Striving for Less: Minimally-Supervised Pseudo-Label Generation for Monocular Road Segmentation

François Robinet*[,1], Yussef Akl[1], Kaleem Ullah[1,2], Farzad Nozarian[3], Christian Müller[3] and Raphaël Frank[1]

***Abstract*—Identifying traversable space is one of the most important problems in autonomous robot navigation and is primarily tackled using learning-based methods. To alleviate the prohibitively high annotation-cost associated with labeling large and diverse datasets, research has recently shifted from traditional supervised methods to focus on unsupervised and semi-supervised approaches. This work focuses on monocular road segmentation and proposes a practical, generic, and minimally-supervised approach based on task-specific feature extraction and pseudo-labeling. Building on recent advances in monocular depth estimation models, we process approximate dense depth maps to estimate pixel-wise road-plane distance maps. These maps are then used in both unsupervised and semi-supervised road segmentation scenarios. In the unsupervised case, we propose a pseudo-labeling pipeline that reaches state-of-the-art Intersection-over-Union (IoU), while reducing complexity and computations compared to existing approaches. We also investigate a semi-supervised extension to our method and find that even minimal labeling efforts can greatly improve results. Our semi-supervised experiments using as little as 1% & 10% of ground truth data, yield models scoring 0.9063 & 0.9332 on the IoU metric respectively. These results correspond to a comparative performance of 95.9% & 98.7% of a fully-supervised model's IoU score, which motivates a pragmatic approach to labeling.**

***Index Terms*—Deep Learning for Visual Perception, Computer Vision for Transportation, AI-Based Methods**

I. Introduction

To be able to safely navigate the world, autonomous robots have to be capable of perceiving their environment to detect traversable free space. This work focuses on road detection for vehicles equipped with a single forward-facing camera. While this task has traditionally been solved using supervised segmentation methods, these techniques suffer the drawbacks of laborious labeling cost (1.5h/frame for fine annotations [1]), as well as test-time distribution shift due to the wide variety of environments and weather conditions [2]. To alleviate these issues, the community has recently focused on unsupervised and semi-supervised alternatives.

Although unsupervised approaches lead to respectable results, they are not reliable enough to enable the safe operation of autonomous agents [3]–[7]. Semi-supervised methods offer a compromise by training on restricted subsets of manually-labeled data and exploiting unlabeled data through self-supervised consistencies or pseudo-labeling [8]–[10]. Our work explores this trend by proposing a practical, effective, and generic pseudo-labeling framework. Our method reduces the quantity of labeled data required by exploiting prior knowledge of road semantics as well as recent advances in monocular depth estimation [11].

Our contributions can be summarized as follows:

1) We introduce the use of a monocular depth network to estimate pixel-wise Road-Plane Distances (RPD) using the v-disparity algorithm with a novel filtering method.
2) We use RPD maps in a pseudo-labeling pipeline to reach state-of-the-art IoU for unsupervised road segmentation, with reduced complexity and computations compared to existing work.
3) We propose a semi-supervised extension, which uses as little as 1% & 10% of ground truth data while reaching 0.9063 & 0.9332 IoU respectively. These results correspond to 95.9% & 98.7% respective performance of a comparable fully-supervised model on the same metric.

In addition, we also make our code and trained models freely available online. While our work focuses on urban scenes, the use of an unsupervised monocular depth network and RPD maps are generic enough to allow the same technique to be used in other robotic navigation scenarios, such as indoors or less-constrained outdoor environments.

The content of this work is organized as follows: In Section II, we review existing works on unsupervised and semi-supervised road segmentation. In Section III, we introduce our method for computing RPD maps, our unsupervised and semi-supervised approaches. In Section IV, we define evaluation metrics and detail our use of the Cityscapes dataset [1]. In Section V, we carry out experiments, and analyze the respective results. Finally, we summarize our contributions and share further research directions.

II. Related works

This section presents a brief overview of road segmentation approaches that learn pixel-wise free space representation from

Manuscript received: February 24th, 2022; Revised June 15th, 2022; Accepted July 18th, 2022.

