Weakly-Supervised Free Space Estimation through Stochastic Co-Teaching

François Robinet
360Lab, SnT, University of Luxembourg
francois.robinet@uni.lu

Claudia Parera
360Lab, SnT, University of Luxembourg*
claudia.parera@uni.lu

Christian Hundt
NVIDIA AI Technology Center Luxembourg
chundt@nvidia.com

Raphaël Frank
360Lab, SnT, University of Luxembourg
raphael.frank@uni.lu

Abstract

Free space estimation is an important problem for autonomous robot navigation. Traditional camera-based approaches train a segmentation model using an annotated dataset. The training data needs to capture the wide variety of environments and weather conditions encountered at runtime, making the annotation cost prohibitively high. In this work, we propose a novel approach for obtaining free space estimates from images taken with a single road-facing camera. We rely on a technique that generates weak free space labels without any supervision, which are then used as ground truth to train a segmentation model for free space estimation. Our work differs from prior attempts by explicitly taking label noise into account through the use of Co-Teaching. Since Co-Teaching has traditionally been investigated in classification tasks, we adapt it for segmentation and examine how its parameters affect performances in our experiments. In addition, we propose Stochastic Co-Teaching, which is a novel method to select clean samples that leads to enhanced results. We achieve an IoU of 82.6%, a Precision of 90.9%, and a Recall of 90.3%. Our best model reaches 87% of the IoU, 93% of the Precision, and 93% of the Recall of the equivalent fully-supervised baseline while using no human annotations. To the best of our knowledge, this work is the first to use Co-Teaching to train a free space segmentation model under explicit label noise. Our implementation and models are freely available online.

1. Introduction

Autonomous navigation is one of the key problems in modern robotics. Before being able to safely plan and execute its motion, an autonomous vehicle should perceive its environment and identify drivable free space in an accurate manner. In this context, free space can be defined as road surfaces that are not occupied by other objects such as vehicles, traffic signs, road dividers or pedestrians [22]. Since collision avoidance requires a fine-grained understanding of the scene, we aim to classify every pixel as belonging to either free space or occupied space.

In this work, we focus our attention on systems using only a single road-facing camera. Although free space segmentation can be approached using classical semantic segmentation techniques, they usually require large quantities of annotated images. While bounding-boxes for object detection can be relatively cheap to obtain, studies have shown that pixel-level annotations are significantly more time consuming [34]. In addition to the 1.5 hour labor cost associated with labeling a single frame [9], a wide variety of environmental and weather conditions need to be captured. This creates a need for very large datasets, and renders fully-supervised semantic segmentation solutions impractical. We tackle this problem in a different way: relying on a method that generates weak, noisy, free space annotations without any supervision [49], we train a neural network to generalize past the label noise using Co-Teaching [16].

Our contributions can be summarized as follows: 1) we adapt Co-Teaching for segmentation tasks and illustrate its effectiveness on the particular case of free space estimation, 2) we study the impact of the Co-Teaching schedule on performances, 3) we propose a refinement called *Stochastic Co-Teaching* and 4) we compare Stochastic Co-Teaching to standard training and traditional Co-Teaching and observe improvements in both IoU and Precision. We also make our code and models available online.

The remainder of this paper is organized as follows: In Section 2, we review the recent literature for both free space estimation and weakly-supervised segmentation. In Section 3, we introduce our weakly-supervised Co-Teaching approach to free space estimation and describe the baseline methods used for benchmarking. In Section 4, we describe our use of the Cityscapes dataset [9] and detail the

[*] At the time of writing

experimental setup of this study. In Section 5, we carry out experiments, detail the qualitative and quantitative results achieved, analyse the limitations of our approach, and share further research directions. Finally, we conclude with a summary of our contributions.

2. Related Work

Over the last decades, free space estimation has been approached with methods that leverage a wide variety of sensors, *e.g*. GNSS [30], LiDAR [53] or cameras [42]. In this work, we place a particular focus on recent camera-based learning methods that use Convolutional Neural Networks. Our method builds on recent advances in network architectures for segmentation, weakly supervised techniques for semantic segmentation or specific to free space estimation, and on training under label noise. We present this background material in the following four sections.

Supervised Learning for Segmentation As a segmentation task, supervised free space estimation has directly benefited from progress in semantic segmentation. Fully-Convolutional architectures such as FCNs [35], SegNets [2] and U-Nets [47] have attracted a lot of attention in recent years. Many refinements to U-Net have been proposed [23, 59, 41], but this work will rely on a simple U-Net architecture. Our choice is motivated by a recent finding that many recent architecture improvements are outperformed by a well-tuned standard U-Net [21]. The efficiency of deep networks for free space segmentation has already been demonstrated in the fully-supervised context [42].

Weakly-Supervised Semantic Segmentation One major drawback of these techniques is their reliance on extensive human-annotated datasets. Such pixel-level annotations are extremely expensive, the total annotation time reaching 1.5 hour per frame in some cases [9]. In cases where fine-grained annotations are available for at least a subset of the data, semi-supervised approaches such as Co-Training can be applied [44]. As noted in the previous section, such pixel-wise annotations are expensive to obtain. In their absence, substantial efforts have been focused on domain adaptation from synthetic data [20], or on leveraging weaker ground truth for semantic segmentation. The main research directions re-purpose coarser labels such as bounding boxes [11, 26, 27, 54], image-level labels [45, 13, 50], class activation maps [6], single points [3], or scribbles [33].

