Carl von Ossietzky Universität Oldenburg

Fakultät II - Informatik, Wirtschafts- und Rechtswissenschaften
Department für Informatik

Inherent Interpretability for Safe Pedestrian Detection with Deep Neural Networks

Von der Fakultät für Informatik, Wirtschafts- und Rechtswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

angenommene Dissertation

von Herrn Patrick Feifel

geboren am 24.06.1991 in Bad Soden am Taunus

Gutachter:	Prof. Dr. Frank Köster
Weitere Gutachter:	Prof. Dr.-Ing. habil. Jorge Marx Gómez

Tag der Disputation: 10.12.2025

Abstract

Camera-based pedestrian detection utilizing deep neural networks (DNNs) represents a challenging milestone for automated driving. Changes in the perceived environment of an automated vehicle can expose functional insufficiencies of embedded DNNs, potentially resulting in harmful behavior. Accomplishing safe pedestrian detection means providing a structured safety argumentation based on verifiable evidence for the correct functionality of DNNs.

Hence, this work proposes a novel methodology for machine learning engineering to support a safety argumentation that addresses functional insufficiencies besides the systematic evaluation of DNNs. The methodology presents verifiable evidence with detection metrics designed for automated driving and assurance metrics to quantify the safety impact of inherent interpretability of DNNs.

This work demonstrates that leveraging inherent interpretability enables a safety argumentation. Modifying the DNN architecture and loss formulation to establish interpretable reasoning enforces a structured latent space that inherently provides evidence. The proposed DNN for concept-based pedestrian detection uses semantic concepts such as body parts to structure the latent space by enforcing a high intra-class concentration and inter-class separation. Ultimately, predictions become interpretable since detected body parts have significant similarities to semantic concept vectors. Moreover, structuring the latent space improves the generalization capabilities of DNNs. Despite training with unlabeled real-world data and labeled synthetic data, the DNN for concept-based pedestrian detection adapts successfully to the real world.

The development of baseline methods and progressive advancement to novel interpretable methods with state-of-the-art performance conforms to a generic DNN for pedestrian detection. This counteracts the prevalent lack of transparency in pedestrian detection and ensures valid benchmarking. Meanwhile, the systematic evaluation of methods focuses on safety-critical detection errors from the perspective of automated driving systems.

The presented experimental results demonstrate the substantial contribution of this work, highlighting the significance of inherent interpretability for safe pedestrian detection with DNNs.

Kurzfassung

Die kamerabasierte Erkennung von Fußgängern mithilfe von tiefen neuronalen Netzwerken ist ein wichtiger Meilenstein für das automatisierte Fahren. Veränderungen in der wahrgenommenen Umgebung eines automatisierten Fahrzeugs können funktionale Unzulänglichkeiten des eingebetteten tiefen neuronalen Netzwerkes aufdecken, was möglicherweise zu schädlichem Verhalten führt. Die sichere Erkennung von Fußgängern erfordert eine strukturierte Sicherheitsargumentation, die auf überprüfbaren Beweisen für die Abschwächung funktionaler Unzulänglichkeiten beruht.

Daher wird in dieser Arbeit eine neuartige Methodik für maschinelles Lernen zur Unterstützung einer Sicherheitsargumentation vorgeschlagen, die neben der systematischen Bewertung von tiefen neuronalen Netzwerken auch funktionale Unzulänglichkeiten berücksichtigt. Die Methodik stellt überprüfbare Beweise mit Erkennungsmetriken vor, die für das automatisierte Fahren entwickelt wurden, sowie Zuverlässigkeitsmetriken zur Quantifizierung der Abschwächung funktionaler Unzulänglichkeiten.

Diese Arbeit zeigt, dass die Nutzung der inhärenten Interpretierbarkeit eine Sicherheitsargumentation ermöglicht. Durch Modifikation der Netzwerkarchitektur und der Zielfunktion werden interpretierbare Argumente und ein strukturierter latenter Raum erzwungen, der inhärente Beweise für die korrekte Funktionalität liefert. Das vorgeschlagene tiefe neuronale Netzwerk für die konzeptbasierte Fußgängererkennung führt semantische Konzepte wie Körperteile ein, um den latenten Raum zu strukturieren, indem es eine hohe Konzentration innerhalb der Klasse und eine Trennung zwischen den Klassen erzwingt. Letztendlich werden die Vorhersagen interpretierbar, da die erkannten Körperteile signifikante Ähnlichkeiten mit semantischen Konzeptvektoren aufweisen.

Außerdem verbessert die Strukturierung des latenten Raums die Fähigkeit zur Generalisierung von tiefen neuronalen Netzwerken. Trotz des Trainings mit nicht annotierten Realdaten und annotierten synthetischen Daten, generalisiert das tiefe neuronale Netzwerk für konzeptbasierte Fußgängererkennung erfolgreich.

Die Entwicklung von Basismethoden und die schrittweise Weiterentwicklung zu vorgeschlagenen Methoden mit wettbewerbsfähiger Erkennungsqualität orien-

tiert sich an einem generischen Netzwerk für die Fußgängererkennung. Damit wird der vorherrschenden Intransparenz in der Fußgängererkennung wirksam entgegengewirkt. Die systematische Evaluierung der Methoden konzentriert sich auf sicherheitskritische Erkennungsfehler aus der Perspektive automatisierter Fahrsysteme.

Die vorgestellten experimentellen Ergebnisse zeigen den wesentlichen Beitrag der vorgeschlagenen Methodik und unterstreichen die Bedeutung der inhärenten Interpretierbarkeit für die Entwicklung von tiefen neuronal Netzwerken zur sicheren Fußgängererkennung.

Publications and Supervised Theses

Publications

Patrick Feifel, Frank Bonarens, and Frank Koster. Reevaluating the safety impact of inherent interpretability on deep neural networks for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29–37, 2021.

Patrick Feifel, Frank Bonarens, and Frank Köster. Leveraging interpretability: Concept-based pedestrian detection with deep neural networks. In *Computer Science in Cars Symposium*, pages 1–10, 2021.

Patrick Feifel, Benedikt Franke, Arne Raulf, Friedhelm Schwenker, Frank Bonarens, and Frank Köster. Revisiting the evaluation of deep neural networks for pedestrian detection. In *Proceedings of the Workshop on Artificial Intelligence Safety 2022*, 2022.

Patrick Feifel, Frank Bonarens, and Frank Köster. Domain adaptive pedestrian detection based on semantic concepts. In *Proceeding of VISIGRAPP (4: VISAPP)*, 2023.

Supervised Theses

Benedikt Franke. Improving pedestrian detection and evaluation by utilizing image segmentation. Master's thesis, Ulm University, Institute of Neural Information Processing, 2022.

Tobias Zielke. Pedestrian detection in occlusion scenarios with depth information. Master's thesis, Ulm University, Institute for Databases and Information Systems (DBIS), 2022.

Contents

1. Introduction

Recent innovations in artificial intelligence encourage the automation of advanced and safety-critical tasks such as automated driving. Generally, automated driving is expected to reduce emissions, ensure everyone's mobility and increase road safety in comparison to human operators [39, 80]. Focusing on vulnerable road user safety is one of the main drivers for developing automated driving systems. It is undisputed that the reliable perception of pedestrians prevents casualties and significantly increases the safety of vulnerable road users. Downstream tasks such as motion planning and acting of the automated driving system depend on accurate object localization and classification in the current driving environment.

Thus, the field of computer vision has developed methods to extract knowledge from images or image sequences [37]. Conventional computer vision methods rely on manual feature engineering and hard-coded rules of an expert to describe a given image or detect multiple objects in it. Especially in the open-world context, conventional methods do not achieve the required performance and generalization capabilities for complex perception tasks.

As a subfield of artificial intelligence, machine learning (ML) describes algorithms that enable machine learning models to learn tasks from collected data. Machine learning provides algorithms that replace conventional feature engineering with a flexible and automatic learning process [43]. Especially methods for complex tasks benefit from this approach since hand-designed features are hard to design. Most of the proposed learning algorithms can be considered either unsupervised or supervised. Supervised learning algorithms use sampled pairs of input data and annotated ground truth data from a training dataset. Supervised machine learning models are allowed to experience the realized error of a prediction compared to the given ground truth. Contrarily, unsupervised learning is conducted without access to manually labeled ground truth annotations.

Data is prepared either in structured or unstructured form. Structured tabular data directly possess informative features. When working with structured data,

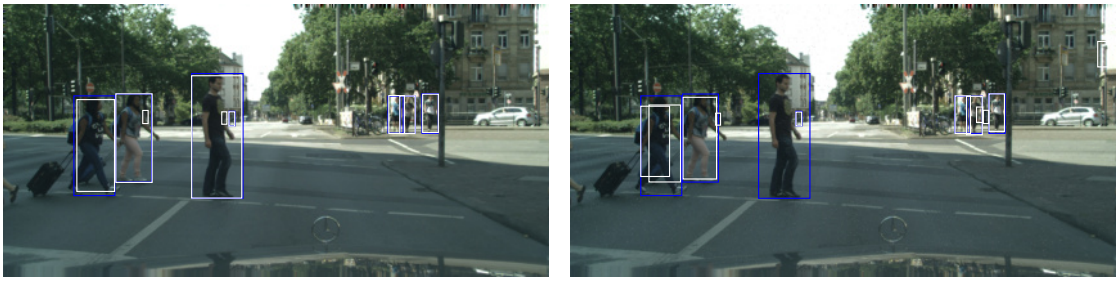
e.g., classifying a disease based on a patient's health status and symptoms, distinct methods must be employed. Machine learning models for structured data range from simple linear regression models to complex decision tree designs. Recently proposed supervised approaches exhibit human-comparable performance for object detection and semantic segmentation, representing important perception tasks of automated driving systems. Solving perception tasks involves processing visual sensor information. RGB images are an example of unstructured data. Learning from images is challenging since extracting discriminative features and creating predictions happen simultaneously. That means raw pixel values are encoded automatically. The resulting features have to be suitable for the respective perception task.

Originally designed for computer vision tasks, convolutional neural networks as a specific class of deep neural networks (DNNs) have been proposed [59]. Convolutional neural networks have a large number of different network layers mixed with convolutional operations. They unfold their full potential by learning from unstructured data.

State-of-the-art DNNs embedded in automated driving systems are mainly trained by supervised deep learning algorithms and real-world data. Deep learning describes an optimization strategy by which a DNN iteratively extracts meaningful patterns and knowledge from experiencing data. Meanwhile, DNN parameters are optimized based on input data samples and corresponding ground truth targets. Unfortunately, scaling supervised DNN methods for real-world applications is complicated since creating datasets with manually labeled ground truth targets is highly labor-intensive and costly.

Although automated driving offers significant benefits and newer prototypes catch the public's attention, no manufacturer has launched fully automated driving systems yet. The major drawback is that the safety of an automated driving system must first be argued and demonstrated. Currently, no conclusive and widely accepted safety argumentation ensures safety from a hardware and software perspective. Meanwhile, it must be assured that no unpredictable hazards arise from the unintentional behavior of automated vehicles.

The problem is that utilizing state-of-the-art DNNs exceeds the considerations of existing safety standards. Since DNNs become safety-critical in automated driving, verifiable evidence of the correct functioning has to be provided. The decision-making process in automated driving bases on a reliable perception of the complex environment. Although any safety consideration starts from the systems perspective, arguing safety for automated diving is first related to perception.



(a) Original image. (b) Image with salt and pepper noise.

Figure 1.1: Natural image distortions cause pedestrian detection errors. Inference results for CityPersons validation dataset of a state-of-the-art DNN for pedestrian detection with ground truth (blue) and detection bounding boxes (white).

1.1 Motivation

To give an intuitive example, Figure 1.1 shows inference results for the CityPersons dataset [123] of a state-of-the-art DNN for pedestrian detection. Simply by adding small perturbations with salt and pepper noise, the detection performance decreases substantially. Surprisingly, the pedestrian right in front of the automated vehicle is not detected anymore. Due to the image distortion, the DNN is no longer able to predict correct bounding boxes for all vulnerable road users in the driving path. Consequently, automated vehicles might consider unsafe control actions with potentially hazardous events.

Generally, the safety of new technologies is argued by following the procedure of well-established safety standards. ISO 26262 [32, 33] is the safety standard for the functional safety of automotive electrical and/or electronic systems. Therein, functional safety is defined as the "absence of unreasonable risk due to hazards caused by malfunctioning behavior" [32]. In the context of DNNs, ISO 26262 can help to assure functional safety for software components such as the deployed model, the pre-processing and post-processing of the detection pipeline.

The loss of functional safety attributes to an "abnormal condition that might lead to an error in a hardware or software component" [32]. However, functional safety only considers if an input image is correctly processed and predictions outputted.

Besides software bugs or sensor defects, automated driving systems could also experience abnormal conditions leading to undetected pedestrians. Simple lighting variations during twilight or motion blur effects can cause serious pixel changes in

an image. In summary, while a DNN processes an image and outputs detections, it may not correctly detect all safety-critical pedestrians. Functional insufficiencies, inherent in the trained DNN, can cause unexpected failures [6]. Moreover, Figure 1.1 demonstrates that functional insufficiencies such as the lack of generalization and lack of robustness are closely related and cannot be clearly separated.

Motivated by performance limitations of machine learning models, ISO 21448 has introduced the safety of the intended functionality (SOTIF) [36]. SOTIF defines safety as the absence of unreasonable risk due to functional insufficiencies. DNN failures depend on functional insufficiencies such as a lack of interpretability and the occurrence of a triggering event. An automated driving system failure manifests in flawed pedestrian detection, where safety-critical pedestrians are missed.

Since state-of-the-art DNNs for pedestrian detection are black boxes, the inner working and extracted features are not intuitively human-understandable. As a consequence, the reason for detection errors cannot be identified and eliminated. There are ongoing initiatives such as KI Absicherung[1] as part of the KI Familie that try to overcome these limitations. KI Absicherung is a German consortium project funded by the Federal Ministry for Economic Affairs and Climate Action of Germany. The consortium attempts to formulate a SOTIF-based safety argumentation for the pedestrian detection use case.

However, the interpretation of SOTIF and its application to perception tasks for automated driving remains an open question. Recent research proposes to provide a structured safety argumentation with verifiable evidence to state safety according to SOTIF [6, 36]. Hence, appropriate methods that mitigate functional insufficiencies must be developed. Verifiable evidence based on application-oriented detection and assurance metrics has to demonstrate successful mitigation.

The motivation of this work can be summarized in the following research problem:

The safety argumentation of DNNs for pedestrian detection in automated driving cannot be formulated since no methods and metrics exist to provide verifiable evidence.

[1] translation: Artificial Intelligence Safeguarding, `https://www.ki-absicherung-projekt.de/en/`

1.2 Research Questions

As indicated by the research problem, the overarching goal of this work is two folded. On the one hand, methods for mitigating functional insufficiencies and validation methods for error analysis must be developed. On the other hand, verifiable evidence must be provided. Evidence must be quantifiable, thus, has to be based on metrics.

Functional insufficiencies can cause pedestrian detection errors such as false negatives. A false negative is given when no detection bounding box (white boxes in Figure 1.1) can be matched to a ground truth bounding box (blue boxes in Figure 1.1). Inevitably, mitigation methods must be developed to guarantee the safety of vulnerable road users. In this work, mitigation methods are considered as DNNs with necessary adaptations in terms of their architecture, but also loss formulation and training strategy. Thus, no measures are applied post-hoc to DNNs that have already been trained. Instead, methods with novel training strategies or DNN architectures mitigating functional insufficiencies are proposed.

The successful mitigation and absence of unreasonable errors must be demonstrated by verifiable evidence with appropriate metrics. In the course of this work, potential detection errors of DNNs for pedestrian detection have been categorized and a systematical evaluation based on application-oriented performance metrics is proposed.

The research questions derived from the consideration are presented below:

I **How can methods and metrics related to the lack of interpretability in pedestrian detection support a safety argumentation?** It is still an open question, how key aspects of existing safety standards can be applied to a safety argumentation for DNNs. Although safety considerations for software define safety elements such as decomposition, traceability and the mitigation of functional insufficiencies theoretically, it remains unclear how these concepts can be used for safe pedestrian detection. Moreover, establishing a requirements-based software development process seems impossible due to the black-box nature of state-of-the-art DNNs. The first research question aims at a methodology to support a safety argumentation concerning the lack of interpretability and lack of generalization.

II **How should the performance of a deep neural network for pedestrian detection be systematically evaluated?** From the perspective of automated driving systems, it makes sense to assign different levels of risk to

vulnerable road users. Intuitively, a pedestrian crossing the street directly in front of the automated vehicle must be detected with higher accuracy than a person sitting on a bench at a great distance. Detection metrics that incorporate a meaningful categorization of pedestrians must be developed to enable a fair and application-oriented comparison of DNNs. The second research question aims to provide verifiable evidence of performance metrics for safety-critical pedestrians.

III **How can methods enable the inherent interpretability of DNNs and what verifiable evidence emerges?** This work focuses on enabling inherent interpretability by adapting state-of-the-art DNN architectures and proposing novel loss formulations and training strategies for DNNs for pedestrian detection. Methods for post-hoc interpretability are out of the scope of this work. The third research question aims to develop methods to mitigate the lack of interpretability and quantify the effect.

IV **How far does inherent interpretability improve the generalization capabilities of DNNs?** From a human perspective, it is undoubted that an arm in a synthetic image is similar to an arm in a real-world image. However, the domain shift between real-world and synthetic data exhibits great performance deviations. Inherent interpretability is expected to have a positive effect on the generalization capabilities of DNNs. The fourth research question aims to develop methods to mitigate the lack of generalization and demonstrate improvements with verifiable evidence.

1.3 Contribution

To answer the identified research questions, the contributions of this work can be summarized as follows.

A failure-driven ML engineering approach for DNNs for pedestrian detection to support a safety argumentation is proposed. The proposed methodology consists of a systematic evaluation of safety-critical detection errors and methods to mitigate functional insufficiencies. This contribution targets the first research question.

A novel systematic evaluation of DNNs for pedestrian detection [29] based on the categorization of ground truth and detection bounding boxes has been developed during the course of this work. Thus, potential DNN detection errors can be formally defined. Based on that, application-oriented detection metrics concerning

safety-critical pedestrians for automated driving are evaluated. This contribution targets the second research question.

A DNN for prototype-based pedestrian detection (PPD) as a novel inherently interpretable DNN architecture for pedestrian detection is proposed [27]. PPD is trained end-to-end and incorporates interpretable reasoning, using squared L^2 distances within the latent space between extracted latent representations and unsupervised prototypes. This contribution targets the third research question.

The idea of PPD has been extended to explicitly-learned and pre-defined semantic concepts. This line of work considers body parts as semantic concepts of humans. The novel DNN architecture for concept-based pedestrian detection (CPD) enables an interpretable reasoning process to predict an auxiliary body part segmentation. CPD's post-processing fuses the segmentation with predicted 2D bounding boxes by applying simple decision rules [26]. This contribution targets the third research question.

It is shown that inherently interpretable DNNs such as PPD and CPD are motivated by techniques from deep metric learning. Consequently, they offer superior generalization capabilities. By proposing a DNN for concept-based domain adaptation (ConDA), the performance on real-world data is substantially increased while only using labeled ground truth annotations for synthetic data and real-world images [28]. This contribution targets the fourth research question.

As an outlook, Figure 1.2 shows inference results for a domain-adapted DNN that is inherently interpretable. In contrast to the previous Figure 1.1, body parts are detected as semantic concepts of a pedestrian. Although the added salt and pepper noise is still a challenge and leads to inaccuracies, the safety-critical pedestrian is correctly detected.

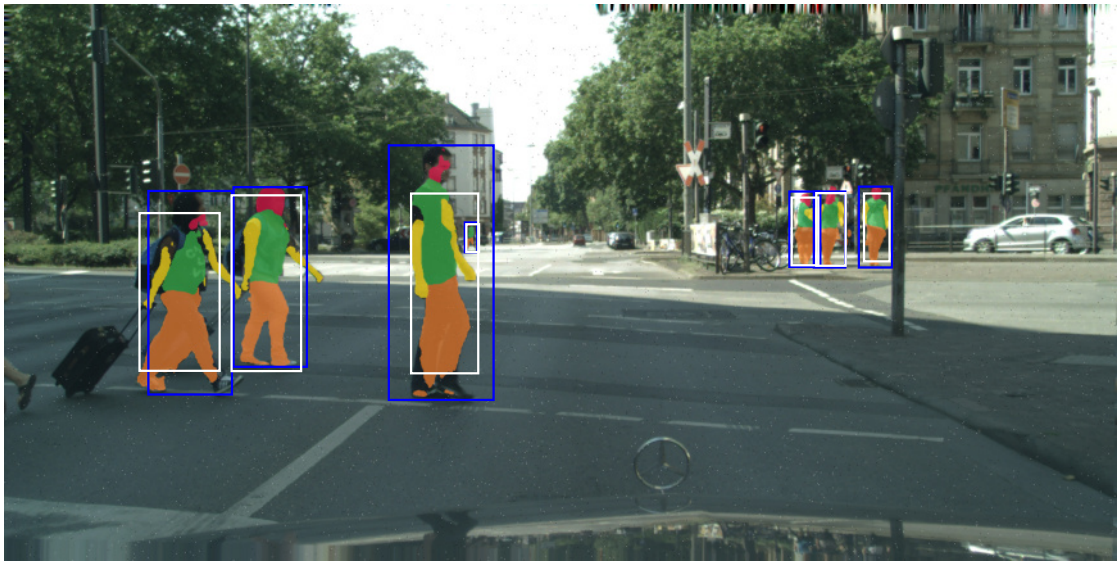


Figure 1.2: The proposed DNN for concept-based domain adaptation (ConDA) for pedestrian detection withstands natural perturbation. ConDA is able to overcome the lack of robustness, although it has been only trained with ground truth for synthetic data and adapted to real-world data with unsupervised training. The DNN is trained on synthetic data from SynPeDS [97] and adapted to real-world data from Cityscapes by unsupervised learning without annotated ground truth data.

1.4 Outline

This work is highly interdisciplinary and addresses safety in automated driving for pedestrian detection using methods for the inherent interpretability of DNNs.

Consequently, the second chapter describes a broad basis of related work. First, existing system architectures for automated driving are investigated and relevant subsystems and components for pedestrian detection are identified. Then, recently proposed approaches for the safety argumentation towards SOTIF and existing safety standards for software development for safety-critical systems are discussed. Furthermore, state-of-the-art DNNs for pedestrian detection are introduced and benchmarked. Finally, preliminary work for the mitigation of functional insufficiency is identified. Here, methods for inherent interpretability and unsupervised domain adaptation are described.

In the following, a methodology for ML engineering to support the safety argu-

mentation of DNNs for pedestrian detection, building on methods and metrics, is proposed in the third chapter. Meanwhile, an initial structured safety argumentation is described where safety goals are linked to proposed methods and verifiable evidence with metrics. In addition, the identification of safety-critical pedestrians based on assumptions from a system view enables the application-oriented systematic evaluation of developed methods [29].

To achieve the identified safety goals, methods for inherent interpretability [27, 26] and domain adaptation [28] for pedestrian detection are developed. The fourth chapter describes in detail how DNNs must be adapted to enable inherent interpretability based on unsupervised prototypes or supervised semantic concept vectors. Eventually, the domain adaptation based on semantic concepts is outlined and the improved generalization capabilities are measured.

The fifth chapter describes the experimental setup and presents the experimental results of all developed baseline methods and proposed methods. The results of the systematic evaluation with detection metrics and mitigation of functional insufficiencies measured by assurance metrics, are linked to verifiable evidence to support the safety argumentation of inherently interpretable DNNs for pedestrian detection.

Finally, the last chapter concludes this work and possible future work is discussed.

2. Related Work

In the following, a comprehensive review of related work is presented. This work focuses on pedestrian detection as an important aspect of automated driving. Due to the emphasis on safety, the investigated system architecture of the automated driving system has to be defined. Although no widespread system description for automated driving exists, there is some related work giving a well-defined starting point.

If relevant subsystems with known interdependencies have been identified, a structured safety argumentation can be proposed. To argue safety of automated driving systems, existing safety standards come into consideration. Despite the description of initial safety aspects, intensive work is still ongoing to standardize various safety aspects of automated driving in the context of machine learning.

There is a large body of research on pedestrian detection with DNNs. Numerous state-of-the-art DNNs for pedestrian detection have already been proposed. Moreover, a widespread benchmarking protocol has been established, promoting a continuous improvement of the state of the art.

This work proposes to leverage inherent interpretability for safe pedestrian detection. Hence, different methods and aspects of interpretability must be investigated and discussed. Currently, there is no widely accepted terminology for this research field, since it ranges from explainable or trustworthy artificial intelligence to machine learning model interpretability. Likewise, interpretability or explainability is hard to distinguish from work on deep metric learning and generalization of DNNs. Meanwhile, the tackled computer vision tasks in this research field are diverse.

As this work points out, similar techniques are used for interpretability as well as to increase the generalization capabilities of DNNs. The unsupervised domain adaptation addresses generalization in terms of utilizing unlabeled data from a target domain. To understand the close connection to interpretability, a review of state-of-the-art methods for unsupervised domain adaptation is given.

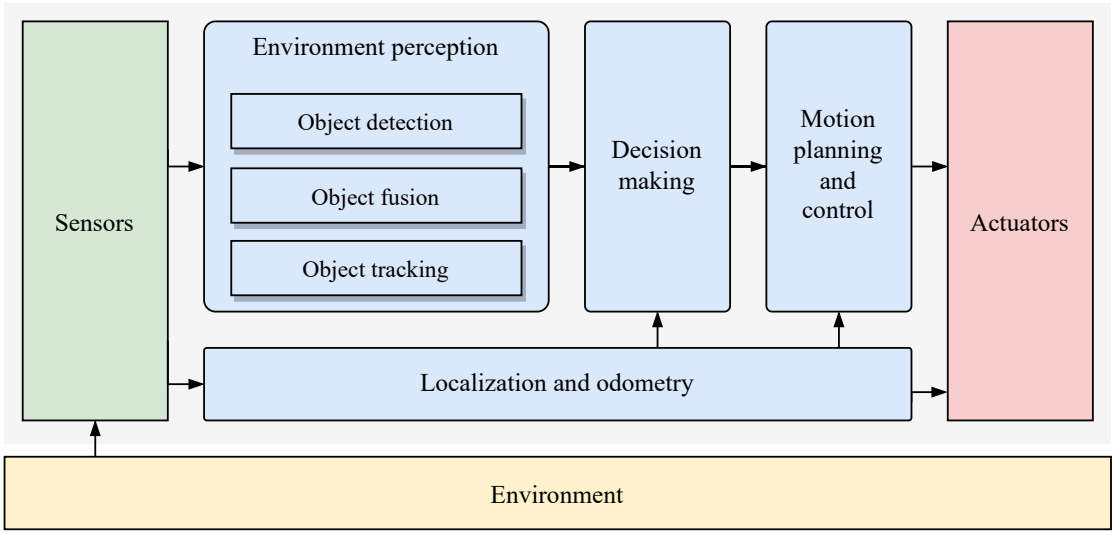


Figure 2.1: Exemplary system architecture of an automated driving system [87, 101].

2.1 Automated Driving System

Automated driving requires additional activities enlarging well-established automotive engineering disciplines. Different sensors are used for the perception of the driving environment. Novel methods for perception and planning enabled by machine learning are combined in the automated driving system. The automated driving system describes relations between subsystems for various tasks regarding sensing, planning and acting of automated vehicles.

Figure 2.1 shows the arrangement of subsystems inspired by the generic vehicle architecture of SET Level [87] and BerthaOne [101]. Most relevant for this work is the perception subsystem which creates an internal representation of the environment by using different sensors such as a monocular camera, stereo camera, LiDAR or radar.

2.2 Pedestrian Detection

Pedestrian detection is a well-established research field and has high relevancy for automated driving. Since methods for pedestrian detection automatically process visual information and extract knowledge, pedestrian detection is a computer vision discipline. As a part of the subsystem for moving object detection, methods

for pedestrian detection predict 2D bounding boxes for possibly shown pedestrians in a given image. Comparable to a sliding window approach, multiple binary classification tasks have to be performed for different image regions.

Pedestrian detection can be considered the binary case of object detection. The term detection always refers to the simultaneous classification and localization of objects. Detecting pedestrians means that the machine learning model only needs to distinguish between the absence and the presence of pedestrians. Only two classes, background and pedestrian, are considered.

In the past, it has been shown that pedestrian detection can be accomplished by conventional methods for computer vision [20, 21]. Conventional computer vision methods are even used in advanced driver assistance systems. However, the limited performance and generalization capabilities for the open-world context of automated driving contradict the application in automated driving systems. Recent research has shown that DNNs for pedestrian detection significantly outperform conventional computer vision methods and meet real-world complexity [121]. DNNs for either object or pedestrian detection can be systematized based on the following characteristics: anchor dependence and DNN architecture.

Anchor-based Deep Neural Networks Early DNNs for object detection such as SSD [110] or Faster R-CNN [93] utilize default anchors to predict 2D bounding boxes. These anchor-based DNNs predict relative deviations of bounding box coordinates from default anchor boxes. Since the optimal configuration of scales and aspect ratios of default anchors is unknown, hyperparameters for their configuration are highly tuned on specific datasets. Hence, the final performance strongly depends on the quality of the default anchor box generation. The application of anchors has been widely applied by DNNs for pedestrian detection, e.g., ALFNet [69], Faster R-CNN [93], Beta R-CNN [116], PRNet [95] and Cascade R-CNN [7].

Anchor-free Deep Neural Networks Recent research focuses more on anchor-free DNNs yielding simpler DNN architectures. Since these methods do not rely on default anchor boxes, the need for a priori mining of training data to initiate anchor scales and sizes is eliminated. Anchor-free DNNs predict the confidence in the existence of key points. Most commonly the center point of an object or pedestrian is used.

Predictions of additional regression heads, which correspond to the position of the estimated key point, are used to decode the scale (height and width) and center offset of 2D bounding boxes. The current trend brings more high-

performing anchor-free approaches that achieve state-of-the-art performance and surpass anchor-based models for object detection, e.g., CenterNet [24], FoveaBox [99], FCOS [102] and CenterNet [125], and pedestrian detection, e.g., CSP [70], BGCNet [61], APD [119] and F2DNet [54].

One-stage and Two-stage Deep Neural Networks With respect to the architecture, DNNs can be further grouped into one-stage or two-stage DNNs. Two-stage DNNs first predict region proposals followed by a classification step. Nonetheless, two-stage DNNs can also be designed as an end-to-end approach. Proposing a large number of interest image regions is time-consuming. Although the number of proposed regions is another hyperparameter, the inference speed of one-stage DNNs is usually higher.

Contrarily, one-stage DNNs predict the confidence in key points and scales of a bounding box without any intermediate steps. For example, CSP [70] extracts and combines high-level semantic features of multiple scales to predict a center confidence, center offsets and height of a pedestrian in one step. A comprehensive categorization of DNNs for pedestrian detection based on the application of anchors or the number of stages can be found in Table 2.1.

Performance Evaluation The most common metric to quantify performance and enable benchmarking for pedestrian detection is the log-average miss rate $LAMR(c)$ [21, 5, 20, 107]. The $LAMR(c)$ describes a compressed performance metric that considers the miss rate $MR(c)$ besides the false positives per image $FPPI(c)$. Both metrics implicitly take the confidence threshold c into account.

To calculate the $LAMR(c)$, all detections for a given validation or test dataset must be sorted in descending order according to the predicted confidence scores. After identifying true positives, false negatives and false positives, the $MR(c)$ and $FPPI(c)$ can be calculated. Then, the $LAMR(c)$ is defined as

$$LAMR(c) = \exp\left(\frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}}\log MR^f(c)\right) \tag{2.1}$$

where $\mathcal{F} = \{10^{-2}, 10^{-1.78}, \ldots, 10^{0}\}$ describes predefined $FPPI(c)$ values at which the miss rate is taken.

The dependence on the confidence threshold is usually avoided by choosing an appropriately low confidence threshold, i.e., $c = 0.01$. Thus, the mentioned metrics are simplified to MR, $FPPI$ and $LAMR$.

Table 2.1: Performance and characteristics of state-of-the-art DNNs for pedestrian detection. The middle columns outlines important characteristics. The right column shows the performance in terms of the log-average miss rate $LAMR$ [%] for the reasonable subset of the CityPersons [123] validation dataset.

Method	Anchor-based	Stages	Reasonable
ALFNet [69]	✓	one	12.0
CSP [70]	×	one	11.0
NOH-NMS [124]	✓	two	10.8
RepLoss [107]	✓	two	10.9
PRNet [95]	✓	one	10.8
Beta R-CNN [116]	✓	two	10.6
NMS-Loss [74]	✓	two	10.1
Cascade R-CNN [7]	✓	two	9.2
BGCNet [61]	×	one	8.8
APD [119]	×	one	8.8
F2DNet [54]	×	two	**8.7**

Subset-based Evaluation Due to small pedestrian heights, various kinds of occlusions and crowd appearances in images, pedestrian detection is subject to high complexity. Pedestrians show a large variety of posture and perspective, in terms of shape, and of appearance and illumination, in terms of texture. To enable a fair performance comparison, Dollar et al. [20] has proposed a reasonable evaluation setting with corresponding subsets of validation datasets. The reasonable subset only considers pedestrians with a height greater than 50 pixels.

Subsequently, Wang et al. [107] has refined the settings for the CityPersons dataset [123] and excluded annotations with an occlusion rate larger than 0.35. Although arbitrary thresholds are used, the subset-based evaluation of DNNs leads to a more realistic view of the application-oriented performance. Combining both evaluation settings, a reasonable bounding box has a height of more than 50 pixels and an occlusion rate of 0.35 or less.

The performance evaluation in terms of the $LAMR$ for the reasonable subset of the CityPersons validation dataset defines the single important benchmark for DNNs for pedestrian detection. Table 2.1 gives an overview of the performance of state-of-the-art DNNs for pedestrian detection.

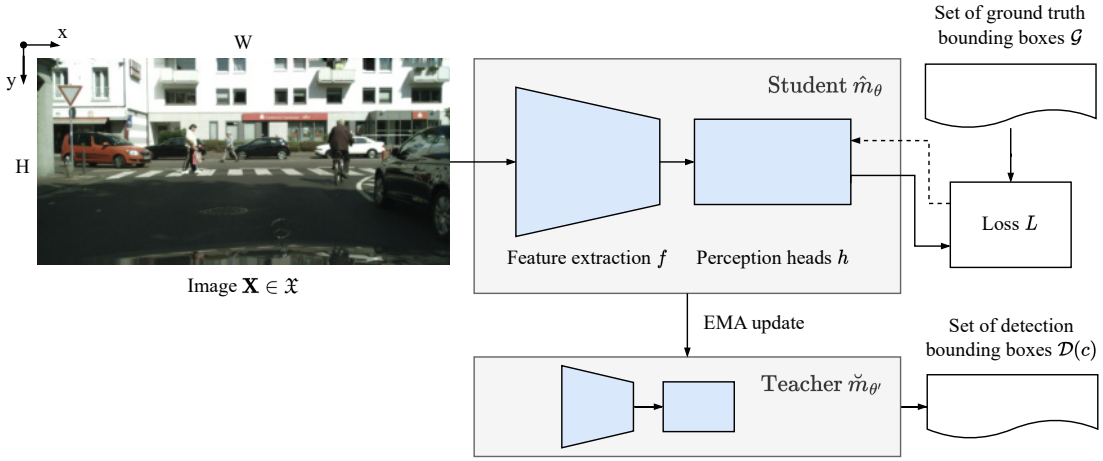


Figure 2.2: Teacher-student framework of state-of-the-art DNNs for pedestrian detection.

Teacher-Student Framework A widespread training strategy in the field of pedestrian detection adopts the work on the performance-improving teacher-student framework for DNNs [100]. The authors have proposed to continuously update the parameters of a teacher DNN $\breve{m}_{\theta'}$ during the training of a student DNN $\hat{m}_\theta$. The teacher is initialized with the same parameters as the student with parameters θ.

As shown in Figure 2.2, the student $\hat{m}$ is trained with a supervised loss L. During the training, only the parameters of the student are optimized by back-propagation. The parameters of the teacher are defined as the exponential moving average of the student's parameters:

$$\breve{m}_{\theta'} : \theta'_{i+1} \leftarrow \alpha\, \theta'_i + (1-\alpha)\, \theta_i \tag{2.2}$$

2.3 Safety Argumentation

Although automated vehicles equipped with automated driving systems have intuitive advantages, it must be argued that the system does not exhibit an unacceptable risk to humans. To establish a common safety understanding in the automotive industry, safety standards are defined.

The functional safety of road vehicles is described in detail by ISO 26262 [32] considering multiple safety aspects such as the definition of automotive integrity

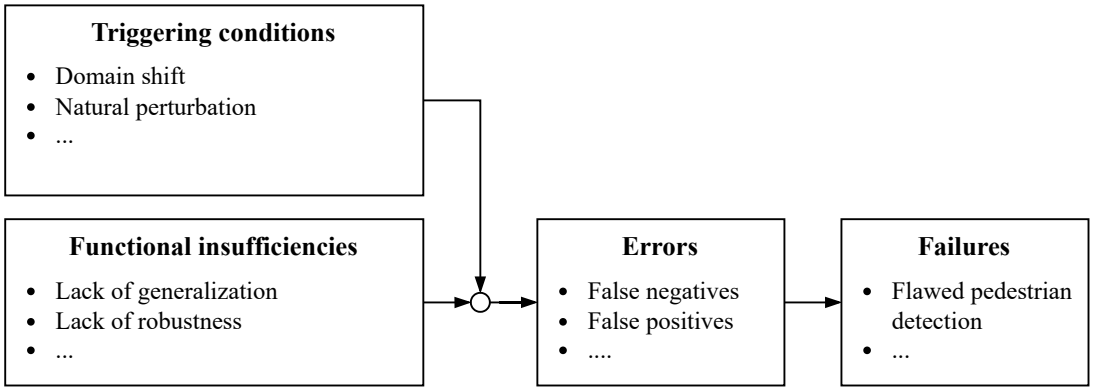


Figure 2.3: Failure model for an automated driving system [6, 112]. Failures are a result of triggering conditions unveiling functional insufficiencies leading to errors.

levels or general guidelines for product development on hardware, software and system level. ISO 26262 can be considered the most widespread safety standard in the automotive industry.

Concerning automated driving, the complexity of the safety argumentation drastically increases. Originally planned as a part of the ISO 26262, ISO 21448 addresses occurring safety issues although the system does not have faults in the sense of the ISO 26262. ISO 21448 construes the safety of the intended functionality (SOTIF) of a system by guaranteeing the absence of unreasonable risk [36]. In that notion, risk results from functional insufficiencies inherent in system components that might use machine learning.

Well-described functional insufficiencies of machine learning models are the lack of robustness and lack of generalization [112, 6]. Figure 2.3 outlines the failure model motivated by SOTIF [6, 112]. Here, functional insufficiencies are exposed by triggering conditions that can cause safety-critical detection errors of pedestrian detection in an automated driving system.

The proposed safety formalism demands risk-reducing activities during the specification, design, verification, validation and operation phase of systems that use machine learning. Eventually, SOTIF provides a coherent failure model and initial safety guidelines for automated driving on the vehicle level. Consequently, there is still a research gap on the component level of the automated driving system. It remains unclear how verifiable evidence for machine learning models that support a structured safety argumentation can be generated.

ISO/AWI PAS 8800 [34] is an ongoing initiative working explicitly on safety and artificial intelligence for road vehicles. It describes a framework for the complete development lifecycle of artificial intelligence. Furthermore, ISO/CD TS 5083 [35] is focused on the top-level safety goals of an automated driving system. ISO/AWI PAS 8800 and ISO/CD TS 5083 are currently under development, thus, no further details can be provided. Especially, ISO/AWI PAS 8800 seems to have a similar focus on mitigating functional insufficiencies of artificial intelligence and evidence-driven safety argumentation as described in this work.

The previously mentioned safety standards have a clear focus on road vehicles in common. Apart from the automotive industry, other fields already use software solutions in safety-critical applications. The DO-178C [31] offers advanced considerations for the development of safety-critical software. Evidence is based on artifacts that are generated throughout a transparent software life cycle. Artifacts describe data from the software life cycle as by-products for every step in the software life cycle. DO-178C argues safety by ensuring a safe software development process. A key concept is the bi-directional traceability for the decomposition of high-level requirements into low-level requirements and software. Thus, source code is directly developed from the definition of low-level requirements.

Despite a vast amount of work that has been already spent to develop a safety argumentation for automated driving systems, no widely accepted safety concept could be established yet. The research agenda of artificial intelligence engineering [4] conceptualizes the development process of artificial intelligence for safety-critical systems. The new field of research is focused on domain-specific characteristics of artificial intelligence that complicate the application of well-established safety considerations such as ISO 26262 [32] or DO-178C [31]. Nonetheless, there is some related work that outlines the potential of a concept embedding analysis for the safety assurance of DNNs [91]. Such an analysis is applied in explainable artificial engineering [44] or post-hoc interpretability.

2.4 Machine Learning Model Interpretability

Doshi-Velez and Kim [22] define interpretability as "... the ability to explain or to present in understandable terms to a human". Interpretability enables humans to understand the reasons for predictions of a machine learning model and contributes to trustworthiness. Interpretability requires individual adaptations for the investigated machine learning model. As one can imagine, even different notions of interpretability might exist in terms of DNNs or decision trees.

Since this work focuses on the use case of pedestrian detection in automated driving systems, only DNNs are considered. Lipton [68] defines two non-absolute dimensions of interpretability: post-hoc interpretability and transparency [68]. Closely related to the notion of transparency, the concept of inherent interpretability is introduced by Rudin [88].

Post-hoc Interpretability Methods for post-hoc interpretability generate explanations for the behavior of already trained models. Thus, explaining means applying surrogate explanatory models. Although interpretability has been introduced as a two-sided concept, recent research activities focus largely on post-hoc interpretability, commonly referred to as explainable artificial engineering [44]. Proposed methods focus on local approximations on parameter level with surrogate models that provide so-called heatmaps or saliency maps [90]. These saliency methods have limited capabilities to represent holistic interpretability [52]. Despite their well-described limitations, methods for post-hoc interpretability can contribute to a versatile understanding of a trained DNN.

Inherent Interpretability In contrast to post-hoc interpretability, interpretability can also be inherently integrated into DNNs. The analysis of the interpretability on the model, parameter and training level is described by the transparency. Transparency focuses on inherent properties of the reasoning process and is most similar to the inherent interpretability mentioned by Rudin [88, 89]. Rudin asserts that the question of whether a model is inherently interpretable is highly domain-specific and must be examined in terms of the applied task.

However, the inherent interpretability of all layers in a DNN seems to be not feasible and not even necessary. Instead, it focuses on the reasoning of a prediction and the involved representations. Representations or prototypes are intermediate tensors that are optimized with specifically designed loss formulations fostering interpretability. The identification of and special training interest in intermediate representations immediately provide human-understandable explanations.

Prototypical Part Network One of the first proposed inherently interpretable DNNs for fine-grained bird classification is the prototypical part network (ProtoPNet) [9]. The general idea of the distance-based reasoning proposed by ProtoPNet is shown in Figure 2.4.

ProtoPNet applies the generalized convolution with squared L^2 distances [41, 79] to estimate the similarity between the l-th prototype $\mathbf{x}^l$ as a vector and an

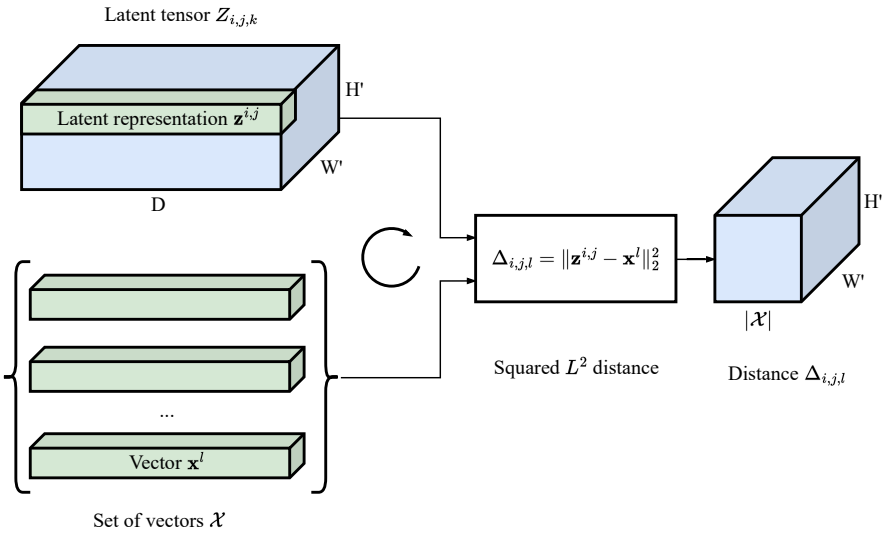


Figure 2.4: Illustration of generalized convolution [41] with calculation of squared L^2 distances between a vector $\mathbf{x}^{\mathrm{l}}$ and latent representations $\mathbf{z}^{\mathrm{i,j}}$.

extracted latent representation $\mathbf{z}^{\mathrm{i,j}}$. Both vectors can also be written in the index notation, $\mathrm{x}_{\mathrm{k}}^{\mathrm{l}}$ and $\mathrm{z}_{\mathrm{k}}^{\mathrm{i,j}}$.

The classification becomes interpretable since predictions are based on the similarity of learned prototypes that are characteristic for specific classes. Effectively, latent representations are clustered and separated within the latent space of the trained DNN. Eventually, the receptive field of a latent representation can be used to identify the image region that is most similar to the characteristics of a bird species.

Deep Metric Learning In contrast to supervised deep learning for object detection or pedestrian detection, computer vision tasks such as face recognition must be able to handle open-set classification problems. That means unseen faces must be classified without seeing them during the training. This one-shot learning problem is solved by building an embedding space with training images and analyzing the position of new samples in the space for a chosen metric. Hence, discriminative features must be learned to identify similar faces and ultimately verify their conformity [111, 72, 71, 103, 104].

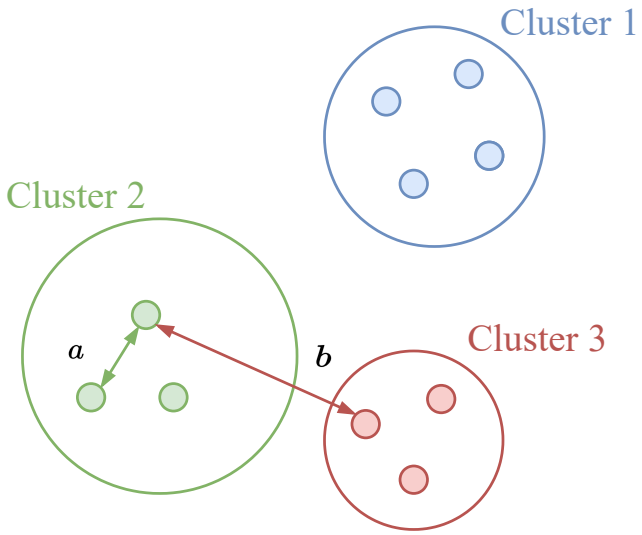


Figure 2.5: Illustration of silhouette coefficient [86] to estimate the cluster performance with intra-class distance a and inter-class distance b.

The effect of deep metric learning and the facilitated structure of the latent space can be evaluated by well-known cluster metrics such as the silhouette coefficient [86]. Figure 2.5 gives an impression of how the silhouette coefficient defines parameters a and b to quantify the cluster performance of the shown data points. The common goal with deep metric learning is to maximize the inter-class distance b and minimize the intra-class distance a. Eventually, well-separated clusters are defined by discriminative features in the latent space.

2.5 Unsupervised Domain Adaptation

The objective of unsupervised domain adaptation (UDA) addresses insufficient generalization of DNNs by enhancing supervised learning with unsupervised learning from unlabeled data. Driven by the easy accessibility of synthetic data and insufficient generalization capabilities of state-of-the-art DNNs, methods for UDA use synthetic data with automatically generated ground truth data and unlabeled real-world data. Since no ground truth is available for real-world data, representing the target domain, methods for unsupervised learning must be developed. Most commonly, UDA methods use self-training techniques. Related work is either focused on semantic segmentation or object detection.

ProDA [122] and SePiCo [114] for semantic segmentation leverage prototypes as feature vectors in the latent space to rectify noisy pseudo labels. By incorporating distances to prototypes, the softmax-based class predictions can be improved, resulting in enhanced classification and generalization capabilities even when faced with class imbalances or outliers. Thus, domain-invariant and well-generalizing representations are learned. Closely related to deep metric learning, these self-training techniques substantially improve the generalization performance on the real-world Cityscapes dataset [14]. In addition to that, DAFormer [48] utilizes data augmentation techniques incorporating a priori knowledge of class distribution and object properties such as shape and texture for the target domain.

Self-paced learning for object detection [96] uses a teacher-student framework for domain adaptation. Additionally, a style-transfer stage is applied in order to transform source images to the target domain and to alleviate the negative influence of noisy pseudo labels in self-training. UMT [19] follows a similar approach wherein both the teacher and student learn from pseudo labels. However, it applies adversarial training for cross-domain distillation in addition.

3. Methodology

Automated driving relies on safety-critical systems utilizing DNNs for various perception tasks. To ensure the safety of the intended functionality (SOTIF), the absence of unreasonable risk due to functional insufficiencies of DNNs has to be argued [36]. Therefore, a novel methodology for ML engineering utilizing methods for inherent interpretability is proposed. Ultimately, the methodology supports a structured safety argumentation of DNNs for pedestrian detection.

First, an exemplary system architecture of automated driving systems following [87, 101] is introduced. It is shown that the subsystem for object detection is the most relevant for this work. Since a DNN is the main component of an object detection subsystem, it must be specifically configured for pedestrian detection.

Related work on a failure model for automated driving [36, 6] represents the starting point for the proposed failure-driven ML engineering. On the one hand, safety-critical failures are evaluated from a system perspective. On the other hand, the mitigation of functional insufficiencies is conducted by introducing novel methods for inherent interpretability for pedestrian detection. The mitigation success is measured by assurance metrics providing verifiable evidence. The ML engineering is enhanced by a structured safety argumentation drawing the connection of proposed methods, metrics and safety goals.

Eventually, methods for the inherent interpretability of DNNs for pedestrian detection and promising safety implications are described. Related work on inherent interpretability for fine-grained classification tasks with ProtoPNet [9] is extended to provide interpretable reasoning for pedestrian detection based on a structured latent space. Meanwhile, it is shown how failure-driven ML engineering, which mainly targets SOTIF aspects [36], can be enriched by safety aspects of the DO-178C [31] that are aligned with inherent interpretability.

Then, safety-critical failures of automated driving systems are investigated. To provide verifiable evidence for the safety argumentation, application-oriented detection metrics must be used for the systematic evaluation of safety-critical

detection errors. The systematical evaluation is based on the prior categorization of ground truth bounding boxes. Furthermore, different error types of detection bounding boxes, resulting from the inference of a trained DNN, are identified. The systematic evaluation proposes different error categories that are used for the definition of application-oriented detection metrics.

The combination of proposed detection metrics and assurance metrics for the interpretable methods creates a set of verifiable evidence for the safety argumentation of DNNs for pedestrian detection.

3.1 Pedestrian Detection in Automated Driving

The high complexity of the real-world driving environment, which must be handled by an automated vehicle, is reflected in the broad-ranging design of automated driving systems. Figure 3.1 shows an exemplary architecture of an automated driving system [87, 101]. It roughly follows a sense-perceive-decide-act scheme and arranges involved subsystems. The perception of the driving environment is just one aspect of many and may involve multiple sensors to leverage individual strengths.

Latest automated vehicles integrate a multitude of different sensors to solve different perception tasks. Due to the high complexity of perceiving the driving environment and the demand for sufficient detection performance, subsystems are equipped with DNNs. A vast amount of research focuses on semantic segmentation [51, 55] and panoptic segmentation [78] for scene understanding or object detection.

General object detection describes a broad field in computer vision where bounding boxes are predicted for relevant objects. In this work, object detection describes a subsystem of an automated driving system. Although the DNN can be seen as the main component, the object detection subsystem considers also other important aspects such as the generation and labeling of training and validation datasets, pre-processing and post-processing steps of a DNN, loss formulation and training algorithm. Due to simplicity, it is only referred to as object detection despite including system aspects. Thus, object detection is not limited to the DNN.

DNNs for object detection predict either 2D or 3D bounding boxes. Well-known DNNs such as the SSD [110] or Faster R-CNN [93] predict 2D bounding boxes for typical objects for automated driving such as cars, trucks, traffic signs or pedestrians. Methods for object detection with 3D bounding boxes use and optionally fuse different sensor modalities, i.e., camera-only [127], LiDAR-only [73],

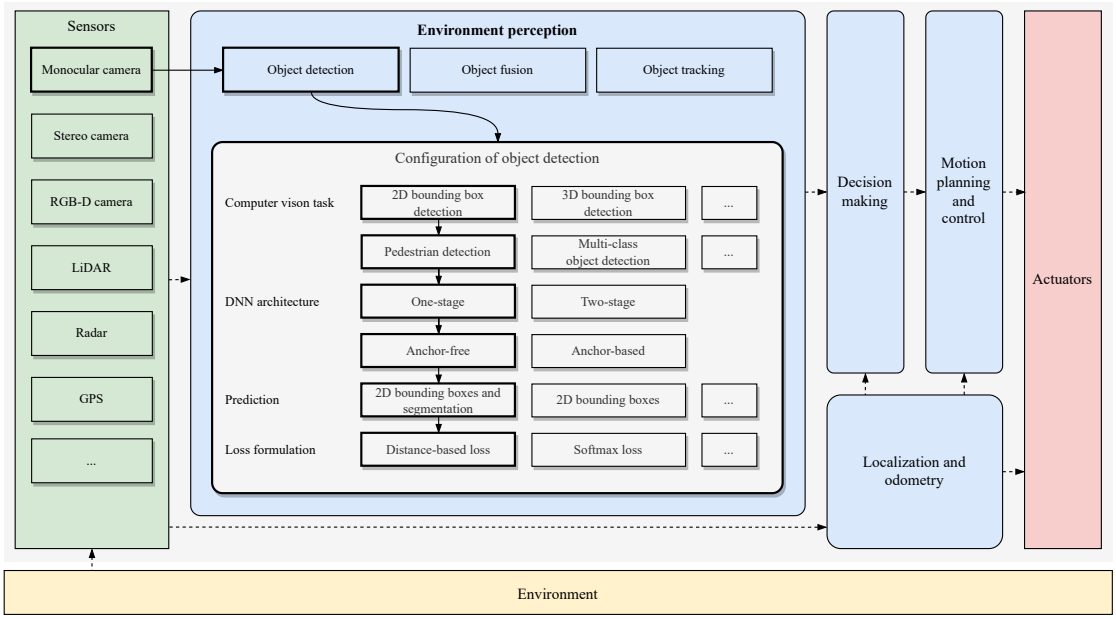


Figure 3.1: Exemplary system architecture of an automated driving system [87, 101]. Configuration of the object detection subsystem used for pedestrian detection. Object detection is part of the environment perception in automated driving using different sensors.

camera-LiDAR fusion [117] or camera-radar fusion [57]. Furthermore, methods for object tracking using image sequences have been introduced [64]. Although the list of mentioned research fields is not complete, this shows the diversity of methods applied in automated driving. Finally, different methods can also interact to enable redundancy, increase robustness or simply improve the performance of pedestrian detection.

However, state-of-the-art DNNs for pedestrian detection use single RGB images from monocular cameras as sensory input [54, 119, 61]. Since only two classes, background and pedestrian, are relevant for pedestrian detection, pedestrian detection can be seen as the binary case of general object detection. Although single RGB images can also be extracted from sequences, in this work no temporal dependencies between consecutive images or additional tracking algorithms are used for prediction.

Although fusion and tracking methods are probably mandatory for automated driving, the initial investigation of safety for pedestrian detection with DNNs has to begin with a simplistic and expandable reference implementation of a DNN for

pedestrian detection with state-of-the-art performance. Since this work proposes a methodology to support a safety argumentation based on verifiable evidence, it is reasonable to decrease complexity. Consequently, considering additional sensors and the resulting interdependencies of multiple subsystems is out of scope.

Instead, this work is focused on the use case of safe pedestrian detection with 2D bounding boxes using single RGB images from a monocular camera in the context of inherent interpretability. In this work, a generic DNN architecture for pedestrian detection is defined and gradually adapted. This is achieved by an in-depth configuration of the object detection subsystem, shown in Figure 3.1. In alignment with the state of the art for pedestrian detection and due to simplicity, one-stage and anchor-free DNNs are used. The DNN architecture and loss formulation are studied in more detail since they are directly related to inherent interpretability. The novel contribution is given by integrating an auxiliary body part segmentation, drawing the connection to inherent interpretability and extending the loss formulation accordingly.

3.2 Safety Argumentation

The following chapter introduces the overarching methodology of this work. First, the ML engineering for safety-critical systems is proposed. It outlines important parts of the safety argumentation such as the analysis of safety-critical failures and mitigation of functional insufficiencies. The main goal of this work is to leverage the inherent interpretability of DNNs for safe pedestrian detection.

In order to clarify the impact of the proposed interpretable methods and corresponding metrics providing evidence, a structured safety argumentation is eventually presented.

3.2.1 Failures of Automated Driving Systems

It is well-studied that DNNs for computer vision applied to perception tasks in automated driving make incomprehensible detection errors [77, 40, 106, 23]. Challenging conditions in the driving environment facilitated by natural perturbations such as sensor noise, brightness variations and blurring might lead to failures of automated driving systems. Referring to SOTIF [36] and Burton [6], a failure is described as "the termination of the intended behavior". A failure of an automated driving system can potentially be safety-critical. Note that failures occur on the level of the automated driving system, not the level of the object detection

subsystem. If a clearly visible pedestrian in the near driving path of an automated vehicle is not correctly detected, the automated driving system becomes a potential source of harm. Due to the inaccuracy and incorrect information, other subsystems cannot decide, plan and act appropriately.

System failures potentially result from detection errors of the object detection designed for pedestrian detection. Thereby the main error sources are the applied DNNs. Looking at state-of-the-art DNNs for pedestrian detection, a detection error is given when no 2D bounding box for the pedestrian has been predicted or detection bounding boxes cannot be matched to the ground truth target since the overlapping area is too small.

These errors are caused by functional insufficiencies that are inherent in the trained and embedded DNN but have remained unknown until now. Eventually, triggering conditions such as natural perturbations reveal functional insufficiencies. Due to limited training and validation data, the source domain of a DNN is also limited to a predefined environment. Hence, the open-world context leads to unknown conditions that may be experienced by an automated driving system. Distorted images may come from a shifted domain, thus, DNNs make detection errors. In conclusion, a detection error of the DNN-based object detection, which results from functional insufficiencies triggered by unknown conditions, can cause a safety-critical failure of an automated driving system.

Functional insufficiencies result from insufficient training configuration, data, DNN architecture or loss formulation. In the scope of this work are the following functional insufficiencies of DNNs: lack of interpretability, lack of generalization and lack of robustness. Here, the properties of interpretability are not absolute and may overlap with aspects of robustness as well as generalization capabilities of DNNs.

The lack of interpretability refers to the fact that state-of-the-art DNNs for pedestrian detection cannot present the reasons for a prediction in understandable terms to a human. Hence, the reasoning process of state-of-the-art DNNs is non-interpretable. The causes for the false classification or incorrect localization of pedestrians also remain unknown.

It can be argued that the decomposability [68] as one of the notions of interpretability is substantially reduced by DNNs for pedestrian detection. There are no parts of state-of-the-art DNNs that have an intuitive explanation. Using high-level semantic features such as center points [70] to detect pedestrians means that semantic concepts of humans representing low-level features must be implicitly used by the DNN. Thus, they cannot be understood or inferred by humans. It

remains unknown what kind of low-level features are extracted and how they are used for the reasoning of a DNN. Especially in the open-world context, it is imperative to evaluate the discrimination strength of the extracted features. Otherwise, natural perturbations may negatively affect the prediction performance and cause detection errors.

Furthermore, the plausibility of predictions is not given since the complex task of pedestrian detection is reduced to the prediction of high-level center points. In the presence of high occlusion, this oversimplification oftentimes leads to matching problems with ground truth bounding boxes. Although low-level features such as body parts may have been detected, the prediction is limited to 2D bounding boxes. Ultimately, the poor localization of detection bounding boxes without additional information on the visibility of body parts can cause a detection error.

3.2.2 Machine Learning Engineering for Safety-Critical Systems

Based on the related failure model, this work proposes a methodology to generate verifiable evidence by developing novel methods and proposing metrics to build a structured safety argumentation. Since methods and corresponding metrics are key aspects, the methodology describes ML engineering for safety-critical systems. Thereby, the contribution is twofold. The described failure model considering functional insufficiencies, errors and failures is visualized on the left-hand side in Figure 3.2. The downstream path highlights related work on a causal model for system failures by Burton [6], highly influenced by SOTIF [36].

First, a method for the systematical evaluation of safety-critical failures from a systems perspective for pedestrian detection is described. The evaluation of DNNs for pedestrian detection should be application-oriented and enable a fair and safety-focused comparison of the performance. The notion of performance is referring to detection metrics evaluating the classification and localization quality of 2D detection bounding boxes compared to ground truth bounding boxes.

Second, the mitigation of functional insufficiencies is a subject of investigation. Without a detailed inspection and insights into the reasoning of a DNN, the true causes for detection errors cannot be identified and consequently cannot be fixed. Since important components with individual responsibilities of a DNN are not identified, no guarantees for unexpected conditions can be given. Although the training is supervised, there is no evidence that discriminative features have been sufficiently learned and used for the prediction. That's why methods must

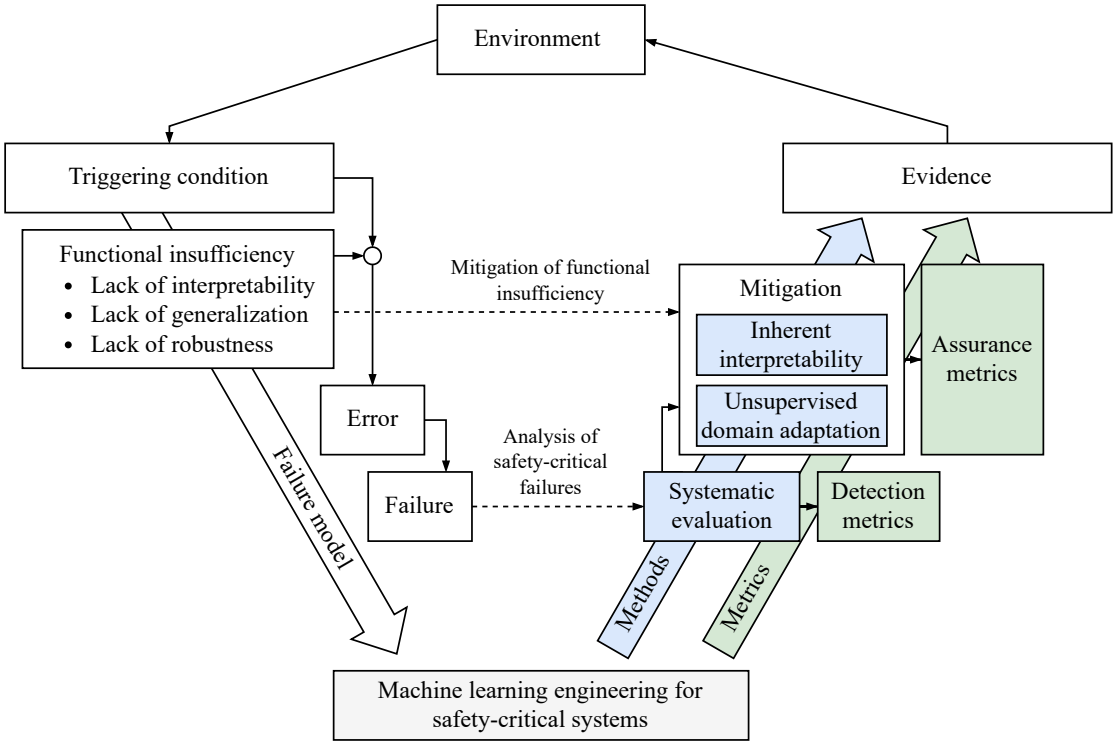


Figure 3.2: Illustration of the proposed machine learning engineering for safety-critical systems. The failure model is based on SOTIF [36] and Burton [6].

be introduced that make the reasoning of a DNN interpretable. Although an auxiliary body part segmentation makes the localization and classification of 2D bounding boxes more plausible, inherent interpretability goes further. Inherent interpretability addresses the structure of the latent space holding relevant features for pedestrian detection. By utilizing additional loss terms for clustering and separation of features, the two opposing mechanisms form a structured latent space for the reasoning process of a DNN [9, 27, 26].

The generalization capabilities of a DNN demonstrate the degree to which discriminative features or concepts for classification have been learned. The domain transfer of knowledge from synthetic to real-world data is an often investigated method to overcome a domain shift. Thereby the distinction between robustness and generalization may overlap and is consequently not absolute. Undisputed, a discriminative set of features should be eligible to detect or segment a *synthetic* hand as well as a *real* hand.

Besides the systematical evaluation and methods to mitigate functional insufficiencies of DNNs, a qualified set of of detection and assurance metrics to provide evidence for the correct functioning has to be defined. The prior goal is to build a structured safety argumentation targeting safety-critical failures and functional insufficiencies based on metrics to generate verifiable evidence. However, the evidence is limited to the predefined environment that may have been extended to other domains.

3.2.3 Inherent Interpretability

Referring to Figure 3.2, the systematical evaluation is followed by methods to mitigate functional insufficiencies of DNNs for pedestrian detection. These methods have to enhance the properties of a DNN such as the interpretability or generalization capabilities. The idea is to provide verifiable evidence based on corresponding assurance metrics. Finally, it is outlined how inherent interpretability for DNNs can be aligned with safety aspects of the DO-178C as an existing safety standard.

As already mentioned, the lack of interpretability is closely related to the lack of robustness and lack of generalization. Moreover, methods concerning the inherent interpretability of DNNs and approaches to improve generalization capabilities like unsupervised domain adaptation use similar ideas coming from the field of deep metric learning. Previous works have outlined severe issues with the application of softmax loss, motivating the development of more robust alternatives, i.e, A-Softmax [71] or I2CS [82]. Their findings are highly relevant for the inherent interpretability of DNNs.

Structured Latent Space Although a general definition of inherent interpretability currently does not exist, a widespread idea is to enforce intra-class concentration and inter-class separability based on distances in the latent space. Here, the latent space is seen as the embedding space of extracted features that are used for prediction. The features in the latent space are encoded by the feature extraction of the DNN.

The feature extraction $f(\mathbf{X}^*;\theta) : \mathfrak{X} \rightarrow \mathfrak{Z}$ with parameters θ describes the encoding of a receptive field $\mathbf{X}^* \in \mathfrak{X}$ with $\mathfrak{X} = \mathbb{R}^{H^* \times W^* \times 3}$ of an input image $\mathbf{X}$. Hence, a latent representation $z_k^{i,j} \in \mathcal{Z}$ in the latent space $\mathfrak{Z} = [0,1]^D$ is linked to an individual receptive field [1]. The encoding is illustrated in Figure 3.3. The receptive field $\mathbf{X}^*$ with width W^* and height H^* describes a cropped image section of $\mathbf{X}$. Throughout the DNN training, the feature extraction learns

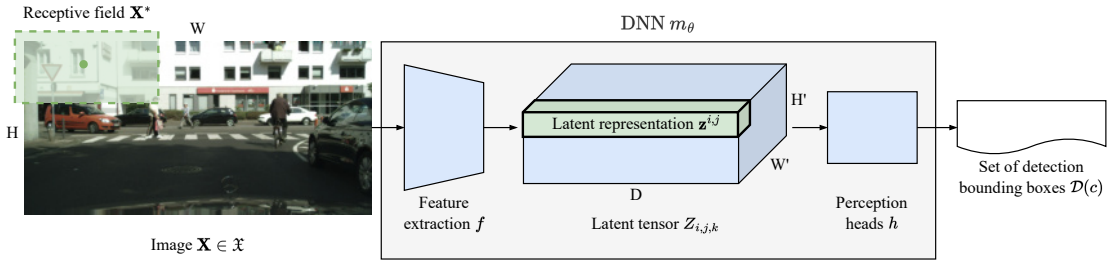


Figure 3.3: DNN architecture with feature extraction, latent tensor as embedding space and perception head to predict 2D bounding boxes for pedestrian detection.

to encode meaningful features. Every latent representation consists of D features in total. Since DNNs process multiple receptive fields simultaneously in a single forward pass, the extracted latent representations $z_k^{i,j} \in Z_{i,j,k}$ are concatenated in the latent tensor $Z_{i,j,k}$. The concatenated latent tensor $Z_{i,j,k}$ can be seen as a set of latent representations, thus, $z_k^{i,j}$ describes a vector with the length D (running over k) at the spatial position (i, j).

Which class is assigned to a latent representation depends on the classification of the center pixel from the receptive field. In the case of bounding box detection from Figure 3.3, latent representations are used to encode pedestrian centers. However, latent representation can be assigned to larger image regions for segmentation tasks. Inherent interpretability highlights the feature structure of latent representations in the latent space.

Most commonly, two opposing mechanisms that structure the latent space are described. On the one hand, *positive* latent representations of the same class are clustered. Here, using cluster centers as abstract class representatives is optional. On the other hand, *negative* representations should be separated. In this context, negative refers to all other classes. The positive or negative attribute must be determined in dependence on the class to be considered. It is important to note that the proposed interpretable methods in this work do not see background as a class. Therefore, the background is always considered negative.

A common understanding of intra-class concentration is that the center of latent representations is characteristic for a certain class and represents a mode in a structured latent space [9]. The center can be either represented by unsupervised prototypes, e.g., ProtoPNet [9], ProtoNCE [62] and PPD [27], or supervised semantic concept vectors as anchors for the training process, e.g., SupCon [56] and

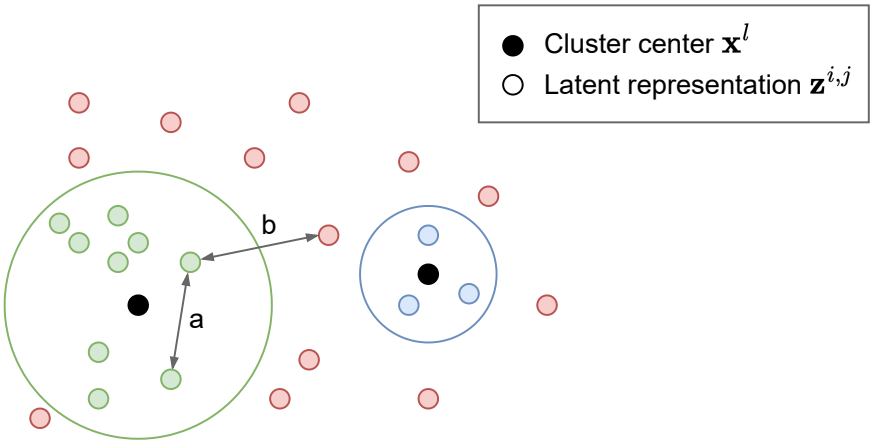


Figure 3.4: Structured latent space that clusters *positive* latent representations (first cluster in green and second cluster in blue) around cluster centers. *Negative* latent representations are separated. Parameter a stands for the intra-class distance and b for the inter-class distance.

CPD [26].

A structured latent space is visualized in Figure 3.4. Generally, latent representations $\mathbf{z}^{\mathrm{i,j}}$ (index notation: $\mathrm{z}_{\mathrm{k}}^{\mathrm{i,j}}$) hold class-typical features and shall be close to the respective l-th cluster center $\mathbf{x}^{\mathrm{l}}$ (index notation: $\mathrm{x}_{\mathrm{k}}^{\mathrm{l}}$). Consequently, new DNN architectures and loss formulations need to be defined.

Structuring the latent space is achieved by minimizing or maximizing distances between representations. According to [9, 41, 79] the squared L^2 distance between all latent representations $\mathrm{z}_{\mathrm{k}}^{\mathrm{i,j}} \in \mathrm{Z}_{\mathrm{i,j,k}}$ and the l-th vector $\mathrm{x}_{\mathrm{k}}^{\mathrm{l}}$ in a set of vectors $\mathcal{X}$ is defined as

$$\Delta_{\mathrm{i,j,l}} = \sum_{k=1}^{D} \left(\mathrm{Z}_{\mathrm{i,j,k}} - \mathrm{x}_{\mathrm{k}}^{\mathrm{l}}\right)^2 \quad \text{for } i = 1, \ldots, H' \text{ and } j = 1, \ldots, W' \text{ and } \mathrm{x}_{\mathrm{k}}^{\mathrm{l}} \in \mathcal{X} \tag{3.1}$$

Hence, the conclusive reasoning leading to a DNN's prediction is based on squared L^2 distances. A class is predicted since the extracted features in the latent representation are similar enough to a learned cluster center.

Deep Metric Learning Methods for inherent interpretability are closely related to deep metric learning. Although the motivation might differ, both research fields

heavily rely on clustering and separation mechanisms within the embedding space. Thereby, the common goal is to establish metric-based or distance-based DNN properties to overcome shortcomings of the widely adapted softmax loss [82].

According to [82], the term softmax loss refers to the combination of the inner product (between layer output and weights) and the cross entropy loss. Referring to Figure 3.4, applying the softmax loss allows higher intra-class than inter-class distances. This behavior seems unacceptable for interpretable methods. Moreover, the application of the softmax loss is associated with a decrease in robustness or generalization [71, 82].

As a result, various works propose loss adaptations to increase the margin to the decision boundary and increase the inter-class distance between clusters [103, 104]. Ultimately, this leads to more discriminative features. Intuitively, distance-based similarity measures are more human-understandable than the inner product, since intra-class distances do not exceed inter-class distances. Insights from deep metric learning are especially valuable for the definition of assurance metrics to generate verifiable evidence.

Unsupervised Domain Adaptation Methods for unsupervised domain adaptation, just like methods for inherent interpretability, benefit from the insights of deep metric learning. This seems reasonable since improvements concerning the interpretability and robustness can oftentimes just formally be separated from methods to improve the generalization capabilities.

Hence, various methods for unsupervised domain adaptation use prototypes to rectify noisy pseudo labels [122, 48]. Prototypes are cluster centroids calculated on a source dataset where ground truth is available, thus, representing the most discriminative features. Meanwhile, prototypes are used to provide a more trustworthy classification for unsupervised training and reduce the domain shift effectively.

3.2.4 Safety Impact of Inherent Interpretability

Although existing safety standards do not require interpretability as a DNN property, the lack of interpretability is highly relevant. Without understanding or actively facilitating the underlying nature of correct detections or detection errors, safety-critical failures cannot be eliminated or verifiable evidence cannot be generated.

Despite following the SOTIF protocol stating the mitigation of functional in-

sufficiencies, methods for the inherent interpretability of DNNs can also be aligned with key aspects of existing safety standards for software and hardware such as the DO-178C [31] following:

> Although interpretability is not directly 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 [27].

Primarily, a DNN can be seen as software since the source code implements the DNN architecture, pre-processing, post-processing, loss formulation and training algorithm. However, the feature extraction is represented by parameters that are trained by algorithms as source code. The DNN structure is hybrid, since it uses parameters besides conventional source code. Hence, the requirements-based decomposition of software according to DO-178C [31] cannot be directly applied.

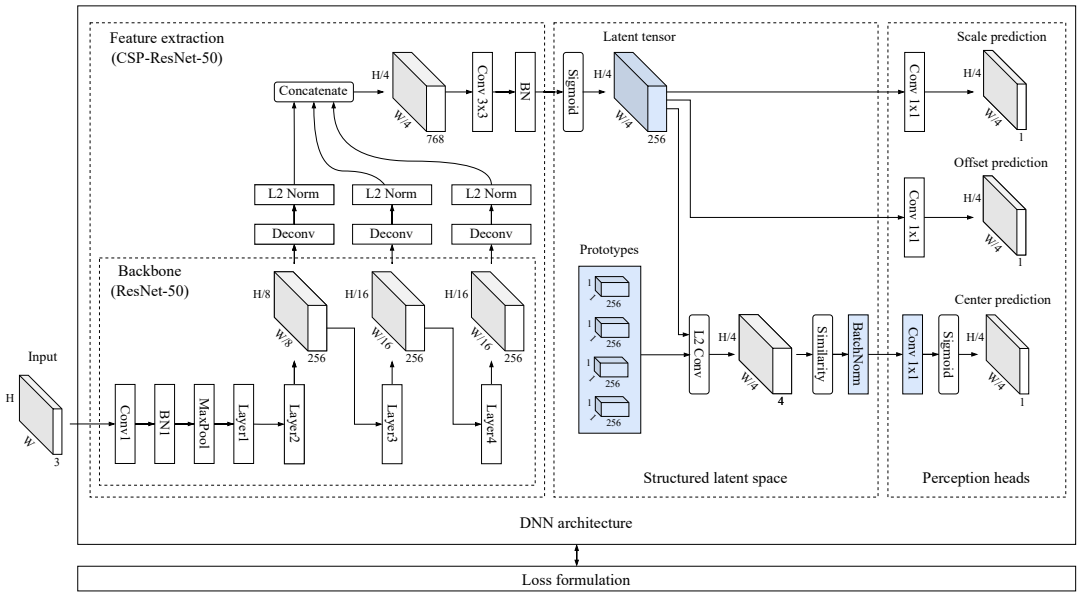
It can be argued that leveraging inherent interpretability for DNNs is aligned with key aspects of the DO-178C as an existing safety standard. Note that in order to enable inherent interpretability adaptations of the DNN architecture and especially the loss formulation are mandatory. Inherent interpretability refers to specific design choices for DNNs that enable a meaningful decomposition of a DNN.

Feifel et al. [27] proposed to rethink decomposition in the context of inherent interpretability. Accordingly, Figure 3.5 illustrates the decomposition of an interpretable DNN for pedestrian detection and the traceability matrix. Here, the DNN for prototype-based pedestrian detection (PPD) [27] is used as a proof of concept. A meaningful decomposition of DNNs isolates important key components, namely DNN artifacts (DAs). DAs have individual responsibilities in the interpretable reasoning of PPD. They are separated components of the DNN architecture and can further be processed by an iterative requirements analysis.

3.2.5 Structured Safety Argumentation

Although the methodology points out important safety aspects, a concrete safety argumentation has not been defined yet. For the purpose of this work, a safety argumentation builds a structured argument by providing verifiable evidence that safety goals have been met. Evidence demonstrates the successful application of methods to achieve safety goals with appropriate metrics.

In this work, safety goals focus on mitigating functional insufficiencies of DNNs for pedestrian detection and the application-oriented systematic evaluation. Consequently, inherent interpretability is identified as the key enabler for generating



Decomposition

Traceability

	Low-level requirement	Source code		DNN artifact (DA)		
		DNN architecture	Loss formulation	Prototypes	BatchNorm	Conv 1x1
I	Latent space shall be explicitly structured	✓	✓	✓		
II	Latent representation shall learn discriminative features of pedestrian center	✓	✓	✓		
III	Confidence scores shall be derived from interpretable reasoning	✓		✓	✓	✓
IV	Ensemble classification with multiple prototypes	✓		✓	✓	

Figure 3.5: Meaningful decomposition of DNN architecture through inherent interpretability. Leveraging inherent interpretability enables the decomposition and the creation of a traceability matrix. The identified low-level requirements can be linked to implemented source code. Particularly, DNN artifacts must be implemented explicitly and traced to the source code. Decomposition is illustrated for the DNN for prototype-based pedestrian detection (PPD) [27] as a proof of concept.

evidence. Functional insufficiencies are mitigated by applying methods to enhance the DNN architecture and loss formulation.

The goal structuring notation [53] provides a nomenclature to build a structured safety argumentation. The overarching safety goal states that pedestrian detection with DNNs shall be acceptably safe to operate. Within the scope of this work,

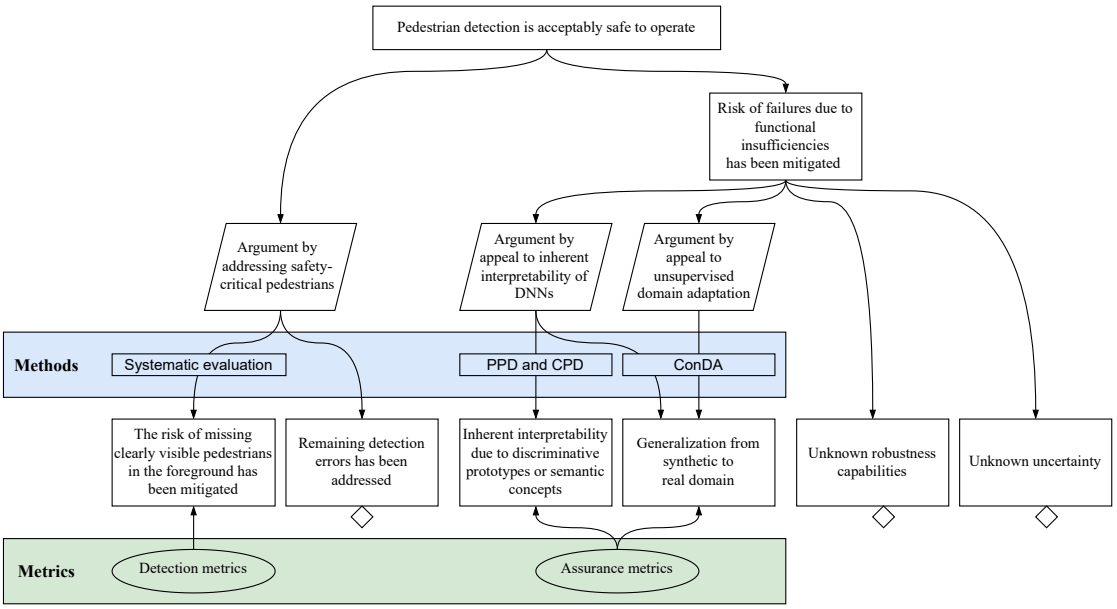


Figure 3.6: Initial structured safety argumentation for pedestrian detection with deep neural networks. The safety argumentation uses the goal structuring notation [53] and is limited to the lack of interpretability and lack of generalization as functional insufficiencies. Safety goals (rectangles) are argued by arguments (rhombuses) using methods and metrics (ellipses) to provide verifiable evidence.

Figure 3.6 shows a structured safety argumentation by using the goal structuring notation.

The safety argumentation, illustrated in Figure 3.6, draws the conclusion between safety goals, functional insufficiencies, methods and metrics more precisely than the proposed ML engineering, shown in Figure 3.2. Here, the argumentation strategy is two-folded.

On the one hand, safety-critical and disruptive detection errors are addressed in order to provide an application-oriented and system-related evaluation of the detection performance [29]. This is achieved by performing a systematic evaluation and presenting experimental results of detection metrics. Due to the systematic evaluation, the risk of minimizing clearly visible pedestrians shall be reduced by selecting the best-performing checkpoint during DNN training that corresponds to a pre-defined safety goal.

On the other hand, functional insufficiencies concerning the lack of interpretabil-

ity and lack of generalization are addressed. Therefore, inherently interpretable DNNs for prototype-based pedestrian detection (PPD) [27] and concept-based pedestrian detection (CPD) [26] are proposed. This line of work has been extended to improve the generalization capabilities utilizing concept-based domain adaptation (ConDA) [28] for pedestrian detection. Eventually, assurance metrics are defined and mapped to a priori formulated safety goals.

Figure 3.6 provides an initial safety argumentation, aligned with the scope of this work. Thus, remaining safety issues, i.e., other detection errors, robustness capabilities and uncertainty of DNNs for pedestrian detection, are only identified.

3.3 Systematic Evaluation

Starting with safety-relevant failures from the perspective of an automated driving system, potential detection errors of DNNs for pedestrian detection can be analyzed. Two types of detection errors can occur, namely false negatives and false positives. A false negative is a ground truth bounding box, representing a pedestrian that has not been correctly detected. The detection is missing, since no detection bounding box could be assigned to the ground truth bounding box. In contrast, a false positive creates the erroneous impression of an imaginary pedestrian. Both detection errors cause flawed pedestrian detection in the driving environment, potentially leading to unsafe control actions of automated vehicles.

Since related work misses a clear reference to different error categories concerning automated driving, a systematic evaluation of false positives and false negatives is proposed. Parts of this section are strongly based on the previously introduced work by Feifel et al. [29] and the preceding master thesis by Benedikt Franke [38].

The initial work does not consider ground truth distances between pedestrians and the automated vehicle. Instead, the depth or distance is estimated based on the height of pedestrians in pixels. Tobias Zielke has extended the systematic evaluation using depth information for pedestrians as part of his master's thesis [126]. Finally, the mathematical notation is adapted to this work and the experiments are extended to the synthetic dataset SynPeDS [97].

Although the systematic evaluation of DNNs for pedestrian detections has been introduced as a method to provide evidence for a safety argumentation, its central position in the methodology requires a deeper insight at this point of the work. The proposed methods for mitigating functional insufficiencies regarding the inherent interpretability and generalization capabilities of DNNs must be evaluated during training. That's why the systematic evaluation is a key element for checkpoint

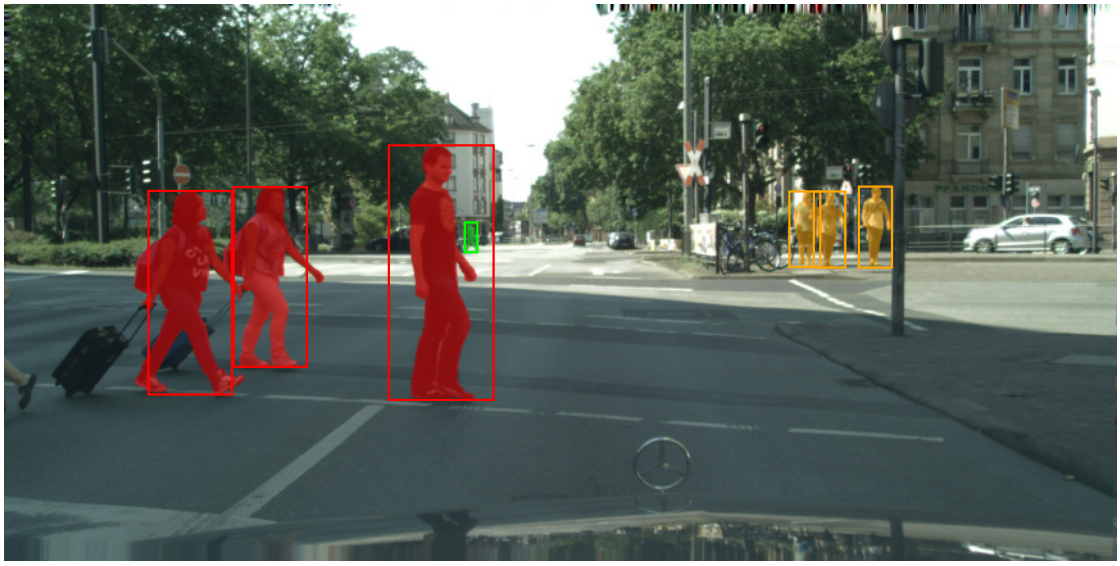


Figure 3.7: Systematic evaluation of safety-critical detection errors of DNN for pedestrian detection. The categorization of ground truth bounding boxes is shown. The proposed systematic evaluation distinguishes between clearly visible and safety-critical pedestrians in the foreground (red) and background (orange). The remaining pedestrians are in the far background (green), which is subordinate to the evaluation with detection metrics [29].

selection during training. It is essential as an evaluation tool that has to be described beforehand.

In an urban driving scenario, it can be argued that pedestrians crossing the street in the near field of an automated vehicle are more relevant for safety than a seated group of people sitting on a park bench further away. False negative close to automated vehicles potentially causes harmful behavior, since pedestrians have not been correctly detected. The impact and potential harm of safety-critical false negatives are intuitive.

Figure 3.7 gives an example of the application-oriented categorization of ground truth bounding boxes for pedestrians. Pedestrians in the foreground (red) or background (orange) should be higher prioritized over pedestrians in the far background (green). The consideration is based on different braking distances at different vehicle speeds and is therefore system-related. Applying different risk levels follows the existing subset-based evaluation methods for DNNs [20].

However, unintended control actions caused by false positives also threaten

other traffic participants, i.e., needless and sudden braking. Post-processing with non-max suppression uses pre-defined hyperparameters to establish a trade-off between the number of false positives and false negatives. The reduction of false negatives will ultimately result in a multitude of detection bounding boxes that can be matched to a single ground truth bounding box - a single pedestrian may become a widespread group for an automated vehicle.

However, a pedestrian is still detected in this area, regardless of whether there are other instances. The automated vehicle is still aware of the situation. This incident can also be resolved with additional sensor input from the automated driving system. In extreme cases, especially with low confidence thresholds, false positives increasingly become ghost detections. Ghost detections are considered disruptive false positives due to their random occurrence without clear reference to existing pedestrians.

In the following, ground truth and detection bounding boxes are systematically evaluated. From a system perspective, it can be argued that errors with different severity levels can be identified. Consequently, the respective bounding boxes are categorized based on quantified assumptions.

3.3.1 Pedestrian Detection Errors

First, the mathematical notation of bounding boxes and conclusively the definition of ground truth and detection bounding boxes have to be introduced. Generally, 2D bounding boxes can be described by four corner coordinates: (x_1, y_1), (x_1, y_2), (x_2, y_1),(x_2, y_2). A bounding box r is defined as a set of all pixels (x, y) enclosed by the corner coordinates:

$$r = \{(x, y) \mid x_1 \leq x < x_2 \wedge y_1 \leq y < y_2\} \tag{3.2}$$

Thus, the width and height of a bounding box are defined as $w(r) = x_2 - x_1$ and $h(r) = y_2 - y_1$. A bounding box describes either a ground truth bounding box $g \in \mathcal{G}$ or a detection bounding box $d \in \mathcal{D}(\mathrm{c})$. The set of detection bounding boxes $\mathcal{D}(\mathrm{c})$ depends on the confidence threshold c. The confidence threshold marks the point at which center predictions are considered real detections. It is important to note, that the confidence threshold is a hyperparameter that must be defined cautiously. It has an immediate effect on the balance between precision and recall.

The set of ground truth bounding boxes, identified as true positives, is defined

as:

$$\begin{aligned}\mathcal{TP}^{\mathrm{g}}(\mathrm{c}) = \{ & g \mid g \in \mathcal{G} \ \wedge \exists\, d \in \mathcal{D}(\mathrm{c}) : \\ & \big[\, IoU(g,d) > 0.5 \wedge \nexists \tilde{g} \in \mathcal{G} : \\ & IoU(\tilde{g},d) > IoU(g,d)\big]\}\end{aligned} \tag{3.3}$$

Following the evaluation of state-of-the-art DNNs for pedestrian detection, a ground truth or detection bounding box is only considered a true positive if a minimum intersection over union (IoU) of 0.5 is achieved.

A false negative always describes a ground truth bounding box g that cannot be matched to any predicted detection bounding box and is given by

$$\mathcal{FN}(\mathrm{c}) = \mathcal{G} \setminus \mathcal{TP}^{\mathrm{g}}(\mathrm{c}) \tag{3.4}$$

Furthermore, errors for detection bounding boxes have to be introduced. A detection bounding box $d \in \mathcal{D}(\mathrm{c})$ that cannot be matched to any $g \in \mathcal{G}$ or can only be matched to an already matched g is assigned to the set of false positives:

$$\begin{aligned}\mathcal{FP}(\mathrm{c}) = \{ & d \mid d \in \mathcal{D}(\mathrm{c}) \ \wedge \nexists\, g \in \mathcal{G} : \\ & IoU(g,d) > 0.5 \ \vee \exists\, g \in \mathcal{G} : \\ & \big[\, IoU(g,d) > 0.5 \ \wedge \ \exists \tilde{d} \in \mathcal{D}(\mathrm{c}) : \\ & IoU(g,\tilde{d}) > IoU(g,d)\big]\}\end{aligned} \tag{3.5}$$

Based on the identification of $\mathcal{TP}^{\mathrm{g}}(\mathrm{c}), \mathcal{FN}(\mathrm{c})$ and $\mathcal{FP}(\mathrm{c})$, safety-critical errors, which potentially cause system failures, are defined. However, ground truth bounding boxes must be categorized first. The application of straightforward set operations isolates safety-critical subsets.

3.3.2 Ground Truth Bounding Boxes

Pedestrian detection is subject to high complexity due to large variations in the height, poses, textures, lighting and visibility. Consequently, it is hard to determine a sufficient level of detail in identifying relevant features for a complete categorization. Rather than listing all possible feature combinations, a manageable set of accessible and relevant features must be identified. Meanwhile, ontologies are a promising tool when applied to dataset engineering in automotive applications equipped with artificial intelligence [47].

In this work, a straightforward approach from the system perspective is taken.

The distance of a pedestrian to the automated vehicle and the visibility are identified as two of the most important features for an application-oriented evaluation of DNNs for pedestrian detection. From a system perspective, two relevant distance ranges for urban driving scenarios have been identified. Thus, three distance regions are defined: (1) foreground, (2) background and (3) far background. Moreover, four categories concerning visibility are proposed: (1) clearly visible/ no occlusion, (2) environmental occlusion, (3) crowd occlusion and (4) ambiguous occlusion.

The definition of different categories for the distance to the automated vehicle and the visibility of a pedestrian enables the identification of clearly visible pedestrians in the foreground or background. This work argues that the highest possible detection performance for these categories is an appropriate metric for the training of DNNs for pedestrian detection. Foremost, it has to be guaranteed that clearly visible pedestrians in the foreground and background are safe. The chosen distance range or depth depends on the investigated driving environment.

In the following, the ground truth categories for three distance regions are described. Then, different forms of occlusion are introduced. Finally, visible ground truth bounding boxes, located either in the foreground or background, are identified with additional information on the pixel-accurate distance to the automated vehicle.

In the sense of this work, missed pedestrians, which are clearly visible in the foreground or background, are regarded as safety-critical. Hence, these pedestrians should be the basis for developing application-related detection metrics. The application-oriented evaluation with metrics provides verifiable evidence for the correct functioning of pedestrian detection.

Foreground The foreground defines the area where an automated vehicle traveling at 30km/h can no longer brake to avoid a collision with a pedestrian. Although higher driving speeds are also considered, 30 km/h seems reasonable as a starting point for the safety-critical error analysis of a DNN.

Table 3.1 defines the necessary parameters to calculate the braking distance of an automated vehicle starting from 30 km/h. Note that a simplified model to calculate the braking distances is used. Nevertheless, it is deemed to be sufficient for the scope of this work.

Based on the predefined parameters, the braking distance of an automated

Table 3.1: Parameters for the simplified braking distance calculation of an automated emergency braking (AEB) d_{AEB} starting from 30 km/h.

Parameter	Value	Description
t_{proc}	0.4 s	Processing time
μ	0.3	Friction coefficient
v	8.33 m/s	Velocity
G	9.81 m s^2	Gravitational acceleration on earth
d_s	2 m	Added distance
d_v	4 m	Distance from rear axis to front

emergency braking (AEB) d_{AEB} is defined as

$$d_{\text{AEB}} = d_s + d_v + \left\lceil \frac{v^2}{2 \cdot \mu \cdot g} \right\rceil + \lceil v \cdot t_{proc} \rceil = 24\,\text{m} \tag{3.6}$$

Considering the braking distance of 24 m, the set of ground truth bounding boxes for pedestrians standing in the foreground is defined as

$$\mathcal{G}_{\text{f}} = \{g \mid g \in \mathcal{G} \ \wedge d(g) \leq 24\,\text{m}\} \tag{3.7}$$

where $d(g)$ is the depth of the center point of the ground truth bounding box g.

Background Subsequently, the driving speed is raised to 50 km/h. Applying Equation 3.6, the braking distance of an automated emergency braking starting with 50 km/h is 48 m. Excluding foreground pedestrians, the set of ground truth bounding boxes in the background is given by

$$\mathcal{G}_{\text{b}} = \{g \mid g \in \mathcal{G} \ \wedge \ (24\,\text{m} < d(g) \leq 48\,\text{m})\} \tag{3.8}$$

Environmental Occlusion In contrast to categorizing ground truth bounding boxes based on the distance of a pedestrian to the automated vehicle, the visibility of pedestrians is another important aspect. The identification of clearly visible or occluded pedestrians is achieved by using the instance segmentation $\text{P}_{\text{i,j}}$ and semantic segmentation $\text{S}_{\text{i,j}}$. Meanwhile, depth masks $\text{D}_{\text{i,j}}$ provide information on the relative distances of objects. The set of classes that potentially occlude pedestrians $\mathcal{O}^{\text{e}}$ explicitly exclude classes such as street and sky.

The Iverson bracket $[\cdot]$ is used to define masks that exhibit different forms of

occlusion. The mask for environmental occlusion considering classes such as trees and cars is defined as

$$\mathrm{E_{i,j}} = [\mathrm{D_{i,j}} \otimes (\mathrm{S_{i,j}} \in \mathcal{O}^{\mathrm{e}}) \leq d(g)] \tag{3.9}$$

Based on that, the visibility with respect to the environment $\mathrm{E_{i,j}}$ is calculated:

$$\nu_e(g) = \frac{\sum_{(i,j)\in g} \mathrm{E_{i,j}}}{\sum_{(i,j)\in g} \mathrm{E_{i,j}} + \sum_{(i,j)\in g} \mathrm{P_{i,j}}} \tag{3.10}$$

Finally, the set of ground truth bounding boxes that are occluded by the environment is given by

$$\mathcal{G}_{\mathrm{e}} = \{g \mid g \in \mathcal{G} \ \wedge \nu_e(g) > 0.5\} \tag{3.11}$$

The threshold to identify environmental occlusion is empirically set to 0.5.

Crowd Occlusion Similar to environmental occlusion, classes that cause crowd occlusion $\mathcal{O}^{\mathrm{c}}$ are identified. Besides the pedestrian class itself, $\mathcal{O}^{\mathrm{c}}$ also considers riders and sitting persons. The resulting mask for crowd occlusion is defined as

$$\mathrm{C_{i,j}} = [\mathrm{D_{i,j}} \otimes (\mathrm{S_{i,j}} \in \mathcal{O}^{\mathrm{c}}) \leq d(g)] \tag{3.12}$$

Based on $\mathrm{C_{i,j}}$, the visibility of a pedestrian instance in a crowd is calculated:

$$\nu_c(g) = \frac{\sum_{(i,j)\in g} \mathrm{C_{i,j}}}{\sum_{(i,j)\in g} \mathrm{C_{i,j}} + \sum_{(i,j)\in g} \mathrm{P_{i,j}}} \tag{3.13}$$

Finally, the set of ground truth bounding boxes that are subject to crowd occlusion is given by

$$\mathcal{G}_{\mathrm{c}} = \{g \mid g \in \mathcal{G} \ \wedge \nu_c(g) > 0.5\} \tag{3.14}$$

The threshold to identify crowd occlusion is empirically set to 0.5.

Ambiguous Occlusion Since environmental and crowd occlusion might happen simultaneously, the two sets are not disjoint. Consequently, a third occlusion category denoted as ambiguous occlusion is proposed. Ambiguously occluded bounding boxes $\mathcal{G}_{\mathrm{a}}$ are occluded by the environment as well as other humans. For more information, please refer to [29].

Clearly Visible Pedestrians in the Foreground and Background Due to the distinction of different distances and forms of occlusion, clearly visible pedestrians in the foreground can easily be identified:

$$\mathcal{G}_{\mathrm{f,v}} = \mathcal{G}_{\mathrm{f}} \setminus (\mathcal{G}_{\mathrm{e}} \cup \mathcal{G}_{\mathrm{c}}) \tag{3.15}$$

Similarly, clearly visible pedestrians in the background are defined as

$$\mathcal{G}_{\mathrm{b,v}} = \mathcal{G}_{\mathrm{b}} \setminus (\mathcal{G}_{\mathrm{e}} \cup \mathcal{G}_{\mathrm{c}}) \tag{3.16}$$

In this work, the two sets of clearly visible pedestrians in the foreground and background ($\mathcal{G}_{\mathrm{f,v}}$ and $\mathcal{G}_{\mathrm{b,v}}$) are considered most relevant for the application-oriented evaluation. Merely appropriate detection metrics are able to support a safety argumentation of DNNs for pedestrian detection.

The following Figure 3.8 gives an overview of the categorization of ground truth bounding boxes.

Figure 3.8: Categorization of ground truth bounding boxes based on distance and visibility. Ground truth and ignored bounding boxes are shown on the left. On the right, the proposed categorization of ground truth bounding boxes is shown. Three distance ranges are investigated: foreground (red), background (orange) and far background (green). Four visibility categories are identified: clearly visible (solid), environmental occlusion (dashed), crowd occlusion (dotted) and ambiguous occlusion (dash-dotted).

3.3.3 Detection Bounding Boxes

In total, three categories for detection errors are introduced: (1) scale errors, (2) detection errors and (3) ghost detection. First, scale and localization errors must be defined. Finally, the categorization of detection bounding boxes is conducted to identify ghost detections. Ghost detections mark the most disruptive false positives in automated driving.

Scale Errors Detection bounding boxes that cannot be matched to ground truth bounding boxes due to a large center point offset are assigned to scale errors:

$$\mathcal{D}_{\mathrm{s}}(\mathrm{c}) = \{d \mid d \in \mathcal{FP}(\mathrm{c}) \ \wedge \exists\, g \in \mathcal{G} : \mathrm{A}(g, d)\} \tag{3.17}$$

The predicate $\mathrm{A}(g, d)$ states whether the center of d is aligned with g, is defined as:

$$\mathrm{A}(g, d) \iff |c_x^g - c_x^d| \leq \mu_s w(g) \wedge |c_y^g - c_y^d| \leq \mu_s h(g) \tag{3.18}$$

For the experiments, the threshold μ_s is empirically set to 0.2.

Localization Errors Contrarily to scale errors, the set of localization errors holds all false positives that fall in close proximity to a g, but the detection bounding box cannot be matched and is not a scale error. Localization errors are defined as

$$\mathcal{D}_{\mathrm{l}}(\mathrm{c}) = \{d \mid d \in (\mathcal{FP}(\mathrm{c}) \setminus \mathcal{D}_{\mathrm{s}}(\mathrm{c})) \ \wedge \ \exists\, g \in \mathcal{G} : IoU(g, d) \geq \mu_l\} \tag{3.19}$$

For the experiments, the threshold μ_l is empirically set to 0.25.

Ghost Detections Based on the definition of $\mathcal{D}_{\mathrm{s}}(\mathrm{c})$ and $\mathcal{D}_{\mathrm{l}}(\mathrm{c})$, ghost detections are defined as

$$\mathcal{D}_{\mathrm{g}}(\mathrm{c}) = \mathcal{FP}(\mathrm{c}) \setminus (\mathcal{D}_{\mathrm{s}}(\mathrm{c}) \ \cup \ \mathcal{D}_{\mathrm{l}}(\mathrm{c})) \tag{3.20}$$

Detections in this category are seen as random and unrelated to the presence of pedestrians. Consequently, these kinds of detection errors are strongly disruptive and have the potential to severely impact the flawless operation of automated driving systems.

The following Figure 3.9 gives an overview of the categorization of detection bounding boxes.

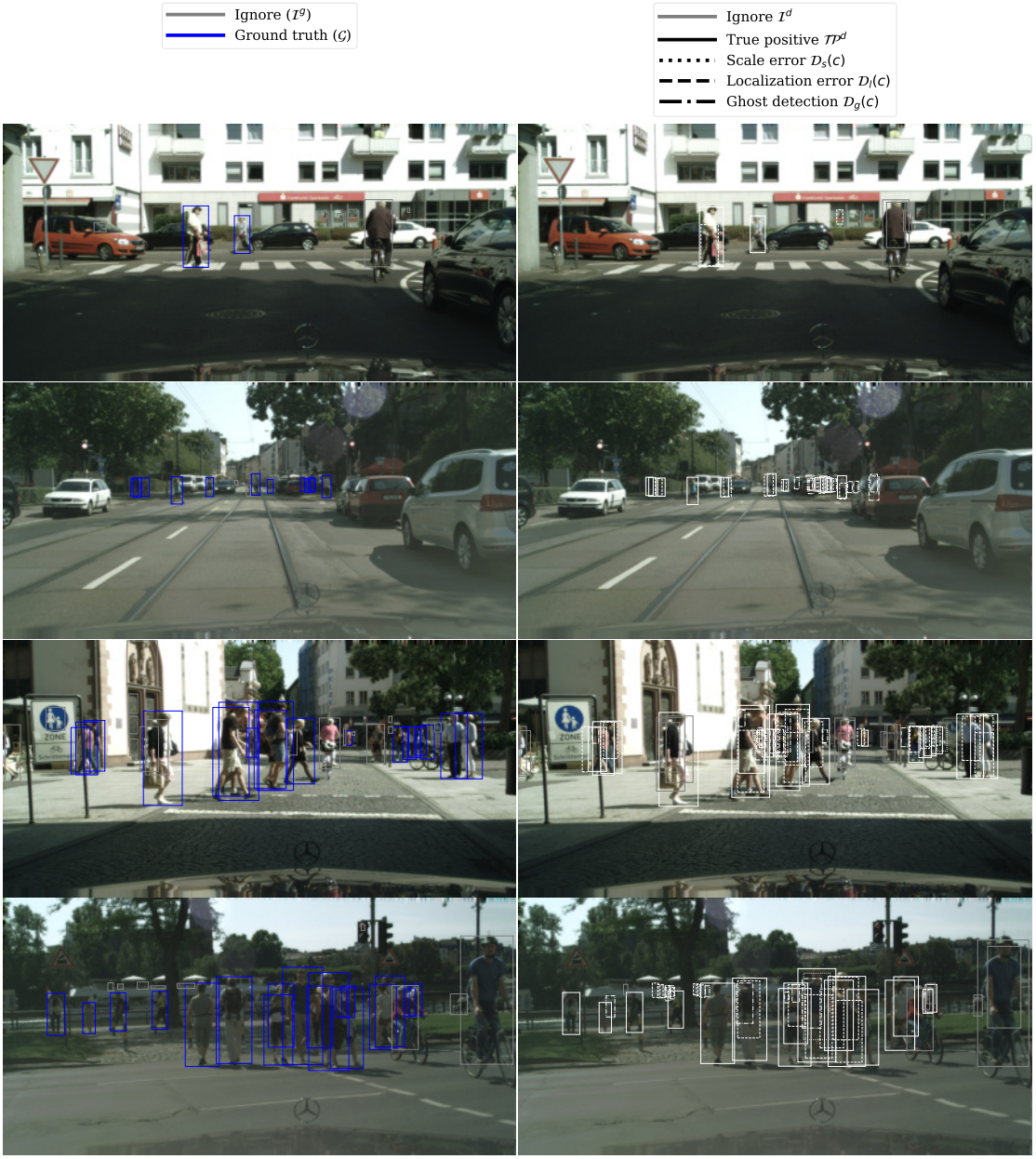


Figure 3.9: Categorization of detection bounding boxes with an empirically defined confidence threshold ($c = 0.1$). Ground truth and ignored bounding boxes are shown on the left. On the right, the proposed categorization of detection truth bounding boxes is shown. The following categories for detection bounding boxes (white) are identified: true positives (solid), sale errors (dotted), localization error (dashed) and ghost detections (dash-dotted). Ignored detection bounding boxes are in grey with a solid line.

3.4 Evidence

For safety-critical applications, the notion of performance can not only refer to the accuracy or quality of detections. Instead, performance must consider metrics with respect to detection quality and additional assurance metrics that can be used in the safety argumentation. In the following, detection metrics and assurance metrics are described. Consequently, the metrics enable valid benchmarking of state-of-the-art methods and proposed methods in this work.

3.4.1 Detection Metrics

The goal of pedestrian detection is to predict 2D detection bounding boxes for pedestrians in an input image. That's why the comparison of detection bounding boxes and ground truth bounding boxes is a key aspect of providing verifiable evidence. Applying the proposed systematic evaluation, detection metrics consider different categories of ground truth bounding boxes and errors of detection bounding boxes. The definition of the detection metrics follows the formulation of the log-average miss rate, which is the most commonly used performance metric for DNNs for pedestrian detection [21, 5, 20, 107].

Filtered Log-Average Miss Rate First, the filtered log-average miss rate $FLAMR(c)$ is proposed to measure the performance concerning the safety-critical error categories. The filtered miss rate $MR_P(c)$ accounts for different categories of ground truth bounding boxes P. Different cases for P might be considered: (1) $P = \mathcal{G}_{\mathrm{f,v}}$ and (2) $P = \mathcal{G}_{\mathrm{b,v}}$, where $\mathcal{G}_{\mathrm{f,v}}$ describes clearly visible pedestrians in the foreground and $\mathcal{G}_{\mathrm{b,v}}$ in the background. The filtered miss rate depends on the ground truth category P and is defined as

$$MR_P(c) = \frac{|\,\mathcal{FN}(\mathrm{c}) \cap P|}{|\,\mathcal{TP}^{\mathrm{g}}(\mathrm{c}) \cap P| + |\,\mathcal{FN}(\mathrm{c}) \cap P|} \tag{3.21}$$

With reference to the $LAMR(c)$, the filtered log-average miss rate $FLAMR(c)$ is defined as

$$FLAMR(c) = \exp\left(\frac{1}{|\,\mathcal{T}\,|}\sum_{c\in\mathcal{T}} \log MR_P(c)\right) \tag{3.22}$$

Here, $\mathcal{T}$ is a set of confidence levels that corresponds to the nine pre-defined $FPPI(c)$ values for calculating the $LAMR(c)$:

$$\mathcal{T} = \left\{ \underset{FPPI \leq f}{\arg\max}\ FPPI(c) \mid f \in \mathcal{F} \right\} \tag{3.23}$$

with $\mathcal{F} = \{10^{-2}, 10^{-1.78}, \ldots, 10^{0}\}$ and $|\mathcal{F}| = 9$. The confidence threshold c has to be fixed to analyze experimental results and is chosen adequately low with $c = 0.01$. $FLAMR(c)$, $FPPI(c)$ and $MR(c)$ are simplified to $FLAMR$, $FPPI$ and MR. The simplification is used for all metrics that are introduced in the following.

Due to the high dependency on hyperparameters, e.g., the confidence threshold c or the IoU threshold for the non-max suppression, detecting individual pedestrians in crowds is difficult. Since the proposed categorization of ground truth bounding boxes does not ignore areas to distinguish other relevant categories, a detection bounding box can be a better match for a crowd-occluded pedestrian instead of the clearly visible pedestrian standing in front of the group. This work argues that it is foremost important to detect clearly visible pedestrians in the foreground. Here, pedestrians that are occluded by detected pedestrians are automatically considered safe, since the trajectory planning of the automated vehicle already considers the detected pedestrian in the foreground.

That is the reason why a double-matching strategy is proposed. There are cases where a detection bounding box can be matched either to a crowd-occluded or clearly visible ground truth bounding box. The strategy enforces that the detection bounding box is always matched to the clearly visible ground truth bounding box. This allows for a more application-oriented evaluation by reducing the complexity of difficult crowd scenarios.

Filtered Log-Average Miss Rate for Ghost Detections It has been argued that not all false positives are equally disruptive. To take this effect into account, the filtered log-average miss rate can be calculated for to ghost detections. Ghost detections per image are defined as

$$GDPI(c) = \frac{1}{N_I} |\, \mathcal{D}_{\mathrm{g}}(\mathrm{c}) \,| \tag{3.24}$$

where N_I is the number of images in the dataset being evaluated.

The set of confidence levels for ghost detections $\mathcal{T}^H$ can be adapted accordingly.

Finally, the filtered log-average miss rate for ghost detections is defined as

$$FLAMR_g = \exp\left(\frac{1}{|\mathcal{T}|}\sum_{c\in\mathcal{T}^H}\log MR_P(c)\right) \tag{3.25}$$

3.4.2 Assurance Metrics

Despite the application-oriented systematic evaluation of DNNs for pedestrian detection, a conclusive safety argumentation has to consider and quantify inherent DNN characteristics that are responsible for the existence of functional insufficiencies, namely assurance metrics. Thus, assurance metrics do not focus on conventional performance aspects, unlike the state-of-the-art benchmarks with the log-average miss rate for pedestrian detection.

Experimental results for assurance metrics have to clearly state that functional insufficiencies have been mitigated. This is achieved by analyzing the segmentation quality and especially the cluster quality of the structured latent space facilitated by inherent interpretability. Proactively analyzing and quantifying the inherent interpretability of DNNs substantially supports a safety argumentation.

Segmentation Quality To evaluate the segmentation quality of body parts, the proposed systematic evaluation is extended to the segmentation task. Hence, the error categories of ground truth bounding boxes are applied to the instance segmentation and further to the body part segmentation. Specifically, $\Omega_{\mathrm{i,j,k}}$ is defined as the one-hot encoded ground truth for the body part segmentation of a pedestrian instance that is assigned to a ground truth bounding box g.

The filtered mean intersection over union $FmIoU$ for the ground truth error category $P \in \{\mathcal{G}_{\mathrm{f,v}}, \mathcal{G}_{\mathrm{b,v}}\}$) is defined as

$$FmIoU = \frac{1}{N_C}\sum_{l=1}^{N_C}\frac{\sum_{g\in P}|\,\Omega_{\mathrm{i,j,k}}\cap\Omega^{'}_{\mathrm{i,j,k}}\,|}{\sum_{g\in P}|\,\Omega_{\mathrm{i,j,k}}\cup\Omega^{'}_{\mathrm{i,j,k}}\,|} \tag{3.26}$$

where $\Omega^{'}_{\mathrm{i,j,k}}$ describes the one-hot encoded predictions of the teacher DNN $\breve{m}_{\theta'}$. Assuming the ground truth bounding box g is assigned to the one-hot encoded ground truth body part segmentation $\Omega_{\mathrm{i,j,k}}$ for the l-th concept, the $FmIoU$ estimates the segmentation quality. Moreover, the predictions are adjusted to only show instances of the respective error category P. Otherwise, false positives are wrongly introduced and distort evaluating the segmentation quality.

The proposed $FmIoU$ is similar to the method of Network Dissection [2] and Net2Vec [30]. However, $FmIoU$ does not follow the holistic approach of evaluating the concept embedding qualities of multiple intermediate DNN layers. Instead, the segmentation quality focuses on the prediction quality of the body part segmentation.

Cluster Quality The segmentation quality purely compares the predicted body part segmentation and ground truth. Here, the nature of the underlying latent space is only considered implicitly. Therefore, the validity of experimental results in terms of supporting a safety argumentation is limited.

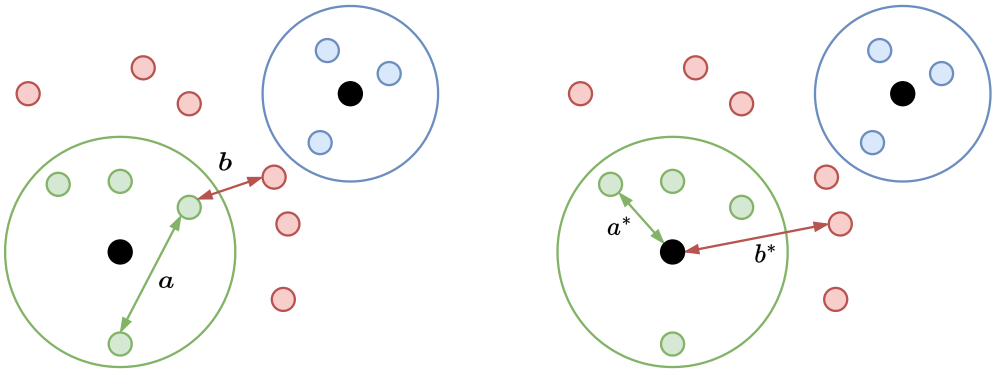
Regarding inherently interpretable DNNs, the latent space is explicitly structured by designing the DNN architecture and loss formulation respectively. Clear assumptions for the latent space claim high intra-class concentration and high inter-class separation. Consequently, these assumptions can be verified by analyzing trained DNNs. Verifiable evidence in terms of inherent interpretability is generated since assumptions concerning the latent space structure are tested by quantifying the quality of cluster concentration and separation. Structuring the latent space is desirable, since previous work indicates that a structured latent space is accompanied by increased robustness and generalization capabilities [71, 82]. Hence, the analysis of the structured latent space substantially contributes to the safety argumentation of DNNs for pedestrian detection.

Although a structure latent space might help provide evidence and argue safety, its natural existence is not guaranteed. Instead, it has to be explicitly enforced by the training strategy. That is why this work is focused on inherent interpretability and not post-hoc interpretability. Due to the application of squared L^2 distances the structure is intuitive and understandable. If a latent representation is close or similar to a prototype or semantic concept vector, it is more likely that the latent representation is a true class realization.

Metrics such as the silhouette coefficient [86], Davies-Bouldin index [17] or Calinski-Harabasz index [8] are commonly used to quantify the clustering quality. The silhouette coefficient is defined as

$$SC = \frac{b - a}{\max\{a, b\}} \tag{3.27}$$

Here, a resembles the average of intra-class distances of a latent representation and b is the minimum inter-class distance. For the proposed inherently interpretable methods and the expected structure of the latent space, the silhouette coefficient



(a) Silhouette coefficient (b) Cluster quality

Figure 3.10: Illustration of the silhouette coefficient and proposed cluster quality with important parameters. To take the spherical shape of the learned concept cluster into consideration, novel parameters for the intra-class distance a^* and inter-class distance b^* are introduced.

has to be adapted to adequately evaluate the cluster quality. Figure 3.10 (a) gives a simple example and shows the implications.

The clustering quality of two clusters (blue, green) shall be evaluated, where a third cluster (red) only holds irrelevant representations and is scattered in the latent space, thus, placed close to or between clusters. This occurs when only task-relevant concepts are learned and irrelevant features are separated. The silhouette coefficient would indicate a weak clustering quality which is counter-intuitive. As a consequence, a^* and b^* are defined according to Figure 3.10 (b).

Meanwhile, the global separation as a more robust extension of the silhouette coefficient has been proposed [3]. Here, only parts of the smallest distances are considered, anticipating an imperfect clustering compared to the ground truth. Similar to the idea of the more robust global separation [3], the parameters a^* or b^* can be reduced to different percentiles Q to account for imperfections. For example, the highly sensitive $CQ_{0.99}$ is defined as:

$$CQ_{0.99} = \frac{Q^{a^*}_{0.99} - Q^{b^*}_{0.01}}{\max\{Q^{a^*}_{0.99}, Q^{b^*}_{0.01}\}} \tag{3.28}$$

Consequently, nearly the maximum of intra-class distances and the minimum of inter-class distances to the background is used to quantify the cluster quality. When applied to the evaluation of multiple clusters, the mean cluster quality can be formalized:

$$mCQ = \sum_{l=1}^{N_C} \frac{b_l - a_l}{\max\{a_l, b_l\}} \tag{3.29}$$

4. Methods

The following section gives a detailed description of all methods developed in this work. In the first step, the DNN for center, scale and offset prediction (CSO) [29] is introduced as a baseline method. CSO is a DNN for pedestrian detection with state-of-the-art performance that predicts 2D bounding boxes.

During the course of this work, DNNs for pedestrian detection, which predict 2D bounding boxes, have been extended to segmentation tasks. The DNN for center, scale and offset prediction and segmentation (CSOS) extends CSO to the prediction of an auxiliary body part segmentation. Based on simple decision rules, an auxiliary body part segmentation is used to make detection bounding boxes plausible. Although CSOS predicts more plausible 2D bounding boxes, it is still considered a baseline method.

Although all proposed method predict 2D bounding boxes to detect pedestrians in single RGB images, the DNN architectures differ substantially. Here, novel loss formulations are proposed and formally defined. Based on that, the proposed interpretable methods for pedestrian detection are introduced, i.e., DNN for prototype-based pedestrian detection (PPD) [70], DNN for concept-based pedestrian detection (CPD) [26] and DNN for concept-based domain adaptation for pedestrian detection (ConDA) [28].

To describe the proposed adaptations of baseline methods, it is necessary to introduce a common mathematical notation. Thus, the gradual advancement of DNN architectures and loss formulations to make DNNs inherently interpretable is transparent. Subsequently, similarities and differences to baseline methods can be demonstrated in this work.

However, the method descriptions are similar to the previously published work [27, 26, 28]. The common mathematical notation is foremost enabled by introducing a generic view on DNNs for pedestrian detection [29]. Figure 4.1 gives an overview of the baseline and proposed methods in this work.

To demonstrate meaningful benchmarks and a valid interpretation of the ex-

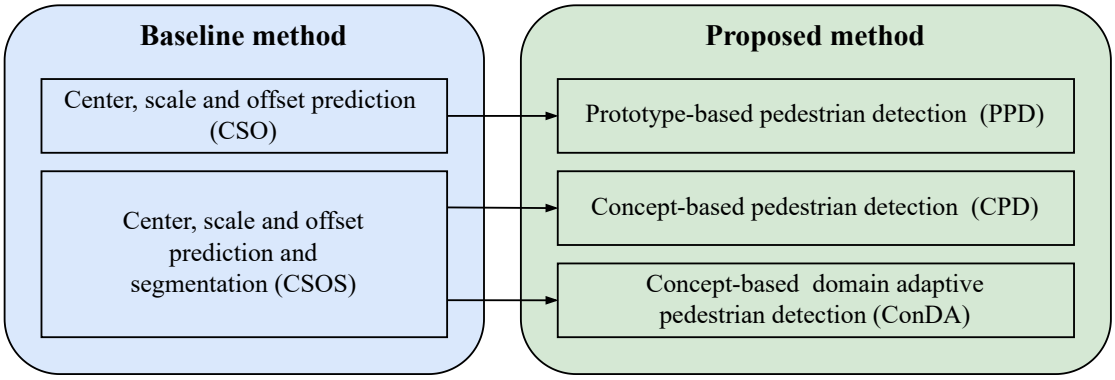


Figure 4.1: Overview of developed baseline methods and proposed methods. Every proposed interpretable DNN for pedestrian detection has a corresponding baseline method.

perimental results, every proposed method is based on a generic and expandable DNN architecture for pedestrian detection. Until this point, the connection to the proposed interpretable methods (PPD, CSP and ConDA) remained unclear. This work fills the gap by using instances of a generic method for pedestrian detection as baselines with state-of-the-art performance to quantify improvements achieved by the proposed methods.

Previously, different possibilities to enable inherent interpretability for pedestrian detection have been identified. On the one hand, unsupervised prototypes characterizing a pedestrian can be learned. On the other hand, predefined semantic concepts of pedestrians can be used to train supervised concept vectors. This chapter describes in detail the necessary adaptations of DNNs to enable inherent interpretability for either unsupervised prototypes (PPD) or supervised semantic concepts (CPD). Finally, a method for concept-based domain adaptation for pedestrian detection (ConDA) based on CPD is proposed.

4.1 Baseline Methods

Research in the field of pedestrian detection is rapidly evolving. Novel backbones [16, 118, 98], loss formulations [107, 119], post-processing methods [74, 124], different DNN architectures (anchor-free or anchor-based, one-stage or two-stage) or additional training data [45] are proposed simultaneously and oftentimes combined in future work. Hence, the conducted benchmarking might be skewed since the

true factors for performance improvement cannot be identified.

Since this work involves the evolution from non-interpretable to inherently interpretable DNNs for pedestrian detection, it is inevitable to define solid baselines beforehand. Reliable benchmarks can only be reported if the baseline method is known in detail. Newly developed methods for pedestrian detection should always be transparent and propose incremental improvements to the baseline method. For this work, a benchmark is given by the performance of a baseline method quantified by detection and assurance metrics. Furthermore, benchmarks are only suitable if the baseline method achieves state-of-the-art performance.

In this work, a method generally corresponds to a DNN for pedestrian detection embedded in the object detection of automated driving systems. Especially, different DNN architectures and loss formulations and their implications are investigated. Nonetheless, other aspects are crucial for the training and validation of a DNN such as the training configuration, post-processing steps for the inference and dataset generation. Although a deeper analysis of these aspects in terms of inherent interpretability is out of scope for this work, applied methods and data must be determined in advance.

Deep Neural Network Architecture DNN architectures largely vary with respect to the computer vision discipline: semantic segmentation, e.g., Deeplabv1 [10] and Deeplabv3+ [11], panoptic segmentation with EfficientPS[78], generative adversarial networks [15] or pedestrian detection, e.g., APD [119] and CSP [70]. Therefore it is important to develop task-specific DNN architectures. For this work, a generic one-stage DNN architecture for pedestrian detection is defined as the composition $m = f \circ h$ of the feature extraction f and a set of perception heads h. For the sake of simplicity and in accordance with the state of the art, an anchor-free and one-stage architecture is chosen. Due to this generic DNN architecture, adaptations can be easily explained and methods with different feature extractions and backbones can be pair-wise investigated. Fair and transparent pair-wise comparisons are inevitable in order to identify reasons for performance improvements.

Figure 4.2 shows a generic DNN architecture for pedestrian detection. A trained DNN is used to predict a set of 2D detection bounding boxes: $\mathcal{D}(\mathrm{c}) = m_\theta(\mathbf{X})$. The feature extraction $f : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{H' \times W' \times D}$ encodes an image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$ by learning relevant semantic features. Task-dependent features are extracted by spatial reduction, i.e, $H \rightarrow H'$ with $H' = \frac{H}{s}$ and downsampling rate $s = 4$ (W' accordingly), eventually increasing the channel size ($3 \rightarrow D$) of the latent tensor

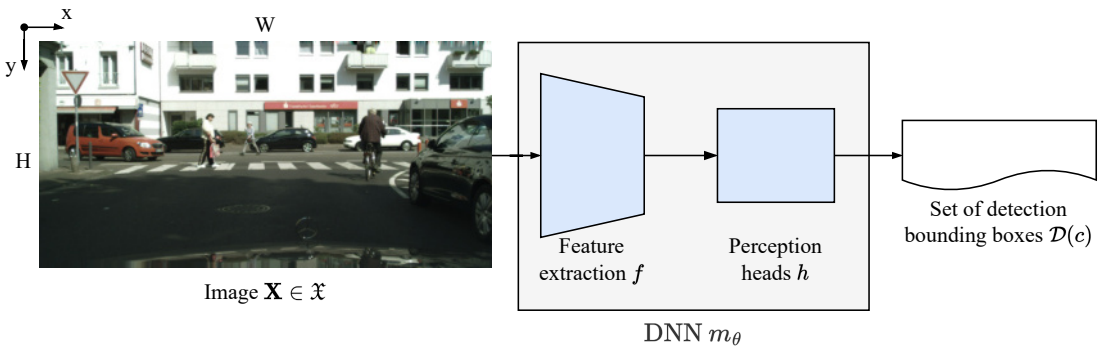


Figure 4.2: Generic DNN architecture for pedestrian detection. The feature extraction encodes a given image $\mathbf{X}$ and predicts a set of detection 2D bounding boxes $\mathcal{D}$(c) with respect to a pre-defined confidence threshold c.

$\mathbf{Z}$. The output tensor is seen as an encoding or intermediate representation of the original input image $\mathbf{X}$.

Multiple perception heads utilize extracted features of the spatially reduced output tensor to solve pre-determined perception tasks. Depending on the perception task, varying perception heads are used. Most commonly in the field of pedestrian detection, perceptions heads consist of convolutional layers and output a center, scale and offset prediction to predict 2D bounding boxes. Although not part of this work, the generic DNN architecture can be easily extended towards 3D bounding boxes, an auxiliary instance segmentation or depth estimation by adding appropriate perception heads.

The main component of feature extraction is the backbone. Most commonly, backbones are pre-trained on the ImageNet [18] dataset. In this work, the following backbones are used: ResNet-50 [46], DLA [118, 119] and HRNet [98]. However, backbones must be embedded in the feature extraction with additional layers to extract knowledge from different resolutions. Hence the applied feature extractions are denoted as modified DLA (MDLA) [118, 16] employed by APD [119], CSP-ResNet-50 employed by CSP [70], FPN-ResNet-50 employed by APD [119, 65] and BGC-HRNet employed by BGCNet [61]. It is important to note that the integration of a backbone (ResNet-50) can be realized in different ways leading to different implementations, i.e., CSP [70] with CSP-ResNet-50 and APD [119] with FPN-ResNet-50. Meanwhile, DNNs for pedestrian detection are often developed

and evaluated with different feature extractions, e.g., APD can be used with either MDLA or FPN-ResNet-50 as feature extraction.

Loss Formulation For every perception head, a suitable loss formulation must be defined. Hence, the sum of the individual loss contributions build the overall loss. Loss terms describe cost functions that estimate the error of a task-specific prediction [60]. Most commonly, classification tasks are solved with a cross entropy loss. Contrarily, L1 losses or related loss definitions are used for regression tasks. Finally, the error is backpropagated and the applied optimization algorithm updates the parameters, weights and biases, of the DNN accordingly [60]. Lately, more advanced and robust optimizers than stochastic gradient descent such as Adam [58] have been proposed and widely applied in computer vision. The number of loss terms is not limited to the number of perception heads. Throughout the work, it is shown how additional loss terms establish inherent interpretability of DNNs for pedestrian detection. This is done by formulating additional training goals and defining auxiliary loss terms.

Training All of the developed methods are implemented with the deep learning library PyTorch[1]. The Adam optimizer [58] with a learning rate of 10^{-4} is used for every training run. Unless otherwise stated, a linear warm-up strategy for 2000 iterations is also applied. Considering other metrics or performing multiple training runs with different seeds or initialization methods may yield different DNNs.

The Xavier initialization [42] is applied to the weights of all additional layers in the feature extraction or perception heads. The stochastic nature of deep learning is increased by applying random data augmentation techniques proposed by CSP [70]. Unless otherwise stated, the data augmentation of CSP is used for every training run.

DNNs are trained for a maximum of 50.000 iterations on two GPUs with a total batch size of 8 and a mini-batch size of 4. With ConDA, the learning rate subsequently decreases linearly. Following the state of the art, DNN parameters θ' of the teacher network $\check{m}_{\theta'}$ are constantly averaged using a student-teacher framework to stabilize training and increase performance [100].

The input image size of training data is reduced to 640x1028 pixels. Ground truth bounding boxes with a height smaller than 50 pixels (w.r.t. original im-

[1] https://pytorch.org/

age size: 1024x2048 pixels) are ignored for training. This approach ensures that training and online evaluation are aligned.

Inference Regarding an anchor-free DNN for pedestrian detection, 2D detection bounding boxes are encoded by the center, scale and offset predictions. To shorten the evaluation time, only center points with a confidence score greater than 0.1 are considered. In alignment with the training, detection bounding boxes with heights smaller than 50 pixels are ignored. As a post-processing step for inference, the greedy non-maximum suppression with a threshold of 0.5 is applied.

Online Evaluation The performance evaluation of a DNN is conducted after every training epoch. Thus, multiple checkpoints of a DNN are created during training. It is important to note, that a checkpoint with the best performance regarding a pre-determined metric represents the final and so-called trained DNN. For the online and offline evaluation and development of state-of-the-art DNNs, the COCO API[2] for the COCO dataset [67] is used. Online refers to evaluation activities during the training of DNN. Contrarily, the offline evaluation concerns the evaluation of already trained DNNs. Therefore, the same evaluation is conducted for both cases.

Regarding the systematic evaluation, substantial changes to the COCO API are implemented resulting in a new evaluation framework[3] [29]. It has been shown that the current limitations or shortcomings of the state-of-the-art evaluation can be mitigated by the use of additional ground truth information. Here, the instance and semantic segmentation of Cityscapes [14] has been used to describe a rule-based categorization of ground truth bounding boxes. Based on that, application-oriented metrics for the (offline or online) evaluation of DNNs for pedestrian detection are defined. In this work, the previously proposed systematic evaluation is extended to the synthetic SynPEDS dataset [97].

Datasets State-of-the-art DNNs for pedestrian detection are most commonly developed for real data such as CityPersons [123], EuroCitypersons [5] or Caltech [20]. Since CityPersons is a part of Cityscapes [14], other sensor information such as disparity maps, instance and semantic segmentation and also panoptic segmentation with Cityscapes-Panoptic-Parts [76] dataset is publicly available. Hence, this work uses the entirety of Cityscapes (CS) [14, 123, 76] as real-world

[2]https://github.com/cocodataset/cocoapi
[3]https://github.com/BeFranke/ErrorCategoriesPedestrianDetection

data for all proposed methods. Cityscapes consists of 2975 images in the training dataset and 500 images in the validation dataset.

In addition, the synthetic dataset SynPeDS [97] created in the course of KI Absicherung is used. Data from SynPeDS has to be pre-processed in order to align with the image size of CityPersons. SynPeDS consists of 9885 images in the training dataset and 696 images in the validation dataset.

4.1.1 Center, Scale and Offset Prediction

The simplest instance of generic DNN architecture for pedestrian detection describes a DNN for center, scale and offset prediction (CSO). CSO is based on CSP [70] which utilizes high-level semantic features to predict the center of a pedestrian as a key point. Meanwhile, the design of the perception heads with convolutional modules (CMs) is inspired by APD [119]. CSO uses a CM for every perception head consisting of a 3x3 convolutional layer followed by a 1x1 convolutional layer.

In conclusion, CSO combines multiple aspects of different previously introduced DNNs for pedestrian detection: (1) keypoint-based pedestrian detection from CSP [70], (2) perception heads from APD [119] and (3) different feature extractions from CSP [70], APD [119] and BGCNet [61] with HRNet [98]. Although ignore areas are explicitly proposed in the CSO loss, the formulation of the center, scale and offset loss is based on the previously introduced CSP [70]. This is consistent with the development of CSO as a baseline method by combining parts from related work. The resulting DNN architecture of CSO is shown in Figure 4.3.

To formulate the loss for the training of CSO, an individual loss term is assigned to every perception head. The overall loss is given by

$$L = L_{\text{center}} + L_{\text{scale}} + L_{\text{offset}} \tag{4.1}$$

In the following, the loss terms for center L_{center}, scale L_{scale} and offset prediction L_{offset} are introduced.

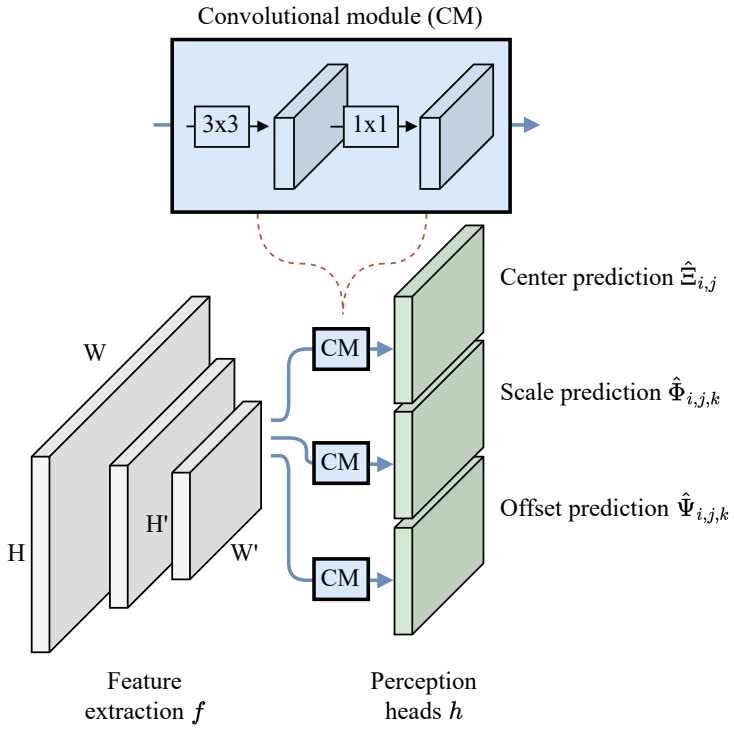


Figure 4.3: Architecture of the DNN center, scale and offset prediction (CSO). CSO is based on CSP [70] and APD [119].

Center Loss CSP introduces Gaussian masks $G_{i,j}$ to account for the problem to estimate the exact center of a pedestrian. Intuitively, there is a high possibility that multiple pixels close to the center predict a high confidence score in $\hat{\Xi}_{i,j}$. Introducing the Gaussian mask ensures that errors outside the vague center area have a higher influence on the loss.

Pedestrian detection faces a high class imbalance since only a few pixels are assigned to pedestrian centers. Generally, the amount of pedestrians N_K is overwhelmed by the number of negative pixels To counteract negative effects on the training, the focal loss puts more weight on correcting false negatives than learning easy true negatives [66]. According to [66], the focal loss parameter is set to $\gamma = 4.0$.

The ignore areas $I_{i,j}^{box}$ account for ground truth labels very similar to pedestrians such as sitting persons or crowds that must be excluded from training to avoid conflicting training impulses.

The final center loss [70] is given by

$$L_{\text{center}} = -\lambda_{\text{center}} \frac{1}{N_K} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \tilde{\mathrm{G}}_{\mathrm{i,j}} (1 - \mathrm{I}_{\mathrm{i,j}}^{\text{box}}) (1 - \tilde{\Xi}_{\mathrm{i,j}})^{\gamma} \log \tilde{\Xi}_{\mathrm{i,j}} \tag{4.2}$$

where

$$\tilde{\mathrm{G}}_{\mathrm{i,j}} = \begin{cases} 1, & \text{if } \Xi_{\mathrm{i,j}} = 1 \\ (1 - \mathrm{G}_{\mathrm{i,j}})^{\beta}, & \text{otherwise} \end{cases}$$

$$\tilde{\Xi}_{\mathrm{i,j}} = \begin{cases} \hat{\Xi}_{\mathrm{i,j}}, & \text{if } \Xi_{\mathrm{i,j}} = 1 \\ 1 - \hat{\Xi}_{\mathrm{i,j}}, & \text{otherwise} \end{cases}$$

$$N_K = \sum_{i=1}^{H'} \sum_{j=1}^{W'} \Xi_{\mathrm{i,j}}$$

According to CSP [70], λ_{center} is set to 0.1 for the experiments.

Scale Loss In contrast to state-of-the-art methods for pedestrian detection that are highly tuned on the most common datasets, the defined scale loss predicts the height as well as the width of a detection bounding box. Most commonly, only the height of a bounding box is predicted and the width is approximated, i.e., $w(g) = 0.41h(g)$ [70] for CityPersons [123].

The loss [70] for scale predictions $\Phi_{\mathrm{i,j,k}}$ and scale ground truth $\hat{\Phi}_{\mathrm{i,j,k}}$ is given by

$$L_{\text{scale}} = \lambda_{\text{scale}} \frac{1}{N_K} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \sum_{k=1}^{2} l\left(\Phi_{\mathrm{i,j,k}}, \hat{\Phi}_{\mathrm{i,j,k}}\right) \tag{4.3}$$

where the smooth L1 loss is defined as

$$l\left(\Phi_{\mathrm{i,j,k}}, \hat{\Phi}_{\mathrm{i,j,k}}\right) = \begin{cases} 0.5\left(\hat{\Phi}_{\mathrm{i,j,k}} - \Phi_{\mathrm{i,j,k}}\right)^2, & \text{if } |\, \hat{\Phi}_{\mathrm{i,j,k}} - \Phi_{\mathrm{i,j,k}} \,| < 1 \text{ and } \Xi_{\mathrm{i,j}} = 1 \\ |\, \hat{\Phi}_{\mathrm{i,j,k}} - \Phi_{\mathrm{i,j,k}} \,| - 0.5, & \text{if } |\, \hat{\Phi}_{\mathrm{i,j,k}} - \Phi_{\mathrm{i,j,k}} \,| \geq 1 \text{ and } \Xi_{\mathrm{i,j}} = 1 \\ 0, & \text{otherwise} \end{cases}$$

According to CSP [70], λ_{scale} is set to 1.0 for the experiments.