This paper was recommended for publication by Editor Cesar Cadena upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by the Fonds National de la Recherche, Luxembourg (MASSIVE Project). The authors also thank Foyer Assurances Luxembourg for their support.

*Corresponding author

[1]Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg

[2]University of Saarland, Saarbrücken

[3]German Research Center for Artificial Intelligence (DFKI), Saarbrücken

Digital Object Identifier (DOI): see top of this page.

data. While related works rely on video sequences to learn sparse point cloud maps [12], or to segment obstacle footprints using structure-from-motion [13], we focus our attention on unsupervised and semi-supervised methods that learn road masks using single images.

A. Unsupervised Road Segmentation

Most recent unsupervised road segmentation methods rely on generating approximate pseudo-labels using prior knowledge about road geometry or semantics. These approximate road masks can then be used as targets to train a statistical model to generalize past the noise that they contain. One successful example, and an inspiration for this work, is the use of the v-disparity algorithm [14]. This method uses disparity maps to estimate a flat road plane by making the generic assumption that the road-camera distance linearly increases as the road recedes to the horizon. Different attempts have exploited this representation of the road, using disparity maps computed from stereo-pairs [3, 5]. Our work differs in that we propose to approximate these disparity maps using a monocular depth estimation network which can be trained without supervision [11]. More direct use of depth measurements can also be beneficial, as shown through the use of RGB-D inputs to improve performance in indoor perception [15]. Also exploiting geometric information present in depth maps, the work from [16] uses a trained network specifically designed to extract features from RGB-D images by fusing depth feature maps at various stages of encoding. Rather than explicitly relying on scene geometry, another successful method is to identify the road by over-segmenting frames into superpixels and extracting feature vectors for each superpixel using a network trained for generic image classification. Superpixels can then be clustered in feature-space before using a spatial prior to identify the cluster corresponding to the road [4, 6, 7]. Our unsupervised approach unifies the geometrical and semantic approaches by combining features extracted from a v-disparity representation with cues obtained from over-segmenting the RGB space into superpixels.

B. Semi-Supervised Segmentation

Recent advances in semi-supervised segmentation can mainly be divided in two distinct categories of approaches: consistency training and self-training. Consistency training obtains a loss by making the assumption that the output of a model should be invariant to perturbations of inputs that do not affect their semantics. Examples that fall into this category are Unsupervised Data Augmentation [17], Virtual Adversarial Training [18] and Cutmix [19]. The semi-supervised part of this work adopts the second approach of self-training, in particular through the use of a mix of ground truth labels and noisy pseudo-labels that are generated for unlabeled data [8, 9]. Recent research has shown that over-parameterized neural networks can generalize past some of the label noise using Stochastic Gradient Descent (SGD) schemes combined with early stopping [20], but more advanced schemes have been devised to explicitly deal with this noise, such as Mean Teacher [10] and Co-Teaching [21]. For a comprehensive overview of semi-supervised techniques that cope with noisy labels in image analysis, we refer the reader to the survey in [22].

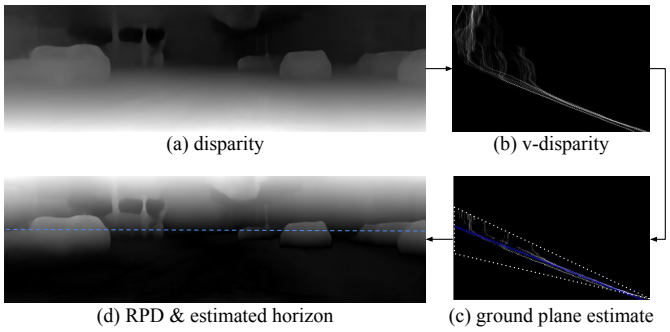

Fig. 1: Estimation of Road Plane Distance (RPD) from dense disparity maps using robust line fitting in v-disparity space.

III. Methodology

In this section, we start by how we obtain our dense Road Plane Distance (RPD) maps. We then show how we leverage these estimated distances to train both unsupervised and semi-supervised models.

