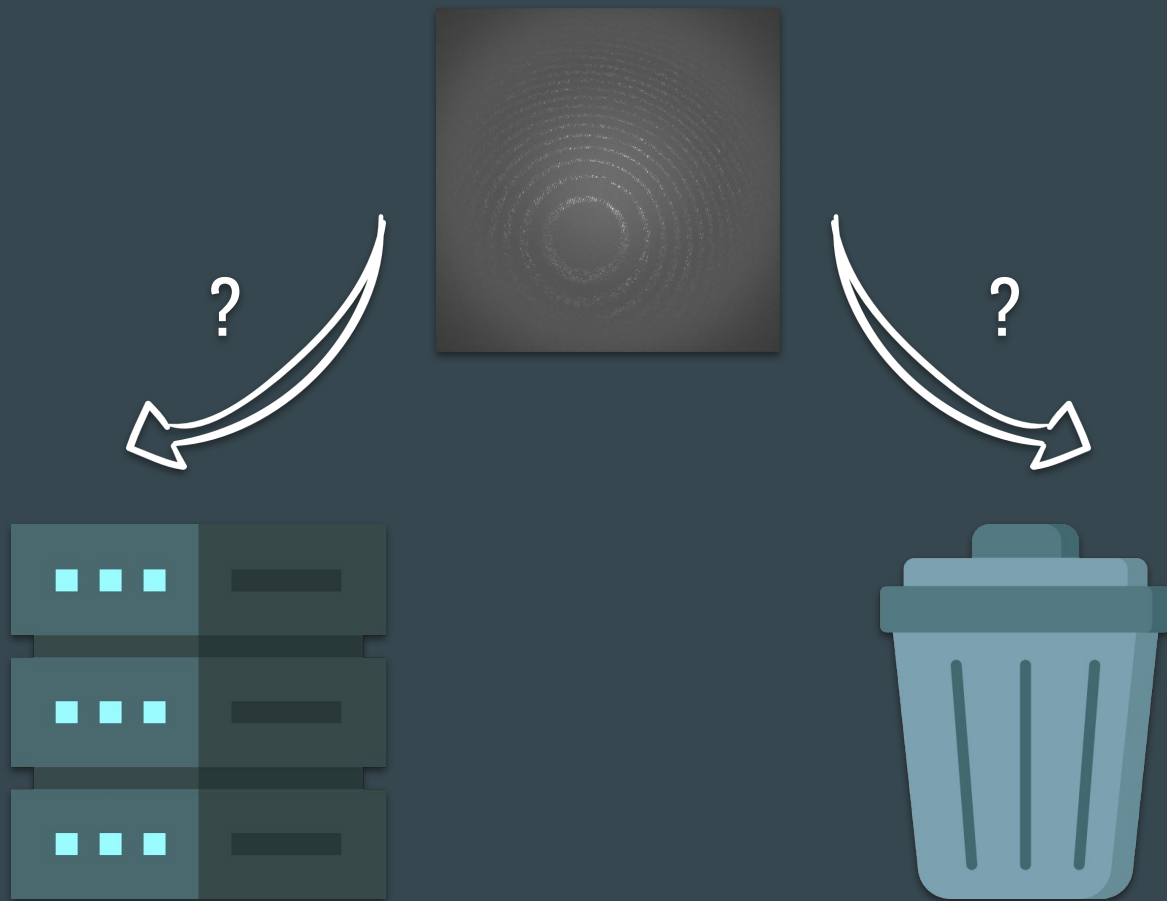
# The aim of Explainable AI



# X-ray diffraction

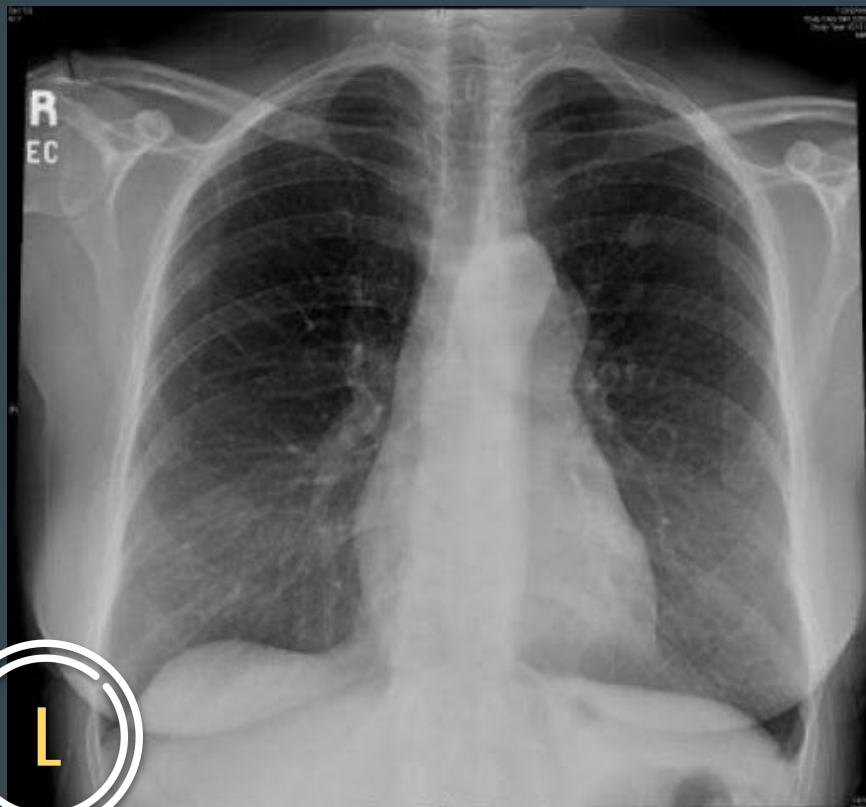


# TASK



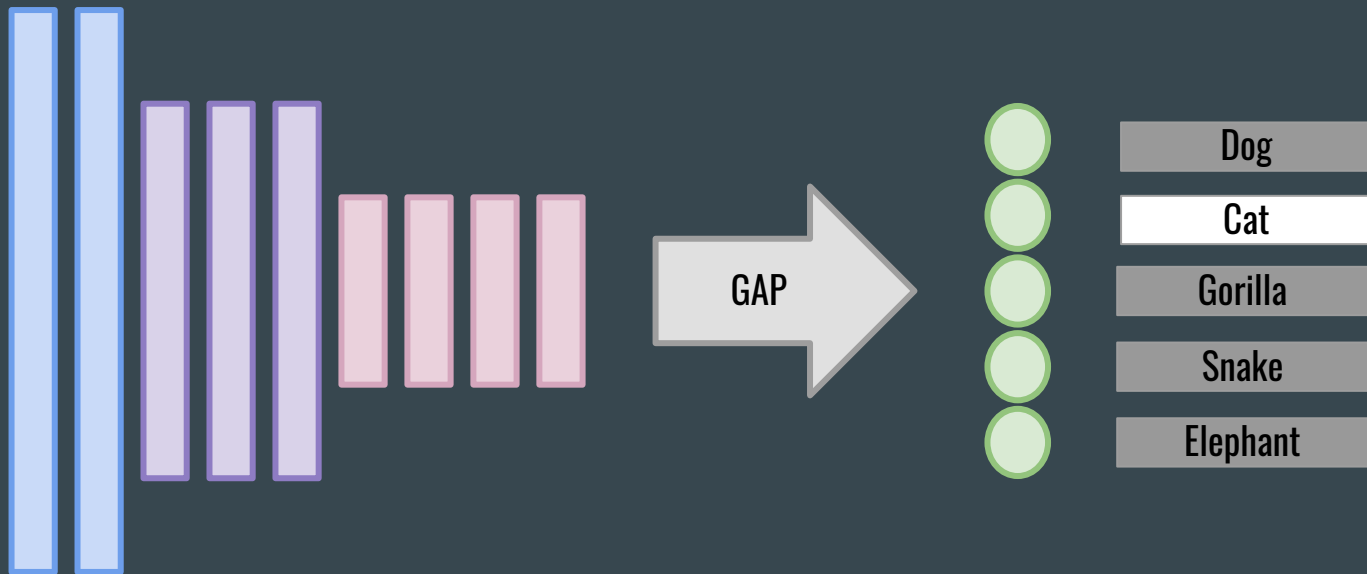
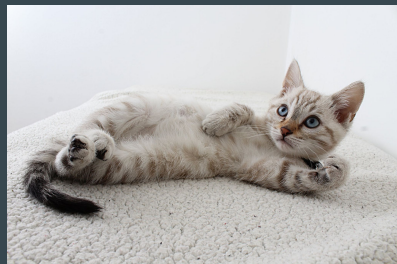
# Grad-CAM

