# Benchmarking the Attribution Quality of Vision Models


Robin Hesse


Simone Schaub-Meyer


Stefan Roth

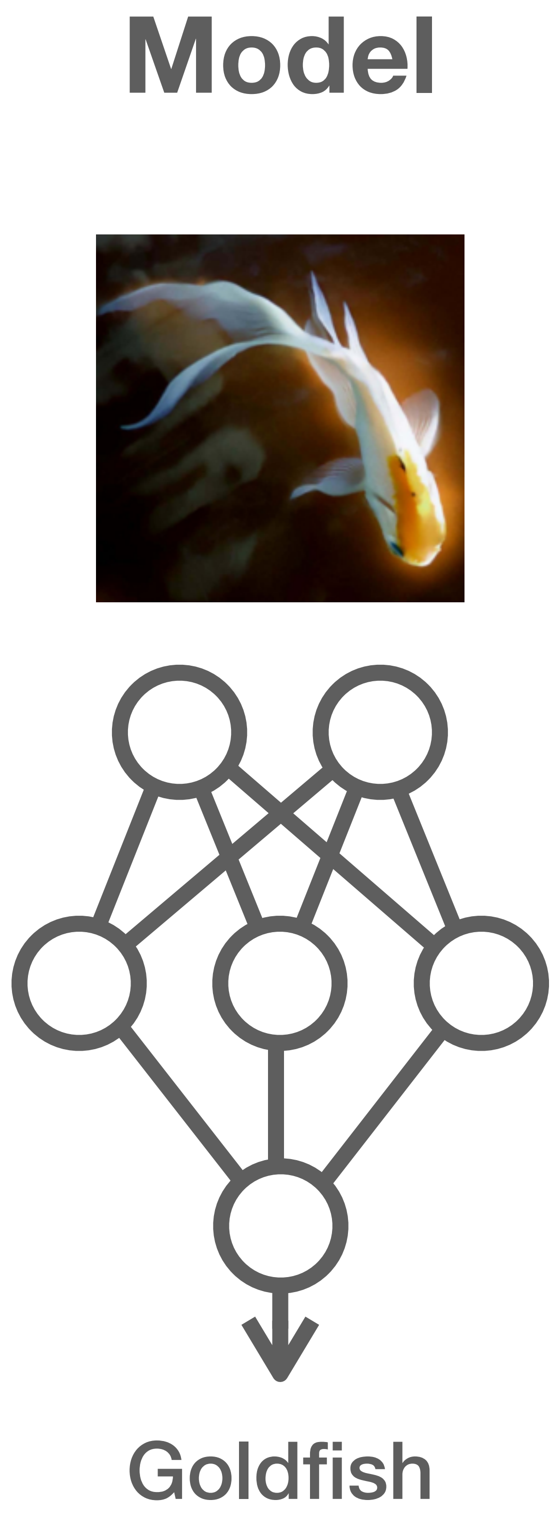
Visual Inference Lab | TU Darmstadt

NEURAL INFORMATION PROCESSING SYSTEMS

TECHNISCHE UNIVERSITÄT DARMSTADT

hessian.AI

# What are attribution maps
## …and why is it hard to evaluate them?

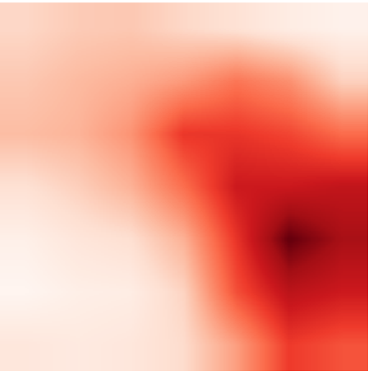



[1] Selvaraju et al. (2017). “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: ICCV
[2] Sundararajan et al. (2017). “Axiomatic attribution for deep networks.” In: ICML

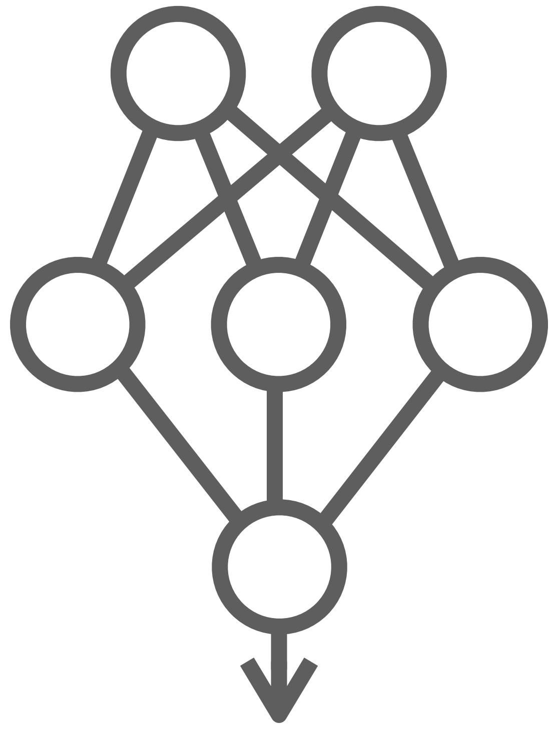
# What are attribution maps
## …and why is it hard to evaluate them?

### Attribution

Grad-CAM [1]



### Model



Goldfish

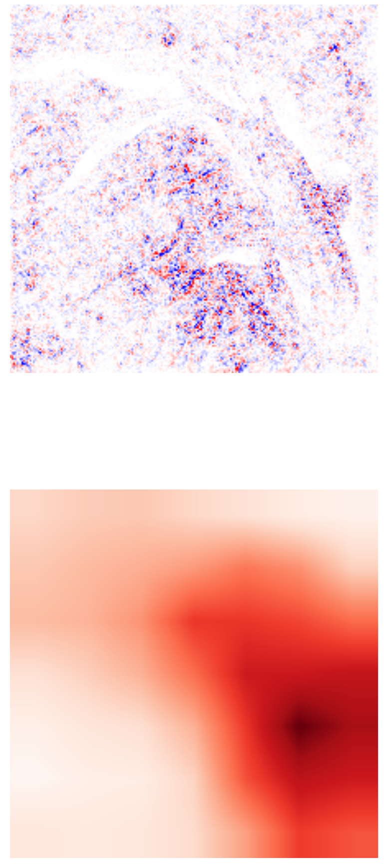
[1] Selvaraju et al. (2017). “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: ICCV
[2] Sundararajan et al. (2017). “Axiomatic attribution for deep networks.” In: ICML

# What are attribution maps
## …and why is it hard to evaluate them?

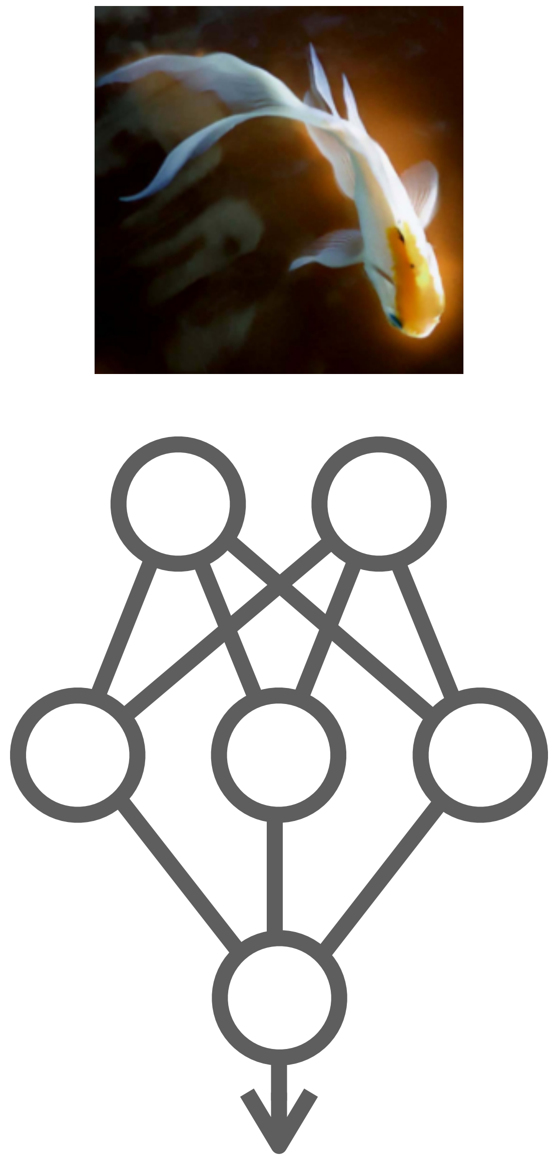
### Attribution

Integrated Gradients [2]

Grad-CAM [1]



### Model



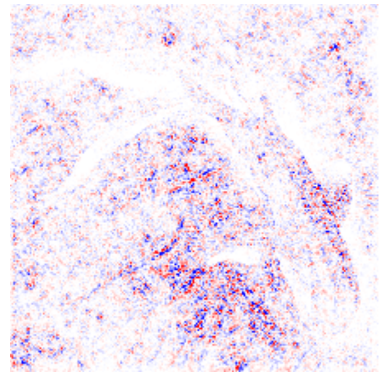
Goldfish

[1] Selvaraju et al. (2017). "Grad-CAM: Visual explanations from deep networks via gradient-based localization." In: ICCV
[2] Sundararajan et al. (2017). "Axiomatic attribution for deep networks." In: ICML

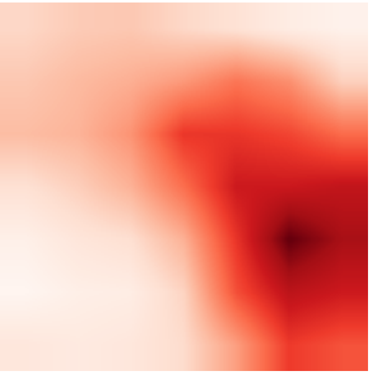
# What are attribution maps
## …and why is it hard to evaluate them?

### Attribution

Integrated Gradients [2]

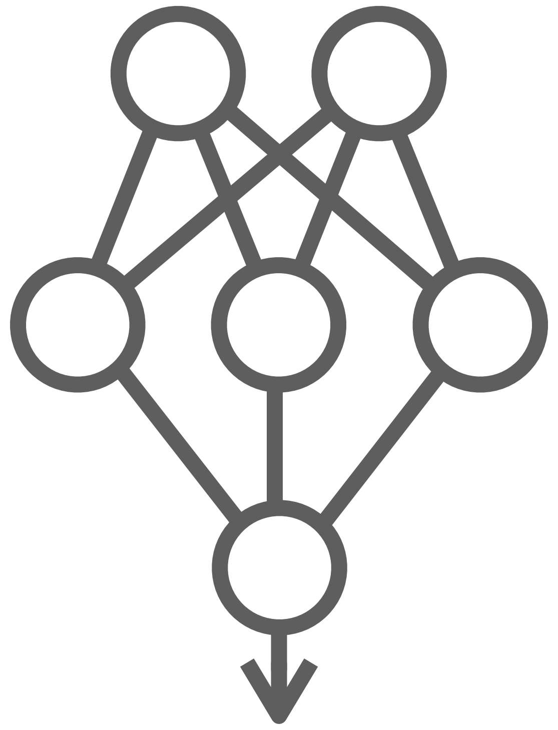


Grad-CAM [1]



