Refining Weakly-Supervised Free Space Estimation through Data Augmentation and Recursive Training ⋆

François Robinet[0000−0001−5158−5561]⋆⋆ and Raphaël Frank[0000−0001−8239−2041]

Interdisciplinary Centre for Security, Reliability and Trust (SnT)
University of Luxembourg
firstname.lastname@uni.lu

Abstract. Free space estimation is an important problem for autonomous robot navigation. Traditional camera-based approaches rely on pixel-wise ground truth annotations to train a segmentation model. To cover the wide variety of environments and lighting conditions encountered on roads, training supervised models requires large datasets. This makes 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. We study the impact of different data augmentation techniques on the performances of free space predictions, and propose to use a recursive training strategy. Our results are benchmarked using the Cityscapes dataset and improve over comparable published work across all evaluation metrics. Our best model reaches 83.64% IoU ($+2.3\%$), 91.75% Precision ($+2.4\%$) and 91.29% Recall ($+0.4\%$). These results correspond to 88.8% of the IoU, 94.3% of the Precision and 93.1% of the Recall obtained by an equivalent fully-supervised baseline, while using no ground truth annotation. Our code and models are freely available online.

Keywords: Weak supervision · Free space estimation · Data augmentation · Recursive training

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

⋆⋆ Corresponding author

1 Introduction

Perception is the first step towards autonomous robot navigation. To be able to safely act in the world, a robot needs to perceive its environment and identify traversable free space. In the context of autonomous driving, free space is usually defined as road areas that are not occupied by either static objects such as traffic signs and road dividers, or by dynamic entities such pedestrians and cars [18]. Since collision-free planning requires a fine-grained understanding of the environment around the vehicle, we attempt to label each pixel of a front-facing camera as traversable or not.

This work focuses on systems that use a single road-facing camera. Monocular free space segmentation has traditionally been approached using supervised segmentation techniques. Although effective, these techniques require vast amounts of pixel-wise annotated frames. Studies have shown that such pixel-level ground truth is significantly more expensive to craft than image-level labels or bounding boxes [27]. In addition to the large labor costs entailed by labeling each frame [7], such approaches are held back by the wide variety of environments and lighting conditions that are present at runtime and need to be captured in training data. This need for ever larger annotated datasets makes supervised learning unsuitable for solving this problem. Instead, we tackle it in a different way: relying on a method that generates weak, noisy, free space annotations without any supervision [42], we train a neural network to generalize past the label noise using data augmentation and recursive training.

Our contributions can be summarized as follows: (1) we study the impact of data augmentation on weakly-supervised free space segmentation, (2) we propose a recursive training scheme that uses a progressively refined ground truth, (3) we establish a new state-of-the-art for weakly supervised free space estimation on the Cityscapes dataset, improving over previous efforts by $+2.3\%$ in IoU, $+2.4\%$ in Precision, and 0.4% in Recall, (4) we discuss the limitations of our simple recursive training approach, and (5) we release our code and models for reproduction and further work.

The remainder of this paper is organized as follows: In Section 2, we review the recent literature for free space estimation, data augmentation in the context of semantic segmentation, and recursive training. In Section 3, we introduce our data augmentation and recursive training schemes. In Section 4, we describe our use of the Cityscapes dataset [7] and detail the experimental setup of this study. In Section 5, we carry out experiments and present the qualitative and quantitative results achieved. Finally, we summarize our contributions and share further research directions.

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 [24], LiDAR [45] or cameras [35]. In this work, we place a particular focus on recent camera-based learning methods that use Convolutional Neural Networks (CNNs). Our work builds on recent advances in network architectures for segmentation and on unsupervised methods specific to free space estimation. We present this background material in the following sections.

2.1 Supervised Learning for Segmentation

As a segmentation task, supervised free space estimation has directly benefited from progress in semantic segmentation. Pixel-level prediction carries a crucial challenge for network design: an optimal prediction can only be achieved by combining fine-grained local information with global contextual cues. Fully Convolutional Networks (FCNs) rely on skip connections to carry these cues in their encoder-decoder architecture [28], while SegNets ease the upsampling task by reusing encoder max-pooling indices in the decoder [3]. Building on similar ideas, U-Nets combine entire encoder feature maps with decoder features at each step of the expansion path of the network [40]. U-Nets have attracted a lot of attention in recent years, and researchers have proposed refinements such as the use of dense connections [19] and dilated convolutions [51], the integration of attention mechanisms [34], or extensions to volumetric images [32]. In this work, we 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 vanilla U-Net [17].