# Examples

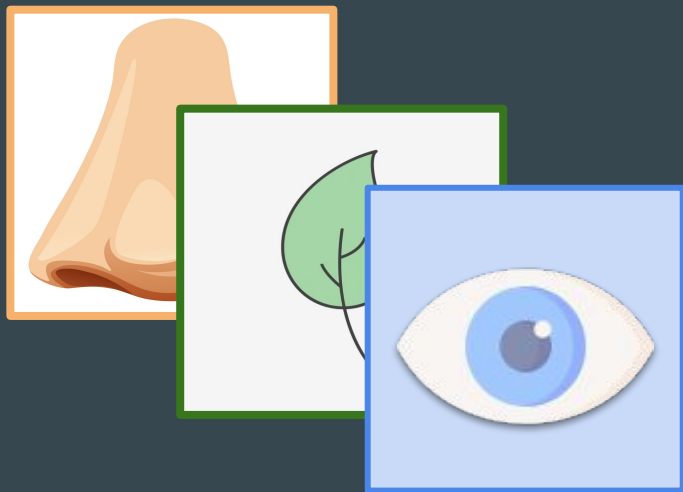




# Class Activation Mapping



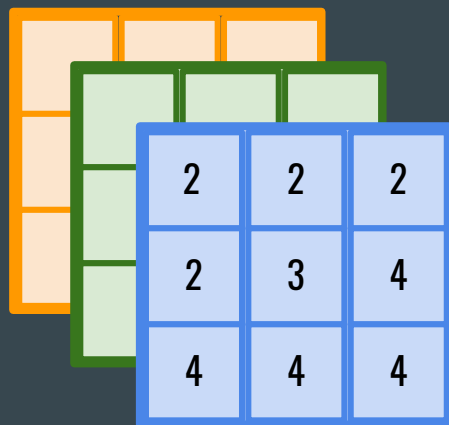
# Global Average Pooling



The success rates of  
extracting the features



# Global Average Pooling

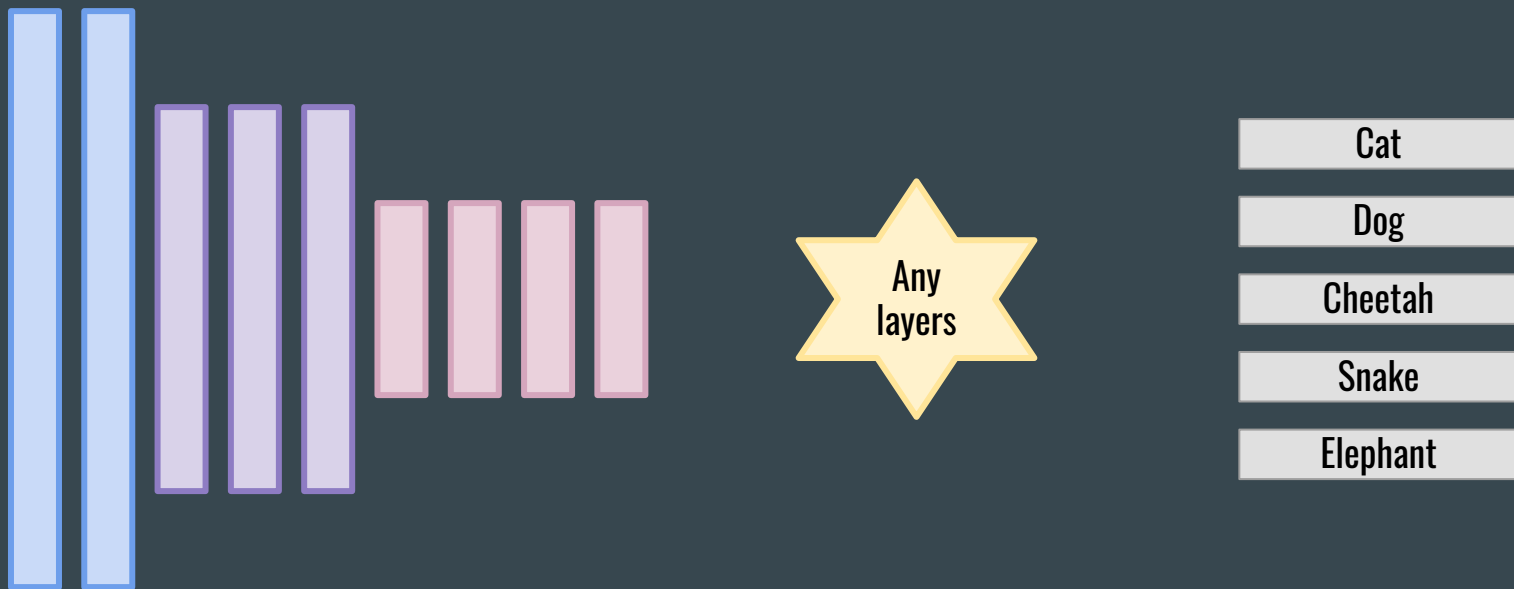


The success rates of  
extracting the features

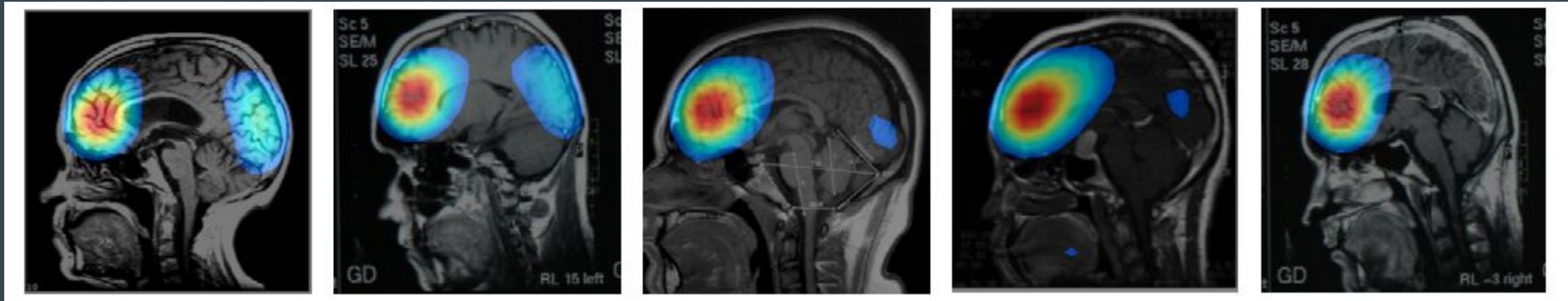


$$\text{AVERAGE} = (2 + 2 + 2 + 2 + 3 + 4 + 4 + 4 + 4) / 9 = 3$$

# Gradient-weighted Class Activation Mapping

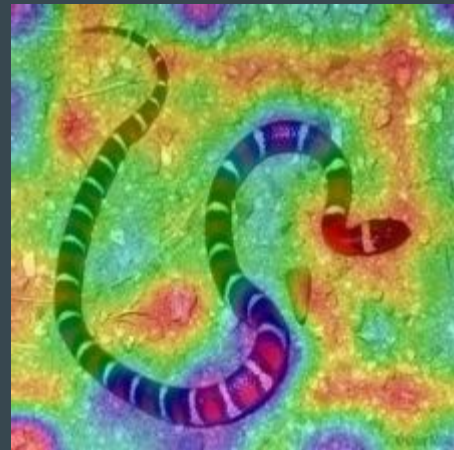


