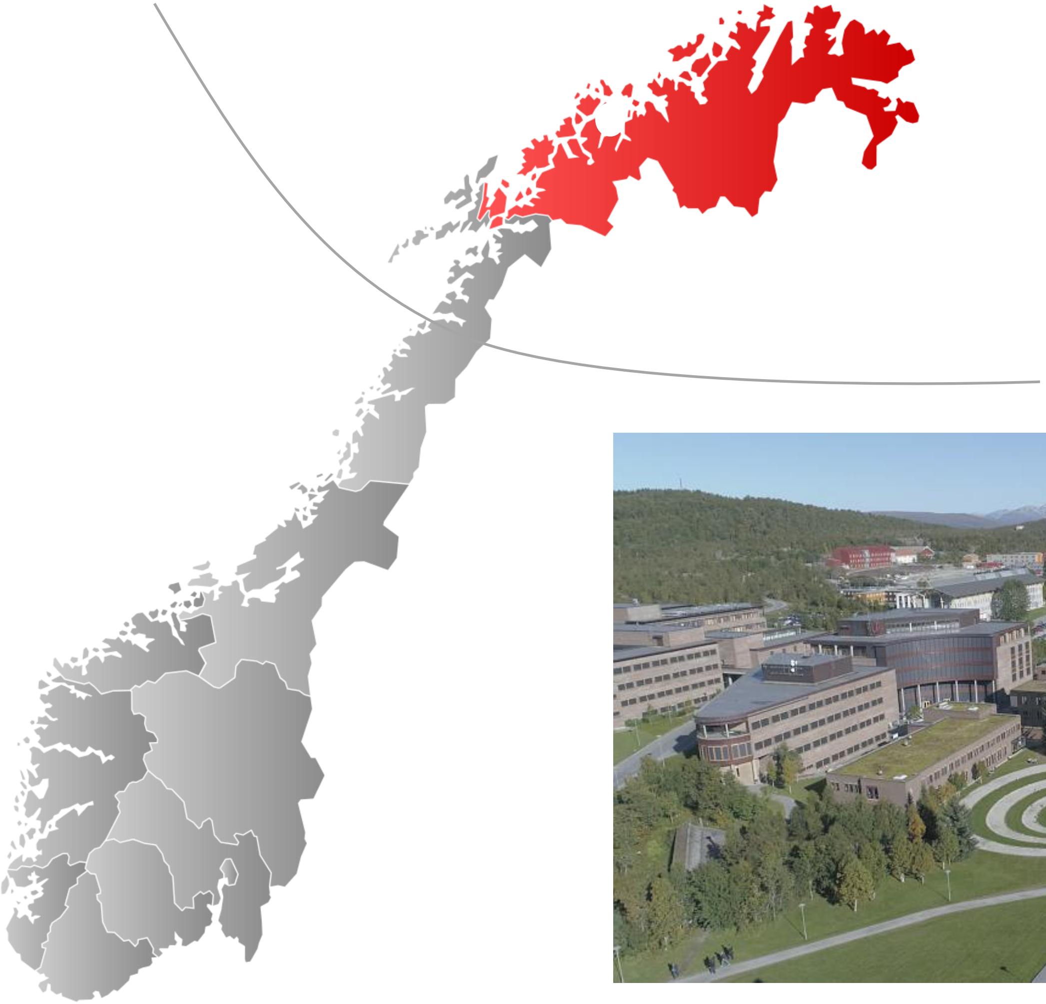
UiT The Arctic University of Norway

From Post-hoc Explainability to Self-Explainable Models

Michael Kampffmeyer
Machine Learning Group
UiT The Arctic University of Norway
machine-learning.uit.no

Where are we?



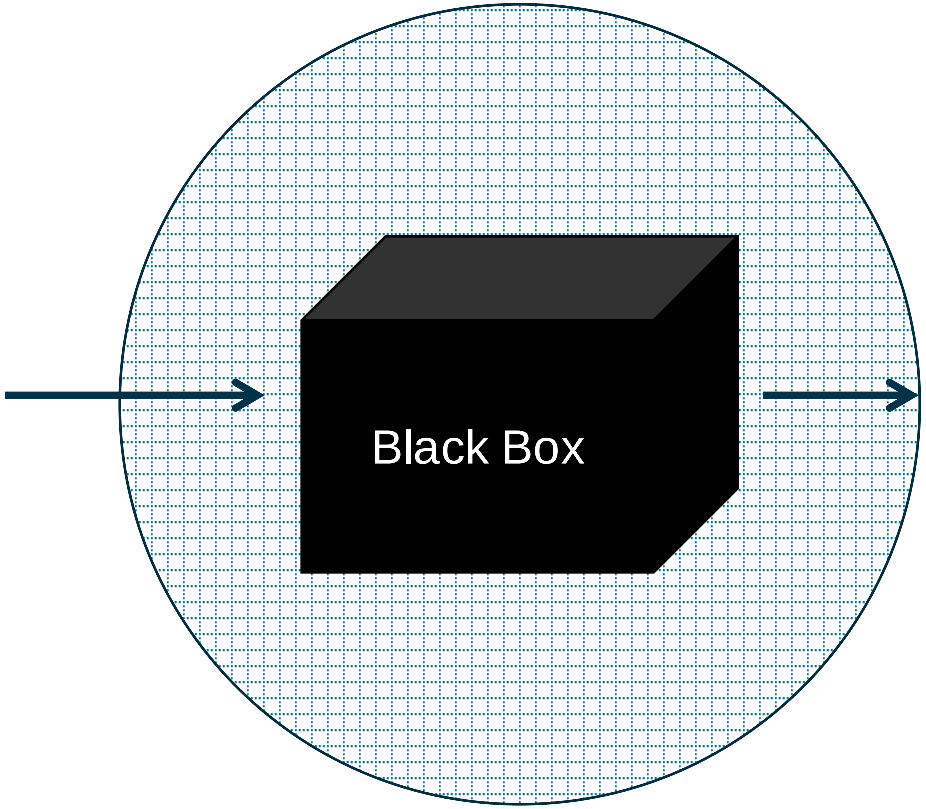
7 faculty/~35 people/machine-learning.uit.no

Need For XAI

- Deep learning models used are mostly black-box models

Suppose this image is classified as pneumonia.

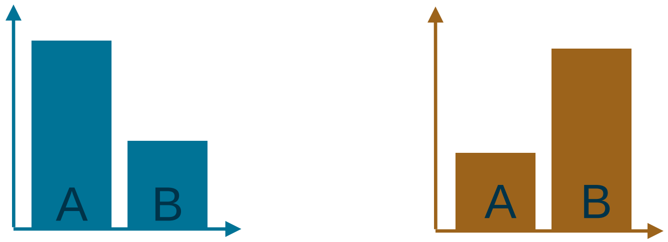
But **why?**



Case-study

- Assumption 1:
  - Multi-source datasets (data scarcity).

- Assumption 2:
  - Class-imbalance from these 2 sources.

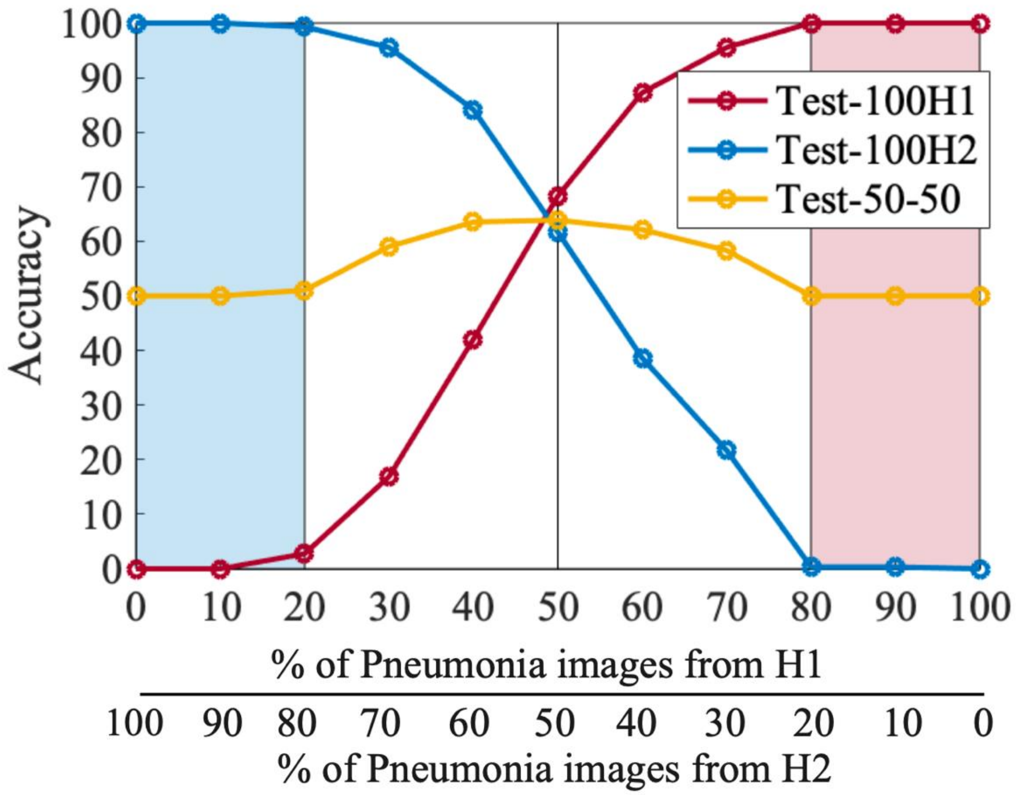


Gautam, et al. "Demonstrating the risk of imbalanced datasets in chest x-ray image-based diagnostics by prototypical relevance propagation." *ISBI 2022*