Unsupervised and Weakly-Supervised Monocular Free Space Segmentation Monocular free space estimation has been approached in many different ways that differ in the representation they use. A popular representation is the stixel world, which approximates the ground plane and represents obstacles as vertical sticks [1, 10]. Another possibility is to represent free space as a single horizontal curve lying on the ground plane [56]. These approaches are however not directly comparable to our work, since they do not annotate free space behind obstacles, and rely on ground truth annotations for training. Monocular SLAM tackles a related problem, but relies on video sequences and results in point clouds that do not explicitly represent free space [14, 40, 12]. Recent work has also used video sequences to jointly learn free space and obstacle footprints using structure-from-motion [52]. Our work explores another avenue: we learn dense free space from single images using approximate masks that can be generated without requiring any supervision. One way of generating such *weak labels* is to obtain depth information from stereo pairs and to extract a ground plane estimate, often using the v-Disparity algorithm [29, 17, 39]. Another possibility is to exploit strong road texture and location priors, by dividing the input into superpixels and clustering them based on saliency maps [51] or semantic features [42]. We note that relying on approximate masks differs from approaches based on coarse labels, since the generated masks contain both false positives and false negatives. Indeed, bounding-boxes contain no false negative, and scribble or point supervisions do not include any false positive.

Learning to Segment under Label Noise Recent research has shown that it is possible to train over-parametrized models to generalize past some of the label noise using Stochastic Gradient Descent (SGD) schemes combined with early stopping [31]. However, approaches that explicitly deal with noisy labels can further improve performances, and have become an important research focus over the past few years. Solutions to this problem include label cleaning [8], noise-aware network architectures [48], or noise reduction through robust loss functions [37, 36, 46]. Another line of research proposes to adapt the training procedure itself. Curriculum learning [4] is based on training a model on samples of increasing difficulty, which can correspond to different noise levels [24, 15]. Knowledge distillation [19] is another procedure that can cope with noise by training the teacher model on a relatively clean subset of the data, and using it to guide the training of the student model on the whole dataset [32]. Decoupling [38] and Co-Teaching [16] are two other approaches where two models are trained simultaneously. Decoupling trains both models only on data where their outputs disagree, while Co-Teaching trains each model on the fraction of the data that the other considers to be clean. For a more comprehensive overview of techniques that cope with noisy labels in image analysis, we refer the reader to the survey in [25].

Our work builds on supervised segmentation research through its use of the U-Net architecture. We address label noise using a Co-Teaching training scheme that we adapt for segmentation tasks. We choose Co-Teaching because it has been shown to perform well under moderate amounts of noise [7]. We present our method in detail in the next section.

3. Methodology

In this section we describe the main steps of our weakly-supervised approach for a free space estimation task using Co-Teaching. We present Co-Teaching and its adaptation for a segmentation task, and we introduce its Stochastic variant. Since we focus on improving the training aspect, we use the weak labels proposed in [49] as targets during training. We benchmark the performances of (Stochastic) Co-Teaching against a fully-supervised model, as well as against unsupervised and weakly-supervised baselines in Section 5.

Co-Teaching for Segmentation The intuition for Co-Teaching is based on memorization properties of deep neural networks trained with a variant of SGD [16]. Although these networks are capable of overfitting random noise in their training set [58], they also learn patterns from clean data first [31]. To exploit this property, Co-Teaching proposes to train two separate student networks f and g, and to have each one select clean instances for the other to train on. Since models tend to learn from clean data first, standard Co-Teaching selects clean labels as the ones with the lowest loss. The main additional meta-parameter of Co-Teaching is a schedule $R(t)$ that defines the fraction of the data that is considered clean and should be used for training at any given iteration t. In early iterations, models have not learned enough to identify noise in the training data, and we should generally set $R(t) \approx 1$. As training goes on, $R(t)$ should tend towards the expected noise rate τ, such that all the noise and only the noise gets discarded. In a weakly-supervised setting, τ is unknown and one must either estimate it or try different schedules. Although it has traditionally been used in classification tasks, Co-Teaching has also recently seen some success when training an object detector from noisy bounding boxes [5]. To the best of our knowledge, this work is the first to adapt it to a segmentation task. Our adaptation is straightforward: we consider each weakly-labeled pixel as an independent label, and therefore train each network on a sample of all pixels in a batch. The entire procedure is detailed in Algorithm 1.

