What Does Really Count? Estimating Relevance of Corner Cases for Semantic Segmentation in Automated Driving

Jasmin Breitenstein[1] Florian Heidecker [2] Maria Lyssenko[3] Daniel Bogdoll[4,5]
Maarten Bieshaar[6] J. Marius Zöllner[4,5] Bernhard Sick[2] Tim Fingscheidt[1]

{j.breitenstein, t.fingscheidt}@tu-bs.de
{florian.heidecker, bsick}@uni-kassel.de
{maria.lyssenko, maarten.bieshaar}@de.bosch.com
{bogdoll, zoellner}@fzi.de

[1]Technische Universität Braunschweig [2]University of Kassel [3]Robert Bosch GmbH [4]KIT
[5]FZI Research Center for Information Technology [6]Bosch Center for Artificial Intelligence

Abstract

In safety-critical applications such as automated driving, perception errors may create an imminent risk to vulnerable road users (VRU). To mitigate the occurrence of unexpected and potentially dangerous situations, so-called corner cases, perception models are trained on a huge amount of data. However, the models are typically evaluated using task-agnostic metrics, which do not reflect the severity of safety-critical misdetections. Consequently, misdetections with particular relevance for the safe driving task should entail a more severe penalty during evaluation to pinpoint corner cases in large-scale datasets. In this work, we propose a novel metric IoU_w that exploits relevance on the pixel level of the semantic segmentation output to extend the notion of the intersection over union (IoU) by emphasizing small areas of an image affected by corner cases. We (i) employ IoU_w to measure the effect of pre-defined relevance criteria on the segmentation evaluation, and (ii) use the relevance-adapted IoU_w to refine the identification of corner cases. In our experiments, we investigate vision-based relevance criteria and physical attributes as per-pixel criticality to factor in the imminent risk, showing that IoU_w precisely accentuates the relevance of corner cases.

1. Introduction

For automated vehicles, it is of utmost importance to guarantee a reliable and accurate detection of vulnerable road users (VRU). For environment perception tasks, such as object detection and semantic segmentation, deep neural networks yield state-of-the-art performance [15]. However, evaluation metrics such as mean intersection over union (mIoU) are designed to treat all pixels in the scene equally [26]. For non-safety-critical applications, task-agnostic evaluation metrics as, e.g., intersection over union (IoU) [21], may be sufficient to identify difficult detections. However, in automated driving, failed detections (false negatives) and ghost objects (false positives) not only corrupt the overall performance but relate to so-called *corner cases* (see Fig. 1). As a consequence, those relevant and non-predictable situations pose a great challenge to visual perception [6]. For example, a VRU crossing the street is hard to detect in a scene affected by heavy overexposure

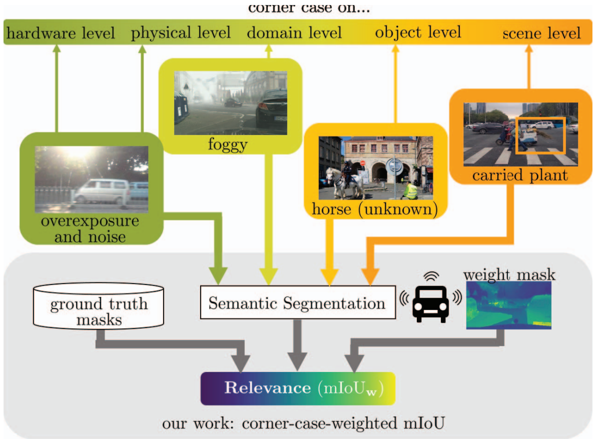


Figure 1: Our novel relevance-adapted intersection over union (IoU_w) is based on a per-pixel relevance weighting captured in an image-wise weight mask. The weight mask aggregates multiple corner case relevance criteria w.r.t. visual attributes and safety. In an offline evaluation process, these weights are applied to misdetections so that a relevance-weighted IoU (IoU_w) gives a more reliable estimate of whether corner case data is relevant for the semantic segmentation method.

or adverse weather conditions. Here, pure vision-based relevance measures for corner cases do not reflect the criticality of the situation as it depends on the associated physical properties like the distance to the ego vehicle.

As corner cases often only appear in limited portions of a scene, they reduce the misdetections' influence on agnostic evaluation measures like IoU. Therefore, to accentuate the relevance of a small pixel number in a high-resolution image, a rigorous definition of corner cases is required to pinpoint challenging scenarios. Recent works present a refined categorization of corner cases on different levels [7, 17], as exemplified in Fig. 1. Bolte et al. [6] extend the definition of detected corner cases by utilizing an object's vertical position as a proxy for distance-based relevance and thus, emphasize the necessity for more elaborate relevance measures. To account for the criticality of corner cases and to mitigate the underspecification of purely vision-based metrics [31], works by [32, 22, 3] include a notion of safety on the basis of distance and time-to-collision (TTC) to identify the safety-relevant objects for the driving task along with the respective performance assessment. Hereby, they consider the VRU's physical properties to rate whether an occurring misdetection is indeed as safety-critical corner case.

In this work, we first define relevance criteria w.r.t. *both* visual attributes and safety, bridging the gap between safety investigations and visual perception to calculate per-pixel relevance for the semantic segmentation task. Second, we propose a relevance-adapted version of the IoU to refine the identification of corner cases based on the accumulated relevance criteria in comparison with the ground truth. Our novel offline evaluation regime for the semantic segmentation task highlights the detection capabilities w.r.t. predefined relevance criteria and thus, measures the effect of occurring corner cases. The proposed methodology helps to establish a link between corner case types and perception methods, facilitating a relevance-driven data selection for training semantic segmentation models.

The remainder of the paper is structured as follows: Sec. 2 discusses related works. Sec. 3 introduces our definition of the relevance-adapted (weighted) IoU for semantic segmentation. Sec. 4 outlines the offline evaluation procedure for the relevance-adapted IoU using estimated corner case weights, and Sec. 5 presents our experimental results.

2. Related Work

2.1. Corner cases in automated driving

The first definition of a corner case for automated driving was phrased by Bolte et al. [6]: "A corner case is given, if there is a non-predictable relevant object/class in relevant location". The authors presented a corner case detection system using relevance as a filtering criterion. As this approach lacked a specific categorization, Breitenstein et al. [7] provided a systematization of corner cases for the camera sensor consisting of multiple levels. Their systematization describes *external* corner cases occurring in the real world. Based on this, they were able to show a strong link to corner case detection approaches [8], bridging the gap between the theoretical taxonomy and practical methods. Heidecker et al. [17] extended the systematization by not only including LiDAR and RaDAR data, but also by simplifying it and adding a new layer for *internal* corner cases introduced by malfunctions in the software, such as false negatives. Bogdoll et al. [4, 5] have shown the applicability of this taxonomy for the description and generation of scenarios, demonstrating its maturity. As our work builds heavily upon the systematization by Heidecker et al. [17], we hereby provide a short overview:

- **Sensor layer:** Corner case introduced by sensor attributes on either the *hardware* or *physical* level.
- **Content layer:** Corner case on a single-frame on either the *domain*, *object*, or *scene* level. Respective examples are changing weather, unknown objects, and known objects in unusual locations or quantities.
- **Temporal layer:** Corner case only detectable on multiple frames, i.e., *scenarios*, with a focus on behavior.

While previous works focused on knowledge-driven corner case types based on scene content and physical acquisition properties and have linked those types of corner cases with sensor modalities in autonomous driving, the notion of *relevance* of corner cases on a perception function has to our knowledge not yet been explored beyond the original concept of Bolte et al. [6]. In our work, we aim to provide a computational model of such relevance by introducing a weighted variant of the mIoU metric, which reflects the relevance of corner cases for the scene evaluation.

2.2. Criticality measures in automated driving

Gannamaneni et al. [16] and Lyssenko et al. [21] leverage synthetic data to show that standard performance measures such as IoU are insensitive to important physical cues such as distance. However, sole distance measures do not consider the velocity or direction of motion of the VRU needed to adequately assess the criticality of the interaction. To account for criticality, Wolf et al. [32] and Bansal et al. [3] introduced risk-aware recall versions on the basis of the direction of motion and distance, respectively. In contrast to aggregated safety-aware performance evaluation, potentially critical interactions between automated vehicles and individual pedestrian sequences can be derived by a reachability-based method in combination with TTC [22]. For semantic segmentation, per-pixel TTC estimation by Badki et al. [2] leveraged consecutive frames from a mono camera to calculate a binary time-to-contact (the time for an object to collide with the observer's plane) via a series of binary per-pixel classifications to estimate whether a vehicle may collide with the camera plane.

Varghese et al. [28] incorporated optical flow to measure the stability of semantic segmentation between consecutive frames, leading to a critical temporal consistency metric to identify unstable predictions during offline evaluation. We follow Varghese et al. [28] by combining safety investigations with visual attributes for our metric in an offline evaluation setting. However, instead of identifying unstable semantic segmentation predictions, we use them to accentuate scene parts affected by corner cases to determine their relevance for semantic segmentation. The weighting of mIoU has been investigated for improved training of object detectors by using the weighted mIoU inside a pixel-position-aware loss function [27, 30] and to weight the influence of semantic classes based on their frequency of appearance [11, 35, 1]. While we follow the intuition of weighting the mIoU, we design our weighting to estimate the relevance of corner cases using multiple corner case criteria instead of just relying on class frequency or position awareness.

3. Relevance Measure for Corner Cases in Semantic Segmentation

The mIoU is a widely used metric for evaluating semantic segmentation, considering the overlap between the ground truth semantic segmentation mask $\overline{\mathbf{m}} \in \mathcal{S}^{H \times W}$ and the predicted semantic segmentation mask $\mathbf{m} \in \mathcal{S}^{H \times W}$, where $\mathcal{S} = \{1, \dots, S\}$ is the set of the S semantic classes, and $H,\ W \in \mathbb{N}$ are the height and width of the masks, respectively. The mIoU calculates the mean of the class-specific intersection over union (IoU) values according to:

$$\text{mIoU} = \frac{1}{S} \sum_{s \in \mathcal{S}} \text{IoU}(s). \tag{1}$$

Most often, the intersection over union (IoU) is described by the true positive (TP), false positive (FP), and false negative (FN) cases or by using the Jaccard score [19]. Here, the overlapping area between the one-hot encoded ground truth per class s, i.e., $\overline{\mathbf{y}}_s \in \{0, 1\}^{H \times W}$, and the one-hot encoded prediction per class s, i.e., $\mathbf{y}_s = (y_{i,s})_{i \in \mathcal{I}} \in \{0, 1\}^{H \times W}$, is calculated and normalized by the area of the union:

$$\text{IoU}(s) = \frac{\sum\limits_{i \in \mathcal{I}} \text{TP}(s, i)}{\sum\limits_{i \in \mathcal{I}} (\text{TP}(s, i) + \text{FP}(s, i) + \text{FN}(s, i))} \tag{2}$$

$$= \frac{|\overline{\mathbf{y}}_s \cap \mathbf{y}_s|}{|\overline{\mathbf{y}}_s \cup \mathbf{y}_s|} = \frac{|\overline{\mathbf{y}}_s \cap \mathbf{y}_s|}{|\overline{\mathbf{y}}_s \cap \mathbf{y}_s| + |\overline{\mathbf{y}}_s \triangle \mathbf{y}_s|}, \tag{3}$$

where $i \in \mathcal{I} = \{1, \dots, H \cdot W\}$ denotes the pixel index for the semantic class s. We split the area of union $\overline{\mathbf{y}}_s \cup \mathbf{y}_s = \overline{\mathbf{y}}_s \cap \mathbf{y}_s + \overline{\mathbf{y}}_s \triangle \mathbf{y}_s$ into two parts, where $\overline{\mathbf{y}}_s \cap \mathbf{y}_s$ is the intersection of both areas as in the numerator and $\triangle$ denotes the symmetric difference, i.e., the union of the non-intersecting areas of $\mathbf{y}_s$ and $\overline{\mathbf{y}}_s$.

In the mIoU equation (1), each classified pixel in the image is given equal weight regardless of the class, position, or relevance. This does not pose a problem for evaluating semantic segmentation models in non-safety-critical domains. However, in safety-critical systems, some areas in the image are too dominantly considered in the metric, while others play a too-small role. For instance, if safety aspects are important and corner cases are estimated according to their criticality, the pixels of the semantic mask must be weighted differently in the metric. We use the Jaccard score formula (2) to compute the *weighted* IoU

$$\text{IoU}_{\mathbf{w}}(s) = \frac{\sum\limits_{i \in \mathcal{I}} \text{TP}(i, s)}{\sum\limits_{i \in \mathcal{I}} \text{TP}(i, s) + (\text{FP}(i, s) + \text{FN}(i, s)) \cdot w(i)}. \tag{4}$$

Weighting the intersection $\overline{\mathbf{y}}_s \cap \mathbf{y}_s$ (note that $|\overline{\mathbf{y}}_s \cap \mathbf{y}_s| = \sum_{i \in \mathcal{I}} \text{TP}(i, s)$) would not be beneficial as TP pixels do not impose a risk (cf. (2)). Therefore, we focus on the symmetric difference $\overline{\mathbf{y}}_s \triangle \mathbf{y}_s$, where $|\overline{\mathbf{y}}_s \triangle \mathbf{y}_s| = \sum_{i \in \mathcal{I}} \text{FP}(i, s) + \text{FN}(i, s)$ counts the segmentation errors of class s pixels in the ground truth. The weighted errors of class s in pixel i are denoted as $E_{\mathbf{w}}(i, s) = (\text{FP}(i, s) + \text{FN}(i, s)) \cdot w(i)$.

For each pixel i in (4), we multiply the corresponding errors $\text{FP}(i, s) + \text{FN}(i, s)$ by the pixel-wise weight, where:

$$w(i) = \frac{1}{N} \sum_{n \in \mathcal{N}} \lambda_n \omega_n(i). \tag{5}$$

To obtain $w(i)$, we consider N corner case criteria weights $\boldsymbol{\omega}_n = (\omega_n(i))_{i \in \mathcal{I}} \in [0, 2]^{H \times W}, n \in \mathcal{N} = \{1, \dots, N\}$, designed to estimate corner case relevance. The weight range $\omega_n(i) \in [0, 2]$ ensures that relevant pixels i ($\omega_n(i) > 1$) lead to a larger $E_{\mathbf{w}}(i, s)$, decreasing $\text{IoU}_{\mathbf{w}}$, and irrelevant pixels i ($\omega_n(i) < 1$) lead to a lower $E_{\mathbf{w}}(i, s)$, increasing $\text{IoU}_{\mathbf{w}}$ in (4). As corner case detection is a highly complex topic, we do not rely on a single weighting procedure to estimate relevance. Instead, we use the aggregation of multiple criteria weights $\omega_n(i), n \in \mathcal{N}$, to benefit from different aspects in each criterion. The N different corner case criteria providing the corresponding weights $\boldsymbol{\omega}_n$ are introduced in detail in Section 4.2. In (5), we apply a weighting factor $\lambda_n > 0$ to emphasize the relative importance of each corner case criterion similar to [30], where we choose $\frac{1}{N} \sum_{n \in \mathcal{N}} \lambda_n = 2$. Our $\text{IoU}_{\mathbf{w}}$ equals IoU if $w(i) = 1$ for pixel indices $i \in \mathcal{I}$. This can be obtained, e.g., by $\omega_n(i) = 1/2$ for all pixel $i \in \mathcal{I}$ and criteria $n \in \mathcal{N}$. The resulting weights $w(i)$ in (5) form the final weight mask $\mathbf{w} = (w(i))_{i \in \mathcal{I}} \in [0, 4N]^{H \times W}$ combining N corner case criteria.

Please note that the integration of the weight mask $\mathbf{w}$ in the *weighted IoU* in (4) keeps the same range $\text{IoU}_{\mathbf{w}} \in [0, 1]$,

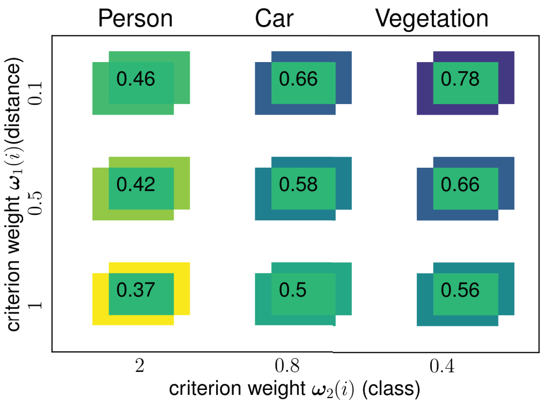


Figure 2: Nine examples of the same semantic masks and the resulting $\mathrm{IoU}_{\mathbf{w}}$ (9 values) under the influence of two corner case criteria weights: distance to the camera (distance weight) $\boldsymbol{\omega}_1(i)$ and class vulnerability (class weight) $\boldsymbol{\omega}_2(i)$. Each example consists of two rectangles representing the ground truth and predicted semantic mask. The intersection is always shown in the same green color, and the weighted error area ($E_{\mathbf{w}}(i,s)$), is colored depending on the weight $\mathbf{w}$, calculated according to (5). For simplicity, in each example, we assign the same weight $w(i)$ to each pixel i.

as $(\mathrm{FP}(i,s)+\mathrm{FN}(i,s))\cdot w(i) \geq 0$ for $w(i) \in [0,4N]$. The mean $\mathrm{IoU}_{\mathbf{w}}$ ($\mathrm{mIoU}_{\mathbf{w}}$) is obtained in analogy to (1).

We visualize the benefit of $\mathrm{IoU}_{\mathbf{w}}$ in a toy example in Fig. 2, where we present nine identical examples and show how relevance-driven weighting can be beneficial for the evaluation. We can see an equal symmetric difference (and overlap) for each of the nine examples, i.e., equal IoU scores (0.47). However, two relevance criteria w.r.t. corner cases govern the horizontal and vertical axis: distance to the camera $\boldsymbol{\omega}_1(i)$ and severity of misclassification of a semantic class $\boldsymbol{\omega}_2(i)$, i.e., person, car, and vegetation. Each criterion assigns a weight to each example in Fig. 2. We obtain the weight $\mathbf{w}$ according to (5) to calculate $\mathrm{IoU}_{\mathbf{w}}$, setting $\lambda_1 = 1, \lambda_2 = 1$. For each example, we weight each pixel i in $E_{\mathbf{w}}(i,s)$ using the same weights $\omega_1(i)$ and $\omega_2(i)$, e.g., for the example in the bottom left of Fig. 2, $\omega_1(i) = 1$ and $\omega_2(i) = 2$ for all pixels $i \in \mathcal{I}$. The values for $\omega_1(i)$ and $\omega_2(i)$ are denoted on the left and lower axes of Fig. 2.

The first column represents the class "person". As we can see, with increasing distance weight $\boldsymbol{\omega}_1(i)$, $\mathrm{IoU}_{\mathbf{w}}$ decreases from 0.46 down to 0.37. The most safety-relevant case for the class "person" has the highest value ($\boldsymbol{\omega}_2(i)$=2) in this toy example. Our corner case criteria for evaluation are introduced in Sec. 4.2. Similarly, closer distance to the ego-vehicle corresponds to a higher distance weight $\boldsymbol{\omega}_1(i)$ and thus higher criticality for the person class, i.e., this leads to a stronger penalty on FP and FN pixels.

The same behavior can be seen in the other two columns (representing classes "car" and "vegetation") of Fig 2. Depending on the relevance of the class, i.e., class weight $\boldsymbol{\omega}_2(i)$, we can see that misdetection at the same distance criticality obtain different $\mathrm{IoU}_{\mathbf{w}}$ scores. Thus, VRUs are treated with higher priority in this evaluation than, e.g., vegetation. Further, using TTC as an auxiliary weight, we are able to extend the pure distance-based criticality weighting by velocity information to account for criticality (see Sec. 4). Consequently, $\mathrm{IoU}_{\mathbf{w}}$ would denote greater changes to the generic IoU for dynamic objects, e.g., cars.

4. Experimental Setup

4.1. Implementation Details

We use the Cityscapes dataset $\mathcal{D}_{\mathrm{CS}}$ [13] and train our semantic segmentation model detailed in Section 4.1 on the training dataset $\mathcal{D}_{\mathrm{CS}}^{\mathrm{train}}$. We use the validation dataset $\mathcal{D}_{\mathrm{CS}}^{\mathrm{val}}$ for evaluation. Furthermore, we define a subset $\mathcal{D}_{\mathrm{CS}}^{\mathrm{cc}}$ of $\mathcal{D}_{\mathrm{CS}}^{\mathrm{val}}$ curated manually to fit the definitions of corner case types [17] consisting of three examples per corner case type. Some examples of our corner case dataset $\mathcal{D}_{\mathrm{CS}}^{\mathrm{cc}}$ are shown in Fig. 8. As we investigate image-based semantic segmentation, corner case types defined on the temporal axis, i.e., scenario-level corner cases, were not included.

We use `OCRNet` [34] for semantic segmentation in Fig. 3. It trained on $\mathcal{D}_{\mathrm{CS}}^{\mathrm{train}}$ using the implementation provided by the `mmsemgmentation` repository [12]. Our trained network reaches mIoU$= 80.23\%$ on $\mathcal{D}_{\mathrm{CS}}^{\mathrm{val}}$ (cf. Table 1).

4.2. Corner Case Criteria

Our proposed relevance estimate for corner cases in semantic segmentation is based on criteria defined based on corner case types in [17]. We first discuss the proposed evaluation procedure shown in Fig. 3. Then, we detail how we obtain each criterion $\boldsymbol{\omega}_n$ in (5).

Fig. 3 shows the proposed offline evaluation procedure to obtain $\mathrm{mIoU}_{\mathbf{w}}$ for an input image $\mathbf{x}_t = (x_{t,i}) \in [0,1]^{H\times W\times C}$ of given data, where $t \in \mathcal{T} = \{1,\ldots,T\}$ is the index of the image in the data with T image frames and $i \in \mathcal{I} = \{1,\ldots,H\cdot W\}$ denotes the pixel position in the image with height H and width W. The number of channels is denoted by C, where, for RGB images, $C = 3$. The semantic segmentation method outputs the corresponding semantic segmentation mask $\mathbf{m}_t \in \mathcal{S}^{H\times W}$ with $\mathcal{S} = \{1,\ldots,S\}$ being the set of S semantic classes. From the semantic segmentation mask $\mathbf{m}_t$ and the corresponding ground truth $\overline{\mathbf{m}}_t \in \mathcal{S}^{H\times W}$, we extract our pre-defined N corner case relevance criteria $\boldsymbol{\omega}_{t,n}$ for $n \in \{1,\ldots,N\}$, which will be detailed in the following. These criteria are then accumulated to the weight mask $\mathbf{w}_t \in [0,4N]^{H\times W}$ according to (5). In the performance evaluation, we calculate $\mathrm{mIoU}_{\mathbf{w}}$ according to (4). Next, we describe how to obtain the various criteria.

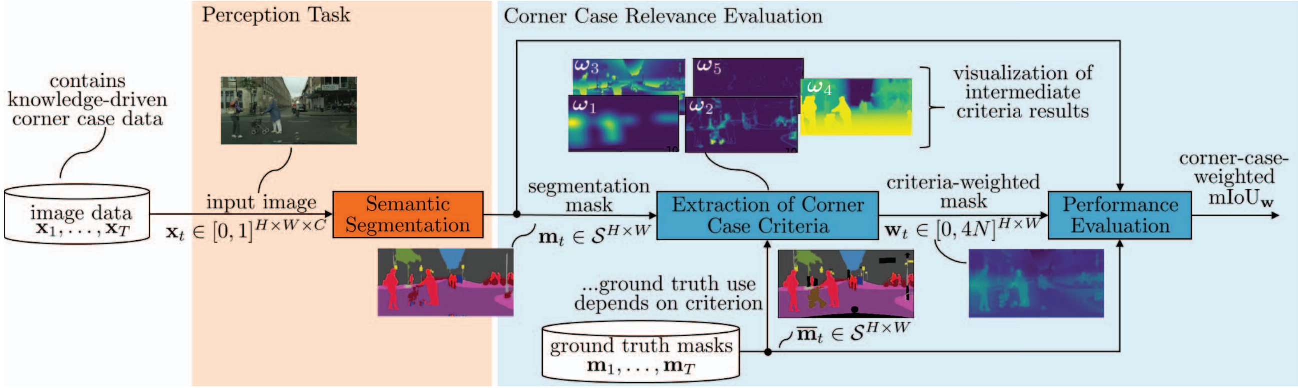

Perception Task
Corner Case Relevance Evaluation
contains knowledge-driven corner case data
image data $\mathbf{x}_1, \dots, \mathbf{x}_T$
input image
$\mathbf{x}_t \in [0,1]^{H\times W\times C}$
Semantic Segmentation
segmentation mask
$\mathbf{m}_t \in \mathcal{S}^{H\times W}$
ω_3
ω_5
ω_4
ω_1
ω_2
visualization of intermediate criteria results
Extraction of Corner Case Criteria
criteria-weighted mask
$\mathbf{w}_t \in [0,4N]^{H\times W}$
Performance Evaluation
corner-case-weighted mIoU$_\mathbf{w}$
...ground truth use depends on criterion
ground truth masks $\mathbf{m}_1, \dots, \mathbf{m}_T$
$\overline{\mathbf{m}}_t \in \mathcal{S}^{H\times W}$

Figure 3: Our proposed evaluation procedure to measure the weighted mIoU (mIoU$_\mathbf{w}$) of given data $\mathbf{x}_1, \dots, \mathbf{x}_T$. We obtain the corresponding semantic segmentation mask $\mathbf{m}_t$ for a given input image $\mathbf{x}_t$. Then, we calculate our corner case criteria, which comprehensively form the weight mask $\mathbf{w}_t$ (5). We use the weight mask to calculate the proposed mIoU$_\mathbf{w}$.

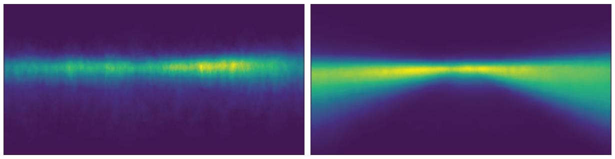


Figure 4: **Location priors** calculated on the Cityscapes training dataset $\mathcal{D}_{\text{CS}}^{\text{train}}$ for the classes "person" (left) and "car" (right).

4.2.1 Crowdedness of the scene

Crowded areas, i.e., with semantic classes person and rider, are more prone to errors and/or safety issues but also generally pose a higher risk for corner cases. Thus, areas with higher crowdedness are assigned higher weights, i.e., a higher relevance. To do so, we consider pixel i in the semantic mask $\mathbf{m}$ and a neighborhood of an experimentally chosen fixed size of 128×256. We count the number of VRU appearances in this neighborhood of a pixel i. By repeating the procedure for all pixels, we obtain the criterion weight $\boldsymbol{\omega}_1$, which we normalize such that $\boldsymbol{\omega}_1 \in [0,2]^{H\times W}$. This emphasizes crowded areas, where $\omega_1(i) > 1/2$ puts more emphasis on pixel i, $\omega_1(i) = 1/2$ for few VRUs, and $\omega_1(i) < 1/2$ describes areas mostly without VRUs, leading to less relevant pixels i.

4.2.2 Confidence

This criterion applies the maximum softmax method [18]. Instead of the usual argmax function, the maximum softmax method extracts the maximum of softmax probabilities to obtain the probabilities associated with the predicted classes s, i.e.,

$$\widetilde{\mathbf{y}}_t = (\widetilde{y}_{t,i}) = \left(\max_{s\in\mathcal{S}} y_{t,i} \right). \tag{6}$$

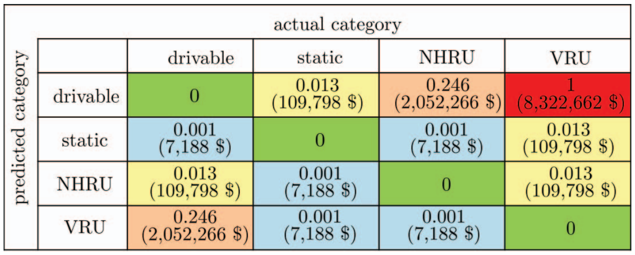
According to [18], we interpret low maximum softmax probabilities as uncertainties of the semantic segmentation network, and we use the inverse of the maximum softmax probabilities as the criterion for the weight mask, i.e., $\omega_2(i) = 1 - \widetilde{y}_{t,i}$. The inverse ensures that higher values in the weight mask $\boldsymbol{\omega}_2$ are related to higher uncertainties. We normalize so that $\omega_2(i) \in [0,2]$ to ensure that high uncertainty lowers IoU$_\mathbf{w}$ to show relevance, and low values increase IoU$_\mathbf{w}$ as errors are, in this case, less relevant.

4.2.3 Spatial diversity

Intuitively, we are interested in predicted classes in unusual locations to detect certain corner case types, e.g., scene-level corner cases. For this, we use location priors for each class calculated on $\mathcal{D}_{\text{CS}}^{\text{train}}$, which we assume to contain mostly normal data and a negligible amount of corner cases. These location priors then contain the information where each class is typically located [29]. We calculate a location prior per semantic class in the training data $\mathbf{x}_t$ for $t \in \mathcal{T}_{\text{train}} = \{1, \dots, T_{\text{train}}\}$ with T_{train} being the amount of training data, by

$$P(i|s) = \frac{\sum\limits_{t\in\mathcal{T}_{\text{train}}} \overline{y}_{t,i,s}}{\max\limits_{j\in\mathcal{I}} \sum\limits_{t\in\mathcal{T}_{\text{train}}} \overline{y}_{t,j,s}} \in [0,1], \tag{7}$$

Fig. 4 shows two example location priors for the classes "person" and "car". Persons are mostly observed in the background and slightly more frequently on the right-hand side of the ego-vehicle (probably due to right-hand traffic in the recorded areas, i.e., sidewalks to the right of the ego-vehicle). Cars are observed left and right of the ego-vehicle, but typically not directly in front of it. During inference, we obtain the criterion weight $\boldsymbol{\omega}_3$ from the location prior corresponding to the predicted semantic class s^*, i.e., $\omega_3(i) = 1 - P(i|s^*)$. For ease of notation, we omit the



		actual category			
		drivable	static	NHRU	VRU
predicted category	drivable	0	0.013 (109,798 $)	0.246 (2,052,266 $)	1 (8,322,662 $)
	static	0.001 (7,188 $)	0	0.001 (7,188 $)	0.013 (109,798 $)
	NHRU	0.013 (109,798 $)	0.001 (7,188 $)	0	0.013 (109,798 $)
	VRU	0.246 (2,052,266 $)	0.001 (7,188 $)	0.001 (7,188 $)	0

Figure 5: Our **cost matrix** for four categories of semantic classes: drivable, static, non-human road users (NHRU), and VRU. We weight misclassifications based on their risk for accidents associated with monetary costs (in $) shown in the cost matrix together with the associated cost c.

dependence of $\omega_3(i)$ on $s = s^*$. To denote relevance, we normalize to $\omega_3(i) \in [0, 2]$. Note that $\omega_3(i) = 1/2$ if the predicted class s appears in green colored areas of Fig. 4.

4.2.4 Time-to-collision

The Cityscapes dataset also provides depth ground truth data. Without the given depth information, the calculation of TTC is an ongoing research field [2, 33, 20]. Since depth information and camera intrinsics are given, we over-approximate TTC[1] while assuming a constant velocity[2] of 50 km/h. For safety considerations, we assume a perception reaction time of 2.5 s. This results in a critical distance of up to 60 meters. We obtain the TTC by normalizing the ground truth depth map of the Cityscapes dataset by the critical distance. We use the inverse of the normalized TTC per pixel as an additional weight $\boldsymbol{\omega}_4$, i.e., $\omega_4(i) = 1 - \frac{TTC(i)}{\max\limits_{j\in\mathcal{I}} TTC(j)}$, driven by the assumption that perception errors far in the background are irrelevant to the driving task and that perception errors become highly relevant when they appear closer to the camera, i.e., within the critical distance. We normalize to $\omega_4(i) \in [0, 2]$, where $\omega_4(i) = 1/2$ if i is approximately at 3/4 of the stopping distance. Using TTC instead of pure distance-based criticality weightings, we also include velocity information that greatly impacts the resulting $\text{mIoU}_\mathbf{w}$ for dynamic objects.

4.2.5 Weighted misclassification

We distinguish between different types of misclassifications: We consider four categories of semantic class types, i.e., `drivable`, `static`, `VRU`, and `non-human road users`, and we categorize the semantic classes of our dataset into those four categories. We develop a cost matrix to punish misclassifications between the four categories. Our costs c are based on the severity of injuries based on the maximal abbreviated injury scale (MAIS). It ranges from no injury (MAIS0) to critical injuries (MAIS5). We include two additional cases: property damage only and fatal injury, and relate the scales with real costs associated with those types of accidents [24, 25, 14]. The misclassification of VRUs as drivable area can have the most serious consequences, i.e., fatal, MAIS5, and MAIS4. The cost matrix in Fig. 5 shows our derived costs c and corresponding monetary costs. The second highest cost (orange in Fig. 5) is associated with misclassifications that can lead to serious injuries, i.e., classifying a non-human road user such as a car as a drivable area. We obtain the associated criterion weight $\boldsymbol{\omega}_5$ by $\omega_5(i) = 1/2 + c$ with the cost c provided by the cost matrix in Fig. 5. Misclassifications within the same category lead to $\omega_5(i) = 1/2$.

While we produce a cost matrix sufficient for our task and could, e.g., also be applied for Bayesian risk minimization in semantic segmentation, it should be noted that a generally accepted and universally valid cost matrix is an active research topic [10, 9], and not the aim of this work.

	$\mathcal{D}_{\text{CS}}^{\text{val}}$	$\mathcal{D}_{\text{CS}}^{\text{cc}}$
mIoU	80.23%	74.12%
$\text{mIoU}_\mathbf{w}$	80.35%	72.33%

Table 1: Results w.r.t. mIoU and weighted mIoU ($\text{mIoU}_\mathbf{w}$) on the considered datasets, i.e., $\mathcal{D}_{\text{CS}}^{\text{val}}$ and $\mathcal{D}_{\text{CS}}^{\text{cc}}$.

5. Experiments and Discussion

In this section, we discuss the experiments based on our proposed evaluation procedure in Fig. 3. We evaluate our results in terms of mIoU and $\text{mIoU}_\mathbf{w}$ to show the benefit of the proposed weighting to evaluate data w.r.t. corner cases.

While $\text{mIoU}_\mathbf{w}$ has more exploitable properties as an image-wise evaluation measure, we first show results on the respective datasets in Table 1. We see that on $\mathcal{D}_{\text{CS}}^{\text{val}}$ there is just a small increase from mIoU (80.23%) to $\text{mIoU}_\mathbf{w}$ (80.35%). As expected, on $\mathcal{D}_{\text{CS}}^{\text{cc}}$ mIoU (74.12%) decreases to an $\text{mIoU}_\mathbf{w} = 72.33\%$, since by design of $\text{mIoU}_\mathbf{w}$, not well-segmented areas in the presence of corner cases are punished. Since $\mathcal{D}_{\text{CS}}^{\text{cc}}$ is designed to contain all corner case types, we expect errors that are punished by $\text{mIoU}_\mathbf{w}$.

Fig. 6 shows the effect of various criteria weights $\boldsymbol{\omega}_n$ on the weight mask $\mathbf{w}$. In the first row, we see the input image $\mathbf{x}_t$, the ground truth $\overline{\mathbf{m}}_t$, the segmentation prediction $\mathbf{m}_t$, and the weight mask $\mathbf{w}_t$. In the bottom row, we show (only) the four $\boldsymbol{\omega}_n$ with the strongest influence on $\mathbf{w}_t$ for this input image. The input image is atypical for $\mathcal{D}_{\text{CS}}$, with no road straight ahead, many people close to the ego-vehicle, and a ship in the background. The obtained weight mask $\mathbf{w}_t$ accentuates the relatively small area ahead where

[1] We define TTC similar to Badki et al. [2], i.e., as the time for an object to collide with the observer's plane.

[2] Our speed assumption is based on standard German speed limits, which for larger urban streets is often set at 50 km/h (approx. 31 mph).

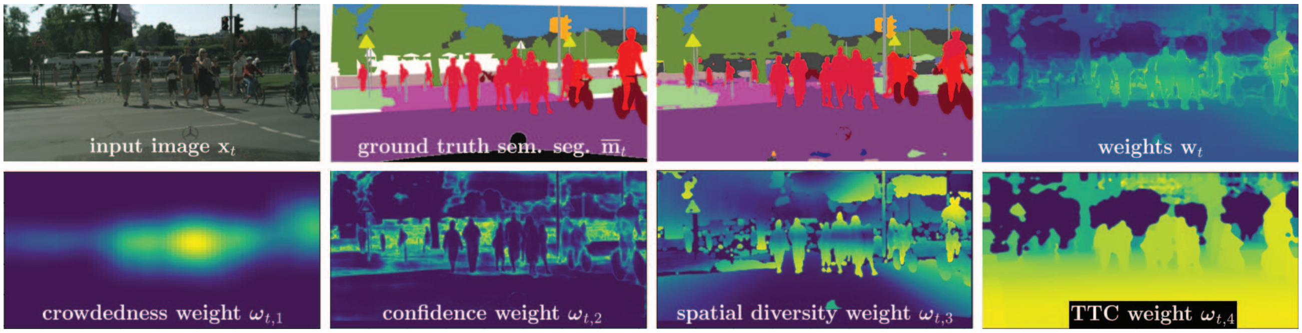

input image $\mathbf{x}_t$
ground truth sem. seg. $\overline{\mathbf{m}}_t$
weights $\mathbf{w}_t$
crowdedness weight $\omega_{t,1}$
confidence weight $\omega_{t,2}$
spatial diversity weight $\omega_{t,3}$
TTC weight $\omega_{t,4}$

Figure 6: Input image $\mathbf{x}_t$, corresponding ground truth segmentation mask $\overline{\mathbf{m}}_t$, segmentation mask $\mathbf{m}_t$ and weights $\mathbf{w}_t$ to calculate mIoU$_\mathbf{w}$ are shown in the top row. In the bottom row, we show the criteria weights $\boldsymbol{\omega}_{t,1}, \ldots, \boldsymbol{\omega}_{t,4}$, which have the highest effect on the weights $\mathbf{w}_t$. Evaluation provides mIoU$_\mathbf{w}$ $= 57.03\%$ and mIoU$= 63.29\%$.

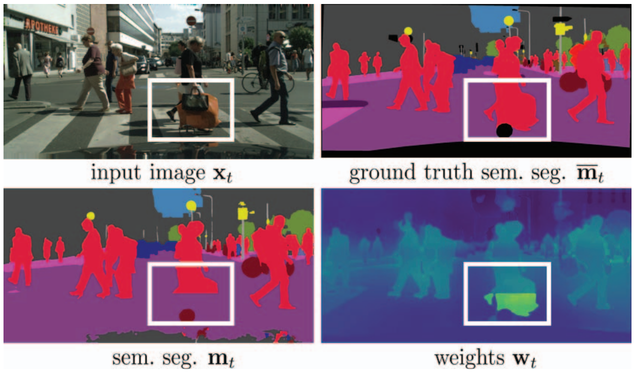

input image $\mathbf{x}_t$
ground truth sem. seg. $\overline{\mathbf{m}}_t$
sem. seg. $\mathbf{m}_t$
weights $\mathbf{w}_t$

Figure 7: Input image $\mathbf{x}_t$ and ground truth $\overline{\mathbf{m}}_t$. We simulate a safety-critical situation by replacing parts of a human in the output segmentation $\mathbf{m}_t$ (bottom left) by the class "road", as can be seen in the white box. The mIoU for this altered prediction $\mathbf{m}_t$ is 60.67%, while mIoU$_\mathbf{w}$ $= 56.03\%$ represents more accurately the safety-critical prediction in the semantic segmentation. Its weights $\mathbf{w}_t$ (bottom right) highlight the areas affected by the corner case.

crowdedness weights ($\boldsymbol{\omega}_1$) reflect many person pixels, TTC ($\boldsymbol{\omega}_4$) weights a small distance to the ego-vehicle, spatial diversity ($\boldsymbol{\omega}_3$) weights the VRU classes close by as these are unusual positions for those classes (see Fig. 4) and the sidewalk. Confidence ($\boldsymbol{\omega}_2$) accentuates the background with the ship, walls, and vegetation. As the parts highlighted by our weights $\mathbf{w}_t$ are not well segmented in $\mathbf{m}_t$, we obtain mIoU$_\mathbf{w}$ $= 57.03\%$ compared to mIoU$= 63.29\%$.

Fig. 7 shows the need for relevance-weighting using an example of a safety-related corner case. On top, Fig. 7 shows the input image $\mathbf{x}_t$ and the ground truth $\overline{\mathbf{m}}_t$. In the corresponding semantic segmentation $\mathbf{m}_t$, we simulate a safety-critical situation: Inside the white frame, we replace pixels of a person with road pixel labels. This is critical since due to the higher base point, the respective person might be located at a larger distance to the ego-vehicle. The mIoU for the altered prediction is 60.67%, which is close to the average mIoU per image on the Cityscapes validation data $\mathcal{D}_{\mathrm{CS}}^{\mathrm{val}}$ (61.47%), which we use as reference for normal data, i.e., not safety-critical. The mIoU is not affected by the safety-relevant misclassification, as small local errors have little effect on this global measure. However, mIoU$_\mathbf{w}$ drops to 56.03%. This reduced value highlights the impact of misclassified, safety-relevant pixels (such as person pixels) on semantic segmentation. The light-colored weights $\mathbf{w}_t$ show that the areas affected by corner cases are emphasized.

Corner cases on...	mIoU	mIoU$_\mathbf{w}$
hardware	79.48%	78.98%
physical	78.34%	79.14%
domain	54.06%	48.98%
object	80.12%	80.46%
collective	78.01%	75.71%
contextual	76.21%	74.92%

Table 2: **Results** in mIoU and mIoU$_\mathbf{w}$ on $\mathcal{D}_{\mathrm{CS}}^{\mathrm{cc}}$ divided into the **different corner case levels** [17]. If mIoU$_\mathbf{w}$ $<$ mIoU, the semantic segmentation is much more affected by that corner case type thereby making it relevant.

As we propose mIoU$_\mathbf{w}$ for corner cases, Table 2 shows the mIoU and mIoU$_\mathbf{w}$ results on $\mathcal{D}_{\mathrm{CS}}^{\mathrm{cc}}$ for six corner case levels. We see that corner cases on physical level lead to improved results, i.e., do not affect the semantic segmentation method much (mIoU 78.34% vs. mIoU$_\mathbf{w}$ 79.14%). The same holds for the object level, where mIoU$_\mathbf{w}$ increases mIoU by 0.34% absolute. However, we observe small decreases for corner cases on a hardware and contextual level ($0.5\%/1.29\%$ absolute, respectively. Larger negative deviations can be observed for collective corner cases (2.3% absolute) and domain-level corner cases (5.08% absolute), making both corner cases especially relevant for our semantic segmentation method. While domain-level corner cases also have a large impact on the entire image [23], the deviations for collective corner cases only impact parts of the image and are strongly emphasized by the mIoU$_\mathbf{w}$.

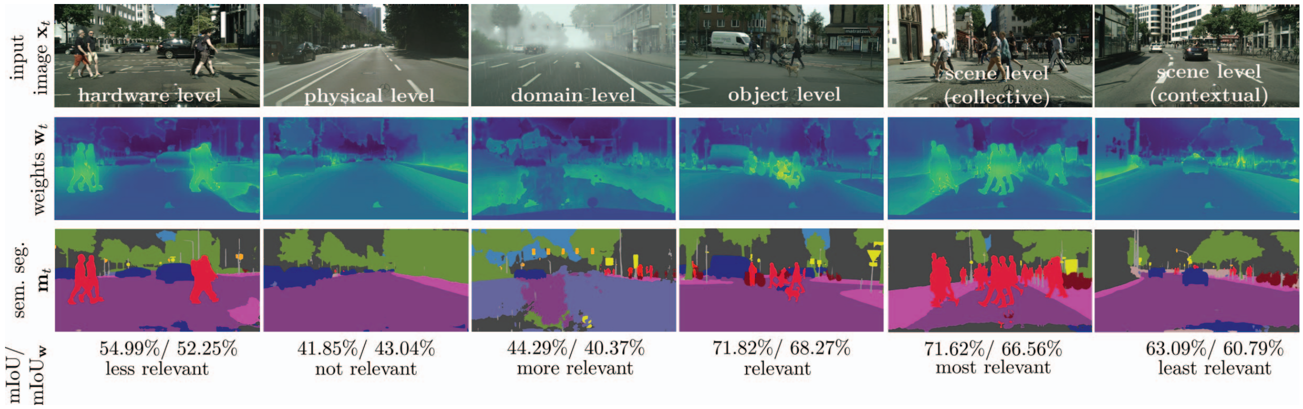


Figure 8: **Example images** $\mathbf{x}_t$ of **corner case** types with weights $\mathbf{w}_t$ (higher values are visualized in brighter yellow) and segmentation masks $\mathbf{m}_t$. For each image, the corner case type is shown as well as mIoU (1) and mIoU$_\mathbf{w}$ (4).

In Fig. 8, we show six images corresponding to different corner case types, i.e., hardware, physical, domain, object, and two times scene level. For each image, we show the weight $\mathbf{w}_t$ as well as the semantic segmentation $\mathbf{m}_t$. We report both mIoU and mIoU$_\mathbf{w}$ per input corner case. The example scene-level (collective) corner case has the strongest influence on the segmentation method, where the discrepancy between mIoU and mIoU$_\mathbf{w}$ is 5.06% absolute, showcasing that the semantic segmentation method is highly vulnerable to the collective scene-level corner case. Similarly, high reductions can be witnessed for the example domain-level corner case, where the discrepancy between mIoU and mIoU$_\mathbf{w}$ is 3.92% absolute, i.e., the domain-level corner case is also relevant for the segmentation method.

The example object-level corner case leads to a decrease by 3.55% absolute from mIoU to mIoU$_\mathbf{w}$ as the crowded scene containing a dog and a stroller is highlighted by the weights $\mathbf{w}_t$. This area contains many FP and FN, which are consequently weighted higher in $E_\mathbf{w}(i, s)$. Since no hardware-level corner cases are available in $\mathcal{D}_{\text{CS}}$, we simulated motion blur for the hardware-level corner case. Interestingly, the motion blur leads to misclassifications between the pedestrians and the street which can be observed as bright pixels in $\mathbf{w}_t$ (best seen digitally). These misclassifications cause a drop from mIoU$= 54.99\%$ to mIoU$_\mathbf{w} = 52.25\%$. On scene level (contextual), mIoU$_\mathbf{w}$ decreases 2.3% absolute compared to the mIoU. In the image, a person is standing inside the convertible. This is also estimated as an area of high relevance by the weight $\mathbf{w}_t$ (more yellow, i.e., high values). The decrease in mIoU$_\mathbf{w}$ is likely due to the limited-quality prediction of this context-level corner case. The example physical-level corner case is caused by overexposure, resulting in heavy blending effects on the cars. However, the segmentation method seems not visibly affected by overexposure, as is also validated by an increased mIoU$_\mathbf{w} = 43.04\%$ compared to mIoU$= 41.85\%$. From our examples, the physical-level corner case is irrelevant to our semantic segmentation.

Our results show that mIoU$_\mathbf{w}$ can estimate the relevance of corner cases for semantic segmentation. However, for future work, we would be interested in adapting this weighting to other perception tasks, e.g., by adjusting average precision or transferring the concept into 3D space for more actionable relevance representations.

6. Conclusion

In this work, we propose a relevance-adapted version of the generic IoU metric, called mIoU$_\mathbf{w}$, to refine the identification of corner cases during evaluation. Therefore, we predefine relevance criteria with respect to visual attributes and safety to estimate a per-pixel weighting for semantic segmentation evaluation. Thus, our novel metric establishes the link between perception methods and existing definitions of corner case types using vision-based attributes and per-pixel criticality. Further, we measure the effect of the relevance criteria showing that domain- and scene-level (collective anomalies) corner cases have the highest relevance for the semantic segmentation method w.r.t. IoU$_\mathbf{w}$. Our proposed metric can be applied in future work to detect data-driven corner cases purely based on large reductions in mIoU$_\mathbf{w}$ compared to mIoU. Additionally, mIoU$_\mathbf{w}$ motivates the design of novel corner case detectors by extending the investigated corner case criteria in this work.

ACKNOWLEDGMENT

This work results from the project KI Data Tooling (19A20001M, 19A20001O, 19A20001J) funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK). The authors thank Fabian Linkerhägner for his contributions to this work.

References

[1] Mohamed Aladem and Samir A Rawashdeh. A Single-stream Segmentation and Depth Prediction CNN for Autonomous Driving. *IEEE Intelligent Systems*, 36(4):79–85, May 2020. 3

[2] Abhishek Badki, Orazio Gallo, Jan Kautz, and Pradeep Sen. Binary TTC: A Temporal Geofence for Autonomous Navigation. In *Proc. of CVPR*, pages 12946–12955, Nashville, TN, USA, June 2021. 2, 6

[3] Ayoosh Bansal, Jayati Singh, Micaela Verucchi, Marco Caccamo, and Lui Sha. Risk Ranked Recall: Collision Safety Metric for Object Detection Systems in Autonomous Vehicles. In *Proc. of MECO*, pages 1–4, Budva, Montenegro, June 2021. 2

[4] Daniel Bogdoll, Jasmin Breitenstein, Florian Heidecker, Maarten Bieshaar, Bernhard Sick, Tim Fingscheidt, and J. Marius Zöllner. Description of Corner Cases in Automated Driving: Goals and Challenges. In *Proc. of ICCV - Workshops*, pages 1023–1028, Montreal, QC, Canada, Oct. 2021. 2

[5] Daniel Bogdoll, Stefani Guneshka, and J. Marius Zöllner. One Ontology to Rule Them All: Corner Case Scenarios for Autonomous Driving. In *Proc. of ECCV - Workshops*, pages 409—-425, Tel Aviv, Israel, Oct. 2023. 2

[6] Jan-Aike Bolte, Andreas Bär, Daniel Lipinski, and Tim Fingscheidt. Towards Corner Case Detection for Autonomous Driving. In *Proc. of IV*, pages 366–373, Paris, France, June 2019. 1, 2

[7] Jasmin Breitenstein, Jan-Aike Termöhlen, Daniel Lipinski, and Tim Fingscheidt. Systematization of Corner Cases for Visual Perception in Automated Driving. In *Proc. of IV*, pages 986–993, Las Vegas, NV, USA, Oct. 2020. 2

[8] Jasmin Breitenstein, Jan-Aike Termöhlen, Daniel Lipinski, and Tim Fingscheidt. Corner Cases for Visual Perception in Automated Driving: Some Guidance on Detection Approaches. *arXiv*, pages 1–8, Feb. 2021. 2

[9] Robin Chan, Radin Dardashti, Meike Osinski, Matthias Rottmann, Dominik Brüggemann, Cilia Cilia Rücker, Peter Schlicht, Fabian Hüger, Nikol Rummel, and Hanno Gottschalk. What Should AI See? Using the Public's Opinion to Determine the Perception of an AI. *AI Ethics*, pages 1–25, Jan. 2023. 6

[10] Robin Chan, Matthias Rottmann, Radin Dardashti, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. The Ethical Dilemma when (not) Setting up Cost-based Decision Rules in Semantic Segmentation. In *Proc. of CVPR-Workshops*, pages 1–9, Long Beach, CA, USA, June 2019. 6

[11] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive Deconvolution Networks with Gated Fusion for RGB-D Indoor Semantic Segmentation. In *Proc. of CVPR*, pages 3029–3037, Honolulu, HI, USA, June 2017. 3

[12] MMSegmentation Contributors. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 4

[13] 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 *Proc. of CVPR*, pages 3213–3223, Las Vegas, NV, USA, June 2016. 4

[14] Sara Ferreira, Marco Amorim, and Antonio Couto. Risk Factors Affecting Injury Severity Determined by the MAIS Score. *Traffic Injury Prevention*, 18(5):515–520, Jan. 2017. 6

[15] Tim Fingscheidt, Hanno Gottschalk, and Sebastian Houben, editors. *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*. Springer Nature, Cham, 2022. 1

[16] Sujan Sai Gannamaneni, Sebastian Houben, and Maram Akila. Semantic Concept Testing in Autonomous Driving by Extraction of Object-Level Annotations from CARLA. In *Proc. of ICCV-Workshops*, pages 1006–1014, Montreal, QC, Canada, Oct. 2021. 2

[17] Florian Heidecker, Jasmin Breitenstein, Kevin Rösch, Jonas Löhdefink, Maarten Bieshaar, Christoph Stiller, Tim Fingscheidt, and Bernhard Sick. An Application-Driven Conceptualization of Corner Cases for Perception in Highly Automated Driving. In *Proc. of IV*, pages 644–651, Nagoya, Japan, July 2021. 2, 4, 7

[18] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proc. of ICLR*, pages 1–12, Toulon, France, Apr. 2017. 5

[19] Paul Jaccard. The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11(2):37–50, Feb. 1912. 3

[20] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-Supervised Monocular Depth Estimation: Solving the Dynamic Object Problem by Semantic Guidance. In *Proc. of ECCV*, pages 582–600, Glasgow, UK, Aug. 2020. 6

[21] Maria Lyssenko, Christoph Gladisch, Christian Heinzemann, Matthias Woehrle, and Rudolph Triebel. From Evaluation to Verification: Towards Task-oriented Relevance Metrics for Pedestrian Detection in Safety-Critical Domains. In *Proc. of CVPR-Workshops*, pages 38–45, Nashville, TN, USA, June 2021. 1, 2

[22] Maria Lyssenko, Christoph Gladisch, Christian Heinzemann, Matthias Woehrle, and Rudolph Triebel. Towards Safety-Aware Pedestrian Detection in Autonomous Systems. In *Proc. of IROS*, pages 293–300, Kyoto, Japan, Oct. 2022. 2

[23] Jonas Löhdefink, Justin Fehrling, Marvin Klingner, Fabian Hüger, Peter Schlicht, Nico M. Schmidt, and Tim Fingscheidt. Self-Supervised Domain Mismatch Estimation for Autonomous Perception. In *Proc. of CVPR - Workshops*, pages 1359–1368, Seattle, WA, USA, June 2020. 7

[24] Fatima Meguellati, Assi N'Guessan, and Thierry Hermitte. Analyzing the Maximum Abbreviated Injury Scale in Vehicle Crashes by Using a Logistic Normal Model. *Transportation Research Record*, 2432(1):74–81, Jan. 2014. 6

[25] James Nunn, Jo Barnes, Andrew Morris, Emily Petherick, Roderick Mackenzie, and Matt Staton. Identifying mais 3+

injury severity collisions in uk police collision records. *Traffic Injury Prevention*, 19(2):142–144, Mar. 2019. 6

[26] Eshed Ohn-Bar and Mohan Manubhai Trivedi. Are All Objects Equal? Deep Spatio-temporal Importance Prediction in Driving Videos. *Pattern Recognition*, 64:425–436, Apr. 2017. 1

[27] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. BASNet: Boundary-Aware Salient Object Detection. In *Proc. of CVPR*, pages 7479–7489, Long Beach, CA, USA, June 2019. 3

[28] Serin Varghese, Yasin Bayzidi, Andreas Bär, Nikhil Kapoor, Sounak Lahiri, Jan David Schneider, Nico Schmidt, Fabian Hüger, Peter Schlicht, and Tim Fingscheidt. Unsupervised Temporal Consistency Metric for Video Segmentation in Highly-Automated Driving. In *Proc. of CVPR - Workshops*, pages 1369–1378, Seattle, WA, USA, June 2020. 3

[29] Qi Wang, Junyu Gao, and Yuan Yuan. Embedding Structured Contour and Location Prior in Siamesed Fully Convolutional Networks for Road Detection. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):230–241, Jan. 2018. 5

[30] Jun Wei, Shuhui Wang, and Qingming Huang. F^{3}Net: Fusion, Feedback and Focus for Salient Object Detection. In *Proc. of AAAI*, pages 12321–12328, New York, NY, USA, Feb. 2020. 3

[31] 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 *SAFECOMP Workshops*, 2020. 2

[32] Mirja Wolf, Luiz R. Douat, and Michael Erz. Safety-Aware Metric for People Detection. In *Proc. of ITSC*, pages 2759–2765, Indianapolis, IN,USA, Sept. 2021. 2

[33] Gengshan Yang and Deva Ramanan. Upgrading Optical Flow to 3D Scene Flow Through Optical Expansion. In *Proc. of CVPR*, pages 1334–1343, Seattle, WA, USA, June 2020. 6

[34] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-Contextual Representations for Semantic Segmentation. In *Proc. of ECCV*, pages 173–190, Glasgow, United Kingdom, Aug. 2020. 4

[35] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open Vocabulary Scene Parsing. In *Proc. of ICCV*, pages 2002–2010, Venice, Italy, Oct. 2017. 3