Reevaluating the Safety Impact of Inherent Interpretability on Deep Neural Networks for Pedestrian Detection

Patrick Feifel[1,2] Frank Bonarens[1] Frank Köster[2,3]

frank.bonarens@stellantis.com

[1] Stellantis, Opel Automobile GmbH

[2] Carl von Ossietzky Universität Oldenburg

[3] Deutsches Zentrum für Luft- und Raumfahrt

Abstract

AI-based perception is a key factor towards the automation of driving systems. A conclusive safety argumentation must provide evidence for safe functioning. Existing safety standards are not suitable to deal with non-interpretable deep neural networks (DNN) learning from unstructured data. This work provides a proof of concept for a comprehensible requirements analysis based on an interpretable DNN. Recent work on interpretability motivates to rethink software considerations of safety standards. We describe the application of established considerations to DNNs by integrating interpretability and identifying artifacts. DNN artifacts result from a meaningful decomposition of requirements and adaptions of the perception pipeline. To prove our concept, we propose an interpretable method for the center, scale and prototype prediction (CSPP) that learns an explicitly structured latent space. The interpretability-based requirements analysis of CSPP is completed by tracing artifacts and source code to decomposed requirements. Finally, qualitative post-hoc evaluations provide evidence for the fulfillment of defined requirements for the latent space.

1. Introduction

AI techniques develop various domains, reaching from the integration in consumer electronics to even more security-related surveillance tasks. The human-comparable performance of machine learning models (MLM) enables the automation of advanced tasks. Besides pushing technological boundaries, automated systems can significantly contribute to safety. Automated driving systems (ADS) utilizing MLMs can improve safety for pedestrians and prevent deadly casualties. Although safety-critical systems can profit from a responsibility transfer away from humans to automated MLM-based decision-making, the safety of involved components must be proven. Regarding the modular ADS pipeline, the decision-making process depends on a reliable perception of the complex environment based on sensory input [12]. Hence, guaranteeing pedestrian safety is in the first instance related to a safe perception.

MLMs integrated into the perception pipeline of ADS are mainly generated by supervised deep learning algorithms. The algorithms unfold their full potential by substituting an expert's manual feature engineering with feature learning from unstructured data. Data used for the perception task is represented by pixel values in images and is therefore unstructured. During training, features are learned as parameters in a deep neural network (DNN). Applied in ADS, DNNs become safety-critical and plausible evidence of reliable functioning has to be stated. Although standards for safety-critical systems are well established, DNNs are not suitable due to non-interpretable behavior in combination with unstructured data sampled from a complex input domain.

In this work, we reevaluate the safety impact of inherent interpretability by rethinking the role of MLM interpretability for existing safety standards and evaluating an interpretable DNN. The *development* of software according to DO-178C [19] is driven by the decomposition of requirements - naturally in contrast to a deep *learning* approach. However, recent work on MLM interpretability introduces mechanisms describing a meaningful decomposition of a DNN. Evolved from the inherent interpretability, we can define requirements and establish explicit traces to DNN artifacts. We propose an interpretable DNN for the center, scale and prototype prediction (CSPP). Our work relates most closely to Aravantinos and Diehl [1] who also present an approach to integrate traceability of artifacts with requirements in the context of a DNN. However, we specifically focus on an adapted and extended DNN architecture, propose its meaningful decomposition and derive comprehensive requirements. We evaluate the performance of CSPP

for pedestrian detection on the CityPersons dataset. Finally, the post-hoc evaluation of CSPP derives qualitative evidence for the fulfillment of defined requirements.

2. Related Work

2.1. Pedestrian Detection

Pedestrian detection can be accomplished by traditional computer vision (CV) algorithms as part of advanced driver assistance systems (ADAS). Nevertheless, the limited capabilities contradicts the application to ADS. DNNs significantly outperform traditional CV techniques and meet real-world complexity.

AI-based pedestrian detection results from reducing 2d object detection to two classes - pedestrian and background. More precisely, the DNN has to solve multiple binary classification tasks for different image regions.

Starting from this fundamental idea, early single-stage and two-stage object detectors utilized default anchor boxes, for instance, SSD [28] and Faster R-CNN [21]. *Anchor-based models* predict relative deviations to the bounding box coordinates of default anchors. Hence, the performance strongly depends on post-processing steps for clustering and calculating the scales and aspect ratios of default anchors boxes.