Stochastic Co-Teaching We further propose a refinement named *Stochastic Co-Teaching*. Rather than systematically selecting the subset of labels that incur the least loss, Stochastic Co-Teaching samples them with weights that are inversely proportional to their loss. With this change, when labels in a batch incur similar losses, they are all similarly likely to be selected for training, regardless of the schedule $R(t)$. We are trusting that low-loss labels are more likely to be clean, but we accept that some higher-loss samples can also be selected with lower probability. Rather than being noisy, some of these high-loss labels may correspond to harder examples that should not be systematically discarded. Figure 1 illustrates the whole process: (a) An identical batch of B images with resolution $H \times W$ is fed to independent networks f and g, (b) pixel-wise losses $L^{(f)}$ and $L^{(g)}$ are computed using the noisy labels for each image, (c) clean indices $I^{(f)}$ and $I^{(g)}$ are independently selected for each student network by sampling a fraction $R(t)$ of indices without replacement and with probability inversely proportional to their loss, (d) networks exchange their clean indices and use them to sub-sample their own losses and obtain $\bar{L}^{(f)}$ and $\bar{L}^{(g)}$, (e) each network only learns from pixels that the other student deems clean. Note that nothing prevents the use of networks f and g with completely different topologies. In this work, students share the same architecture, but their weights are randomly initialized independently from each other.

4. Experimental Setup

Dataset The Cityscapes dataset provides pixel-wise human labels for 30 visual classes in 5000 frames [9]. Since the test set has no public annotation, we treat the 500 frames of its validation set as our test set and randomly split the Cityscapes training set into 2380 training and 595 validation frames. In the context of autonomous robot navigation, we consider free space to correspond to the *road* object class. Cityscapes also contains 1.6% of frames where no pixel is labeled as *road*. For these frames only, we use the *ground* class to denote free space. Visual inspection confirmed that *ground* corresponds to free space in these frames. Finally, the semantic labels include 6 *void* classes such as *unlabeled*, *out of the region of interest* or *ego-vehicle*. Following Cityscapes semantic segmentation benchmarks, pixels that correspond to such classes are ignored at evaluation time using a mask $m \in \{0, 1\}^{H \times W}$.

Evaluation Metrics Our evaluation relies on three evaluation metrics: the Intersection-over-Union (IoU), Precision and Recall of the free space class. IoU reflects the overall quality of the prediction, but does not immediately capture the fraction of pixels that are labeled as part of the road when they are actually occupied. Since these *false free space positives* are extremely harmful to a robot navigation scenario, we also emphasize the importance of measuring the Precision of our predictions, *i.e.* the fraction of our free space prediction that is indeed free space. Although it is also interesting to monitor Recall, we note that missing free

Algorithm 1: (Stochastic) Co-Teaching for Segmentation

```
Inputs: Models $f$ and $g$, training data generator $\mathcal{G}$, loss $\mathcal{L}$, max iterations $T$ and schedule $R(t)$
forall $t \in \{1, \ldots, T\}$ do
    Obtain next batch $(x, y_{weak}) \in \left(\mathbb{R}^{B \times H \times W \times 3} \times \mathbb{R}^{B \times H \times W}\right)$ from training data $\mathcal{G}$
    Compute per pixel losses $L^{(f)} = \mathcal{L}(f(x), y_{weak})$ and $L^{(g)} = \mathcal{L}(g(x), y_{weak})$
    Compute $n = R(t) \times B \times H \times W$, the number pixel samples to keep
    if stochastic co-teaching then
        Let $\mathcal{S}(w, k)$ randomly sample $k$ unique indices of $w$, using values of $w$ as weights
        Sample $n$ clean indices $I^{(f)} = \mathcal{S}(1/L^{(f)}, n)$ and $I^{(g)} = \mathcal{S}(1/L^{(g)}, n)$
    else
        Let $TopK(z, k)$ select the indices of the $k$ largest elements of $z$
        Compute $n$ clean indices $I^{(f)} = TopK(1/L^{(f)}, n)$ and $I^{(g)} = TopK(1/L^{(g)}, n)$
    Compute clean losses $\bar{L}^{(f)} = \left\{L_i^{(f)} : i \in I^{(g)}\right\}$ and $\bar{L}^{(g)} = \left\{L_i^{(g)} : i \in I^{(f)}\right\}$
    Update $f$ using $\nabla \bar{L}^{(f)}$ and $g$ using $\nabla \bar{L}^{(g)}$
```

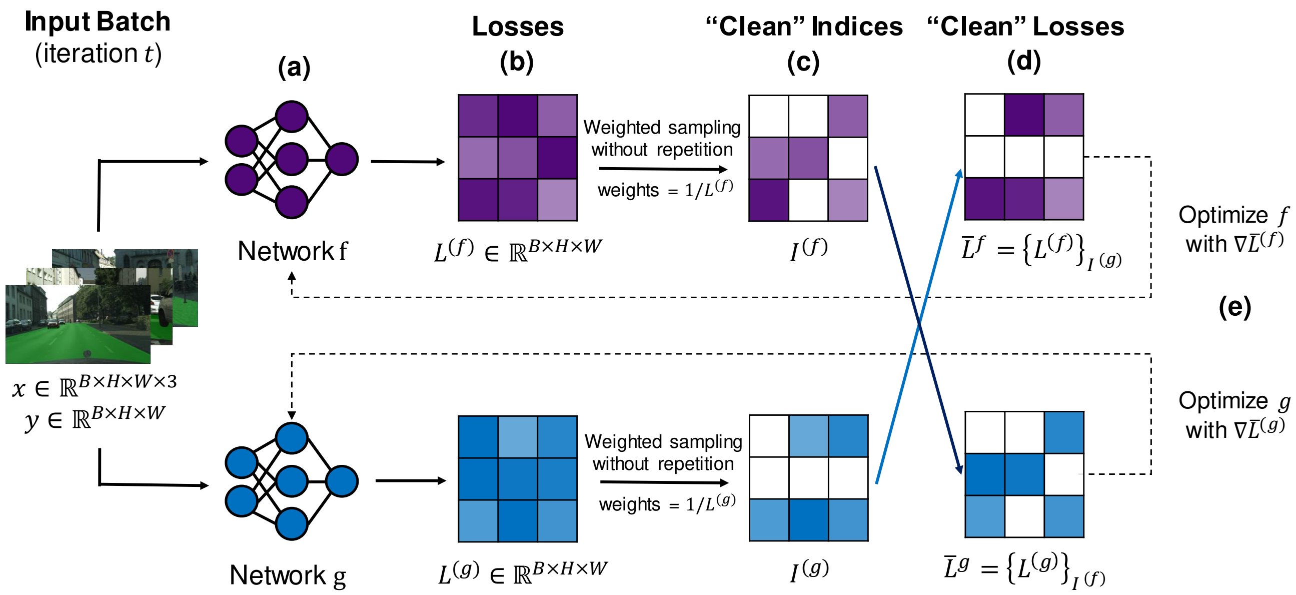
Input Batch
(iteration t)
(a)
Losses
(b)
"Clean" Indices
(c)
"Clean" Losses
(d)
Weighted sampling without repetition
weights = $1/L^{(f)}$
Network f
$L^{(f)} \in \mathbb{R}^{B\times H\times W}$
$I^{(f)}$
$\bar{L}^f = \{L^{(f)}\}_{I^{(g)}}$
Optimize f with $\nabla\bar{L}^{(f)}$
(e)
$x \in \mathbb{R}^{B\times H\times W\times 3}$
$y \in \mathbb{R}^{B\times H\times W}$
Weighted sampling without repetition
weights = $1/L^{(g)}$
Optimize g with $\nabla\bar{L}^{(g)}$
Network g
$L^{(g)} \in \mathbb{R}^{B\times H\times W}$
$I^{(g)}$
$\bar{L}^g = \{L^{(g)}\}_{I^{(f)}}$

Figure 1: Stochastic Co-Teaching. Student networks compute their own pixel-wise loss, and randomly select a subset of clean pixels to train on, based on loss values of the other student.

space has less impact than false positives in an autonomous driving scenario. Given a single free space prediction $\hat{y}$, ground truth y, and evaluation mask m, the metrics for a single frame of shape $H \times W$ are computed with Equation 1 to 3, where $\hat{y},\ y,\ m \in \{0, 1\}^{H \times W}$.

$$IoU = \frac{\sum_i \hat{y}_i y_i m_i}{\sum_i (\hat{y}_i + y_i - \hat{y}_i y_i) m_i} \tag{1}$$

$$Precision = \frac{\sum_i \hat{y}_i y_i m_i}{\sum_i \hat{y}_i m_i} \tag{2}$$

$$Recall = \frac{\sum_i \hat{y}_i y_i m_i}{\sum_i y_i m_i} \tag{3}$$

Network architectures Following recent research that shows that a well-tuned vanilla U-Net can outperform many variants on segmentation tasks [21], we opt for a U-Net structure based on a ResNet18 backbone (14.3M parameters) [47, 18, 55]. To compare with prior art, we also implement and train the SegNet model described in [49]. For computational reasons, we use a 512 × 1024 input resolution in all experiments. Outputs are however re-scaled using nearest neighbor interpolation in order to compute metrics in the original 1024 × 2048 resolution.

Training procedure Since the authors of [49] do not share trained weights for their SegNet architecture, all models are trained from randomly initialized weights to allow for fair comparison. We use the PyTorch framework [43] and train models with the Adam optimizer [28], a batch size of 6 and a learning rate of 0.001. The models are trained on a single NVIDIA K80 until the training loss plateaus, which occurs after 25 epochs for U-Nets and 50 epochs for SegNets. We keep all intermediate models and perform model selection post-training.

Model selection In the context of weakly-supervised learning, we must be careful when performing model selection. This is especially important since Cityscapes provides ground truth annotations for all training and validation frames used in this study. We stress that these frames are never used for training, picking hyper-parameters, or to perform early stopping. We therefore evaluate ground truth IoU, Precision and Recall only once on the test set, after all these steps have been performed. Models trained with standard training loops are selected to minimize the validation loss. This approach has to be slightly adapted when using Co-Teaching, since the scale of the loss varies with the value of $R(t)$, the fraction of data to be considered clean over time. As explained in Section 3, $R(t)$ usually starts at 1, before decreasing to a minimum, and again increasing to plateau at $R(t) = \tau$. When $R(t)$ is smaller, a larger fraction of high-loss samples is discarded and the loss value is deflated. To account for this, we only select models in the *final plateau* of our schedules, where $R(t) = \tau$. For more information about the schedules $R(t)$ we consider, see Appendix A.

5. Results

This section outlines the set of experiments carried out to benchmark our proposed method, using IoU, Precision and Recall. Five main types of approaches were tested: 1) Fully-supervised upper-bounds, 2) unsupervised baselines, 3) standard training, 4) (Stochastic) Co-Teaching, and 5) Ensembling (Stochastic) Co-Teaching students. The quantitative results are summarized in the five categories of Table 1. In this section, we present results for each category, before analyzing the limitations of our approach and presenting qualitative results.

Fully-Supervised Upper-bound Since Cityscapes provides human annotations for all of the data, it is natural to compare our unsupervised approach with its supervised counterpart. To this end, we train a *Fully-Supervised U-Net* using the ground truth labels and observe that it is able to reach high IoU (0.9412) and Precision (0.9726). Since this is the only method that uses the ground truth labels for training or validation, we expect it to provide an upper-bound for unsupervised results. To account for the effect of potential noise in the ground truth, we also train the same network using (Stochastic) Co-Teaching. Since ground truth data is assumed to contain only a small amount of noise, we use a specific $R(t)$ schedule that trains on the whole training data for one epoch, before progressively discarding up to 4% of the training data and slowly incorporating 3% back in to finish training with $R(t) = 0.99$. Examples of similar schedules are illustrated in Appendix A. We observe that both Co-Teaching and Stochastic Co-Teaching result in degraded performance in the fully-supervised case. This indicates that we are discarding valuable data and that ground truth label noise, although always present in pixel-wise annotations, is likely negligible for our purposes.

Unsupervised Baselines In order to compare our approach to other algorithms, we use two simple methods that do not need training and should act as lower bounds. The *Bottom Half* model is a trivial baseline that classifies the entire lower half of the image as free space. Bottom-Half is able to reach a decent IoU of 0.7550 and a high Recall, which is not surprising since free space indeed covers a large portion of the lower half of most frames. The Precision of this model is however only of 0.7798, which is poor compared to the 0.8778 achieved by our second unsupervised baseline, the raw *Weak Labels* from [49]. This second baseline also yields a large IoU improvement, reaching 0.7900. Competing unsupervised approaches often tackle the more general problem of semantic segmentation, for which other datasets are preferred to Cityscapes [11, 54, 45, 13, 6]. Furthermore, papers that use the Cityscapes benchmark seldom report road-class IoU. Recent weakly-supervised free space estimation works also use varied datasets [39, 17, 56]. Two exceptions are presented in [51] and [20], which respectively obtain an IoU of 0.8 and 0.704, but do not report Precision and Recall.

Standard Training We train both our own U-Net and the SegNet model described in [49], using the weak labels as targets in a standard training loop. We stress the fact that these models do not use any human-annotated ground truth at any point during training or validation. As previously observed in [31], standard training is robust to noise to some degree, and our models are able to generalize beyond the noise in their training targets. Compared to raw weak labels, U-Net is able to improve in IoU ($+2.52\%$), Precision ($+1.58\%$), and Recall ($+2.34\%$). Since SegNet yields slightly worse IoU results than U-Net ($+2.3\%$) and is slower at training and inference time, we focus our Co-Teaching experiments on U-Nets.

(Stochastic) Co-Teaching Training We first report results from the best student models, which are selected as having the lowest validation loss among the two students of each Co-Teaching experiment. Note that all models

	Training/Validation Labels	**Test IoU**	**Test Precision**	**Test Recall**
Supervised U-Net (standard training)	ground truth	0.9412	0.9726	0.9727
Supervised U-Net (Co-Teaching ensemble)	ground truth	0.9311	0.9621	0.9646
Supervised U-Net (Stochastic Co-Teaching ensemble)	ground truth	0.9360	0.9664	0.9655
Bottom Half	no training	0.7550	0.7798	0.9616
Weak Labels [49]	no training	0.7900	0.8778	0.8924
Unsupervised Domain Adaptation [20]	synthetic data	0.7040	not reported	not reported
Distant Supervision [51]	image labels	0.8000	not reported	not reported
Standard SegNet	weak labels	0.8130	0.8936	0.9015
Standard U-Net	weak labels	0.8152	0.8854	**0.9138**
Co-Teaching U-Net (best student)	weak labels	0.8214	0.8995	0.9074
Stochastic Co-Teaching U-Net (best student)	weak labels	0.8237	0.9076	0.9017
Co-Teaching U-Net (ensemble)	weak labels	0.8219	0.9028	0.9047
Stochastic Co-Teaching U-Net (ensemble)	weak labels	**0.8261**	**0.9093**	0.9027

Table 1: Quantitative results on the Cityscapes validation set, which we treat as our test set.

trained with Co-Teaching in the 4th and 5th sections of Table 1 use a tuned $R(t)$, whose effect on the performance is explored in Appendix A. Co-Teaching is able to improve over standard training in both IoU ($+0.62\%$) and Precision ($+0.76\%$), while our stochastic variant results in an additional improvement of 0.23% in IoU and 0.81% in Precision.

Ensembling Models trained with (Stochastic) Co-Teaching To assess their convergence, we run Co-Teaching trained students over the training data and we observe a high agreement over the free space predictions (99.2% of pixels are predicted the same), and over which pixels should be considered clean for training (99.4% agreement). Due to its additional sampling step, we observe a slightly lower convergence when Stochastic Co-Teaching is used (97.6% agreement on predictions, 97.8% on clean indices). Since the students exhibit similar validation losses but are not completely equivalent, it is natural to ensemble them by averaging their confidence outputs before thresholding for a prediction. We obtain our best model using this strategy with the Stochastic Co-Teaching students, yielding an IoU of 82.61%, a 0.24% improvement over the best student, and a 0.42% improvement over the Co-Teaching equivalent. These results amount to 87% of the IoU, 93% of the Precision, and 93% of the Recall of the fully-supervised baseline while not using any human labels.

Limitations of (Stochastic) Co-Teaching The introduction of sampling during the training process in the stochastic variant results in performance gains. However, these gains are limited to a few percentage points. To understand why, we take advantage of the availability of ground truth labels in training data from Cityscapes. We train the same U-Net model used in previous experiments for 2 warm-up epochs using the entire weak labels and depict the distribution of pixel-wise losses on Figure 2. This allows us to observe the distribution of noise with respect to the loss at the beginning of training, and analyze the impact of applying different training strategies in subsequent epochs. Figure 2 illustrates an absence of noisy labels at low loss values, and a much larger proportion of wrong labels at high loss values. The Co-Teaching assumption that almost all noisy labels incur high loss values is not completely respected in this case. Indeed, non-negligible noise is also observed at median loss values. This empirical example validates the idea of sampling rather than using a fixed loss cutoff to reject likely noisy samples. Table 2 presents the noise statistics for training a third epoch using different strategies. When discarding 5% of pixels with the highest loss in each batch, Co-Teaching is able to discard 3.3% of the noise, while only removing 1.7% of the clean data. By sampling the training losses 10000 times and reporting mean noise statistics along with their standard deviation, we observe that Stochastic Co-Teaching is able to discard slightly more noise. Using the knowledge that our training weak labels contain 15.67% of noise, we show that using $R(t) = 0.85$ rejects a larger fraction of the noise, but also rejects more clean data. The fact that both Co-Teaching methods invariably sacrifice a small fraction of clean data explains why their Recall results are slightly worse than for Standard training in Table 1. We remind the reader that our previous experiments were conducted without any use of the ground truth on the training and validation data, which prevents the use of such noise level estimates to set optimal co-teaching schedules in practice.

Qualitative Results Figure 3 shows test set predictions of our Stochastic Co-Teaching U-Net, and compares them against the Cityscapes ground truth and raw weak labels. The first three columns of images illustrate the higher Precision of our learned model. It is able to classify regions

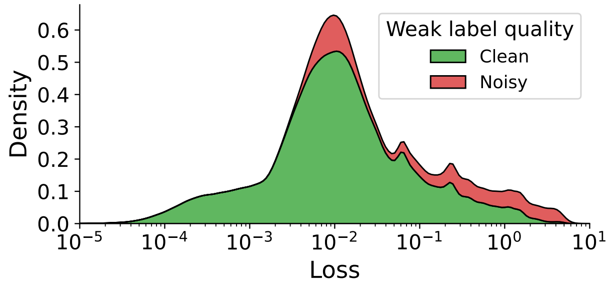

Figure 2: Distribution of U-Net pixel-wise loss after 2 training epochs on weak labels.

Training Strategy	Discarded Noise	Discarded Clean
Standard	0%	0%
Co-Teaching ($R(t) = 0.95$)	3.3%	1.7%
Stochastic Co-Teaching ($R(t) = 0.95$)	$3.6\% \pm 0.04\%$	$1.4\% \pm 0.04\%$
Co-Teaching ($R(t) = 0.85$)	7.2%	7.8%
Stochastic Co-Teaching ($R(t) = 0.85$)	$7.5\% \pm 0.11\%$	$7.5\% \pm 0.11\%$

Table 2: Noise statistics for training epoch 3 using different training strategies.

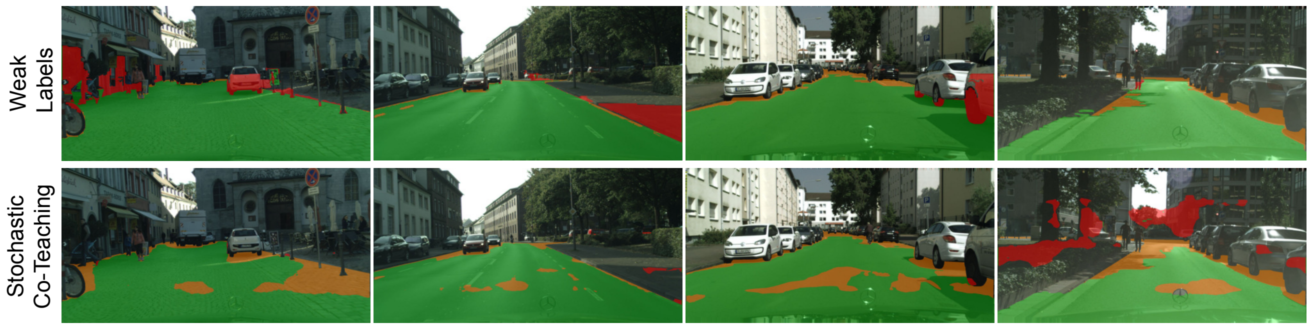

Figure 3: Qualitative results from the test set. Green, red and orange respectively indicate correct, incorrect and missing free space predictions. Note that we display the raw outputs of our model, without masking the ego-vehicle or *void* classes discussed in Section 4.

such as cars or pedestrians as occupied, even though weak labels mark them as free space. Improvements in Precision happen at the cost of Recall, and our predictions tend to be less homogeneous than weak labels. Finally, the last row illustrates a failure case: the weak labels are almost flawless but the model fails to segment free space correctly.

Future Research Directions Because weak labels generation relies on clustering, wrong labels tend to occur in entire spatial regions. Rather than attempting the hard task of discarding individual pixels, future work may investigate clustering approaches to ignore entire superpixels at training time. Since the output of our model can be seen as stronger labels, another promising research direction would be to iteratively train models to refine them.

6. Conclusion

In this work, we introduce a novel approach for training a neural network to predict free space from images taken with a single road-facing camera. We train our models using weak labels that are generated without expensive human annotations, and adapt Co-Teaching to our segmentation task in order to cope with label noise. To the best of our knowledge, our method is the first free space estimation approach to explicitly take label noise into account during training by using an adaptation of Co-Teaching. We also propose *Stochastic Co-Teaching*, a refinement that allows us to improve over results obtained with standard training and classical Co-Teaching procedures. By ensembling students trained with Stochastic Co-Teaching, we improve over standard training in both IoU ($+1.1\%$) and Precision ($+2.4\%$). Our best model reaches 87% of the IoU and 93% of the Precision of the fully-supervised competitor that trains from ground truth pixel-wise labels. Future work will investigate improvements to weak label generation, superpixel-level Co-Teaching, iterative training of successive models, and applications for more general segmentation scenarios.

A. Co-Teaching Schedule Impact

The most important hyper-parameter of Co-Teaching is the schedule $R(t)$, which controls the fraction of the training data that should be considered clean at any epoch t. Following recent research, our $R(t)$ starts at one, decrease to a minimum and then increase to plateau at a final value $R(T)$ [57]. We choose piecewise linear schedules, and vary the length of their *warmup phase* where $R(t) = 1$, their minimum and their final value. Figure 4 illustrates the schedules we consider, and the two parts of Table 3 presents the corresponding test set metrics for the Co-Teaching U-Net model presented in Section 5. In the first part of Table 3, we alter both the minimum and final values of $R(t)$. The 70%-85% schedule discards so much data that many clean labels are also ignored. This lowers the IoU to 0.3749, while the decreased noise allows the Precision to rise to 0.9352. As more data is kept in rows 2 to 4, IoU increases, while Precision slightly decreases. Since our goal is to balance IoU with Precision, we select 90%-95% for further investigation. In the second part of Table 3, we keep the schedule bounds fixed to 90%-95% and vary the length of the initial warm-up phase, where $R(t) = 1$ and all the data is kept. We observe little impact on IoU, but Precision gradually rises with shorter warm-up phases. This indicates that few iterations are enough for the student models to identify and discard noise, and we select the 90%-95% with a single warm-up epoch as our best $R(t)$ schedule, for which we report our results in Section 5.

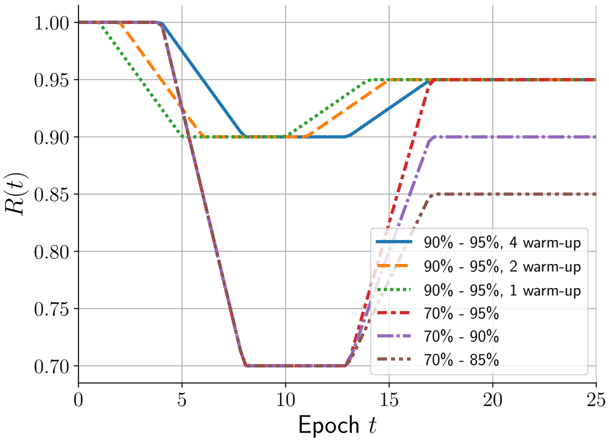

Figure 4: Tested $R(t)$ schedules

$R(t)$ **Bounds**		**Warm-up**	**Test IoU**	**Test Precision**
70%	85%	4	0.3749	**0.9352**
70%	90%	4	0.8081	0.9021
70%	95%	4	0.8079	0.8961
90%	95%	4	**0.8214**	0.8940
90%	95%	2	0.8210	0.8970
90%	95%	1	**0.8214**	**0.8995**

Table 3: Co-Teaching U-Net results

Acknowledgements

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

References

[1] Hernán Badino, Uwe Franke, and David Pfeiffer. The stixel world - a compact medium level representation of the 3d-world. In Joachim Denzler, Gunther Notni, and Herbert Süße, editors, *Pattern Recognition*, pages 51–60, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[3] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science (LNCS), pages 549–565. Springer International Publishing, Sept. 2016. 14th European Conference on Computer Vision 2016, ECCV 2016 ; Conference date: 08-10-2016 Through 16-10-2016.

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.

[5] Simon Chadwick and Paul Newman. Radar as a teacher: Weakly supervised vehicle detection using radar labels. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 222–228, 2020.

[6] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan 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.

[7] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1062–1070. PMLR, 09–15 Jun 2019.

[8] F Chiaroni, M-C Rahal, N. Hueber, and Frédéric Dufaux. Hallucinating a Cleanly Labeled Augmented Dataset from a Noisy Labeled Dataset Using GANs. In IEEE, editor, *26th IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, Sept. 2019.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision*, volume 2, 2015.

[10] Marius Cordts, Timo Rehfeld, Lukas Schneider, David Pfeiffer, Markus Enzweiler, Stefan Roth, Marc Pollefeys, and

Uwe Franke. The stixel world: A medium-level representation of traffic scenes. *Image and Vision Computing*, 68, 02 2017.

[11] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015.

[12] Andrew J. Davison, Ian D. Reid, Nicholas D. Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[13] 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)*, pages 5957–5966, 2017.

[14] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 834–849, Cham, 2014. Springer International Publishing.

[15] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[16] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537, 2018.

[17] Ali Harakeh, Daniel Asmar, and Elie 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.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[20] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR*, abs/1612.02649, 2016.

[21] Fabian Isensee, Jens Petersen, André Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian J. Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *CoRR*, abs/1809.10486, 2018.

[22] J. Janai, F. Güney, A. Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. *ArXiv*, abs/1704.05519, 2020.

[23] S. Jégou, M. Drozdzal, David Vázquez, A. Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1175–1183, 2017.

[24] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2304–2313. PMLR, 10–15 Jul 2018.

[25] Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

[26] Hoel Kervadec, Jose Dolz, Shanshan Wang, Eric Granger, and Ismail ben Ayed. Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. In *Medical Imaging with Deep Learning*, 2020.

[27] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1665–1674, 2017.

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[29] Raphael Labayrade, Didier 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*, volume 2, pages 646–651. IEEE, 2002.

[30] A. Laddha, M. K. Kocamaz, L. E. Navarro-Serment, and M. Hebert. Map-supervised road detection. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 118–123, 2016.

[31] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 4313–4324. PMLR, 2020.

[32] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1928–1936, 2017.

[33] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3159–3167, 2016.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[36] Zhiwu Lu, Zhenyong Fu, Tao Xiang, Peng Han, Liwei Wang, and Xin Gao. Learning from weak and noisy labels for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:486–500, 03 2017.

[37] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *Trans. Img. Proc.*, 17(1):53–69, Jan. 2008.

[38] Eran Malach and Shai Shalev-Shwartz. Decoupling "when to update" from "how to update". In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[39] Jakob Mayr, Christian Unger, and Federico Tombari. Self-supervised learning of the drivable area for autonomous vehicles. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 362–369. IEEE, 2018.

[40] Richard Newcombe, Steven Lovegrove, and Andrew Davison. Dtam: Dense tracking and mapping in real-time. pages 2320–2327, 11 2011.

[41] Ozan Oktay, Jo Schlemper, Loic Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. 04 2018.

[42] G. L. Oliveira, W. Burgard, and T. Brox. Efficient deep models for monocular road segmentation. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4885–4891, 2016.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[44] Jizong Peng, Guillermo Estrada, Marco Pedersoli, and Christian Desrosiers. Deep co-training for semi-supervised image segmentation, 2019.

[45] 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)*, pages 1713–1721, 2015.

[46] François Robinet, Antoine Demeules, Raphaël Frank, Georgios Varisteas, and Christian Hundt. Leveraging privileged information to limit distraction in end-to-end lane following. In *2020 IEEE 17th Annual Consumer Communications Networking Conference (CCNC)*, pages 1–6, 2020.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[48] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. Jan. 2015. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

[49] Satoshi Tsutsui, Tommi Kerola, Shunta Saito, and David J Crandall. Minimizing supervision for free-space segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 988–997, 2018.

[50] S. Tsutsui, S. Saito, and T. Kerola. Distantly supervised road segmentation. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 174–181, 2017.

[51] Satoshi Tsutsui, Shunta Saito, and Tommi Kerola. Distantly supervised road segmentation. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 174–181, 2017.

[52] Jamie Watson, Michael Firman, Aron Monszpart, and Gabriel J. Brostow. Footprints and free space from a single color image. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[53] Liang Xiao, Bin Dai, Daxue Liu, Tingbo Hu, and Tao Wu. Crf based road detection with multi-sensor fusion. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 192–198, 2015.

[54] Wenbin Xie, Qiaoqiao Wei, Zheng Li, and Hui Zhang. Learning effectively from noisy supervision for weakly supervised semantic segmentation. In *BMVC*, 2020.

[55] Pavel Yakubovskiy. Segmentation models. `https://github.com/qubvel/segmentation_models`, 2019.

[56] Jian Yao, Srikumar Ramalingam, Yuichi Taguchi, Yohei Miki, and Raquel Urtasun. Estimating drivable collision-free space from monocular video. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 420–427, 2015.

[57] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10789–10798. PMLR, 13–18 Jul 2020.

[58] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 64, 11 2016.

[59] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017.