2.2 Weakly-Supervised Semantic Segmentation

The major drawback of supervised techniques is their reliance on extensive human-annotated datasets. The cost of labeling is particularly important in segmentation tasks, where the total time required to annotate every pixel in a single frame can reach 1.5 hours in some cases [7]. The reuse of models pre-trained on very large datasets such as ImageNet [11] partially alleviates this problem, but several thousands of training images are still routinely needed to reach adequate performance. In recent years, researchers have devised strategies to reduce or eliminate the need for human annotations during training.

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 [37]. In the complete absence of pixel-wise ground truth labels, researchers have proposed to use domain adaptation from synthetic datasets [16], or to rely on weaker ground truth. Existing techniques rely on coarser labels, such as bounding boxes [9,20,21,46], image-level labels [38,12,43], class activation maps [5], single points [4], or scribbles [26].

2.3 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. Stixel-like approaches represent obstacles as verti-

cal sticks [2,8] or horizontal curves [48], but ignore free space lying behind obstacles. Monocular SLAM relies on video sequences to obtain point-clouds which do not explicitly represent free space [13,33,10]. Using temporal sequences and structure-from-motion to jointly learn an explicit representation of free space and obstacle footprints has also been recently proposed [44]. Our work uses a different strategy: we learn dense free space estimates from single frames using approximate masks that are obtained without human-supervision. Such *weak labels* have historically been generated using depth information from stereo pairs before localizing the ground plane, for example using the v-Disparity algorithm [23,14,31]. Other attempts exploit strong road texture and location priors, by dividing the input into superpixels and clustering them based on saliency maps [43] or semantic features [35]. We stress that using weak labels departs from previously mentioned approaches that leverage coarse ground truth, since weak labels contain false positives and negatives.

2.4 Training Strategies for Weakly-Supervised Segmentation

Recent research shows 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 [25]. Dealing with label noise at training time has become an important research area over the past few years. Solutions to this problem include label cleaning [6], noise-aware network architectures [41], or noise reduction through robust loss functions [30,29,39].

Besides work on training algorithms themselves, researchers have also largely explored regularization through data augmentation in unsupervised settings. Traditional augmentation strategies (scaling, color jittering, flipping, cropping, *etc.*) change pixel values in a single input image without altering its semantic content. More recently, researchers have proposed augmentations that combine several images and their labels. Two notable examples are MixUp [50] and CutMix [49]. MixUp is a method that augments the training set using convex combinations of image pairs and labels, while CutMix overlays random crops of other samples on top of original frames.

3 Methodology

In this work, we train U-Net models to predict dense free space from RGB images by learning on approximate labels that can be generated without any supervision. Since our focus is on improving training aspects rather than on improving weak labels generation, we will reuse the weak labels from [42]. We look at improving training across two dimensions: data augmentation and recursive training.

3.1 Data Augmentation

We study the impact of data augmentation on weakly-supervised free space estimation. We cover both traditional augmentation techniques that operate on single images, as well as MixUp and CutMix, which are more recent and combine multiple samples.

Color-Flip-Crop To represent traditional augmentation techniques, we use a combination of color jittering, horizontal flips and random cropping, which we will refer to as *Color-Flip-Crop* or *CFC* in the remainder of the text. Each augmentation is independently applied with a 50% probability. The color jittering randomly affects brightness, contrast, saturation, and hue using the bounds defined in the Torchvision implementation [1]. In order to preserve most of the original image, cropping is performed with a randomly chosen rectangle that occupies between 25% and 50% of the image area. The aspect ratio is also randomly chosen, with the constraint that the height is at least 10% of the height of the original image. Figure 1 shows some examples of the effect of CFC on a single randomly chosen training image.

Fig. 1: Seven possible Color-Flip-Crop augmentations on a random training sample. The original sample is on the top-left. We show ground truth mask for illustration purposes, they are not used during training.

MixUp Rather than augmenting isolated images, Mixup trains models on convex combinations of samples [50]. By training on synthesized samples that lie between the original training samples, MixUp encourages the network to exhibit a linear behavior between samples and helps preventing memorization. During training, each sample (x_1, y_1) is combined with another random sample (x_2, y_2) from the batch using Equations 1 and 2, where we sample λ uniformly in $[0, 1]$. The effect of combining input samples is illustrated on Figure 2.

$$x_{mixup} = \lambda x_1 + (1 - \lambda) x_2 \tag{1}$$

$$y_{mixup} = \lambda y_1 + (1 - \lambda) y_2 \tag{2}$$

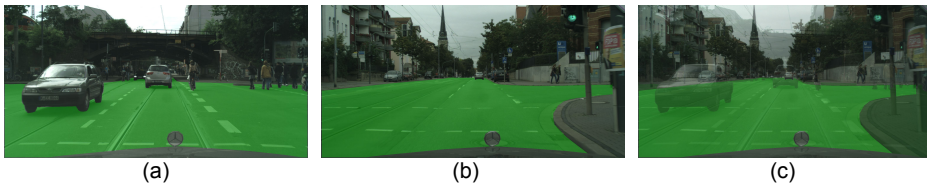

Fig. 2: MixUp augmentation combining two random samples (a) and (b) from the training set. The convex combination using $\lambda = 0.5$ is shown as (c). We show ground truth mask for illustration purposes, they are not used during training.