A. Estimating Road Plane Distance (RPD)

To estimate the road plane, we rely on the v-disparity algorithm, which was first proposed in the context of obstacle detection [14]. The algorithm does not operate on raw RGB images, but instead takes dense disparity maps as inputs. Rather than relying on depth reconstruction using stereo pairs [5, 23], we propose to use a monocular depth estimation network to estimate such disparity maps from a single view of the scene. For this purpose, we have chosen a Monodepth2 network [11] that was trained on the KITTI dataset [24] without any ground truth labels using Structure-from-Motion [25].

Starting from a disparity map with dimensions (H, W) in Figure 1(a), a v-disparity map with dimension (H, B) is obtained by creating a B-bin histogram of disparity values in each row, as shown on Figure 1(b). The road detection procedure builds on the following intuition: oblique planes in the disparity input are mapped to straight lines in v-disparity space. Assuming that free space is the dominant planar region in the disparity input, one can therefore approximate it through line-fitting in v-disparity space. This process is illustrated on Figure 1(c). Since obstacles appear as vertical lines in v-disparity space, it is important to filter data before attempting line fitting. We follow [3] and only keep bin values belonging to the 95% percentile in each row. Since the road plane maps to a line in v-disparity space, the intercept of that line corresponds to an estimate of the horizon line in the original image. Exploiting prior knowledge about camera placement, we further restrict the area of interest to only match common horizons lines appearing on the frame between 0.8 and 0.4 relative heights. We fit the filtered v-disparity points using RANSAC linear regression [26].

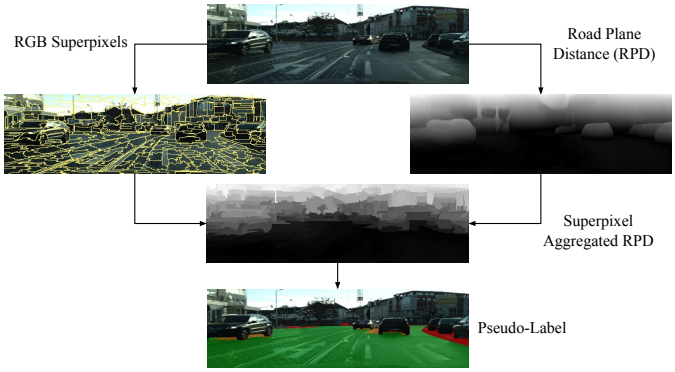

Fig. 2: Unsupervised pseudo-labels generation. RPD quantiles are computed over each RGB superpixels, and the result is thresholded and filtered to obtain a road estimate.

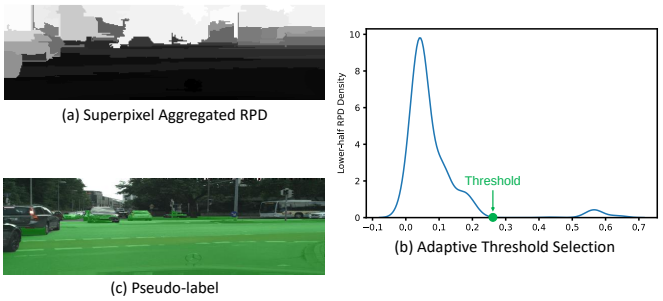

Fig. 3: Adaptive thresholding procedure.

As opposed to prior work [3, 5, 14], we don't back-project the fitted plane to the disparity image by reverting the histograms to obtain a road plane estimate. Rather, for each pixel in the disparity space, we estimate its elevation relative to the ground plane by computing the horizontal distance between its v-disparity projection and the fitted ground line. We call the result a Road Plane Distance map (RPD) and represent it on Figure 1(d).

B. Unsupervised Road Segmentation

In this section, we propose the use of RPD maps to generate approximate pseudo-labels of the road class. Our unsupervised pseudo-label generation method is illustrated on Figure 2.

Superpixel cues: Since object parts that touch the ground will contain low RPD values near the ground, directly thresholding RPD values cannot recover precise class boundaries. Instead, we rely on superpixel segmentation in RGB space to recover crisp boundaries. For a given RGB input, we generate a normalized RPD map and compute a superpixel segmentation using the Felzenszwalb method with a scale of 50 and a minimal size of 500 pixels [27]. We then aggregate the RPD values over each RGB superpixel using its 90% RPD quantile. We select the 90% quantile rather than the maximum value in an effort to capture the highest RPD values without being overly affected by possible outliers. An analysis of the impact of quantile choice is provided in Appendix A.

