MUSES: The Multi-Sensor Semantic Perception Dataset for Driving under Uncertainty

Tim Brödermann*[1], David Bruggemann*[1], Christos Sakaridis[1], Kevin Ta[1], Odysseas Liagouris[1], Jason Corkill[1], and Luc Van Gool[1]

* Equal contribution [1] Computer Vision Lab, ETH Zürich, Switzerland

Abstract. Achieving level-5 driving automation in autonomous vehicles necessitates a robust semantic visual perception system capable of parsing data from different sensors across diverse conditions. However, existing semantic perception datasets often lack important non-camera modalities typically used in autonomous vehicles, or they do not exploit such modalities to aid and improve semantic annotations in challenging conditions. To address this, we introduce MUSES, the MUlti-SEnsor Semantic perception dataset for driving in adverse conditions under increased uncertainty. MUSES includes synchronized multimodal recordings with 2D panoptic annotations for 2500 images captured under diverse weather and illumination. The dataset integrates a frame camera, a lidar, a radar, an event camera, and an IMU/GNSS sensor. Our new two-stage panoptic annotation protocol captures both class-level and instance-level uncertainty in the ground truth and enables the novel task of uncertainty-aware panoptic segmentation, along with standard semantic and panoptic segmentation. MUSES proves both effective for training and challenging for evaluating models under diverse visual conditions, and it opens new avenues for research in multimodal and uncertainty-aware dense semantic perception. Our dataset and benchmark are publicly available at `https://muses.vision.ee.ethz.ch/`.

1 Introduction

Experimental evidence [53] suggests that dense pixel-level semantic perception is one of the most central tasks for embodied intelligent agents such as autonomous cars. It densely parses the scene that surrounds the car into segments belonging to different semantic classes and/or distinct instances of these classes. The resulting high-level visual representation is an essential input to the downstream driving modules involving planning and control.

Due to their excellent spatial resolution, frame cameras are the standard sensor of choice in datasets for 2D pixel-level semantic perception of driving scenes [13, 21, 27, 34, 41, 49, 51]. However, their signal deteriorates severely in adverse visual conditions, such as nighttime, fog, rain, snow, or sun glare. This implies that the semantic perception module of an autonomous car can benefit from additional sensor inputs such as lidars, radars, and event cameras. These can complement frame cameras in different ways, such as via the robustness

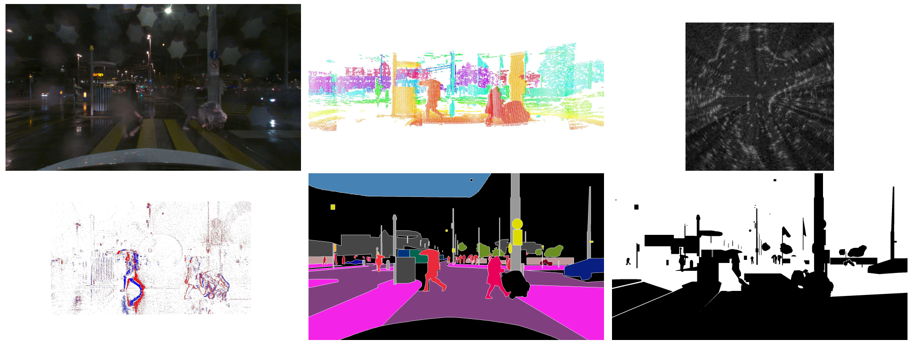

Fig. 1: Annotated scene of MUSES. First row left to right: frame camera, lidar, radar; second row left to right: event camera, panoptic annotation, and difficulty annotation.

of lidars to ambient illumination, the robustness of radars to weather, and the high dynamic range and low latency of event cameras. Through sensor fusion, complementary information from the additional sensors can be exploited in parts of the scene where the camera signal is unintelligible.

To develop sensor fusion models for dense semantic perception in driving scenes, appropriate datasets are required. The central features of such datasets are (i) the inclusion of multiple sensor modalities that are relevant for parsing driving scenes, (ii) the annotation of the input images of the dataset with fine pixel-level labels that facilitate panoptic segmentation, (iii) proper annotation of the aleatoric uncertainty in the semantic content of the input owing to the potentially low signal-to-noise ratio of the latter, and (iv) the representation of a diverse set of visual conditions, including adverse ones. Existing driving datasets, however, feature only some of these attributes while missing others [8, 9, 12, 20, 28, 44]. In particular, multimodal datasets targeting adverse weather only feature coarse bounding boxes for traffic-related objects [1,4,30,35], pixel-level annotations just for the road [1], or limited object classes [16]. These annotations preclude the training and evaluation of general pixel-level segmentation methods that parse the entire driving scene (e.g. sidewalk, pole, and other important classes for driving). On the other hand, adverse-condition datasets that feature dense semantic labels are based solely on camera inputs [34, 51], which limits the number and extent of pixels that can reliably receive manual semantic labels. There is a need for a multimodal driving dataset enabling dense semantic perception where adverse conditions are sufficiently represented.

To respond to this need, we construct MUSES, the MUlti-SEnsor Semantic perception dataset for driving under uncertainty. MUSES consists of driving sequences that include synchronized, calibrated, multimodal recordings with a normal frame-based camera, a new-generation microelectromechanical-system (MEMS) mirror-based lidar, a frequency-modulated continuous wave (FMCW) scanning radar, a high-definition (HD) event camera, and an IMU/GNSS sensor (see Fig. 1). These sequences are recorded under a diverse set of weather and illumination conditions, covering various combinations of time of day and precipitation. To the best of our knowledge, MUSES is the first adverse conditions-focused dataset that includes event camera or MEMS lidar data.