CutMix Similar to Mixup in spirit, CutMix also combines two random input samples (x_1, y_1) and (x_2, y_2) from the same batch [49]. Rather than combining them over the entire image, CutMix overlays a crop of x_2 over x_1, and the same crop of y_2 over y_1. Equations 3 and 4 formalize this process using a random binary mask $M \in \{0, 1\}^{H \times W}$ to denote the cropped area ($\circ$ denotes the element-wise product). Like for the CFC augmentation, the cropping mask M occupies between 25% and 50% of the image area with a random aspect ratio. Figure 3 illustrates four different instances of CutMix augmentation on a chosen training sample. CutMix generates more natural images than MixUp and allows the network to learn more localizable features since the transformation is only applied to a fraction of the input image.

$$x_{cutmix} = (1 - M) \circ x_1 + M \circ x_2 \tag{3}$$

$$y_{cutmix} = (1 - M) \circ y_1 + M \circ y_2 \tag{4}$$

3.2 Recursive Training

We are training neural networks to estimate free space by learning on approximate labels y_{weak}. Since neural networks trained with SGD variants are partially robust to noise in their training targets [25], the outputs y will tend to approximate the unknown ground truth y^* better than y_{weak}. Assuming the outputs y are better estimates of free space than y_{weak}, it is natural to treat them as cleaner targets for a second round of training. This process can in principle be iterated to obtain progressively cleaner outputs y_2, y_3, *etc*. This approach was already attempted in the context weakly-supervision free space segmentation [43], but we revisit its impact in the presence of data augmentation and with different weak labels. Figure 4 illustrates the process for a given training round.

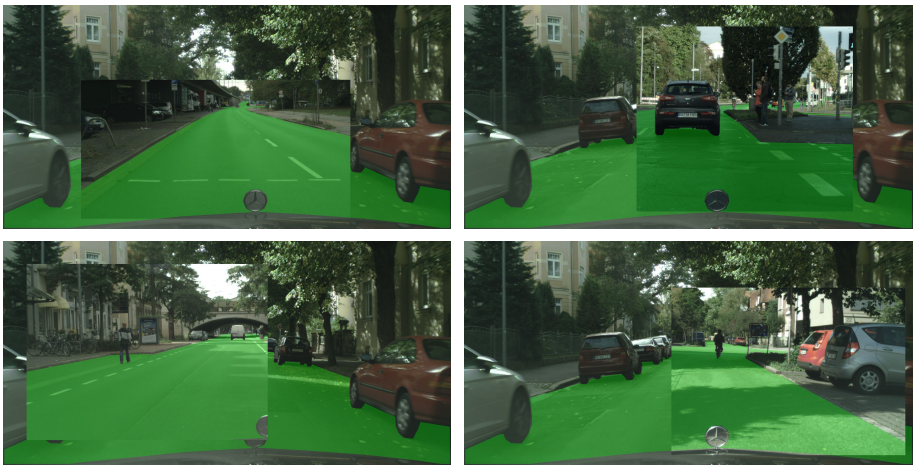

Fig. 3: Four instances of the CutMix augmentation on a random training sample.
We show ground truth mask for illustration purposes, they are not used during training.