# Examples - Grad-CAM

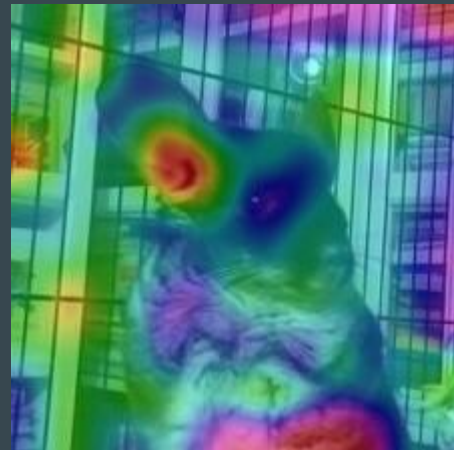


Source model: <https://doi.org/10.3390/diagnostics9020038>

# Examples - Grad-CAM



# Examples - Grad-CAM



# Guided Grad-CAM

Guided Backpropagation



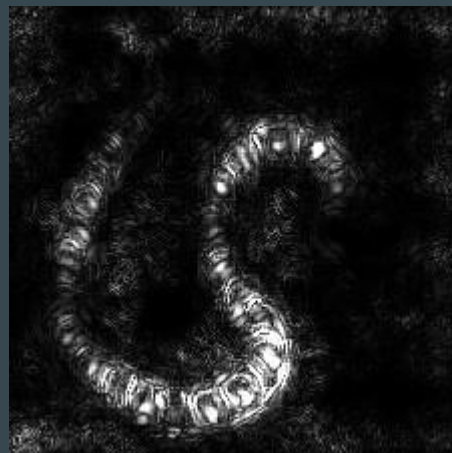
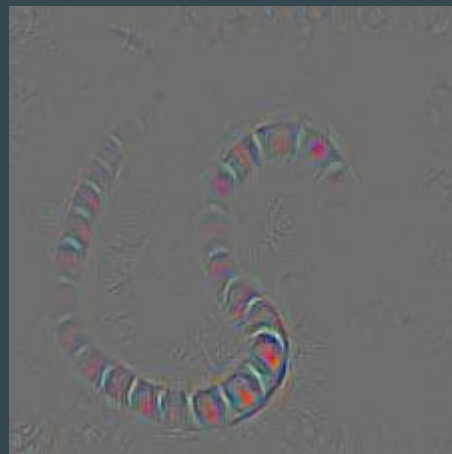
GradCAM



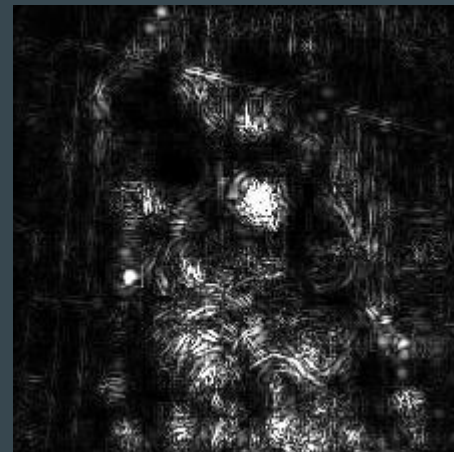
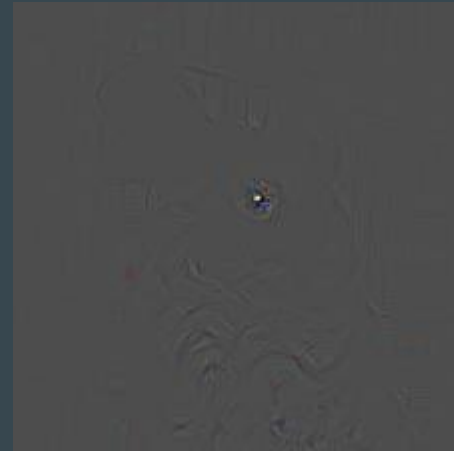
# Guided Grad-CAM

Compute gradients and zero out negatives.