We annotate 2500 selected camera images of MUSES with high-quality fine 2D pixel-level panoptic annotations, which directly afford semantic and instance annotations. The annotated images are evenly distributed over different combinations of time of day, visibility, and precipitation, in each of the training, validation, and test sets. Each of the panoptically annotated camera images is accompanied by the respective readings of the other four sensors, which render MUSES the first large-scale multimodal diverse-condition driving dataset for dense semantic perception tasks, including semantic, instance, and panoptic segmentation. Moreover, our specialized image annotation protocol leverages the respective readings of the additional modalities and a corresponding normal-condition sequence, allowing the annotators to also reliably label degraded image regions that are still discernible in other modalities but would otherwise be impossible to label only from the image itself. This results in better pixel coverage in the annotations and creates a more challenging evaluation setup.

As uncertainty quantification of the perception module is crucial for certifiable downstream control [3, 5, 26, 46], our specialized annotation protocol accounts for high aleatoric uncertainty in the semantic content of images in MUSES. In particular, we extend the uncertainty-aware annotation protocol of ACDC [34] to model *instance-level uncertainty*, besides ACDC's class-level uncertainty. This allows MUSES to support the novel task of *uncertainty-aware panoptic segmentation*, which constitutes a generalization of both panoptic segmentation [22] and uncertainty-aware semantic segmentation [34]. Ground-truth class-level and instance-level difficulty maps are used in uncertainty-aware panoptic segmentation to evaluate respective class- and instance-level confidence maps jointly with the standard panoptic predictions via the novel average uncertainty-aware panoptic quality (AUPQ) metric, which rewards labels and confidence values that are consistent with their human counterparts.

We present a detailed analysis of MUSES, highlighting the challenges for annotation in adverse conditions, notably fog and nighttime. We further show that our two-stage annotation indeed enhances label coverage and difficulty. The investigation of sensor impact on semantic perception highlights the intricacies of multimodal fusion, emphasizing the need for specialized fusion methods. Also, we show that models trained on MUSES excel in cross-domain evaluation, which is attributable to the high diversity and annotation quality of our dataset. Overall, our experimental analysis establishes MUSES as a challenging benchmark for uni- and multimodal dense semantic perception tasks, including standard and uncertainty-aware semantic/panoptic segmentation.

MUSES opens up the following new research directions: (1) Sensor fusion for pixel-level semantic perception in adverse conditions, (2) exploring event camera utility in adverse weather for automated driving, (3) examining challenges and opportunities of new-generation automotive MEMS lidars for semantic perception, and (4) researching the novel uncertainty-aware panoptic segmentation task. MUSES promises to advance the field by enabling more robust and accurate perception systems, particularly in challenging conditions, thus enhancing the safety and efficiency of autonomous driving technology.

2 Related Work

Two of the earliest and most influential datasets for autonomous driving are KITTI [19], which provides high-resolution stereo images, lidar point clouds, and IMU/GNSS data, and Cityscapes [13], which offers high-quality pixel-level semantic annotations for RGB images. However, neither KITTI nor Cityscapes include any samples for adverse weather or illumination. To address this limitation, several datasets have been proposed that explicitly incorporate visual hazards [50] in their data collection. These hazards can be caused by natural phenomena, such as rain, snow, fog, or night, or by artificial factors, such as lens flare, motion blur, or sensor noise. Some of these datasets are Oxford RobotCar [24], Oxford Radar RobotCar [2], DENSE [4], CADC [30], AWARE [29] and Boreas [7]. These datasets provide multi-sensor recordings from different modalities, such as RGB cameras, lidar scanners, radar sensors, or thermal cameras. However, most of them do not provide *fine*, pixel-level *semantic* annotations, which are essential for training semantic segmentation models that are crucial for downstream action tasks of embodied agents [53]. Another set of datasets provides dense semantic annotations for RGB images under adverse conditions. These include the small-scale Foggy Driving [32], Foggy Zurich [31], Dark Zurich [33] and Raincouver [40] and the large-scale ACDC [34], Wilddash2 [51] and BDD100K [47]. Although these datasets focus on diverse weather and illumination conditions, they are limited to vision-only data and are not usually sufficient for reliable perception under adverse conditions. Some recent datasets like InfraParis [17] provide multimodal data with dense semantic annotations, some include adverse lighting conditions like [42] and the panoptic extension [25] of Waymo Open [38], or even adverse weather conditions [1, 16]. They combine RGB frames mainly with lidar point clouds to provide a more comprehensive representation of the scene but do not include further sensors or specialized labeling for adverse conditions. Another relevant dataset is DSEC-Semantic [18,39], which adds event cameras to the frame-lidar combination, but it only provides semantic pseudo-labels.

Our dataset, MUSES, aims to fill the gap between the existing datasets for semantic perception under adverse conditions (see Tab. 1). MUSES provides diverse-condition multimodal data from a normal frame camera, a lidar, a radar, and an event camera, and corresponding high-quality 2D pixel-level panoptic annotations. We extend the uncertainty-aware semantic segmentation task introduced in [34] to panoptic segmentation, by accounting for both class- and instance-level aleatoric uncertainty in predictions and ground truth. The resulting novel task of uncertainty-aware panoptic segmentation is enabled by a specialized annotation protocol, and it is fundamentally different from previous work [37] that investigates epistemic uncertainty for panoptic segmentation.

Table 1: Comparison of MUSES to densely labeled semantic adverse-condition datasets. “Sem./Pan. seg.”: semantic/panoptic segmentation, “Corr. normal”: includes image-level correspondences, “Class/Inst. uncert.”: captures class/instance uncertainty, *: segmentation only for road/lanes, **: pseudo-labels.

Dataset	Frame camera	Lidar	Event camera	Radar	Sem. seg.	Pan. seg.	Corr. normal	Class uncert.	Inst. uncert.
Foggy Zurich [14]	✓	×	×	×	✓	×	×	×	×
Dark Zurich [33]	✓	×	×	×	✓	×	✓	✓	×
Wilddash2 [51]	✓	×	×	×	✓	✓	×	×	×
BDD100K [47]	✓	×	×	×	✓	✓	×	×	×
ACDC [34]	✓	×	×	×	✓	✓	✓	✓	×
Zenseact Open Dataset [1]	✓	✓	×	×	*	×	×	×	×
WOD: PVPS [25,38]	✓	✓	×	×	✓	✓	×	×	×
DSEC-Semantic [18,39]	✓	✓	✓	×	**	×	×	×	×
MUSES (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓

3 The MUSES Dataset

MUSES is a multi-sensor dataset with 2D panoptic annotations for autonomous driving that emphasizes adverse conditions. The dataset was recorded in Switzerland from November 2022 to July 2023. We recorded multiple driving sequences under adverse conditions (night, rain, snow, fog, or a combination thereof) and recorded each adverse-condition route again in clear-weather daytime conditions. The multi-sensor readings serve two purposes: 1) they provide corroborative evidence during annotation to increase label coverage; 2) they support multimodal methods for dense semantic perception in adverse visual conditions. This section provides a detailed overview of the dataset. We detail the sensors in Sec. 3.1, the annotation protocol in Sec. 3.2, and the splits and benchmarks in Sec. 3.3.

3.1 Sensor Suite

Our sensor suite consists of a high-resolution frame camera, an event camera, a MEMS lidar, an FMCW scanning radar, and an IMU/GNSS module, all mounted on the front part of the roof of our recording car (see supplement). Since we recorded highly adverse scenes, the two cameras and the GNSS were placed inside a waterproof box with a transparent acrylic glass window. To manage lens occlusion due to water droplets or snow, the acrylic glass window and lidar cover glass were treated with a hydrophobic coating and regularly wiped clean. The geometric calibration and synchronization of all the sensors are described in the supplement, and the sensor specifications are summarized in Tab. 2.

Frame camera. We use a TRI023S-CC camera from Lucid Vision Labs with a global shutter and automatic exposure time. The RGB images are undistorted and provided as 8-bit PNG files. We use the frame camera as our frame-of-reference as our panoptic annotations pertain to the undistorted RGB images. To protect the privacy of all persons identifiable from our dataset, we used a

Table 2: Overview of sensor specifications for MUSES.

Modality	Name	Specifications
Frame camera	TRI023S-CC	8-bit RGB, 30 Hz, 1920×1080, HFOV: 77°, VFOV: 43°
Event camera	Prophesee GEN4.1	1280×720, 15M events/s, HFOV: 64°, VFOV: 39°
Lidar	RS-LiDAR-M1	10 Hz, avg. angular resolution: 0.2°, range: 200 m, HFOV: 120°, VFOV: 25°, 75K points/scan
Radar	Navtech CIR-DEV	4 Hz, range resolution: 43.8 mm, horizontal angular resolution: 0.9°, range: 330 m
IMU/GNSS	simpleRTK2B Fusion	RTK accuracy: <10cm, 30 Hz

semi-automatic anonymization pipeline based on [43] to segment and blur faces and license plates. More details on anonymization are in the supplement.

Event camera. Event cameras exhibit high dynamic range and low latency, but the potential of event cameras in adverse conditions remains unexplored due to a lack of available datasets. To foster research in this direction, we use a Prophesee Gen4.1 event camera with HD resolution. The contrast sensitivity is manually adjusted depending on the lighting conditions, to record more details in brightly lit scenes and limit noise under low illumination. To further avoid peaks of events, we used an event rate controller. The raw events—*i.e.*, the 4-tuples (x, y, timestamp, polarity)—are undistorted, truncated to 3-second chunks, and stored in compressed HDF5 files.

Lidar. Lidars are a core sensing modality for autonomous cars, providing high-resolution 3D maps irrespective of lighting conditions. New-generation MEMS lidars offer advantages such as a compact form, minimal moving parts, and cost-effectiveness, but their irregular scanning pattern poses compatibility issues with standard motorized scanning lidar frameworks. To facilitate further research on such sensors, we use the new-generation automotive RS-LiDAR-M1 lidar from RoboSense. The raw lidar point cloud is stored as a binary file with six entries (x, y, z, intensity, mirror #, timestamp) for each point.

Radar. Due to its robustness to any type of weather condition, the radar is an important component of reliable automotive perception stacks. We use a Navtech CIR-DEV FMCW radar with a cosec dish that directs part of the emitted power below the horizontal plane for better short-range detection. Inspired by [2], we store the radar data as a range-azimuth PNG file, containing per azimuth the timestamp, sweep counter, and raw received power readings in dB.

IMU/GNSS. We use the simpleRTK2B board from ArduSimple as our IMU/GNSS device. It offers dead reckoning and real-time kinematics (RTK). An extended Kalman filter (EKF) fuses GNSS and IMU, resulting in a global 6 DoF pose with respective velocities.

3.2 Annotation

Along with the processed multi-sensor data, MUSES contains high-quality panoptic labels. Figure 2 illustrates that lidar point clouds provide insufficient information for accurate 3D annotation of objects in adverse conditions. We thus label the 2D frame camera images, due to their high resolution and ease of interpretability for annotators. The RGB images are labeled for panoptic segmentation with 19 classes, following the taxonomy of Cityscapes evaluation classes [13]. Even 2D panoptic labeling is highly challenging in adverse conditions because these images contain regions with indiscernible semantic content: the image may be too dark or blurry to recognize an object, or a rain droplet on the lens may partially occlude the field-of-view. To account for such cases, we asked annotators to follow the decision hierarchy shown in Fig. 4 for panoptic labeling. If the object class is indiscernible, the pixel is labeled as `unknown_class` (and by extension `unknown_instance`), otherwise it belongs to either a `thing` class or a `stuff` class. Pixels outside the 19 defined classes (*e.g.*, a wheelchair) are given the fallback label `other_class` and treated similarly to `stuff`. If the pixel belongs to a `thing` class, the annotator can either assign it to an instance or label it as `unknown_instance`. This decision hierarchy results in two levels of uncertainty: class and instance uncertainty. Instance uncertainty occurs when we know the class but not its instance membership, *e.g.*, for pixels at the blurry boundary between two adjacent cars.

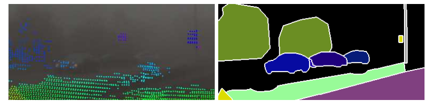

Fig. 2: 3D vs. 2D. This example shows that the lidar point cloud (left, projected onto RGB image) can deteriorate in adverse weather, yielding insufficient information for 3D annotation. By contrast, the distant cars are captured with 2D annotations (right).

We design a two-stage annotation protocol. In stage 1, the RGB image is labeled with an initial ground-truth panoptic label H1 following the hierarchical protocol in Fig. 4. In this stage, the annotator only has access to the RGB image under annotation. In stage 2, H1 is refined with the same decision hierarchy to obtain the final ground truth panoptic label H2 (see Fig. 3). The refinement is made possible by additionally inspecting auxiliary data, which includes enhanced versions of the original image [36], event and lidar data, temporally adjacent

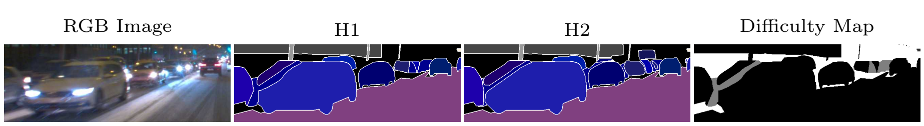

Fig. 3: Example of stage 1 and stage 2 panoptic annotations H1 and H2. The auxiliary data available in stage 2 allows better separation between the three car instances on the left, reducing the `unknown_instance` area from H1 to H2, but keeping a `difficult_instance` label (grey) in the difficulty map. Notice the additional class labels added to H2 for distant cars; the corresponding regions keep the `difficult_class` (white) label in the difficulty map.

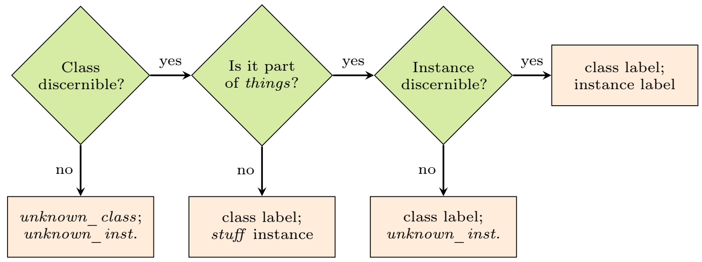

Fig. 4: Flowchart of uncertainty-aware panoptic annotation.

frames of the sequence, and a corresponding sequence from the same location recorded under normal conditions. This auxiliary data enables the annotators to fill in previously uncertain regions (*i.e.*, droplets on RGB image in Fig. 5) and correct possible mistakes from stage 1. In all stages, we asked annotators to prioritize traffic-related classes, *i.e.*, `things` classes plus road, traffic light, and traffic sign. The total annotation time was 11,827 hours, averaging 4.6 hours per image. It was divided equally between stage 1 and stage 2.

The two-stage annotation process allows us to assign ternary *difficulty* levels to our labels (see last column in Fig. 3). Pixels with valid labels that are fully consistent in H1 and H2 are labeled `not_difficult`. Pixels with a consistent class label in H1 and H2, but an inconsistent instance label or the `unknown_instance` label, receive a `difficult_instance` label. All other pixels, *i.e.*, pixels with an inconsistent class label in H1 and H2 or pixels that are `unknown_class`, receive a `difficult_class` label.

3.3 Splits and Benchmarks

MUSES consists of 2500 annotated multimodal samples, 500 of which are recorded in daytime and clear weather. The remaining 2000 adverse-condition samples are split between 1000 daytime (333/334/333 samples in fog/rain/snow) and 1000 nighttime (250/250/250/250 samples in clear/fog/rain/snow) samples. Figure 5

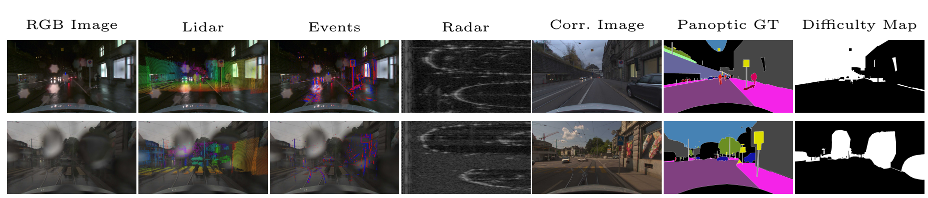

Fig. 5: Visualization of two adverse-condition samples from MUSES. From left to right: RGB image; motion-compensated lidar points projected and overlaid with the image; events projected onto the image (assuming infinite distance); azimuth-range radar scan (with ranges above a threshold cropped out); corresponding normal-condition image; panoptic ground truth; difficulty map. Best viewed zoomed in.

shows two example adverse-condition scenes with a visualization of each recorded modality, the corresponding normal-condition images, and the annotation outputs. We split the full dataset into 1500 training, 250 validation, and 750 test samples, stratified by condition. The individual splits are strictly geographically separated across conditions. The annotations of the test set are withheld, and the test set evaluation is only accessible through a submission system. Public benchmarks will be provided for three tasks: 1) semantic segmentation, 2) panoptic segmentation, and 3) uncertainty-aware panoptic segmentation (see Sec. 4). Each of the three benchmarks will have two different tracks: i) using RGB input only, and ii) using multimodal input.

4 Uncertainty-Aware Panoptic Segmentation

The ternary difficulty levels assigned to each pixel during the annotation of MUSES (see Sec. 3.2) enable a novel task: uncertainty-aware panoptic segmentation. In this task, a panoptic segmentation model is compensated for errors in difficult image regions if it predicts the difficulty level correctly. To incorporate this idea into evaluation, we introduce the uncertainty-aware panoptic quality (UPQ) metric, an extension of panoptic quality [22].

4.1 Uncertainty-Aware Panoptic Quality

In addition to the panoptic prediction and ground truth, UPQ takes as input a binary class confidence prediction and a binary instance confidence prediction, which are evaluated against the ternary difficulty maps (see Sec. 3.2). This confidence comparison transforms some pixels in the panoptic prediction to either `ANY` or `VOID`, as follows: First, the binary class confidence prediction divides all pixels into class-unconfident (CU) and class-confident (CC) sets. CU pixels are compared with the ground truth difficulty map: If they are indeed `class_difficult`, they are converted to `ANY` pixels in the panoptic prediction, otherwise, they are converted to `VOID`. The CC pixels are instead divided into instance-unconfident (IU) and instance-confident (IC) sets according to the binary instance confidence prediction. IU pixels with a correct class prediction are compared with the difficulty map: If they are `difficult_instance` or `difficult_class`, they are converted to `ANY` pixels, otherwise, they are converted to `VOID`. Thus, ANY masks pixels that are correctly predicted as uncertain (class or instance level), while VOID masks pixels that are erroneously predicted as uncertain. Note that ANY/VOID masks apply to the predictions, not the ground truth.

Analogously to the standard PQ [22], UPQ is composed of two steps: segment matching and UPQ computation given the matched segments. Compared to PQ, UPQ modifies the first step, by incorporating `ANY` pixels as wild cards and `VOID` pixels as wrong predictions. Hence, UPQ forgives errors for difficult pixels when predictions are correctly uncertain but penalizes predicting "easy" pixels as uncertain. Note that, if a model predicts all pixels as confident, no `ANY` or `VOID` pixels are created and UPQ reduces to the standard PQ.

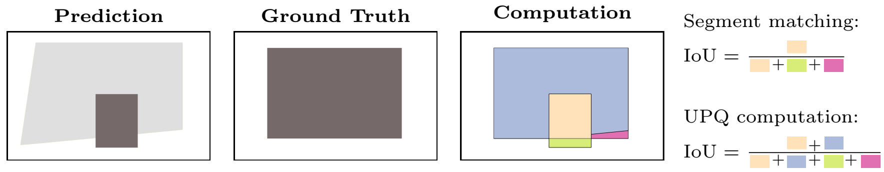

Fig. 6: UPQ computation between a prediction containing an instance (black), `ANY` region (gray), and background (white), and a ground-truth instance (black). For segment matching, the `ANY` regions are ignored, which leads to an IoU>0.5 in this example. Before UPQ computation after matching, the matched `ANY` regions are replaced with the ground truth.

Segment Matching. In standard PQ, matches between predicted and ground-truth segments are formed when their intersection over union (IoU) is strictly greater than 0.5, which guarantees unique matching. To maintain this property while accounting for `ANY` pixels, UPQ matches segments in two steps:

1. Ignore all `ANY` pixels in both prediction and ground truth, then match remaining segments with IoU > 0.5. After matching, copy the ground truth class/instance labels to the `ANY` pixels within matched segments.
2. Check remaining unmatched ground-truth segment for $> 50\%$ overlap with ANY pixels. For matches, replace the `ANY` label with the ground truth.

Figure 6 conveys the intuition for this process: by ignoring the `ANY` pixels in step 1, segments surrounded by accurate confidence predictions are more easily matched. Even though the predicted segment has an IoU of less than 0.5 with the ground-truth one, it is still matched because the confidence level of the surrounding pixels is correctly predicted as `difficult_class` or `difficult_instance` (which results in `ANY` pixels). Note that ANY pixels never form segments and thus cannot create FP or FN segments. If the confidence level of the entire instance were correctly classified as difficult, it would still be matched in step 2.

UPQ Computation. Given segment matches, UPQ is computed as

$$\mathrm{UPQ} = \frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}, \tag{1}$$

where p and g are matched segments, and $|TP|$, $|FP|$, and $|FN|$ are the numbers of true positive, false positive, and false negative segments respectively. See [22] for details and the motivation behind this formulation.

4.2 Threshold-Agnostic Evaluation

UPQ evaluation requires binary class confidence and instance confidence predictions. For a model that produces confidence *scores* for both cases, UPQ must thus be evaluated at a specific operating point, by defining confidence thresholds for class and instance confidences respectively. To obtain a threshold-agnostic summary metric, we define a linear grid of 16×16 thresholds for class and instance

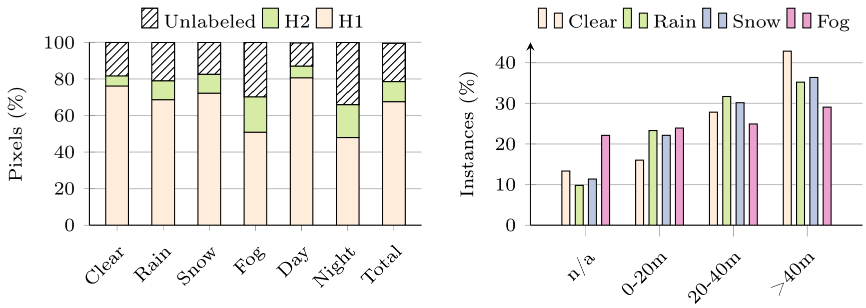

Fig. 7: (left) Percentages of pixels labeled in the two annotation stages for different conditions. **(right)** Distance distribution of instances across different weather conditions. H1: stage 1, H2: stage 2, "n/a": instances without any lidar returns.

confidences, evaluate UPQ for each of the 256 threshold pairs, and report the average value—AUPQ.