**No ground truth explanation!**

### Model

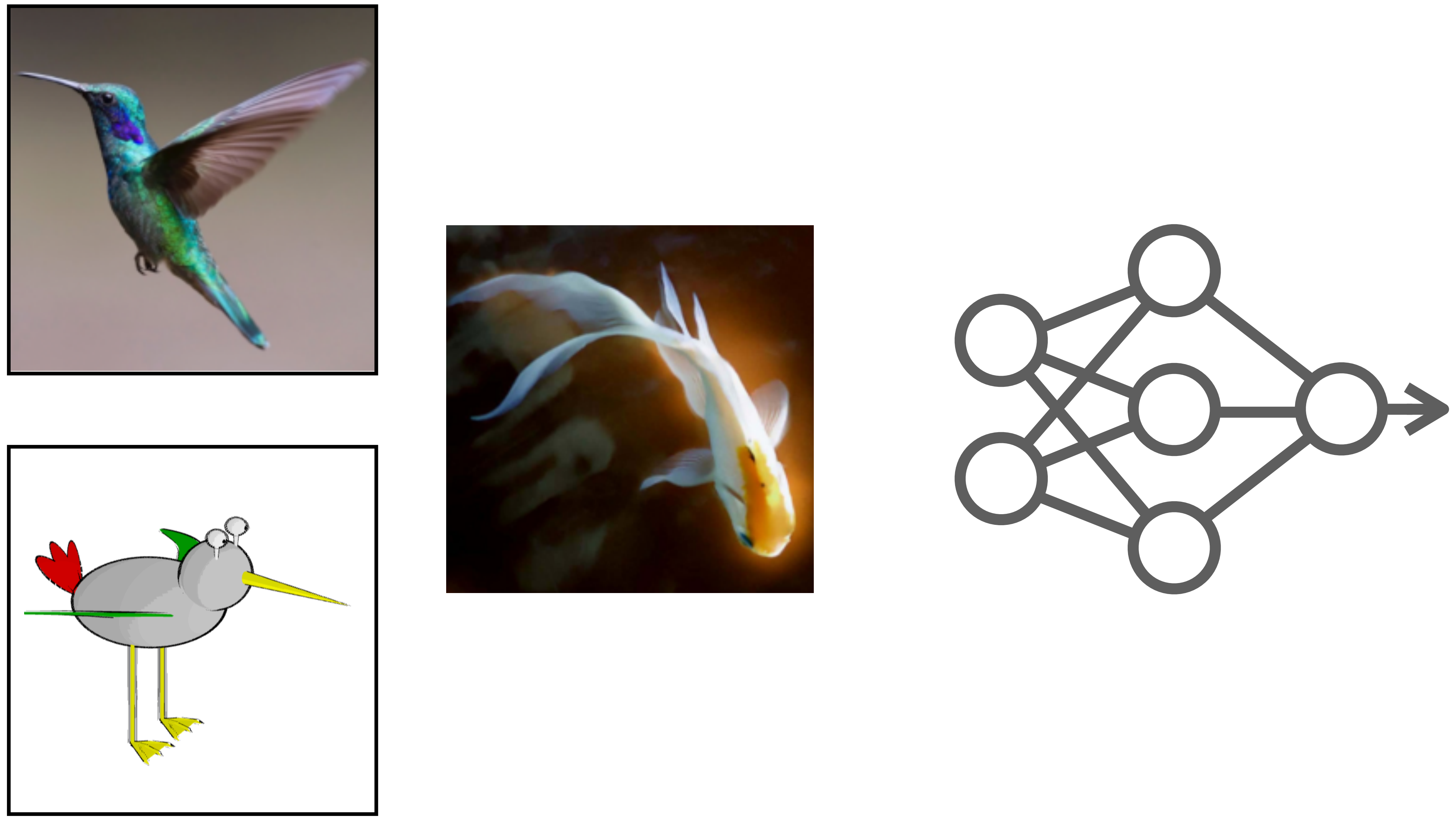


Goldfish

[1] Selvaraju et al. (2017). “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: ICCV
[2] Sundararajan et al. (2017). “Axiomatic attribution for deep networks.” In: ICML

# Related work

## …and its limitations

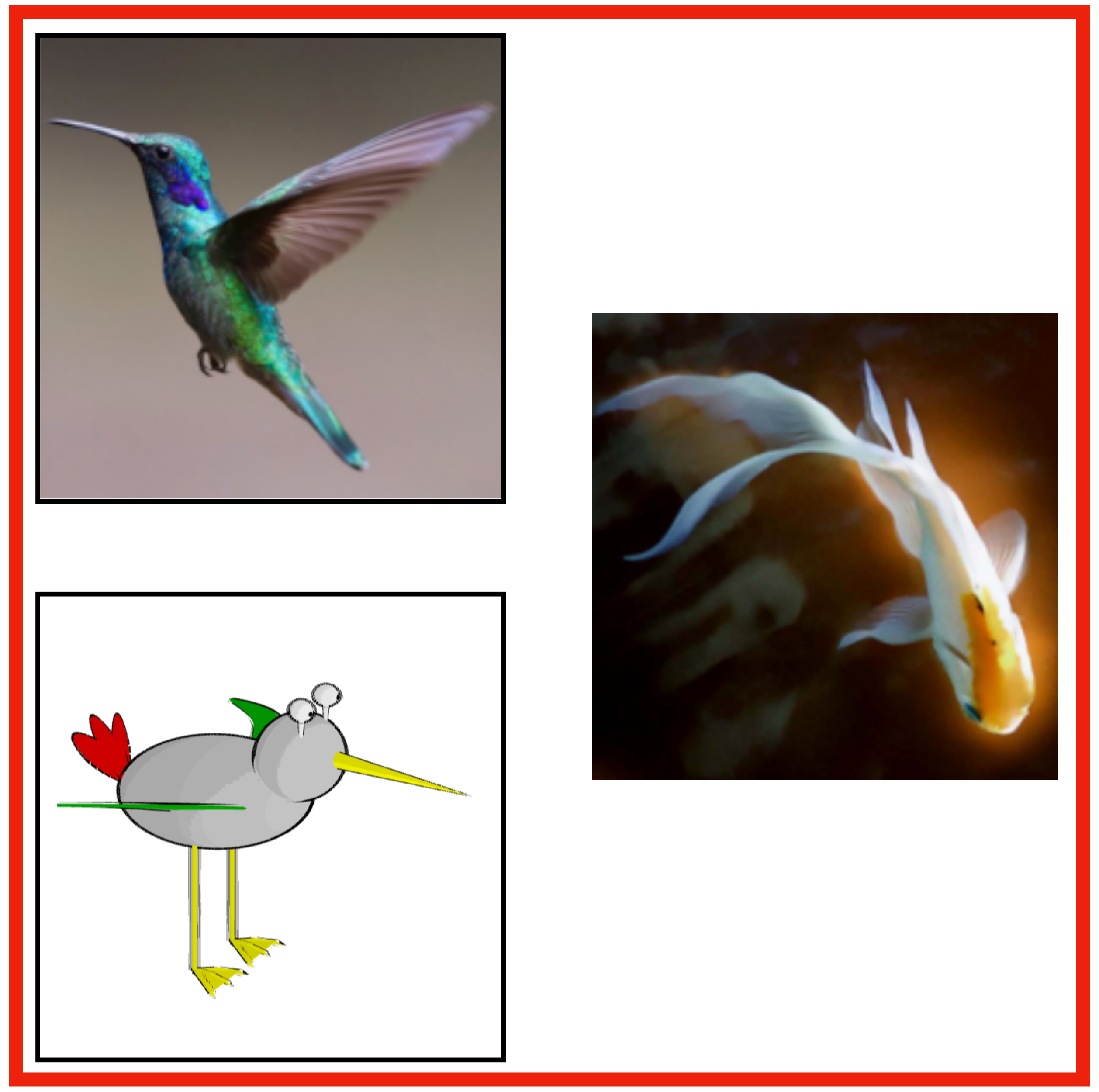


0.9 — Goldfish

[1] Selvaraju et al. (2017). “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: ICCV
[4] Samek et al. (2017). “Evaluating the visualization of what a deep neural network has learned.” In: IEEE Trans. Neural Networks Learn. Syst.
[5] Hesse et al. (2023). “FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods.” In: ICCV

# Related work

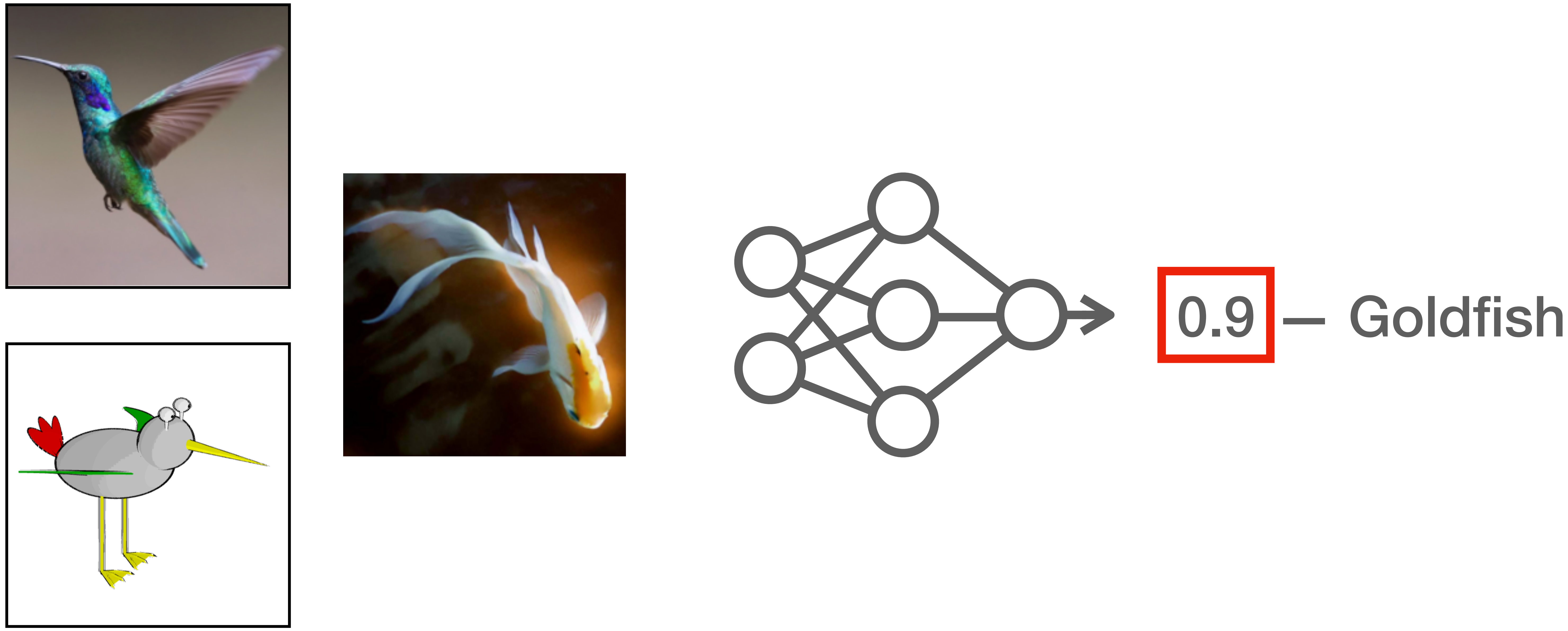
## …and its limitations




[1] Selvaraju et al. (2017). "Grad-CAM: Visual explanations from deep networks via gradient-based localization." In: ICCV
[4] Samek et al. (2017). "Evaluating the visualization of what a deep neural network has learned." In: IEEE Trans. Neural Networks Learn. Syst.
[5] Hesse et al. (2023). "FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods." In: ICCV