Recently, research has focused on *anchor-free models* yielding simpler DNN architectures. The absence of default anchor boxes eliminate the need for a-priori mining of training data to initiate anchor scales and sizes. Anchor-free models achieve state-of-the-art performance and surpass anchor-based models. Thereby, center-oriented anchor-free models simplify the 2d object detection in terms of predicting center points and the scale of an object, for instance, CenterNet [7], FoveaBox [22], FCOS [24] and CenterNet [30].

The anchor-free approach is also applied to pedestrian detection. CSP [16] extracts and combines semantic features of multiple scales to predict the center and height of a pedestrian. Regarding the evaluation on CityPersons [29] and Caltech [5], it achieves state-of-the-art performance in terms of the log-average miss rate (LAMR).

The LAMR [3, 5] takes the false positives per image (FPPI) next to the miss rate (MR) into account. Thereby, all detections for a given test dataset are sorted in descending order according to the predicted confidence scores. By applying a threshold of 0.5 to the intersection over union, predicted bounding boxes are matched with ground truth bounding boxes. Positive matches determine true positives (TP) in contrast to false positives (FP) or false negatives (FN).

The complexity of pedestrian detection can be explained by the small pixel height of pedestrians, occlusion and crowd appearances in images. Therefore, Dollar *et al*. [5] propose an initial definition of a *reasonable* evaluation setting: Pedestrians with a height smaller than 50 pixels shall be ignored. Wang *et al*. [27] refine the settings for CityPersons and exclude annotations with an occlusion rate larger than 0.35. Taking both settings into consideration, a reasonable subset with $\text{height} > 50$ pixels and $\text{occlusion rate} \leq 0.35$ can be defined. The reasonable evaluation leads to a more realistic view of the actual DNN capabilities.

2.2. Safety Argumentation

The assessment of safety-critical DNNs is challenging considering the non-interpretability in combination with unstructured data. In terms of ADS, practitioners see overwhelming complexity since the camera-based perception is based on pixel values. Till now, the inner working of a DNN remains largely non-interpretable and designing comprehensible test cases for individual DNN components seems impossible.

A straightforward verification approach based on input-output relations and formulation of general assumptions is problematic due to the dependency on pixel values. Nonetheless, one might make assumptions for the input domain from a functional safety perspective. A reduced domain might be used for the description of a safe operating state. It is still pending whether useful safety statements can be derived from such an argumentation.

In the first step towards a safety argumentation, functional insufficiencies (FI) such as incomprehensible behavior, unreliable confidence information and brittleness [18] were identified. Other safety concerns like the intrinsic oracle problem of perception can be added. Although each FI highlights slightly different issues, they jointly refer to the incomprehensible behavior - the non-interpretability of a DNN.

The lack of interpretability in combination with a high-dimensional input domain prevent the direct application of automotive-related standards like ISO 26262 [10] or SOTIF [11]. SOTIF ensures the 'absence of unreasonable risk' [18], but resilient evidence for the absence can yet not be extracted from DNNs. The verification formalism is largely targeting the function itself and not the software which represents the inner working of the function.

Safety standards, specifically developed for software, might be more applicable. The DO-178C [19] offers advanced considerations for safety critical software and focuses on the safety argumentation for source code development. The DO-178C relies on software life cycle data for generating evidence. Software life cycle data comprises *artifacts*, as by-products, for every step in the software life cycle. The wholesome DO process builds evidences by ensuring a safe software development process. A key concept is the *bi-directional traceability* for a *decomposition* of software high level requirements (HLR) into low level require-

ments (LLR). Hence, source code can be directly developed from the definition of LLRs - the decomposition of HLRs equally affects source code and LLRs.

The software development process of AI techniques can be conceptualized by the research agenda of *AI Engineering* [2]. The new field of research is focused on AI-related characteristics that complicate the application of well-known traditional safety considerations for software. Bosch *et al*. [2] identify explainability as a domain-specific research topic for safety critical systems.

2.3. Machine Learning Model Interpretability

Recently developed methods focusing on MLM interpretability can help to generate insights and increase trust in DNN decisions. According to Doshi-Velez and Kim [6], interpretability can be defined as '... the ability to explain or to present in understandable terms to a human'. Whether a model can be seen as interpretable is highly domain-specific and needs to be evaluated with respect to the task to be accomplished [20].

Regarding the assessment of interpretability, Lipton [15] defines two non-absolute dimensions: (1) *transparency* and (2) *post-hoc interpretability*. Transparency describes the evaluation on model, parameter and training level and focuses on inherent properties of the reasoning process. Additionally, post-hoc interpretability generates explanations for a given model behavior by applying surrogate explanatory models.