Adaptive RPD thresholding: In order to obtain pseudo-labels, we apply a threshold on superpixel aggregated RPD maps. Rather than choosing a fixed threshold to apply over all frames, we propose an adaptive procedure based on the distribution of RPD values in each frame. We make the assumption that most road pixels lie in the lower half of the frame, and compute a frame-specific threshold based on their distribution. We compute a smooth approximation of the distribution of RPD values in the lower portion of the frame using gaussian kernel density estimation. Since road superpixels have uniform and low RPD values, they will tend to form the largest peak in this distribution, and we identify the threshold as the first local minimum following this peak. In order to obtain our final pseudo-labels, we also remove any pixel lying above the estimated horizon from the road prediction.

Unsupervised model training: We use the pseudo-labels as targets to train a model to segment the road directly from an RGB input. Recent work has shown that deep neural networks trained via stochastic gradient descent exhibit surprising robustness to noise in their training targets [20], and such procedure has repeatedly been proven beneficial to road segmentation [4]–[7]. The network architecture and training procedure are detailed in Section V.

C. Semi-Supervised Road Segmentation

Although RPD maps are useful to detect the ground plane, the road class is more restrictive since it should not contain things like sidewalks or grass patches, even though they might lie in the same plane. Since RPD maps do not contain this information, we propose to combine them with RGB frames and to learn from a minimal number of ground truth annotations.

Pseudo-Label Generator (PLG): To generate pseudo-labels in the semi-supervised scenario, we train a PLG network to predict road masks from RPD maps and RGB frames. We concatenate them and use them as inputs to our PLG, which will allow it to learn to segment the road from only dozens of ground truth samples. This process is illustrated on Figure 4.

Training from Semi-Supervised Pseudo-Labels: As in the unsupervised case from Section III-B, an additional model can be trained using the semi-supervised pseudo-labels as targets in order to obtain road masks from RGB inputs. For frames for which ground truth was used at pseudo-label generation time, we also use pixel-wise ground truth as our target. Details about the model and training procedure used are available in Section V.

IV. Evaluation

A. Dataset

Our experiments leverage the Cityscapes dataset, which provides pixel-wise ground truth labels for 30 visual classes in 5000 frames [1]. Since the official test set has no public annotation, we follow prior work [5]–[7] and use the 500 frames of the validation set as our test set. Our training and validation sets are obtained by randomly splitting the Cityscapes training set into 2380 training and 595 validation

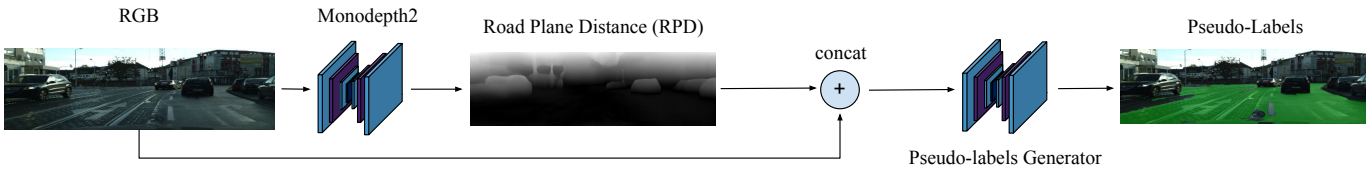

Fig. 4: Semi-supervised pseudo-labels generation procedure. We train a pseudo-label generator to predict pseudo-labels using RGB frames and RPD maps. The network learns using a small amount of ground-truth annotated frames.

frames. For frames that have pixels labelled as *road*, we consider only this class to denote the road. The dataset also contains 1.6% of frames with no *road* pixel. For these frames only, we verified that the *ground* annotation was used in its place and used it instead of *road*.

