

Revisiting the Evaluation of Deep Neural Networks for Pedestrian Detection

Patrick Feifel^{1,2,†}, Benedikt Franke^{3,4,†}, Arne Raulf³, Friedhelm Schwenker⁴, Frank Bonarens¹ and Frank Köster^{3,2}

¹Stellantis, Opel Automobile GmbH

²Carl von Ossietzky Universität Oldenburg

³Deutsches Zentrum für Luft- und Raumfahrt

⁴Universität Ulm

Abstract

The reliable DNN-based perception of pedestrians represents a crucial step towards automated driving systems. Currently applied metrics for a subset-based evaluation prohibit an application-oriented performance evaluation of DNNs for pedestrian detection. We argue that the current limitation in evaluation can be mitigated by the use of image segmentation. In this work, we leverage the instance and semantic segmentation of Cityscapes to describe a rule-based categorization of potential detection errors for CityPersons. Based on our systematic categorization, the filtered log-average miss rate as a new performance metric for pedestrian detection is introduced. Additionally, we derive and analyze a meaningful upper bound for the confidence threshold. We train and evaluate four backbones as part of a generic pedestrian detector and achieve state-of-the-art performance on CityPersons by using a rather simple architecture. Our results and comprehensive analysis show benefits of the newly proposed performance metrics. Code for evaluation is available at <https://github.com/BeFranke/ErrorCategories>.

Keywords

pedestrian detection, evaluation, safety

1. Introduction

Pedestrian detection is a crucial perception task for automated driving systems (ADS). Due to high complexity of the ADS environment, supervised machine learning models such as deep neural networks (DNNs) outperform traditional computer vision models and meet the high performance standards. Hence, traditional methods such as HOG [1] have been replaced by DNNs, which can be designed single-staged and anchor-free [2, 3] or two-staged and anchor-based [4].

Avoiding false negatives is the main objective for pedestrian detection in an ADS. A critical scene as shown in Figure 1 outlines the key task: A group of pedestrians cross the street right in front of the automated vehicle (AV). Intuitively, the evaluation should focus on pedestrians in the immediate vicinity of an AV, rather than distant pedestrians standing on the sidewalk in the middle of a crowd. The goal is to build relevant subsets of an evaluation dataset that contains these highly safety-critical pedestrians. This enables a more meaningful performance evaluation of DNNs. Motivated by the Caltech evaluation protocol [5], occlusion-related



Figure 1: State-of-the-art DNNs for pedestrian detection are benchmarked with the log-average miss rate on the **reasonable** subset of the CityPersons validation dataset (left). From a safety perspective, particularly safety-critical pedestrians, such as the one standing directly in front of the automated vehicle, must be included in the evaluation and not be **ignored**. Our proposed error categories (right) correctly distinguish between **foreground** and **background**, among others. Based on them, we perform an application-oriented performance evaluation of DNNs for pedestrian detection.

or height-based subsets were proposed [6, 7, 8, 9].

The reasonable subset is most commonly used to benchmark DNNs for pedestrian detection. It is based on the visibility and pixel height of ground truth bounding boxes. As shown in Figure 1, using this sparse information can result in particularly safety-critical pedestrians being ignored in the reasonable subset. As a consequence, currently used metrics only give limited information on the application-oriented performance.

Despite efforts to address highly occluded and therefore very difficult pedestrian detection cases, we argue that a realistic performance evaluation of a DNN for pedestrian detection should primarily address pedestri-

The IJCAI-ECAI-22 Workshop on Artificial Intelligence Safety (AISafety 2022), July 24-25, 2022, Vienna, Austria

[†]These authors contributed equally.

✉ patrick.feifel@external.stellantis.com (P. Feifel);

benedikt.franke@dlr.de (B. Franke)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

ans in the near field of an AV. In this sense, we think that a missed pedestrian in a distant crowd is less significant than a missed pedestrian standing directly in front of the AV.

Although a high recall is the primary objective for a DNN in a safety-critical application, the precision strongly influences the ADS operation in a complex environment. The current subset-based evaluation neglects the impact of different forms of false positives. We argue that multiple detections of the same pedestrian are less problematic for an ADS than false positives randomly scattered in the scene without reference to pedestrian-like features. Thus, a clear distinction between false positives must be found.

In our work, we introduce a systematically derived categorization of errors that can leverage a safety argumentation for the DNN-based perception of pedestrians. Our contribution can be summarized as follows:

1. We propose a rule-based categorization that describes potential errors of a DNN for pedestrian detection.
2. We define novel performance metrics focusing on safety-critical pedestrians that enable application-oriented DNN evaluation.
3. We report results and analyze 44 different DNNs for pedestrian detection, divided into 11 training runs for four backbones.

2. Background

Pedestrian detection is usually done by locating a 2d bounding box and assigning the correct class. Most commonly, DNNs for pedestrian detection are evaluated with the log-average miss rate (LAMR) [10] on the *reasonable* subset of the CityPersons [11] validation dataset. We refer to this performance metric as $LAMR_r$. Since pedestrian detection is highly safety-critical and relevant to an ADS, the $LAMR_r$ aggregates the miss rate (MR) and false positives per image (FPPI).