Although interpretability is described as a two-sided concept, latest research activities are largely focused on explainable AI (XAI) proposing local approximations on parameter level with surrogate models. Popular representatives are saliency methods. Adebayo *et al*. [13] show their vulnerability and limited capabilities to represent holistic interpretability. However, techniques modeling post-hoc explanations can contribute to a versatile understanding of a trained DNN.

In contrast to post-hoc methods, interpretability can also be inherently integrated into the DNN. *Inherent interpretability* addresses transparency but must be defined task-dependent. The inherent approach gives a new perspective on the reasoning process of a DNN. Although interpretability for all convolutional operations seems to be not feasible and maybe not even necessary, the development of intermediate representations that provide human-understandable explanations might be possible.

Our work builds upon the interpretable ProtoPNet [4]. The interpretable DNN bases its classification on the linear combination of similarities between *latent representations* and learned *prototypes*. A latent representation is defined as a vector $\mathbf{z} \in [0,1]^D$ with D feature channels. Extracted features depend on the spatial position of the latent representation in the latent space. The *latent space* $\mathbf{Z} \in [0,1]^{\frac{H}{r} \times \frac{W}{r} \times D}$ is the output of the AI-based feature extraction with scaling factor r. It represents the spatially compressed and encoded information regarding the input image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$ with H and W describing the width and height of an image. A prototype $\mathbf{p} \in [0,1]^D$ acts as a class-specific counterpart for a latent representation. Similarities are derived from the euclidean distance between a prototype and a latent representation. The classification task is solved in terms of an inference to the best explanation. If extracted features have high similarities to a class-specific prototype, it is most likely that the given image shows an object with the specific class.

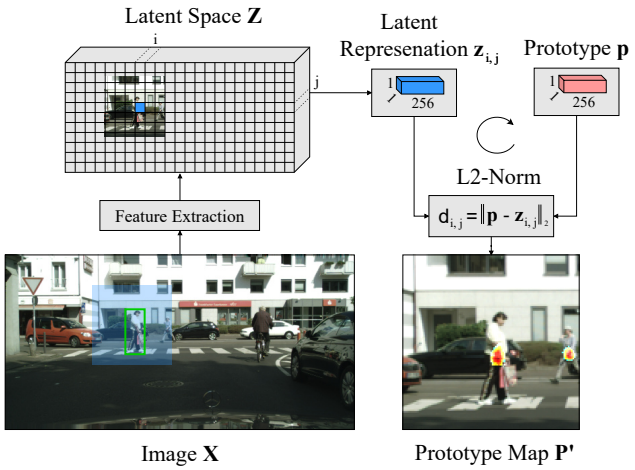

Figure 1. Pipeline to generate prototype map. Features within a receptive field (light blue) of an image $\mathbf{X}$ are encoded in a latent representation $\mathbf{z}_{i,j}$ (blue). Ground truth (green) is matched with a grid cell (blue) in the latent space $\mathbf{Z}$. The prototype map $\mathbf{P}'$ (reduced to the receptive field) highlights areas of small distances $d_{i,j}$ for a given prototype $\mathbf{p}$.

The focus on an interpretable DNN architecture, explicitly structured latent space and understandable reasoning process strengthens the inherent interpretability. ProtoPNet generates inherent explanations without applying surrogate models.

3. Proposed Methodology

3.1. Enable Decomposition through Interpretability

In our work, we apply key concepts introduced by the DO-178C as an existing safety standard for software considerations to a DNN. Even though a DNN can be seen as software, the source code primarily implements the network architecture, pre-, post-processing steps and learning process. Extracted features are represented by parameters. The hybrid DNN structure is not compatible with the decomposition approach of DO-178C.

We propose to rethink decomposition in the context of inherent interpretability. Although interpretability is not di-

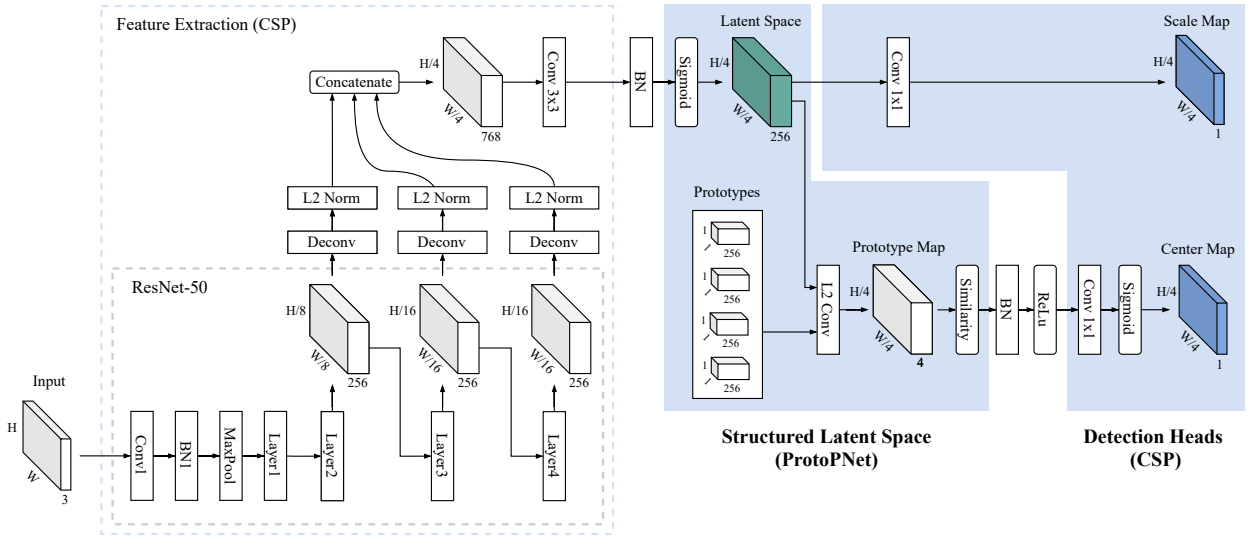

Figure 2. Interpretable DNN for center, scale and prototype prediction (CSPP) of pedestrians. The prototype-based structure of the latent space is introduced by ProtoPNet [4]. Prototypes are reference points in the latent space representing features of a pedestrian center. The structured latent space (green tensor) represents the inner understanding of a given image. The center and scale heads of CSP [16] complete the detection of a pedestrian center and prediction of the height. Considering an interpretable DNN, the detection heads (blue tensors) must reason from an explicitly structured latent space. We develop CSPP to prove our concept. In order to derive comprehensive low level requirements, we integrate prototypes and a comprehensible reasoning process into the detection pipeline. CSPP thus describes a novel interpretable DNN for pedestrian detection based on CSP (DNN for pedestrian detection) and ProtoPNet (interpretable DNN for classification).

rectly implied by the DO process, it can conceptualize the idea of decomposition. A *meaningful decomposition* of a DNN leads to components with individual responsibilities which can further be processed by an iterative requirements analysis. Hereby, interpretable DNNs are capable to decompose the internal reasoning process and assign meaning to separated components. With respect to the interpretable ProtoPNet, the mandatory extensions have to explicitly address the learning process and architecture of the DNN.

Regarding the inference of a DNN, the latent space encodes a given image, thus represents the internal representation and inner understanding of the input. Final detections are directly processed from the latent space.

We reformulate the optimization of an interpretable DNN for pedestrian detection: *Every prototype shall minimize its euclidean distance to the latent representation which can be matched to a pedestrian center*, shown in Figure 1 for one prototype. Thereby, the matching strategy determines which latent representation is responsible for an annotated pedestrian center (positive match) contrary to the background (negative match). High similarities result from small distances and increase confidence in a detection. The final prototype map is built upon the distance estimation between a given prototype and every latent representation.

In comparison to ProtoPNet, we ease constraints on the task-dependent interpretability and adapt class-specific prototypes to center-focused prototypes. Rather than implementing case-based reasoning for classification, our goal is to detect pedestrians based on learned features that can be separated from background information. The reasoning process can be understood as: *A pedestrian is detected because its extracted features in the latent representation are close enough to a prototype for a pedestrian center*.

Consequently, prototypes describe learned reference points for pedestrian features in the latent space. Learning from sampled images determines (1) how an abstract prototype is finally shaped and (2) how the extracted latent representations are clustered around prototypes. Both mechanisms describe the final structure of the latent space - the inner understanding of the given data. As a result of the distance-based reasoning process, decisions made by the interpretable DNN are quantifiable.

We would like to emphasize the enforcement of interpretability by explicitly structuring the latent space through prototypes. However, intermediate outputs of convolutional layers in the previous feature extraction remain non-human understandable. The limitation seems reasonable since machine-learned features motivates the use of DNNs for complex perception tasks. The proposed level of interpretability is balanced with respect to the complexity of the perception task and the characteristics of human perception [26]. Extracted features in latent representations are

initially not comprehensible to humans. Nonetheless, latent representations can be individually mapped to receptive fields that contain the compressed information [4].

3.2. An Interpretable DNN for Pedestrian Detection

The proposed method for **c**enter, **s**cale and **p**rototype **p**rediction (CSPP), illustrated in Figure 2 combines the anchor-free and center-oriented concept of CSP [16] for pedestrian detection and the interpretability of ProtoPNet [4]. In doing so, we extend the interpretable approach from the classification task to pedestrian detection. CSPP learns prototypes representing extracted features characteristic for different types of pedestrians. Due to the fact that pedestrian detection faces high complexity in terms of poses, appearances, apparels and occlusion, we see center-oriented prototypes for pedestrians as reasonable to propose a general view on interpretable DNNs for pedestrian detection. However, estimating the optimal number of prototypes remains non-trivial.

The feature extraction of CSP also builds the basis for CSPP and combines information from different scales, drawn from various layers of the ResNet-50 [9] backbone. The compatibility of the concatenated latent space with ProtoPNet is achieved by the sigmoid activation function. Every step based on the latent space contributes to the interpretable reasoning process of CSPP. However, the feature extraction as a significant part of CSPP remains non-interpretable.

The CSPP architecture is divided into two paths starting from the latent space to generate the center map and the scale map. The center path initially computes prototype maps applying a generalized L2 convolution [8, 17]. Distances $d_{i,j,k} = \|\mathbf{p}_k - \mathbf{z}_{i,j}\|_2$ for k prototypes are reinterpreted as similarities $s_{i,j,k} = log(\frac{d_{i,j,k}+1}{d_{i,j,k}+10^{-4}})$. The non-affine batch norm layer (Center-BN) acts as a domain-specific discriminator that distinguish whether one of the prototypes is close enough to the latent representation to be treated as a pedestrian center. The 1x1 center convolution (Center-Conv) for the center map represents a linear weighting without bias for all distances to prototypes. Parallel to that, the 1x1 scale convolution (Scale-Conv) calculates the height for a detected pedestrian center. The pedestrian width is approximated by $\text{width} = 0.41 \cdot \text{height}$ [29]. During inference, non-maximum suppression is applied as a post-processing step.

The integration of prototypes has to be explicitly formulated in the CSPP loss. On the one hand, latent representations and prototypes learn features that are characteristic for pedestrians, represented by minimizing the distance to all prototypes (L_{cluster}). On the other hand, the distance between prototypes and latent representations receipting background features is maximized ($L_{\text{separation}}$). Furthermore, we apply the loss terms according to CSP [16] to formulate the total loss:

$$\begin{aligned} L_{\text{total}} &= \lambda_{\text{cluster}} L_{\text{cluster}} + \lambda_{\text{separation}} L_{\text{separation}} \\ &\quad + \lambda_{\text{center}} L_{\text{center}} + \lambda_{\text{scale}} L_{\text{scale}} \end{aligned} \tag{1}$$

LLR	Definition	Rationale
1	The latent space shall be explicitly structured.	Structuring the high-dimensional latent space demands for the training of domain-specific reference points or prototypes.
2	A latent representation shall learn features that describe a pedestrian center.	The feature extraction must learn consistent latent representations that are aligned with prototypes. A latent representation encodes features of its receptive field.
3	Confidence scores shall be derived from an interpretable reasoning process.	Euclidean distances between latent representations and prototypes enable an interpretable reasoning process behind a detection.
4	The threshold for accepting or declining a detection shall be learned from statistics on the training dataset.	A domain-specific discriminator must learn statistics representing the distribution of the prototype maps for a given training dataset.
5	Outputs of ensemble classifiers shall be linearly resolved.	Multiple prototypes are involved in the joint decision and their contributions must be weighted.
6	Scales of a detection shall be drawn from latent representations.	Solving multiple tasks simultaneously must rely on a common latent space.

Table 1. Decomposed low level requirements (LLR). Inherent interpretability enables a meaningful decomposition of a DNN which can be processed by a requirements analysis.

3.3. Requirements Analysis with DNN Artifacts

According to the safety assessment of DO-178C, low level requirements (LLR) are derived from high level requirements (HLR) and source code can be directly developed from LLRs. The decomposition is driven by requirement refinement. In the chapter before, we have described an interpretable DNN to enable a meaningful decomposi-

tion. The interpretability approach is focused on the latent space and the subsequent reasoning process.

As a result, layers that process information from the latent space become meaningful. In the following, we refer to parameters of these layers as *DNN artifacts*. We define DNN artifacts as learnable parameters of an interpretable DNN that are explicitly formed by source code. Due to the already mentioned reasonable limitation, we exclude parameters of the feature extraction. DNN artifacts have a significant influence on the final detection pipeline attached to the feature extraction. The identification of DNN artifacts allows for the decomposition of HLRs.

Although this paper is not supposed to provide a complete safety argumentation for a DNN, we base our work on the following high level requirement (HLR): *The latent space of a DNN for pedestrian detection shall be interpretable*. The HLR relates directly to a meaningful decomposition. Interpretability targeting the latent space and reasoning process of a DNN enables a starting point to derive LLRs from a HLR. A meaningful decomposition allows to assign multiple DNN artifacts to specific requirements. We propose a non-exhaustive list of LLRs in Table 1.

The proposed methodology aims at the decomposition of an interpretable DNN to establish the bi-directional traceability of HLR, LLR, DNN artifacts and source code. Traces can finally be made explicit by creating a traceability matrix, shown in Table 2. Learned prototypes mark the most influential DNN artifacts. Furthermore, we identify parameters of the batch normalization (Center-BN) and 1x1 convolution (Center-Conv) considering the center path leading to the center map (Center-). The DNN artifact of the scale path (Scale-) is given by the 1x1 convolution (Scale-Conv). In contrast to learned parameters, CSPP also implements clearly separated source code: DNN architecture, distance measures, loss formulation, matching strategy and non-maximum suppression (NMS).

	Low level requirements					
	1	**2**	**3**	**4**	**5**	**6**
DNN artifacts						
Prototypes	✓	✓	✓	✓	✓	✓
Center-BN				✓		
Center-Conv					✓	
Scale-Conv						✓
Source code						
DNN architecture	✓		✓			✓
Distance measure	✓	✓		✓		
Matching strategy		✓				
Loss formulation	✓	✓				
NMS			✓		✓	

Table 2. Traceability matrix. Low level requirements (LLR) are traced to implemented DNN artifacts and source code. The identified DNN artifacts are learnable parameters of the interpretable CSPP. Explicit traces and full coverage are key factors for a conclusive safety argumentation.

4. Evaluation

4.1. Performance Evaluation

We evaluate the performance of CSPP compared to CSP on the reasonable validation dataset of CityPersons [29]. We emphasize the state-of-the-art 11.0% LAMR on CityPersons [29] achieved by CSP (with offset). For a fair comparison with CSPP, we reevaluate a simplified CSP implementation that neglects the strategy of moving average weights [23] and the additional offset prediction for centers. Furthermore, we apply the same thresholds for the confidence score of 0.1 and NMS of 0.5 and limit the number of detections to 1000. Table 3 compares results from various subsets of CityPersons of different methods. The performance of the proposed CSPP with 4 prototypes (13.78%) ranks between the simplified CSP (15.52%) and CSP with offset (11.00%). Hence, CSPP provides competitive performance while being inherently interpretable.

We would like to point out that our work is focused on enabling a safety argumentation, rather than defining a new state-of-the-art for pedestrian detection. That is why we do not cross-validate parameters such as the number of prototypes and channels in latent space. We arbitrarily choose the number of 1 and 4 prototypes for our experiments. Performance improvements should be part of future work.

4.2. Analyzing the Latent Space Structure

According to the traceability matrix in Table 2, all LLRs target the integration of prototypes. As pointed out, prototypes describe reference points structuring the high-dimensional latent space. The latent space is constrained and structured by quantifiable mechanisms (distance-based clustering and separation). Hence, CSPP learns consistent areas that hold features for pedestrian centers.

Latent representations can be directly mapped to individual detections or ground truth annotations (see TP, FP and FN in Figure 3). In contrast to a non-interpretable DNN, the mapping gives insights into the behavior of a DNN. The analysis of the latent space and the behavior of latent representations can be used in a safety argumentation since confidence scores are distance-based. A pedestrian is detected if the latent representation is within a certain distance to a prototype. Thus, the influence of each prototype is quantifiable and can be directly monitored. The critical distance threshold is learned.