# Related work

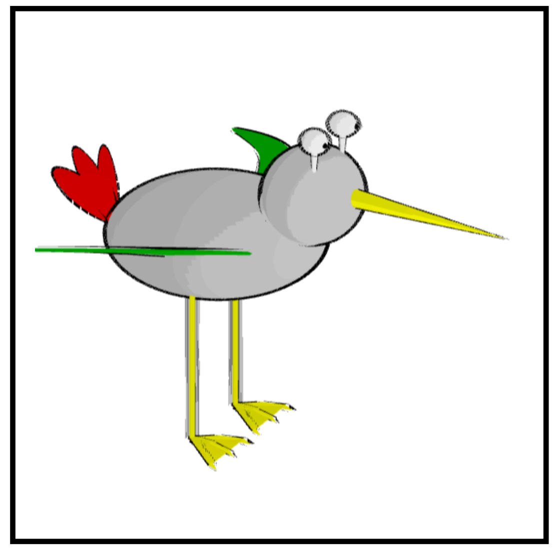
## …and its limitations



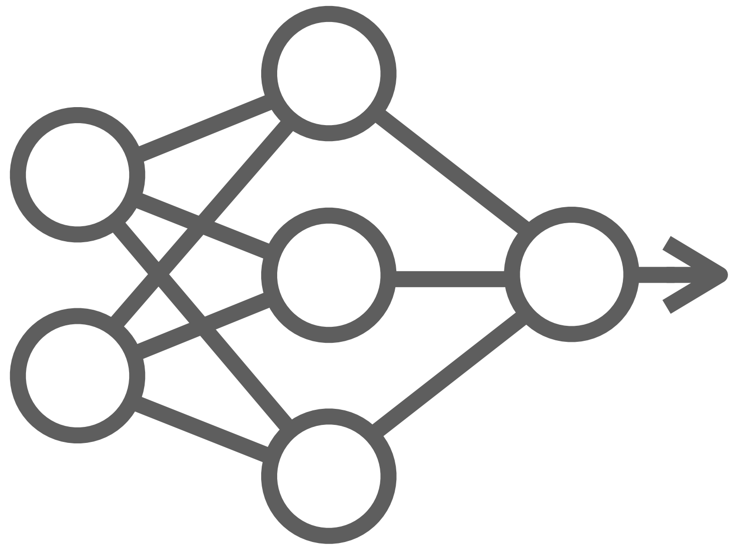


[1] Selvaraju et al. (2017). "Grad-CAM: Visual explanations from deep networks via gradient-based localization." In: ICCV
[4] Samek et al. (2017). "Evaluating the visualization of what a deep neural network has learned." In: IEEE Trans. Neural Networks Learn. Syst.
[5] Hesse et al. (2023). "FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods." In: ICCV

# Related work
## …and its limitations



[1]



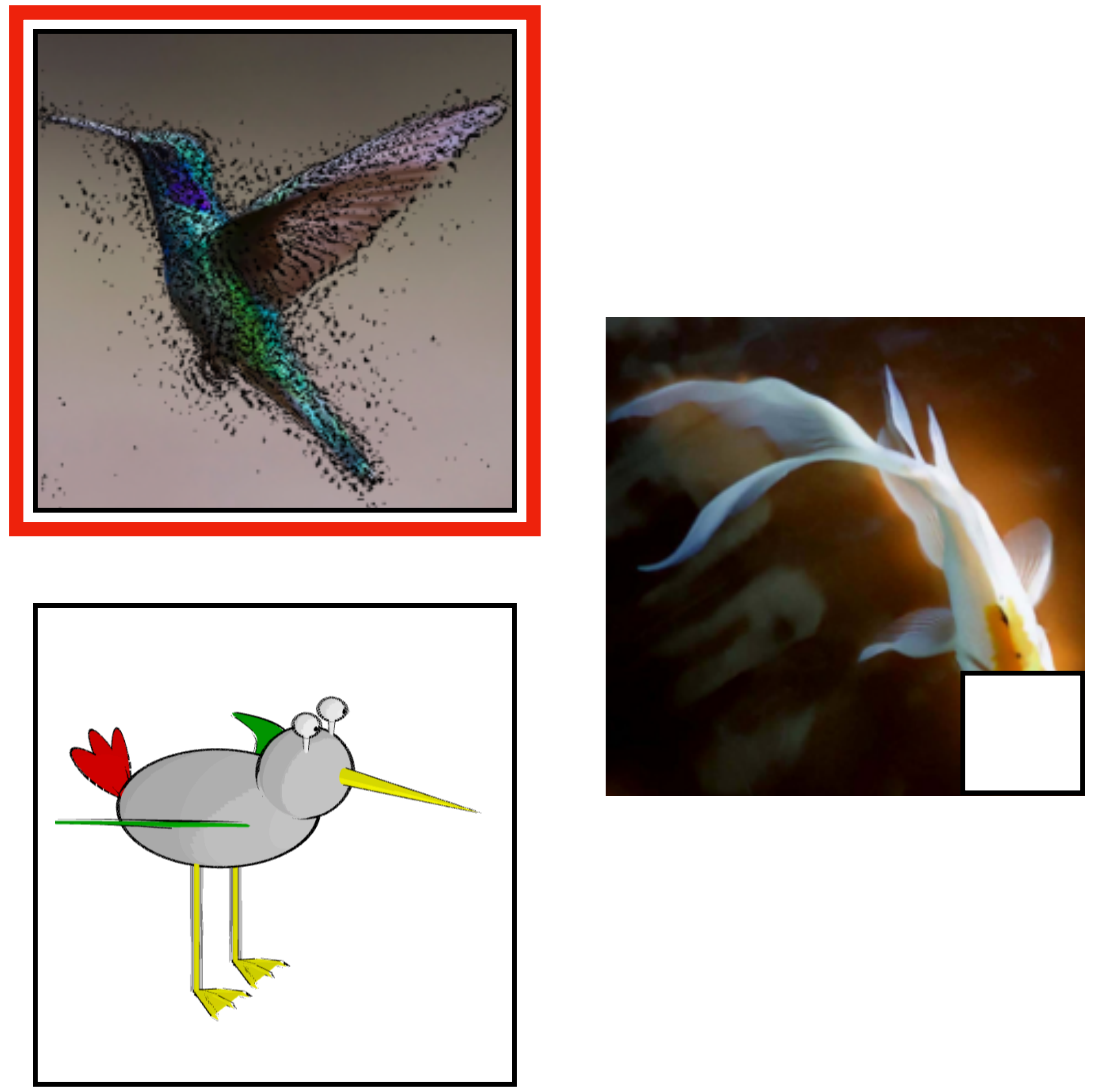
— Goldfish

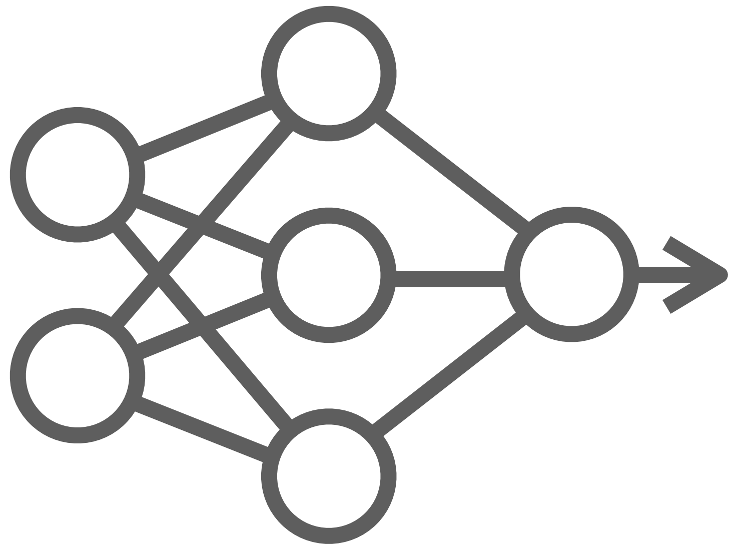
[1] Selvaraju et al. (2017). “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: ICCV
[4] Samek et al. (2017). “Evaluating the visualization of what a deep neural network has learned.” In: IEEE Trans. Neural Networks Learn. Syst.
[5] Hesse et al. (2023). “FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods.” In: ICCV

# Related work

## …and its limitations

[4]



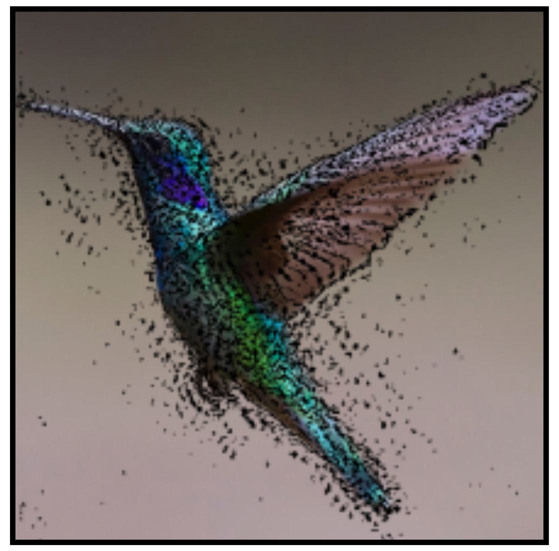


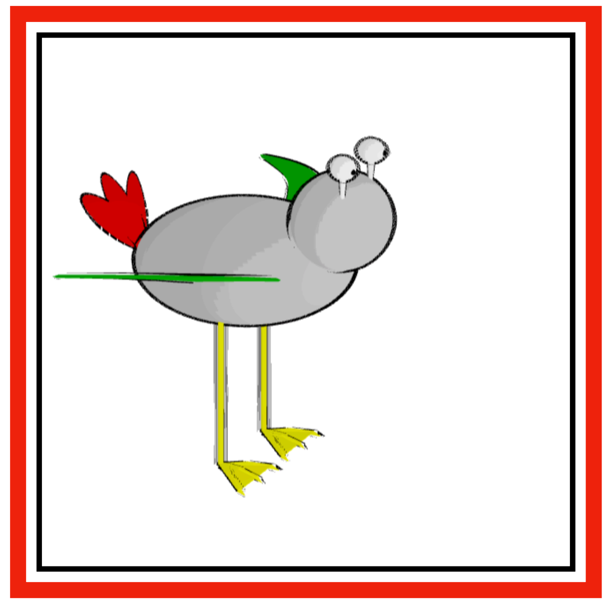
— Goldfish

[1] Selvaraju et al. (2017). “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: ICCV
[4] Samek et al. (2017). “Evaluating the visualization of what a deep neural network has learned.” In: IEEE Trans. Neural Networks Learn. Syst.
[5] Hesse et al. (2023). “FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods.” In: ICCV

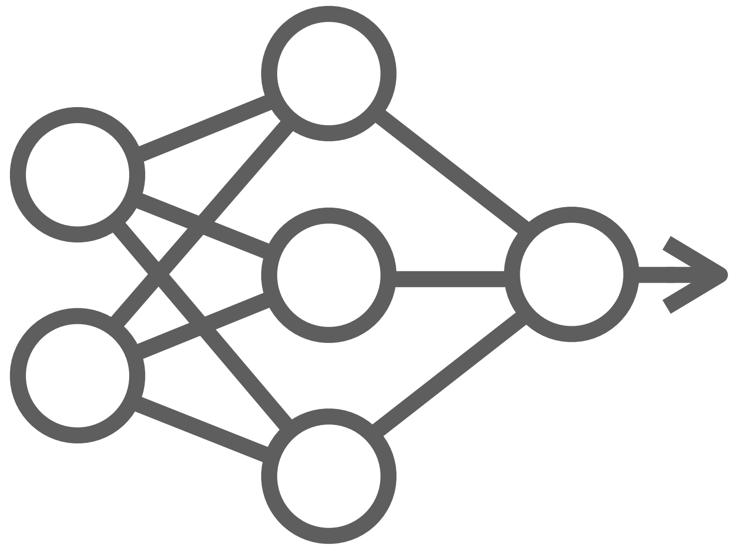
# Related work

## …and its limitations





[5]