# Examples - Guided Grad-CAM

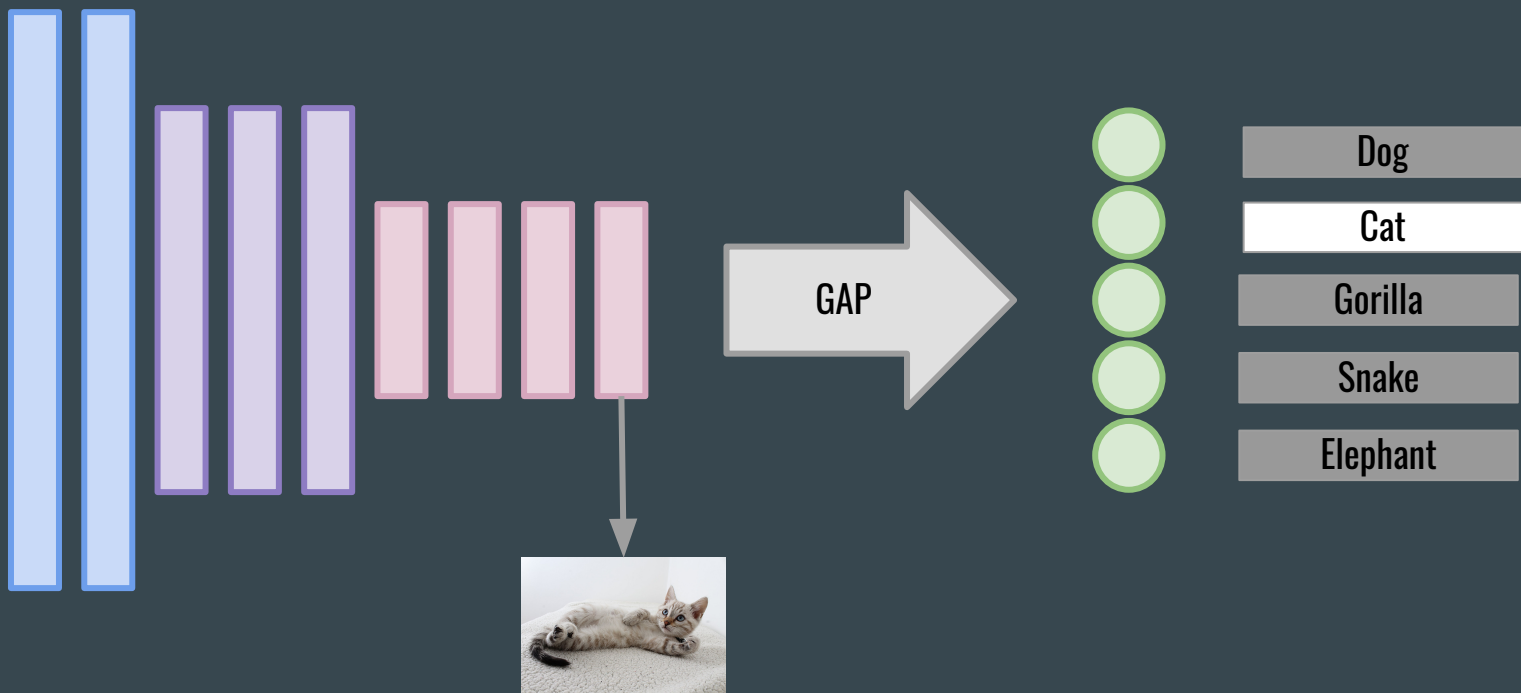


# Examples - Guided Grad-CAM

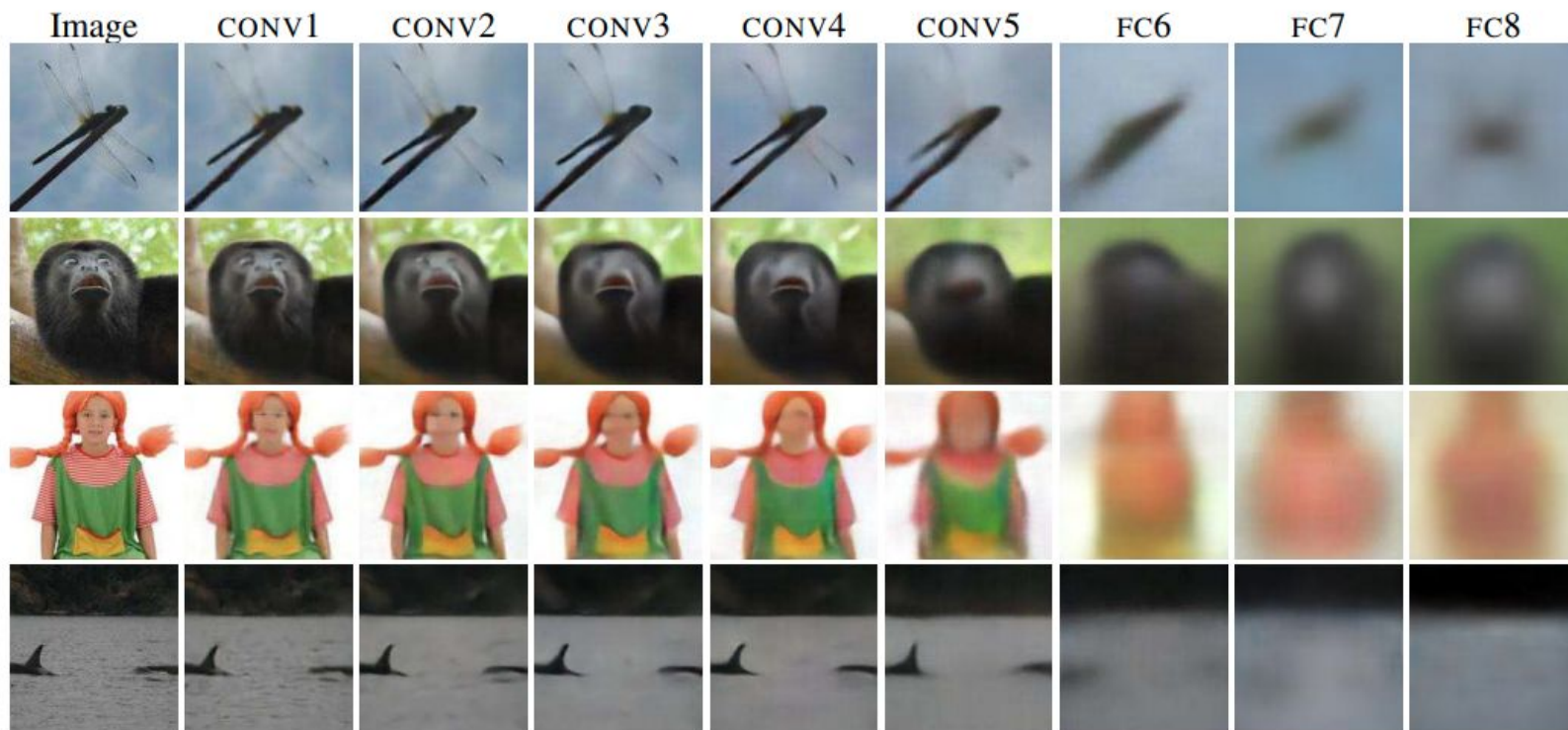


# Inverting Visual Representations

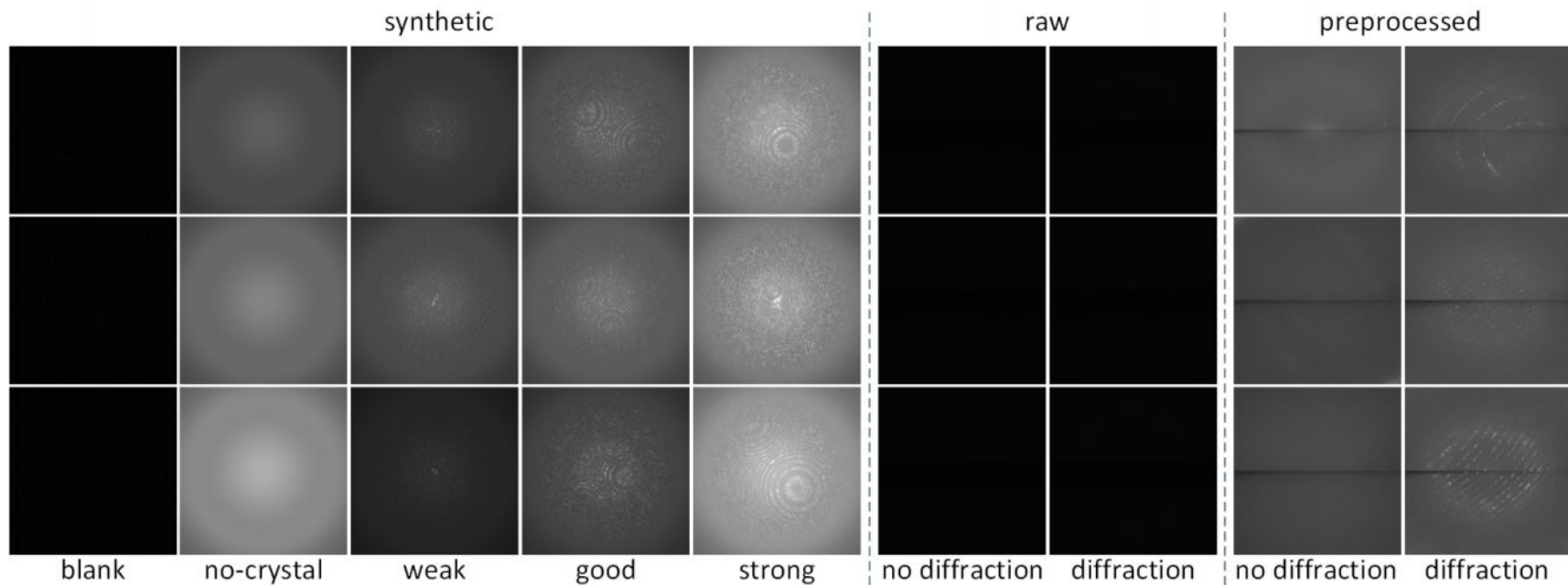
# Inverting Visual Representations



# Examples - Inverting visual representations

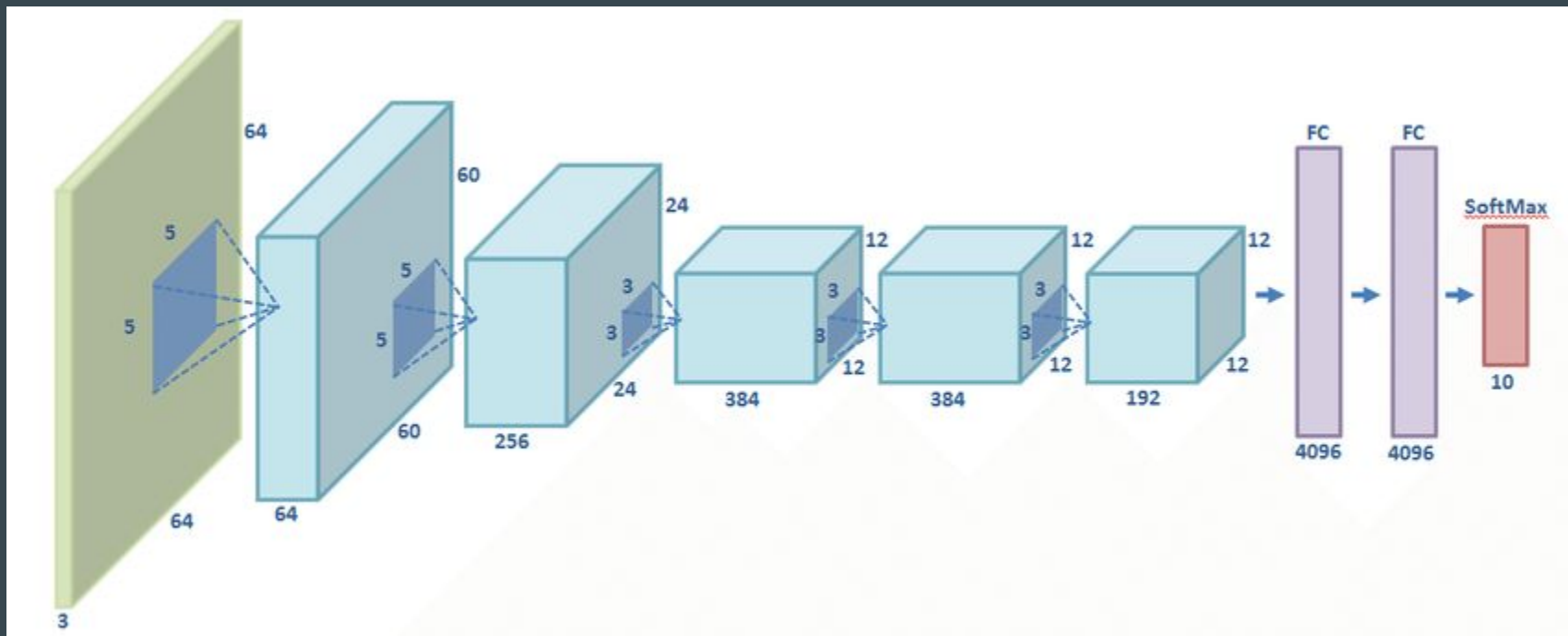


# Experiments





# AlexNet



# Confusion Matrix

TRUE  
CLASSES

|            |       |            |      |      |        |
|------------|-------|------------|------|------|--------|
| BLANK      | 2069  | 0          | 0    | 0    | 0      |
| NO-CRYSTAL | 0     | 3268       | 0    | 0    | 0      |
| WEAK       | 1     | 136        | 3120 | 89   | 0      |
| GOOD       | 0     | 0          | 128  | 2183 | 133    |
| STRONG     | 0     | 0          | 0    | 118  | 1354   |
|            | BLANK | NO-CRYSTAL | WEAK | GOOD | STRONG |