Method	Reasonable	Bare	Partial	Heavy	Large	Medium	Small
CSP (with offset) [16]	**11.00**	**7.30**	**10.40**	49.30	6.50	**3.70**	**16.00**
CSP (w/o offset) [16]	11.40	8.10	10.80	49.90	**6.00**	3.90	18.20
CSP (simplified)	14.90	10.53	13.08	53.85	9.04	5.20	21.18
CSPP (1 prototype)	14.72	10.13	13.40	52.31	8.93	5.09	19.82
CSPP (4 prototypes)	13.78	9.31	12.70	**49.18**	8.13	4.83	20.68

Table 3. Log-average miss rates [%] for various validation subsets of CityPersons [29]. Evaluation was conducted with the original image size (1024x2048 pixels).

The fulfillment of LLRs for the latent space structure has to be demonstrated by post-hoc analysis. Although the loss formulation structures the latent space, the correct functioning of the intended learning process must be demonstrated. Clustering and separation is enforced by L_{cluster} and $L_{\text{separation}}$ as part of the loss formulation. Therefore, we expect a prototype to have smaller distances to TPs than to latent representations for background (see TP and background in Figure 3). For a given image, latent representations can be clustered according to their distances to the four prototypes (see prototype clusters in Figure 3). Due to the normalization of distances, the rejection area may differ between prototypes.

The transformation of latent representations into the 2d plane with PCA [25] and t-SNE [14] is shown in Figure 3. The t-SNE plot highlights the clusters of latent representations for TPs around trained prototypes. In contrast to that, we see latent background representations in the outer regions of a prototype cluster. The position of FPs and FNs describes a transition area where the risk of wrong decisions increases. The visualization enables the identification and analysis of failure modes. Wrong DNN detections can be attributed to an incorrect position of a latent representation relative to a prototype. Consequently, systematical failures can be analyzed from a DNN perspective. Dense clusters of FNs or FPs indicate potential improvement or areas of high uncertainty.

The qualitative analysis can be enriched by descriptive statistics since the clustering and separation of latent representations are distance-based. Figure 4 shows boxplots for four prototype clusters. It can be seen that prototypes have in general smaller distances to the background (blue) than TPs (green). FNs (red) and FPs (orange) are characterized by small distances to prototypes. A decision is a collective process. Hence, TPs can be a result of small distances to multiple prototypes. Overlapping boxes of TPs and FNs result from the weighted sum of distances that may not exceed the predefined confidence threshold (0.1). Furthermore, overlaps indicate problems in matching detections to ground truth. A detected pedestrian may be based on a latent representation within close distance to a prototype, but the predicted height and width of the bounding box differ significantly from the ground truth bounding box. The questionable detection is declared as a FN.

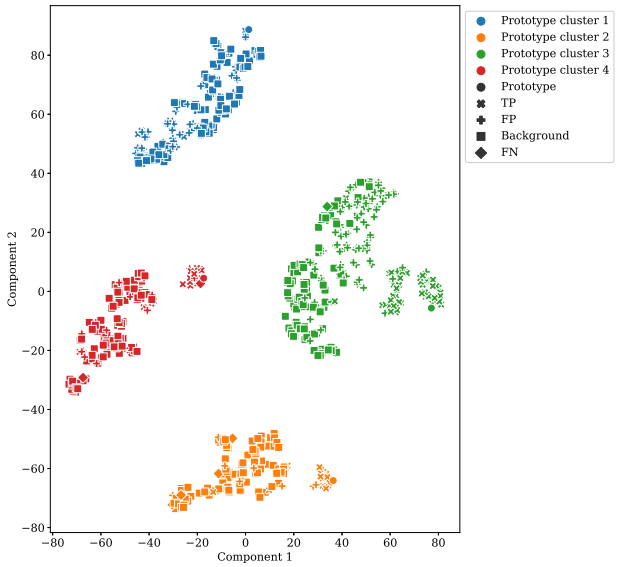

Figure 3. Visualization of latent representations with t-SNE in 2d plane. Due to the reasoning process of CSPP, a detected pedestrian is related to one specific latent representation. We can evaluate and assign the detection to TPs, FPs, FNs or background. The evaluation and visualization of detections and additional background representations are done for the first 16 images of the reasonable CityPersons validation dataset. The qualitative analysis is supported by a distance-based evaluation in Figure 4. The explicitly learned structure of the latent space is shown. Latent representations of true pedestrian centers (TPs) are close to prototypes but separated from the background.

In terms of a safety argumentation for a DNN, post-hoc analysis of an interpretable DNN are of great value because of their accessibility through requirements analysis. Nonetheless, the starting point must be given by an interpretable DNN.