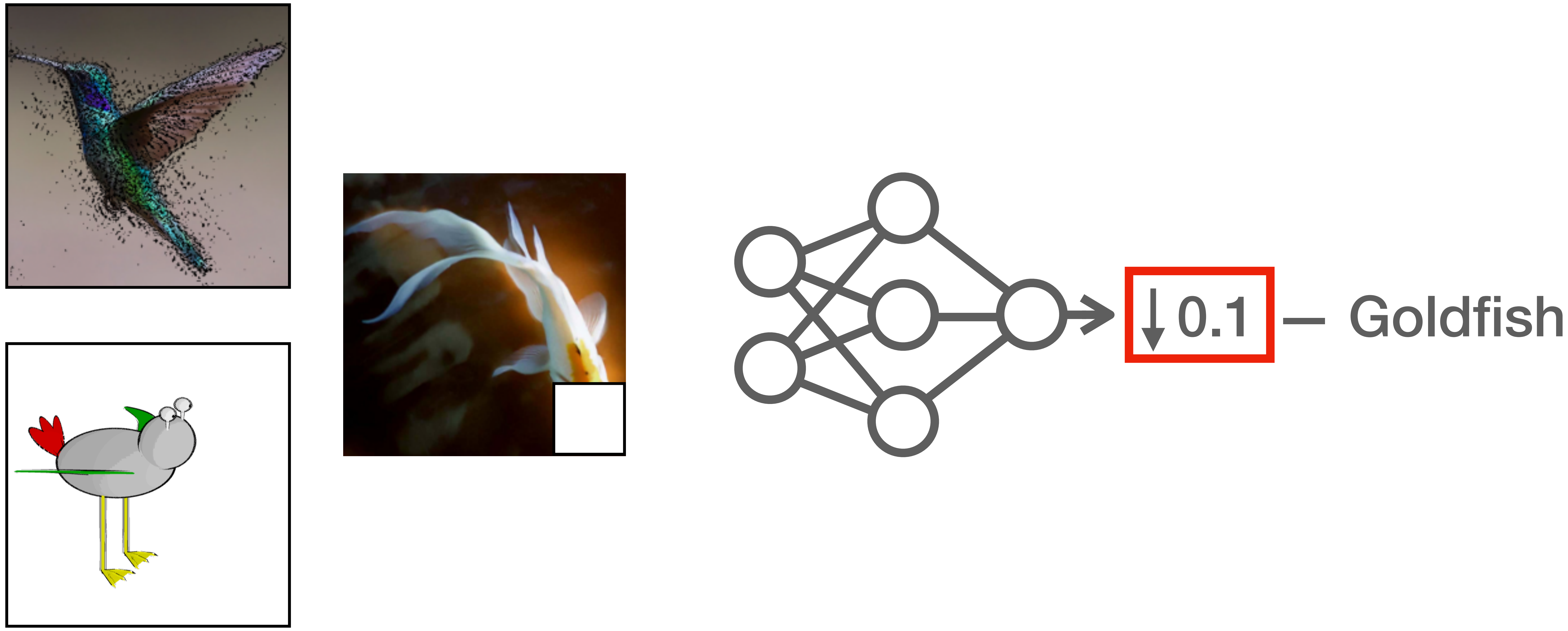


– Goldfish

[1] Selvaraju et al. (2017). “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: ICCV
[4] Samek et al. (2017). “Evaluating the visualization of what a deep neural network has learned.” In: IEEE Trans. Neural Networks Learn. Syst.
[5] Hesse et al. (2023). “FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods.” In: ICCV

# Related work

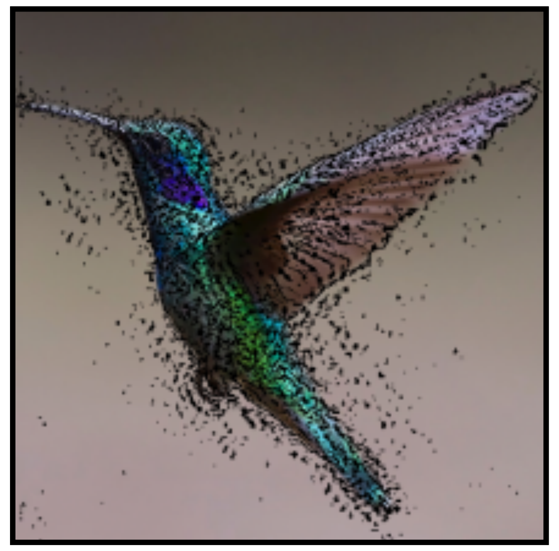
## …and its limitations

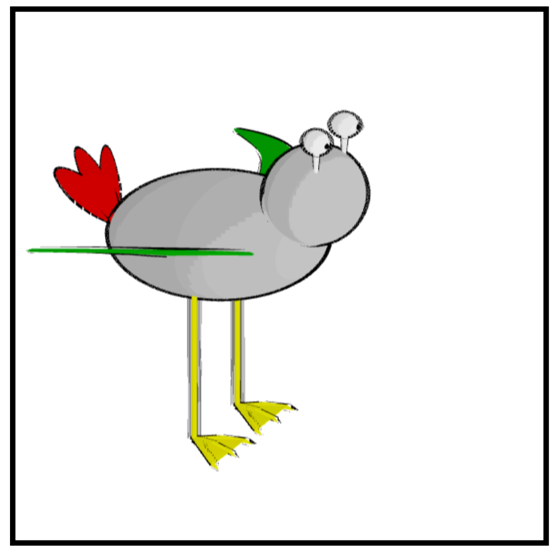


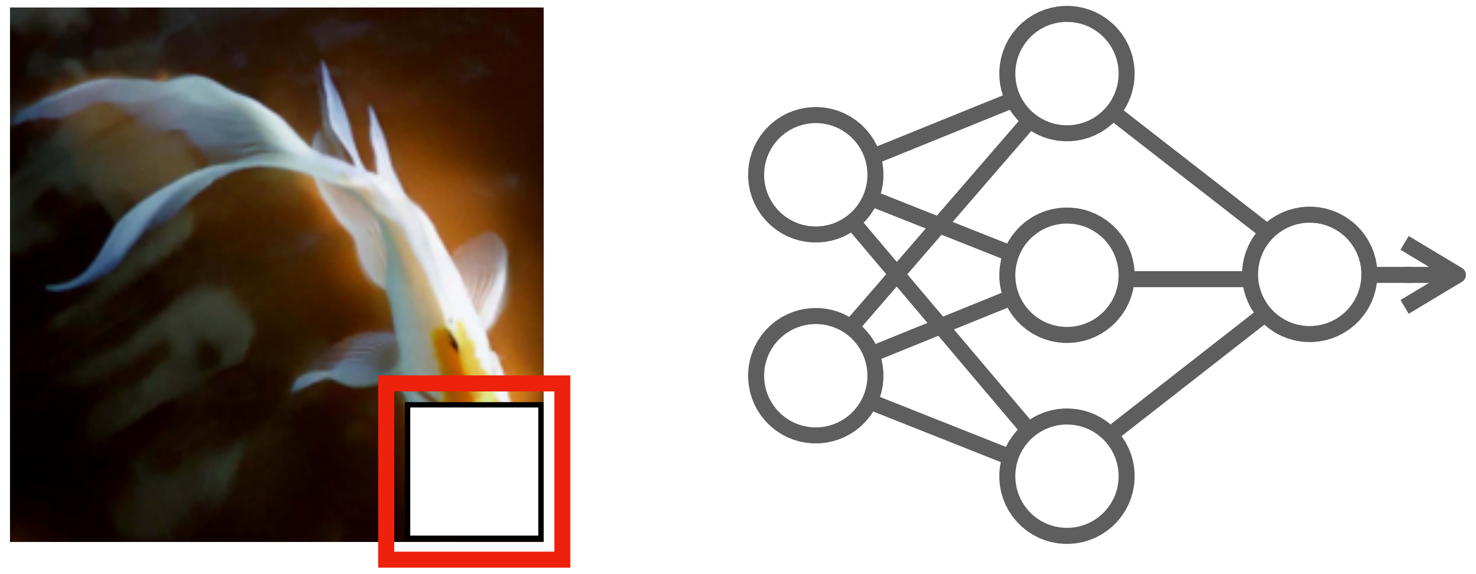


[1] Selvaraju et al. (2017). “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: ICCV
[4] Samek et al. (2017). “Evaluating the visualization of what a deep neural network has learned.” In: IEEE Trans. Neural Networks Learn. Syst.
[5] Hesse et al. (2023). “FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods.” In: ICCV

# Related work
## …and its limitations





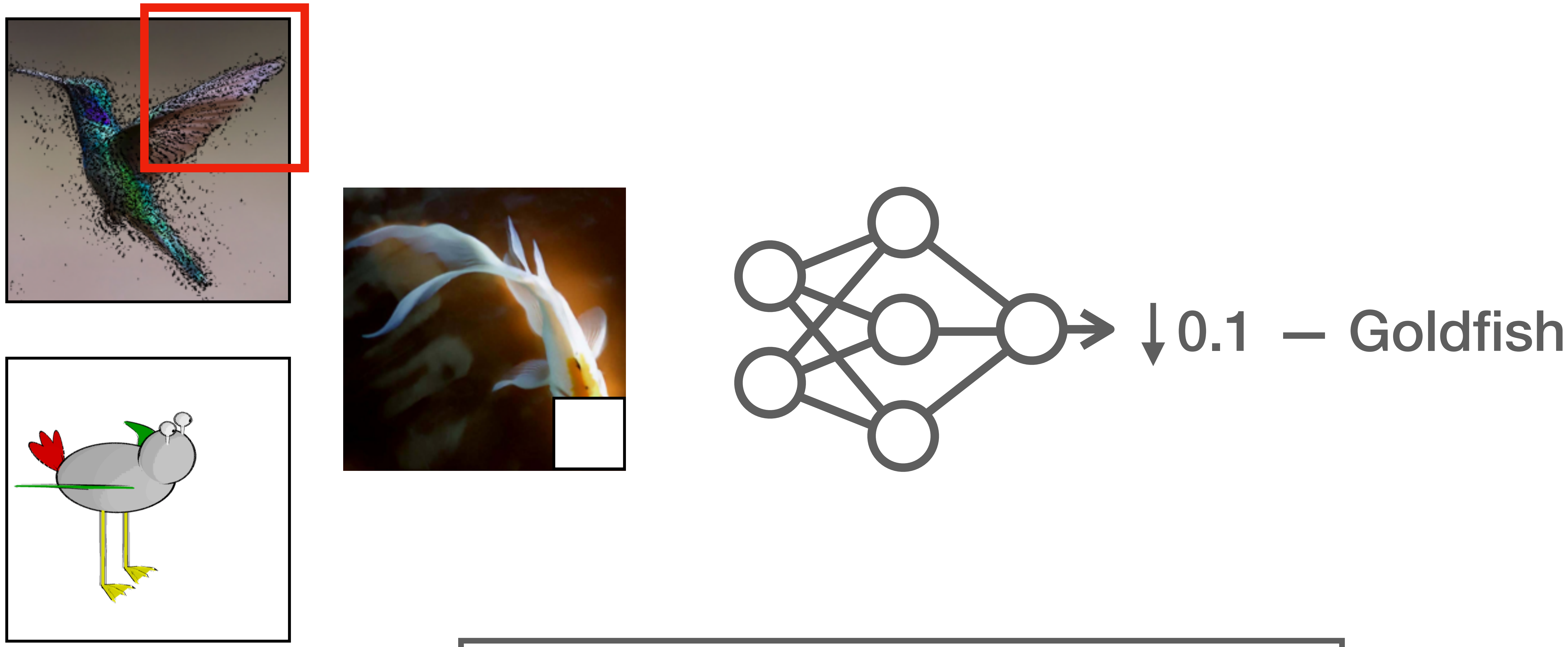



→ Out-of-domain issues

[1] Selvaraju et al. (2017). “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: ICCV
[4] Samek et al. (2017). “Evaluating the visualization of what a deep neural network has learned.” In: IEEE Trans. Neural Networks Learn. Syst.
[5] Hesse et al. (2023). “FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods.” In: ICCV