Case-study

- 2 sources: ChestX-Ray14 (H1) & CheXPert (H2)
  Problem: Pneumonia (P) vs Non-Pneumonia (NP)
- Label-imbalance? : source-related disease imbalance

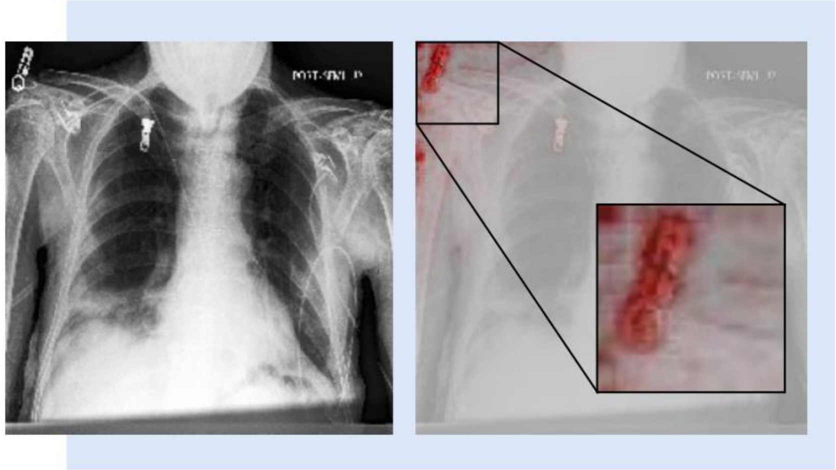
Training	x% P H1 + (100-x)% P H2 (100-x)% NP H1 + x% NP H2	
Testing	Test-100H1	100% Pnuemonia H1 + 100% Non-Pneumonia H2
	Test-100H2	100% Pneumonia H2 + 100% Non-Pneumonia H1
	Test-50-50	50% Pneumonia + 50% Non-Pneumonia from both



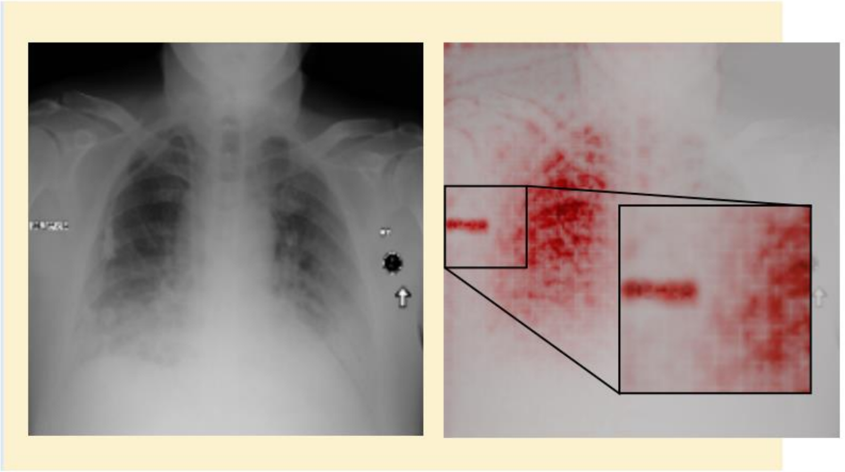
Gautam, et al. "Demonstrating the risk of imbalanced datasets in chest x-ray image-based diagnostics by prototypical relevance propagation." *ISBI 2022*

What is the model looking at?

Acts as hospital classifier
Instead of disease classifier



90H1-10H2



60H1-40H2

Gautam et al. "Demonstrating the risk of imbalanced datasets in chest x-ray image-based diagnostics by prototypical relevance propagation." *ISBI 2022*

Explainability & Interpretability[1]

Post-hoc methods

- Model-agnostic: LIME[2]
- Model-aware: LRP[3]

Self-explaining models

- Aligning latent to known visual concepts[4] : Prototypes

Post-hoc explanations:
Provides prediction first, then why!

Self-explainable models:
Provides prediction and why at the same time!

[1] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019

[2] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015.

[4] Alina Barnett, Jonathan Su, Cynthia Rudin, Chaofan Chen, Oscar Li. This looks like that: Deep learning for interpretable image recognition. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2019.

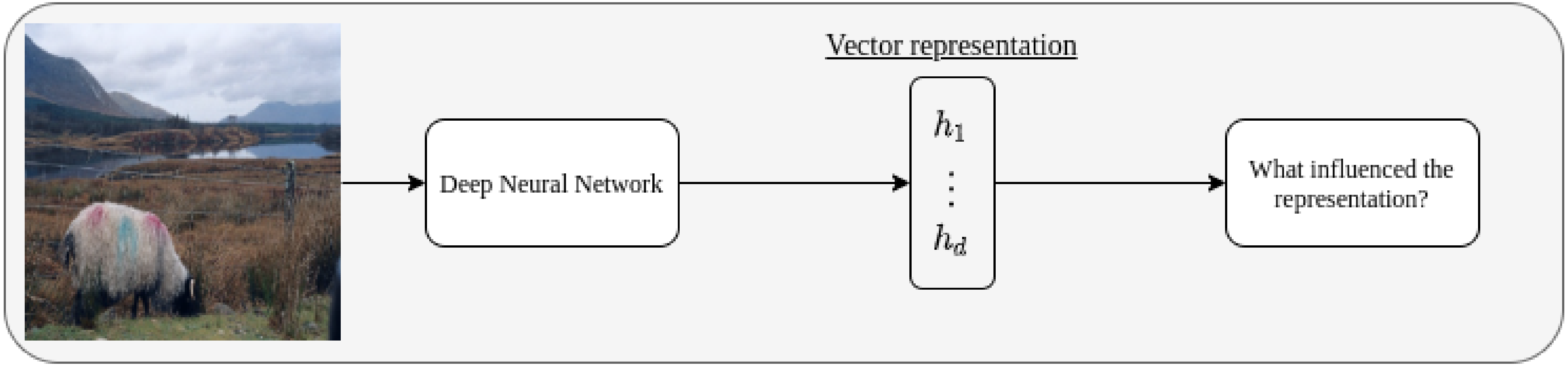
Post-hoc XAI for Representations

Towards Self-Explainable Models

Post-hoc XAI for Representations

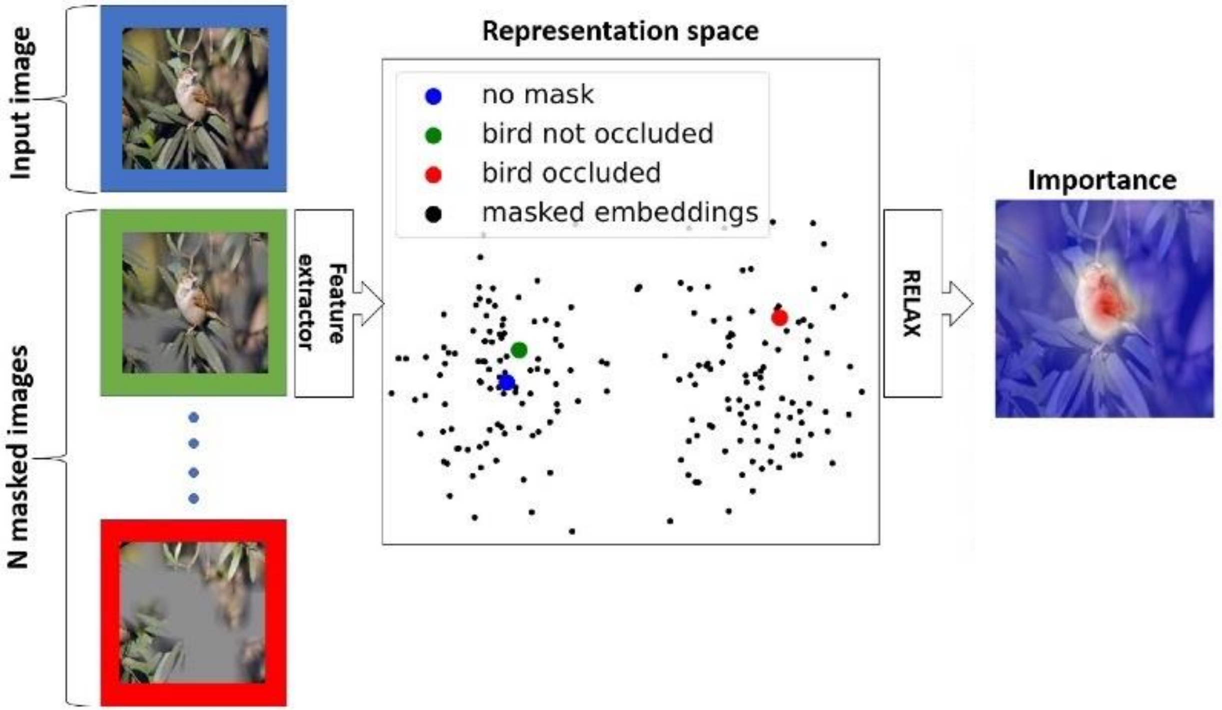
Post-hoc Methods

- Many methods exists to explain predictions.
- How to handle unlabeled vectorial outputs?
- Increasingly important with improvements in representation learning.