Table 1 gives an overview of state-of-the-art DNNs for pedestrian detection that are evaluated on different subsets of CityPersons. The definitions of the subsets are based on the height interval $h = [50, 1024]$ and a varying visibility range $v = \frac{|R^G_{vis}|}{|R^G|}$ of a pedestrian: reasonable ($v = [0.65, 1]$), bare ($v = [0.90, 1]$), partial ($v = [0.65, 0.90]$) and heavy ($v = [0, 0.65]$).

3. Generic Pedestrian Detector

In this work, we provide a comprehensive analysis of different *backbones* that are commonly used for DNNs for pedestrian detection. To achieve comparable results,

Table 1

LAMR [%] for different subsets of the CityPersons validation dataset: reasonable (r), bare (b), partial (p) and heavy (h).

Method	r	b	p	h
CSP [2]	11.0	7.3	10.4	49.3
NOH-NMS [12]	10.8	6.6	11.2	53.0
RepLoss [7]	10.9	6.3	13.4	52.9
PRNet [13]	10.8	6.8	10.0	53.3
Beta R-CNN [14]	10.6	6.4	10.3	47.1
NMS-Loss [15]	10.1	-	-	-
Cascade R-CNN [16]	9.2	-	-	36.9
BGCNet [17]	8.8	6.1	8.0	43.9
APD [3]	8.8	5.8	8.3	46.6
F2DNet [4]	8.7	-	-	32.6

we propose a DNN-based and generic pedestrian detector (GPD) consisting of *feature extraction* and *perception heads*.

Feature Extraction Pre-trained image classification networks form the backbone of the feature extraction. To utilize backbones for pedestrian detection, additional layers (ALs) must be implemented. Based on computed features for various scales by the backbone, the feature extraction outputs a representation for a given input image. In our work, we use the following feature extractions:

- **CSP-ResNet-50:** CSP [2] creates high-level semantic features based on ResNet-50 [18] and deconvolutions.
- **FPN-ResNet-50:** Feature pyramid network (FPN) [3] that adds a pyramidal decoder to ResNet-50 to combine features from different scales.
- **MDLA-UP-34:** Modified DLA (MDLA) [19] augments DLA-34 [20] with deformable convolutions from lower layers to the output.
- **BGC-HRNet-w32:** BGC [17] adds deconvolutions to a HRNet-w32 [21] concatenating the outputs to form the final representation.

Perception Heads In total, we have three perception heads taking extracted features as inputs and outputting a center, scale (height w/o width) and offset map. Similar to APD [3], we apply 3x3 convolutions for each perception head.

Training We train and evaluate different GPD instances with varying pre-trained backbones on the CityPersons dataset. In the following, an instance of GPD is simply referred to as a pedestrian detector (PD). All PDs are trained with the Adam optimizer [22] without weight decay and a reduced image size of 640x1028 pixels. A linear warm up strategy is employed that increases the

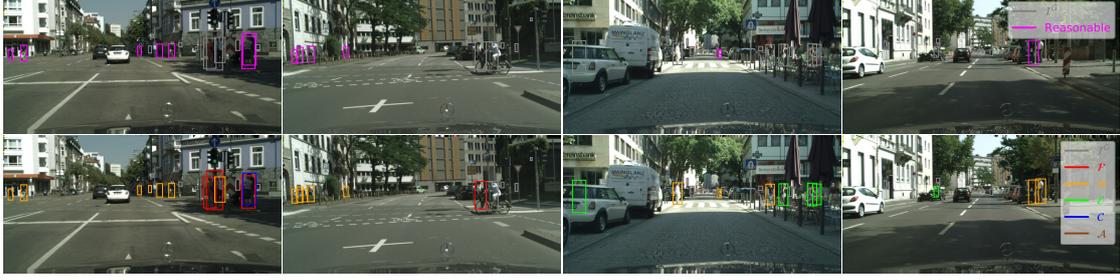


Figure 2: Incorrectly ignored bounding boxes from the **reasonable** subset of CityPersons are recovered by our proposed error categories.

learning rate from $5 \cdot 10^{-8}$ to the final learning rate of 10^{-4} over 2k iterations. We train for a maximum of 50k iterations on 2 GPUs with a batch size of 8. The final PD is given by the best checkpoint with the lowest LAMR score on the reasonable subset of the CityPersons validation dataset. ResNet-50¹, DLA-34² and HRNet-w32³ are used as pre-trained backbones on ImageNet. Furthermore, we apply the center, scale and offset loss terms according to CSP [2]. Common data augmentation techniques like modifying brightness, contrast or saturation are applied.

Inference For post-processing, we apply a confidence threshold of 0.01 and use NMS with a threshold of 0.5. The inference of PDs is conducted with the original image size of 1024x2048 pixels. Ground truth and detection bounding boxes are clipped to the image size.

4. Methodology

In the following, we introduce different categories for ground truth bounding boxes R^G of the CityPersons validation dataset. Matching ground truth with detection bounding boxes R^D based on our systematic categorization identifies errors for false negatives and false positives. Reducing false negatives is the primary safety-related objective during PD training. Intuitively, we expect false negatives to be positively correlated with pedestrian occlusion by other pedestrians and other environmental objects. That’s why categories regarding false negatives build upon the description of different forms of occlusions. As shown in Figure 2 and Figure 3, our proposed error categorization recovers ignored pedestrians for the reasonable and bare subset of the CityPersons validation dataset. Finally, we categorize false positives to identify the most disruptive ones for an ADS.



Figure 3: Incorrectly ignored bounding boxes from the **bare** subset of CityPersons are re-grouped to **background**.

Bounding Boxes We define a bounding box R as set of all pixels with (x, y) corner coordinates that fall into the bounding box: $R = \{(x, y) \mid x_1 \leq x < x_2 \wedge y_1 \leq y < y_2\}$. Therefore, the width of the bounding box is defined as $w(R) = x_2 - x_1$ and similarly the height $h(R) = y_2 - y_1$. A ground truth bounding box is denoted as $R^G \in \mathcal{G}$ in contrast to a detection bounding box $R^D \in \mathcal{D}$ which is associated with a confidence score $p(R^D)$. Because of highly overlapping detections, post-processing methods such as non-maximum suppression (NMS : $\mathcal{D} \rightarrow \mathcal{D}$) are applied to reduce the number of detections to $R^D \in \mathcal{D}$. Based on a predefined confidence threshold c , less confident detections are ignored: $D(c) = \{R^D \mid R^D \in \mathcal{D} \wedge p(R^D) > c\}$.

Generally, a pixel-precise match between bounding boxes can not be expected. Therefore the intersection over union (IoU) is used to measure the localization quality of R^D for R^G . The set of true positives is defined as:

$$\begin{aligned} TP^G(c) = & \left\{ R^G \mid R^G \in \mathcal{G} \wedge \exists R^D \in D(c) : \right. \\ & \left. [IoU(R^G, R^D) > 0.5 \wedge \nexists \tilde{R}^G \in \mathcal{G} : \right. \\ & \left. IoU(\tilde{R}^G, R^D) > IoU(R^G, R^D)] \right\} \end{aligned} \quad (1)$$

A ground truth bounding box R^G that can not be matched is a false negative $FN^G(c) = \mathcal{G} \setminus TP^G(c)$. A detection bounding box R^D that can not be matched or can only be matched to an already matched R^G is assigned to the set of false positives:

¹https://pytorch.org/hub/pytorch_vision_resnet/

²<http://dl.yf.io/dla/models/imagenet/dla34-ba72cf86.pth>

³<https://github.com/HRNet/HRNet-Image-Classification>



Figure 4: Categories for ground truth bounding boxes: foreground \mathcal{F} , background \mathcal{B} , environmental occlusion \mathcal{E} , crowd occlusion \mathcal{C} , ambiguous occlusion \mathcal{A} and ignored \mathcal{S}^G .

$$\text{FP}^D(c) = \left\{ \text{R}^D \mid \text{R}^D \in \text{D}(c) \wedge \exists \text{R}^G \in \mathcal{G} : \right. \\ \text{IoU}(\text{R}^G, \text{R}^D) > 0.5 \vee \exists \text{R}^G \in \mathcal{G} : \\ \left. \left[\text{IoU}(\text{R}^G, \text{R}^D) > 0.5 \wedge \exists \tilde{\text{R}}^D \in \mathcal{D} : \right. \right. \\ \left. \left. \text{IoU}(\text{R}^G, \tilde{\text{R}}^D) > \text{IoU}(\text{R}^G, \text{R}^D) \right] \right\} \quad (2)$$

Image Segmentation In this work, we employ ground truth for semantic segmentation \mathcal{S} and instance segmentation \mathfrak{S} to refine the subset-based evaluation of DNNs for pedestrian detection. $\mathcal{S}[x, y]$ = person means that the pixel at position (x, y) belongs to a pedestrian. $\mathfrak{S}[x, y] = i$ means that the pixel at position (x, y) has the instance ID i .

4.1. Error Categories for False Negatives

We define five error categories that separate ground truth bounding boxes of occluded pedestrians as well as highly safety-relevant pedestrians standing in the foreground or background. Examples of our categorization are shown in Figure 4. We propose a two-stage process to detect occlusion. First, potentially occluded pedestrians are identified based on the segmentation-based visibility ϕ , where i represents the instance ID belonging to the pedestrian:

$$\phi(\text{R}^G, i) = \frac{|\text{R}^G \cap \{(x, y) \mid \mathfrak{S}[x, y] = i\}|}{|\text{R}^G|} \quad (3)$$

The set of occlusion candidates $\tilde{\mathcal{O}}$ builds upon the threshold λ_ϕ : $\tilde{\mathcal{O}} = \{\text{R}^G \mid \text{R}^G \in \mathcal{G} \wedge \phi_c(\text{R}^G) < \lambda_\phi\}$. For our experiments, we empirically set $\lambda_\phi = 0.6$.

Environmental Occlusion Environmental occlusion occurs when a pedestrian is partially hidden behind objects in the scene, e.g. traffic signs, vegetation or cars. We

define \mathcal{O} as 20 selected classes of the semantic segmentation \mathcal{S} of Cityscapes [23] that can potentially cause occlusion. Truncated bounding boxes belong to this category, as the area that extends beyond the image is understood as environmental occlusion. We define the visibility with respect to the environment ϕ_e as

$$\phi_e(\text{R}^G) = \frac{|\text{R}^G \cap \{(x, y) \mid \mathcal{S}[x, y] \in \mathcal{O}\}|}{|\text{R}^G|} \quad (4)$$

For our experiments, we empirically set $\lambda_e = 0.7$. We define the intermediate set of environmentally occluded ground truth bounding boxes as $\tilde{\mathcal{E}} = \{\text{R}^G \mid \text{R}^G \in \tilde{\mathcal{O}} \wedge \phi_e(\text{R}^G) > \lambda_e\}$.

Crowd Occlusion Crowd occlusion (also intra-class occlusion [7]) occurs when a pedestrian is occluded by other pedestrians. We define the intra-class visibility ϕ_c that describes the relation of the instance area of a pedestrian to the semantic area occupied by the person class:

$$\phi_c(\text{R}^G, i) = \frac{|\text{R}^G \cap \{(x, y) \mid \mathfrak{S}[x, y] = i\}|}{|\text{R}^G \cap \{(x, y) \mid \mathcal{S}[x, y] = \text{person}\}|} \quad (5)$$

We introduce the threshold λ_c and define the intermediate set of crowd occluded ground truth bounding boxes as $\tilde{\mathcal{C}} = \{\text{R}^G \mid \text{R}^G \in \tilde{\mathcal{O}} \wedge \phi_c(\text{R}^G, i) > \lambda_c\}$. For our experiments, we empirically set $\lambda_c = 0.5$.

Ambiguous Occlusion Ambiguous occlusion occurs when pedestrians are simultaneously occluded by the environment and other pedestrians. We introduce the ambiguity factor $\lambda_a \in (0, 1)$ to relax the thresholds for crowd and environment occlusion and define $\mathcal{A}^E = \{\text{R}^G \mid \text{R}^G \in \tilde{\mathcal{E}} \wedge \phi_e(\text{R}^G) > \lambda_e \cdot \lambda_a\}$ and $\mathcal{A}^C = \{\text{R}^G \mid \text{R}^G \in \tilde{\mathcal{C}} \wedge \phi_c(\text{R}^G) > \lambda_c \cdot \lambda_a\}$.

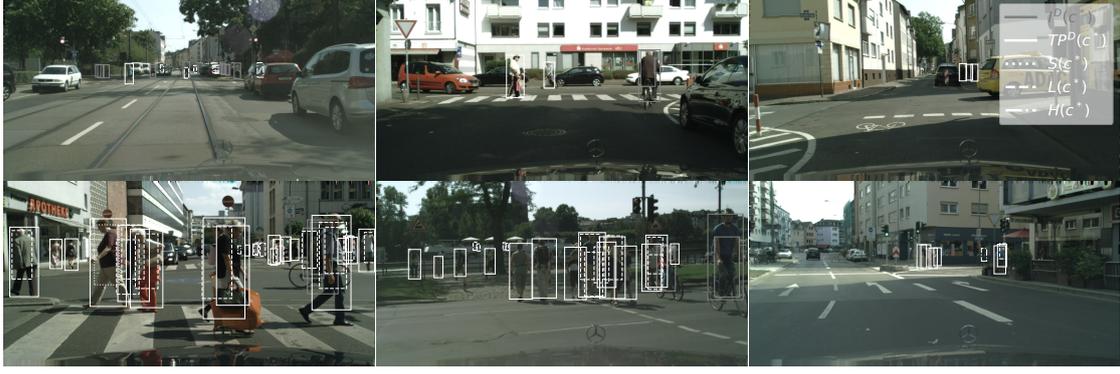


Figure 5: Categories for detection bounding boxes with an empirically defined confidence threshold ($c^* = 0.1$): true positives $TP^D(c)$ (solid), ghost detections $H(c)$ (dash dotted), localization errors $L(c)$ (dashed), scale errors $S(c)$ (dotted) and ignored detections $IP^D(c)$.

Ground truth bounding boxes with ambiguous occlusion are defined as $\mathcal{A} = \mathcal{A}^E \cup \mathcal{A}^C$. Based on that, the set of environmentally occluded ground truth bounding boxes is reduced to $\mathcal{E} = \mathcal{E} \setminus \mathcal{A}^E$ and the set of crowd occluded ground truth bounding boxes is $\mathcal{C} = \mathcal{C} \setminus \mathcal{A}^C$. For our experiments, we empirically set $\lambda_a = 0.75$.

Foreground and Background After defining occluded ground truth bounding boxes as $\mathcal{O} = \mathcal{E} \cup \mathcal{C} \cup \mathcal{A}$, the clearly visible bounding boxes are given by $\mathcal{V} = \mathcal{G} \setminus \mathcal{O}$. By applying a height threshold λ_f , we can further divide \mathcal{V} into foreground \mathcal{F} or background \mathcal{B} . First, we define the foreground $\mathcal{F} = \{R^G \mid R^G \in \mathcal{V} \wedge \text{height}(R^G) \geq \lambda_f\}$. Then, $F(c) = FN^G(c) \cap \mathcal{F}$ defines errors in the foreground. In order to define a reasonable λ_f , the braking distance of an automated emergency braking d_{AEB} is defined as $d_{AEB} = d_s + d_v + \left\lceil \frac{v^2}{2 \cdot \mu \cdot g} \right\rceil + \left\lceil v \cdot t_{proc} \right\rceil$.

Table 2
Parameters for the simplified braking distance calculation of an automated emergency braking d_{AEB} with $30 \frac{km}{h}$.

Parameter	Value	Description
h	1.7 m	Pedestrian height
t_{proc}	0.4 sec	Processing time
μ	0.3	Friction coefficient
v	$8.33 \frac{m}{s}$	Velocity
g	$9.81 \frac{m}{s^2}$	Gravitational constant
d_s	2 m	Added distance
d_v	4 m	Distance from rear axis to front

Applying the parameters shown in Table 2, the separating distance is $d_{AEB} = 22 m$. Based on camera calibration parameters of Cityscapes, the corresponding pixel height λ_f is approximately 190 pixels. Finally, the background is

specified as $\mathcal{B} = \mathcal{V} \setminus \mathcal{F}$ and potential background errors are defined as: $B(c) = FN^G(c) \cap \mathcal{B}$.

Due to highly crowd-occluded pedestrians that introduce doubtful false negatives into the evaluation, we relax the matching strategy in Equation 1: For all $R^G \in \mathcal{F} \cup \mathcal{B}$, we see R^G as a true positive if there exists a detection with $IoU > 0.5$ irrespective of another $R^G \in \mathcal{E}$ that could be matched with a higher IoU .

4.2. Error Categories for False Positives

A detection bounding box $R^D \in FP^D(c)$ is a false positive. We argue that false positives that coincide with pedestrian crowds do not disrupt the operation of an ADS as much as unrelated and random false positives. Hence, we propose three error categories with respect to false positives in order to identify the most disruptive. For examples see Figure 5.

Scale Errors This category includes detections that fail only with respect to the scale of the bounding box. Let $cX(R)$, $cY(R)$ give the x- and y-center coordinates of any bounding box R and λ_o the maximum permitted center offset. The predicate that states whether the center of R^D is aligned with R^G , is defined as:

$$D(R^G, R^D) \iff |cX(R^G) - cX(R^D)| \leq \lambda_o w(R^G) \wedge |cY(R^G) - cY(R^D)| \leq \lambda_o h(R^G) \quad (6)$$

For our experiments, we empirically set $\lambda_o = 0.2$. Scale errors are defined as $S(c) = \{R^D \mid R^D \in FP^D(c) \wedge \exists R^G \in \mathcal{E} : D(R^G, R^D)\}$.

Localization Errors Holds all false positives that fall in close proximity to a R^G , but the detection can not be matched and is not a scale error. Localization errors

Table 3

Results of our experiments with different metrics.

Feature Extraction	LAMR			FLAMR _p				FLAMR _p ^ℱ			
	reasonable			ℱ		ℬ		ℱ		ℬ	
	best	μ	CI _{0.95}	μ	CI _{0.95}	μ	CI _{0.95}	μ	CI _{0.95}	μ	CI _{0.95}
FPN-ResNet-50	10.9	11.6	[11.2, 12.1]	4.5	[4.2, 4.9]	12.4	[11.7, 13.1]	1.9	[1.2, 2.5]	6.8	[6.3, 7.3]
CSP-ResNet-50	10.6	11.0	[10.7, 11.3]	5.2	[4.8, 5.5]	11.2	[10.8, 11.6]	2.2	[1.9, 2.4]	6.3	[6.2, 6.5]
MDLA-UP-34	9.6	10.5	[10.1, 10.8]	4.7	[4.2, 5.2]	10.4	[10.0, 10.8]	2.8	[2.5, 3.1]	6.6	[6.3, 6.9]
BGC-HRNet-w32	8.8	9.1	[9.0, 9.2]	3.8	[3.2, 4.4]	9.0	[8.7, 9.4]	1.6	[1.2, 2.0]	5.6	[5.3, 5.8]

are defined as $L(c) = \{R^D \mid R^D \in (FP^D(c) \setminus S(c)) \wedge \exists R^G \in \mathcal{G} : \text{IoU}(R^G, R^D) \geq \lambda_i\}$. For our experiments, we empirically set $\lambda_i = 0.25$.

Ghost Detections Inspired by a term from automotive radar systems [24], we define ghost detections as $H(c) = FP^D(c) \setminus (S(c) \cup L(c))$. Detections in this category are random and unrelated to the presence of pedestrians. Thus, these are strongly disruptive that severely impact the operation of an ADS.

4.3. Filtered Log-average Miss Rate

In the following, we derive metrics to measure the performance of a PD over the proposed error categories. Table 4 shows the number of ground truth bounding boxes for each category. The filtered miss rate $MR_p(c)$ accounts for ground truth bounding boxes with $P \in \{\mathcal{F}, \mathcal{B}, \mathcal{E}, \mathcal{C}, \mathcal{A}\}$:

$$MR_p(c) = \frac{|FN^G(c) \cap P|}{|TP^G(c) \cap P| + |FN^G(c) \cap P|} \quad (7)$$

Table 4

Allocation of ground truth bounding boxes for the CityPersons validation dataset.

Subset	ℱ	ℬ	ℰ	℄	℄
Cardinality	348	1269	364	438	130

False Positives per Image With reference to the LAMR, the filtered log-average miss rate (FLAMR_p) is defined as

$$FLAMR_p = \exp\left(\frac{1}{|C|} \sum_{c \in C} \log MR_p(c)\right) \quad (8)$$

Here, C is a set of confidence levels that correspond to the nine pre-defined FPPI(c) values for calculating the LAMR:

$$C = \left\{ \arg \max_{FPPI(c) \leq f} FPPI(c) \mid f \in F \right\} \quad (9)$$

with $F = \{10^{-2}, 10^{-1.78}, \dots, 10^0\}$ and $|F| = 9$.