B. Evaluation Metrics

We assess prediction quality using the standard Intersection-over-Union (IoU) metric, as well as Precision and Recall. The Cityscapes dataset includes 6 *void* classes which cannot be assigned to road or non-road, such *unlabeled*, *out of the region of interest* or *ego-vehicle*. We follow the standard Cityscapes benchmark procedure and we ignore pixels corresponding to such classes at evaluation time using binary evaluation masks. These masks are considered part of the ground truth and are only used for evaluation.

V. Results

This section outlines the set of experiments carried out to benchmark the proposed methods. We detail our network architectures and training procedure before presenting results.

A. Network Architectures

Competing semi-supervised approaches are often focused on generic semantic segmentation rather than road segmentation, or use other datasets than Cityscapes as benchmarks [28]–[32]. The few semi-supervised approaches that publish metrics for the *road* class use a wide variety of network architectures, input resolutions, computational requirements, annotations (all classes or road only) and data splits, making direct comparison difficult. To allow for fair comparison and analysis of the impact of using unsupervised or semi-supervised pseudo-labels instead of training from only limited ground truth, all our experiments use the same architecture. Recent work has shown that properly tuned standard U-Nets can outperform more advanced variants in many segmentation scenarios [33], and we therefore opt for a standard U-Net architecture based on a ResNet18 residual network backbone [34, 35]. The unsupervised, fully-supervised and semi-supervised models trained to predict road masks from RGB frames all have 14.3M parameters. We also use the same architecture for the semi-supervised PLG from Figure 4, but we slightly adapt its first layer to accept the additional channel from RPD in addition to the three RGB channels.

B. Training Procedure

We train randomly initialized models to minimize a binary cross-entropy loss using a batch size of 4. We set an initial learning rate of 10^{-4} and decay it by half when the training loss plateaus for at least 25 epochs. We train our PyTorch models using automatic mixed precision [36], on a single NVIDIA V100 for up to 500 epochs, with an early stopping strategy that halts training when the validation loss has not improved by at least 0.0003 for 75 consecutive epochs.

We use the Cutmix data augmentation strategy proposed for road segmentation in [7] for all of our models. For models that use RGB as inputs, we also apply the Color-Crop-Flip (CFC) augmentation strategy, which applies a color jitter, takes a random crop of appropriate aspect ratio, and randomly perform an horizontal flip of the input [7]. Each augmentation is applied with 50% probability.

For each experiment, we select the model that minimizes the validation loss. For computational reasons and to match Monodepth2 input shape, we use a 192×640 input resolution in all experiments. Outputs are however re-scaled using nearest neighbor interpolation in order to compute IoU and Precision in the original 1024×2048 resolution.

C. Unsupervised Approaches

Road Classes: only road	**IoU**	**Precision**	**Recall**
Superpixel Clustering [4, 7]	0.8152	0.8854	0.9138
Co-Teaching Superpixel Clustering [6]	0.8261	0.9093	0.9027
Cutmix Superpixel Clustering [7]	0.8377	**0.9193**	0.9129
Ours (unsupervised)	**0.8529**	0.8827	**0.9623**
Road Classes: road, ground & parking	**IoU$_3$**	**Precision$_3$**	**Recall$_3$**
Stereo v-disparity [5]	0.8001	**0.9283**	0.8529
Ours (unsupervised)	**0.8600**	0.8943	**0.9595**

TABLE I: Test set results for unsupervised road segmentation. The second part of the table evaluates using a different definition of road that includes the *road*, *parking* and *ground* classes to allow comparison with [5]. Metrics are suffixed to emphasize that three classes are used as part of the road.

In Table I, we compare our model trained on unsupervised pseudo-labels (described in Section III-B) to other recent unsupervised methods. All approaches use a similar strategy of generating pseudo-labels before training a neural network to generalize part of the label noise away. The Stereo v-disparity method [5] is closest in spirit to our approach since it also exploits disparity maps. The notable differences are that

Superpixel aggregation	Adaptive thresholding	IoU	Precision	Recall
–	–	0.7456	0.7535	**0.9875**
✓	–	0.8432	0.8683	0.9664
✓	✓	**0.8529**	**0.8827**	0.9623

TABLE II: Ablation study of trained models test results for unsupervised free space estimation.

they obtain disparity maps through stereo depth reconstruction rather than from a monocular depth estimation network, and they simply threshold in v-disparity space to obtain a road mask, while we use cues from the RGB image in the form of superpixel segments. The other three methods [4, 6, 7], rely on the same pseudo-labeling strategy based on computing superpixel features from an ImageNet pretrained-network. Our method does not need stereo pairs, but still achieves the highest IoU (0.8529) and Recall (0.9623). The large increase in Recall however comes at the cost of Precision, as expected for a method that detects the ground plane rather than the road. The source of imprecision is indeed mostly attributable to ground-level pixels being wrongly classified as *road* when they are actually *sidewalks* or *parking*. When including such *flat* classes from Cityscapes as part of our definition of the road, the Precision rises to over 98%. This shows the potential of our approach for different robotics applications where any flat surfaces can be traversed.

To identify the contributions of superpixel aggregation and adaptive thresholding, Table II presents an ablation study. In the absence of adaptive thresholding, a single fixed threshold is selected for all the frames by visually examining RPD maps for 10 random training frames. The manual threshold is set to 0.075 when no superpixel aggregation is used, and to 0.1 when it is enabled. Note that although this manual selection technically introduces human-supervision, it is only used for investigating the impact of adaptive thresholding. Due to noise in the estimated depth and imprecisions in the road plane fit used to build RPD maps, the absence of superpixel aggregation causes a severe lack of Precision.

D. Semi-Supervised Approaches

Cityscapes provides pixel-wise ground truth annotations for the 2380 training and 595 validation frames, which allows us to examine the fully-supervised scenario in the first section of Table III. Our supervised baseline that uses 100% of the ground truth reaches an IoU of 0.9454. Using a higher input resolution, a larger model, and pseudo-labels generated for additional frames not included in our dataset, the Naive-Student Video Sequence model is able to further improve IoU to 0.9882 [8]. We do not include the latter approach in Table III since it uses ground truth annotations for all Cityscapes classes on 2975 frames and pseudo-labels generated for an additional 109 thousand frames, making a direct comparison with our results impossible.

The second and third sections of Table III present models trained on a small fraction of ground truth labels, with and

	GT % (frames)	PLG pseudo-labels	IoU	Precision	Recall
Supervised baseline	100% (2975)	0	**0.9454**	**0.9837**	**0.9427**
Supervised baseline	10%	0	0.9032	0.9614	0.9353
Ours (semi-supervised)	(297)	2678	**0.9332**	**0.9807**	**0.9493**
Supervised baseline	1%	0	0.8338	0.9174	0.9009
Ours (semi-supervised)	(29)	2946	**0.9063**	**0.9701**	**0.9310**

TABLE III: Test set results for fully-supervised and semi-supervised road segmentation. The best results for each level of supervision are reported in bold.

GT %	PLG Inputs RGB	PLG Inputs RPD	IoU	Precision	Recall
1%	✓	–	0.8193	0.9277	0.8743
	–	✓	0.8600	0.9274	0.9218
	✓	✓	**0.9063**	**0.9701**	**0.9310**
10%	✓	–	0.9149	0.9665	0.9435
	–	✓	0.9006	0.9645	0.9303
	✓	✓	**0.9332**	**0.9807**	**0.9493**

TABLE IV: Test results of our semi-supervised models using different PLG inputs. The best results for a given level of supervision are reported in bold.

without adding pseudo-labels for the remaining frames. Using 10% annotated samples, our semi-supervised model is able to achieve an IoU of 0.9332. This corresponds to a +3% increase over a supervised baseline trained on the same ground-truth frames without pseudo-labels. When dropping the fraction of labeled samples to 1%, our proposed method improves IoU by +7.25% compared to its supervised counterpart and reaches 0.9063 IoU.

These results motivate a pragmatic approach to data annotation for free-space segmentation tasks: even a minimal labeling effort can greatly improve results and enable rapid prototyping for robotics applications. Indeed, annotating only 29 frames allows to increase Precision by +8.74% and IoU by +5.34% over the best unsupervised method of Table I. Although annotating more data is always beneficial, it is also important to notice that semi-supervised models using 1% and 10% of labeled data respectively achieve 95.9% and 98.7% of the IoU obtained using a comparable model trained on 100% of the ground truth.