# Related work

## …and its limitations




→ Out-of-domain issues

→ Information leakage

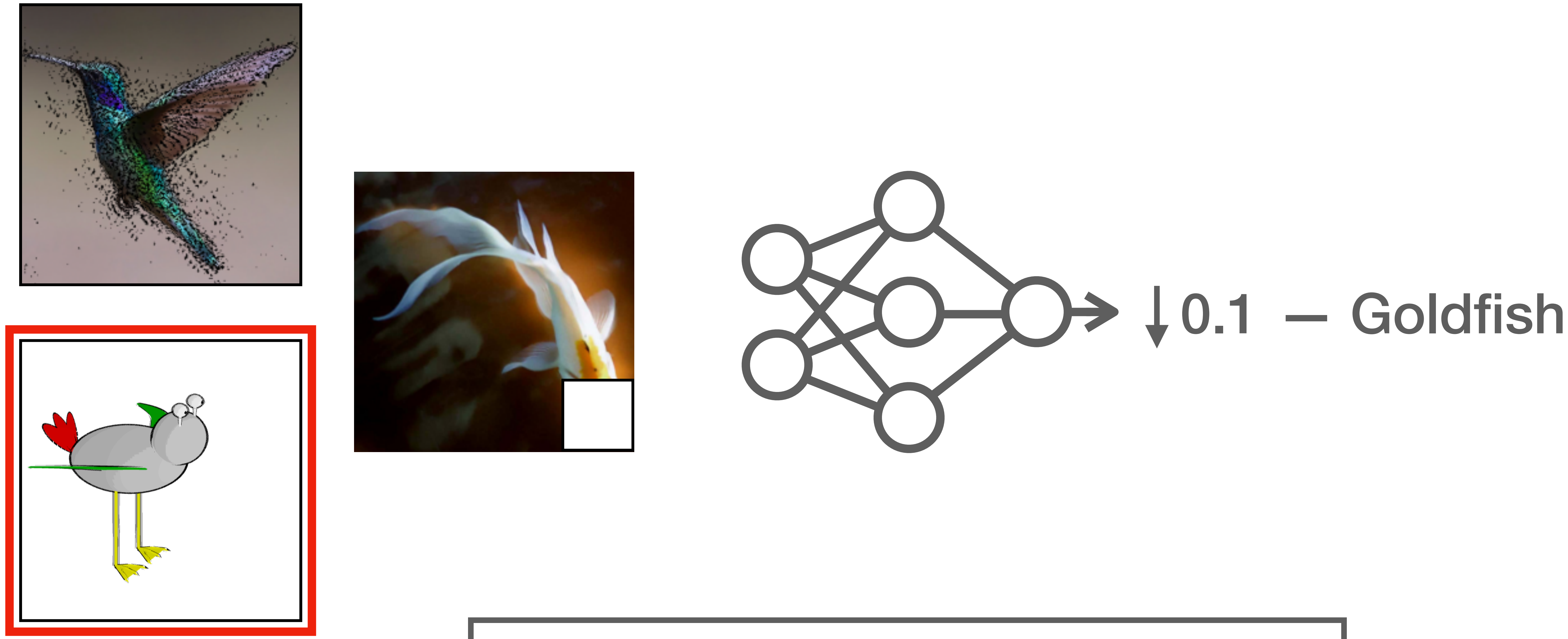

[1] Selvaraju et al. (2017). “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: ICCV

[4] Samek et al. (2017). “Evaluating the visualization of what a deep neural network has learned.” In: IEEE Trans. Neural Networks Learn. Syst.

[5] Hesse et al. (2023). “FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods.” In: ICCV

# Related work

## …and its limitations




→ Out-of-domain issues
→ Information leakage
→ Synthetic data

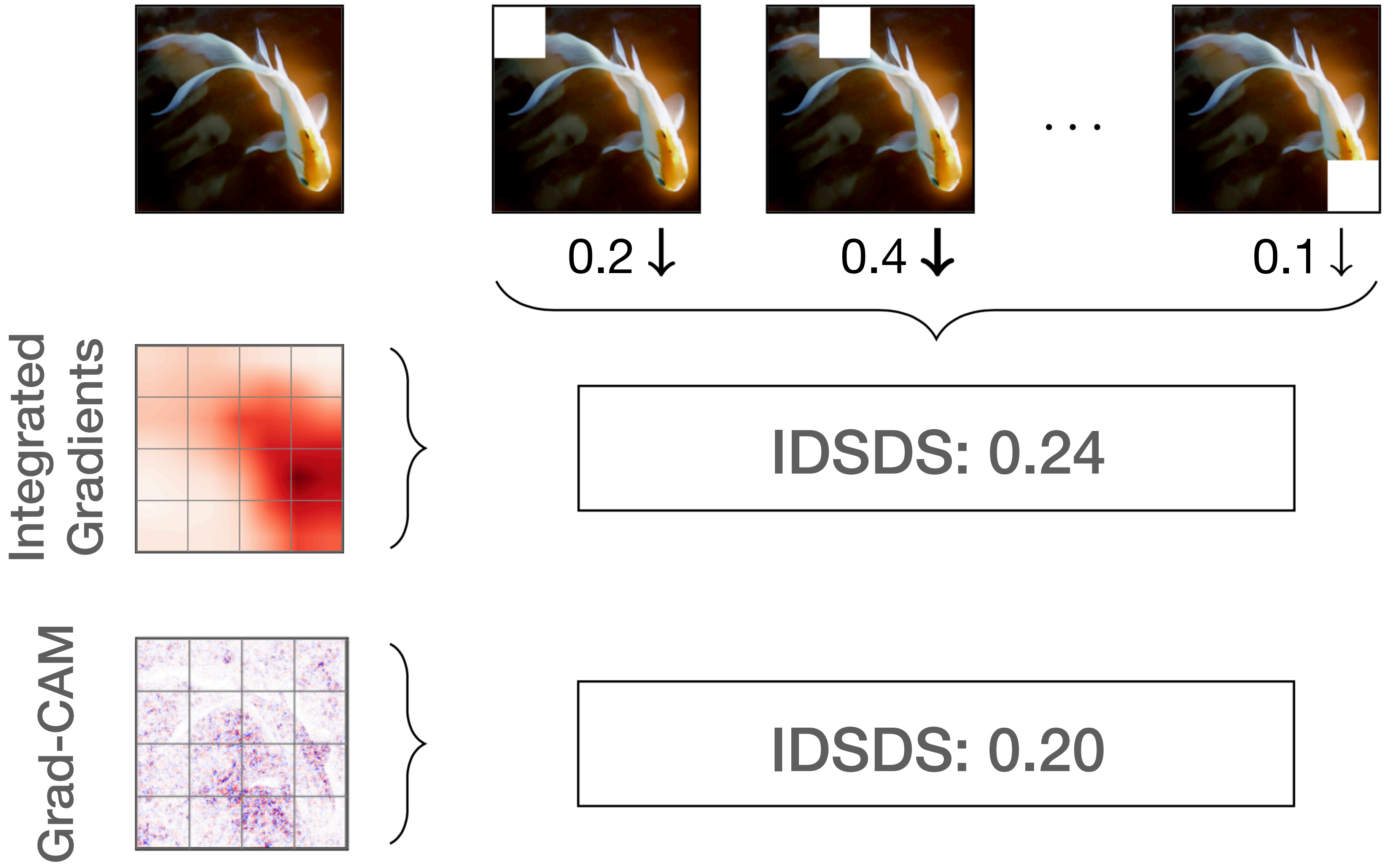

[1] Selvaraju et al. (2017). “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: ICCV
[4] Samek et al. (2017). “Evaluating the visualization of what a deep neural network has learned.” In: IEEE Trans. Neural Networks Learn. Syst.
[5] Hesse et al. (2023). “FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods.” In: ICCV

# In-domain single deletion score (IDSDS)

1. **Train the model on images with deleted patches**

# In-domain single deletion score (IDSDS)

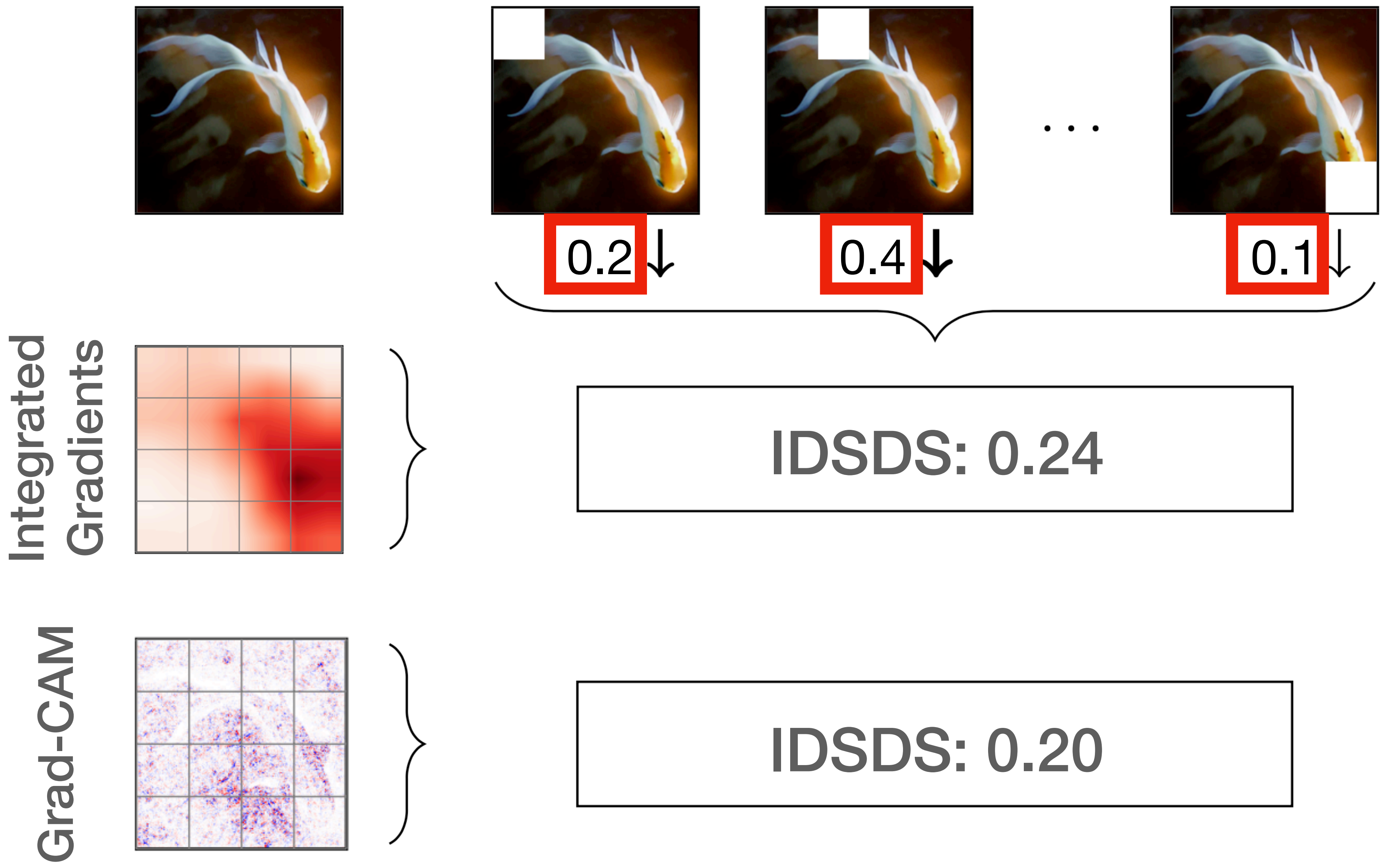
1. **Train the model on images with deleted patches**
2. **Rank correlation between output drops and attribution strength for each patch**