RELAX: A representation learning explainability framework

- Key idea: mask out parts of the image and monitor how the representation changes.

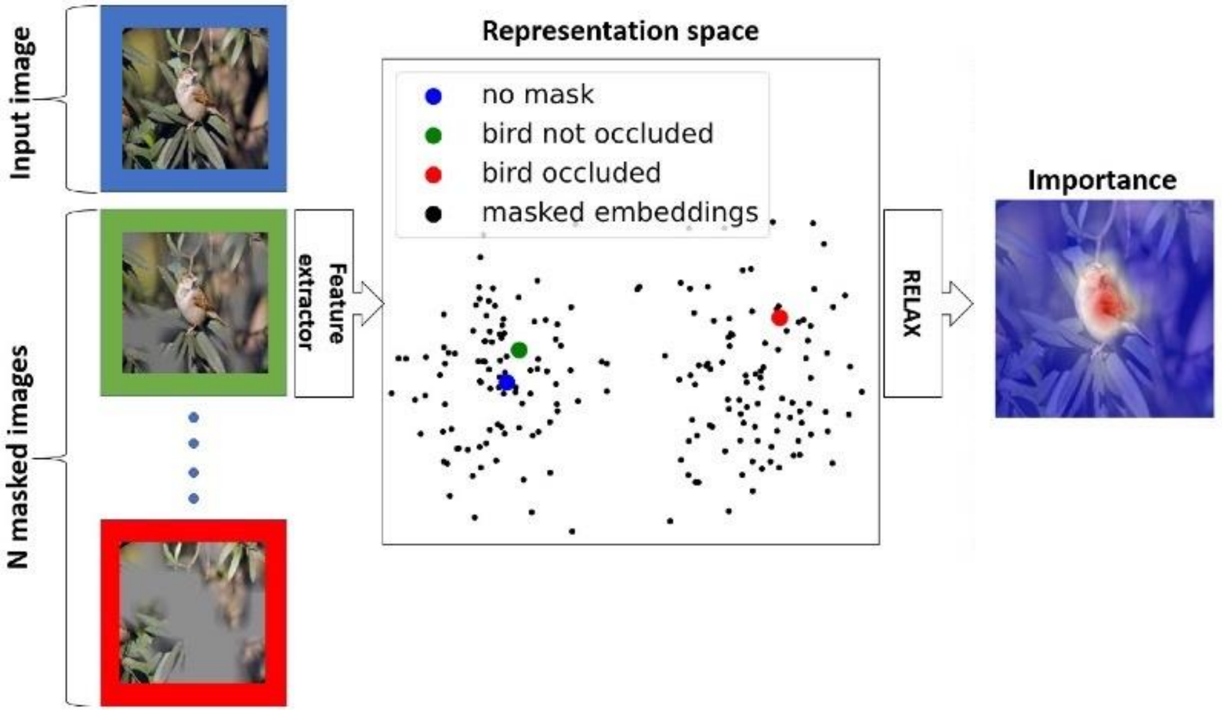


Wickstrøm et al. "RELAX: Representation learning explainability." *IJCV 2023*

RELAX: A representation learning explainability framework

$$R_{ij} = \mathrm{E}_{\mathbf{M}}\left[s(\mathbf{h}, \bar{\mathbf{h}})M_{ij}\right]$$

$$\bar{R}_{ij} = \frac{1}{N}\sum_{n=1}^{N} s(\mathbf{h}, \bar{\mathbf{h}}_n)M_{ij}(n)$$



Wickstrøm et al. "RELAX: Representation learning explainability." *IJCV 2023*

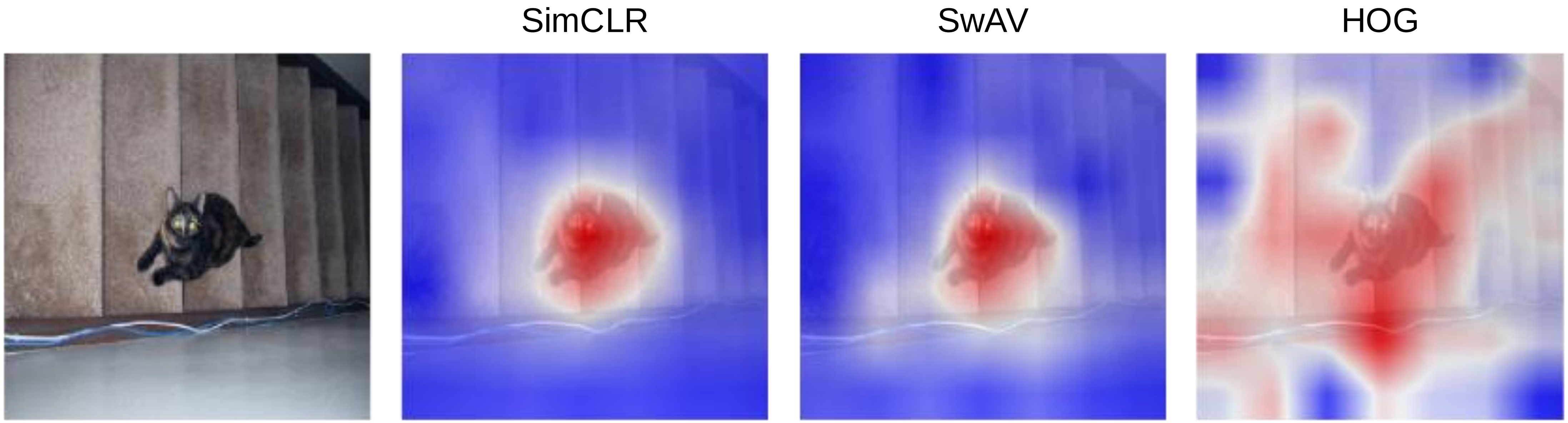
RELAX gives highest quality explanations of representations

Scores	Methods	Supervised		SimCLR		SwAV	
		COCO	VOC	COCO	VOC	COCO	VOC
pointing game	Saliency	67.1±0.0	82.8±0.0	59.9±0.0	75.9±0.0	60.0±0.0	76.3±0.0
	Smooth Saliency	62.8±0.0	79.5±0.0	60.1±0.0	75.9±0.0	59.8±0.0	76.4±0.0
	Guided Saliency	66.6±0.0	82.9±0.0	58.4±0.0	73.3±0.0	59.5±0.0	75.8±0.0
	Integrated Gradients	47.8±0.0	59.1±0.0	32.9±0.0	48.2±0.0	36.5±0.0	51.5±0.0
	Grad CAM	66.8±0.4	78.7±0.5	47.7±0.7	57.0±0.6	48.7±1.0	58.6±0.8
	RELAX	**72.6±0.1**	**86.6±0.2**	**68.7±0.3**	**85.2±0.3**	**67.8±0.2**	**84.7±0.2**
	U-RELAX	72.1±0.3	86.4±0.4	68.6±0.2	85.0±0.5	66.7±0.7	84.1±0.4
top k	Saliency	62.2±0.0	80.1±0.0	56.5±0.0	71.3±0.0	56.5±0.0	71.4±0.0
	Smooth Saliency	59.2±0.0	74.1±0.0	56.4±0.0	71.1±0.0	56.4±0.0	71.3±0.0
	Guided Saliency	62.2±0.0	80.2±0.0	55.1±0.0	69.0±0.0	56.3±0.0	71.1±0.0
	Integrated Gradients	47.7±0.0	61.0±0.0	35.4±0.0	52.8±0.0	33.2±0.0	49.0±0.0
	Grad CAM	64.0±0.0	78.3±0.0	43.6±0.0	55.3±0.0	43.1±0.1	54.8±0.0
	RELAX	**72.8±0.4**	**86.9±0.1**	**69.0±0.3**	**85.6±0.2**	**68.1±0.4**	**85.1±0.2**
	U-RELAX	72.2±0.4	86.5±0.2	68.8±0.4	85.3±0.1	66.6±0.4	84.2±0.3
relevance rank	Saliency	46.8±0.0	59.5±0.0	41.2±0.0	53.6±0.0	40.9±0.0	53.4±0.0
	Smooth Saliency	42.6±0.0	54.6±0.0	41.1±0.0	53.4±0.0	40.9±0.0	53.3±0.0
	Guided Saliency	46.8±0.0	59.8±0.0	40.6±0.0	53.0±0.0	40.9±0.0	53.3±0.0
	Integrated Gradients	38.4±0.0	51.9±0.0	31.9±0.0	47.2±0.0	32.3±0.0	48.3±0.0
	Grad CAM	46.0±0.0	60.2±0.0	37.5±0.0	50.7±0.0	37.8±0.0	50.9±0.0
	RELAX	**56.4±0.0**	**70.2±0.1**	**54.2±0.2**	**69.8±0.1**	**52.4±0.1**	**69.1±0.0**
	U-RELAX	52.4±0.0	64.7±0.1	50.7±0.1	63.3±0.1	46.2±0.1	59.5±0.0

Table 1 Pointing game, top k, and relevance rank scores in percentages and averaged over 3 runs. Higher is better and bold numbers highlight the top performance. Results show that our method improves on the baseline across all scores.