To assess the impact of using both RGB frames and RPD maps as inputs to our semi-supervised PLG, we conduct an ablation in Table IV. The results shows that using RGB alone decreases IoU by 8.7% when using 1% labeled frames. The impact of RPD maps is less extreme when using 10% ground truth annotations, but adding them still increases IoU by 1.8% over RGB frames only. Table IV also shows that using RPD maps alone does not yield the best results. This is expected since RPD values are approximate and don't contain information to distinguish between the ground and object parts lying close to it.

E. Inference Time

The models used in our unsupervised and semi-supervised methods are also computationally efficient, since they use

the same architecture but a lower input resolution of 192×640 compared to the 512×1024 inputs used in other approaches [4]–[7]. Table V illustrates the differences in both inference time and Multiply-Accumulate operations (MACs) required in a forward pass for a single input frame. The inference times include the copy of the frame to GPU memory and of the result back to CPU memory.

Input Resolution	Inference Time	GMACs
192×640	3.56 ms $\pm$ 5.65 µs	10.16
512×1024	10.42 ms $\pm$ 4.29 µs	43.37

TABLE V: Inference times and Mutiply-Accumulate operations on a NVIDIA Tesla V100. For time measurements, we report mean and standard deviation over 1000 runs.

VI. Conclusion

This work investigates several unsupervised and semi-supervised pseudo-labeling strategies to train a neural network to segment the road using frames captured by a single road-facing camera with little to no manual annotations.

In order to minimize the labeling efforts required to train these models, we devise a novel feature extraction method based on the v-disparity algorithm. To the best of our knowledge, our approach is the first to compute approximate v-disparity maps using a depth prediction network and use them to compute dense Road Plane Distance (RPD) maps. These RPD maps constitute task-specific features and can be used for both unsupervised and semi-supervised road segmentation.

We show that RPD maps can be combined with RGB superpixels and processed in order to obtain pseudo-labels, which are then used as targets to train a neural network in a completely unsupervised way. Unlike previous work, our unsupervised method does not require the use of stereo-pairs but still reaches state-of-the-art results while also being less computationally-intensive. These results can also be further improved if using a small fraction of labeled data is acceptable.

By combining RPD maps with features extracted from a depth estimation network and using them to train pseudo-label generators with minimal supervision, we are able to greatly improve IoU over the unsupervised case. The semi-supervised results obtained using 1% (resp. 10%) of ground truth labels improve IoU by 5.34% (resp. 8.03%) over the unsupervised approach. These results correspond to 95.9% (resp. 98.7%) of the IoU achieved by a comparable fully-supervised model. Considering that 1% of the Cityscapes annotations correspond to only 29 frames, these results motivate a pragmatic approach to labeling for segmentation tasks: even minimal labeling efforts can greatly improve results.

This work mainly focuses on generating high-quality, task-specific, pseudo-labels. Although our results illustrate the known ability of neural networks to generate past some of the label noise when trained using Stochastic Gradient Descent, future work will investigate training strategies that take label noise into account to further improve results. Examples of such techniques are Co-Teaching or Mean Teacher [10, 21]. Another interesting future line of work lies in the exploitation of temporal consistency in video data rather than single isolated frames [8, 9]. Finally, although this work emphasizes an application to road segmentation, our approaches are not specific to urban scenes. Indeed, the depth network used for RPD computation and features extraction in this work can be trained without any label on other datasets, enabling future work to explore applications to monocular robots operating in less-contrained indoor and outdoor environments.

Appendix A
Impact of RPD quantile choice

The unsupervised pseudo-label generation described in Section III-B relies on RGB superpixel cues to reach the best performance. The values of RPD maps are aggregated over RGB superpixels using a quantile function in order to discard some of the noise. In Table VI, we study the impact of changing the quantile value on IoU, Precision and Recall. Note that these results are for raw pseudo-labels. They can therefore not be directly compared to the IoU results presented in Table I and Table II, which correspond to a trained model. They are however indicative of relative performance between quantile choices. Since evaluating different quantile choices on the test set technically constitutes a fit, this analysis is only provided for additional insights. The decision to use the 90% quantile in our methodology was made prior to such evaluation, by visually inspecting the RPD maps of a few random training frames to decide on an appropriate value.

The exact choice of the quantile does not have a large impact on IoU, but influences the Precision-Recall trade-off, with lower values favoring Recall since more superpixels tend be classified as part of the road. The choice of the 100% quantile is equivalent to aggregating a superpixel by its maximum RPD value, and is detrimental to IoU since it is highly sensitive to large RPD outliers.

RPD Quantile	IoU	Precision	Recall
50%	0.7836	0.8140	**0.9532**
60%	0.7890	0.8283	0.9422
70%	0.7939	0.8307	0.9471
80%	**0.7984**	0.8357	0.9475
85%	0.7983	0.8441	0.9360
90%	0.7957	0.8548	0.9214
95%	0.7963	0.8653	0.9104
100%	0.7625	**0.8661**	0.8779

TABLE VI: Test results for unsupervised raw pseudo-labels using adaptive thresholding with different quantile values.

REFERENCES

[1] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset," in *CVPR Workshop on the Future of Datasets in Vision*, 2015.

[2] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 969–977.

[3] A. Harakeh, D. Asmar, and E. Shammas, "Identifying good training data for self-supervised free space estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] S. Tsutsui, T. Kerola, S. Saito, and D. J. Crandall, "Minimizing supervision for free-space segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 988–997.

[5] J. Mayr, C. Unger, and F. Tombari, "Self-supervised learning of the drivable area for autonomous vehicles," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 362–369.

[6] F. Robinet, C. Parera, C. Hundt, and R. Frank, "Weakly-supervised free space estimation through stochastic co-teaching," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, January 2022, pp. 618–627.

[7] F. Robinet and R. Frank, "Refining weakly-supervised free space estimation through data augmentation and recursive training," in *Artificial Intelligence and Machine Learning*. Springer International Publishing, 2022, pp. 30–45.

[8] L.-C. Chen, R. G. Lopes, B. Cheng, M. D. Collins, E. D. Cubuk, B. Zoph, H. Adam, and J. Shlens, "Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation," in *European Conference on Computer Vision*, 2020, pp. 695–714.

[9] L. Hoyer, D. Dai, Y. Chen, A. Koring, S. Saha, and L. Gool, "Three ways to improve semantic segmentation with self-supervised depth estimation," 06 2021, pp. 11 125–11 135.

[10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[11] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.

[12] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[13] J. Watson, M. Firman, A. Monszpart, and G. J. Brostow, "Footprints and free space from a single color image," in *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] R. Labayrade, D. Aubert, and J.-P. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through" v-disparity" representation," in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 2. IEEE, 2002, pp. 646–651.

[15] H. Wang, Y. Sun, and M. Liu, "Self-supervised drivable area and road anomaly segmentation using rgb-d data for robotic wheelchairs," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4386–4393, 2019.

[16] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient rgb-d semantic segmentation for indoor scene analysis," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 525–13 531.

[17] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," *arXiv preprint arXiv:1904.12848*, 2019.

[18] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, 04 2017.

[19] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6022–6031, 2019.

[20] M. Li, M. Soltanolkotabi, and S. Oymak, "Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 4313–4324.

[21] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8527–8537.

[22] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101759, 2020.

[23] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[25] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[26] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[27] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[28] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1635–1643.

[29] W. Xie, Q. Wei, Z. Li, and H. Zhang, "Learning effectively from noisy supervision for weakly supervised semantic segmentation," in *BMVC*, 2020.

[30] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[31] T. Durand, T. Mordan, N. Thome, and M. Cord, "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5957–5966.

[32] Y. Chang, Q. Wang, W. Hung, R. Piramuthu, Y. Tsai, and M. Yang, "Mixup-cam: Weakly-supervised semantic segmentation via uncertainty regularization," in *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[33] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. J. Wirkert, and K. H. Maier-Hein, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," *CoRR*, vol. abs/1809.10486, 2018.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf