KI WISSEN
Automotive AI Powered by Knowledge

3.2 Model Agnostic Local Analysis with Latent Attacks (MALALA)

Daniel Kaulbach, Luca Bruder, Dina Krayzler, Leonard Rosen
Jan-Hendrik Clausen | Alexander Thamm GmbH

Formal Methods and Knowledge Representation

Problem Formulation

We employ a strategy for conducting a local and model-agnostic analysis of black-box models (BBM) specifically designed for image processing. Our approach produces relevant local neighborhoods by harnessing the generative characteristics of variational autoencoders (VAE). [1] Through these neighborhoods, we effectively depict the decision logic of the BBM using counter-exemplars with corresponding saliency maps.

Latent Attacks

The goal of latent attacks is changing an image in a meaningful way, such that it will lead to a different BBM prediction than the original image. Unlike adversarial attacks [2] the changes should consider the interrelations between the different pixels of the image. The ABELE method [3] is generating a neighborhood of images with a VAE. The new images are altered by changing values in the latent space of the VAE, which makes them more likely to be related to semantic concepts than changes on a pixel level. However, one drawback of VAEs is that they create blurry images. We improve the sharpness of generated images locally by intentionally overfitting the decoder on the original image.

Generating Pedestrian Counter-Exemplars

Our approach is designed to generate local explanations for the predictions of a given pedestrian detection model, following several distinct steps:

1. A detected pedestrian is extracted from a traffic scene, resized and inserted into the VAE model.
2. Using the VAE, samples with similar latent feature values to the original are generated, i.e., a neighborhood of similar examples.
3. The decoded examples are cropped back into the original traffic scene to create labels.
4. The initial neighborhood is optimized and augmented by sampling additional examples in a heuristically guided fashion.
5. The counter-exemplars (novel examples that are not detected by the BBM) from the neighborhood are used to create a saliency map which represents the extracted knowledge. (Fig. 1)

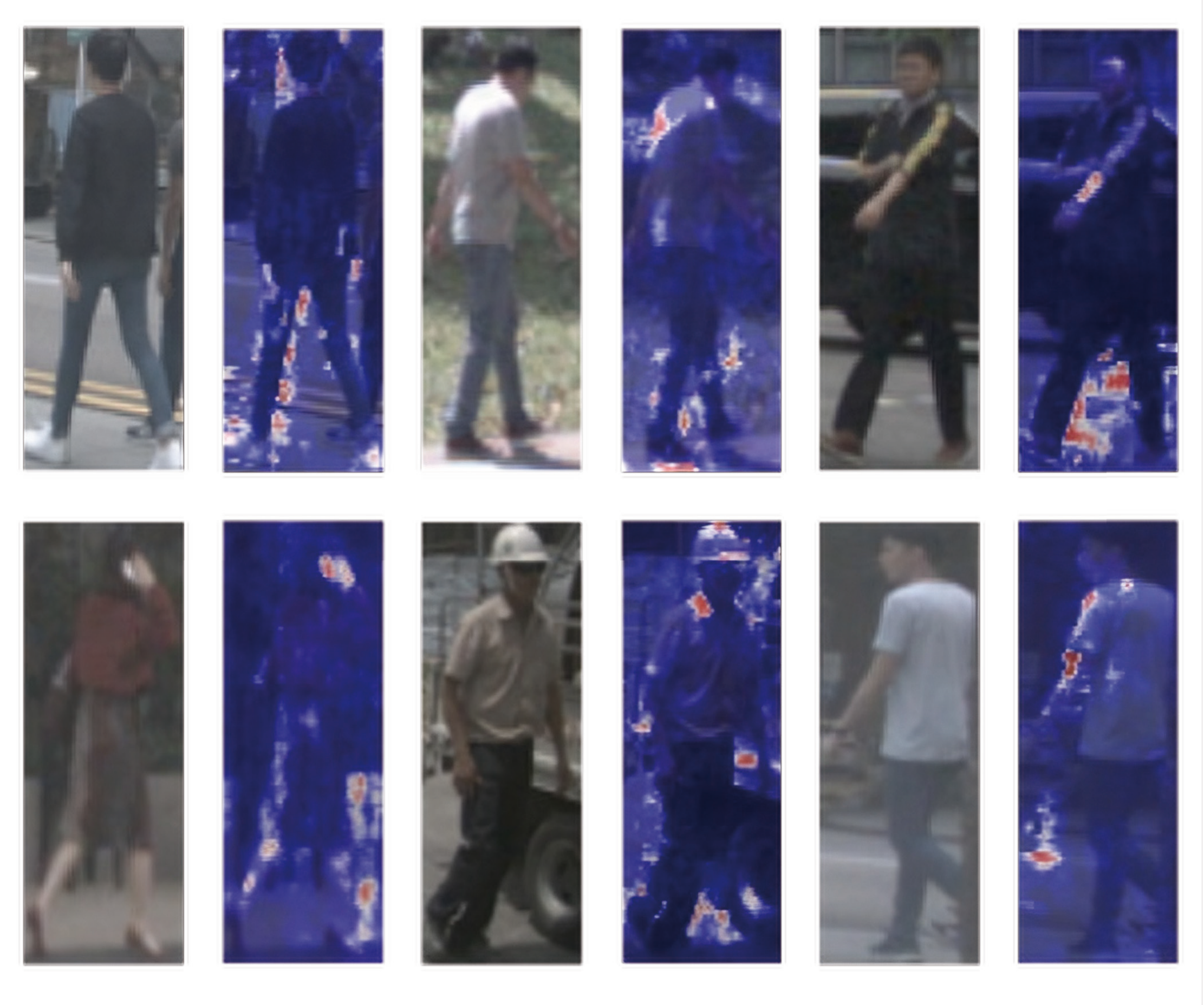

Figure 1: Detected pedestrians and their saliency maps that highlight image areas where influential to the BBM prediction change of counter-exemplars (© Alexander Thamm GmbH)

Undetected Pedestrians

MALALA is also applicable on undetected pedestrians for finding out, in which minimal way their image needs to change, for making them detectable. Fig. 2 shows a counterfactual, that added legs to an undetected person whose legs were originally occluded behind a barrier. Since the counterfactual is detected and the saliency map highlights the area around the legs, one can conclude that the BBM is sensitive to legs.

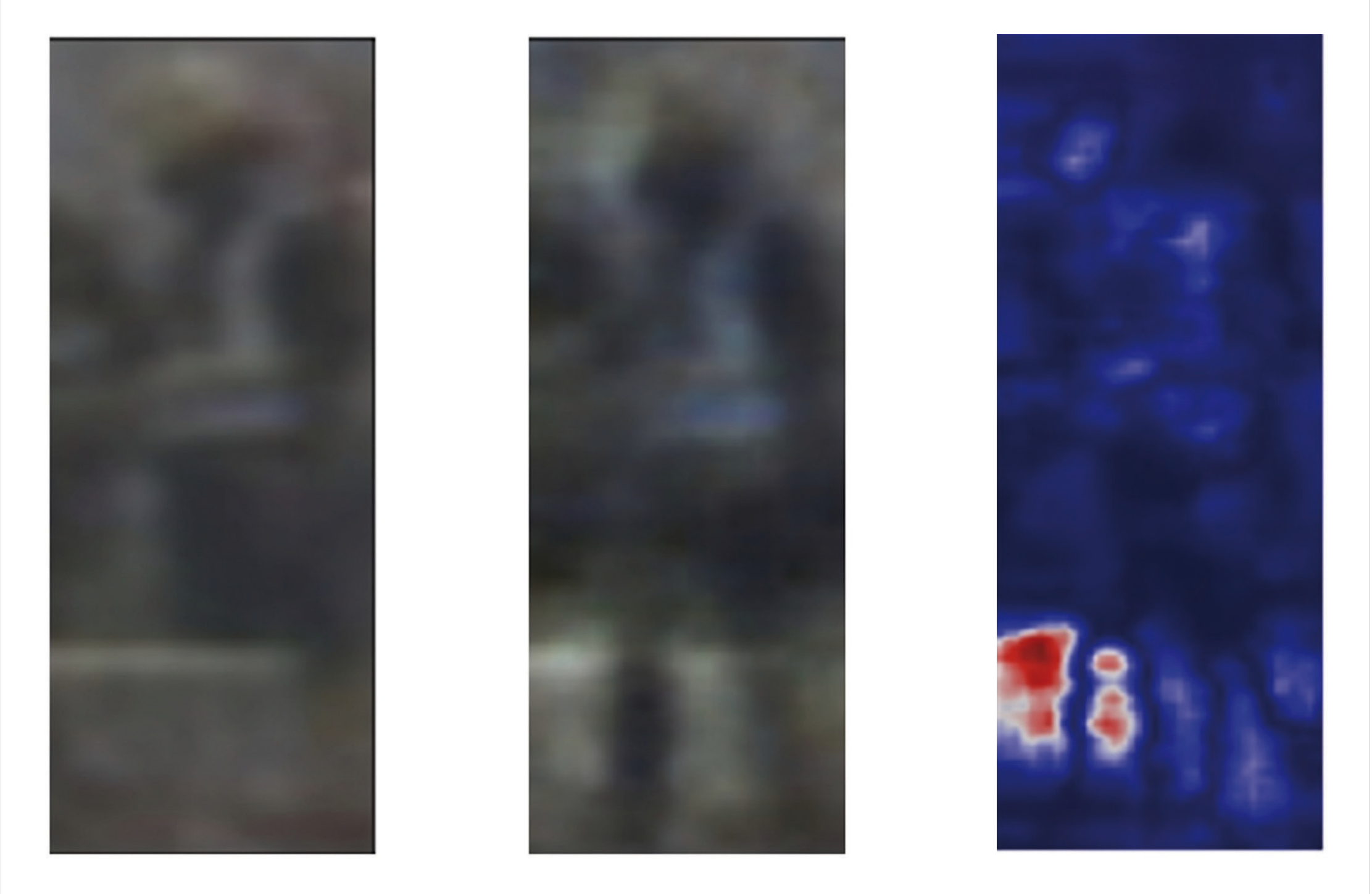

Figure 2: An undetected pedestrian with occluded legs (left), a counter-exemplar that added legs (middle) and its saliency map that highlights the area around the legs (right) (© Alexander Thamm GmbH)

Future Work

The counter-exemplars generated by MALALA are random and highly depending on the VAE and its training data. Text to image frameworks can potentially be used for creating changed images in a more controlled way to obtain more diverse scenarios.

References

[1] Y. Pu et al.: Variational autoencoder for deep learning of images, labels and captions. Advances in neural information processing systems, 2016
[2] Szegedy et al.: Intriguing properties of neural networks, 2013
[3] Guidotti et al.: Black box explanation by learning image exemplars in the latent feature space, 2020

Partners

External partners

For more information contact:
daniel.kaulbach@alexanderthamm.com
luca.bruder@alexanderthamm.com

KI Wissen is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

www.kiwissen.de @KI_Familie KI Familie

KI FAMILIE
VDA LEITINITIATIVE

Funded by the European Union
NextGenerationEU

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag