Aalborg Universitet

AALBORG UNIVERSITY

Navigation-Oriented Scene Understanding for Robotic Autonomy: Learning to Segment Driveability in Egocentric Images

Humblot-Renaux, Galadrielle; Marchegiani, Letizia; Moeslund, Thomas B.; Gade, Rikke

Navigation-Oriented Scene Understanding for Robotic Autonomy: Learning to Segment Driveability in Egocentric Images

Galadrielle Humblot-Renaux[1], Letizia Marchegiani[2], Thomas B. Moeslund[1], and Rikke Gade[1]

Abstract—This work tackles scene understanding for outdoor robotic navigation, solely relying on images captured by an on-board camera. Conventional visual scene understanding interprets the environment based on specific descriptive categories. However, such a representation is not directly interpretable for decision-making and constrains robot operation to a specific domain. Thus, we propose to segment egocentric images directly in terms of how a robot can navigate in them, and tailor the learning problem to an autonomous navigation task. Building around an image segmentation network, we present a generic affordance consisting of 3 driveability levels which can broadly apply to both urban and off-road scenes. By encoding these levels with soft ordinal labels, we incorporate inter-class distances during learning which improves segmentation compared to standard "hard" one-hot labelling. In addition, we propose a navigation-oriented pixel-wise loss weighting method which assigns higher importance to safety-critical areas. We evaluate our approach on large-scale public image segmentation datasets ranging from sunny city streets to snowy forest trails. In a cross-dataset generalization experiment, we show that our affordance learning scheme can be applied across a diverse mix of datasets and improves driveability estimation in unseen environments compared to general-purpose, single-dataset segmentation.

Index Terms—Deep learning for visual perception, semantic scene understanding, computer vision for transportation.

I. INTRODUCTION

A ROBOT roaming outdoors "in the wild" needs to know where to go, and what to avoid. It may traverse vast areas with unfamiliar terrain, unexpected obstacles or challenging environmental conditions which degrade its view, yet should still be able to identify safe and suitable terrain to drive on. In this work, our aim is to parse images captured by an outdoor robot and interpret them at the pixel level in order to inform navigation decisions [1], without constraining scene understanding to a specific domain. We rather consider an "open-world" navigation task spanning on-road and off-road scenes, from grassy fields to city traffic, or from forest trails to pedestrian areas. In this context, it is beneficial to know not only where the road/path is (if there is one), but also what other areas are driveable, although perhaps not ideally so.
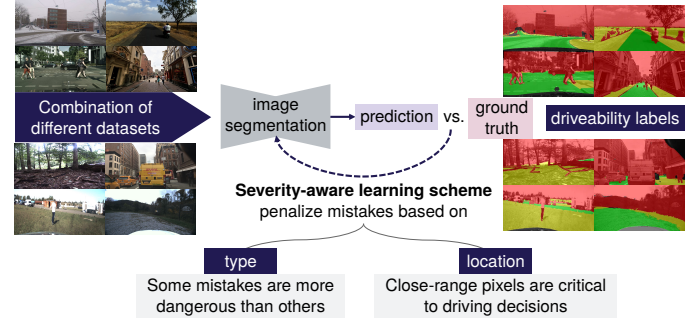
Fig. 1. Overview of our navigation-oriented learning scheme for learning 3-level driveability across diverse outdoor scenes from pixel-annotated datasets.

To learn a consistent and useful representation across diverse scenes, we interpret images directly in terms of potential action rather than object descriptions, following the concept of visual *affordance* [2]. Existing affordance learning approaches for outdoor navigation rely on sensor data recorded on a real platform to generate self-supervised or weakly-supervised labels [3], [4], [5] - this is an impractical and resource-intensive process, which limits the diversity of images seen during training. In contrast, we approach affordance learning as a fully supervised image segmentation problem, leveraging the abundance of large-scale scene understanding datasets. We present a 3-level driveability affordance which is directly interpretable for robotic decision-making and applies across arbitrary outdoor environments (not just roads as in [3], or static off-road scenes as in [5], [4]), while explicitly tailoring the learning problem to navigation. Our contributions include:

1) a navigation-oriented framework which enables cross-dataset training, bypassing the need for real-world exploration or additional labelling effort;
2) a soft label encoding which incorporates the ambiguity and order between levels of driveability, penalizing some mis-classifications more than others during learning;
3) a loss weighting scheme which, rather than treating all pixels as equally important for navigation, concentrates learning in safety-critical areas while allowing leniency around object outlines and distant scene background;
4) a challenging experimental procedure: beyond same-dataset testing, we evaluate the generalizability of our approach on three unseen datasets, including the Wild-Dash benchmark [6] which captures a large variety of difficult driving scenarios across 100+ countries.

Figure 1 illustrates the core idea of our approach. This learning scheme is, to the best of our knowledge, the first attempt at incorporating an inter-class ranking in a scene understanding task, taking both the *type* and *location* of mistakes into account during learning to improve affordance segmentation.

II. RELATED WORK

Semantic segmentation for outdoor scene understanding is extensively studied [7]. The bulk of existing approaches either segment images at the object level [8], or reduce the problem to binary segmentation (e.g. road vs. rest [9] or free space vs. obstacles [10]). Object-based approaches are dataset- and domain-specific, unnecessarily descriptive for navigation, hinder generalisation [11], and scale poorly to unseen obstacles or unstructured scenes [12]. Conversely, binary segmentation is much more generic, but does not capture the kinds of degrees of driveability which are relevant for off-road robotic applications traversing diverse terrain [5], [4], [13].

Instead, we take a probabilistic affordance segmentation approach to scene understanding. Existing works in this direction are either contained to simulation [14], indoor environments [15], [16] or static outdoor scenes [4]. In contrast, our approach considers challenging, dynamic outdoor scenes captured by a real robot or vehicle. Closely related to our work, [3] proposes to segment obstacles and a proposed path in driving scenes, with weakly-supervised labels generated from Lidar and odometry data, and unlabeled pixels assigned to a 3rd "unknown" class. While [3] achieves remarkable performance in structured urban scenes, the driveable area is limited to the current lane, and the method's applicability to off-road scenarios is unclear. Like [3], we adopt a 3-class definition for scene understanding, but as recommended by [17], we introduce a *degree* of driveability. This allows us to generate viable predictions beyond on-road driving scenes, with the aim of enabling open-world robot navigation.

More generally, our method contrasts with existing affordance learning approaches for outdoor navigation which require additional sensor data to be collected by a navigation platform (eg. Lidar [10], [3], [5], odometry [3], [4], IMU [5], or force-torque measurements [4]). Instead, our method leverages the wide range of readily available image segmentation datasets at no annotation cost, and only requires monocular images at training-time.

In addition, as illustrated in Figure 1, our approach places particular emphasis on generalization and mistake severity for safe robotic perception. This contrasts with all the aforementioned works, which are limited to single-dataset training/evaluation and treat all pixels and classes as interchangeable during learning. For bridging the gap across different segmentation datasets, rather than expanding the label space to accommodate an ever-increasing number of object labels as in [12], we reduce the label space down to 3 generic driveability levels. Our "severity-aware" segmentation framework builds upon the findings in [18], which show that encoding the severity of different misclassifications in ground truth labels significantly reduces the risk of collision. However, we also consider the *location* of mistakes during learning, and

propose a multi-domain affordance-based representation which is tailored to robotic navigation.

III. APPROACH

Our approach primarily revolves around how we formulate the learning problem. First, we generate driveability labels by mapping object-based pixel annotations from existing semantic segmentation datasets to a 3-level affordance. "Hard" driveability labels are then softened to model inter-class severity. Lastly, our loss weighting scheme selectively emphasizes the areas most relevant to navigation during learning.

A. From object semantics to driveability labels

We define a 3-level affordance to characterize the driveability [19] of a pixel:

- 🟩 *Preferable*: where we expect the robot to drive (paved roads or paths);
- 🟨 *Possible*, but not preferable: areas which are technically navigable but more challenging or less suitable, and would not be chosen as a first resort (e.g. grass, sand);
- 🟥 *Impossible* or undesirable: any part of the scene which is unreachable (e.g. the sky) or should be unconditionally avoided (obstacles, hazardous terrain).

This taxonomy is inspired by the action plausibility ratings proposed in [21]. For each pixel, we generate an affordance label by mapping its original semantic label (eg. car, sidewalk, tree, road) to a driveability level {🟥, 🟨, 🟩}. As discussed in [22], such a mapping from descriptive semantic labels to affordance is somewhat reductive as it does not take any contextual information into account - however, it remains a common approach [15], [16], since it enables fully-supervised learning without the need for manual affordance labelling.

In our experiments, we compare this 3-level definition to two common binary segmentation approaches mentioned in Section II: road vs. rest segmentation (🟨 level mapped to 🟥) and free space vs. obstacles segmentation (🟨 mapped to 🟩). A comparison is shown in Figure 2. The intermediate level 🟨 can serve as a fallback in the absence of a clear path in the scene, and leaves more flexibility at the planning level (e.g. if a robot has an off-road navigation target, or if an autonomous vehicle needs to park, change lanes, or overturn a car).

Fig. 2. Example of a pixel-annotated outdoor scene from the IDD dataset [20]. We map the original object classes to driveability levels.

B. From one-hot to soft ordinal labels

Intuitively, mis-classifying an area which is ■ *preferable* (e.g. the path) to drive on as ■ *impossible* should be penalized more heavily than classifying it as ■ *possible*. However, a standard one-hot encoding (Figure 3a) coupled with a categorical loss function do not capture this distinction during learning: mis-classifications are treated as equally severe regardless of the target. To incorporate a notion of pair-wise distance or severity between driveability levels, we opt for a soft labelling approach, which does not require any architectural changes and has shown to improve generalization in a wide range of tasks [23]. Specifically, we implement the Soft Ordinal vectors (or SORD) labelling scheme proposed in [24]: standard one-hot encoded labels are converted to a softmax-normalized probability distribution based on a ranking definition, such that the target class has the highest probability and the other probabilities encode a distance from the target class. Given a set of ranks $R = \{r_{impossible}, r_{possible}, r_{preferable}\}$ (one per driveability level), a SORD ground truth label $\mathbf{y}$ is generated based on a target rank r_t as follows:

$$y_i = \frac{\exp\left(-\phi(r_t, r_i)\right)}{\sum_{k \in R} \exp\left(-\phi(r_t, r_k)\right)} \quad \forall r_i \in R \qquad (1)$$

where $\phi(r_t, r_i)$ is a metric function which penalizes deviation from the target rank r_t. As inter-rank distances approach infinity, $\hat{\mathbf{y}}$ reduces to a one-hot encoded vector; as the distances approach 0, $\hat{\mathbf{y}}$ approaches a uniform probability distribution.

For this application, we consider a simple ranking definition between driveability levels: $R = \{$■ 1, ■ 2, ■ 3$\}$ (least to most driveable). We define the metric penalty function ϕ as the square log difference (SLD) $\phi(r_t, r_i) = |\log_e(r_i) - \log_e(r_t)|^2$, which reduces the penalty with increasing rank. Compared to absolute difference for instance, SLD shifts the middle rank ■ *possible* much "closer" to ■ *preferable* than to ■ *impossible*: in other words, the distinction between obstacles and driveable areas is much more clear-cut than the blurry line between driveable areas which are ■ *preferable* or not. Intuitively, this mirrors the ambiguity that a human annotator would face when labelling images: we are much less likely to hesitate when categorizing an area as obstacle vs. non-obstacle than when determining whether a driveable area is preferable or not.

Figure 3 shows the resulting asymmetrical SORD label encoding $\mathbf{y}$ for each of the 3 possible driveability targets, compared to a one-hot categorical encoding. Following [24], we then take the loss per pixel as the Kullback-Leibler (KL) di-
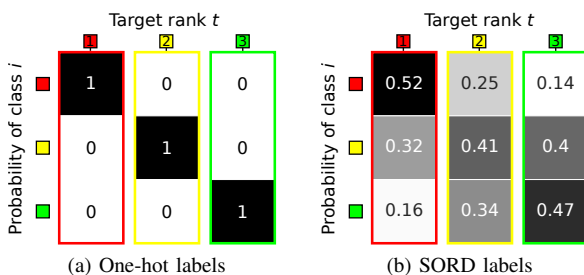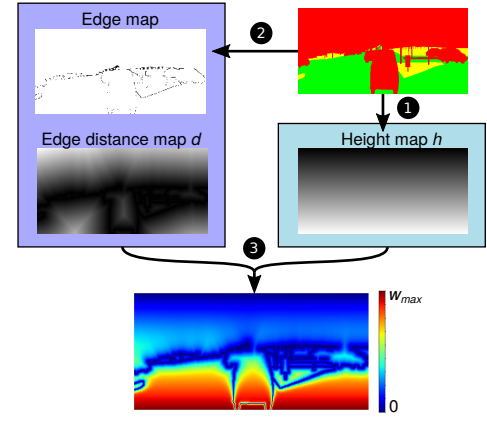

Fig. 4. Steps in the loss weight map computation, numbered and illustrated with a ground truth sample from the Kitti dataset [25].

vergence between the predicted class probability vector $\hat{\mathbf{y}}$ and the SORD label $\mathbf{y}$ from (1): $L_{KL}(\mathbf{y}||\hat{\mathbf{y}}) = \sum_{\forall r_i \in R} y_i \log_e \frac{y_i}{\hat{y}_i}$

C. Loss weighting

We argue that for navigation, detailed understanding of the entire scene is not necessary. Rather than giving each pixel equal contribution, we focus learning *away* from object contours which are difficult to learn [26], and towards areas in the camera's immediate vicinity which are critical to driving decisions. For selectively emphasizing relevant pixels during learning, we adapt the loss weighting scheme proposed in [27]. We adapt its formulation to our task such that boundary pixels are assigned the lowest weight, and we introduce a notion of image depth to distinguish between close-range and background elements. Given a pixel location $\mathbf{p} = [p_x, p_y]^T$ in the ground truth mask, we compute a weight map $w(\mathbf{p})$ which is applied to the loss per pixel via element-wise multiplication. The weight of a pixel depends on its Euclidean distance $d(\mathbf{p})$ to the closest segmentation boundary and on its vertical position (height) in the image $h(\mathbf{p})$:

$$w(\mathbf{p}) = h(\mathbf{p}) \cdot \left[1 - \exp\left(-\frac{d(\mathbf{p})}{1 + \beta(1 - h(\mathbf{p})^2)^2}\right)\right] \qquad (2)$$

where β is an experimentally defined constant. The height map $h(\mathbf{p})$ is used to scale the rate at which the pixel weight increases when moving away from a boundary, and as a pixel-wise multiplication factor which assigns higher weight to lower pixels. It serves as a naive placeholder for depth data, under the simple assumption that the lower a pixel in the image, the closer it is to the camera.

As illustrated in Figure 4, we generate a weight map $w(\mathbf{p})$ from a ground truth mask in three steps:

1) the height map $h(\mathbf{p})$ is pre-computed for all possible pixel locations based on the image height H as: $h(\mathbf{p}) = p_y/H$ such that pixels in the lowest row of the image have the value 1 and top row pixels have a value of 0.

2) for computing the edge distance map $d(\mathbf{p})$, we first perform edge detection on the gray-scaled ground truth

Fig. 3. Label class probabilities with a standard one-hot encoding (left) vs. the SORD labelling scheme (right).

mask, binarize the edge map, and apply a distance transform [28] with a 5×5 kernel.

3) the weight map $w(\mathbf{p})$ is then computed following (2), and min-max normalized to lie within $[0, w_{max}]$.

IV. EXPERIMENTAL SET-UP

A. Architecture and hyper-parameters

For pixel-wise classification, we pick SegNet [8] as a base network, similarly to [3]. Our variant applies drop-out (rate of 0.5) in the six deepest encoder and decoder blocks for regularization, and reduces the number of convolutional layers in each block to 2 (as opposed to 3 in the deepest blocks of VGG-16 [29]), resulting in a total of 20 convolutional layers. We measure a forward pass time of 32ms on the NVIDIA TITAN X for a single sample. In all our experiments, we train SegNet using Adam optimization [30] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) to minimize the KL divergence. Unlabeled/void pixels are ignored: the batch loss is computed as the sum of the loss per non-void pixel, divided by the number of non-void pixels in the batch. Samples are fed to the network in shuffled mini-batches of size 8, and the best model is selected based on minimal validation loss.

B. Cross-domain datasets

Our approach is entirely data-driven: accurate estimates of driveability in unconstrained environments require challenging samples to be included during training. For evaluating generalization to new environments with our method, we adopt a similar zero-shot cross-dataset strategy to [12]: models are trained on a combination of cross-domain datasets, and evaluated on a separate combination of datasets which have never been seen during training or validation.