5 Analysis and Experiments

In this section, we present a thorough analysis of MUSES. We analyze its annotations in Sec. 5.1, corroborate the benefit of the non-camera modalities of our dataset for dense semantic perception in Sec. 5.2, demonstrate the increased difficulty MUSES presents for state-of-the-art semantic segmentation approaches compared to competing central benchmarks in Sec. 5.3, and evaluate baselines and oracles for the novel uncertainty-aware panoptic segmentation task in Sec. 5.4. Performance is always reported on the respective test sets.

5.1 Analysis of Annotations

Fig. 8 summarizes the number of annotated pixels per class in MUSES. Overall, 78.5% of all pixels receive a valid label after stage 2 of the annotation (in H2), an increase of 10.3% compared to H1. Notably, H2 contains roughly 1.5× more instances than H1. The increased level of detail and completeness in H2 supports the design choice of complementing the annotation with multi-sensor information. When comparing the annotation density in H2 for images of different visual conditions in Fig. 7 (left), we observe that fog and nighttime constitute the most challenging conditions to annotate, showing the greatest increases in annotation density when incorporating auxiliary data in

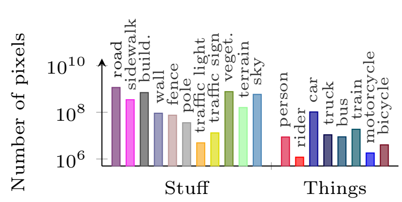

Fig. 8: Number of annotated pixels per class in MUSES.

Table 3: PQ of Mask2Former [11] models trained on different sets of input modalities. A Swin-T [23] backbone is used in all cases.

Frame camera	Event camera	Radar	Lidar	Clear	Fog	Rain	Snow	Day	Night	All
✓				48.8	46.5	45.4	42.2	49.4	39.4	46.9
✓	✓			52.1	49.4	48.2	42.6	51.7	42.2	49.5
✓		✓		52.9	49.5	49.9	46.1	52.9	44.8	51.3
✓			✓	54.2	49.9	52.2	47.6	53.7	48.0	52.7
✓	✓	✓	✓	**55.3**	**50.3**	**53.8**	**47.9**	**54.1**	**49.7**	**53.6**

stage 2 of the annotation. To measure the added *difficulty* of stage 2 annotations, we train Mask2Former [11] with H2 and evaluate it on stage 1 (H1) and stage 2 (H2) labels. We observe a 11.1% drop in PQ, from 58.0% (H1) to 46.9% (H2), which indicates the substantial increase in the difficulty of the labels that are created thanks to our multimodal annotation. This highlights the challenging nature of ground truth in MUSES compared to the ground truth in 2D semantic perception datasets that lack auxiliary data—such as our additional modalities—for adverse-condition annotations. We provide a quantitative verification of this argument via a cross-dataset comparison in Sec. 5.3. As a result of our two-stage annotation protocol, 24.5% of pixels are labeled as `difficult_class`, and 1.5% of `things` pixels are labeled as `difficult_instance`.

We investigate the range distribution of annotated `things` instances in MUSES by estimating their distances via the motion-compensated and outlier-filtered lidar point cloud. Figure 7 (right) shows that annotated instances in adverse conditions are closer on average. However, those results need to be interpreted cautiously, as the lidar can be heavily affected by adverse conditions, *e.g.*, in fog 43.4% of instances have less than 10 lidar returns and 22.3% have none. This underscores the advantage of 2D image labeling over point cloud labeling for adverse scenes. More statistics for MUSES are provided in the supplement.

5.2 Benefit of Additional Modalities for Dense Semantic Perception

In Tab. 3, we show the benefit of the additional non-camera modalities which MUSES offers compared to previous dense semantic driving perception benchmarks [13, 34, 47], by comparing multimodal models to a camera-only model. In particular, we train different Mask2Former-based networks [11] on MUSES for panoptic segmentation, each using different modalities as inputs. For this, we construct simple multimodal networks by employing separate Swin-T [23] backbones for each modality and fusing their outputs with a parallel cross-attention block [6]. For more implementation details, please refer to the supplement.

The key observation on the results of this comparison is that the quadrimodal network outperforms the baseline camera-only network across all visual conditions. This finding proves the value of our novel multimodal dense semantic perception dataset for developing better models for such pixel-level tasks. In addition, we observe that all three examined bimodal networks, which fuse the camera input with one of the additional modalities respectively, also deliver consistent improvements over the camera-only network across conditions and trail the performance of the quadrimodal network. Each non-camera modality provides valuable complementary information for semantic perception on top of the camera. However, there is significant potential to improve upon our simple fusion method. Future multimodal dense semantic perception works based on MUSES could develop more specialized fusion architectures and enhance model performance across all domains.

5.3 State of the Art in Semantic Image Segmentation on MUSES

We evaluate the performance of state-of-the-art semantic segmentation architectures on MUSES in Tab. 4. All backbones are pre-trained on ImageNet [15]. High-capacity models tend to perform better: Mask2Former [11] achieves the highest mIoU.

Table 4: State of the art in semantic segmentation on MUSES.

Architecture	mIoU
DeepLabv3+ (ResNet101-D8) [10]	70.5
OCRNet (HRNetV2p-W48) [48]	71.9
SETR (ViT-L) [52]	71.1
SegFormer (MiT-B2) [45]	72.5
SegFormer (MiT-B5) [45]	74.7
Mask2Former (Swin-T) [11]	70.7
Mask2Former (Swin-L) [11]	**77.1**

Because of its diverse range of conditions and high-quality ground truth, MUSES constitutes an ideal testbed for model adaptation and generalization experiments. To compare different datasets along this axis, we train Mask2Former on three datasets—Cityscapes [13], ACDC [34], and MUSES—and cross-evaluate the resulting models. Note that, of the three datasets, MUSES has the smallest training set. The results in Tab. 5 show that, despite this handicap, the model trained on MUSES achieves the highest average performance over all datasets, demonstrating its robustness. We hypothesize that this can be attributed to the superior diversity and annotation quality of MUSES samples. Nevertheless, the performance of all models drops significantly under domain shift, motivating future research efforts in this field.

Table 5: Generalization of semantic segmentation models trained on different datasets. Mask2Former [11] with a Swin-L [23] backbone and camera-only input is used. We train and test on different combinations of datasets. Δ: mIoU difference to in-domain model.

Training Dataset	Cityscapes-val [13]		ACDC-test [34]		MUSES-test		Mean
	mIoU	Δ	mIoU	Δ	mIoU	Δ	mIoU
Cityscapes [13]	83.7	—	65.7	-10.6	58.9	-18.2	69.4
ACDC [34]	70.9	-12.8	76.3	—	66.9	**-10.2**	71.4
MUSES	73.1	**-10.6**	72.0	**-4.3**	77.1	—	**74.1**

Table 6: Uncertainty-aware panoptic segmentation baselines and oracles. Mask2Former [11] with Swin-T [23] is used.

Confidence	AUSQ↑	AURQ↑	AUPQ↑		
			Stuff	Things	All
Constant 100%	78.3	58.0	58.2	31.3	46.9
Marginalization	80.2	53.6	54.3	30.7	44.3
Oracle	89.7	83.7	78.0	71.2	75.2

5.4 Uncertainty-Aware Panoptic Segmentation

For the novel uncertainty-aware panoptic segmentation task introduced in Sec. 4, we present in Tab. 6 two mask-classification baselines and an oracle approach for the unimodal setting. The first baseline simply predicts 100% class and instance confidence for all pixels. In this case, the AUPQ is equivalent to the standard PQ. The second baseline aims to extract class and instance confidence from the trained mask-classification model: we obtain the class confidence by marginalizing over probability-mask pairs, whereas the instance confidence is obtained by dividing the mask class score through the class confidence. A more detailed description and motivation for this procedure is given in the supplement. The oracle approach is a confidence oracle: it uses ground-truth class and instance confidence maps for confidence prediction. Its performance sets an upper bound for the examined models. As shown in Tab. 6, the first baseline outperforms the marginalization approach, meaning that the confidence scores of those models are not well calibrated to the difficulty labels. Furthermore, there is a substantial performance gap between the two baselines and the oracle, indicating that future custom-trained models have great room for improvement in this novel task.

6 Conclusion

We have presented MUSES, the first multi-sensor semantic perception dataset for autonomous driving in adverse conditions. MUSES provides 2500 samples, each consisting of synchronized and calibrated RGB images, MEMS lidar scans, HD event sequences, FMCW radar scans, IMU/GNSS readings, and an image-level corresponding normal-condition image. Our new two-stage panoptic annotation protocol yields high-quality 2D panoptic annotations for each image and class- and instance-level difficulty maps, which enable the novel task of uncertainty-aware panoptic segmentation. Our experiments highlight the benefits of the additional non-camera modalities for dense semantic perception and prove that MUSES is effective for training and provides a challenging evaluation for models under diverse visual conditions. acrossThis dataset motivates many possible lines of research: sensor fusion for dense semantic perception tasks, the exploration of strengths and weaknesses of individual sensors, *e.g.*, event cameras, for semantic perception in diverse visual conditions, and model generalization and/or adaptation across conditions or modalities.

Acknowledgements

This work was supported by the ETH Future Computing Laboratory (EFCL), financed by a donation from Huawei Technologies.

References

1. Alibeigi, M., Ljungbergh, W., Tonderski, A., Hess, G., Lilja, A., Lindström, C., Motorniuk, D., Fu, J., Widahl, J., Petersson, C.: Zenseact Open Dataset: A large-scale and diverse multimodal dataset for autonomous driving. In: ICCV (2023)
2. Barnes, D., Gadd, M., Murcutt, P., Newman, P., Posner, I.: The Oxford Radar RobotCar Dataset: A radar extension to the Oxford RobotCar Dataset. In: ICRA (2020)
3. Bemporad, A., Morari, M.: Robust model predictive control: A survey. In: Robustness in identification and control, pp. 207–226. Springer (2007)
4. Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., Heide, F.: Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In: CVPR (2020)
5. Bonzanini, A.D., Mesbah, A., Di Cairano, S.: Perception-aware chance-constrained model predictive control for uncertain environments. In: 2021 American Control Conference (ACC). pp. 2082–2087. IEEE (2021)
6. Broedermann, T., Sakaridis, C., Dai, D., Van Gool, L.: HRFuser: A multi-resolution sensor fusion architecture for 2D object detection. In: ITSC (2023)
7. Burnett, K., Yoon, D.J., Wu, Y., Li, A.Z., Zhang, H., Lu, S., Qian, J., Tseng, W.K., Lambert, A., Leung, K.Y., et al.: Boreas: A multi-season autonomous driving dataset. The International Journal of Robotics Research **42**(1-2), 33–42 (2023)
8. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
9. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al.: Argoverse: 3D tracking and forecasting with rich maps. In: CVPR (2019)
10. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
11. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR (2022)
12. Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J.S., An, K., Kweon, I.S.: KAIST multi-spectral day/night data set for autonomous and assisted driving. IEEE Transactions on Intelligent Transportation Systems **19**(3), 934–948 (2018)
13. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
14. Dai, D., Sakaridis, C., Hecker, S., Van Gool, L.: Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. International Journal of Computer Vision **128**(5), 1182–1204 (2020)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
16. Diaz-Ruiz, C.A., Xia, Y., You, Y., Nino, J., Chen, J., Monica, J., Chen, X., Luo, K., Wang, Y., Emond, M., et al.: Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In: CVPR (2022)