PREDICTED CLASSES

ACCURACY: 95.2 %

LABEL  
**strong**

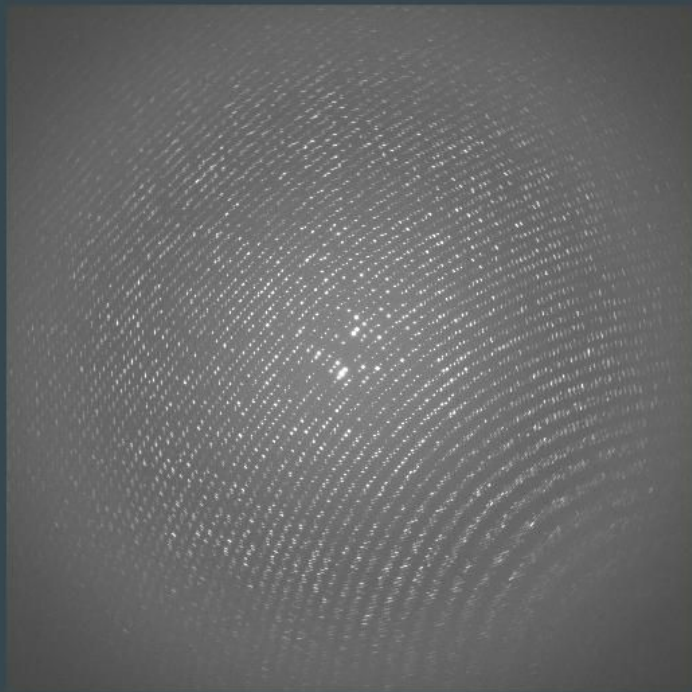
BLANK  
1.55e-06

NO-CRYSTAL  
1.09e-07

WEAK  
2.09e-06

GOOD  
0.0408

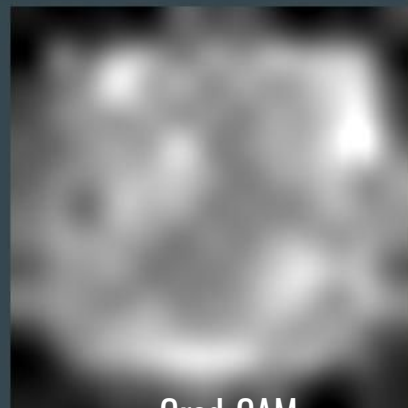
STRONG  
**0.959**



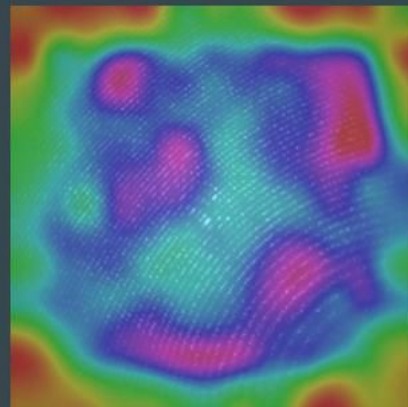
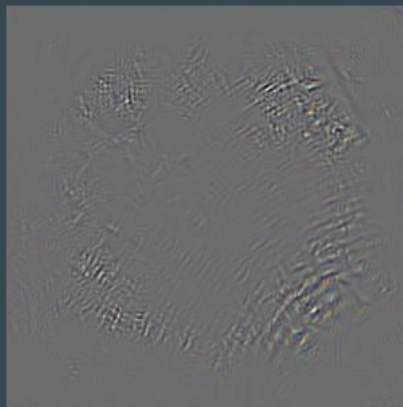
INPUT IMAGE



Guided Grad-CAM



Grad-CAM



LABEL  
**good**

BLANK  
1.01e-05

NO-CRYSTAL  
6.54e-06

WEAK  
0.000825

GOOD  
**0.99**

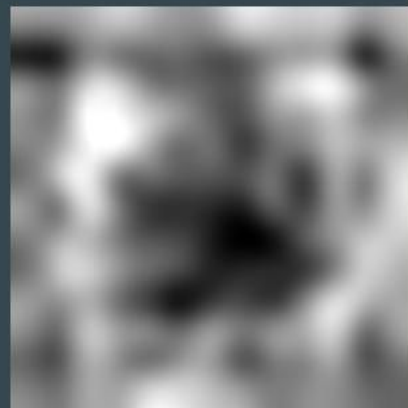
STRONG  
0.00934



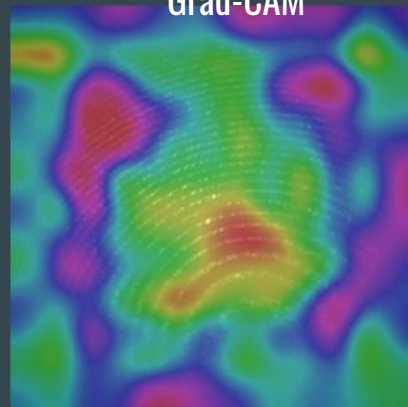
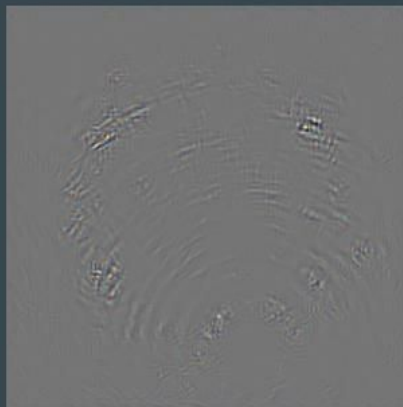
INPUT IMAGE