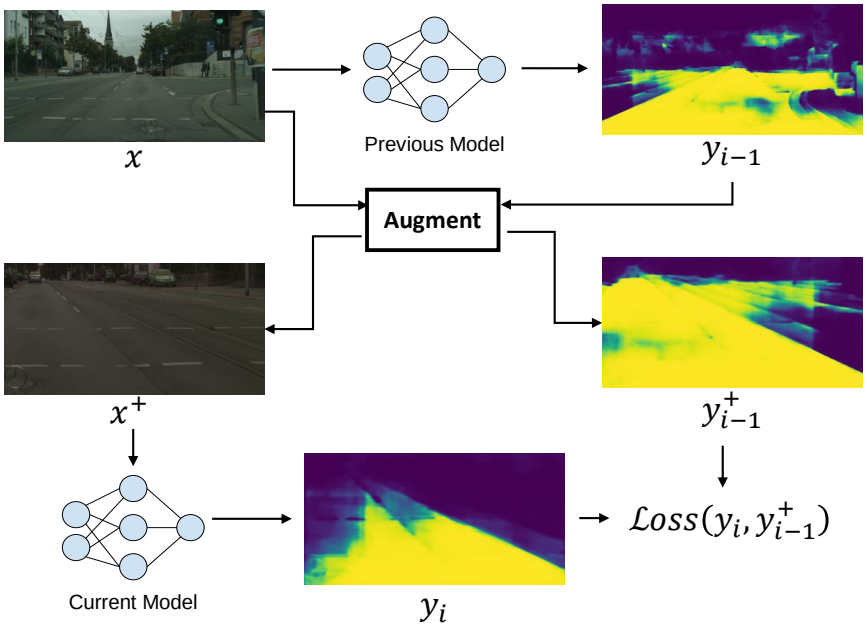

Fig. 4: Recursive training procedure. The current model is trained on augmented outputs from the model obtained at the previous training round. In this example, CFC is used for augmentation. The process is similar for other augmentation strategies.

4 Experimental Setup

4.1 Dataset

Our experiments leverage the Cityscapes dataset, which provides pixel-wise ground truth labels for 30 visual classes in 5000 frames [7]. The official test set has no public annotation, and we therefore 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. Since we are interested in estimating drivable free space in the context of autonomous vehicle navigation, we consider free space equivalent to the *road* class. Cityscapes also contains 1.6% of frames with no *road* pixel. For these frames, visual inspection confirmed that free space correspond to the *ground* class, and that label was used for free space instead of *road*. Finally, the semantic labels include 6 *void* classes such as *unlabeled*, *out of the region of interest* or *ego-vehicle*. Following official Cityscapes segmentation benchmarks, we ignore pixels corresponding to such classes at evaluation time using a binary evaluation mask. We note that this evaluation mask is never used during training or validation, only to evaluate models on the test set.

4.2 Evaluation Metrics

We use three evaluation metrics: the Intersection-over-Union (IoU), Precision and Recall. IoU is a standard metric in segmentation tasks to reflect the overall quality of the predictions. However, IoU does not immediately capture *false free space positives*. These pixels that are labeled as part of the road but are actually occupied are extremely harmful in robotic path-planning scenarios. For this reason, we also monitor the Precision of the free space class, *i.e.* the fraction of our free space prediction that is indeed free space. To obtain a complete picture of prediction quality, we also monitor Recall. We however note that missing free space in predictions has less impact than false free space positives in robot navigation contexts. 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 Equations 5, 6 and 7, where $\hat{y},\ y,\ m \in \{0,1\}^{H \times W}$.

$$IoU = \frac{\sum_i \hat{y}_i y_i m_i}{\sum_i (\hat{y}_i + y_i - \hat{y}_i y_i) m_i} \tag{5}$$

$$Precision = \frac{\sum_i \hat{y}_i y_i m_i}{\sum_i \hat{y}_i m_i} \tag{6}$$

$$Recall = \frac{\sum_i \hat{y}_i y_i m_i}{\sum_i y_i m_i} \tag{7}$$

4.3 Network architectures

Following recent research that shows that a well-tuned vanilla U-Net can outperform many refined variants on most segmentation tasks [17], we opt for a U-Net structure based on a ResNet18 residual network backbone [40,15,47]. To allow for comparison with prior art, we also implement and train the SegNet model described in [42]. 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 IoU and Precision in the original 1024×2048 resolution.

4.4 Training procedure

We use the PyTorch framework [36] and train randomly initialized models to minimize a binary cross-entropy loss using the Adam optimizer [22], a batch size of 8 and an initial learning rate of 0.001. We train our models on single NVIDIA V100 for up to 200 epochs, with an early stopping strategy that halts training when the validation loss has not improved by at least 10^{-4} for 50 consecutive epochs. For each experiment, we select the model that minimizes the validation loss.

4.5 Use of ground truth data

The Cityscapes dataset provides ground truth annotations for all training and validation frames used in this study. We stress that these annotations are only used to train the fully-supervised baseline for comparison with our weakly-supervised approach. Outside of the fully-supervised experiment, ground truth labels are never used for training, hyperparameter tuning, or to perform early stopping. Ground truth IoU, Precision and Recall are computed only once on the test set, after all these steps have been performed.