Offset Loss The center of a pedestrian is detected based on confidence scores in the down-sampled center matrix $\Xi'_{\mathrm{i,j}}$. Hence, floats of pedestrian center coordinates

$\left(\frac{x_i}{s}, \frac{y_i}{s}\right)$ with the downsampling rate s must be assigned to integers $\left(\left\lfloor \frac{x_i}{r} \right\rfloor, \left\lfloor \frac{y_i}{r} \right\rfloor\right)$. The resulting offsets [70] are formally defined as:

$$o_i^x = \frac{x_i}{r} - \left\lfloor \frac{x_i}{r} \right\rfloor - 0.5$$
$$o_i^y = \frac{y_i}{r} - \left\lfloor \frac{y_i}{r} \right\rfloor - 0.5$$

The offset loss is described as the smooth L1 loss [70] for predicted offsets $\hat{\Psi}_{\mathrm{i,j,k}}$ and offsets targets $\Psi_{\mathrm{i,j,k}}$ as:

$$L_{\mathrm{offset}} = \lambda_{\mathrm{offset}} \frac{1}{N_K} \sum_{i=1}^{H'} \sum_{j=1}^{W'} l\left(\Psi_{\mathrm{i,j,k}}, \hat{\Psi}_{\mathrm{i,j,k}}\right) \tag{4.4}$$

According to CSP [70], $\lambda_{\mathrm{offset}}$ is set to 0.01 for the experiments.

Post-Processing In order to output a detection bounding box for an inferred image, the boxes must be decoded by extracting valid center points from the center predictions $\hat{\Xi}_{\mathrm{i,j}}$. What pixels are accepted depends on the predefined confidence threshold. Height, width and offset of a bounding box are determined based on the position of the pixel of an accepted center in the center prediction $\hat{\Xi}_{\mathrm{i,j}}$. Due to the spatial reduction of the prediction maps, the box properties must be upsampled accordingly. Finally, non-max suppression is applied to suppress strongly overlapping bounding boxes.

4.1.2 Center, Scale and Offset Prediction and Segmentation

Involving a body part segmentation alongside 2D bounding boxes enhances the plausibility of the bounding boxes. The auxiliary segmentation of body parts is intuitively beneficial to unravel critical occlusions for pedestrian detection. State-of-the-art DNNs for pedestrian detection largely rely on high-level semantic features for pedestrian centers, which may be invisible under certain environmental or crowd occlusions. Identified body parts can be used as a supporting source of information to mitigate inaccurate detections. Simple heuristics can be applied to check whether a bounding box actually holds predefined body parts. If the condition does not hold the bounding box can be rejected. Based on CSO, a DNN for center, scale and offset prediction with an auxiliary body part segmentation

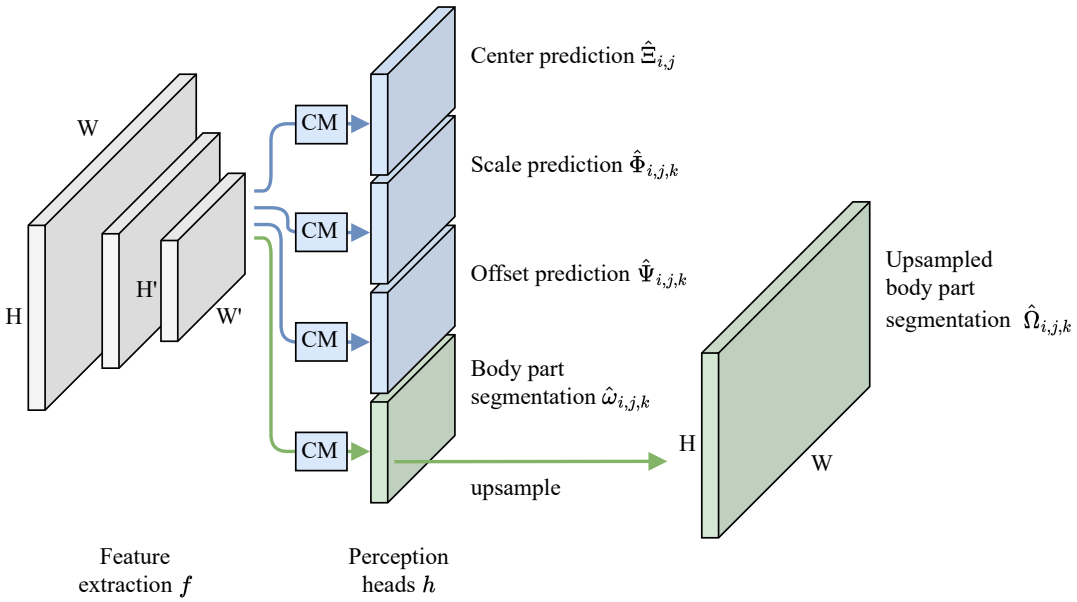


Figure 4.4: Architecture of the DNN for center, scale and offset prediction and segmentation (CSOS). CSOS introduces an auxiliary body part segmentation.

(CSOS) can be described. The DNN architecture of CSOS is shown in Figure 4.4.

Due to the modular and generic approach, the auxiliary body part segmentation can be realized by simply adding another perception head connected with a convolutional module (CM). Inspired by DLA [118] as a light-weight option for semantic segmentation, the segmentation prediction $\omega_{\mathrm{i,j,k}}$ is upsampled by fixed deconvolution layers to compute the final confidence scores:

$$\omega_{\mathrm{i,j,k}} \in [0,1]^{H' \times W' \times N_S} \rightarrow \Omega_{\mathrm{i,j,k}} \in [0,1]^{H \times W \times N_S}$$

while N_S depends on the available ground truth data and describes the number of different body parts that shall be predicted including a background class.

In addition to the relevant loss terms for center, scale and offset prediction, a loss term for the body part segmentation has to be added for CSOS. The final loss formulation is given by

$$L = L_{\mathrm{center}} + L_{\mathrm{scale}} + L_{\mathrm{offset}} + L_{\mathrm{segmentation}} \tag{4.5}$$

To learn an auxiliary body part segmentation, a focal binary cross entropy loss

is defined:

$$L_{\text{segmentation}} = -\lambda_{\text{segmentation}} \frac{1}{N_S} \sum_{k=1}^{N_S} \sum_{i=1}^{H} \sum_{j=1}^{W} (1 - \mathrm{I}_{\mathrm{i,j}}^{\text{inst}})(1 - \tilde{\Omega}_{\mathrm{i,j,k}})^{\gamma} \log(\tilde{\Omega}_{\mathrm{i,j,k}}) \quad (4.6)$$

where

$$\tilde{\Omega}_{\mathrm{i,j,k}} = \begin{cases} \hat{\Omega}_{\mathrm{i,j,k}} & \text{if } \Omega_{\mathrm{i,j,k}} = 1 \\ 1 - \hat{\Omega}_{\mathrm{i,j,k}} & \text{otherwise} \end{cases}$$

Here, the ignore mask $\mathrm{I}_{\mathrm{i,j}}^{\text{inst}}$ is used to exclude ignored instances such as too small pedestrians or cyclists in order to not distort the segmentation loss. For the experiments, $\lambda_{\text{segmentation}}$ is set to 2.0.

Post-Processing In addition to the decoding and non-max suppression of detection bounding boxes, simple decision rules for the acceptance of detection bounding boxes are established. The rule-based approach enhances the plausibility of bounding boxes. It is required that the area of a bounding box shows at least one body part segment.

4.2 Prototype-based Pedestrian Detection

When it comes to the inference of a DNN, the latent tensor, which refers to the output of the feature extraction process, is responsible for encoding a given image. Meanwhile, latent representations hold decisive features for perception tasks. In the case of the interpretable ProtoPNet [9], emphasizing the training of latent representations in the latent shape requires explicit adaptations of the DNN architecture and loss formulation of a DNN.

In order to establish interpretable reasoning for pedestrian detection, the constraints on task-dependent interpretability are relaxed. The goal is to detect pedestrians based on learned features that can be separated from background information, rather than implementing case-based reasoning for classification. Therefore, the class-specific prototypes of ProtoPNet are remodeled into prototypes that are focused on the center of a pedestrian.

The following reasoning from [27] is established: A pedestrian is detected since its extracted features of the latent representation are similar to a prototype for a pedestrian center. The DNN for prototype-based pedestrian detection (PPD) enforces inherent interpretability by explicitly structuring the latent space through

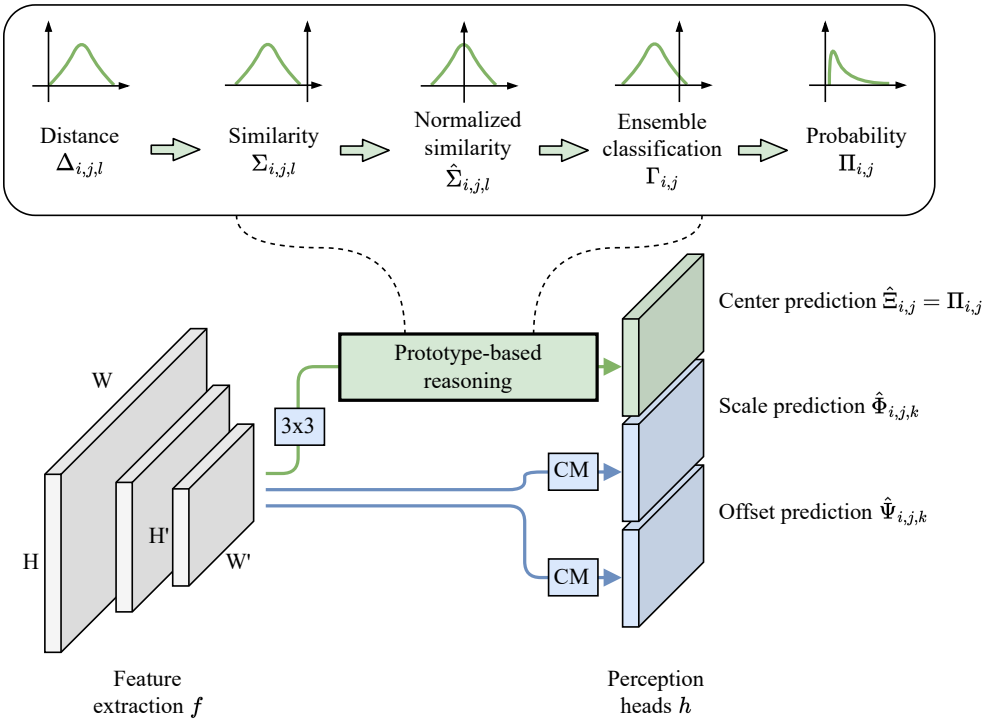


Figure 4.5: Architecture of the DNN for prototype-based pedestrian detection (PPD). PPD introduces an interpretable reasoning using unsupervised prototypes for pedestrian centers.

prototypes for a pedestrian center.

4.2.1 Deep Neural Network Architecture

Building upon the DNN architecture of CSO as a baseline method, PPD integrates prototype-based reasoning as shown in Figure 4.5. To be comparable with the convolutional module (CM) of CSO's center path, a prior 3x3 convolutional layer is used.

PPD learns a set of prototypes $\mathcal{P}$ as cluster centers. Here, the number of prototypes is arbitrarily set to 4. For every latent representation, the distance to a prototype vector is calculated. Due to the application of multiple prototypes, an ensemble of classifiers is used for the center prediction.

Interpretable Reasoning based on Prototypes Based on Equation 3.1, the squared L^2 distance between latent representations $z_k^{i,j} \in \mathbf{Z}$ and the l-th prototype

$\mathrm{p}_{\mathrm{k}}^{\mathrm{l}}$ in a set of prototype $\mathcal{P}$ is defined as

$$\Delta_{\mathrm{i,j,l}} = \sum_{k=1}^{D} \left(\mathrm{Z}_{\mathrm{i,j,k}} - \mathrm{p}_{\mathrm{k}}^{\mathrm{l}}\right)^2 \quad \text{for } i = 1, \ldots, H' \text{ and } j = 1, \ldots, W' \text{ and } \mathrm{p}_{\mathrm{k}}^{\mathrm{l}} \in \mathcal{P} \quad (4.7)$$

Negative squared L^2 distances are a simple similarity measure, leading to

$$\Sigma_{\mathrm{i,j,l}} = -\Delta_{\mathrm{i,j,l}} \quad (4.8)$$

The normalization of similarities allows to identify the distance range for a given training dataset acting as a domain-specific discriminator. Normalized similarities are defined as

$$\hat{\Sigma}_{\mathrm{i,j,l}} = \frac{\Sigma_{\mathrm{i,j,l}} - \mathrm{E}[\Sigma_{\mathrm{i,j,l}}]}{\sqrt{\mathrm{Var}[\Sigma_{\mathrm{i,j,l}}] + \epsilon}} \quad (4.9)$$

Since multiple prototypes are part of the reasoning, the normalized similarities $\hat{\Sigma}_{\mathrm{i,j,l}}$ must be interpreted by a linear weighting to gain an ensemble classification defined as

$$\Gamma_{\mathrm{i,j}} = \alpha \cdot \hat{\Sigma}_{\mathrm{i,j,l}} + \eta \quad (4.10)$$

Based on the transformed and normalized similarities $\Gamma_{\mathrm{i,j}}$, the final probability of a predicted pedestrian center is given by

$$\Pi_{\mathrm{i,j}} = \frac{1}{1 + e^{-\Gamma_{\mathrm{i,j}}}} \quad (4.11)$$

Figure 4.5 arranges all of the five important steps that are summarized as the interpretable prototype-based reasoning: (1) squared L^2 distance $\Delta_{\mathrm{i,j,l}}$, (2) similarity $\Sigma_{\mathrm{i,j,l}}$, (3) normalized similarity $\hat{\Sigma}_{\mathrm{i,j,l}}$, (4) ensemble classification $\Gamma_{\mathrm{i,j}}$ and (5) probability $\Pi_{\mathrm{i,j}}$.

4.2.2 Loss Formulation

As shown in Figure 4.5, the proposed prototype-based deep neural network for pedestrian detection extends the architecture of the deep neural network for center, scale and offset prediction. The final loss formulation for PPD is defined as

$$L = L_{\text{center}} + L_{\text{scale}} + L_{\text{offset}} + L_{\text{cluster}}^{\mathrm{P}} + L_{\text{separation}}^{\mathrm{P}} + L_{\text{regularizer}} \quad (4.12)$$

Thus, loss terms to penalize low intra-class concentration $L_{\text{cluster}}^{\mathrm{P}}$ and low inter-

class separation $L^{\mathrm{P}}_{\mathrm{separation}}$ to structure the latent space are introduced. Additionally, a regularization term $L_{\mathrm{regularizer}}$ is proposed to increase the diversity of the involved prototypes in the decision-making process. All loss terms are described in detail in the following.

Prototype Clusters Clustering latent representations is achieved by minimizing distances $\Delta_{\mathrm{i,j,l}}$ to prototypes. The loss term is crucial for the structure of the latent space and is defined as

$$L^{\mathrm{P}}_{\mathrm{cluster}} = \frac{\lambda^{\mathrm{P}}_{\mathrm{cluster}}}{N_K} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \Xi_{\mathrm{i,j}} \min_{l:\mathbf{p}^{\mathrm{l}} \in \mathcal{P}} \Delta_{\mathrm{i,j,l}} \tag{4.13}$$

Since the distance depends on the features of latent representations $\mathrm{z}^{\mathrm{i,j}}_{\mathrm{k}} \in \mathrm{Z}_{\mathrm{i,j,k}}$ and prototypes, the loss term can be reformulated:

$$L^{\mathrm{P}}_{\mathrm{cluster}} = \frac{\lambda^{\mathrm{P}}_{\mathrm{cluster}}}{N_K} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \Xi_{\mathrm{i,j}} \min_{\mathbf{p}^{\mathrm{l}} \in \mathcal{P}} \|\mathbf{z}^{\mathrm{i,j}} - \mathbf{p}^{\mathrm{l}}\|_2^2 \tag{4.14}$$

The number of given pedestrians in an image is given by N_K.

Latent representations of the latent tensor and prototypes are optimized simultaneously throughout the training. The multiplication with $\Xi_{\mathrm{i,j}}$ guarantees that only positive samples for pedestrian centers are clustered around the assigned prototype. For every pedestrian in every training step, the prototype with minimal distance is identified and the distance is further minimized. For the experiments, $\lambda^{\mathrm{P}}_{\mathrm{cluster}}$ is set to 10^{-3}.

Prototype Separation In contrast to learning clusters with high intra-class concentration, the distance of prototypes to negative samples in the background shall be maximized. This is achieved by applying the mask $(1 - \Xi_{\mathrm{i,j}})$ that excludes all pixels within 2D bounding boxes including ignored bounding boxes.

The loss formulation to learn high inter-class separation is given by

$$L^{\mathrm{P}}_{\mathrm{separation}} = -\frac{\lambda^{\mathrm{P}}_{\mathrm{separation}}}{N\text{-}} \sum_{i=1}^{H'} \sum_{j=1}^{W'} (1 - \Xi_{\mathrm{i,j}}) \min_{\mathbf{p}^{\mathrm{l}} \in \mathcal{P}} \|\mathbf{z}^{\mathrm{i,j}} - \mathbf{p}^{\mathrm{l}}\|_2^2 \tag{4.15}$$

with

$$N^{-} = \sum_{i=1}^{H'} \sum_{j=1}^{W'} 1 - \Xi_{\mathrm{i,j}} \tag{4.16}$$

N^{-} describes the number of negative samples for an image and $\lambda^{\mathrm{P}}_{\mathrm{separation}}$ is the applied weight for the separation loss. In that case, negative refers to pixels that are outside of bounding boxes and clearly have no reference to a pedestrian center. For the experiments, $\lambda^{\mathrm{P}}_{\mathrm{separation}}$ is set to 10^{-3}. Considering minimal distances guarantees that only the worst-performing prototype is updated. Hence, prototypes that are already well-clustered are not considered.

Inequality Regularization To prevent trivial solutions for prototype clusters, a regularization term is introduced. Since the best-fitting prototype with the smallest distance is chosen and updated for every training iteration, it might be the case that only one prototype is learned. Such trivial solutions relate to the random initialization of DNN parameters and prototypes. Consequently, the training must be guided in the sense that all predefined prototypes are somehow equally used throughout the training.

However, learning discriminative features for different appearances, lighting conditions or poses shall not be hindered. Thus, the loss contribution of the regularization must be balanced. The Gini coefficient for a sample $X = \{x_i\}_{i=1}^{N}$ is defined as

$$GC = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|}{2N^2 \bar{x}} \tag{4.17}$$

Transferred to the inequality of responsible prototypes and their distances, the regularization term is given by

$$L_{\mathrm{regularizer}}(\Delta_{\mathrm{i,j,l}}, \Xi_{\mathrm{i,j}}) = \lambda_{\mathrm{regularizer}} \frac{\sum_{i=1}^{H'} \sum_{j=1}^{W'} \tilde{\Delta}_{i,j}}{2N^2 \bar{\Delta}} \tag{4.18}$$

with

$$\tilde{\Delta}_{i,j} = \begin{cases} \sum_{k=1}^{N_P} \sum_{l=1}^{N_P} |\Delta_{i,j,k} - \Delta_{i,j,l}| & \text{if } \Xi_{\mathrm{i,j}} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.19}$$

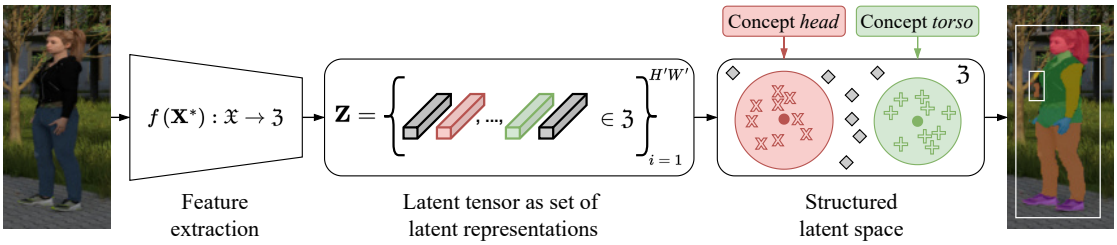


Figure 4.6: Structured latent space of the DNN for concept-based pedestrian detection (CPD). Body parts are considered semantic concepts of humans. Latent representations that extract features from the receptive fields of an image are compared to learnable concept vectors. These concept vectors are cluster centers in a structured latent space.

while the mean distance is defined as

$$\bar{\Delta} = \frac{\sum_{i=1}^{H'} \sum_{j=1}^{W'} \sum_{l=1}^{N_P} \Xi_{\mathrm{i,j}} \, \Delta_{\mathrm{i,j,l}}}{\sum_{i=1}^{H'} \sum_{j=1}^{W'} \Xi_{\mathrm{i,j}}} \tag{4.20}$$

For the experiments, $\lambda_{\mathrm{regularizer}}$ is set to 0.01.

4.3 Concept-based Pedestrian Detection

An obvious extension towards a safety-critical application of DNNs for pedestrian detection is considering additional information. The proposed DNN for concept-based pedestrian detection predicts pedestrians based on the semantic concepts of humans. Eventually, the interpretable body part segmentation is aligned with a parallel detection branch for 2D bounding boxes.

CPD is closely related to the previously introduced DNN for prototype-based pedestrian detection. However, the method has been enhanced towards predicting an auxiliary body part segmentation building upon predefined semantic concepts such as body parts, illustrated in Figure 4.6. CPD learns multiple perception tasks simultaneously in a shared latent space.

The reasoning is considered to be interpretable since extracted latent representations are seen as concept realizations due to their high similarity to learned

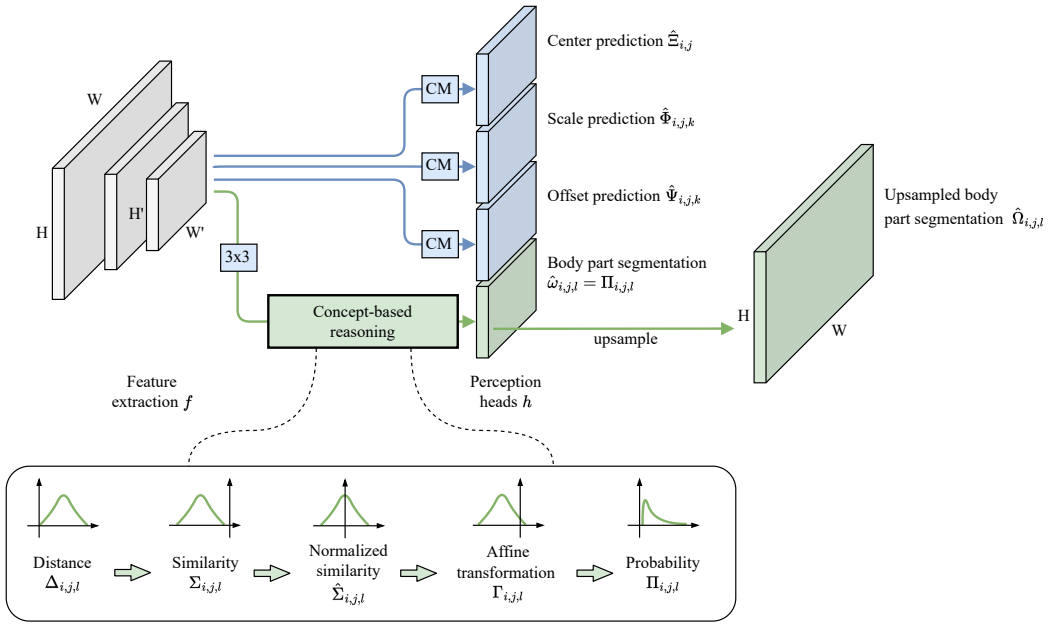


Figure 4.7: Architecture of the DNN for concept-based pedestrian detection (CPD). CPD introduces an interpretable reasoning considering semantic concepts such as body parts of humans.

semantic concept vectors in a structured latent space. Intuitively, it is more interpretable to humans than the common application of softmax loss based on the inner product as a similarity measure of features from a latent space with an unknown structure.

4.3.1 Deep Neural Network Architecture

Building upon the baseline method CSOS and interpretable prototype-based reasoning of PPD, CPD integrates concept-based reasoning. Figure 4.7 illustrates the developed DNN architecture of CPD with an auxiliary body part segmentation and interpretable concept-based reasoning.

Interpretable Reasoning based on Concepts Based on Equation 3.1, the squared L^2 distance between latent representations $z_k^{i,j} \in \mathbf{Z}$ and the l-th semantic

concept vector $\mathrm{c}_{\mathrm{k}}^{\mathrm{l}}$ in a set of concept vectors $\mathcal{C}$ is defined as

$$\Delta_{\mathrm{i,j,l}} = \sum_{k=1}^{D} \left(\mathrm{Z}_{\mathrm{i,j,k}} - \mathrm{c}_{\mathrm{k}}^{\mathrm{l}}\right)^2 \quad \text{for } i = 1, \ldots, H' \text{ and } j = 1, \ldots, W' \text{ and } \mathrm{c}_{\mathrm{k}}^{\mathrm{l}} \in \mathcal{C} \quad (4.21)$$

Similar to PPD, similarities $\Sigma_{\mathrm{i,j,l}} = -\Delta_{\mathrm{i,j,l}}$ are given by the negative squared Euclidean distances. The normalization

$$\hat{\Sigma}_{\mathrm{i,j,l}} = \frac{\Sigma_{\mathrm{i,j,l}} - \mathrm{E}[\Sigma_{\mathrm{i,j,l}}]}{\sqrt{\mathrm{Var}[\Sigma_{\mathrm{i,j,l}}] + \epsilon}} \quad (4.22)$$

is part of learning the decision boundary of concept acceptance.

In contrast to PPD, similarities must not be weighted by an ensemble classification. Instead, the transformation is applied for every l-th concept and is given by

$$\Gamma_{\mathrm{i,j,l}} = \beta \cdot \hat{\Sigma}_{\mathrm{i,j,l}} + \eta \quad (4.23)$$

Finally, the probability of a concept realization is defined as the sigmoid following

$$\Pi_{\mathrm{i,j,l}} = \frac{1}{1 + e^{-\Gamma_{\mathrm{i,j,l}}}} \quad (4.24)$$

4.3.2 Critical Distance

The segmentation of body parts is modeled as a concept-wise binary decision. Only if $\Pi_{\mathrm{i,j,l}} > 0.5$ is true, a latent representation is considered a semantic concept. In this case, the probability is higher than 50%. For the previously defined transformations of the interpretable reasoning, a critical distance δ_{l}^* for a concept realization can be calculated:

$$\delta_{\mathrm{l}}^* = \frac{\eta}{\beta} \sqrt{\mathrm{Var}[\Delta_{\mathrm{i,j,l}}] + \epsilon} - \mathrm{E}[\Delta_{\mathrm{i,j,l}}] \quad (4.25)$$

The critical distance determines the decision boundary for the l-th semantic concept vector $\mathrm{c}_{\mathrm{k}}^{\mathrm{l}}$. Consequently, the distance range of a valid concept realization for a concept vector $\mathrm{c}_{\mathrm{k}}^{\mathrm{l}}$ is $[0, \delta_{\mathrm{l}}^*)$. In conclusion, a latent representation with $\delta < \delta^*$ is seen as a concept realization.

4.3.3 Loss Formulation

Extending the loss formulation of CSOS, novel loss terms for clustering and separation of semantic concepts are proposed. The overall loss of CPD is defined as

$$L = L_{\text{center}} + L_{\text{scale}} + L_{\text{offset}} + L_{\text{segmentation}} + L^{\text{C}}_{\text{cluster}} + L^{\text{C}}_{\text{separation}} \tag{4.26}$$

In the following, the loss terms for the opposing mechanisms of clustering and separation are described in detail.

Semantic Concept Clustering The cluster loss of CPD increases intra-class concentration by pulling positive samples of latent representations closer to the concept vector. Based on the one-hot encoded concept mask $\omega_{\text{i,j,l}}$ the loss term is defined as

$$L^{\text{C}}_{\text{cluster}} = \frac{\lambda^{\text{C}}_{\text{cluster}}}{N^{+}} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \sum_{l=1}^{N_C} \omega_{\text{i,j,l}} \left\| \mathbf{z}^{\text{i,j}} - \mathbf{c}^{\text{l}} \right\|_2^2 \tag{4.27}$$

with

$$N^{+} = \sum_{i=1}^{H'} \sum_{j=1}^{W'} \sum_{l=1}^{N_C} \omega_{\text{i,j,l}} \tag{4.28}$$

for the l-th concept vector $\text{c}^{\text{l}}_{\text{k}}$ in the set of concept vectors $\mathcal{C}$. For the experiments, $\lambda^{\text{C}}_{\text{cluster}}$ is set to 0.1.

Semantic Concept Separation In contrast to building clusters by minimizing intra-class distances, the separation loss encourages the extraction of discriminative features with high inter-class distances to separated clusters. Hence, clusters for different semantic concepts such as body parts and background shall be separated in the structured latent space. Based on the one-hot encoded separation mask $\rho_{\text{i,j,l}}$ and safety margin δ'_{l} the general loss term for separation is defined as

$$L^{\text{C}}_{\text{separation}} = \frac{\lambda^{\text{C}}_{\text{separation}}}{N^{\text{-}}} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \sum_{l=1}^{N_C} \rho_{\text{i,j,l}} \left(\psi_l(\delta) \left(\delta'_{\text{l}} - \left\| \mathbf{z}^{\text{i,j}} - \mathbf{c}^{\text{l}} \right\|_2^2 \right) \right)^2 \tag{4.29}$$

for the l-th concept vector $\text{c}^{\text{l}}_{\text{k}}$ in the set of concept vectors $\mathcal{C}$. The separation mask $\rho_{\text{i,j,l}}$ is used to either separate concepts from each other or concepts from the background. In the case of background separation, the one-hot encoded mask describes pixels that do not belong to any body part. Regarding the concept separation, the mask must be adjusted channel-wise to show every other remaining

concept except the one being examined.

Since separation is only necessary at a close distance to the decision boundary marked by the critical distance the loss contribution is decreasingly weighted:

$$\psi_l(\delta) = \exp\left(\frac{\log(z)}{\delta_\mathrm{l}' - \delta_\mathrm{l}^*} \cdot \left(\delta - \delta_\mathrm{l}^*\right)\right) \tag{4.30}$$

Based on the safety margin δ_l' as a multiple of the critical distance (arbitrarily set to $\delta_\mathrm{l}' = 3\,\delta_\mathrm{l}^*$), the separation demand is weighted. Here, an adequately small number z determines the weight that is assigned to already well-separated latent representations. For the experiments, z is set to 10^{-4}. The amount of negative pixels is given by

$$N^{\text{-}} = \sum_{i=1}^{H'} \sum_{j=1}^{W'} \sum_{l=1}^{N_C} \rho_{i,j,l} \tag{4.31}$$

Figure 4.8 gives an impression of how different distances δ of negative latent representations are weighted and contribute to the separation loss. Finally, two separation losses have to be distinguished, one for separation to other semantic concepts $L^{\mathrm{C}}_{\mathrm{separation,C}}$ and one for background separation $L^{\mathrm{C}}_{\mathrm{separation,B}}$. This depends on the applied separation mask $\rho_{\mathrm{i,j,l}}$. For the experiments, $\lambda^{\mathrm{C}}_{\mathrm{separation}}$ for concept separation is set to 0.1 and for background separation to 10.0.

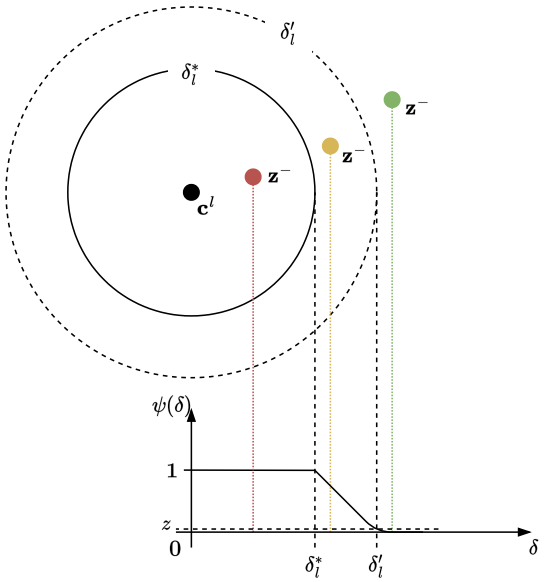


Figure 4.8: Separation of color-coded negative latent representations $\mathbf{z}^-$. Different weights ψ_l are applied depending on the remaining safety margin δ_l' to the critical distance δ_l^*. Red demands strong separation, yellow is already outside the decision boundary but still close and green is acceptable.

4.4 Domain Adaptation based on Semantic Concepts

State-of-the-art DNNs for pedestrian detection are specifically designed for and trained on real datasets [20, 123, 5]. Hence, a limiting factor is the expensive manual generation of ground truth annotations. Labeling 2D bounding boxes or generating a semantic, instance or body part segmentation for real-world images is labor-intensive and time-consuming. Furthermore, the manual labeling of ground truth data is prone to error since annotations are created with incomplete information. In particular, errors or inaccuracies can easily occur in the case of highly occluded or small pedestrians. Intuitively, it is hard to position full bounding boxes correctly for pedestrians standing in a large group or behind a car. Possible annotation errors can cause optimization issues due to loss divergence or ill-defined optimization goals of DNNs.

Contrarily, synthetic data generation has the potential to be more cost-effective,

customizable and scalable. Nevertheless, the comprehensive specification of synthetic data and developing a data generation process have its difficulties. The synthetic data generation must be specified carefully and experts for the investigated computer vision tasks and machine learning models should be consulted early on. However, the rendering process enables full control over training and validation data. Data can be specifically generated for different scenarios or use cases for automated driving.

Due to the accessibility of synthetic data, additional task-specific information such as body part segmentation, instance segmentation or depth maps can easily be provided. There are initiatives focusing on synthetic data for semantic segmentation [84, 85] or pedestrian detection such as MOTSynth [25] or SynPeDS [97]. However, even photo-realistic synthetic data exhibits an immense domain shift to real-world data. The naive source-only training with synthetic data of DNNs generalizes poorly to the real world. A large body of work focuses either on unsupervised domain adaptation (UDA) for semantic segmentation [122, 48, 12, 114, 49]) or object detection [96, 120, 108, 105, 113, 13, 96, 81].