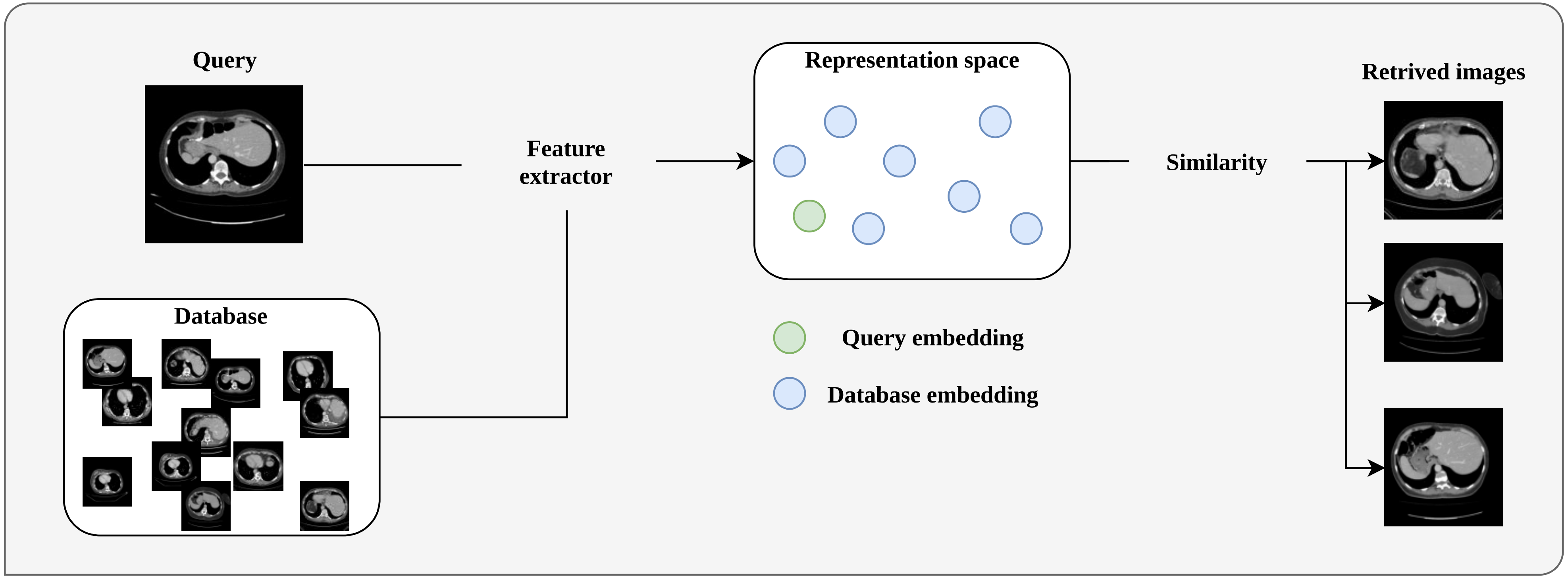
Compare feature extractor trained with and without supervision

Compare deep learning feature extractors



Content-based image retrieval of CT liver images

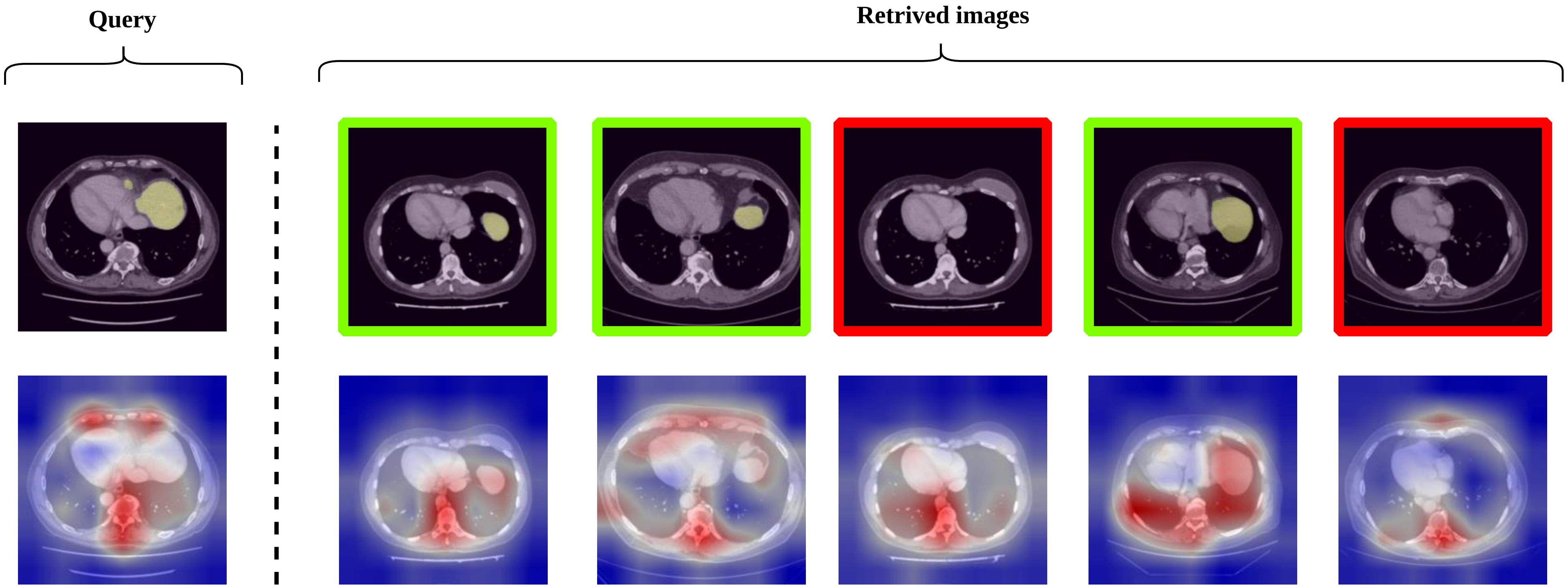
- Simple idea: retrieve images in large database based on image content.
- Use self-supervised learning to train feature extractor without labeled data.



Wickstrøm et al. "A clinically motivated self-supervised approach for content-based image retrieval of CT liver images." CMIG 2023.

RELAX analysis of feature extractor

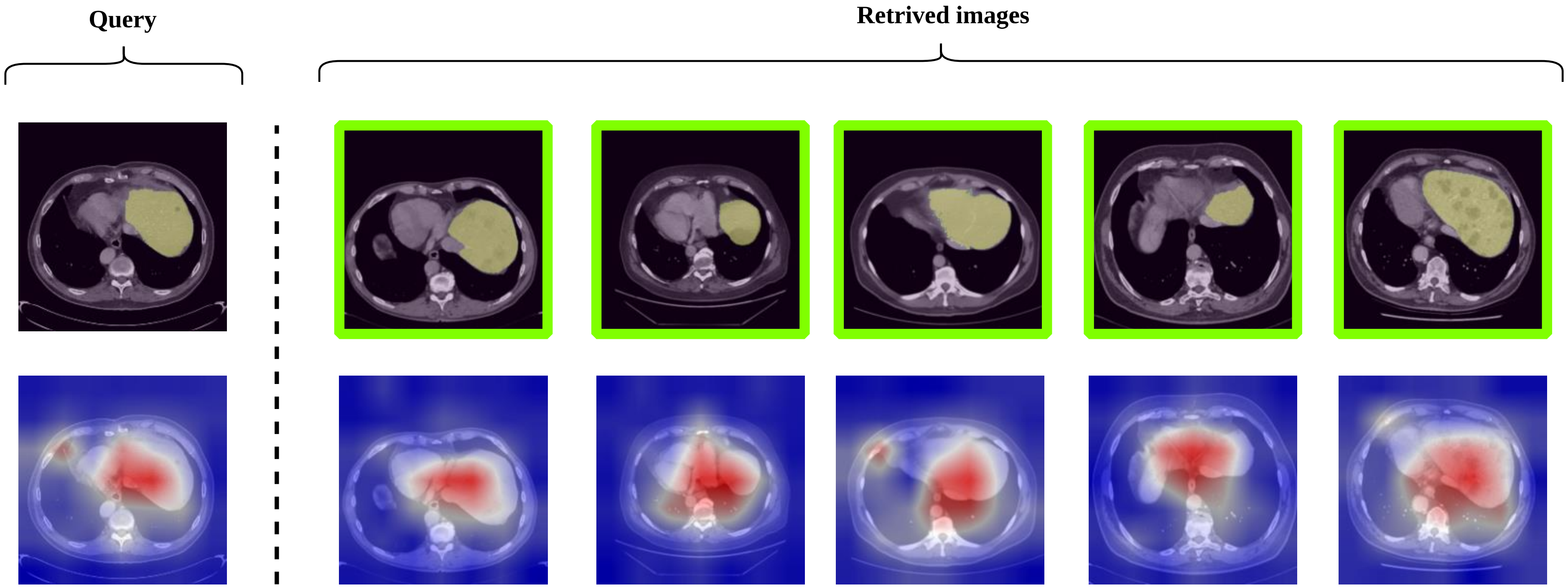
- Imagenet feature extractor focus on edge information.



Wickstrøm et al. "A clinically motivated self-supervised approach for content-based image retrieval of CT liver images." CMIG 2023.

RELAX analysis of feature extractor

- Feature extractor trained using our method focus on liver features.



Wickstrøm et al. "A clinically motivated self-supervised approach for content-based image retrieval of CT liver images." CMIG 2023.

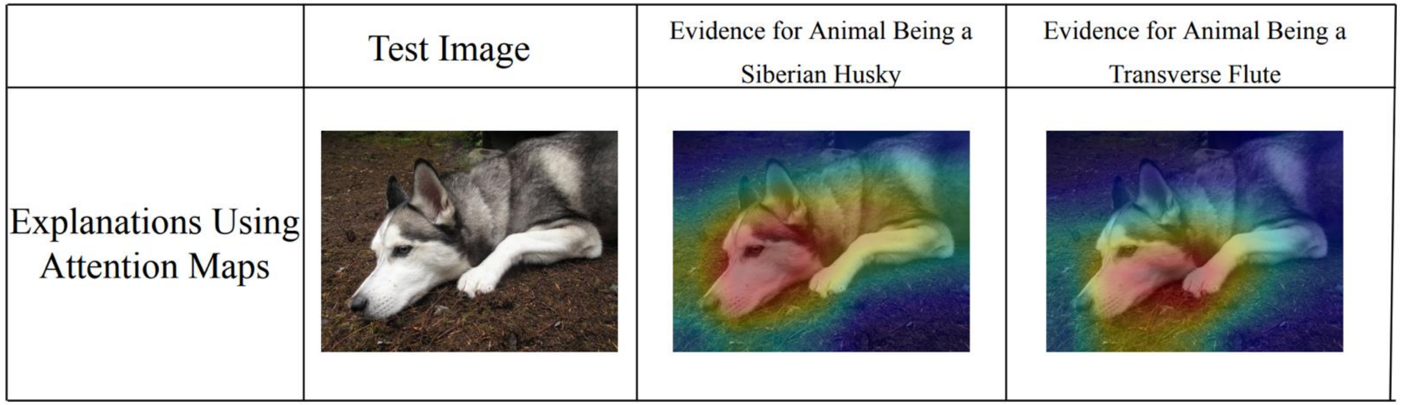
Summary

- Explainability for representation learning
- RELAX – A simple approach
- Model agnostic

Towards Self-Explainable Models

Why self-explaining models?

- Want inherently interpretable models
- Ensure faithfulness to computation
- Want to go beyond what the model is looking at



Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 2019

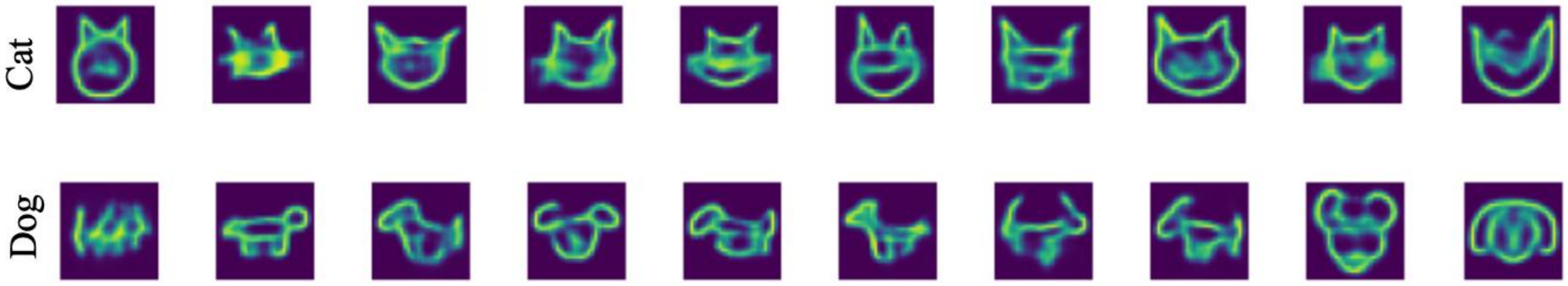
ProtoVAE: A Trustworthy Self-Explainable Prototypical Variational Model

Gautam, et al. "Protovae: A trustworthy self-explainable prototypical variational model." NeurIPS 2022.

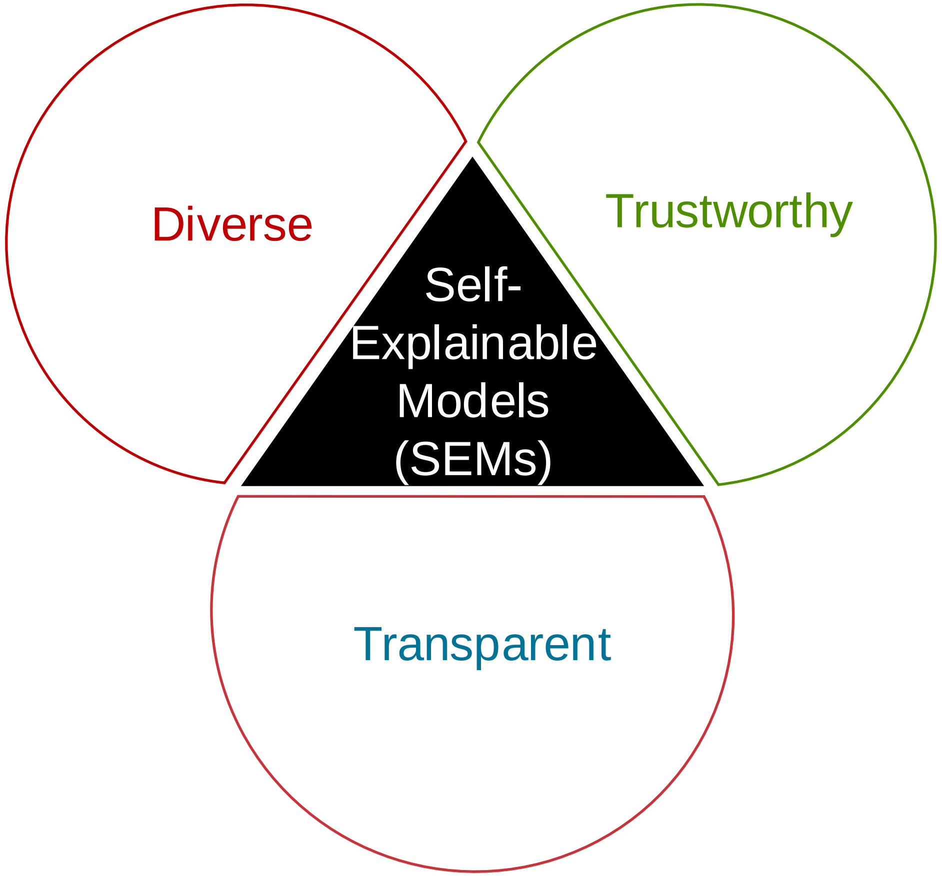
Concept/Prototypical Self-Explainable Models

Self-Explainable Models: Provides predictions and explanations at the same time.