# In-domain single deletion score (IDSDS)

1. Train the model on images with deleted patches

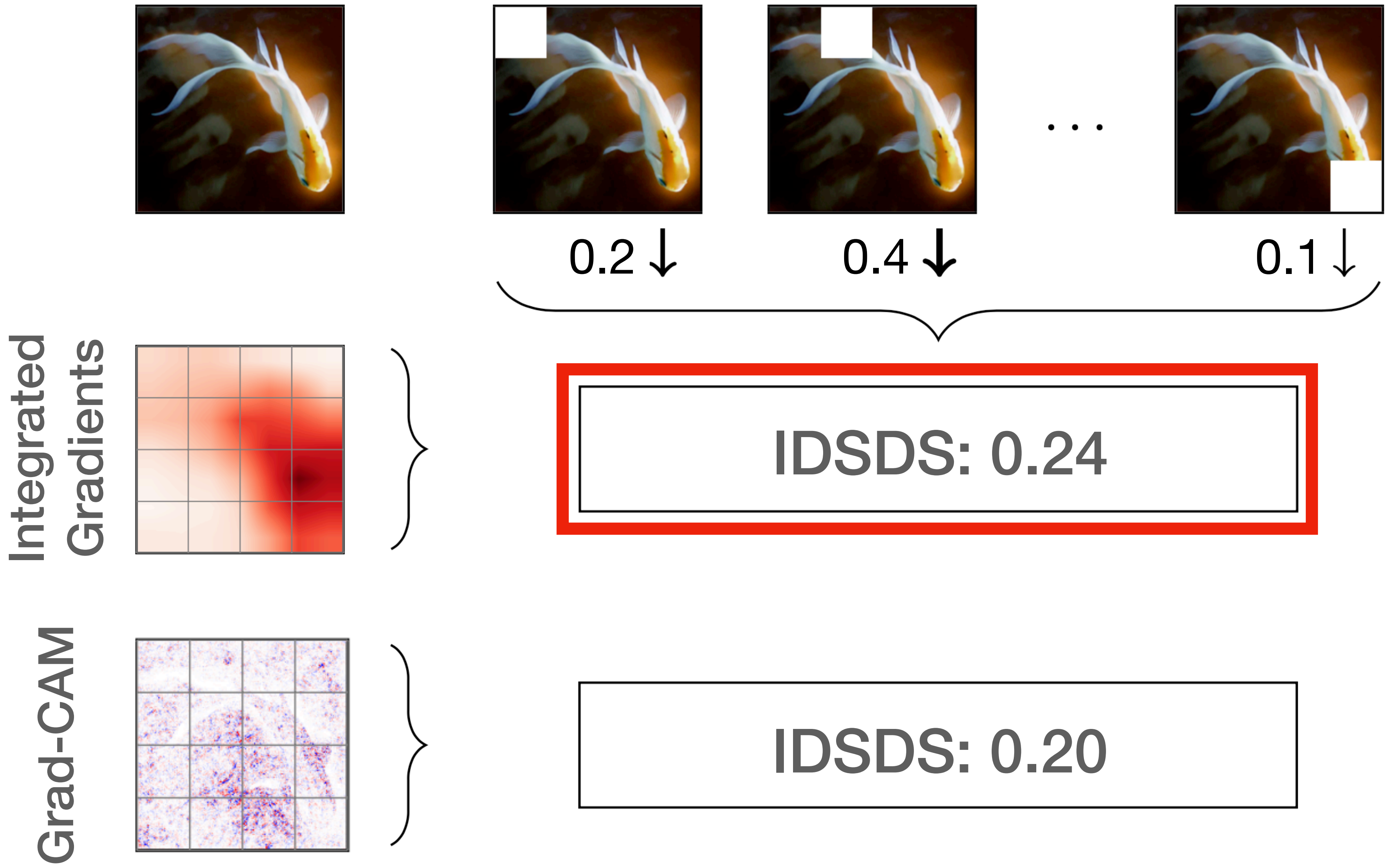
2. Rank correlation between output drops and attribution strength for each patch