We select outdoor scene understanding datasets with pixel-level annotations and RGB images captured by a vehicle or mobile robot, as outlined in Table I. For training, we include Cityscapes [31], a widespread benchmark featuring "clean" scenes, as well as more recent driving datasets covering a wide range of environmental conditions, sensor characteristics and geographical contexts including Mapillary [33], Berkeley DeepDrive (BDD) [32] and ACDC [34]. Outside of urban landscapes, RUGD [35], YCOR [37] and TAS500 [38] cover off-road grassy environments. Lastly, IDD [20] brings a unique challenge since it was captured in unstructured Indian traffic and rural scenes. For evaluation, we select 3 datasets with

TABLE I
CROSS-DOMAIN COMBINATION OF IMAGE SEGMENTATION DATASETS USED IN OUR ZERO-SHOT CROSS-DATASET EXPERIMENT.

scene type	**Training & validation** (# images)		**Testing** (# images)	
urban driving	Cityscapes [31]	(3,484)	Kitti [25]	(200)
	BDD [32]	(8,000)		
	Mapillary [33]	(20,000)		
	ACDC [34]	(2,006)		
unstructured / off-road	RUGD [35]	(5492)	Freiburg Forest [36]	(366)
	YCOR [37]	(1076)		
	TAS500 [38]	(540)		
mixed	IDD [20]	(8089)	WildDash [6]	(4256)

varying levels of difficulty. Kitti [25] is a small-scale benchmark of "clean" city driving scenes. Freiburg Forest [36] was captured by a mobile robot traversing forested trails, with some challenging illumination conditions, but no dynamic obstacles. WildDash [6] was specifically designed as a difficult test set for evaluating robustness to visual driving hazards in diverse environments. We use each dataset's official train/validation split during learning, and full datasets during testing, resulting in a total of 42,759 / 5,939 / 4,822 images for training/validation/testing respectively.

Note that these 11 datasets were annotated under different sets of semantic classes, but mapping their original object labels to a generic notion of driveability allows us to bridge this semantic gap during training and evaluation. During learning, each driveability level is informed by samples from all 8 training datasets. To counteract the imbalance in dataset size, similarly to [12], mini-batches are constructed with an equal number of samples (1 in our case) from each dataset.

C. Data preparation

Input color: While it is commonplace to preserve color information in input images for scene understanding [39], [7], we speculate that color may add unnecessary or distracting information when trying to learn such an abstract concept as driveability. Thus, we investigate the importance of color in our experiments by comparing the standard RGB representation with a single-channel grayscale input.

Input size: This is also an important consideration, with a trade-off between computational cost and segmentation detail. Resizing images to fixed dimensions is common practice, especially when learning from a combination of datasets [12]. For our affordance learning task, retaining a high level of detail is not a primary concern, but incorporating global context is crucial [16]. Therefore, we opt for a small input resolution of 240×480 - the same width as in [8], but with a wider aspect ratio to accommodate wide-FOV datasets.

Data augmentation: During training, input samples are randomly augmented on-the-fly with geometric (horizontal flip, rotation, crop, perspective transform, grid-based distortion) and photometric (brightness, contrast, tone curve and color manipulation) transformations, each having a probability of 0.5. See [40] for a detailed description.

D. Training procedure

Pre-training: As a starting point, we train SegNet on Cityscapes to predict the 30 original object classes in the dataset [31], using an initial learning rate of 10^{-3}. We refer to this model as Cityscapes$_{obj}$. Note that this model is trained under a standard learning scheme (one-hot labels, uniformly weighted loss), and thus can be substituted with other pre-trained segmentation models.

Driveability via transfer learning: To learn 3-level driveability from a combination of datasets, the last convolutional layer of Cityscapes$_{obj}$ is re-initialized with 3 output channels and trained under the SORD labelling scheme with an initial learning rate of 10^{-4} until convergence.

Loss weighting (LW) is implemented as a final training stage to consolidate the segmentation, while maintaining the same labelling scheme. Weight maps are generated with $w_{max} = 10$ (same as [27]) and $\beta = 30$.

E. Evaluation metrics

Classification metrics: In the context of autonomous navigation, under-segmentation of obstacles and over-segmentation of driveable areas pose particular safety risks. Therefore, aligning with [3], we select two segmentation metrics of interest: pixel-wise recall (R) for the 🟥 level, and precision (P) for 🟩. We also introduce a *weighted* version of these metrics R_w and P_w which weighs each pixel based on the LW map, thus emphasizing the most navigation-relevant areas.

Regression metrics: The segmentation metrics above do not capture inter-rank distances. Therefore, similarly to [24], we report Root-mean-square error (RMSE) to evaluate the degree of error between predicted and ground truth driveability levels, with heavier penalty for large error (confusion between 🟩 and 🟥 levels). Based on [41], we also compute a measure of *mistake severity* (MS) as the mean absolute error of *incorrect* predictions; note that MS is fully decoupled from accuracy. We normalize MS per pixel, such that it ranges from 0 to 1: mis-classifying a pixel as 🟨 yields a MS of 0, while confusing the 🟥 and 🟩 levels yields a MS of 1.

V. Results

We first validate our 3-level driveability definition and learning scheme. We then benchmark our approach against the state-of-the-art and an object-based baseline, and comment on the effect of input color to motivate the use of grayscale images in our experiments. Lastly, we show some failure cases and probabilistic predictions of our model.

3-level driveability vs. binary segmentation: For comparison, we train a cross-domain model with standard one-hot labels scheme for each of the three class definitions presented in Figure 2. Table II compares the models' performance in terms of segmentation error (calculated with ranks $R = \{$ 🟥 1, 🟩 3 $\}$ for the binary segmentation baselines). The driveability model consistently achieves the lowest segmentation error, followed by free space segmentation in urban scenes and road/path segmentation in mixed or forested scenes. Figure 6 shows the qualitative advantage of our 3-level driveability definition over these binary segmentation approaches. The driveability model identifies 🟨 *possible* driveable ground in off-road or open areas, while also distinguishing 🟩 *preferable* areas when there is a clear path in the scene.

TABLE II
RMSE OF ONE-HOT CROSS-DOMAIN DRIVEABILITY MODELS COMPARED TO BINARY SEGMENTATION BASELINES.

Segmentation class definition	Cross-domain (val)	Kitti	Freiburg	WildDash
Road/path 🟩 vs. 🟥 rest	0.412	0.423	0.310	0.490
Free space 🟩 vs. 🟥 obstacles	0.437	0.377	0.445	0.407
3-level driveability 🟩🟨🟥	**0.283**	**0.311**	**0.284**	**0.402**

Navigation-oriented learning scheme: Figure 5 shows predictions by our proposed cross-domain SORD+LW model on out-of-dataset samples from the three unseen test sets, and we include a video showing additional qualitative results as supplementary material. The model produces sensible driveability estimates across a wide range of navigation scenarios, with variations in scene layout and contents, lighting and weather conditions, as well as camera characteristics and perspectives. Table III reports quantitative performance, with a comparison between a model trained under our proposed training scheme (Section IV-D), and a standard model trained with one-hot labels in the transfer learning stage and uniformly-weighted pixel-wise loss. Table III shows our navigation-oriented learning scheme to be effective at bringing down RMSE on the validation set and on every unseen test dataset, with SORD labelling reducing mistake severity by almost 50% compared to a standard one-hot model. The addition of LW consistently improves segmentation in safety-critical areas, as indicated by the weighted obstacle 🟥 recall and 🟩 precision scores. We note the most significant quantitative improvement in generalization performance for Freiburg Forest's highly unstructured environments, where our method helps disambiguate the fuzzy transitions between path, grass and surrounding vegetation without getting lost in details. Looking at the overall aspect of the segmentation across test samples, we find that SORD labelling produce smoother contours and less spotty segmentation, and encourages cautious, low-stakes predictions especially for ambiguous border pixels. As can be seen in the examples of Figure 7, this visually manifests as a layer of 🟨 pixels around non-driveable areas, rather than sharp transitions between 🟥 and 🟩 levels. While this deviates from what ground truth masks look like, we consider it beneficial for navigation, since it essentially adds a safe margin around obstacles. LW, which concentrates learning away from distant details towards close-range and non-border areas, results in a more approximate but cohesive segmentation.

TABLE III
SAME-DATASET AND ZERO-SHOT GENERALIZATION PERFORMANCE OF OUR CROSS-DOMAIN DRIVEABILITY MODELS.

Test data	Learning	🟥 R %	🟥 R_w %	🟩 P %	🟩 P_w %	MS %	RMSE
Cross-domain (validation)	one-hot	**98.41**	97.77	93.20	94.32	15.28	0.283
	SORD	98.12	97.48	93.04	93.70	**7.94**	**0.276**
	SORD + LW	98.33	**97.88**	**93.75**	**94.71**	9.15	0.278
Kitti	one-hot	98.72	98.17	87.86	90.18	10.79	0.311
	SORD	98.42	97.95	**89.55**	90.82	**5.79**	**0.293**
	SORD + LW	**98.79**	**98.64**	89.44	**91.09**	7.43	0.304
Freiburg	one-hot	94.15	89.98	85.19	88.27	1.75	0.284
	SORD	96.12	94.07	80.38	83.15	**0.50**	0.269
	SORD + LW	**97.57**	**96.98**	**86.29**	**89.26**	0.69	**0.258**
WildDash	one-hot	**98.71**	98.07	91.63	93.68	30.27	0.402
	SORD	98.18	97.46	93.95	95.25	**15.64**	**0.369**
	SORD + LW	98.58	**98.08**	**94.01**	**95.48**	18.54	0.380

Comparison to the state-of-the-art: The closest candidate for comparison are the segmentation results in [3] for the general *obstacle* class, defined similarly to our 🟥 *impossible* level. The authors train a SegNet model on weakly labelled images from Kitti Raw [25], and evaluate it on the Kitti Object & Tracking datasets (over 85k obstacle bounding boxes). We evaluate our cross-domain SORD+LW model with the same procedure and metrics in Table IV. Note that the pixel recall metric considers the whole bounding box area, while instance
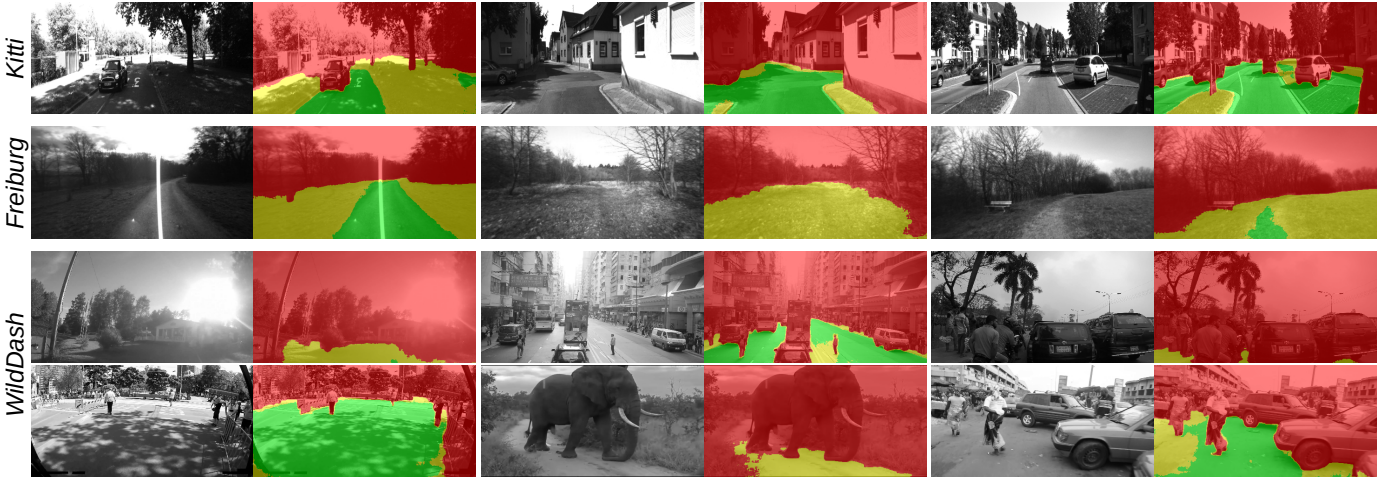
Fig. 5. Examples of out-of-dataset predictions by the proposed model, trained on the cross-domain dataset with soft driveability labels and loss weighting.

recall requires a certain percentage of pixels in a box to be predicted as ■ *obstacle* for it to be considered detected. Our model achieves state-of-the-art object detection ($> 50\%$ instance recall) on this dataset, despite not having seen any

Road segmentation fails to produce viable predictions in open areas

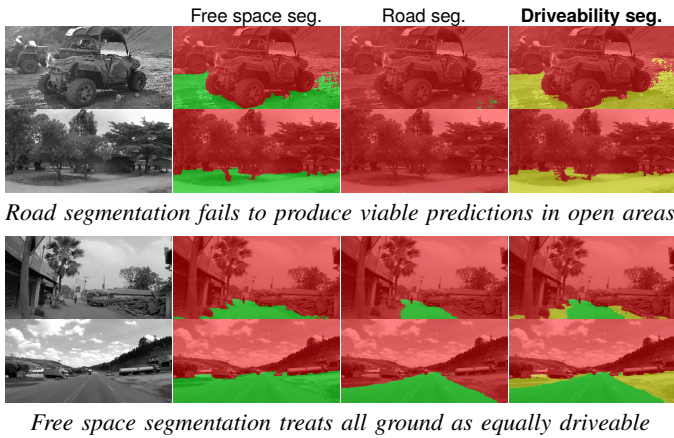*Free space segmentation treats all ground as equally driveable*

Fig. 6. Out-of-dataset predictions by one-hot cross-domain models on WildDash samples, trained under different class definitions.
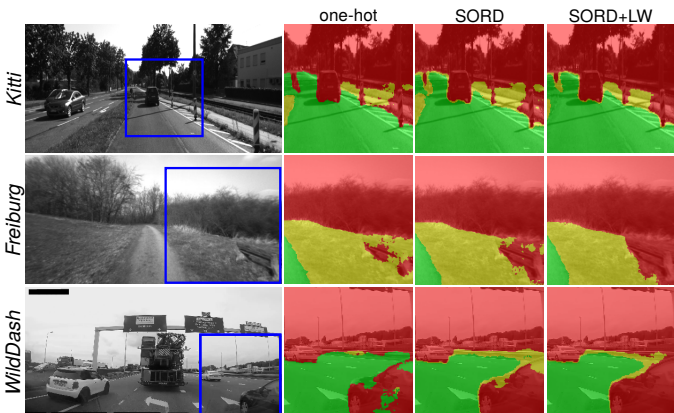

Fig. 7. Selected crops of out-of-dataset predictions by the cross-domain driveability model, showing the qualitative effect of the soft labelling and loss weighting training schemes.

Kitti images during training. The lower pixel recall and $> 75\%$ instance recall scores of our model can be attributed to the granularity of our labels and predictions compared to [3]'s coarse, "boxy" segmentations, which naturally take up a larger portion of the ground truth bounding boxes on this benchmark.

In terms of qualitative results, while [3] fails to predict viable path segmentations in ambiguous road configurations (e.g. tight turns in intersections) and does not show results in road-free scenes, the examples in Figures 5 and 6 show that our model successfully identifies *preferable* ■ driveable areas even in the absence of a structured lane, while falling back to the ■ level in open unstructured areas.

Comparison to an object-based single-dataset baseline: The conventional approach to semantic scene segmentation consist of learning object descriptions on a single dataset. In contrast, our driveability definition allows us to combine heterogeneously labelled datasets during training. To show the benefit of our approach for generalization to new scenes, we take Cityscapes$_{obj}$ as an experimental baseline, and map its object-based predictions to driveability levels for evaluation. We then apply the transfer learning and LW stages to learn driveability on Cityscapes (Cityscapes$_{driv}$). Table V compares our cross-domain model with these two single-dataset base-

TABLE IV
OBSTACLE SEGMENTATION RESULTS ON THE KITTI OBJECT & TRACKING DATASETS (EVALUATION PROCEDURE FROM [3]).

	Seen Kitti images?	Input	Pixel recall	Instance recall ($>50\%$)	Instance recall ($>75\%$)
[3]	24,443	RGB	**93.53%**	99.55%	**97.93%**
ours	✗	Gray	88.09%	**99.74%**	96.34%

TABLE V
RMSE OF OUR MODEL COMPARED TO SINGLE-DATASET BASELINES.

Train data	Learning	Cityscapes (val)	Kitti	Freiburg	WildDash
Cityscapes$_{obj}$	one-hot	0.210	0.353	0.660	0.491
Cityscapes$_{driv}$	SORD+LW	**0.207**	0.317	0.491	0.402
Cross-domain$_{driv}$	SORD+LW	0.226	**0.304**	**0.258**	**0.380**

lines. Comparing the two Cityscapes models, we note that learning driveability with SORD+LW consistently reduces same-dataset and generalization error compared to a one-hot object-based approach, with the most notable improvement for Cityscapes → Freiburg Forest transfer. Extending the findings in [12], our results show cross-domain learning to be beneficial for segmenting driveability in out-of-dataset images: learning driveability across a diverse 8-dataset combination reduces generalization error across all 3 unseen test datasets. While the performance of Cityscapes models drops when faced with Freiburg Forest's unstructured scenes, the cross-domain models maintain an RMSE below 0.4 (and pixel accuracy above 90%) across all test sets.

Does color matter? On unseen samples from a known dataset or from a dataset captured in ideal urban scenarios (Cityscapes and Kitti in Table VI and Figure 8), color brings a small improvement in segmentation. However, interestingly, we note a significant performance gap in favour of grayscale models when evaluating zero-shot generalization to challenging new scenes (Freiburg Forest and WildDash). While grayscale models are blind to dataset-specific color palettes, RGB models seem to make color-class associations (e.g. dark gray for the driveable road, bright red for cars) which may not hold in different outdoor environments (e.g. brown paths in Freiburg Forest, red reflections on the road).

TABLE VI
EFFECT OF INPUT COLOR ON ■ RECALL FOR ONE-HOT MODELS.

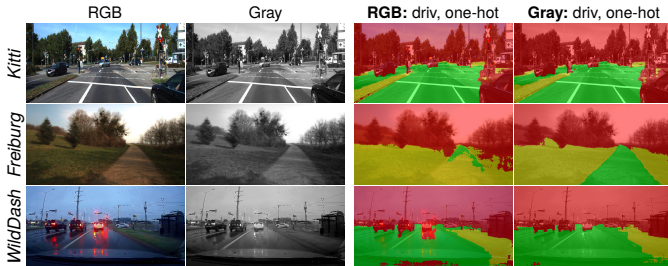| *Train data*	*Input*	Cityscapes (val)	Kitti	Freiburg	WildDash
Cityscapes$_{obj}$	RGB	**99.51**	**99.30**	89.11	92.62
	Gray	98.91	97.96	**92.53**	**93.36**
Cross-domain$_{driv}$	RGB	**99.33**	**98.91**	91.55	92.36
	Gray	99.32	98.72	**94.11**	**98.71**

Fig. 8. Qualitative comparison of out-of-dataset predictions by the cross-domain model trained on RGB vs. grayscale input images.

Failure cases: As indicated by its performance on the Kitti Object & Tracking datasets (Table IV), our model reliably detects common obstacles in ideal road scenes. However, looking at predictions on the challenging WildDash benchmark, we note that the model inherits the limitations of RGB vision, with poor results in extreme weather or illumination conditions. In the first row of Figure 9, the images are too dark and foggy (left) or rainy/snowy (right) to estimate driveability, especially through a windshield with the car's dashboard blocking a large portion of the image. In addition, the bottom row examples of Figure 9 suggests that distinguishing flat textures with obstacles can be tricky in case of small, unusual

Blinded by the weather
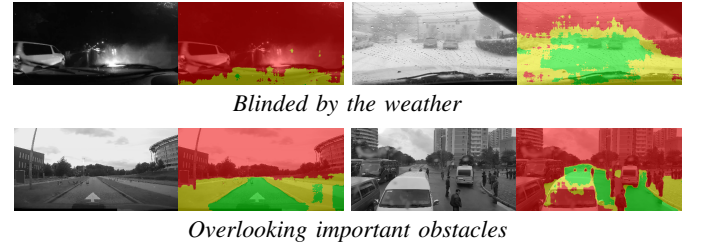

Overlooking important obstacles

Fig. 9. Examples of unacceptable predictions on WildDash [6] images.

objects (eg. ducks in the bottom left), or structures aligning with the road configuration (bottom right). We expect that the incorporation of depth cues for driveability estimation would help disambiguate between road irregularities such as manholes, shadows, lane markings (all of which are considered driveable ■) from actual hazards on the robot's path.

Probabilistic affordance maps for planning: In our evaluation, we have taken predictions as the argmax of the output layer, resulting in crisp 3-level segmentation. Instead, since our model predicts ordered *ranks*, predictions can also be taken as the expected value $\sum_{\forall r_i \in R} r_i \hat{y}_i$ - resulting in probabilistic affordance maps as shown in Figure 10, with smooth transitions between driveability levels.
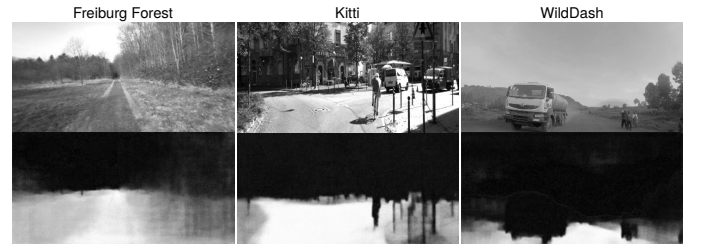

Fig. 10. Probabilistic driveability estimation by the Cross-domain SORD+LW model on out-of-dataset samples.

VI. DISCUSSION AND FUTURE WORK

Driveability estimation: By defining a simple ground truth mapping between object classes and driveability, we bridge the semantic gap between datasets to allow joint cross-domain training while bypassing the need for manual labelling. However, while this mapping can easily be adapted to the task at hand and robot capabilities, it remains blind to contextual information: the sidewalk may be the *preferable* path for a "pedestrian" robot, but only a *possible* last resort for an autonomous vehicle driving on the road; a dirt path may be *preferable* to drive on in a forest, but not a route of choice in a city. Incorporating such scene- and application-dependent context during learning is an important direction for further research. Future work will also investigate how driveability can be learned from multi-modal image data to improve scene understanding in poor visibility.

Towards robot navigation: To investigate the practical implications of our approach for open-world robotic navigation, future work will incorporate our probabilistic driveability maps (Figure 10) into a severity-aware planning module, which aims to maximize driveability along sampled trajectories. To this

end, our pixel-wise affordances could be projected into 3D space using depth and odometry data, and used as a cost for graph-based path planning - as demonstrated in [4] and [14]. For safe navigation in urban environments, our method should also be complemented with recognition of traffic cues [7]. Extending our 3-level definition to distinguish between static background and moving obstacles may also be beneficial.

VII. Conclusion

We have presented a simple yet effective method for learning pixel-wise driveability across outdoor scenes for open-world robotic navigation. Compared to existing approaches which treat all pixels and mistakes equally and are constrained to a specific domain, our severity-aware affordance learning framework allows cross-dataset training and tailors the label and loss formulation to navigation, with quantitative and qualitative improvements in segmentation of unseen environments.

References

[1] B. Zhou, P. Krähenbühl, and V. Koltun, "Does computer vision matter for action?" *Science Robotics*, vol. 4, 2019.

[2] M. Hassanin, S. Khan, and M. Tahtali, "Visual affordance and function understanding: A survey," *ACM Comput. Surv.*, vol. 54, no. 3, Apr. 2021.

[3] D. Barnes, W. Maddern, and I. Posner, "Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017, pp. 203–210.

[4] L. Wellhausen, A. Dosovitskiy, R. Ranftl, K. Walas, C. Cadena, and M. Hutter, "Where should i walk? predicting terrain properties from images via self-supervised learning," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1509–1516, 2019.

[5] G. Kahn, P. Abbeel, and S. Levine, "Badgr: An autonomous self-supervised learning-based navigation system," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1312–1319, 2021.

[6] O. Zendel, K. Honauer, M. Murschitz, D. Steininger, and G. F. Dominguez, "Wilddash - creating hazard-aware benchmarks," in *Proc. Eur. Conf. Comput. Vis.*, 9 2018.

[7] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Foundations and Trends® in Computer Graphics and Vision*, vol. 12, no. 1–3, pp. 1–308, 2020.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[9] M. Teichmann, M. Weber, M. Zöllner, R. Cipolla, and R. Urtasun, "Multinet: Real-time joint semantic reasoning for autonomous driving," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 1013–1020.

[10] D. Levi, N. Garnett, E. Fetaya, and I. Herzlyia, "Stixelnet: A deep convolutional network for obstacle detection and road segmentation." in *Proceedings of the British Machine Vision Conference 2015, BMVC*, vol. 1, no. 2, 2015, p. 4.

[11] A. Behl, K. Chitta, A. Prakash, E. Ohn-Bar, and A. Geiger, "Label efficient visual abstractions for autonomous driving," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2020, pp. 2338–2345.

[12] J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, "Mseg: A composite dataset for multi-domain semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2876–2885.

[13] M. Broome, M. Gadd, D. De Martini, and P. Newman, "On the road: Route proposal from radar self-supervised by fuzzy lidar traversability," *AI*, vol. 1, no. 4, pp. 558–585, 2020.

[14] W. Qi, R. T. Mullapudi, S. Gupta, and D. Ramanan, "Learning to move with affordance maps," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[15] A. Roy and S. Todorovic, "A multi-scale cnn for affordance segmentation in rgb images," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 186–201.

[16] T. Lüddecke, T. Kulvicius, and F. Wörgötter, "Context-based affordance segmentation from 2d images for robot actions," *Robotics and Autonomous Systems*, vol. 119, pp. 92–107, 2019.

[17] H. Min, C. Yi, R. Luo, J. Zhu, and S. Bi, "Affordance research in developmental robotics: A survey," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 4, pp. 237–255, 2016.

[18] X. Liu, W. Ji, J. You, G. El Fakhri, and J. Woo, "Severity-aware semantic segmentation with reinforced wasserstein training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12 563–12 572.

[19] J. Guo, U. Kurup, and M. Shah, "Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3135–3151, 2020.

[20] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. V. Jawahar, "Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *Proc. IEEE Winter Conf. App. Comput. Vis (WACV)*, 2019, pp. 1743–1751.

[21] T. Lüddecke and F. Wörgötter, "Fine-grained action plausibility rating," *Robotics and Autonomous Systems*, vol. 129, p. 103511, 2020.

[22] P. Ardón, Èric Pairet, K. S. Lohan, S. Ramamoorthy, and R. P. A. Petrick, "Affordances in robotic tasks - a survey," 2020.

[23] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[24] R. Díaz and A. Marathe, "Soft labels for ordinal regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4733–4742.

[25] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *Int. J. Robot. Res.*, 2013.

[26] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8848–8857.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Med. Image. Comput. Comput. Assist. Interv. (MICCAI)*, 2015, pp. 234–241.

[28] G. Borgefors, "Distance transformations in digital images," *Comput. Vis. Graph. Image Process.*, vol. 34, pp. 344–371, 1986.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds., 2015.

[31] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[32] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, June 2020.

[33] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017.

[34] C. Sakaridis, D. Dai, and L. V. Gool, "ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding," 2021.

[35] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2019, pp. 5000–5007.

[36] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, "Deep multi-spectral semantic scene understanding of forested environments using multimodal fusion," in *2016 International Symposium on Experimental Robotics (ISER)*, 2017, pp. 465–477.

[37] D. Maturana, P.-W. Chou, M. Uenoyama, and S. Scherer, "Real-time semantic mapping for autonomous off-road navigation," in *Field and Service Robotics*. Springer, 2018, pp. 335–350.

[38] K. A. Metzger, P. Mortimer, and H.-J. Wuensche, "A fine-grained dataset and its efficient semantic segmentation for unstructured driving scenarios," in *Proc. Int. Conf. Pattern Recog. (ICPR)*, 2021.

[39] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017, pp. 4644–4651.

[40] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020.

[41] L. Bertinetto, R. Mueller, K. Tertikas, S. Samangooei, and N. A. Lord, "Making better mistakes: Leveraging class hierarchies with deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12 503–12 512.