Ghost Detections per Image Since not all false positives are equally disruptive, we propose to focus on the number of ghost detections per image $GDPI(c) = \frac{1}{N} |H(c)|$. Based on $GDPI(c)$ and Equation 9, we denote the set of confidence levels for ghost detections as $C^\mathcal{H}$. The filtered log-average miss rate with respect to ghost detections (FLAMR_p^ℱ) is defined as:

$$FLAMR_p^\mathcal{H} = \exp\left(\frac{1}{|C^\mathcal{H}|} \sum_{c \in C^\mathcal{H}} \log MR_p(c)\right) \quad (10)$$

4.4. Upper Bound for Confidence Threshold

From a safety perspective, we are interested in finding an operating point for a PD where no safety-critical pedestrian is missed. It is still open to what extent this requirement can be relaxed for DNNs for object tracking. In this work, we are focused on single images and define a safety-critical pedestrian as any pedestrian who is in the foreground. Furthermore, we assign the operating point to a confidence threshold that must be determined post-hoc to PD training. We define the confidence threshold $c \in [0, c_{\mathcal{F}}^*]$ and the upper bound $c_{\mathcal{F}}^*$ as the confidence threshold with the lowest miss rate for foreground \mathcal{F} : $c_{\mathcal{F}}^* = \arg \min MR_{\mathcal{F}}(c)$. If $MR_{\mathcal{F}}(c_{\mathcal{F}}^*) = 0$ holds for a given validation dataset, it can be ensured that every safety-critical pedestrian in the foreground is correctly detected. Lowering the confidence threshold so that $c < c_{\mathcal{F}}^*$ may improve performance for other error categories, but is not capable of causing foreground errors. Consequently, we see $c_{\mathcal{F}}^*$ as a reasonable choice for an operating point. The corresponding amount of ghost detections per image is given by $GDPI(c_{\mathcal{F}}^*)$.

5. Results

In total, we trained and analyze results for 44 PDs i.e. 11 PDs for each of the four different feature extractions and

backbones. PDs with the same feature extraction also differ since the randomly initialized AL parameters in the feature extraction and perception heads change for every training run. We report confidence intervals ($CI_{0.95}$), using a student’s t-distribution due to the small sample size. This accounts for randomness and improves transparency, although satisfactory sample sizes are difficult when working with large DNNs.

5.1. Log-average Miss Rate

LAMR scores for the reasonable subset of the CityPersons validation dataset are reported in Table 3. BGC-HRNet-w32 has the best LAMR performance, which confirms the reported benchmarks listed in Table 1. In summary, our experiments show overlapping $CI_{0.95}$, indicating that randomness in initialization and training influences performance. Interestingly, our PD with BGC-HRNet-w32 as backbone achieves a score very similar to the results reported for BGCNet [17] despite using a simpler architecture that does not employ box-guided convolutions.

5.2. Bias of Reasonable Subset

Compared to the $LAMR_r$ scores on the reasonable subset, we see a corresponding order of the $FLAMR_p$ scores for background in Table 3. This indicates that the reasonable subset holds a vast amount of smaller pedestrians in the background. BGC-HRNet-w32 performs best for all performance metrics and subsets. In contrast, $FLAMR_{\mathcal{F}}$ scores contradict the $LAMR_r$ results with a different ranking of PDs and strongly overlapping $CI_{0.95}$. The inherent bias of the $LAMR_r$ evaluation leads to underestimation of the true potential of certain feature extractions and backbones for the highly safety-critical foreground category.

Figure 6 analyzes the dependence of $LAMR_r$ scores and $FLAMR_p$ and $FLAMR_{\mathcal{F}}$ for foreground and background. Whereas $FLAMR_p$ scores are strongly correlated with $LAMR_r$ scores in the background, the dependence in the foreground is lower. Our analysis shows that the reasonable subset is dominated by pedestrians in the background, which are less safety-critical.

5.3. Application-oriented Evaluation

Evaluating the miss rate in foreground \mathcal{F} and background \mathcal{B} while only considering ghost detections per image (GDPI) combines the evaluation of opposed critical cases: Missing a safety-critical pedestrian or predicting non-existing pedestrians. The evaluation of the best run in terms of $LAMR_r$ and $FLAMR_p$ in Table 3 shows BGC-HRNet-w32 as the far superior choice. However, $FLAMR_{\mathcal{F}}$ -scores based on our systematic error categories

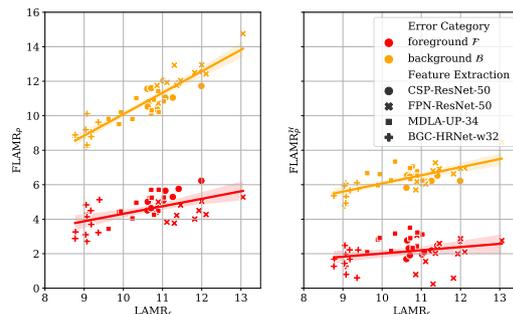


Figure 6: LAMR scores for the reasonable subset compared to the filtered log-average miss rate with and w/o respect to ghost detections ($FLAMR_p$, $FLAMR_{\mathcal{F}}$).

reveal that performance of FPN-ResNet-50 is comparable by achieving the same lower bound of $CI_{0.95}$. This is contradictory to the fact that BGC-HRNet-w32 outperforms FPN-ResNet-50 by nearly 2% in $LAMR_r$. We observe the reversed effect for MDLA-UP-34 which performs second-best in $LAMR_r$ but achieves the worst result for $FLAMR_{\mathcal{F}}$.

Figure 6 shows how the proposed focus on ghost detections almost resolves the weak dependence between $FLAMR_{\mathcal{F}}$ to $LAMR_r$ in the foreground. Hence, the $FLAMR_{\mathcal{F}}$ effectively measures performance differently and considers factors that are ignored by the $LAMR_r$. The results show that PDs optimized for $LAMR_r$ do not necessarily perform best with respect to $FLAMR_p$ or $FLAMR_{\mathcal{F}}$. Controversially, there are PDs (with FPN-ResNet-50) that have a lower $FLAMR_{\mathcal{F}}$ score despite a much higher $LAMR_r$ score. These models have a lower miss rate for pedestrians in the foreground and produce fewer ghost detections per image. Thus, the $LAMR_r$ for the reasonable subset has limits in terms of an application-oriented evaluation. The problem arises from the training strategy for PDs that is not focused on safety. The selection of the best performing checkpoint in terms of the $LAMR_r$ is disconnected from the evaluation of safety-critical pedestrians.

Based on the large deviations between $FLAMR_{\mathcal{F}}$ to $LAMR_r$, we conclude that $FLAMR_{\mathcal{F}}$ introduces a new application-oriented perspective for the evaluation of DNNs for pedestrian detection. This conclusion seems reasonable due to the systematic categorization of errors. Here, safety-critical pedestrians are identified as the complement of highly occluded pedestrians and distant pedestrians.

5.4. Operating Point

Towards an application-oriented analysis of DNNs for pedestrian detection, we determine upper bounds on the

confidence threshold $c_{\mathcal{F}}^*$ as operating points for individual PDs. Results can be seen in Figure 7. We see that between training runs of PDs with the same feature extraction, the upper bound of the confidence threshold $c_{\mathcal{F}}^*$ and the required GDPI($c_{\mathcal{F}}^*$) vary greatly. Thus, operating points must be determined individually for the PDs and cannot be specified in general for a particular feature extraction.

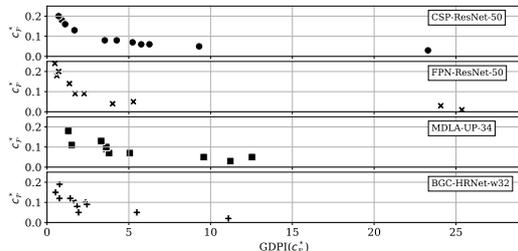


Figure 7: Upper bound for confidence threshold ($c_{\mathcal{F}}^*$) for all tested PDs with the required amount of GDPI.

Furthermore, our evaluation shows that not every PD is capable of detecting every pedestrian in the foreground with a confidence threshold of 0.01. This means that there are PDs with $MR_{\mathcal{F}}(0.01) \neq 0$ (CSP-ResNet-50: 2, FPN-ResNet-50: 4, MDLA-UP-34: 0, BGC-HRNet-w32: 1). In general, foreground pedestrians are missed with a maximum of $MR_{\mathcal{F}}(c_{\mathcal{F}}^*) = 0.29\%$. Up to this point, the subset-based evaluation of DNNs for pedestrian detection has largely focused on benchmarking. Due to the limited informative value, it was difficult to derive guidelines for the application-oriented development process of DNNs. We take the stance that aggregated performance metrics such as $FLAMR_{\mathcal{F}}^{\mathcal{A}}$ must be collated with metrics such as $MR_{\mathcal{F}}(c_{\mathcal{F}}^*)$ and $GDPI(c_{\mathcal{F}}^*)$.

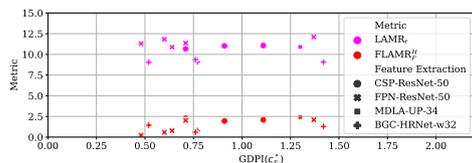


Figure 8: $LAMR_T$ and $FLAMR_P^{\mathcal{A}}$ are independent of GDPI.

Figure 8 shows that performance metrics ($LAMR_T$ and $FLAMR_P^{\mathcal{A}}$) are unrelated to the required number of ghost detections $GDPI(c_{\mathcal{F}}^*)$. Surprisingly, FPN-ResNet-50 achieves with 0.48 the lowest value of $GDPI(c_{\mathcal{F}}^*)$. The reason for the unexpected behavior can be seen in Figure 9. Although the sorted miss rate curves of the selected PDs are close in the middle range, they diverge the most in the head and tail ranges. The vertical lines mark the common values for which the miss rate is averaged. As a consequence, aggregated performance metrics such as $LAMR_T$,

$FLAMR_P$ and $FLAMR_P^{\mathcal{A}}$ average over multiple confidence thresholds and put less weight on safety-relevant ranges towards $c_{\mathcal{F}}^*$.

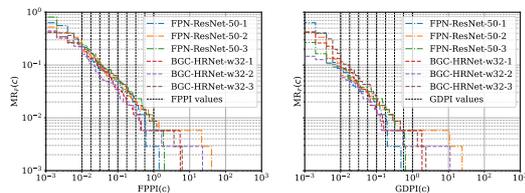


Figure 9: Comparison of selected PDs for false positives per image (FPPI, left) and ghost detections per image (GDPI, right). The filtered miss rate $MR_{\mathcal{F}}(c)$ is calculated for pedestrians in the foreground \mathcal{F} .