# In-domain single deletion score (IDSDS)

1. Train the model on images with deleted patches

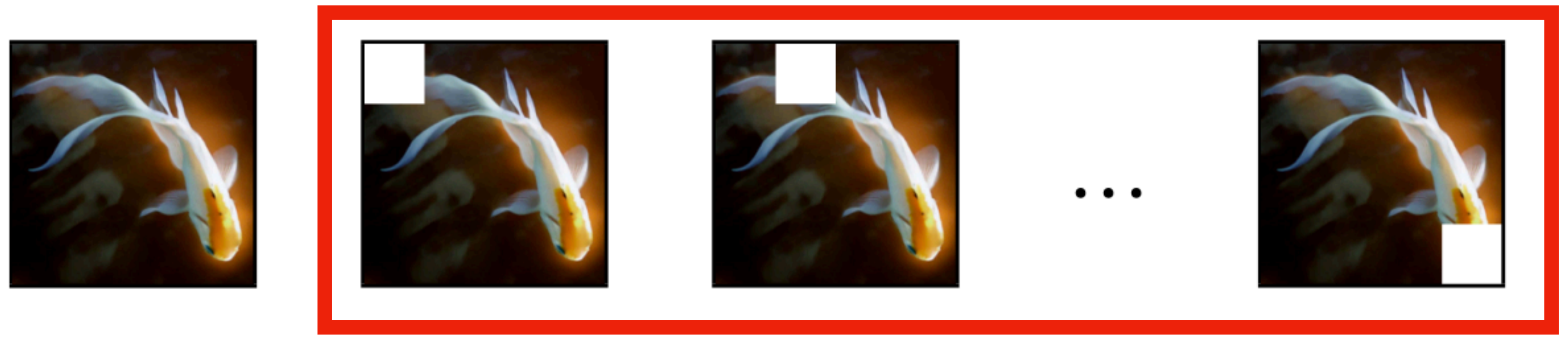

2. Rank correlation between output drops and attribution strength for each patch

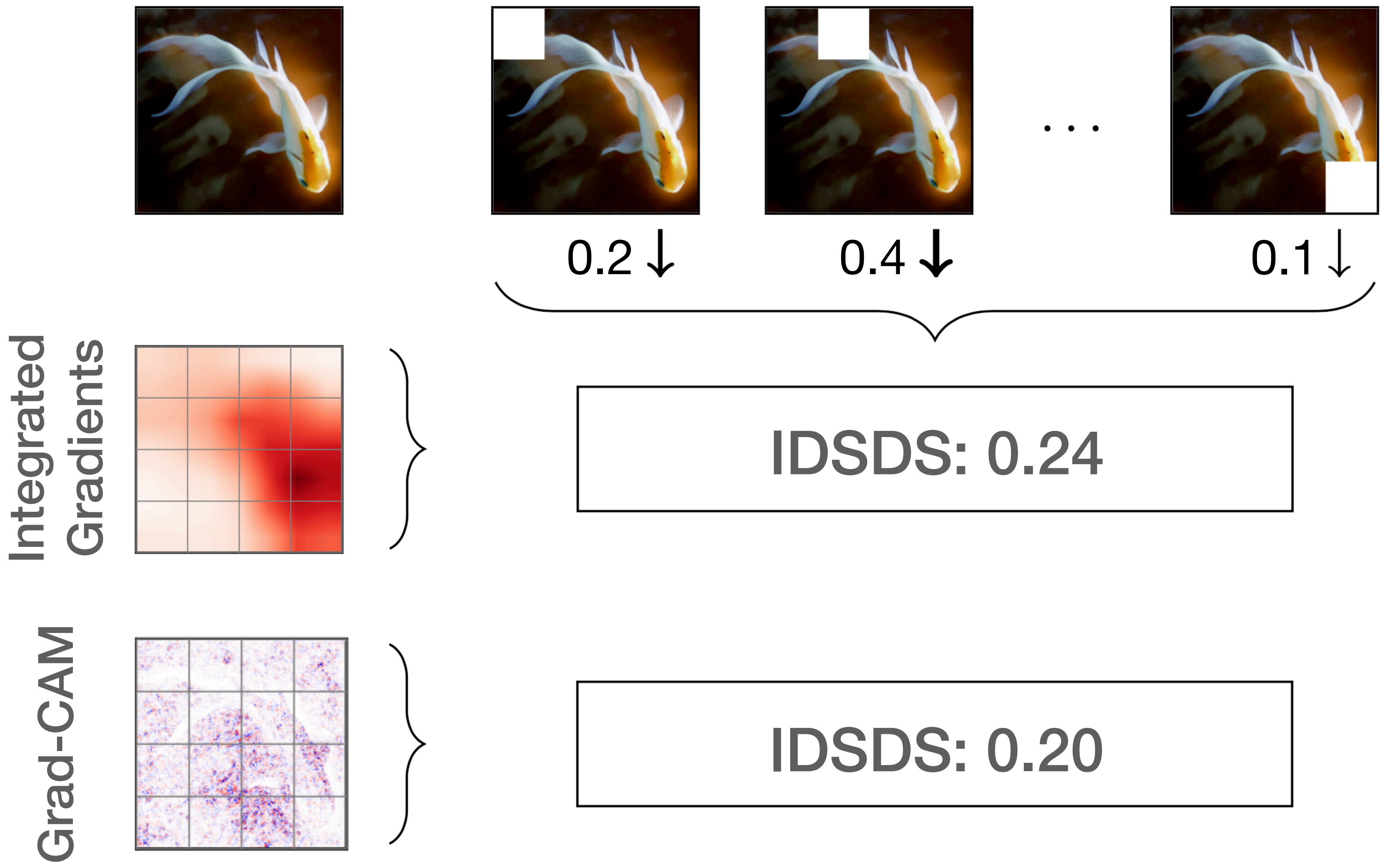


# In-domain single deletion score (IDSDS)

1. Train the model on images with deleted patches

2. Rank correlation between output drops and attribution strength for each patch



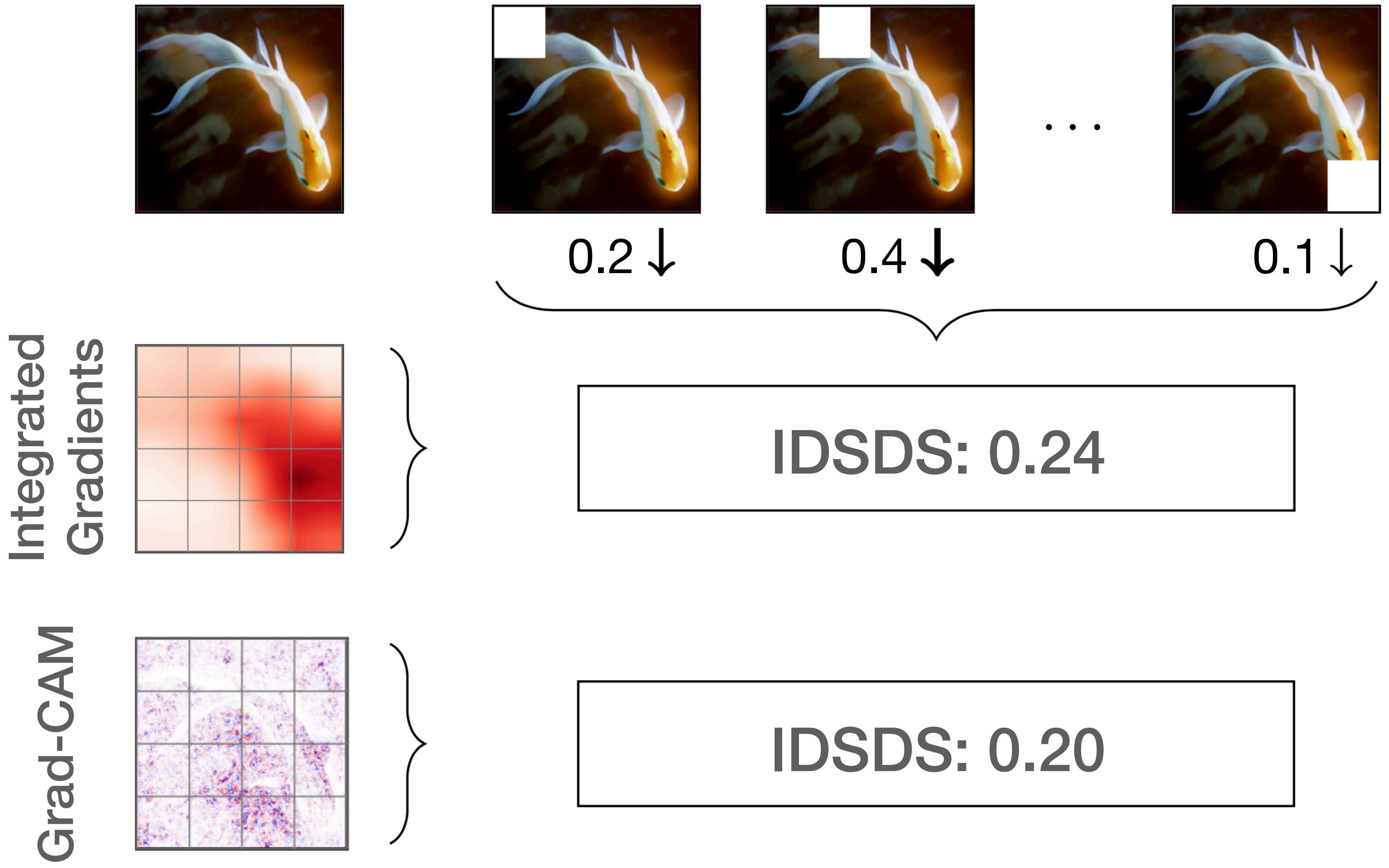



→ Aligned train and test domains

# In-domain single deletion score (IDSDS)

1. Train the model on images with deleted patches

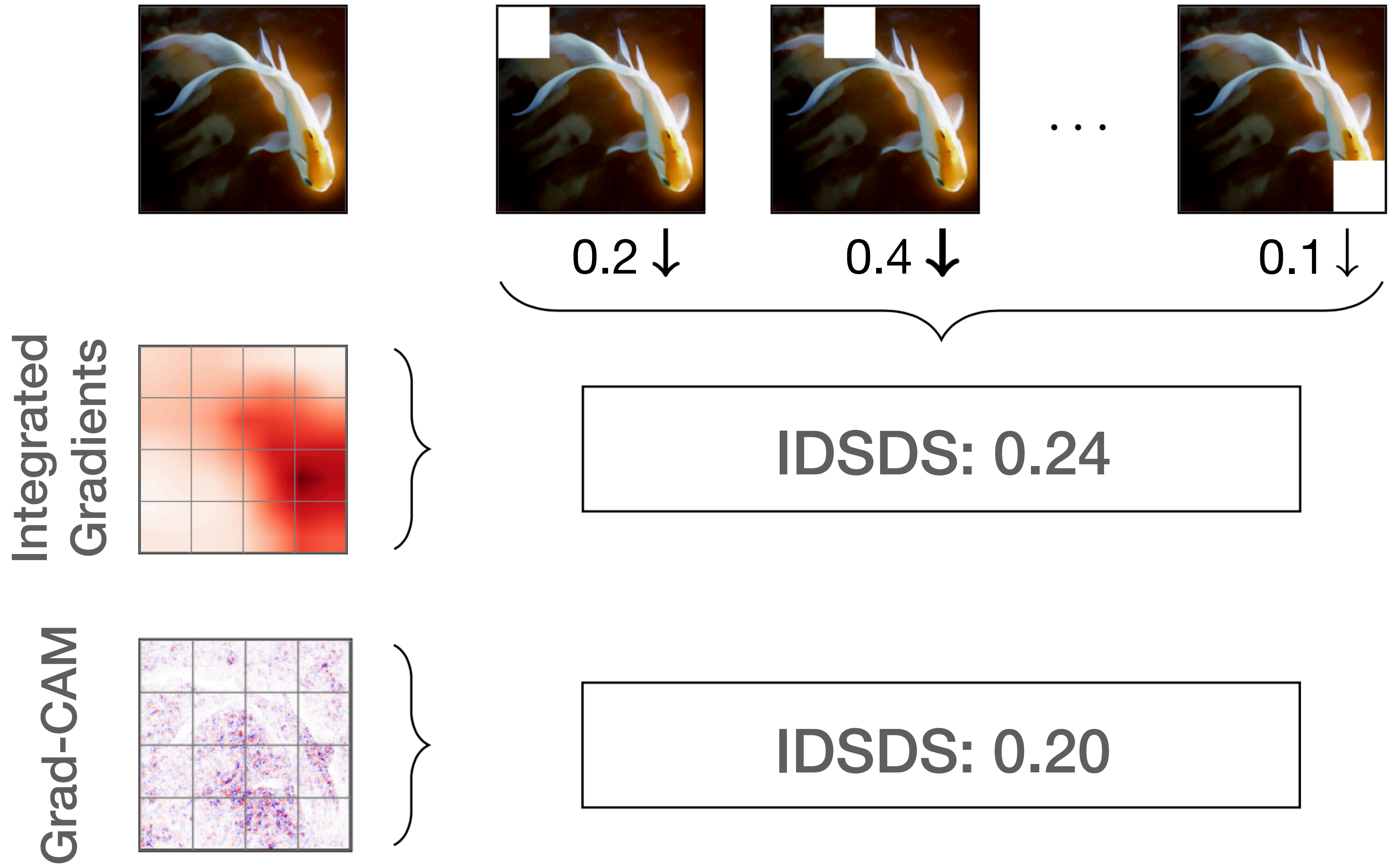
2. Rank correlation between output drops and attribution strength for each patch




→ Aligned train and test domains
→ Provably no information leakage

# In-domain single deletion score (IDSDS)

1. Train the model on images with deleted patches
2. Rank correlation between output drops and attribution strength for each patch

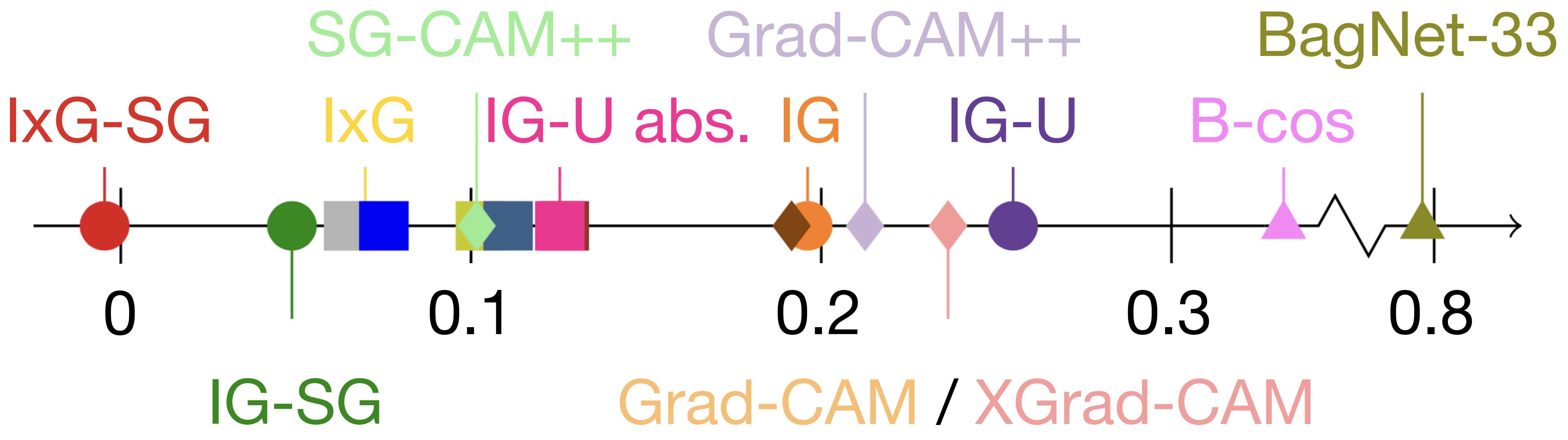



→ Aligned train and test domains
→ Provably no information leakage
→ Allows for inter-model comparison

# Results
## Ranking attribution methods

→ **Taking the absolute attributions (abs.) impairs performance**

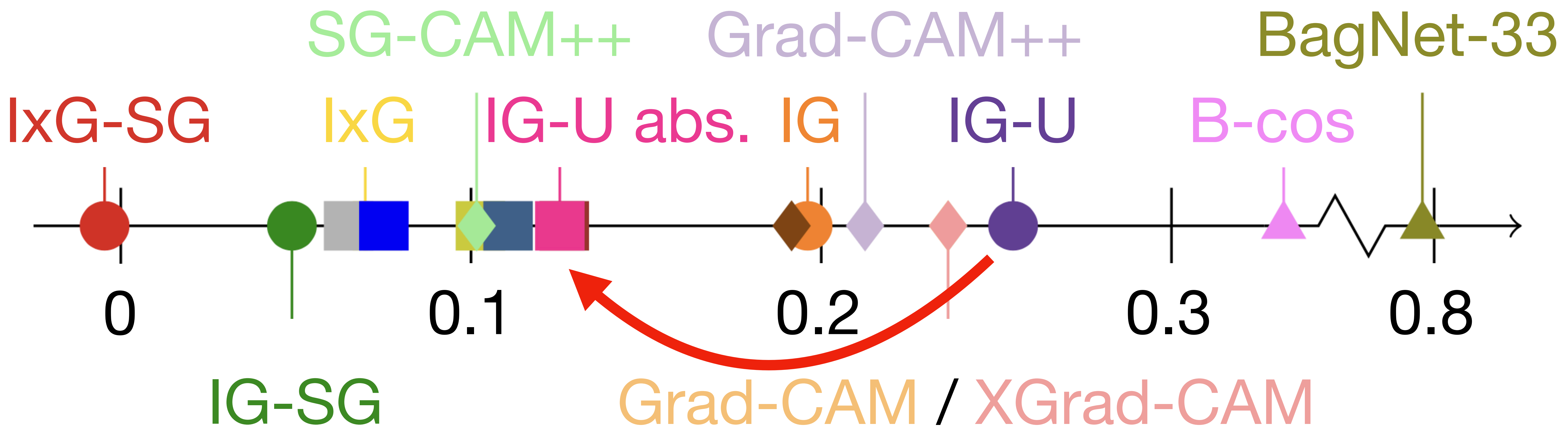
→ Intrinsically explainable models (▲) achieve the best results