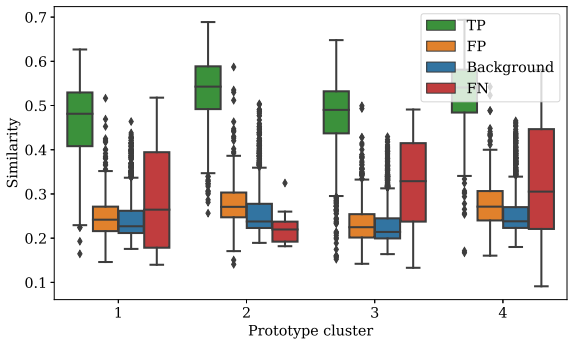

Figure 4. Boxplots for similarities of latent representations. In contrast to t-SNE analyzing the localization of latent representations in the latent space, the analysis with boxplots is distance-based. The distributions of similarities demonstrate that TPs (green) and background (blue) are well separated. The evaluation is conducted for the reasonable CityPersons validation dataset.

5. Conclusion

In this work, we provide a concept for a comprehensive requirements analysis in the context of DNNs. The meaningful decomposition of requirements and identification of DNN artifacts are enabled by inherent interpretability. To prove our concept, we propose the interpretable CSPP for pedestrian detection and establish explicit traces between defined requirements and DNN artifacts. The reevaluation of inherent interpretability is completed by providing qualitative evidence for the fulfillment of requirements.

However, quantifiable evidence derived from traceable and requirements-based test cases is crucial for a conclusive safety argumentation in the context of DNNs. Future work should focus on formulating applicable test cases for the defined requirements. Consequently, metrics must be identified that demonstrate test fulfillment.

Extending CSPP towards case-based reasoning of a pedestrian can further strengthen inherent interpretability. Multiple prototypical parts of pedestrians should be learned and clustered into instances in post-processing steps. Semi-supervised learning of parts can help with heavily occluded pedestrians. Case-based reasoning motivates the formulation of additional requirements and test cases.

Acknowledgment

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project “KI-Absicherung – Safe AI for automated driving”. We thank Seyed Eghbal Ghobadi and Philipp Heidenreich for the suggestions and discussions.

References

[1] Vincent Aravantinos and Frederik Diehl. Traceability of deep neural networks. *arXiv preprint arXiv:1812.06744*, 2018. 1

[2] Jan Bosch, Helena Holmström Olsson, and Ivica Crnkovic. Engineering ai systems: A research agenda. In *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems*, pages 1–19. IGI Global, 2021. 3

[3] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[4] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 8930–8941, 2019. 3, 4, 5

[5] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2011. 2

[6] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. 3

[7] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6569–6578, 2019. 2

[8] Kamaledin Ghiasi-Shirazi. Generalizing the convolution operator in convolutional neural networks. *Neural Processing Letters*, 50(3):2627–2646, 2019. 5

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[10] International Standards Organisation (ISO). Road vehicles - functional safety (iso 26262), 2018. 2

[11] International Standards Organisation (ISO). Road vehicles — safety of the intended functionality (iso/pas 21448), 2019. 2

[12] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1-3):1–308, 2020. 1

[13] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9525–9536, 2018. 3

[14] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 7

[15] Zachary C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2018. 3

[16] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new

perspective for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5187–5196, 2019. 2, 4, 5, 7

[17] Keivan Nalaie, Kamaledin Ghiasi-Shirazi, and Modhammad-R. Akbarzadeh-T. Efficient implementation of a generalized convolutional neural networks based on weighted euclidean distance. In *7th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2017. 5

[18] Oliver Willers, Sebastian Sudholt, Shervin Raafatnia, and Stephanie Abrecht. Safety concerns and mitigation approaches regarding the use of deep learning in safety-critical perception tasks. In *International Conference on Computer Safety, Reliability, and Security*, pages 336–350, 2020. 2

[19] RTCA. Do-178c: Software considerations in airborne systems and equipment certification, 2012. 1, 2

[20] 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. 3

[21] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015. 2

[22] T. Kong, Fu-Chun Sun, H. Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020. 2

[23] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 6

[24] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9627–9636, 2019. 2

[25] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. 7

[26] Fabian Utesch, Alexander Brandies, Paulin Pekezou Fouopi, and Caroline Schießl. Towards behaviour based testing to understand the black box of autonomous cars. *European Transport Research Review*, 12(1):1–11, 2020. 4

[27] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 2

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2016. 2

[29] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 2, 5, 6, 7

[30] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2