Position: Do Not Explain Vision Models Without Context

Paulina Tomaszewska [1] Przemysław Biecek [1,2]

[1]Warsaw University of Technology [2]University of Warsaw

Does the stethoscope in the picture make the adjacent person a doctor or a patient?

It depends on the contextual relationship of the two objects. If it's obvious, why don't explanation methods for vision models use contextual information?

Our position:
we should take into account contextual information when explaining vision models – we need *spatial XAI* (shift from *where* to *how* the objects in the image are oriented towards each other).

Contextual information within images

The importance of context within input data is vastly studied in Time Series and Natural Language Processing, yet much less explored in Computer Vision. The spatial context can matter in street surveillance, autonomous cars, healthcare, land analysis (agriculture, archeology), IQ tests.

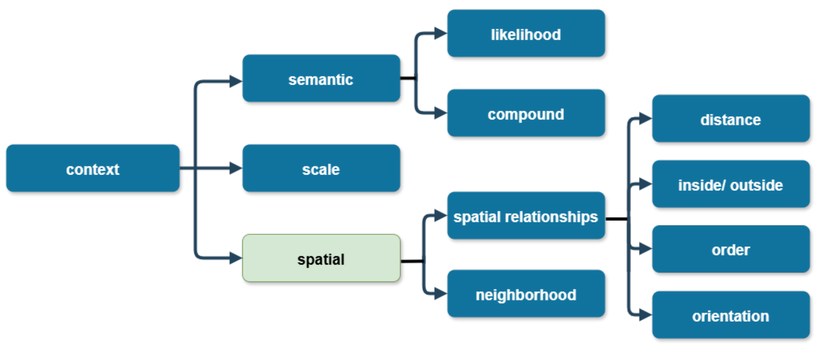
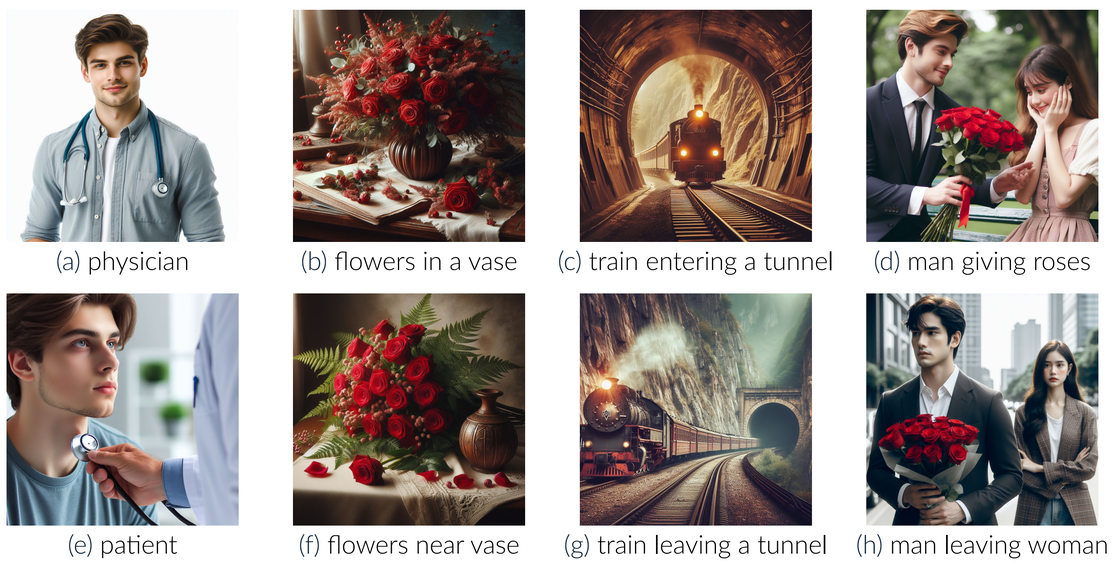


Figure 1. Taxonomy of contextual information within images (extended version of the one in [2]).



(a) physician (b) flowers in a vase (c) train entering a tunnel (d) man giving roses

(e) patient (f) flowers near vase (g) train leaving a tunnel (h) man leaving woman

Figure 2. Examples of images where the ground truth labels depend on spatial relationships between objects: *distance* (a, e), *inside/outside* (b, f), *order* (c, g), *orientation* (d, h). The images were created with the assistance of DALL-E 3.

Modelling community/ XAI community vs spatial context

Spatial context can be valuable when performing Deep Learning (DL) tasks. It was addressed in: (1) model architecture families, (2) context-oriented solutions, (3) pretraining, (4) context relationships as output.

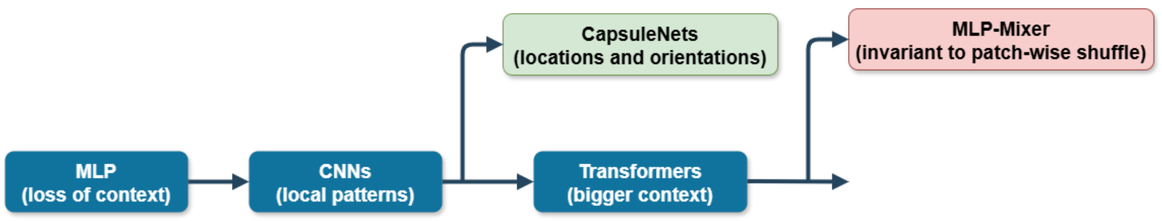


Figure 3. The evolution of model architectures for image processing in regards to context coverage.

The spatial context was considered when training neural networks but it was of much smaller interest in XAI. This is a niche that should be fulfilled (*spatial XAI*). Note that without proper XAI tools, we would not be able to explain the predictions of models when the same features are in input images but placed differently in the scene (Figure 4). We distinguish four main approaches: (1) intrinsically explainable models, (2) measures of spatial context, (3) leveraging current XAI methods, (4) input-output relationships.

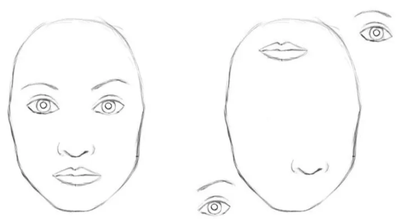
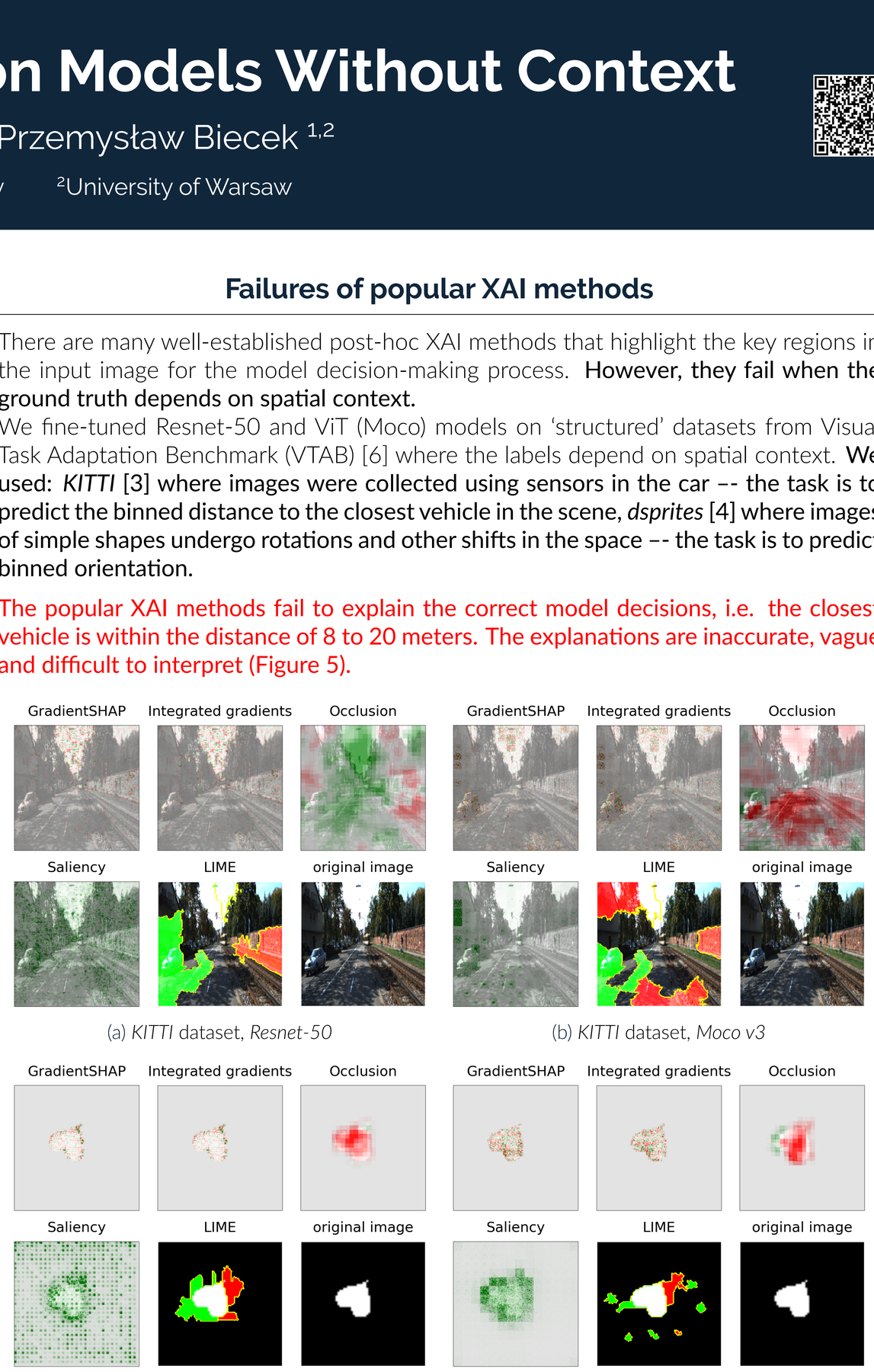


Figure 4. *Face* and *deformed face* [7] depending on spatial relationships between the elements. It was a motivation behind CapsuleNets [5]. Even in the case of two good predictions, we will not get proper explanations.

Failures of popular XAI methods

There are many well-established post-hoc XAI methods that highlight the key regions in the input image for the model decision-making process. **However, they fail when the ground truth depends on spatial context.**
We fine-tuned Resnet-50 and ViT (Moco) models on 'structured' datasets from Visual Task Adaptation Benchmark (VTAB) [6] where the labels depend on spatial context. **We used:** *KITTI* [3] where images were collected using sensors in the car –- the task is to predict the binned distance to the closest vehicle in the scene, *dsprites* [4] where images of simple shapes undergo rotations and other shifts in the space –- the task is to predict binned orientation.

The popular XAI methods fail to explain the correct model decisions, i.e. the closest vehicle is within the distance of 8 to 20 meters. The explanations are inaccurate, vague and difficult to interpret (Figure 5).



(a) *KITTI* dataset, *Resnet-50* (b) *KITTI* dataset, *Moco v3*

(c) *dsprites* dataset, *Resnet-50* (d) *dspites* dataset, *Moco v3*

Figure 5. The explanations of correct model predictions on the samples from the *KITTI* and *dsprites* datasets. Five popular XAI methods were used for the study. The color spectrum from red to green depicts the extent to which a particular image part contributed to the model's prediction (from negative to positive impact).

Our position: from *where* to *how* (*spatial XAI*)

We postulate we should take into account contextual information when explaining vision models – we need *spatial XAI*.

Recently, a change of a paradigm in XAI from *where* to *what* was proposed in [1]. It means that instead of simply highlighting the regions in the input images that are key for the model's prediction, we should focus on extracting what semantic features within the images are important. We propose to shift the approach from *where* to *how* so that instead of only operating on the image pixel space where the pixels are highlighted, we should also analyze how the objects within images are oriented towards each other in the space (Figure 6).

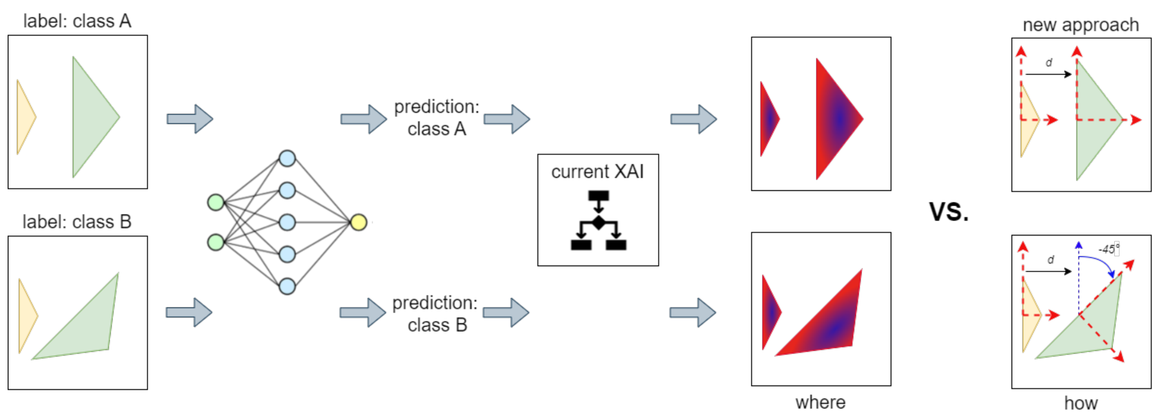


Figure 6. Need for a change of paradigm in XAI – we should focus on from *where* to *how*.

References

[1] Achtibat et al. From attribution maps to human-understandable explanations through Concept Relevance Propagation. *Nature Machine Intelligence*, 2022.

[2] Galleguillos et al. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 2010.

[3] Geiger et al. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012.

[4] Matthey et al. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[5] Sabour et al. Dynamic routing between capsules. In *NeurIPS*, 2017.

[6] Zhai et al. The visual task adaptation benchmark. *ArXiv*, 2019.

[7] Max Pechyonkin. Understanding Hinton's Capsule Networks. Part I: Intuition. Medium (online), 2017.

This research was financially supported by the Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme (grant: 1820/97/Z01/2023).