# Results
## Ranking attribution methods

→ **Taking the absolute attributions (abs.) impairs performance**

→ Intrinsically explainable models (▲) achieve the best results

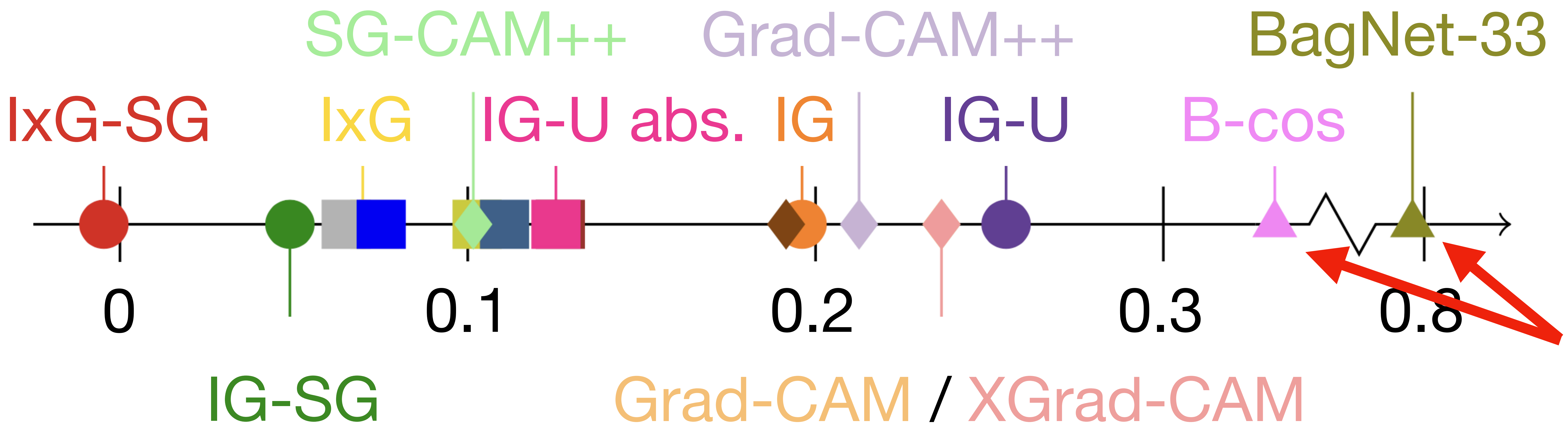


# Results

## Ranking attribution methods

→ Taking the absolute attributions (abs.) impairs performance

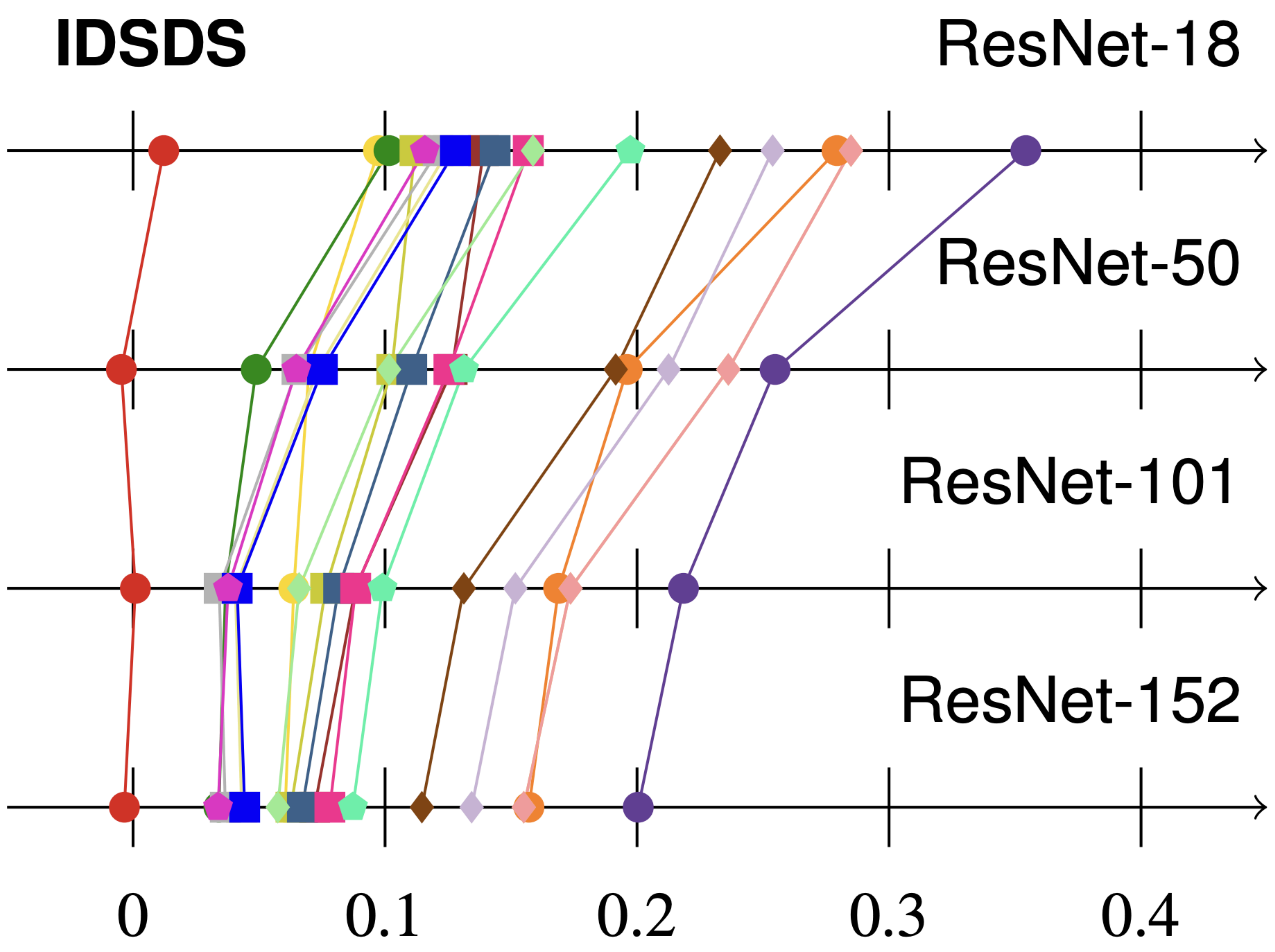
→ **Intrinsically explainable models (▲) achieve the best results**



# Results

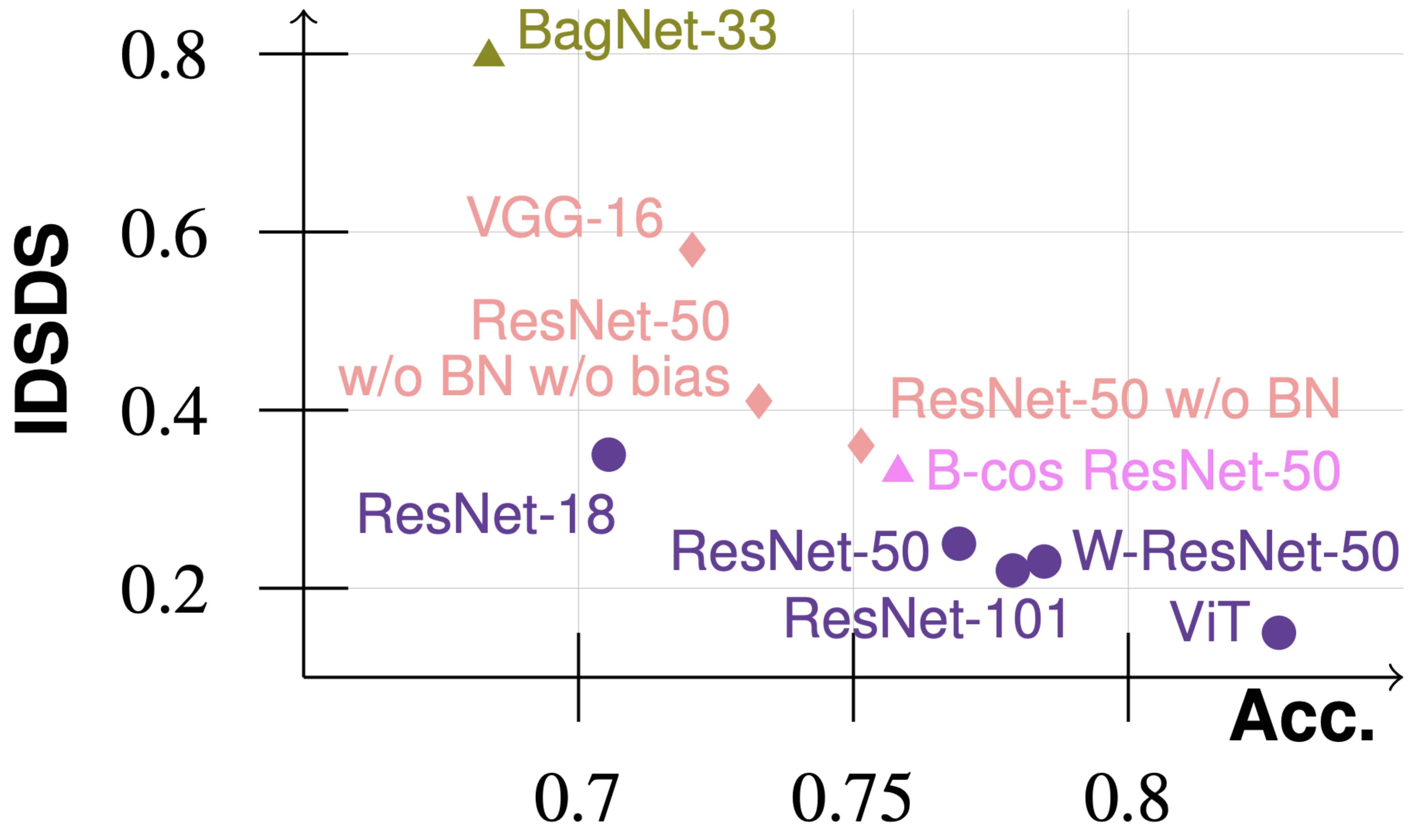
## How design choices affect attribution quality

→ Deeper models have lower attribution quality



# Results
## How design choices affect attribution quality

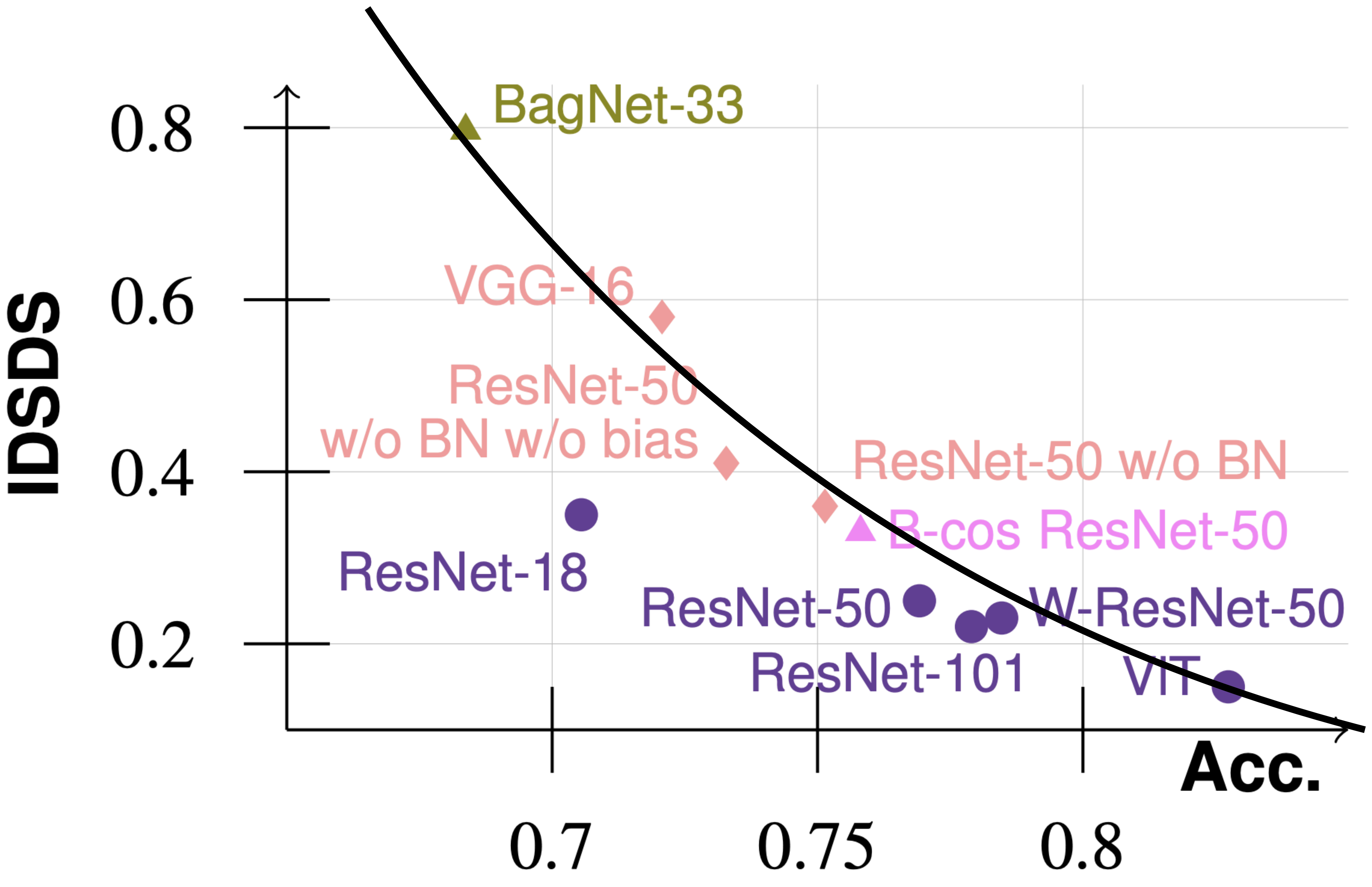
→ There is an accuracy-attribution quality tradeoff



# Results

## How design choices affect attribution quality

→ There is an accuracy-attribution quality tradeoff



TECHNISCHE UNIVERSITÄT DARMSTADT

hessian.AI

NEURAL INFORMATION PROCESSING SYSTEMS

European Research Council
Established by the European Commission

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008). The project has also been supported in part by the State of Hesse through the cluster projects "The Third Wave of Artificial Intelligence (3AI)" and "The Adaptive Mind (TAM)".

# Benchmarking the Attribution Quality of Vision Models



Robin Hesse | Simone Schaub-Meyer | Stefan Roth

Visual Inference Lab | TU Darmstadt

Project page

https://github.com/visinf/idsds