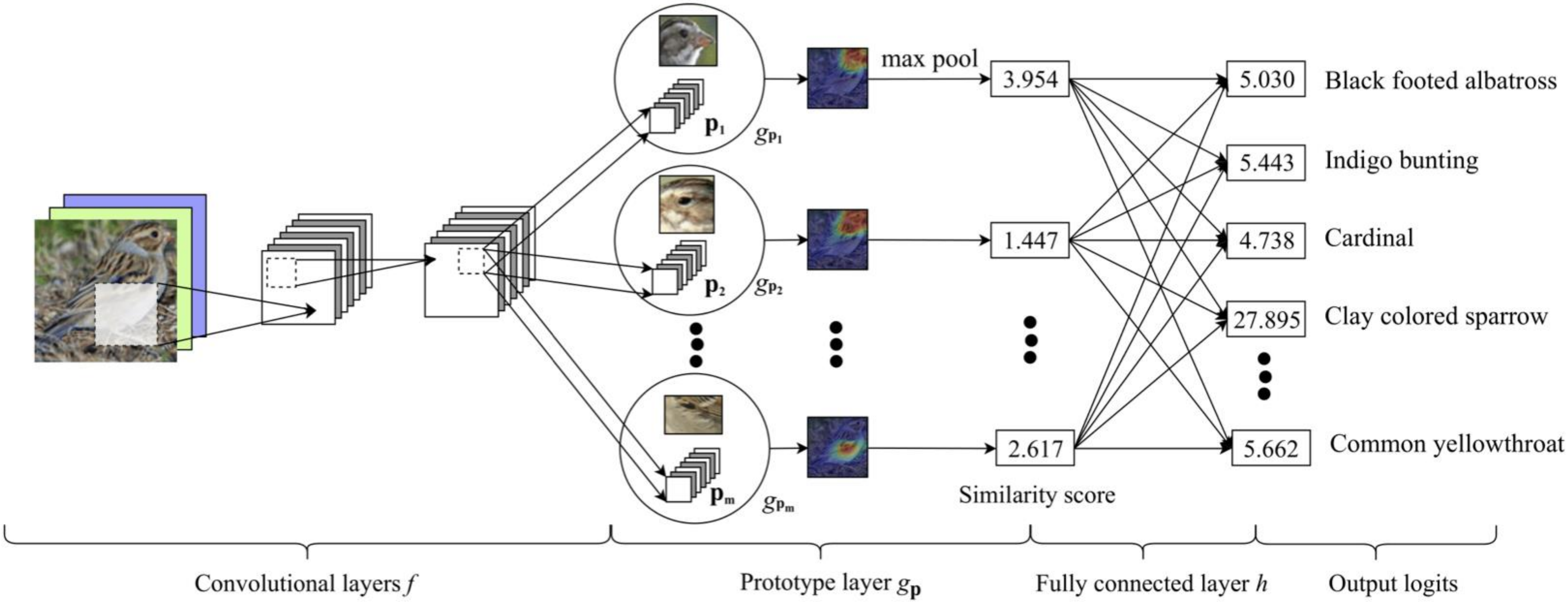
Prototypical Self-Explainable Models: Learns <u>representatives of the class</u>



Predicates for a self-explainable model



Revisit Prior Self-Explainable Models



Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." *NeurIPS* 2019.

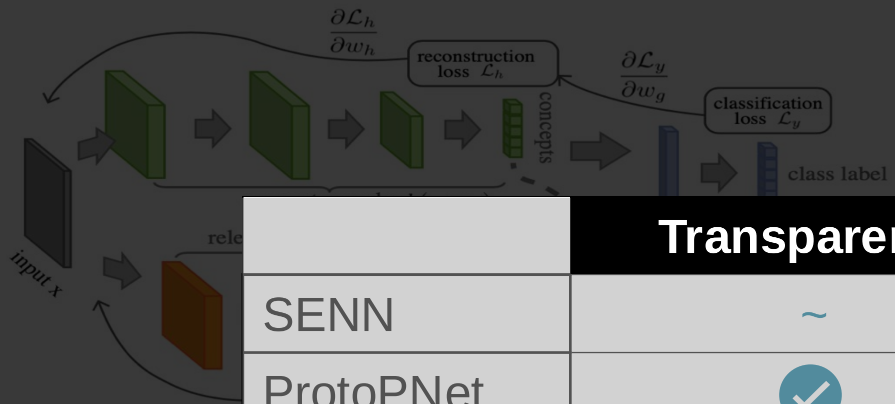
Revisit Prior Self-Explainable Models

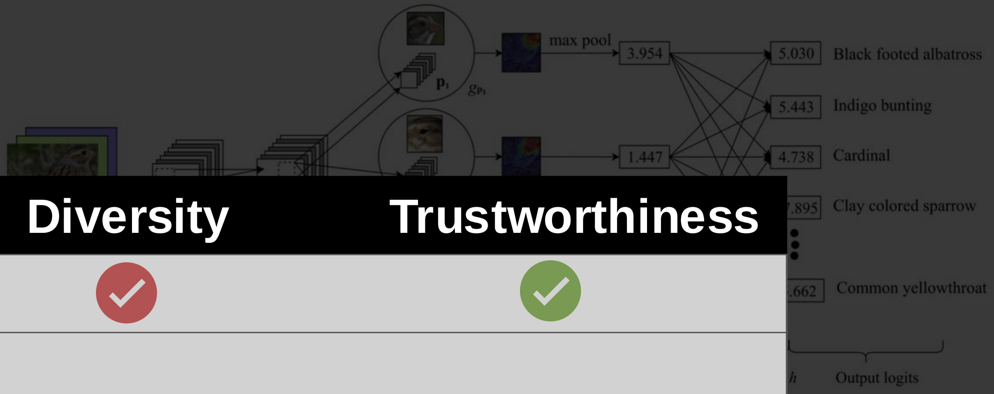
$$\min_{\mathbf{P}, w_{\text{conv}}} \frac{1}{n} \sum_{i=1}^{n} \text{CrsEnt}(h \circ g_{\mathbf{p}} \circ f(\mathbf{x_i}), \mathbf{y_i}) + \lambda_1 \text{Clst} + \lambda_2 \text{Sep}$$

$$\text{Clst} = \frac{1}{n} \sum_{i=1}^{n} \min_{j: \mathbf{p}_j \in \mathbf{P}_{y_i}} \min_{\mathbf{z} \in \text{patches}(f(\mathbf{x}_i))} \|\mathbf{z} - \mathbf{p}_j\|_2^2$$

$$\text{Sep} = -\frac{1}{n} \sum_{i=1}^{n} \min_{j: \mathbf{p}_j \notin \mathbf{P}_{y_i}} \min_{\mathbf{z} \in \text{patches}(f(\mathbf{x}_i))} \|\mathbf{z} - \mathbf{p}_j\|_2^2$$

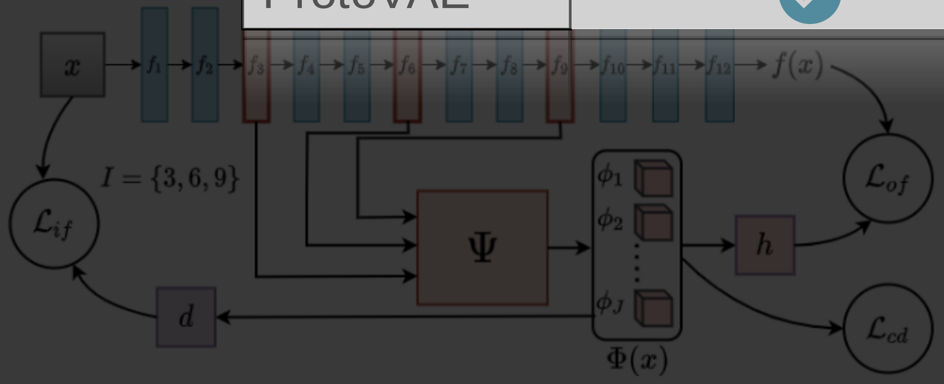
Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." *NeurIPS* 2019.



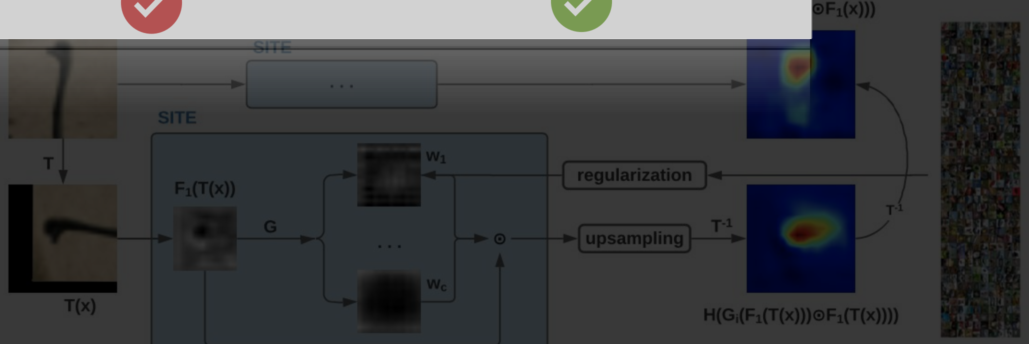


	Transparency	Diversity	Trustworthiness
SENN	~	✓	✓
ProtoPNet	✓		
SITE		✓	✓
FLINT	~	✓	
ProtoVAE	✓	✓	✓

SENN (Al



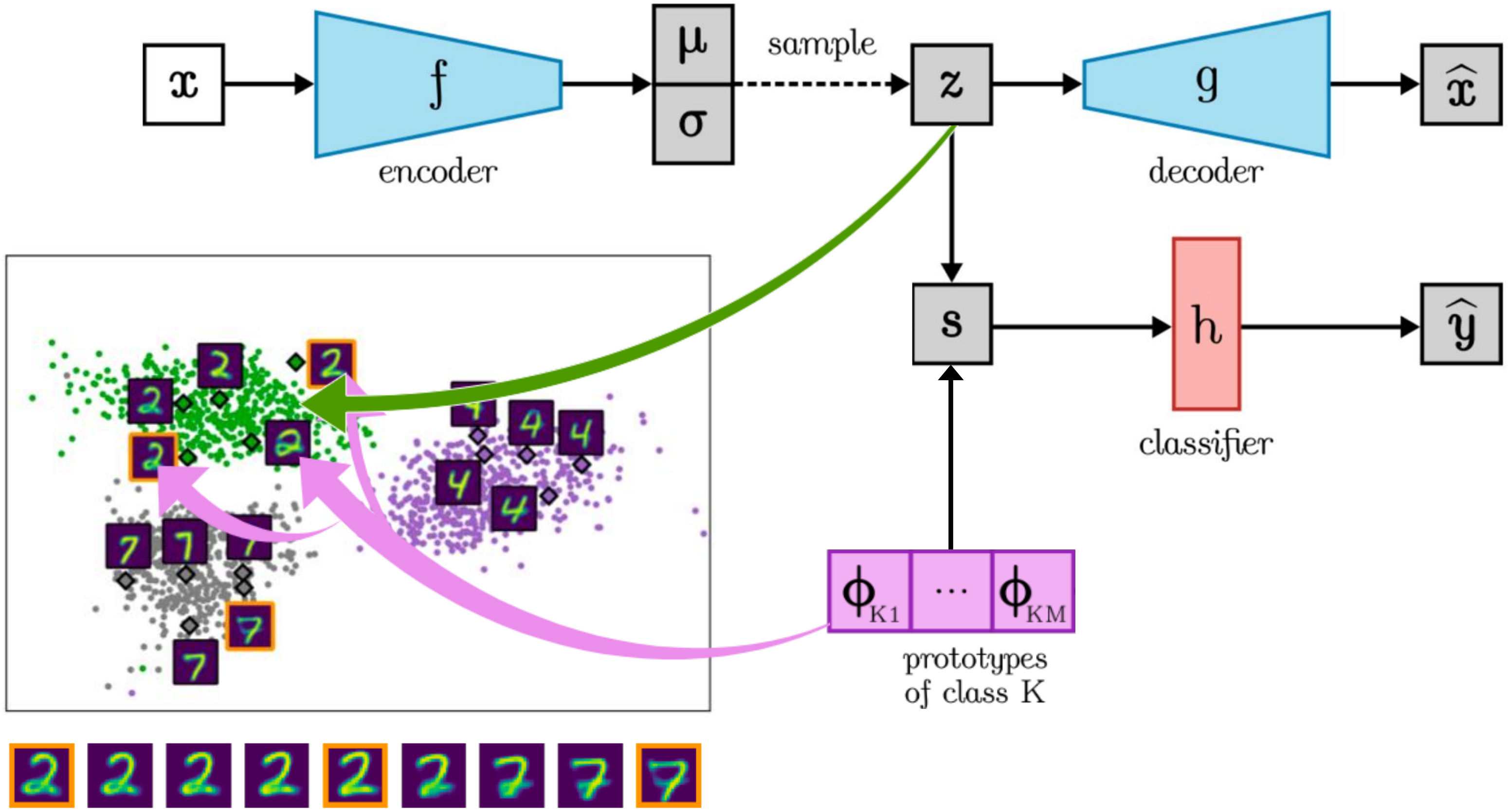
FLINT (Parekh et al. NeurIPS 2021)



SITE (Wang et al. NeurIPS 2021)

ProtoVAE

Transparent architecture



ProtoVAE

Diversity and trustworthiness through loss

$$\mathcal{L}_{\text{ProtoVAE}} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{orth}} + \mathcal{L}_{\text{VAE}}$$

$$\mathcal{L}_{\text{pred}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{CE}(h(\boldsymbol{s}_i); \boldsymbol{y}_i)$$

Inter-class diversity

$$\mathcal{L}_{\text{orth}} = \sum_{k=1}^{K} ||\bar{\boldsymbol{\Phi}}_k^T \bar{\boldsymbol{\Phi}}_k - \boldsymbol{I}_M||_F^2$$

Intra-class diversity

$$\mathcal{L}_{\text{VAE}} = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i||^2 + \sum_{k=1}^{K}\sum_{j=1}^{M} \boldsymbol{y}_i(k) \frac{\boldsymbol{s}_i(k,j)}{\sum_{l=1}^{M} \boldsymbol{s}_i(k,l)} D_{\text{KL}}\big(\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)||\mathcal{N}(\boldsymbol{\phi}_{kj}, \mathbf{I}_d)\big)$$

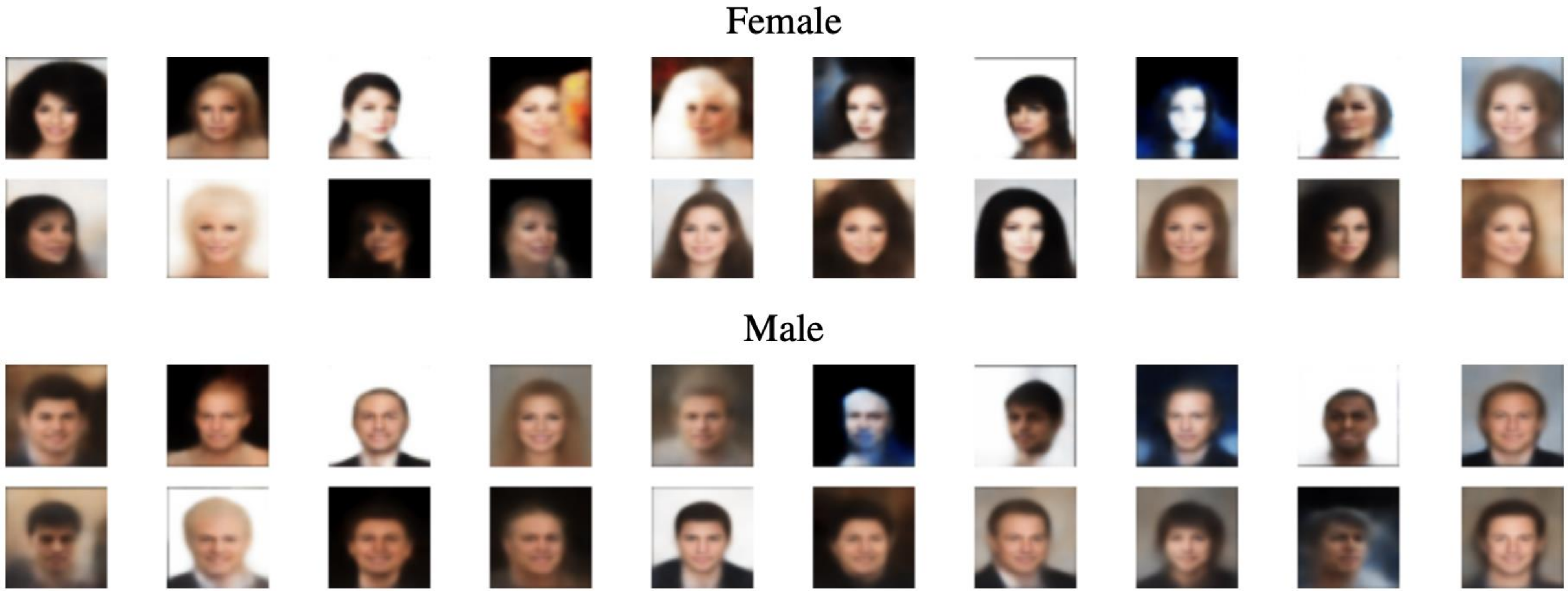
Robust classification and reconstruction

Predictive performance

	Black-box encoder	FLINT	SENN	*SITE	ProtoPNet	ProtoVAE
MNIST	99.2±0.1	**99.4±0.1**	98.8±0.7	98.8	94.7±0.6	**99.4±0.1**
fMNIST	91.5±0.2	91.5±0.2	88.3±0.3	-	85.4±0.6	**91.9±0.2**
CIFAR-10	83.9±0.1	79.6±0.6	76.3±0.2	84.0	67.8±0.9	**84.6±0.1**
QuickDraw	86.7±0.4	82.6±1.4	79.3±0.3	-	58.7±0.0	**87.5±0.1**
SVHN	**92.3±0.3**	90.8±0.4	91.5±0.4	-	88.6±0.3	**92.2±0.3**