5 Results

This section describes the experiments carried out to benchmark our proposed method, using Precision, IoU and Recall. We present results for three main categories of models: 1) a fully-supervised upper-bound, 2) unsupervised and weakly-supervised baselines, and 3) U-Nets trained on the weak labels using recursive training and different augmentation strategies. The quantitative results for each category are summarized in Table 1. In this section, we analyze the results of each category, discuss the limitations of recursive training, and present qualitative results.

5.1 Fully-Supervised Results

Since Cityscapes provides pixel-wise ground truth annotations for our training and validation data, we use it to train a fully-supervised U-Net for comparison with its unsupervised counterpart. When trained on ground-truth labels, our U-Net model reaches high IoU (94.12%), Precision (97.26%) and Recall (97.27%). Since this fully-supervised model is the only one that uses ground truth labels at any point during training and validation, it is expected to produce an upper-bound for our unsupervised experiments.

5.2 Unsupervised and Weakly-Supervised Baselines

Competing unsupervised approaches are often focused on generic semantic segmentation rather than free space estimation, and use other datasets than Cityscapes as benchmarks [9,46,38,12,5]. Among weakly-supervised approaches that tackle free space estimation [14,48,43,16], only two publish results for Cityscapes. *Distant Supervision* [43] and *Unsupervised Domain Adaptation* [16] respectively obtain an IoU of 80% and 70.4%, but do not report Precision or Recall values.

We generate approximate labels without supervision using the technique described in [42]. Evaluating these raw weak labels, we obtain an IoU of 79%, a Precision of 87.78% and a Recall of 89.24%. These results can be further improved by training a neural network to generalize beyond the noise in these labels. This was already attempted using the SegNet architecture in [42], which we also implement and train for comparison. SegNet is able to improve results over raw weak labels in IoU ($+2.3\%$), Precision ($+1.58\%$) and Recall ($+0.91\%$).

5.3 Data Augmentation & Recursive Training

We train the same U-Net model using different data augmentation strategies. Since the outputs of our different augmented U-Nets are better than the initial weak labels, we use them as target for a second round of training. We iterate this recursive training process four times for each of the data augmentation strategies under study. We limit training to four rounds for computational reasons and because it is enough for IoU values to reach their peak.

No Augmentation We start by training a U-Net with the weak labels as targets and without any data augmentation. We observe that it compares favorably with the results from SegNet, reaching an IoU of 81.85%, a Precision of 90.65%, and a Recall of 89.76%. Without resorting to data augmentation, recursive training over several rounds is unable to meaningfully improve IoU, and slightly decreases Precision in favor of Recall.

MixUp Applying MixUp allows to improve Precision compared to not using data augmentation by 0.5% in the first training round. IoU is maintained, but Recall decreases by 0.45%. Iterative training is however not effective when combined with MixUp, since we observe a drop in Precision after each round. As discussed in Section 4.2, free space IoU and Precision are more important than Recall in an autonomous navigation scenario. In this case, increases in Recall are not enough to compensate this effect, and we observe a steady decrease in IoU.

Color-Flip-Crop Traditional data augmentation consisting of color jittering, horizontal flips and random cropping is able to improve IoU over not using augmentation and over using MixUp. After a single training round, CFC allows to reach an IoU of 81.99% through increasing Recall by 1.47% compared to the first round without augmentation. Subsequent training rounds are able to improve both Precision and IoU. After 3 iterations, the model reaches an IoU of 82.34% and a Precision of 90.75%.

CutMix The CutMix augmentation can be seen as providing the advantages of cropping and MixUp. Like MixUp, it synthesizes new input samples by combining pairs of existing ones. However, CutMix produces more natural images and its effect is localized since it only affects the area of a random crop. The locality of CutMix has been shown to allow models to learn more localizable features in classification scenarios [49], and it is not surprising that such features are helpful in this segmentation context. Indeed, models trained with CutMix augmentation outperform all other models by a wide margin. After a single training round, CutMix improves over not using augmentations in IoU ($+1.2\%$), Precision ($+0.5\%$), and Recall ($+0.26\%$).

Since our application scenario favors Precision over Recall, our best overall model is obtained after the fourth training round, reaching an IoU of 83.64% and a Precision of 91.75%. Compared to the prior state-of-the-art results from SegNet [42], it improves IoU by 2.3%, Precision by 2.4% and Recall by 0.4%. Although our model does not rely on any human-annotated ground truth, its relative performance compared to the fully-supervised variant is impressive: we reach 88.8% of its IoU, 94.3% of its Precision, and 93.1% of its Recall.

5.4 Limits of Recursive Training

While CutMix results are impressive, we note that the success of recursive training is limited. When not applying data augmentation or when using MixUp, recursive training does not improve on IoU or Precision. In the case of CFC and CutMix augmentations, results are more encouraging, but the improvements are limited to three rounds of training. Starting with the fourth round of training, IoU results start to degrade, sometimes getting worse than those obtained after a single round of training. Explaining

this effect is not straightforward: given that target labels on round 4 are superior to those used on round 3 in both IoU and Precision, we would expect to either observe improved or plateauing results. Such recursive training strategy has been successfully used in foreground class segmentation contexts with results improving over more than 10 rounds [21]. As opposed to our completely unsupervised approach, the authors of [21] could exploit coarser ground truth in the form of bounding boxes in order to refine predictions after each round. We postulate that the absence of such refinement step in our approach is the reason we are unable to further leverage recursive training. Designing such a prediction refinement step will be the topic of future work.

	Training/Validation Labels	**Test IoU**	**Test Precision**	**Test Recall**
Fully-Supervised U-Net	ground truth	94.12%	97.26%	97.27%
Unsup. Domain Adaptation [16]	synthetic data	70.40%	not reported	not reported
Distant Supervision [43]	image labels	80.00%	not reported	not reported
Weak Labels [42]	no training	79.00%	87.78%	89.24%
SegNet (repr. from [42])	weak labels	81.30%	89.36%	90.15%
U-Net (no augmentation)				
Round 1	weak labels	81.85%	90.65%	89.76%
Round 2	output of round 1	81.79%	89.53%	90.80%
Round 3	output of round 2	81.86%	90.15%	90.27%
Round 4	output of round 3	81.82%	90.11%	90.25%
U-Net + MixUp				
Round 1	weak labels	81.89%	91.14%	89.31%
Round 2	output of round 1	81.97%	90.89%	89.60%
Round 3	output of round 2	81.62%	90.13%	89.97%
Round 4	output of round 3	81.45%	89.91%	90.02%
U-Net + Color-Flip-Crop				
Round 1	weak labels	81.99%	88.80%	91.23%
Round 2	output of round 1	82.12%	89.71%	90.64%
Round 3	output of round 2	82.34%	90.75%	90.69%
Round 4	output of round 3	81.91%	90.21%	90.27%
U-Net + CutMix				
Round 1	weak labels	83.05%	91.19%	90.51%
Round 2	output of round 1	83.58%	91.20%	91.12%
Round 3	output of round 2	**83.77%**	91.23%	**91.29%**
Round 4	output of round 3	83.64%	**91.75%**	90.62%

Table 1: Results on the Cityscapes validation set, which we treat as our test set. The best results for a given data augmentation strategy are underlined, and the best overall results are reported in bold.

5.5 Qualitative Results

We compare the free space estimates from weak labels with the predictions of our best model on test set samples on Figure 5.

The ability of our learned model to generalize past some of the noise present in the weak labels that were used during training is clearly visible in the first two rows of Figure 5. Indeed, the cars and side walks that were wrongly considered free space in the weak labels are correctly predicted by our trained model. In addition to its higher Precision, our model also has higher IoU and Recall, as illustrated by the near-absence of orange areas in its predictions.

The third row shows a more contrasted situation. Although our model is able to cover more free space, it still shows some signs of overfitting to noise in the weak labels. Shadows are especially problematic because they are likely to impact the superpixel segmentation that the weak labels are based on, resulting in missed free space areas such as the one present in front of the cyclist. Since this effect happens fairly consistently over the training set, our model is incapable of completely addressing it.

Finally, the fourth row illustrates another partial failure of our model in a particularly crowded scene. Compared to the corresponding weak labels, the trained model correctly rejects pedestrians, but is unable to produce a clean segmentation around them and considers the pavement as occupied space. Although the prediction still contains errors, we note that red areas in our prediction are much more acceptable from a semantics point-of-view than the ones from the corresponding weak labels.

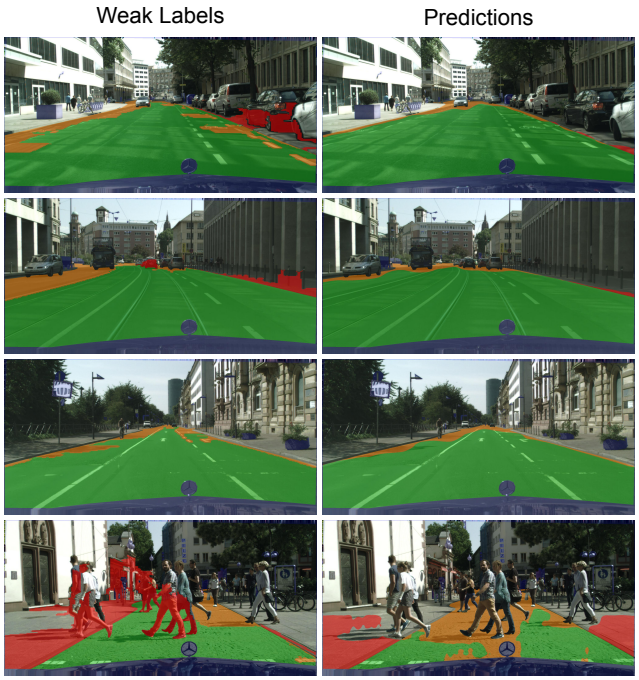

Fig. 5: Qualitative results from the test set obtained from a U-Net trained with CutMix for 4 rounds. Predictions are color-coded using the ground truth: green and red respectively corresponds to correct and incorrect predictions, orange represents missing free space, and areas that are ignored at evaluation time are denoted in blue (see Section 4.1).

6 Conclusion

In this work, we investigate different weakly-supervised training strategies for teaching a neural network to predict free space from images taken with a single road-facing camera. Our models are trained using weak labels that are generated without human intervention, and we investigate the impact of recursive training with several data augmentation schemes. We show that the CutMix augmentation is particularly efficient for free space estimation, especially when combined with recursive training. We benchmark our results on the Cityscapes dataset and improve over unsupervised and weakly-supervised baselines, reaching 83.64% IoU ($+2.3\%$), 91.75% Precision ($+2.4\%$) and 91.29% Recall ($+0.4\%$). Our best model obtains 88.8% of the IoU, 94.3% of the Precision and 93.1% of the Recall of the fully-supervised competitor that trains from expensive pixel-wise labels. Finally, we show that simple recursive training is limited in its ability to increase performances, and suggest directions to improve the approach. Future work will also investigate improvements to weak label generation and applications to more general segmentation scenarios.

References

1. Torchvision: Datasets, transforms and models specific to computer vision. https://github.com/pytorch/vision (2021)
2. Badino, H., Franke, U., Pfeiffer, D.: The stixel world - a compact medium level representation of the 3d-world. In: Denzler, J., Notni, G., Süße, H. (eds.) Pattern Recognition. pp. 51–60. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
4. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: Computer Vision – ECCV 2016. pp. 549–565. Lecture Notes in Computer Science (LNCS), Springer International Publishing (Sep 2016). https://doi.org/10.1007/978-3-319-46478-7_34, http://www.eccv2016.org/, 14th European Conference on Computer Vision 2016, ECCV 2016 ; Conference date: 08-10-2016 Through 16-10-2016
5. Chang, Y., Wang, Q., Hung, W., Piramuthu, R., Tsai, Y., Yang, M.: 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), https://www.bmvc2020-conference.com/assets/papers/0367.pdf
6. Chiaroni, F., Rahal, M.C., Hueber, N., Dufaux, F.: Hallucinating a Cleanly Labeled Augmented Dataset from a Noisy Labeled Dataset Using GANs. In: IEEE (ed.) 26th IEEE International Conference on Image Processing (ICIP). Taipei, Taiwan (Sep 2019), https://hal.archives-ouvertes.fr/hal-02054836
7. Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset. In: CVPR Workshop on the Future of Datasets in Vision. vol. 2 (2015)
8. Cordts, M., Rehfeld, T., Schneider, L., Pfeiffer, D., Enzweiler, M., Roth, S., Pollefeys, M., Franke, U.: The stixel world: A medium-level representation of traffic scenes. Image and Vision Computing **68** (02 2017). https://doi.org/10.1016/j.imavis.2017.01.009
9. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1635–1643 (2015)
10. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: Monoslam: Real-time single camera slam. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(6), 1052–1067 (2007). https://doi.org/10.1109/TPAMI.2007.1049
11. Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
12. Durand, T., Mordan, T., Thome, N., Cord, M.: 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). pp. 5957–5966 (2017). https://doi.org/10.1109/CVPR.2017.631
13. Engel, J., Schöps, T., Cremers, D.: Lsd-slam: Large-scale direct monocular slam. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 834–849. Springer International Publishing, Cham (2014)
14. Harakeh, A., Asmar, D., Shammas, E.: 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)

15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. CoRR **abs/1612.02649** (2016), http://arxiv.org/abs/1612.02649
17. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S.J., Maier-Hein, K.H.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. CoRR **abs/1809.10486** (2018), http://arxiv.org/abs/1809.10486
18. Janai, J., Güney, F., Behl, A., Geiger, A.: Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. ArXiv **abs/1704.05519** (2020)
19. Jégou, S., Drozdzal, M., Vázquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 1175–1183 (2017)
20. Kervadec, H., Dolz, J., Wang, S., Granger, E., ben Ayed, I.: Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. In: Medical Imaging with Deep Learning (2020), https://openreview.net/forum?id=VOQMC3rZtL
21. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1665–1674 (2017). https://doi.org/10.1109/CVPR.2017.181
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2015)
23. Labayrade, R., Aubert, D., Tarel, J.P.: Real time obstacle detection in stereovision on non flat road geometry through" v-disparity" representation. In: Intelligent Vehicle Symposium, 2002. IEEE. vol. 2, pp. 646–651. IEEE (2002)
24. Laddha, A., Kocamaz, M.K., Navarro-Serment, L.E., Hebert, M.: Map-supervised road detection. In: 2016 IEEE Intelligent Vehicles Symposium (IV). pp. 118–123 (2016). https://doi.org/10.1109/IVS.2016.7535374
25. Li, M., Soltanolkotabi, M., Oymak, S.: Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In: International Conference on Artificial Intelligence and Statistics. pp. 4313–4324. PMLR (2020)
26. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3159–3167 (2016). https://doi.org/10.1109/CVPR.2016.344
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
29. Lu, Z., Fu, Z., Xiang, T., Han, P., Wang, L., Gao, X.: Learning from weak and noisy labels for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 486–500 (03 2017). https://doi.org/10.1109/TPAMI.2016.2552172
30. Mairal, J., Elad, M., Sapiro, G.: Sparse representation for color image restoration. Trans. Img. Proc. **17**(1), 53–69 (Jan 2008). https://doi.org/10.1109/TIP.2007.911828, https://doi.org/10.1109/TIP.2007.911828
31. Mayr, J., Unger, C., Tombari, F.: Self-supervised learning of the drivable area for autonomous vehicles. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 362–369. IEEE (2018)

32. Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571 (2016). https://doi.org/10.1109/3DV.2016.79
33. Newcombe, R., Lovegrove, S., Davison, A.: Dtam: Dense tracking and mapping in real-time. pp. 2320–2327 (11 2011). https://doi.org/10.1109/ICCV.2011.6126513
34. Oktay, O., Schlemper, J., Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas (04 2018)
35. Oliveira, G.L., Burgard, W., Brox, T.: Efficient deep models for monocular road segmentation. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4885–4891 (2016). https://doi.org/10.1109/IROS.2016.7759717
36. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
37. Peng, J., Estrada, G., Pedersoli, M., Desrosiers, C.: Deep co-training for semi-supervised image segmentation (2019)
38. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1713–1721 (2015). https://doi.org/10.1109/CVPR.2015.7298780
39. Robinet, F., Demeules, A., Frank, R., Varisteas, G., Hundt, C.: Leveraging privileged information to limit distraction in end-to-end lane following. In: 2020 IEEE 17th Annual Consumer Communications Networking Conference (CCNC). pp. 1–6 (2020). https://doi.org/10.1109/CCNC46108.2020.9045110
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
41. Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., Fergus, R.: 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
42. Tsutsui, S., Kerola, T., Saito, S., Crandall, D.J.: Minimizing supervision for free-space segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 988–997 (2018)
43. Tsutsui, S., Saito, S., Kerola, T.: Distantly supervised road segmentation. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) pp. 174–181 (2017)
44. Watson, J., Firman, M., Monszpart, A., Brostow, G.J.: Footprints and free space from a single color image. In: Computer Vision and Pattern Recognition (CVPR) (2020)
45. Xiao, L., Dai, B., Liu, D., Hu, T., Wu, T.: Crf based road detection with multi-sensor fusion. In: 2015 IEEE Intelligent Vehicles Symposium (IV). pp. 192–198 (2015). https://doi.org/10.1109/IVS.2015.7225685
46. Xie, W., Wei, Q., Li, Z., Zhang, H.: Learning effectively from noisy supervision for weakly supervised semantic segmentation. In: BMVC (2020)
47. Yakubovskiy, P.: Segmentation models. https://github.com/qubvel/segmentation_models (2019)
48. Yao, J., Ramalingam, S., Taguchi, Y., Miki, Y., Urtasun, R.: Estimating drivable collision-free space from monocular video. In: 2015 IEEE Winter Conference on Applications of Computer Vision. pp. 420–427 (2015). https://doi.org/10.1109/WACV.2015.62

49. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6022–6031 (2019)
50. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. International Conference on Learning Representations (2018)
51. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6230–6239 (2017). https://doi.org/10.1109/CVPR.2017.660