Guided Grad-CAM



Grad-CAM



LABEL  
**weak**

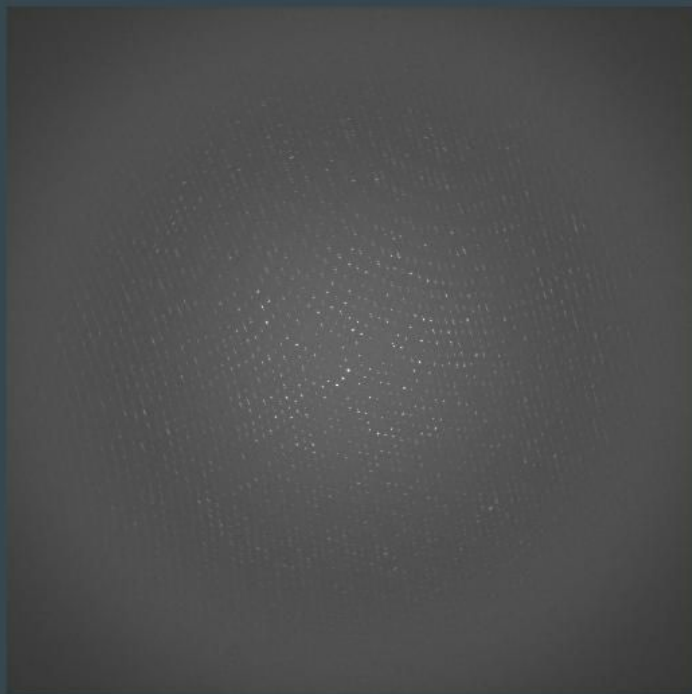
BLANK  
1.06e-08

NO-CRYSTAL  
2.16e-05

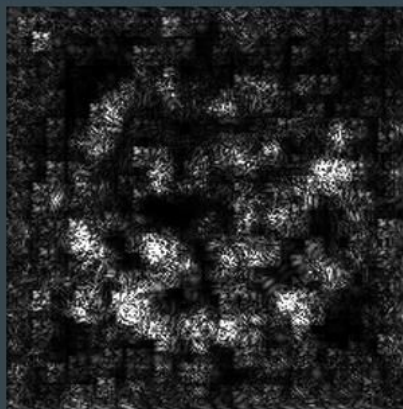
WEAK  
**1.0**

GOOD  
8.23e-06

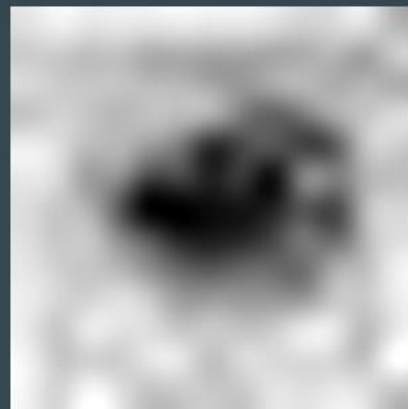
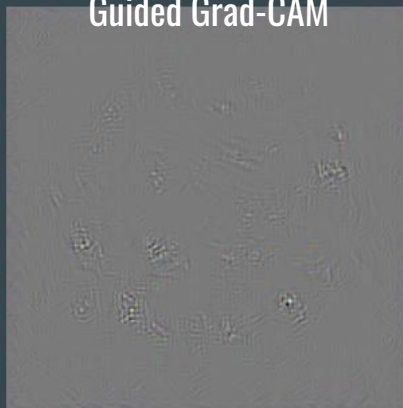
STRONG  
2.97e-14



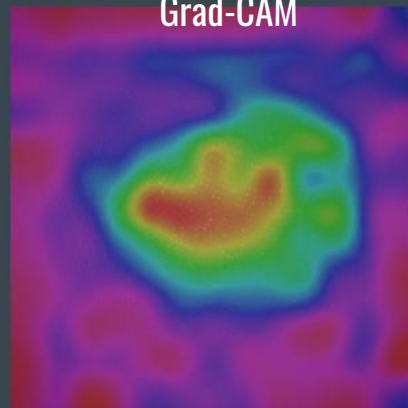
INPUT IMAGE



Guided Grad-CAM



Grad-CAM



LABEL  
**no-crystal**

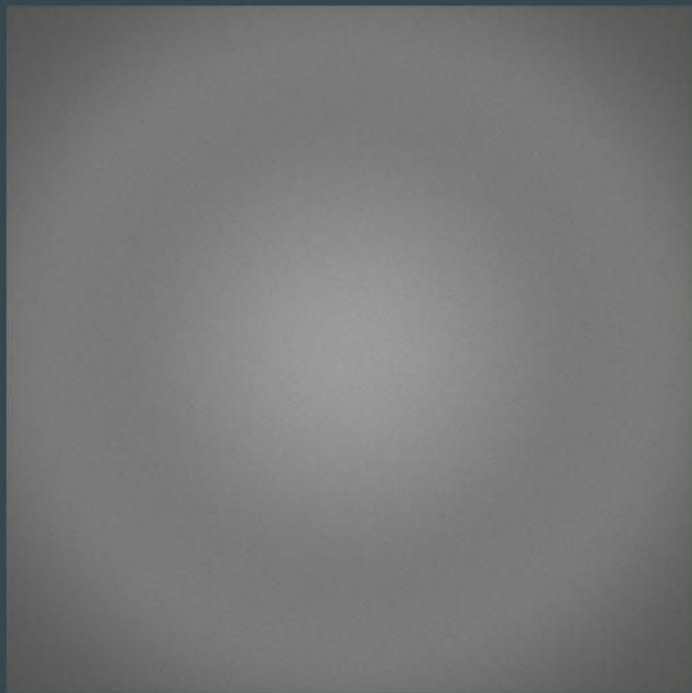
BLANK  
0.000222

NO-CRYSTAL  
**0.985**

WEAK  
0.0146

GOOD  
2.89e-06

STRONG  
3.65e-09



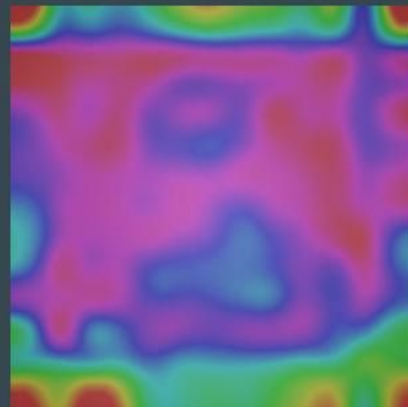
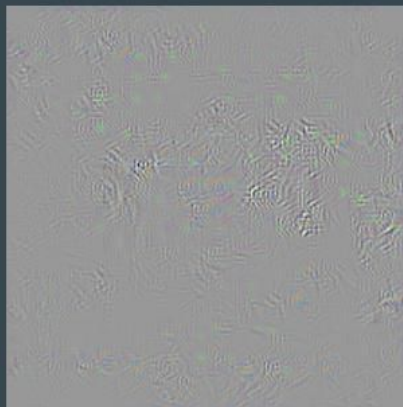
INPUT IMAGE



Guided Grad-CAM



Grad-CAM





LABEL  
**blank**

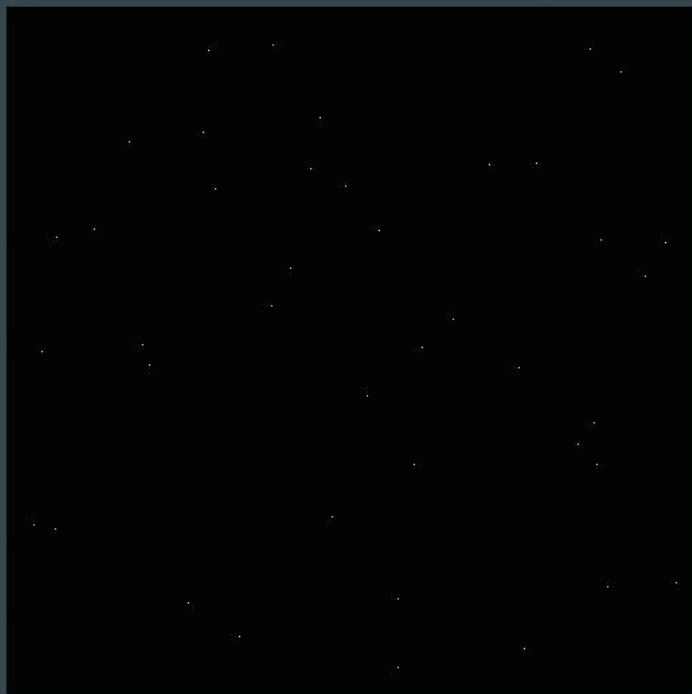
BLANK  
**0.986**

NO-CRYSTAL  
0.0105

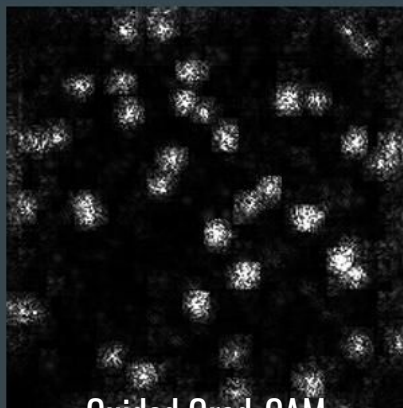
WEAK  
0.00385

GOOD  
1.24e-09

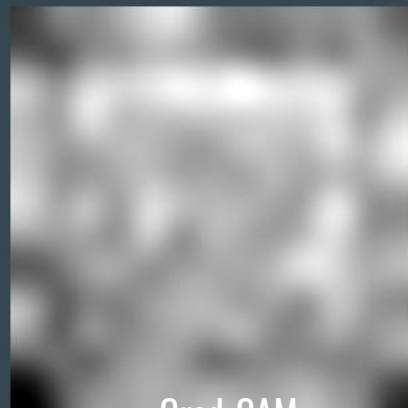
STRONG  
6.01e-11



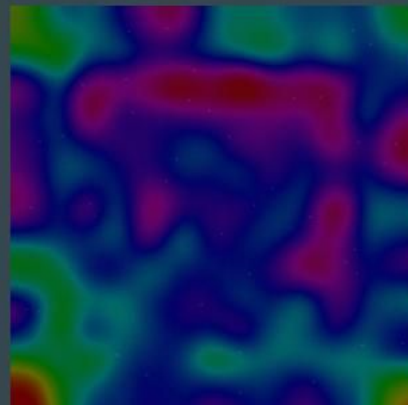
INPUT IMAGE

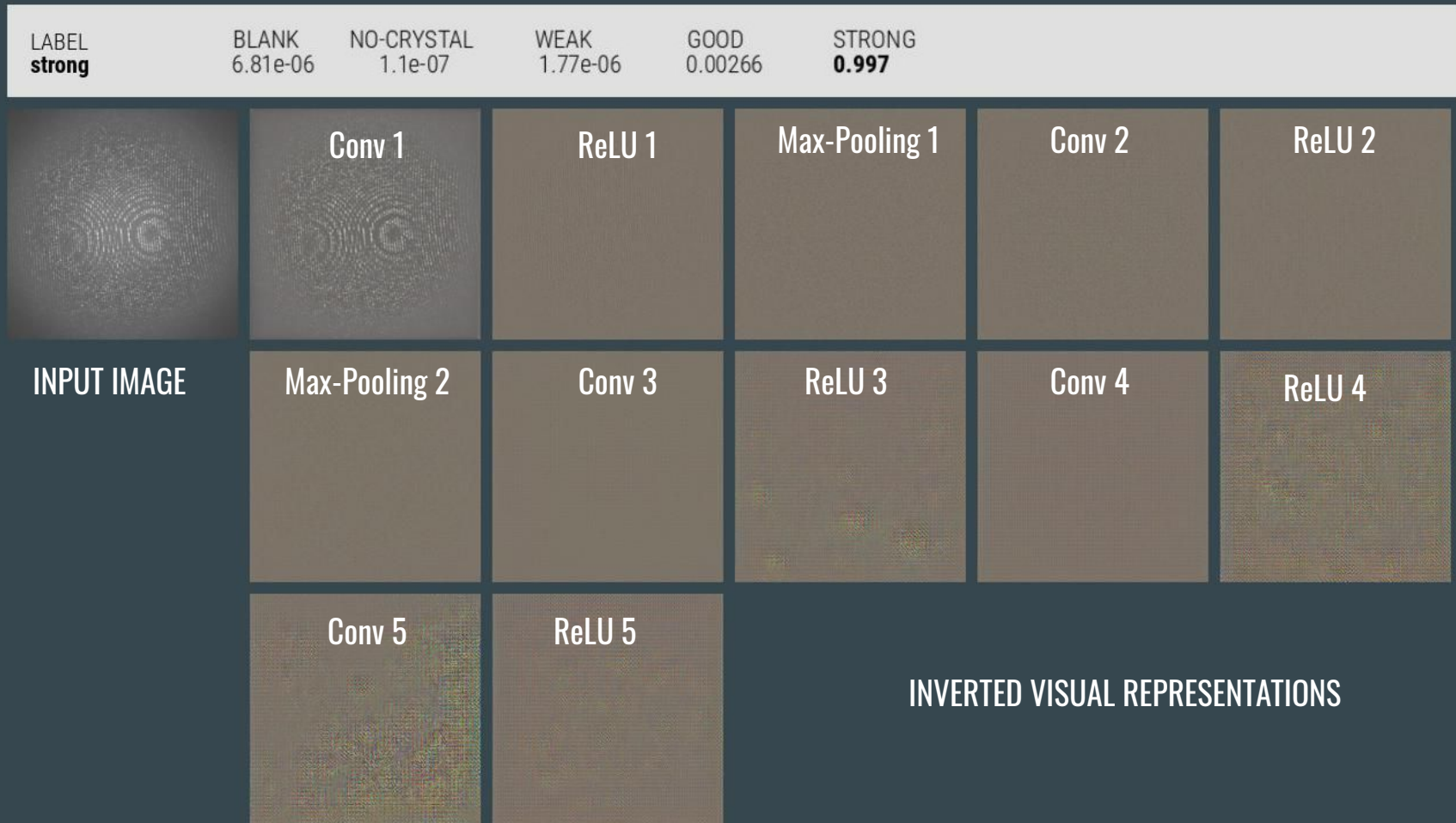


Guided Grad-CAM



Grad-CAM







# Thank you.

...

?? || /\*\*/