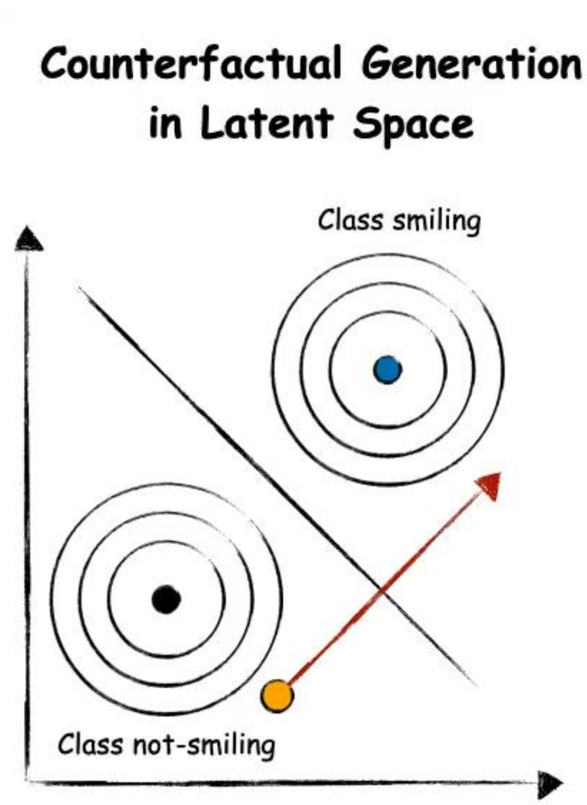
Results for accuracy (in %) for ProtoVAE and comparison with other state-of-the-art methods. *Results for SITE are taken from the original paper and thus based on more complex architectures.

Prototypes



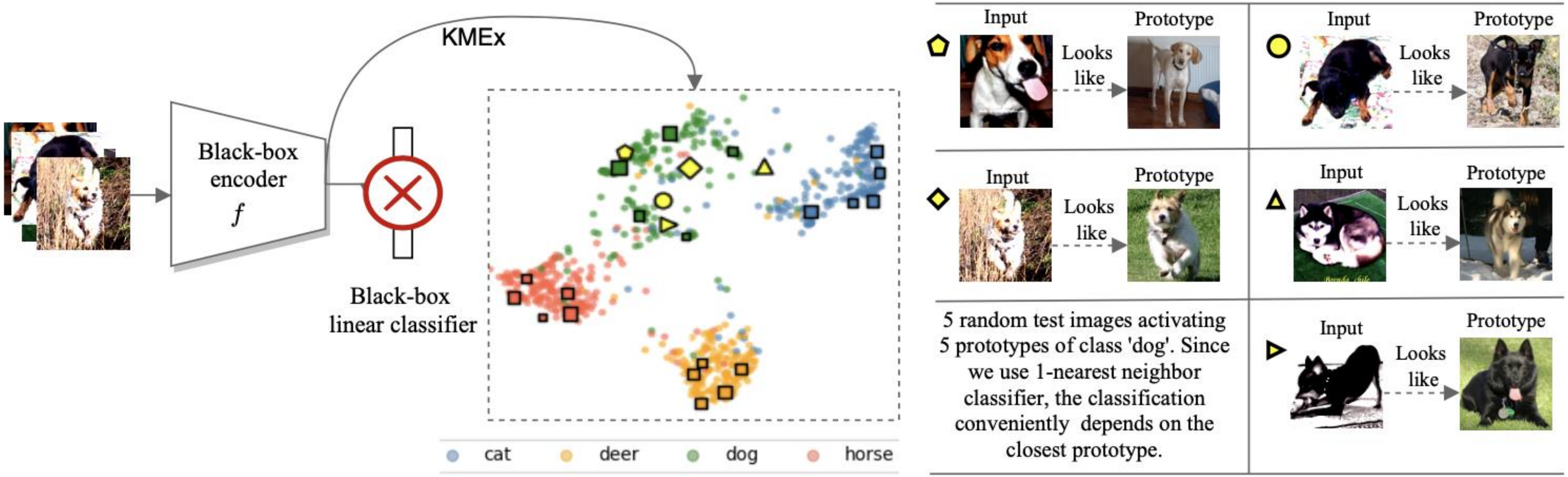
Prototypes learned for CelebA dataset with ProtoVAE

Self-explainable Model with high-res prototypes



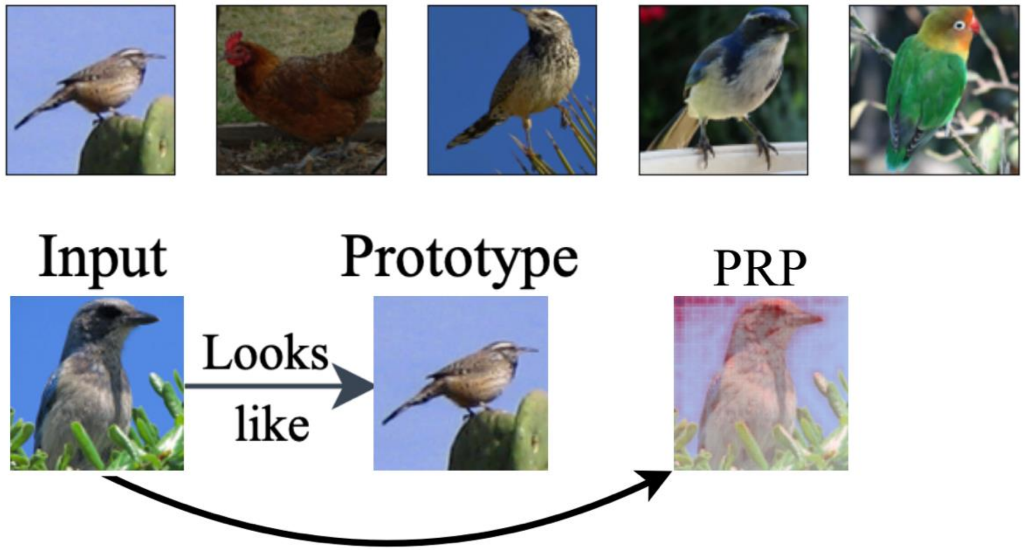
Haselhoff, Anselm, et al. "The Gaussian Discriminant Variational Autoencoder (GdVAE): A Self-Explainable Model with Counterfactual Explanations." *ECCV 2024.*

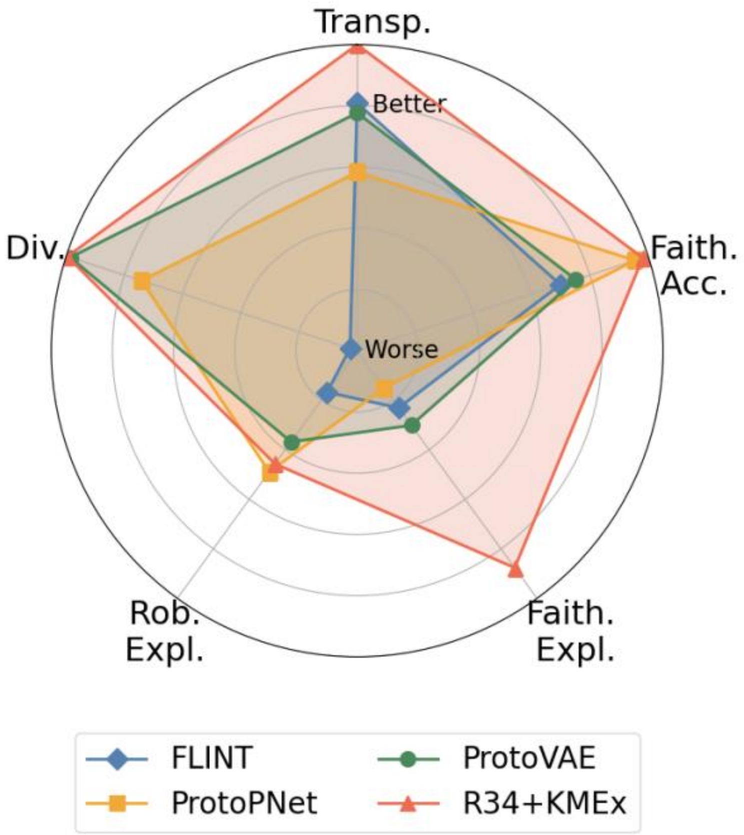
Bridging post-hoc and self-explainable models



Gautam et al. "Prototypical Self-Explainable Models Without Re-training" TMLR 2024.

Bridging post-hoc and self-explainable models





Gautam et al. "Prototypical Self-Explainable Models Without Re-training" TMLR 2024.
Gautam et al. "This looks more like that: Enhancing self-explaining models by prototypical relevance propagation." *Pattern Recognition* 2023

Conclusion

- Opening up the black-box
- Self-explainable deep learning models
- Active area of development
  - Best of both worlds

Northern Lights Deep Learning Conference

Tromsø, January

Main Conference: January 7th-9th
Winter School: January 6th-10th

nldl.org

- International conference
- Great Keynote Speakers
- Winter school
- Diversity in AI, industry event