The concept-based domain adaptation for pedestrian detection (ConDA), presented in the following, can be seen as a first step towards UDA for pedestrian detection leveraging synthetic data. It is important to note that ConDA does not access ground truth for real-world data as illustrated in Figure 4.9.

4.4.1 Training Strategy

In contrast to single-stage and end-to-end approaches such as DAFormer [48] or SePiCo [114], the proposed ConDA opts for a conventional self-training approach similar to ProDA [122]. This involves separate stages for the training of a well-generalizing DNN and unsupervised domain adaptation as finetuning.

The proposed method for concept-based domain adaptation has two stages: (1) training on the source domain $\mathfrak{X}_\mathrm{S}$ (naive source-only training) and (2) self-training with pseudo ground truth. Moreover, DAFormer and SePiCo focus on semantic segmentation and not pedestrian detection. However, ConDA uses CPD as a base model that predicts an auxiliary body part segmentation.

In the first stage, CPD m is trained on images $\mathbf{X}_\mathrm{S} \in \mathfrak{X}_\mathrm{S}$ from a source domain $\mathfrak{X}_\mathrm{S}$ with assigned ground truth labels Y_S. The source domain is represented by synthetic data from SynPeDS. The naive source-only training is done with a student-and-teacher approach. Details on the source-only training of the first ConDA stage are given by Algorithm 4.1. When the training is finished, the best-

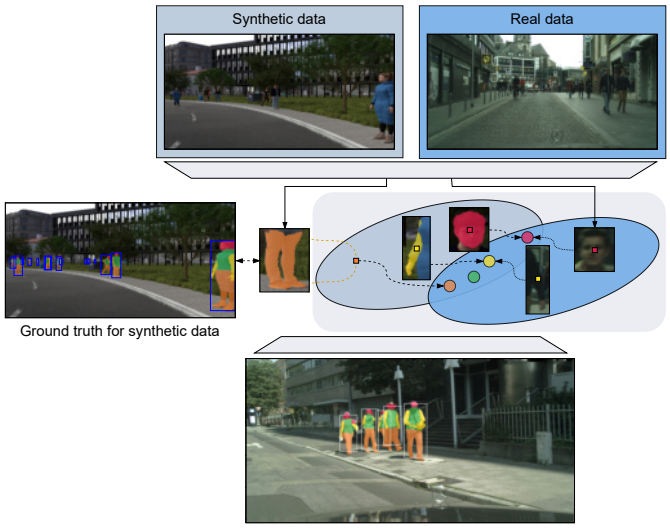


Figure 4.9: Illustration of the concept-based domain adaptation (ConDA) for pedestrian detection. ConDA is based on concept-based pedestrian detection (CPD) that considers body parts as semantic concepts of humans. Latent representations of concepts are clustered in the structured latent space. ConDA enhances the supervised learning (dashed) from synthetic data with conventional self-training (dotted) from unlabeled real data.

Algorithm 4.1: Pseudocode for naive source-only training in the first stage of the DNN for concept-based domain adaptation (ConDA).

Input: Source dataset $(\mathfrak{X}_\mathrm{S}, Y_\mathrm{S})$ and DNN m.
Output: Pre-trained DNN m_θ.

Instantiate student DNN m and initialize parameters θ: $\hat{m}_\theta$
Instantiate teacher network $\breve{m}_{\theta'}$ and set parameters θ': $\breve{m}_{\theta'} \leftarrow \hat{m}_\theta$
while $i \leq$ iterations **do**
    Get source images $\mathbf{X}_\mathrm{S}$ and ground truth Y_S
    Train student $\hat{m}_\theta$ with supervised loss L^S
    EMA update of teacher $\breve{m}_{\theta'}$: $\theta'_{i+1} \leftarrow \alpha\,\theta'_i + (1-\alpha)\,\theta_i$
    Update batch norm statistics of teacher: $\breve{m}_{\theta'} = \breve{m}_{\theta'}(\mathbf{X}_\mathrm{S})$
$m_\theta \leftarrow \breve{m}_{\theta'}$

performing teacher DNN $\breve{m}$ with average weights represents the final DNN of the first training stage of ConDA.

Algorithm 4.2: Pseudocode for self-training in the second stage of the DNN for concept-based domain adaptation (ConDA).

Input: Dataset $(\mathfrak{X}_\mathrm{S}, Y_\mathrm{S}, \mathfrak{X}_\mathrm{T})$ and pre-trained DNN m_θ with parameters θ.
Output: Domain-adapted DNN m_θ.

```
Instantiate student network m̂_θ and initialize parameters θ: m̂_θ ← m_θ
Set oracle network m̄_θ̄ and freeze parameters θ̄: m̄_θ̄ ← m_θ
Instantiate teacher network m̆_θ' and set parameters θ': m̆_θ' ← m_θ
while i ≤ iterations do
    Get source images X_S and ground truth Y_S
    Train student m̂_θ with supervised loss L_S
    Get target images X_T
    Get pseudo ground truth Ȳ_T from oracle network m̄_θ̄
    Train student m̂_θ with unsupervised loss L_T
    EMA update of teacher m̆_θ'
    Update batch norm statistics of teacher: m̆_θ' = m̆_θ'(X_T)
m_θ ← m̆_θ'
```

In the second stage, the trained CPD m_θ with trained parameters θ is adapted to the target domain in a student-teacher-oracle approach. The oracle DNN $\bar{m}$ is initialized with parameters θ that are not optimized in the whole stage. The oracle predictions are used as pseudo ground truth for the domain adaptation applying a conventional self-training approach. The final domain-adapted DNN m is trained in a student-teacher framework. Algorithm 4.2 gives all details on the different parameter updates and the dynamic view on the second stage of ConDA.

Additionally, Figure 4.10 gives an overview of ConDA's training strategy and DNN architecture and all involved components such as the oracle $\bar{m}$, student $\hat{m}$ and teacher DNN $\breve{m}$ of the second training stage.

As a result, ConDA adapts the teacher DNN $\breve{m}$ to images $\mathbf{X}_\mathrm{T} \in \mathfrak{X}_\mathrm{T}$ of the real target domain of Cityscapes $\mathfrak{X}_\mathrm{T}$. Thus, predictions such as detection bounding boxes $\breve{\mathcal{D}}_T(c)$ and body part segmentation $\breve{\Omega}^T_{i,j,l}$ perform better than predictions of a DNN with naive source-only training.

Note that the final domain-adapted DNN is the teacher DNN $\breve{m}$ and not the student DNN that is actually optimized. The weights of the teacher are just the exponential moving average of the student DNN's weights.

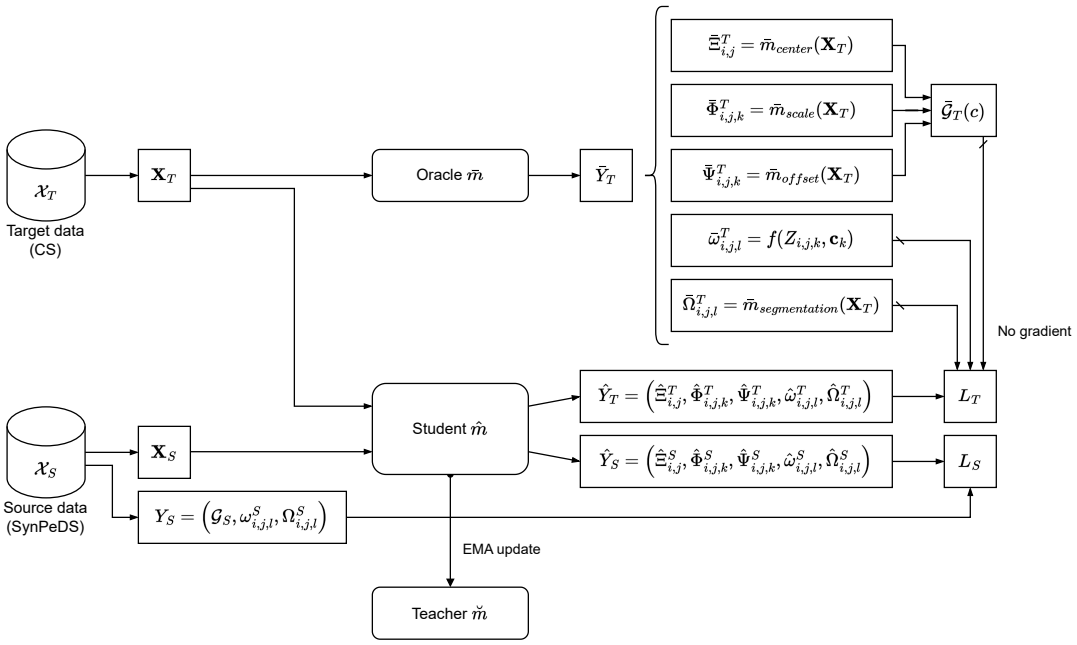


Figure 4.10: Student-teacher-oracle approach of the DNN for concept-based domain adaptation (ConDA). ConDA uses a conventional self-training approach for unsupervised domain adaptation.

4.4.2 Loss Formulation

Due to ConDA's domain adaptation approach the loss formulation consists of two parts:

$$L = L_S + L_T \tag{4.32}$$

where L_S corresponds to the supervised loss of CPD in Equation 4.26 for the source domain with SynPeDS data. In the following, the different loss terms of L_T for the self-training on the target domain with unlabeled data from Cityscapes are explained.

Pseudo Bounding Boxes Pseudo ground truth bounding boxes are encoded by the center, scale and offset predictions of the oracle network $\bar{m}$. Center predictions for a sample $\mathbf{X}_\mathrm{T} \in \mathfrak{X}_\mathrm{T}$ drawn from the target domain are defined as $\bar{\Xi}_{\mathrm{i,j}} = \bar{m}_{center}\left(\mathbf{X}_\mathrm{T}\right)$. Generating pseudo ground bounding boxes $\bar{\mathcal{G}}_\mathrm{T}$ is done by applying a decoding function decode$(\bar{\Xi}_{\mathrm{i,j}}, \bar{\Phi}_{\mathrm{i,j,k}}, \bar{\Psi}_{\mathrm{i,j,k}})$ that also includes the non-max

Figure 4.11: Pseudo ignore areas for self-training. Generated ignore areas $\tilde{o}_{i,j}$ (right) for an image from the target domain (left). The center, scale and offset loss ignores the red area and only considers pixels within the blue area. The bounding boxes (blue) show positive samples for pseudo bounding boxes with a high confidence score.

suppression:

$$\bar{\mathcal{G}}_{\mathrm{T}} = \mathrm{decode}\left(\left[\bar{\Xi}_{\mathrm{i,j}} > \tau_{\mathrm{center}}\right], \bar{\Phi}_{\mathrm{i,j,k}}, \bar{\Psi}_{\mathrm{i,j,k}}\right) \tag{4.33}$$

Applying a center threshold τ_{center} guarantees that only bounding boxes with high certainty are used as pseudo ground truth bounding boxes. τ_{center} is empirically set to 0.6. To efficiently denoise pseudo labels, ignored pseudo bounding boxes are defined as

$$\bar{\mathcal{I}} = \mathrm{decode}\left(\left[\bar{\Xi}_{\mathrm{i,j}} > \tau_{\mathrm{ign}} \wedge \bar{\Xi}_{\mathrm{i,j}} \leq \tau_{\mathrm{center}}\right], \bar{\Phi}_{\mathrm{i,j,k}}, \bar{\Psi}_{\mathrm{i,j,k}}\right) \tag{4.34}$$

where τ_{ign} is empirically set to 0.1. Examples of ignored pseudo ground truth bounding boxes are visualized in Figure 4.11.

Finally, the loss terms for the center (Equation 4.2), scale (Equation 4.3) and offset (Equation 4.4) prediction can be built based on the definition of pseudo ground truth bounding boxes $\bar{\mathcal{G}}_{\mathrm{T}}$ and pseudo ignore areas $\bar{\mathcal{I}}$.

Body Part Segmentation The cluster threshold τ_{cluster} (empirically set to 0.6) is applied to define the upsampled hard pseudo concepts masks based on soft predictions of the oracle network $\bar{\omega}_{i,j,k} = \bar{m}(\mathbf{X}_{\mathrm{T}})$ and $\bar{\Omega}_{\mathrm{i,j,k}} = \bar{m}(\mathbf{X}_{\mathrm{T}})$:

$$\bar{\Omega}^{\mathrm{hard}}_{i,j,l} = \left[l = \arg\max_{l'} \bar{\Omega}_{i,j,l'}\right] \tag{4.35}$$

To reduce misleading training impulses due to predictions with low confidence, only highly confident pixels of the body part segmentation are considered. This

done by applying a separation threshold $\tau_{\text{separation}}$. For the experiments, $\tau_{\text{separation}}$ is empirically set to 0.4.

With respect to Equation 4.6, $L_{\text{segmentation}}$ is calculated based on pseudo ground truth segmentation $\Omega_{i,j,l} = \bar{\Omega}^{\text{hard}}_{i,j,l}$, ignore ares $\bar{I}_{i,j,l}$ and predictions of the student DNN $\hat{m}_{\text{segmentation}}(\mathbf{X}_{\text{T}})$ with

$$\bar{I}_{i,j,l} = \left[\tau_{\text{separation}} \leq \bar{\Omega}_{i,j,l} \leq \tau_{\text{cluster}}\right] \tag{4.36}$$

Semantic Concept Clusters Similar to the upsampled hard pseudo concepts masks $\bar{\Omega}^{\text{hard}}_{i,j,l}$, the cluster mask for semantic concepts is defined:

$$\bar{\omega}^{\text{hard}}_{i,j,l} = \left[l = \arg\max_{l'} \bar{\omega}_{i,j,l'} \wedge \bar{\omega}_{i,j,l} > \tau_{\text{cluster}}\right] \tag{4.37}$$

Only highly confident predictions are used to further structure the latent space and improve the clustering of semantic concepts. Finally, $\bar{\omega}^{\text{hard}}_{i,j,l}$ is used to define the loss term for concept clustering according to Equation 4.27.

Semantic Concept Separation The self-training approach for concept separation is limited to the separation against the background. Here, the background mask with negative latent representations is defined based on the separation threshold $\tau_{\text{separation}}$.

$$\bar{\rho}^{\text{hard}}_{i,j,l} = [\bar{\omega}_{i,j,l} < \tau_{\text{separation}}] \tag{4.38}$$

Finally, $\bar{\rho}^{\text{hard}}_{i,j,l}$ is used to define the loss term for concept separation according to Equation 4.29.

5. Results

The proposed methodology of this work aims at the quantitative evaluation with detection metrics and assurance metrics to generate evidence for the safety argumentation of DNNs for pedestrian detection. The following chapter provides the results of the conducted experiments. By comparing the results of baseline methods and proposed methods for predefined metrics, the improvements in terms of inherent interpretability are quantified. Furthermore, the inherent interpretability, which enforces a structured latent space, offers multiple additional possibilities to generate evidence. Eventually, verifiable evidence supports the safety argumentation of DNNs for pedestrian detection.

Since the structure of the latent space is explicitly enforced by the loss formulation and DNN architecture, the effectiveness of the proposed inherently interpretable DNNs can be evaluated. Thus, experimental results for the cluster quality can show the mitigation of functional insufficiencies and contribute to the safety argumentation.

In the following, the experimental setup of this work is described. The used datasets together with different evaluation and training scenarios are introduced. Analyzing the proposed categorization of ground truth bounding boxes of the Cityscapes dataset [14] and SynPeDS [97] gives an impression of data distribution concerning the distance and visibility of pedestrians. The results of this work are presented in a structured manner with references to individual baselines for every proposed method. In the first step, baseline methods are evaluated to demonstrate state-of-the-art performance. Baseline methods are the DNN for center, scale and offset prediction (CSO) and the DNN for center, scale and offset prediction and segmentation (CSOS).

Although the development of baseline methods can be seen as an individual contribution [29, 28], the focus is on how the proposed methods improve the performance in terms of detection and assurance metrics. Step by step, results for the DNN for prototype-based pedestrian detection (PPD), the DNN for concept-based

pedestrian detection (CPD) and the DNN for concept-based domain adaptation for pedestrian detection (ConDA) are presented.

5.1 Experimental Setup

Since training and validation datasets vary between experiments, different evaluation schemes have to be introduced. Based on that, various benchmarks are proposed to evaluate the proposed methods with the defined metrics.

The description of the training, inference and evaluation details have been already introduced as part of the systematic evaluation and the design of a generic DNN architecture for pedestrian detection. Since results must be comparable, it is essential to keep parameters for DNN training such as the learning rate, batch size or number of epochs consistent. However, performance improvements are expected by optimizing the training configurations, especially for unsupervised domain adaptation.

5.1.1 Datasets

Most commonly, DNNs for pedestrian detection are trained and evaluated on the CityPersons dataset [123]. CityPersons offers additional ground truth 2D bounding boxes for the Cityscapes dataset [14] which originally only provided ground truth for the instance and semantic segmentation of real-world images. In the following, it is always referred to as Cityscapes (CS), including CityPersons. CS consists of 500 images showing 2939 pedestrians.

Since this work focuses on the safety of DNNs for pedestrian detection, more advanced ground truth is extremely valuable for considerations regarding inherent interpretability. SynPeDS [97] is a synthetic dataset specifically created for pedestrian detection with 696 images. In total 11.700 pedestrians are shown.

Table 5.1 shows the distribution of 2D bounding boxes for pedestrians with respect to the distance or depth (foreground, background and far background) and visibility (clearly visible, environmental occlusion, crowd occlusion and ambiguous occlusion). Experiments are conducted for clearly visible pedestrians in the foreground ($\mathcal{G}_{\mathrm{f,v}}$) or background ($\mathcal{G}_{\mathrm{b,v}}$).

Due to the broad selection of experiments for the proposed methods, a notation for different evaluation schemes is introduced. The notation establishes the connection between the training dataset and the validation dataset for evaluation. Note that the notation only refers to the generally used dataset. Meanwhile, the

Table 5.1: Dataset statistics for Cityscapes (CS) [14, 123] and SynPEDS [97]. Pedestrians are grouped into different distance categories and their visibility is divided into four categories: clearly visible $\mathcal{G}_\mathrm{v}$, environmental occlusion $\mathcal{G}_\mathrm{e}$, crowd occlusion $\mathcal{G}_\mathrm{c}$ and ambiguous occlusion $\mathcal{G}_\mathrm{a}$.

Dataset	Distance category	Pedestrians	Visibility			
			$\mathcal{G}_\mathrm{v}$	$\mathcal{G}_\mathrm{e}$	$\mathcal{G}_\mathrm{c}$	$\mathcal{G}_\mathrm{a}$
CS	Foreground ($\mathcal{G}_\mathrm{f}$)	656	**416**	97	91	52
	Background ($\mathcal{G}_\mathrm{b}$)	1148	**662**	286	115	85
	Far background	1135	680	346	63	46
SynPeDS	Foreground ($\mathcal{G}_\mathrm{f}$)	2639	**2083**	238	273	45
	Background ($\mathcal{G}_\mathrm{b}$)	3589	**2207**	609	443	330
	Far background	5472	2618	1338	628	888

training and validation dataset are non-overlapping. In this work, the following evaluation schemes are of interest:

- CS $\rightarrow$ CS: Supervised training with Cityscapes (CS) training data and evaluation with the CS validation dataset.
- SynPeDS $\rightarrow$ SynPeDS: Supervised training with SynPeDS training data and evaluation with the SynPeDS validation dataset.
- SynPeDS $\rightarrow$ CS: Unsupervised domain adaptation (UDA), i.e., supervised training on SynPeDS training data combined with unsupervised training on CS training data and (online) evaluation with CS validation data.

If the training and validation dataset is not explicitly mentioned in the notation, the same dataset, but not the same subset, is used for training and validation.

5.1.2 Benchmarking

In the course of this work, various methods for different datasets have been developed, each requiring individual baseline methods. Due to the increased complexity of benchmarking, a short overview of all conducted experiments with different datasets for baseline methods and proposed methods is shown in Figure 5.1. Furthermore, different combinations of training and validation data are illustrated.

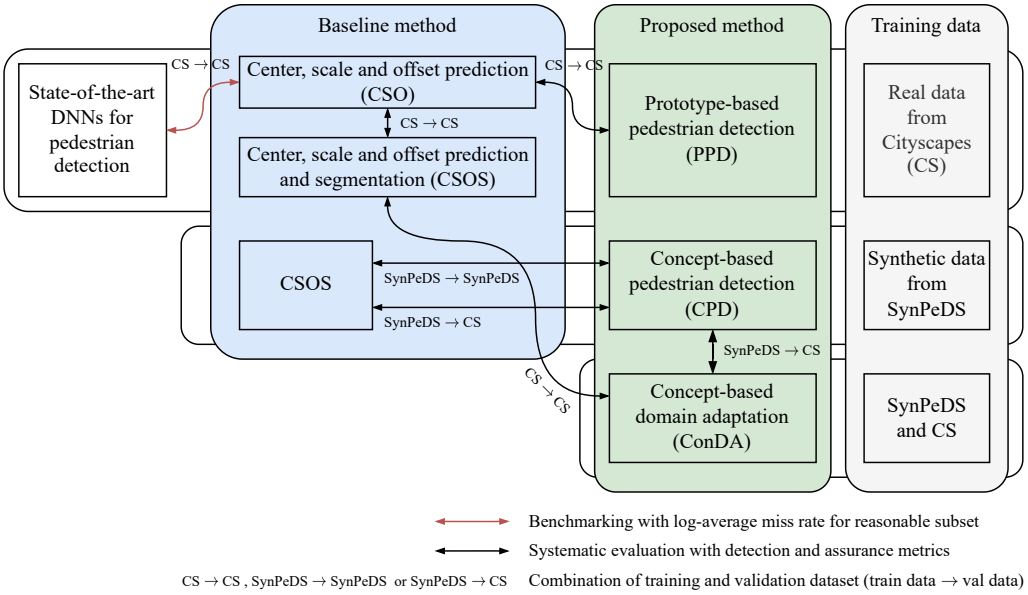


Figure 5.1: Benchmarking of baseline methods and proposed methods with links that describe the used data for training and validation. The illustration outlines the structure of the performed experiments and how the results are compared.

Linking proposed methods to baseline methods highlights how experimental results are used for meaningful benchmarking. Every proposed method must be connected to a baseline method to quantify improvements in performance with detection metrics or inherent interpretability with assurance metrics.

In contrast to related work on pedestrian detection [54, 61, 119, 70], this work is not focused on benchmarking with the log-average miss rate [20, 5] for the reasonable subset of the CityPersons [123] validation dataset. Instead, benchmarks for the introduced detection and assurance metrics of baseline methods and proposed methods are compared. However, to demonstrate the relevance of the proposed methods and the validity of the experimental results, the baseline DNN for center, scale and offset prediction (CSO) is compared to state-of-the-art benchmarks.

5.2 Baseline Methods

The development and evaluation of baseline methods enable valid benchmarks. Baseline methods are the DNN for center, scale and offset prediction (CSO) and the

DNN for center, scale and offset prediction and segmentation (CSOS). Benchmarks are used to argue the advantages of inherently interpretable DNNs for pedestrian detection in order to support the safety argumentation. Initially, it must be shown that the baseline methods, CSO and CSOS, achieve state-of-the-art performance.

5.2.1 Center, Scale and Offset Prediction

The baseline method for pedestrian detection with 2D bounding boxes is given by the DNN for center, scale and offset prediction (CSO). It has a simple anchor-free and one-stage DNN architecture. Figure 5.2 gives a qualitative impression with inference results on Cityscapes. CSO predicts 2D bounding boxes for pedestrians shown in an input image.

Figure 5.2: Inference results, including detection bounding boxes (white) for the DNN for center, scale and offset prediction (CSO, right column). The left column shows the original images from Cityscapes and the middle column adds the ground truth bounding boxes (blue). Gray bounding boxes are ignored. The inference is conducted with a confidence threshold of 0.4.

State-of-the-Art Comparison In the first evaluation step, it is shown that CSO achieves comparable performance to state-of-the-art DNNs for pedestrian detection. Table 5.2 depicts the qualitative performance results for the subset-based evaluation[1] with respect to the log-average miss rate ($LAMR$) for the CityPersons validation dataset. Although results for multiple subsets are reported, the state-of-the-art performance depends on the LAMR for the reasonable subset ($LAMR_r$).

It can be shown that CSO with HRNet [98] achieves state-of-the-art performance with 8.8% $LAMR$ compared to the best performance of 8.7% with F2DNet [54]. The results indicate that the performance strongly depends on the applied

[1] reasonable $(h = [50, 1024], v = [0.65, \infty))$, bare $(h = [50, 1024], v = [0.90, \infty))$, partial $(h = [50, 1024], v = [0.65, 0.90])$, heavy $(h = [50, 1024], v = [0.00, 0.65])$ [123, 107].

Table 5.2: Performance results for the log-average miss rate $LAMR$ [%] of state-of-the-art DNNs for pedestrian detection for different subsets of the Cityscapes validation dataset: reasonable (R), bare (B), partial (P) and heavy (H). Different feature extractions (FPN-ResNet-50, CSP-ResNet-50, MDLA and BGC-HRNet) with varying backbones (ResNet-50, DLA and HRNet) have been investigated for the DNN for center, scale and offset prediction (CSO).

Method	Backbone	R	B	P	H
ALFNet [69]	ResNet-50 [46]	12.0	8.4	11.4	51.9
CSP [70]	ResNet-50	11.0	7.3	10.4	49.3
NOH-NMS [124]	ResNet-50	10.8	6.6	11.2	53.0
RepLoss [107]	ResNet-50	10.9	6.3	13.4	52.9
PRNet [95]	ResNet-50	10.8	6.8	10.0	53.3
Beta R-CNN [116]	ResNet-50	10.6	6.4	10.3	47.1
APD [119]	ResNet-50	10.6	5.8	8.3	46.6
NMS-Loss [74]	ResNet-50	10.1	-	-	-
BGCNet [61]	HRNet [98]	8.8	6.1	8.0	43.9
APD [119]	DLA [118, 16]	8.8	5.8	8.3	46.6
F2DNet [54]	HRNet	8.7	-	-	32.6
CSO [29]	ResNet-50 (FPN-ResNet-50)	10.9	7.5	10.0	48.3
	ResNet-50 (CSP-ResNet-50)	10.6	7.5	9.6	45.6
	DLA (MDLA)	9.6	6.5	8.5	45.2
	HRNet (BGC-HRNet)	8.8	6.2	7.4	41.6

backbone and feature extraction. HRNet [98] has substantially more parameters than the other backbones such as ResNet-50 [46] and MDLA [16, 118]. Meanwhile, CSO represents a rather simple DNN architecture with a straightforward one-stage and anchor-free approach for pedestrian detection. In contrast to F2DNet [54], CSO is a more eligible baseline method since no specially customized components such as the focal detection network and fast suppression head are introduced.

Systematic Evaluation After demonstrating the state-of-the-art performance, CSO is evaluated with the detection metrics of the systematic evaluation: (1) filtered log-average miss rate ($FLAMR$) and (2) filtered log-average miss rate with respect to ghost detections ($FLAMR_g$).

Table 5.3 shows how CSO variants with different feature extractions from Table 5.2 perform in terms of the $FLAMR_g$ and $FLAMR$ for clearly visible pedestrians

Table 5.3: Post-hoc systematic evaluation [%] for Cityscapes of the DNN for center, scale and offset prediction (CSO) with different feature extractions.

Feature extraction	$LAMR_r$	$FLAMR_g$		$FLAMR$	
		$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$
FPN-ResNet-50	10.9	1.56	3.18	3.53	6.13
CSP-ResNet-50	10.6	2.30	2.97	4.57	4.99
MDLA	9.6	2.61	2.87	3.75	4.10
HRNet	8.8	0.13	2.57	1.49	3.89

Table 5.4: Systematic evaluation [%] for Cityscapes of the DNN for center, scale and offset prediction (CSO).

Feature extraction	$FLAMR_g$		$FLAMR$	
	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$
CSP-ResNet-50	1.47	2.59	3.74	5.14
MDLA	1.16	3.07	2.18	5.07
FPN-ResNet-50	0.82	3.79	3.93	7.25
HRNet	0.08	2.71	1.00	4.93

in the foreground ($\mathcal{G}_{\mathrm{f,v}}$) and background ($\mathcal{G}_{\mathrm{b,v}}$). The results of the systematic evaluation with $FLAMR_g$ for $\mathcal{G}_{\mathrm{f,v}}$ contradict the performance measured by the log-average miss rate ($LAMR$) for the reasonable subset. Consequently, both metrics measure different aspects of the detection performance. In contrast to the reasonable subset, the categorization of ground truth bounding boxes as part of the systematical evaluation is transparent and conclusive from the perspective of an automated driving system. This effect has been previously shown by [29].

It is important to note that this discrepancy can be resolved if the DNNs are retrained and the online evaluation is conducted with the novel detection metrics ($FLAMR_g$, $FLAMR$). Online is referring to the ongoing evaluation during training, conducted after every training epoch. Retraining has a massive performance impact, indicating that the design and purpose of the metrics differ. After repeating the experiments, results are given in Table 5.4, it can be shown that retrained DNNs have the potential to perform substantially better than shown in Table 5.3. Finally, CSO with HRNet achieves the best performance in terms of the $FLAMR_g$ for clearly visible pedestrians in the foreground ($\mathcal{G}_{\mathrm{f,v}}$) with 0.08%.

5.2.2 Center, Scale and Offset Prediction and Segmentation

In the following, the DNN for center, scale and offset prediction (CSO) has been extended with an auxiliary body part segmentation. The method is called DNN for center, scale and offset prediction and segmentation (CSOS) accordingly. CSOS ties the prediction of 2D bounding boxes for pedestrians to the segmentation of body parts. Figure 5.3 gives some inference results for the Cityscapes dataset.

Since this work also focuses on the generalization of DNNs from synthetic to real data, a second instance of CSOS has been trained and evaluated on synthetic data from SynPeDS. To give also an impression on the used synthetic data, inference results are shown in Figure 5.4.

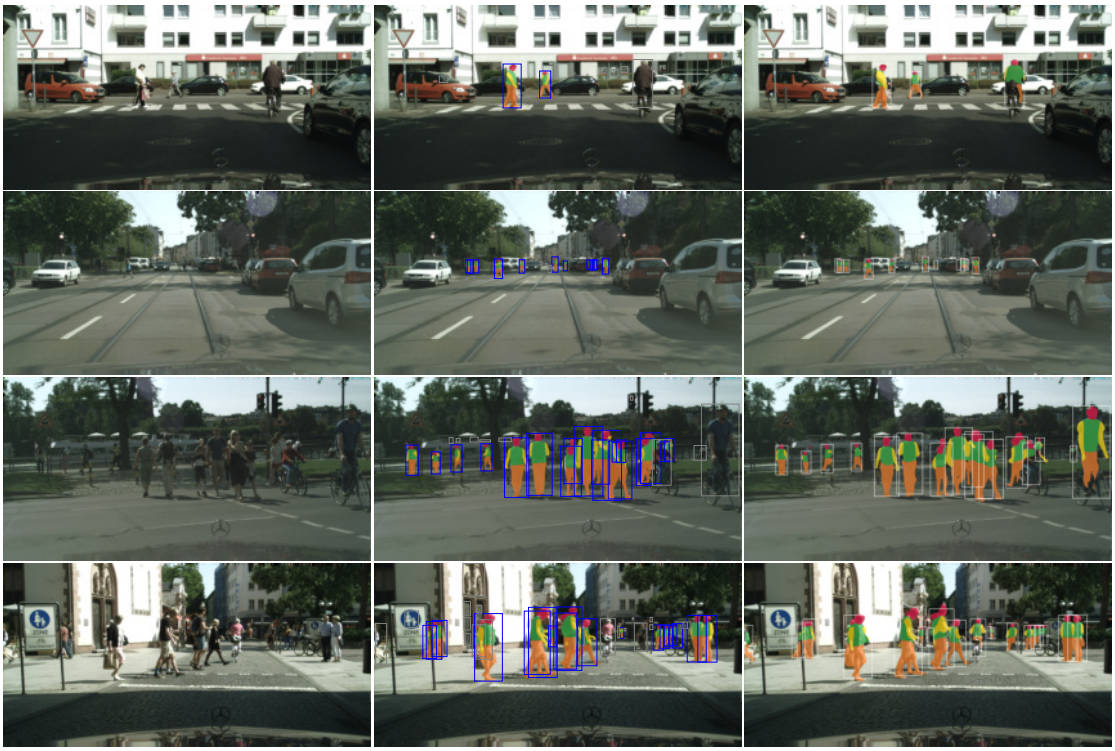


Figure 5.3: Inference results, including detection bounding boxes (white) for the DNN for center, scale and offset prediction and body part segmentation (CSOS, right column). The left column shows the original images from Cityscapes and the middle column adds the ground truth bounding boxes (blue) and body part segmentation. Gray bounding boxes are ignored. The inference is conducted with a confidence threshold of 0.4.

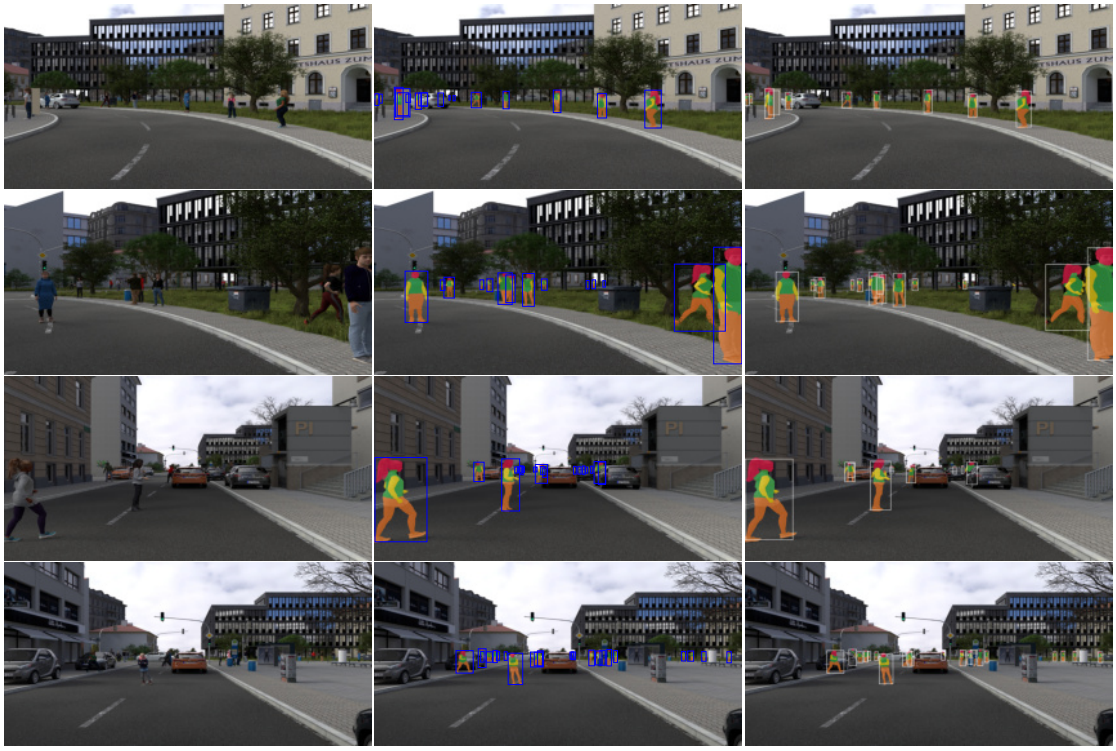


Figure 5.4: Inference results, including detection bounding boxes (white) for the DNN for center, scale and offset prediction and body part segmentation (CSOS, right column). The left column shows the original images from SynPeDS and the middle column adds the ground truth bounding boxes (blue) and body part segmentation. The inference is conducted with a confidence threshold of 0.4.

Table 5.5: Systematic evaluation [%] for Cityscapes of DNN for center, scale and offset prediction and segmentation (CSOS).

Method	Feature extraction	$FLAMR_g$		$FLAMR$	
		$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$
CSO	FPN-ResNet-50	0.82	3.79	3.93	7.25
	CSP-ResNet-50	1.47	2.59	3.74	5.14
	MDLA	1.16	3.07	2.18	5.07
	HRNet	0.08	2.71	1.00	4.93
CSOS	FPN-ResNet-50	2.64	4.61	8.87	13.25
	CSP-ResNet-50	0.31	3.07	5.02	6.11
	MDLA	0.24	3.58	3.03	5.70
	HRNet	0.07	2.66	2.65	5.10

Systematic Evaluation CSOS with an auxiliary body part segmentation achieves superior performance compared to CSO in terms of the detection metrics ($FLAMR_g$, $FLAMR$) for clearly visible pedestrians in the foreground ($\mathcal{G}_{\mathrm{f,v}}$).

As shown in Table 5.5 CSOS generally reduces the number of ghost detections, thus, decreasing $FLAMR_g$ (for $\mathcal{G}_{\mathrm{f,v}}$). However, the number of scale and localization errors increases. This effect is shown by the increase of the $FLAMR$ ($\mathcal{G}_{\mathrm{f,v}}$ as well as $\mathcal{G}_{\mathrm{b,v}}$). In the case of MDLA, the performance improves by absolutely 0.92% and for HRNet by 0.01%. Only FPN-ResNet-50 fails when auxiliary segmentation tasks are solved simultaneously. Whilst improving performance compared to CSO, CSOS also makes 2D detection bounding boxes more plausible. Simple decision rules check if at least one body part is detected within all detection bounding boxes.

It is worth mentioning that the performance for $FLAMR_g$ with $\mathcal{G}_{\mathrm{f,v}}$ of CSOS with HRNet decreases substantially when trained and validated on SynPeDS. In contrast to Table 5.6, MDLA performs best. Presumably, HRNet tends to focus more on pedestrians in the background.

Segmentation Quality In the following the segmentation quality of CSOS for supervised training and evaluation on real data from Cityscapes in Table 5.7 is evaluated. The filtered mean intersection over union ($FmIoU$) measures the segmentation quality of clearly visible pedestrians in the foreground $\mathcal{G}_{\mathrm{f,v}}$ and background $\mathcal{G}_{\mathrm{b,v}}$.

Furthermore, experimental results for synthetic data from SynPeDS are pro-

Table 5.6: Systematic evaluation [%] for SynPeDS of DNN for center, scale and offset prediction and segmentation (CSOS).

Feature extraction	$FLAMR_g$		$FLAMR$	
	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$
FPN-ResNet-50	7.49	6.24	12.22	12.02
CSP-ResNet-50	4.73	2.81	6.87	4.29
MDLA	2.71	1.95	3.76	3.29
HRNet	3.24	1.79	4.69	3.22

Table 5.7: Segmentation quality [%] for Cityscapes of DNN for center, scale and offset prediction and segmentation (CSOS). Detailed results are provided for all four body parts in the foreground.

Feature extraction	$FmIoU$		$\mathcal{G}_{\mathrm{f,v}}$			
	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$	Head	Torso	Arm	Leg
FPN-ResNet-50	65.07	44.80	66.74	54.90	40.14	63.76
CSP-ResNet-50	72.80	55.32	73.35	64.29	55.43	71.08
MDLA	74.94	55.87	74.49	66.26	59.59	74.49
HRNet	76.54	58.76	77.33	67.41	61.62	76.45

Table 5.8: Segmentation quality [%] for SynPeDS of DNN for center, scale and offset prediction and segmentation (CSOS). Detailed results are provided for all four body parts in the foreground.

Feature extraction	$FmIoU$		$\mathcal{G}_{\mathrm{f,v}}$			
	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$	Head	Torso	Arm	Leg
FPN-ResNet-50	79.94	50.98	80.46	74.74	61.76	82.95
CSP-ResNet-50	88.04	63.62	86.01	85.85	78.60	89.88
MDLA	90.00	67.18	87.85	87.96	82.82	91.49
HRNet	90.35	68.78	87.92	88.25	83.82	91.85

vided in Table 5.8. Clearly, HRNet performs best in terms of the filtered mean intersection over union ($FmIoU$) for the clearly visible foreground as well as the background. Contrarily to the detection metrics, this is the case for the real-world and synthetic datasets. Generally, the performance is substantially higher than on real-world data in Table 5.7.

5.3 Prototype-based Pedestrian Detection

The DNN for prototype-based pedestrian detection (PPD) is evaluated as the first proposed method in this work. PPD utilizes unsupervised prototypes as part of interpretable reasoning to detect pedestrians with 2D bounding boxes. Hence, PPD can be seen as a proof of concept for interpretable DNNs for pedestrian detection. Due to the promising results, further methods such as the DNN for concept-based pedestrian detection (CPD) and DNN for concept-based domain adaptation for pedestrian detection (ConDA) are developed.

In the following evaluation, PPD is compared to CSO as its baseline method since both methods predict 2D bounding boxes. CSO is considered a baseline method with state-of-the-art performance that predicts 2D bounding boxes for pedestrian detection.

Figure 5.5 shows some inference results for the validation dataset of Cityscapes.

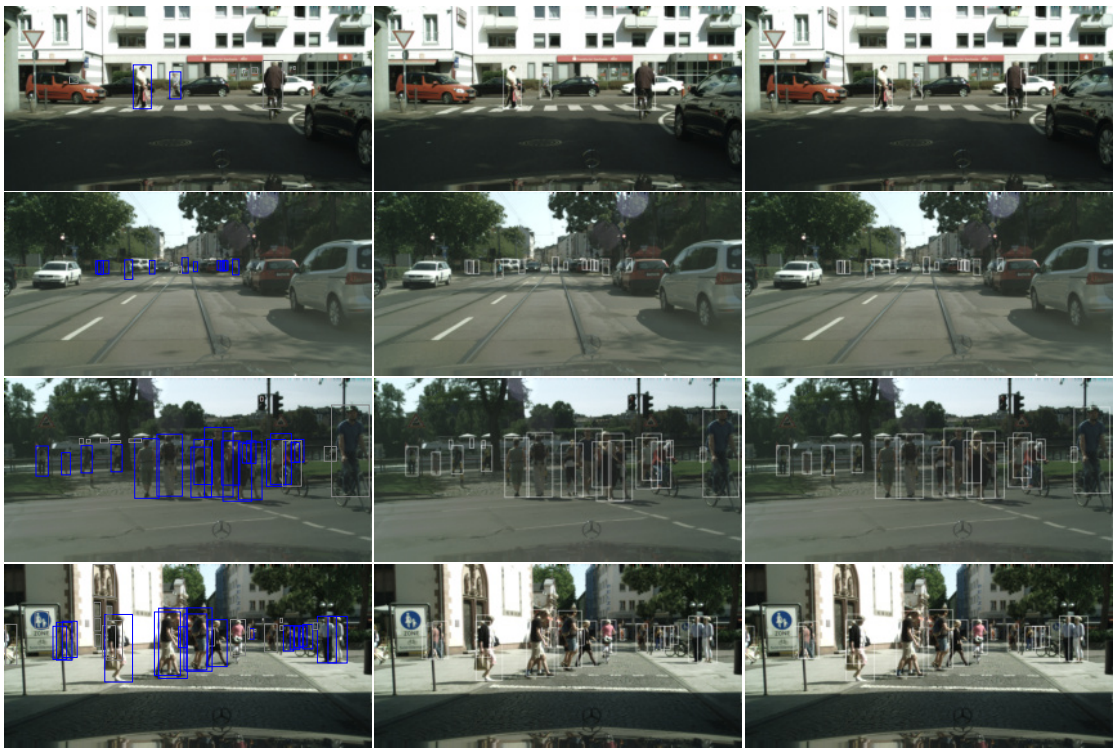


Figure 5.5: Inference results, including detection bounding boxes (white) for the DNN for prototype-pedestrian detection (PPD, right column) and the DNN for center, scale and offset prediction (CSO, middle column). The left column shows the original images from Cityscapes with ground truth bounding boxes (blue) and ignored bounding boxes (red). The inference is conducted with a confidence threshold of 0.4.

Systematic Evaluation Table 5.9 gives an overview of the systematic performance evaluation of PPD with respect to the proposed detection metrics. Since PPD is a proof of concept, only MDLA is investigated for feature extraction. MDLA offers the best trade-off between training time and performance. The results demonstrate that the adapted DNN architecture and loss formulation do not have a negative impact on the detection performance for clearly visible pedestrians in the foreground ($\mathcal{G}_{\mathrm{f,v}}$). Instead, PPD achieves a superior 0.16% $FLAMR_g$ compared to CSO. Despite the decreasing performance in the background ($\mathcal{G}_{\mathrm{fb}}$), the foreground ($\mathcal{G}_{\mathrm{f,v}}$) performance substantially improves.

Table 5.9: Systematic evaluation [%] for Cityscapes of the DNN for prototype-based pedestrian detection (PPD) with MDLA.

Method	$FLAMR_g$		$FLAMR$	
	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$
CSO	1.16	3.07	2.18	5.07
PPD	0.16	3.66	0.92	6.89

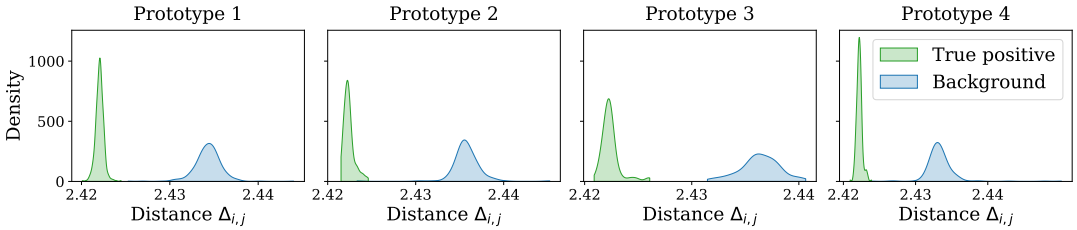


Figure 5.6: Relation of distances between prototypes and latent representations as density plots. Distances of latent representations which are true positives and background are shown.

Structured Latent Space Besides improving the performance, PPD structures the latent space by clustering and separating latent representations. Although the DNN architecture is adapted accordingly and introduces prototypes, the two opposing mechanisms are purely a result of additional loss terms.

Figure 5.6 shows the distribution of latent representations that are either true positives, i.e., pedestrian center, or true negatives, i.e., background pixels. It is important to point out that background refers in this case simply to pixels that are false negatives. Thus, the background describes pixels that are outside of ground truth bounding boxes. It is important to note that 'background' does not refer to pedestrians standing in the background.

It can be shown that PPD is capable to distinguish between centers of pedestrians and background based on distances. To prove this capability, distances of latent representations from the background are randomly sampled from the latent tensor. The clear separation of distributions in Figure 5.6 indicates that the latent space is indeed structured.

Limitations Although the reasoning becomes interpretable since distances in the latent space are utilized, the decision-making process behind the prototype as-

Figure 5.7: Top-5 nearest latent representations of pedestrian centers (rows) for four trained prototypes (columns).

signment remains hidden. Figure 5.7 provides a visual inspection of the prototype assignment. For each of the four prototypes, the top-five latent representations with the assigned pedestrians are shown. Intuitively, individual characteristics of prototypes are hard to distinguish. Nonetheless, PPD is able to identify discriminative features and cluster latent representations. The results are in line with work that focuses on the interpretability of prototypes [94].

5.4 Concept-based Pedestrian Detection

To overcome the limitations of the previously evaluated prototype-based pedestrian detection, the proposed DNN for concept-based pedestrian detection (CPD) predicts 2D bounding boxes and an auxiliary body part segmentation. In contrast to its baseline method, the DNN for center, scale and offset prediction and segmentation (CSOS), CPD's segmentation relies on interpretable reasoning based on high intra-class concentration and inter-class separation of semantic concepts such as human body parts. In contrast to PPD, semantic concepts of a human are

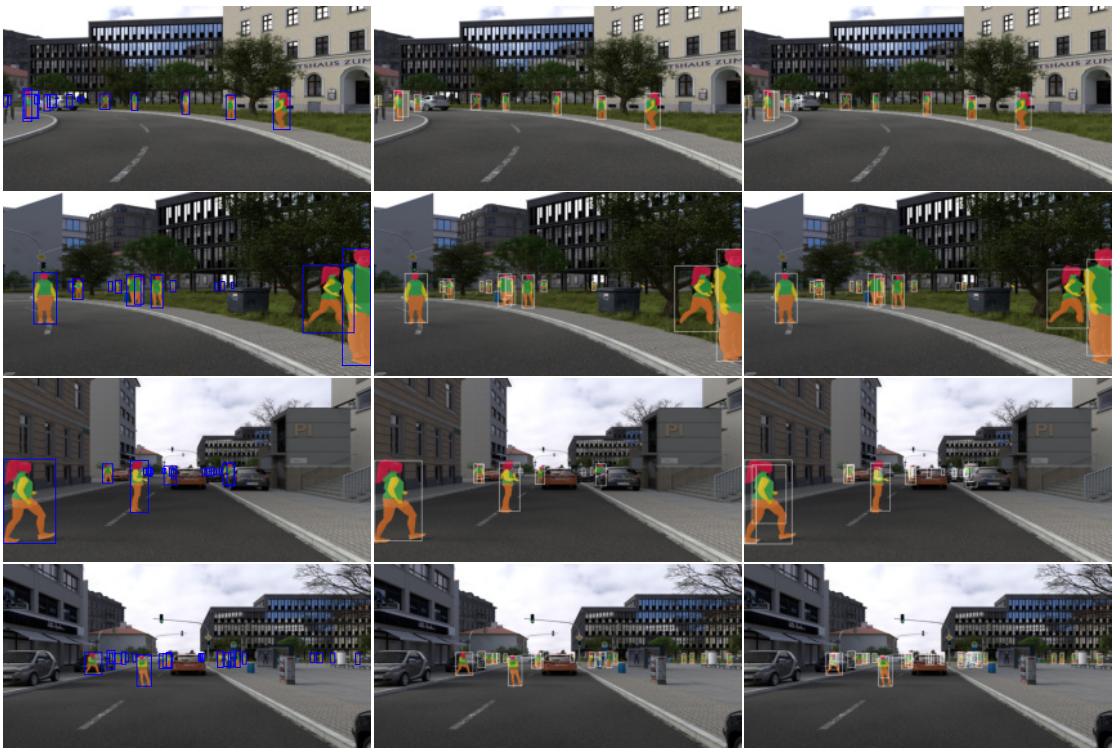


Figure 5.8: Inference results, including detection bounding boxes (white) for the DNN for concept-pedestrian detection (CPD, right column) and the DNN for center, scale and offset prediction and segmentation (CSOS, middle column). The left column shows the original images from SynPeDS with ground truth bounding boxes (blue) and the body part segmentation. The inference is conducted with a confidence threshold of 0.4.

predefined and known due to labeled ground truth. Hence, the clustered latent representations can easily be assigned to a semantic concept and visually inspected and consequently understood by humans. Figure 5.8 shows inference results for CPD trained on the synthetic SynPeDS dataset.

Systematic Evaluation In the first step, the application-oriented systematic evaluation with the proposed detection metrics is conducted. Table 5.10 provides the experimental results in comparison to the baseline method CSOS. CPD achieves better performance results for safety-critical and clearly visible pedestrians in the foreground ($\mathcal{G}_{\mathrm{f,v}}$) and background ($\mathcal{G}_{\mathrm{b,v}}$) in terms of all proposed metrics

Table 5.10: Systematic evaluation [%] for SynPeDS of DNN for concept-based pedestrian detection (CPD).

Method	Feature extraction	$FLAMR_g$		$FLAMR$	
		$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$
CSOS	MDLA	2.71	1.95	3.76	3.29
	HRNet	3.24	1.79	4.69	3.22
CPD	MDLA	2.44	1.69	3.45	3.18
	HRNet	3.38	1.92	4.68	3.24

($FLAMR_g$, $FLAMR$).

Meanwhile, CPD enables inherent interpretability. It can be clearly shown that inherent interpretability not necessarily reduces the DNN performance. This assumption has been formerly postulated by the explainable artificial engineering announcement of the Defense Advanced Research Projects Agency [44] and is since criticized by Rudin [88]. The feature extraction with MDLA provides the best performance ($FLAMR_g$ for $\mathcal{G}_{\mathrm{f,v}}$) for CSOS as well as CPD. CPD outperforms CSOS with 2.44% $FLAMR_g$ compared to 2.71%.

The detection metric $FLAMR_g$ provides a dense performance metric that averages multiple miss rate (MR) values for predefined maximum numbers of false positives per image ($FPPI$). This allows comparing the performance, without determining the exact confidence threshold of the DNN. However, the applied confidence threshold has a substantial impact on the trade-off between the miss rate for a certain subset and the accepted amount of false positives.

Figure 5.9 plots the MR for different subsets over $FPPI$ and ghost detections per image ($GDPI$). The data points are sorted in descending order of the confidence scores for every detection. Detections resemble steps, marking either a true positive (step down) or a false positive (step right). Hence, the miss rate decreases by lowering the confidence threshold but also increases the $FPPI$.

In order to evaluate the actual detection performance of CSOS and CPD, Table 5.11 defines fixed values for false positives, i.e. 1 $FPPI$ and 0.1 $GDPI$. Note that the values are chosen arbitrarily, although related work suggested 1 $FPPI$ [20].

The analysis shows that CPD achieves the lowest miss rate (MR) whilst accepting 0.1 ghost detections per image ($GDPI$). CSOS is outperformed by an absolute margin of 0.20%. However, considering all possible false positives per image ($FPPI$) and accepting only 1 $FPPI$, CSOS offers better performance. The

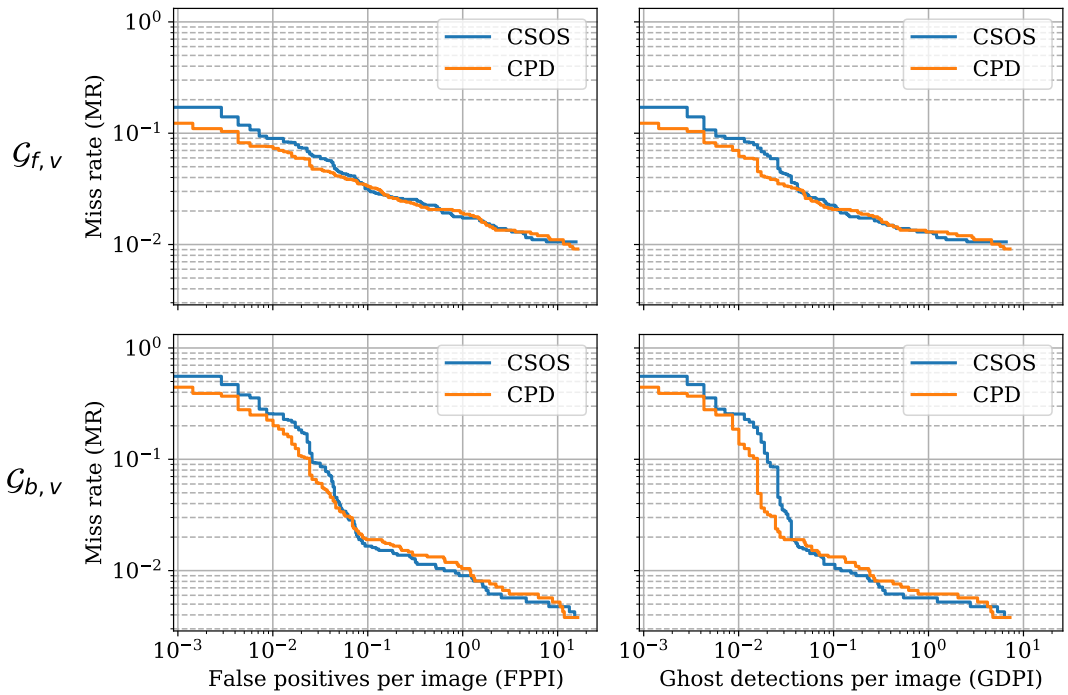


Figure 5.9: Relation of miss rate to false positives of DNN for concept-based pedestrian detection (CPD) for SynPeDS.

Table 5.11: Miss rate [%] analysis of the DNN for concept-based pedestrian detection (CPD) for SynPeds.

Method	$MR@0.1GDPI$		$MR@1FPPI$	
	$\mathcal{G}_{\text{f,v}}$	$\mathcal{G}_{\text{b,v}}$	$\mathcal{G}_{\text{f,v}}$	$\mathcal{G}_{\text{b,v}}$
CSOS	2.26	1.14	1.73	0.90
CPD	2.06	1.33	1.92	1.38

experimental results show how important it is to accurately define application-oriented detection metrics that are used for the online evaluation. Ultimately, the detection metrics define what checkpoint of a training run is chosen.

Segmentation Quality Revisiting the proposed methodology of this work - creating evidence for a safety argumentation based on metrics relies on detection

Table 5.12: Segmentation quality [%] for SynPeDS of DNN for concept-based pedestrian detection (CPD). Detailed results are provided for all four body parts in the foreground.

Method	Feature extraction	$FmIoU$		$\mathcal{G}_{\mathrm{f,v}}$			
		$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$	Head	Torso	Arm	Leg
CSOS	MDLA	90.00	67.18	87.85	87.96	82.82	91.49
	HRNet	90.35	68.78	87.92	88.25	83.82	91.85
CPD	MDLA	90.96	67.68	87.86	90.08	85.12	91.84
	HRNet	89.78	64.31	86.46	88.91	82.83	90.85

and assurance metrics. Detection metrics have been formerly addressed and results are intensively studied. The aim of assurance metrics is to measure the effects of establishing inherent interpretability.

The segmentation quality of CSOS and CPD is presented in Table 5.12. CSOS and CPD have comparable performance with either neglectable advantages for the foreground ($\mathcal{G}_{\mathrm{f,v}}$) or background ($\mathcal{G}_{\mathrm{b,v}}$). It can be shown, that the proposed additional loss terms enforcing dense concept clusters and the separation of background representations have no negative effect on the segmentation quality of body parts. Please note, that the number of concepts is generally not limited to only four body parts (head, torso, arm and leg). As shown in [26], the set of semantic concepts can also include hands or feet.

Structured Latent Space Figure 5.10 shows density plots for the calculated distances between concept vectors and latent representations for either background or ground truth body parts. Here, the label ground truth includes every latent representation for a body part despite the actual prediction, thus, false negatives are possibly included. Similar to the PPD analysis of the structured latent space, note that 'background' does not refer to pedestrians standing in the background. Instead, 'background' refers to pixels in an image that do not belong to pedestrians. Again, the latent representations for the background are randomly sampled.

Since CSOS does not have concept vectors, cluster centroids are calculated post-hoc. That means, for every sample ground truth annotations are used to extract a set of positive latent representations and build the center vector simply by averaging. Note that this only enforces a critical form of post-hoc interpretability, if any. Latent representations with small distances are more likely to predict the

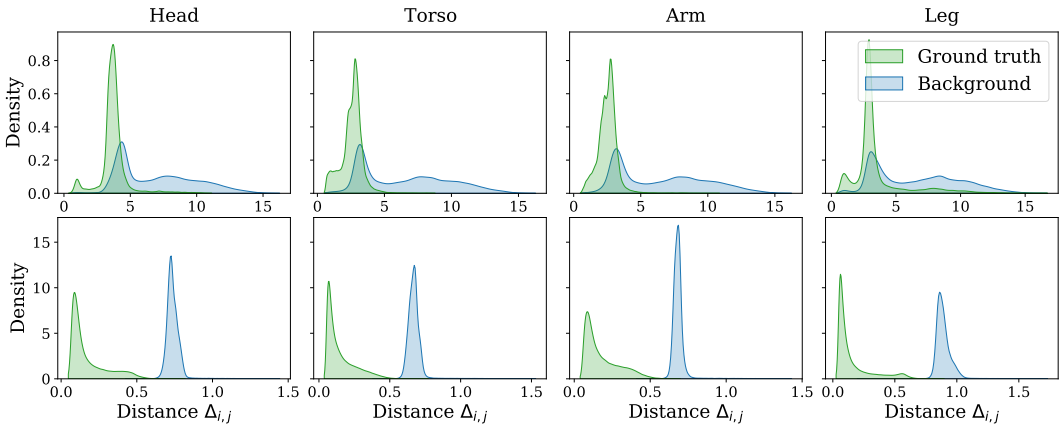


Figure 5.10: Relation of distances between concept vectors and latent representations as density plots for SynPeDS. Two methods are investigated: DNN for center, scale and offset prediction and segmentation (CSOS, top row) and the proposed DNN for concept-based pedestrian detection (CPD, bottom row).

same class as the cluster centroids, but there are no mechanisms that enforce this behavior.

As it can be shown, CSOS (first row) has no interpretable structure of the learned latent space. Distances of ground truth samples for every body part and randomly sampled distances of background latent representations are highly overlapping. Contrarily, CPD is able to significantly separate the distributions for every semantic concept. These findings are supported by the initial investigation of PPD. The results for CPD clearly accentuate the high intra-class concentration of semantic concepts and inter-class separation.

Cluster Quality Since CPD introduces interpretable reasoning based on the squared L^2 distance various metrics such as the silhouette coefficient (SC) [86], the Calinski-Harabasz index (CHI) [8] and the Davies–Bouldin index (DBI) [17] can be applied to estimate the cluster quality. Due to the extensive computational cost of SC calculation, random sub-sampling is applied and the mean and standard deviation of SC are presented.

Table 5.13 shows results for different cluster performance metrics including the proposed metric for mean concept quality (mCQ). The results demonstrate that CPD substantially achieves better performance in structuring the latent space

Table 5.13: Cluster quality for SynPeDS of the DNN for concept-based pedestrian detection (CPD).

Method	SC		$CHI(\times 10^6)$	DBI	$mCQ_{0.99}$	$mCQ_{0.95}$	$mCQ_{0.50}$
	mean	std					
CSOS	0.10	0.01	0.30	1.61	-0.19	-0.03	0.52
CPD	0.29	0.01	0.70	1.60	0.91	0.91	0.85

Table 5.14: Detailed cluster quality for SynPeDS of the DNN for concept-based pedestrian detection (CPD).

Method	$mCQ_{0.50}$	Head	Torso	Arm	Leg
CSOS	0.52	0.43	0.56	0.58	0.53
CPD	0.85	0.85	0.85	0.82	0.90

compared to the baseline method CSOS.

According to the chosen percentile (99%, 95% and 50%) of distances, mCQ becomes more robust to occurring annotation errors of ground truth SynPeDS data. Here, distances of latent representations to ground truth body parts and background are considered.

Although SC and mCQ are closely related, mCQ gives a better view of the concept cluster performance since it incorporates the centroidal integration of concept vectors. Note that both metrics are bound between $[-1; 1]$, where 1 determines perfect separation.

CHI and DBI can only be considered relative indicators. Generally, a higher CHI score means better cluster separation. Contrarily, a lower DBI score indicates better partitioning. The results confirm the demonstrated results with SC and mCQ.

In conclusion, CSOS follows a completely different way to structure the latent space due to the utilization of the softmax loss. Just the fact is not problematic, but in terms of the safety argumentation, it is not possible to define a set of quantifiable evidence based on cluster performance metrics. The latent space structure of CSOS remains intransparent. Contrarily, the DNN architecture and loss formulation of CPD are designed to fill that gap by design and to enable the collection of evidence.

Table 5.14 gives a more detailed overview of the individual performance of every predicted body part. It can be seen that all parts achieve nearly the same cluster quality and no semantic concept is disregarded during training.

Undoubtedly, the cluster quality of non-interpretable baseline methods could be technically analyzed in the same way. But the implications for a safety argumentation would be completely different. Regarding the interpretable methods, an assumption is formulated stating the existence of a structured latent space based on specifically developed, thus, known mechanisms for clustering and separation within the latent space. Experimental results empirically indicate that the assumption is indeed true. This form of argumentation is not applicable to non-interpretable methods since the mechanisms for structuring the latent space are unknown. Instead, the relation between mechanisms and structure can only be guessed.

Hypothesis Testing In contrast to considering latent representations for ground truth body parts, the distances between true positive latent representations and concept vectors are more similar to a normal distribution. This allows the application of a two-sided t-test of dependent samples. Latent representations are seen as true positives if they are similar enough to a concept vector to predict the correct body part. Intuitively, the distance distribution has a right tail due to the distance-based training approach of CPD. Focusing on true positives and their relation to concept vectors enriches the evaluation. Solely relying on performance or detection metrics shifts to generating verifiable evidence for a safety argumentation.

False positives especially occur on the fine border between different body parts or background. Naturally, it is hard to distinguish between concepts on a pixel basis. Distances and the amount of false positives increase till the critical distance is reached.

The goal is to provide evidence that the distance distributions of true positives and background are well separated. Figure 5.11 shows density plots of distances for every respective concept vector and matched latent representation in the SynPeDS validation dataset.

The paired t-test is applied to test whether the mean distances for concepts and background are significantly different. The null hypothesis is defined as

$$H_0 : \frac{\bar{x}_l}{\sigma_l / \sqrt{N_l}} = 0 \tag{5.1}$$

and is tested for every l-th concept.

Due to the high amount of available background samples, we randomly sample N_l distances of latent representations for background. To guarantee equal sample

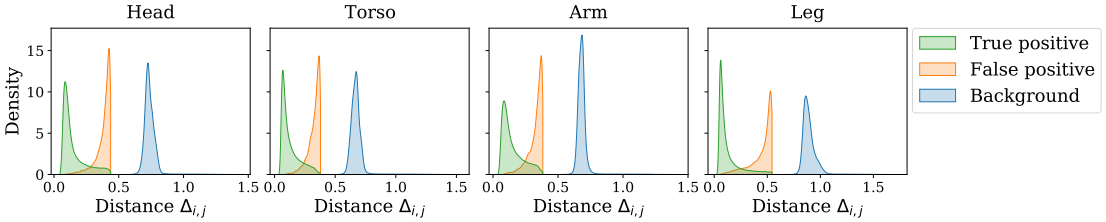


Figure 5.11: Fine-grained analysis of distances between concept vectors and latent representations as density plots for SynPeDS.

sizes for every concept, N_l equals the number of distances of latent representations for the respective body part.

The t-statistic is calculated with the average distance deviation given by

$$\bar{\Delta}_i = \frac{1}{N_l} \sum_{i=1}^{N_l} \left(\Delta_i^{TP} - \Delta_i^{B} \right) \tag{5.2}$$

where the standard deviation of differences is

$$\sigma_l = \sqrt{\frac{1}{N_l} \sum_{i=1}^{N_l} \left(\Delta_i - \bar{\Delta}_i \right)^2} \tag{5.3}$$

As a result, the null hypothesis H_0 can be rejected for all concepts and is statistically significant at a significance level of 0.01. Hence, CPD effectively separates latent representations by leveraging semantic concept vectors.

Concept Testing The formerly introduced analysis techniques targeting distances in the latent space can be extended to quantify the robustness of CPD. Here, natural perturbations such as varying lighting conditions and other image characteristics are exemplarily shown in Figure 5.12.

Although a set of natural perturbations are already considered in the data augmentation strategy of the DNN training, the increase in robustness remains unknown. Based on the concept quality, deviations compared to results presented in Table 5.13 for certain perturbations can be used to quantify the robustness.

The results in Table 5.15 indicate that CPD on average is fairly robust against standard perturbations. However, hue variations have the highest negative impact and decrease the concept quality by 0.11. Nonetheless, the final concept quality is

Figure 5.12: Samples of natural perturbations of images from SynPeDS. Variations of image characteristics from left to right: brightness, contrast, saturation and hue.

Table 5.15: Test results for the robustness of the DNN for concept-based pedestrian detection (CPD). Low negative differences (δ) demonstrate the robustness of CPD. Data augmentation during training only affects brightness variations.

Perturbation	Training	$\Delta mCQ_{0.99}$	$\Delta mCQ_{0.95}$	$\Delta mCQ_{0.50}$
Brightness	✓	0.00	0.00	0.00
Contrast	-	-0.02	+0.06	-0.01
Saturation	-	0.00	0.00	0.00
Hue	-	-0.11	0.00	-0.04

still high and close to ideal clustering marked by $mCQ = 1.0$.

Figure 5.13: Inference results, including detection bounding boxes (white) for the DNN for concept-based domain adaptation for pedestrian detection (ConDA, right column) and the DNN for concept-based pedestrian detection (CPD, middle column). The left column shows the original images from Cityscapes with ground truth bounding boxes (blue) and the body part segmentation. The inference is conducted with a confidence threshold of 0.4.

5.5 Domain Adaptation based on Semantic Concepts

In the last step, the DNN for concept-based pedestrian detection (CPD) has been investigated in terms of generalization capabilities compared to the baseline method, the DNN for center, scale and offset prediction and segmentation (CSOS). This work focuses on generalization from synthetic data from SynPeDS to real data from Cityscapes (CS). Figure 5.13 shows the inference results of the proposed DNN for concept-based domain adaptation (ConDA) for pedestrian detection.

Table 5.16: Systematic evaluation [%] for Cityscapes of the DNN for concept-based domain adaptation (ConDA).

Method	Data	$FLAMR_g$		$FLAMR$		$FmIoU$	
		$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$	$\mathcal{G}_{\mathrm{f,v}}$	$\mathcal{G}_{\mathrm{b,v}}$
CSOS	SynPeDS → CS	11.74	19.73	19.79	25.82	47.88	31.06
CPD		9.98	17.53	17.44	25.70	60.07	40.56
ConDA	UDA	3.09	9.27	8.26	14.78	66.72	48.84
CSOS	CS → CS	0.24	3.58	3.03	5.70	74.94	55.87

Systematic Evaluation ConDA describes a two-stage approach: (1) supervised training on synthetic source data and (2) unsupervised domain adaptation with additional real target data without ground truth annotations. That's why in a first step the generalization of the CPD and CSOS for naive source-only training is evaluated. The columns for CSOS and CPD in Table 5.16 show results for training on synthetic data from SynPeDS and evaluation on real data from CS (SynPeDS → CS).

The results demonstrate that CPD substantially improves the generalization with respect to the detection metrics ($FLAMR_g, FLAMR$) but especially for the segmentation quality ($FmIoU$). In this case, generalization refers to the SynPeDS → CS setting. CPD outperforms CSOS by a large margin of roughly 13% absolutely in terms of the segmentation quality of clearly visible pedestrians in the foreground ($FmIoU$ for $\mathcal{G}_{\mathrm{f,v}}$).

Finally, CPD is used as the best possible starting point for unsupervised domain adaptation with ConDA. The proposed ConDA substantially reduces the performance gap compared to the supervised baseline of CSOS and CS → CS. Surprisingly, the segmentation quality of ConDA nearly matches the supervised performance leaving a gap of roughly 8% in terms of $FmIoU$ for $\mathcal{G}_{\mathrm{f,v}}$. The results are even closer for the background.

Structured Latent Space Similar to the analysis conducted for CPD, the distances of CPD and ConDA are shown in Figure 5.14. Although the separation of semantic concepts in the latent space of CPD for unseen real data is already established, ConDA is able to strengthen the inherent structure of the latent space with unsupervised domain adaptation.

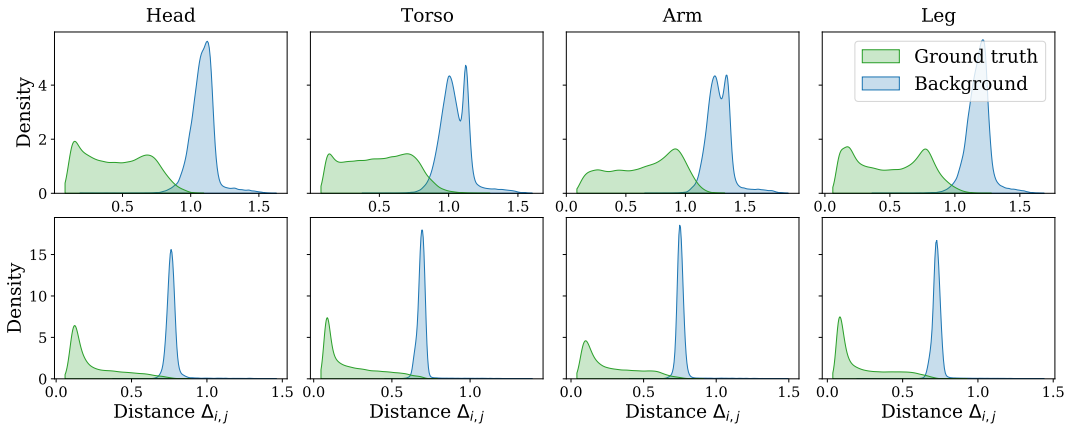


Figure 5.14: Relation of distances between concept vectors and latent representations as density plots for the unsupervised domain adaptation on Cityscapes. The top row shows results for the naive source-only training of the DNN for concept-based pedestrian detection (CPD) with SynPeDS. As shown in the bottom row, ConDA is able to increase the separation of true latent representations for four semantic concepts and background.

Table 5.17: Cluster quality for Cityscapes of the DNN for concept-based domain adaptation for pedestrian detection (ConDA).

Method	$mCQ_{0.99}$	$mCQ_{0.95}$	$mCQ_{0.50}$
CPD	0.82	0.87	0.69
ConDA	0.87	0.89	0.83

Cluster quality Although the performance improvements for the detection metrics and segmentation quality already undermine the substantial contribution, assurance metrics can be analyzed. Table 5.17 shows that the cluster quality for semantic concepts of ConDA adapts to the real domain as the increased scores indicate.

6. Conclusion

This work investigates the inherent interpretability of DNNs for safe pedestrian detection for automated driving to support a safety argumentation. Enforcing inherent interpretability is motivated by a comprehensive methodology. The methodology highlights the interdependencies and motivation of the individual contributions in this work.

6.1 Contribution

In this work, a methodology for failure-driven ML engineering for safe pedestrian detection is proposed. ML engineering aims at generating verifiable evidence for the correct functioning of a DNN for pedestrian detection to support a safety argumentation. Hence, the development of various methods and metrics is motivated. By developing methods for inherently interpretable and domain-adapted DNNs, functional insufficiencies are mitigated. The mitigation of functional insufficiencies is argued by novel assurance metrics. The proposed method for the systematical evaluation enables the definition of application-oriented detection metrics. Both elements, mitigation of functional insufficiencies and systematical evaluation, provide metrics, detection and assurance metrics, to support a safety argumentation with verifiable evidence.

The methodology is motivated by well-established safety standards and promotes rethinking safety aspects such as decomposition and traceability in the context of DNNs. An initial safety argumentation, which incorporates key aspects of this work, is presented. The structured argument allocates the proposed methods and metrics.

This works demonstrate the importance of application-oriented detection metrics and focus on ensuring safety for safety-critical pedestrians. The systematical evaluation of DNNs for pedestrian detection embedded in automated driving systems describes detection metrics to ensure the highest possible detection perfor-

mance of safety-critical pedestrians. This work argues that clearly visible pedestrians in the foreground and background are safety-critical pedestrians. The distinction is based on the distances between pedestrians and the automated vehicle and the visibility of pedestrians. Hence, the categorization of ground truth bounding boxes directly depends on the system perspective for automated driving. As a consequence, weaknesses of state-of-the-art evaluation protocols for pedestrian detection are resolved. The proposed detection metrics guarantee valid benchmarks. Since the checkpoint selection during the training depends on the results of the systematical evaluation, the detection metrics support the safety argumentation of DNNs.

Inherent interpretability is the centerpiece of this work. The inherent interpretability of DNNs has been gradually extended from handling classification tasks to pedestrian detection with an auxiliary body part segmentation. A novel interpretable reasoning within DNNs for pedestrian detection is proposed. Here, predictions are endorsed by unsupervised prototypes or supervised semantic concept vectors. This works outlines the close relationship between inherent interpretability, deep metric learning and unsupervised domain adaptation. These research fields propose methods to enforce the training of discriminate features in a structured latent space. High inter-class separation and high intra-class concentration characterize a structured latent space. This work delivers verifiable evidence based on the structured latent space enforced by specifically designed DNN architectures and loss formulations. Assurance metrics support a safety argumentation by quantifying segmentation and cluster quality and enabling advanced testing of robustness and generalization capabilities. Consequently, this work identifies inherent interpretability as an enabler of the safety argumentation of DNNs for pedestrian detection.

A generic pedestrian detector with state-of-the-art performance is defined to establish valid benchmarks. A great emphasis is on the strict development of gradually adapted DNN architectures and loss formulations. Based on that, assurance and detection metrics quantify the improvements of inherently interpretable DNNs for pedestrian detection compared to non-interpretable baseline methods.

Similar to the close relation of the relevant research fields, the lack of interpretability, robustness and generalization are highly overlapping functional insufficiencies. One possibility for mitigation is to leverage additional training data. However, manually labeled ground truth annotations are generally labor-intensive and costly. Here, synthetic data has rich potential to overcome these weaknesses and provide scalable methods to adapt DNNs to different domains. Most com-

monly, domains are either represented by real-world or synthetic data. Unsupervised domain adaptation motivates base training of DNNs with easily available labeled synthetic data and adaptations for automated driving with real-world data without ground truth annotations. This works highlights that inherent interpretability improves the generalization capabilities of DNNs for pedestrian detection. Moreover, the application of assurance metrics is extended to other domains and used to quantify the robustness and generalization capabilities of interpretable DNNs.

6.2 Future Work

The proposed methods CPD and ConDA require at least partly annotated ground truth data and the pre-definition of semantic concepts. Methods for unsupervised semantic segmentation or object discovery, i.e., SegDiscover [50] and ACSeg [63], try to overcome this shortcoming. Both methods classify semantic concepts based on clustered representations in the latent space without supervision. Meanwhile, DINOSAUR [92] and deep spectral methods [75] propose a finer and more object-centric perspective on unsupervised representation learning. In future work, methods for unsupervised semantic segmentation or representation learning can be used to automatically extract important and relevant semantic concepts for pedestrian detection. Extracted concepts could immediately be inspected and verified. However, the general importance of prototypes or concept vectors for supporting a safety argumentation is already outlined in this work.

Although CPD and ConDA use simple decision rules to make predicted bounding boxes more plausible, no module within the DNN assembles pedestrian instances by connecting key points or segmented body parts. Instead, interpretable reasoning is achieved on a pixel level. Future work should investigate possibilities to extend the reasoning to segment pedestrian instances based on predicted semantic concepts of humans. The application of graph neural networks for human pose estimation seems promising in this context [115]. As a result, pedestrian instances are built from semantic concepts and should be mapped to detection bounding boxes. Hence, inherent interpretability gains strength, providing additional evidence to support a safety argumentation. Especially in the case of high occlusions or dense human crowds, these approaches can contribute substantially to understanding complex scenes.

The generic DNN architecture can easily be extended with perception heads for depth estimation as an additional regression task and instance segmentation [109].

Since the generic pedestrian detector allows the easy addition of perception heads, some encouraging experiments have been conducted. Presumably, a robust fusion of predictions of 2D detection bounding boxes, depth and instance segmentation can contribute to safety. Here, the development of potent post-processing steps has to be encouraged.

Finally, the proposed methods do currently not consider image sequences. Instead, CPD and ConDA only process single RGB images, thus, valuable temporal information is not incorporated. Applying tracking algorithms [83] might improve the detection performance and consistency of pedestrian detections. Nevertheless, inherent interpretability lays the foundation for safe pedestrian detection with DNNs presented in this work.

Bibliography

[1] André Araujo, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. *Distill*, 2019. https://distill.pub/2019/computing-receptive-fields.

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[3] Aleksandar Bojchevski, Yves Matkovic, and Stephan Günnemann. Robust spectral clustering for noisy data: Modeling sparse corruptions improves latent embeddings. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 737–746, 2017.

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

[5] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1844–1861, 2019.

[6] Simon Burton. A causal model of safety assurance for machine learning. *arXiv preprint arXiv:2201.05451*, 2022.

[7] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6154–6162, 2018.

[8] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[9] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[12] Runfa Chen, Yu Rong, Shangmin Guo, Jiaqi Han, Fuchun Sun, Tingyang Xu, and Wenbing Huang. Smoothing matters: Momentum transformer for domain adaptive semantic segmentation. *arXiv preprint arXiv:2203.07988*, 2022.

[13] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive faster r-cnn. *International Journal of Computer Vision*, 129(7):2223–2243, 2021.

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[16] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[17] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, pages 224–227, 1979.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021.

[20] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 304–311, 2009.

[21] 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.

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

[23] Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv preprint arXiv:2112.00639*, 2021.

[24] 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.

[25] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10849–10859, 2021.

[26] Patrick Feifel, Frank Bonarens, and Frank Köster. Leveraging interpretability: Concept-based pedestrian detection with deep neural networks. In *Computer Science in Cars Symposium*, pages 1–10, 2021.

[27] Patrick Feifel, Frank Bonarens, and Frank Koster. Reevaluating the safety impact of inherent interpretability on deep neural networks for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29–37, 2021.

[28] Patrick Feifel, Frank Bonarens, and Frank Köster. Domain adaptive pedestrian detection based on semantic concepts. In *Proceeding of VISIGRAPP (4: VISAPP)*, 2023.

[29] Patrick Feifel, Benedikt Franke, Arne Raulf, Friedhelm Schwenker, Frank Bonarens, and Frank Köster. Revisiting the evaluation of deep neural networks for pedestrian detection. In *Proceedings of the Workshop on Artificial Intelligence Safety 2022*, 2022.

[30] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8730–8738, 2018.

[31] Radio Technical Commission for Aeronautics (RTCA). Do-178c. Technical report, RTCA Special Committee 205 (SC-205) and EUROCAE Working Group 71 (WG-71), 2012.

[32] International Organization for Standardization. Iso 26262-1:2018 road vehicles — functional safety — part 1: Vocabulary. Technical report, ISO/TC 22/SC 32 Electrical and electronic components and general system aspects, 2018.

[33] International Organization for Standardization. Iso 26262-2:2018 road vehicles — functional safety — part 2: Management of functional safety. Technical report, ISO/TC 22/SC 32 Electrical and electronic components and general system aspects, 2018.

[34] International Organization for Standardization. Iso/awi pas 8800 road vehicles — safety and artificial intelligence. Technical report, ISO/TC 22/SC 32 Electrical and electronic components and general system aspects, 2021.

[35] International Organization for Standardization. Iso/cd ts 5083 road vehicles — safety for automated driving systems - design, verification and validation. Technical report, ISO/TC 22/SC 32 Electrical and electronic components and general system aspects, 2021.

[36] International Organization for Standardization. Iso 21448:2022 road vehicles — safety of the intended functionality (sotif). Technical report, ISO/TC 22/SC 32 Electrical and electronic components and general system aspects, 2022.

[37] David Forsyth and Jean Ponce. *Computer vision: A modern approach.* Prentice hall, 2011.

[38] Benedikt Franke. Improving pedestrian detection and evaluation by utilizing image segmentation. Master's thesis, Ulm University, Institute of Neural Information Processing, 2022.

[39] Bundesministerium für Digitales und Verkehr (BMDV). BMDV - Automatisiertes und vernetztes Fahren im Überblick, 2022.

[40] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[41] Kamaledin Ghiasi-Shirazi. Generalizing the Convolution Operator in Convolutional Neural Networks. *Neural Processing Letters*, 50(3):2627–2646, December 2019. arXiv: 1707.09864.

[42] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[43] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[44] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

[45] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11328–11337, 2021.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[47] Martin Herrmann, Christian Witt, Laureen Lake, Stefani Guneshka, Christian Heinzemann, Frank Bonarens, Patrick Feifel, and Simon Funke. Using ontologies for dataset engineering in automotive ai applications. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 526–531. IEEE, 2022.

[48] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022.

[49] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. *arXiv preprint arXiv:2204.13132*, 2022.

[50] Haiyang Huang, Zhi Chen, and Cynthia Rudin. Segdiscover: Visual concept discovery via unsupervised semantic segmentation. *arXiv preprint arXiv:2204.10926*, 2022.

[51] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masked transformers for semantic segmentation. *arXiv preprint arXiv:2112.12782*, 2021.

[52] 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.

[53] Tim P Kelly. *Arguing safety - a systematic approach to safety case management*. PhD thesis, York University, York, 1999.

[54] Abdul Hannan Khan, Mohsin Munir, Ludger van Elst, and Andreas Dengel. F2dnet: Fast focal detection network for pedestrian detection. *arXiv preprint arXiv:2203.02331*, 2022.

[55] Rawal Khirodkar, Brandon Smith, Siddhartha Chandra, Amit Agrawal, and Antonio Criminisi. Sequential ensembling for semantic segmentation. *arXiv preprint arXiv:2210.05387*, 2022.

[56] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[57] Youngseok Kim, Sanmin Kim, Juyeb Shin, Jun Won Choi, and Dongsuk Kum. Crn: Camera radar net for accurate, robust, efficient 3d perception. *arXiv preprint arXiv:2304.00670*, 2023.

[58] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

[59] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[60] Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer, 2002.

[61] Jinpeng Li, Shengcai Liao, Hangzhi Jiang, and Ling Shao. Box Guided Convolution for Pedestrian Detection. In *28th ACM International Conference on Multimedia*, pages 1615–1624, 2020.

[62] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations ICLR*, 2021.

[63] Kehan Li, Zhennan Wang, Zesen Cheng, Runyi Yu, Yian Zhao, Guoli Song, Chang Liu, Li Yuan, and Jie Chen. Acseg: Adaptive conceptualization for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7162–7172, 2023.

[64] Qing Lian, Tai Wang, Dahua Lin, and Jiangmiao Pang. Dort: Modeling dynamic objects in recurrent for multi-camera 3d object detection and tracking. *arXiv preprint arXiv:2303.16628*, 2023.

[65] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[66] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2980–2988, 2017.

[67] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

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

[69] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. Learning Efficient Single-Stage Pedestrian Detectors by Asymptotic Localization

Fitting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–634, 2018.

[70] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5187–5196, 2019.

[71] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[72] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016.

[73] Tao Lu, Xiang Ding, Haisong Liu, Gangshan Wu, and Limin Wang. Link: Linear kernel for lidar-based 3d perception. *arXiv preprint arXiv:2303.16094*, 2023.

[74] Zekun Luo, Zheng Fang, Sixiao Zheng, Yabiao Wang, and Yanwei Fu. Nms-loss: Learning with non-maximum suppression for crowded pedestrian detection. In *International Conference on Multimedia Retrieval*, pages 481–485, 2021.

[75] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8364–8375, 2022.

[76] Panagiotis Meletis, Xiaoxiao Wen, Chenyang Lu, Daan de Geus, and Gijs Dubbelman. Cityscapes-panoptic-parts and pascal-panoptic-parts datasets for scene understanding. *arXiv preprint arXiv:2004.07944*, 2020.

[77] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

[78] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision*, 129:1551 – 1579, 2020.

[79] Keivan Nalaie, Kamaledin Ghiasi-Shirazi, and Modhammad-R. Akbarzadeh-T. Efficient implementation of a generalized convolutional neural networks based on weighted euclidean distance. In *International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 211–216. IEEE, 2017.

[80] U.S. National Highway Traffic Safety Administration (NHTSA). Automated Vehicles for Safety | NHTSA, 2022.

[81] Poojan Oza, Vishwanath A Sindagi, Vibashan VS, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. *arXiv preprint arXiv:2105.13502*, 2021.

[82] Hanyang Peng and Shiqi Yu. Beyond softmax loss: Intra-concentration and inter-separability loss for classification. *Neurocomputing*, 438:155–164, May 2021.

[83] Mohammed Razzok, Abdelmajid Badri, Ilham El Mourabit, Yassine Ruichek, and Aïcha Sahel. Pedestrian detection and tracking system based on deep-sort, yolov5, and new data association metrics. *Information*, 14(4):218, 2023.

[84] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[85] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[86] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[87] Stefan Rude. Set level abschlusspräsentation. `https://setlevel.de/en/news/final-presentation-slides-and-videos`. Accessed: 2023-04-12.

[88] 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.

[89] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.

[90] Gesina Schwalbe. Concept embedding analysis: A review. *arXiv preprint arXiv:2203.13909*, 2022.

[91] Gesina Schwalbe. *Concept Embedding Analysis Based Methods for the Safety Assurance of Deep Neural Networks : towards safe automotive computer vision applications*. PhD thesis, University of Bamberg, Bamberg, 2022.

[92] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022.

[93] 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.

[94] Poulami Sinhamahapatra, Lena Heidemann, Maureen Monnet, and Karsten Roscher. Towards human-interpretable prototypes for visual assessment of image classification models. *arXiv preprint arXiv:2211.12173*, 2022.

[95] Xiaolin Song, Kaili Zhao, Wen-Sheng Chu, Honggang Zhang, and Jun Guo. Progressive refinement network for occluded pedestrian detection. In *European Conference on Computer Vision (ECCV)*, pages 32–48. Springer, 2020.

[96] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum self-paced learning for cross-domain object detection. *Computer Vision and Image Understanding*, 204:103166, 2021.

[97] Thomas Stauner, Frederik Blank, Michael Fürst, Johannes Günther, Korbinian Hagn, Philipp Heidenreich, Markus Huber, Bastian Knerr, Thomas Schulik, and Karl-Ferdinand Leiß. Synpeds: A synthetic dataset for pedestrian detection in urban traffic scenes. In *Proceedings of the 6th ACM Computer Science in Cars Symposium*, pages 1–10, 2022.

[98] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019.

[99] 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.

[100] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[101] Ömer Şahin Taş, Niels Ole Salscheider, Fabian Poggenhans, Sascha Wirges, Claudio Bandera, Marc René Zofka, Tobias Strauss, J Marius Zöllner, and Christoph Stiller. Making bertha cooperate–team annieway's entry to the 2016 grand cooperative driving challenge. *IEEE Transactions on Intelligent Transportation Systems*, 19(4):1262–1276, 2017.

[102] 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.

[103] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[104] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[105] Hongsong Wang, Shengcai Liao, and Ling Shao. Afan: Augmented feature alignment network for cross-domain object detection. *IEEE Transactions on Image Processing*, 30:4046–4056, 2021.

[106] Jiahang Wang, Sheng Jin, Wentao Liu, Weizhong Liu, Chen Qian, and Ping Luo. When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11855–11864, 2021.

[107] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7774–7783, 2018.

[108] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, Yangyang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9603–9612, 2021.

[109] Yuqing Wang, Zhaoliang Xu, Hao Shen, Baoshan Cheng, and Lirong Yang. Centermask: single shot instance segmentation with point representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9313–9321, 2020.

[110] 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.

[111] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[112] 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. Springer, 2020.

[113] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[114] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *arXiv preprint arXiv:2204.08808*, 2022.

[115] Chris Xie, Arsalan Mousavian, Yu Xiang, and Dieter Fox. Rice: Refining instance masks in cluttered environments with graph neural networks. In *Conference on Robot Learning*, pages 1655–1665. PMLR, 2022.

[116] Zixuan Xu, Banghuai Li, Ye Yuan, and Anhong Dang. Beta r-cnn: Looking into pedestrian detection from another perspective. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:19953–19963, 2020.

[117] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *arXiv preprint arXiv:2208.11112*, 2022.

[118] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2403–2412, 2018.

[119] Jialiang Zhang, Lixiang Lin, Jianke Zhu, Yang Li, Yun-chen Chen, Yao Hu, and CH Steven Hoi. Attribute-aware pedestrian detection in a crowd. *IEEE Transactions on Multimedia*, 2020.

[120] Jingyi Zhang, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, and Shijian Lu. Da-detr: Domain adaptive detection transformer by hybrid attention. *arXiv preprint arXiv:2103.17084*, 2021.

[121] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 443–457. Springer, 2016.

[122] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12414–12424, 2021.

[123] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3221, 2017.

[124] Penghao Zhou, Chong Zhou, Pai Peng, Junlong Du, Xing Sun, Xiaowei Guo, and Feiyue Huang. Noh-nms: Improving pedestrian detection by nearby objects hallucination. In *ACM International Conference on Multimedia*, pages 1967–1975, 2020.

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

[126] Tobias Zielke. Pedestrian detection in occlusion scenarios with depth information. Master's thesis, Ulm University, Institute for Databases and Information Systems (DBIS), 2022.

[127] Zhuofan Zong, Dongzhi Jiang, Guanglu Song, Zeyue Xue, Jingyong Su, Hongsheng Li, and Yu Liu. Temporal enhanced training of multi-view 3d object detector via historical object prediction. *arXiv preprint arXiv:2304.00967*, 2023.

List of Acronyms and Symbols

Acronyms

AEB	Automated emergency braking
APD	DNN for attribute-aware pedestrian detection
API	Application programming interface
BGC	DNN with box guided convolution for pedestrian detection
CM	Convolutional module
COCO	Common objects in context
ConDA	Deep neural network for concept-based domain adaptation for pedestrian detection
CPD	Concept-based pedestrian detection
CS	Cityscapes dataset
CSO	DNN for center, scale and offset prediction
CSOS	DNN for center, scale and offset prediction and segmentation
CSP	DNN for center, scale prediction
DA	Deep neural network artifact
DLA	Deep layer aggregation
DNN	Deep neural network
FPN	Feature pyramid network
GPS	Global positioning system
HRNet	High resolution network
ISO	International Organization for Standardization
LiDAR	Light detection and ranging

MDLA	Modified deep layer aggregation
ML	Machine learning
PPD	Deep neural network for prototype-based pedestrian detection
ProtoPNet	Prototypical part network
Radar	Radio detection and ranging
ResNet	Residual network
RGB	Red, green and blue
SOTIF	Safety of the intended functionality
SynPeDS	Synthetic pedestrian dataset
UDA	Unsupervised domain adaptation

Symbols

Symbols are divided in basic symbols, metrics, sets and tensors.

Basic Symbols

Basic symbols are denoted with greek and latin letters. The single letters can be either lower and upper case and can be equipped with multiple indices or an accent.

a	Intra-class distance
$\mathrm{A}(g, d)$	Alignment of ground truth g and detection d bounding box
α	Exponential moving average parameter
b	Inter-class distance
β	Weight for transformation of normalized concept similarities
c	Confidence threshold
c_x^g, c_x^d, c_y^g, c_y^d	Center coordinate in x- and y-direction of ground truth g and detection d bounding box
d	Detection bounding box
$d(g)$	Depth of ground truth bounding box g (m)
d_{AEB}	Distance of automated emergency braking (m)
d_s	Additional distance for braking (m)
d_v	Distance from rear axis of vehicle to front (m)
D	Channel size of 3D tensor
δ	Distance
ϵ	Value for numerical stability of normalization
η	Bias for transformation of normalized concept similarities
f	Feature extraction
g	Ground truth bounding box
G	Gravitational acceleration on earth ($\mathrm{m\,s^2}$)
γ	Focal loss parameter
h	Perception head
$h(r)$	Height of bounding box r
H	Image height
H'	Spatially reduced height of latent tensor
H^*	Height of receptive field
l	Smooth L1 loss
L	Loss value
L_S	Supervised loss values for source domain
L_T	Unsupervised loss value for target domain

Basic Symbols

λ_{center}, λ_{scale}, λ_{offset}	Weight for center, scale and offset loss
$\lambda^{\text{P}}_{\text{cluster}}$, $\lambda^{\text{C}}_{\text{cluster}}$	Weight for clustering loss of prototypes, concepts
$\lambda^{\text{P}}_{\text{separation}}$, $\lambda^{\text{C}}_{\text{separation}}$	Weight for separation loss of prototypes, concepts
m, m_θ	DNN with parameters θ
$\hat{m}$, $\hat{m}_\theta$	Student DNN (with parameters θ)
$\breve{m}$, $\breve{m}_{\theta'}$	Teacher DNN (with averaged parameters θ')
$\bar{m}$, $\bar{m}_{\bar{\theta}}$	Oracle DNN (with parameters θ)
μ	Friction coefficient
μ_s	Threshold for scale error
μ_l	Threshold for localization error
N_B	Number of background pixels
N_C	Number of concepts
N_I	Number of images
N_k	Sample size for t-test
N_K	Number of pedestrians
N_P	Number of prototypes
N^+, N^-	Number of positive and negative pixels
$\nu_c(g)$, $\nu_e(g)$	Visibility w.r.t. crowd and environmental occlusion
P	Ground truth category
$\psi(\delta)$	Separation weight for latent representation
r	2D bounding box
s	Downsampling rate
σ	Standard deviation
t_{proc}	Processing time to initiate braking (s)
τ	Thresholds for center, ignore, cluster and separation
θ	Parameters of DNN
v	Velocity (m/s)
$w(r)$	Width of bounding box r
W	Image width
W'	Spatially reduced width of latent tensor
W^*	Width of receptive field
x	Direction along the width of an image
$\mathfrak{X}$	Image domain
$\mathfrak{X}_{\text{S}}$, $\mathfrak{X}_{\text{T}}$	Source and target domain
y	Direction along the height of an image
z	Small number for weighting the separation
$\mathfrak{Z}$	Latent space

Metrics

Metrics are written as acronyms and mark important performance measures.

CHI	Calinski-Harabsz index
DBI	Davis-Boulding index
$FLAMR$, $FLAMR(c)$	Filtered log-average miss rate
$FLAMR_g$	Filtered log-average miss rate w.r.t. ghost detections
$FmIoU$	Filtered mean intersection over union
$FPPI$, $FPPI(c)$	False positives per image
$GDPI$, $GDPI(c)$	Ghost detections per image
GC	Gini coefficient
IoU	Intersection over union
$LAMR$, $LAMR(c)$	Log-average miss rate
MR, $MR(c)$	Miss rate
MR_P, $MR_P(c)$	Miss rate for ground truth category P
SC	Silhouette coefficient

Sets

Sets are denoted with upper case calligraphic letters.

$\mathcal{C}$	Set of concept vectors $\mathrm{c}_\mathrm{k}^\mathrm{l}$
$\mathcal{D}(\mathrm{c})$	Set of detection bounding boxes d
$\mathcal{D}_\mathrm{g}(\mathrm{c})$	Set of ghost detections
$\mathcal{D}_\mathrm{l}(\mathrm{c})$	Set of detection errors related to localization
$\mathcal{D}_\mathrm{s}(\mathrm{c})$	Set of detection errors related to scale
$\mathcal{G}$	Set of ground truth bounding boxes g
$\mathcal{G}_\mathrm{a}$	Set of ambiguous occluded ground truth bounding boxes
$\mathcal{G}_\mathrm{b}$	Set of background ground truth bounding boxes
$\mathcal{G}_\mathrm{b,v}$	Set of clearly visible background ground truth bounding boxes: $\mathcal{G}_\mathrm{b} \cup \mathcal{G}_\mathrm{v}$
$\mathcal{G}_\mathrm{e}$	Set of environmental occluded ground truth bounding boxes
$\mathcal{G}_\mathrm{f}$	Set of foreground ground truth bounding boxes
$\mathcal{G}_\mathrm{f,v}$	Set of clearly visible foreground ground truth bounding boxes: $\mathcal{G}_\mathrm{f} \cup \mathcal{G}_\mathrm{v}$
$\mathcal{G}_\mathrm{v}$	Set of clearly visible ground truth bounding boxes
$\bar{\mathcal{G}}_\mathrm{T}$	Set of pseudo ground truth bounding boxes
$\mathcal{F}$	Set of values for false positives per image
$\mathcal{FN}(\mathrm{c})$	Set of false negatives
$\mathcal{FP}(\mathrm{c})$	Set of false positives
$\bar{\mathcal{I}}$	Set of ignored pseudo bounding boxes
$\mathcal{O}^\mathrm{c}$	Set of crowd occluding classes
$\mathcal{O}^\mathrm{e}$	Set of environmental occluding classes
$\mathcal{P}$	Set of prototype vectors $\mathrm{p}_\mathrm{k}^\mathrm{l}$
$\mathcal{T}$	Set of confidence thresholds
$\mathcal{TP}^\mathrm{d}(\mathrm{c})$	Set of detection true positives
$\mathcal{TP}^\mathrm{g}(\mathrm{c})$	Set of ground truth true positives
$\mathcal{Z}$	Set of latent representations $\mathrm{z}_\mathrm{k}^\mathrm{i,j}$

Tensors

For readability, tensors are used in both index and matrix notation. 1D tensors are denoted with lower case letters (a_i or $\mathbf{a}$) and 2D or 3D tensors with upper case letters ($A_{i,j}$, $A_{i,j,k}$ or $\mathbf{A}$). Tensors with greek letters mark predictions (output tensors) of deep neural networks.

$\mathrm{c}^{\mathrm{l}}_{\mathrm{k}}$, $\mathbf{c}^{\mathrm{l}}$	Concept vector in index and matrix notation
$\mathrm{C}_{\mathrm{i,j}}$	Crowd occlusion
$\mathrm{D}_{\mathrm{i,j}}$	Depth map
$\Delta_{\mathrm{i,j,l}}$, $\boldsymbol{\Delta}$	Distance in index and matrix notation
δ^*_{l}, δ^*	Critical distance in index and matrix notation
$\mathrm{E}_{\mathrm{i,j}}$	Environmental occlusion
$\mathrm{G}_{\mathrm{i,j}}$	Gaussian mask
$\tilde{\mathrm{G}}_{\mathrm{i,j}}$	Intermediate gaussian mask
$\Gamma_{\mathrm{i,j}}$	Transformed similarity for prototype-based reasoning
$\Gamma_{\mathrm{i,j,l}}$	Transformed similarity for concept-based reasoning
$\mathrm{I}^{\mathrm{box}}_{\mathrm{i,j}}$	Ignore mask with excluded bounding box
$\mathrm{I}^{\mathrm{inst}}_{\mathrm{i,j}}$	Ignore mask with excluded instances
$\mathrm{M}_{\mathrm{i,j,k}}$	Concept mask (one-hot encoded)
$\omega_{\mathrm{i,j,k}}$	Downsampled body part segmentation (ground truth, one-hot encoded)
$\Omega_{\mathrm{i,j,k}}$	Body part segmentation (ground truth, one-hot encoded)
$\hat{\Omega}_{\mathrm{i,j,k}}$, $\Omega^{'}_{\mathrm{i,j,k}}$, $\bar{\Omega}_{\mathrm{i,j,k}}$	Body part segmentation (prediction, one-hot encoded) of student $\hat{m}_\theta$, teacher $\breve{m}_{\theta'}$ and oracle $\bar{m}_{\bar{\theta}}$
$\mathrm{p}^{\mathrm{l}}_{\mathrm{k}}$, $\mathbf{p}^{\mathrm{l}}$	Prototype vector in index and matrix notation
$\mathrm{P}_{\mathrm{i,j}}$	Insance segmentation
$\Pi_{\mathrm{i,j}}$	Probability of prototype-based reasoning
$\Pi_{\mathrm{i,j,l}}$	Probability for concept-based reasoning
$\Phi_{\mathrm{i,j,k}}$	Scale (ground truth)
$\hat{\Phi}_{\mathrm{i,j,k}}$, $\Phi^{'}_{\mathrm{i,j,k}}$, $\bar{\Phi}_{\mathrm{i,j,k}}$	Scale (prediction) of student $\hat{m}_\theta$, teacher $\breve{m}_{\theta'}$ and oracle $\bar{m}_{\bar{\theta}}$
$\Psi_{\mathrm{i,j,k}}$	Offset (ground truth)
$\hat{\Psi}_{\mathrm{i,j,k}}$, $\Psi^{'}_{\mathrm{i,j,k}}$, $\bar{\Psi}_{\mathrm{i,j,k}}$	Offset (prediction) of student $\hat{m}_\theta$, teacher $\breve{m}_{\theta'}$ and oracle $\bar{m}_{\bar{\theta}}$
$\rho_{\mathrm{i,j,l}}$	Background for either concept or background separation
$\mathrm{S}_{\mathrm{i,j}}$	Semantic segmentation
$\Sigma_{\mathrm{i,j,l}}$	Similarity
$\hat{\Sigma}_{\mathrm{i,j,l}}$	Normalized similarity

Tensors

x_k^l, $\mathbf{x}^l$	Vector in index and matrix notation
$\mathbf{X}$	Image
$\mathbf{X}^*$	Receptive field
$\mathbf{X}_S$	Source image
$\mathbf{X}_T$	Target image
$\underset{\sim}{\Xi}_{i,j}$	Center (ground truth)
$\underset{\sim}{\tilde{\Xi}}_{i,j}$	Intermediate center
$\underset{\sim}{\hat{\Xi}}_{i,j}$, $\underset{\sim}{\Xi}'_{i,j}$, $\underset{\sim}{\bar{\Xi}}_{i,j}$	Center (prediction) of student $\hat{m}_\theta$, teacher $\breve{m}_{\theta'}$ and oracle $\bar{m}_{\bar{\theta}}$
$z_k^{i,j}$, $\mathbf{z}^{i,j}$	Latent representation in index and matrix notation
$Z_{i,j,k}$, $\mathbf{Z}$	Latent tensor in index and matrix notation

List of Figures

List of Tables

List of Algorithms