17. Franchi, G., Hariat, M., Yu, X., Belkhir, N., Manzanera, A., Filliat, D.: Infraparis: A multi-modal and multi-task autonomous driving dataset. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2973–2983 (2024)
18. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: DSEC: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters **6**(3), 4947–4954 (2021)
19. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: CVPR (2012)
20. Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., et al.: A2d2: Audi autonomous driving dataset. arXiv preprint arXiv:2004.06320 (2020)
21. Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The ApolloScape open dataset for autonomous driving and its application. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(10), 2702–2719 (2020)
22. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: CVPR (2019)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
24. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The Oxford RobotCar dataset. The International Journal of Robotics Research **36**(1), 3–15 (2017)
25. Mei, J., Zhu, A.Z., Yan, X., Yan, H., Qiao, S., Chen, L.C., Kretzschmar, H.: Waymo open dataset: Panoramic video panoptic segmentation. In: ECCV (2022)
26. Mesbah, A.: Stochastic model predictive control: An overview and perspectives for future research. IEEE Control Systems Magazine **36**(6), 30–44 (2016)
27. Neuhold, G., Ollmann, T., Rota Bulò, S., Kontschieder, P.: The Mapillary Vistas dataset for semantic understanding of street scenes. In: ICCV (2017)
28. Patil, A., Malla, S., Gang, H., Chen, Y.T.: The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes. In: ICRA (2019)
29. Pinchon, N., Cassignol, O., Nicolas, A., Bernardin, F., Leduc, P., Tarel, J.P., Brémond, R., Bercier, E., Brunet, J.: All-weather vision for automotive safety: which spectral band? In: Advanced Microsystems for Automotive Applications 2018: Smart Systems for Clean, Safe and Shared Road Vehicles 22nd. pp. 3–15. Springer (2019)
30. Pitropov, M., Garcia, D.E., Rebello, J., Smart, M., Wang, C., Czarnecki, K., Waslander, S.: Canadian adverse driving conditions dataset. The International Journal of Robotics Research **40**(4-5), 681–690 (2021)
31. Sakaridis, C., Dai, D., Hecker, S., Van Gool, L.: Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In: ECCV (2018)
32. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**(9), 973–992 (2018)
33. Sakaridis, C., Dai, D., Van Gool, L.: Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020). `https://doi.org/10.1109/TPAMI.2020.3045882`
34. Sakaridis, C., Dai, D., Van Gool, L.: ACDC: The Adverse Conditions Dataset with Correspondences for semantic driving scene understanding. In: ICCV (2021)
35. Sheeny, M., De Pellegrin, E., Mukherjee, S., Ahrabian, A., Wang, S., Wallace, A.: RADIATE: A radar dataset for automotive perception in bad weather. In: ICRA (2021)

36. Shi, Z., Zhu, M.M., Guo, B., Zhao, M., Zhang, C.: Nighttime low illumination image enhancement with single image using bright/dark channel prior. EURASIP Journal on Image and Video Processing **2018**, 1–15 (2018)
37. Sirohi, K., Marvi, S., Büscher, D., Burgard, W.: Uncertainty-aware panoptic segmentation. IEEE Robotics and Automation Letters **8**(5), 2629–2636 (2023)
38. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020)
39. Sun, Z., Messikommer, N., Gehrig, D., Scaramuzza, D.: ESS: Learning event-based semantic segmentation from still images. In: ECCV (2022)
40. Tung, F., Chen, J., Meng, L., Little, J.J.: The Raincouver scene parsing benchmark for self-driving in adverse weather and at night. IEEE Robotics and Automation Letters **2**(4), 2188–2193 (2017)
41. Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C.V.: IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In: WACV (2019)
42. Vertens, J., Zürn, J., Burgard, W.: Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8461–8468. IEEE (2020)
43. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. In: CVPR (2019)
44. Xiao, P., Shao, Z., Hao, S., Zhang, Z., Chai, X., Jiao, J., Li, Z., Wu, J., Sun, K., Jiang, K., et al.: Pandaset: Advanced sensor suite dataset for autonomous driving. In: ITSC (2021)
45. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS (2021)
46. Yang, S., Pappas, G.J., Mangharam, R., Lindemann, L.: Safe perception-based control under stochastic sensor uncertainty using conformal prediction. arXiv preprint arXiv:2304.00194 (2023)
47. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: CVPR (2020)
48. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: ECCV (2020)
49. Zendel, O., Honauer, K., Murschitz, M., Steininger, D., Fernandez Dominguez, G.: WildDash - creating hazard-aware benchmarks. In: ECCV (2018)
50. Zendel, O., Murschitz, M., Humenberger, M., Herzner, W.: How good is my test data? introducing safety analysis for computer vision. International Journal of Computer Vision **125**, 95–109 (2017)
51. Zendel, O., Schörghuber, M., Rainer, B., Murschitz, M., Beleznai, C.: Unifying panoptic segmentation for autonomous driving. In: CVPR (2022)
52. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR (2021)
53. Zhou, B., Krähenbühl, P., Koltun, V.: Does computer vision matter for action? Science Robotics **4**(30) (2019). `https://doi.org/10.1126/scirobotics.aaw6661`