Figure 10: Inference results for BGC-HRNet-w32 (first run, best $LAMR_T$ score with 8.8%) and FPN-ResNet50 (first run).

Since $c_{\mathcal{F}}^*$ determines an operating point for a PD, it can serve as a meaningful confidence threshold to visually assess inference samples (see Figure 10). This provides practitioners with a reliable basis of information and allows them to evaluate applicability more intuitively.

6. Conclusion

In this work, we propose a rule-based error categorization to evaluate the performance of a DNN for pedestrian detection. Multiple disjoint categories for false negatives are defined in order to identify safety-critical errors in the foreground. The distinction is based on three occlusion-related categories and the braking distance of an automated driving system. We expect that the inclusion of depth information would improve the separation between foreground, background, occluding pedestrians and environment. In future work, we would like to reevaluate the performance of DNNs specifically designed for the occlusion problem using our proposed error categories. We identify three categories of false positives, with ghost detections being the most disruptive. For our experiments, we use a simple and generic framework to build DNNs for pedestrian detection. In consequence, we train 44 DNNs based on four commonly used backbones, achieving state-of-the-art performance in terms of $LAMR_T$. The goal of our application-oriented evaluation is two-folded. To account for safety-critical false

negatives as well as disruptive false positives, we propose FLAMR \mathcal{F} as a new performance metric. Finally, we determine an operating point as the confidence threshold where no pedestrian in the foreground is missed. By revisiting and refining the current evaluation, we contribute to a safety-focused development process of DNNs for pedestrian detection.

Acknowledgments

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project “Methoden und Maßnahmen zur Absicherung von KI basierten Wahrnehmungsfunktionen für das automatisierte Fahren (KI-Absicherung)”. The authors would like to thank the consortium for the successful cooperation. The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

References

- [1] Z.-R. Wang, Y.-L. Jia, H. Huang, S.-M. Tang, Pedestrian detection using boosted HOG features, in: IEEE Conference on Intelligent Transportation Systems (ITSC), 2008, pp. 1155–1160.
- [2] W. Liu, S. Liao, W. Ren, W. Hu, Y. Yu, High-level semantic feature detection: A new perspective for pedestrian detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5187–5196.
- [3] J. Zhang, L. Lin, J. Zhu, Y. Li, Y.-c. Chen, Y. Hu, C. S. Hoi, Attribute-aware pedestrian detection in a crowd, IEEE Transactions on Multimedia (2020).
- [4] A. H. Khan, M. Munir, L. van Elst, A. Dengel, F2dnet: Fast focal detection network for pedestrian detection, arXiv preprint arXiv:2203.02331 (2022).
- [5] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: A benchmark, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 304–311.
- [6] J. Mao, T. Xiao, Y. Jiang, Z. Cao, What can help pedestrian detection?, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6034–6043.
- [7] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, C. Shen, Repulsion loss: Detecting pedestrians in a crowd, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7774–7783.
- [8] C. Ning, L. Menglu, Y. Hao, S. Xueping, L. Yunhong, Survey of pedestrian detection with occlusion, Complex & Intelligent Systems 7 (2021) 577–587.
- [9] M. He, H. Luo, Z. Chang, B. Hui, Pedestrian detection with semantic regions of interest, Sensors 17 (2017) 2699.
- [10] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (2011) 743–761.
- [11] S. Zhang, R. Benenson, B. Schiele, Citypersons: A diverse dataset for pedestrian detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3213–3221.
- [12] P. Zhou, C. Zhou, P. Peng, J. Du, X. Sun, X. Guo, F. Huang, Noh-nms: Improving pedestrian detection by nearby objects hallucination, in: ACM International Conference on Multimedia, 2020, pp. 1967–1975.
- [13] X. Song, K. Zhao, W.-S. Chu, H. Zhang, J. Guo, Progressive refinement network for occluded pedestrian detection, in: European Conference on Computer Vision (ECCV), Springer, 2020, pp. 32–48.
- [14] Z. Xu, B. Li, Y. Yuan, A. Dang, Beta r-cnn: Looking into pedestrian detection from another perspective, Advances in Neural Information Processing Systems (NeurIPS) 33 (2020) 19953–19963.
- [15] Z. Luo, Z. Fang, S. Zheng, Y. Wang, Y. Fu, Nms-loss: Learning with non-maximum suppression for crowded pedestrian detection, in: International Conference on Multimedia Retrieval, 2021, pp. 481–485.
- [16] Z. Cai, N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6154–6162.
- [17] J. Li, S. Liao, H. Jiang, L. Shao, Box Guided Convolution for Pedestrian Detection, in: 28th ACM International Conference on Multimedia, 2020, pp. 1615–1624.
- [18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
- [19] X. Zhou, D. Wang, P. Krähenbühl, Objects as points, arXiv preprint arXiv:1904.07850 (2019).
- [20] F. Yu, D. Wang, E. Shelhamer, T. Darrell, Deep layer aggregation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2403–2412.
- [21] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5693–5703.
- [22] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization (2015).
- [23] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele,

The cityscapes dataset for semantic urban scene understanding, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3213–3223.

- [24] F. Kraus, N. Scheiner, W. Ritter, K. Dietmayer, Using machine learning to detect ghost images in automotive radar, in: IEEE Conference on Intelligent Transportation Systems (ITSC), 2020, pp. 1–7.