2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Mark Yourself: Road Marking Segmentation via Weakly-Supervised Annotations from Multimodal Data

Tom Bruls, Will Maddern, Akshay A. Morye, and Paul Newman

***Abstract*— This paper presents a weakly-supervised learning system for real-time road marking detection using images of complex urban environments obtained from a monocular camera. We avoid expensive manual labelling by exploiting additional sensor modalities to generate large quantities of annotated images in a weakly-supervised way, which are then used to train a deep semantic segmentation network. At run time, the road markings in the scene are detected in real time in a variety of traffic situations and under different lighting and weather conditions without relying on any preprocessing steps or predefined models. We achieve reliable qualitative performance on the Oxford RobotCar dataset, and demonstrate quantitatively on the CamVid dataset that exploiting these annotations significantly reduces the required labelling effort and improves performance.**

I. Introduction

Autonomous vehicles need to understand their workspace for informed decision making and safe navigation in complex urban settings. In contrast to recently developed end-to-end approaches for autonomous driving [1], mediated approaches detect important objects in the scene separately to build a combined, real-time model of the environment that can be employed for navigation and operational purposes. In urban environments, the collection of all painted road markings (e.g. Fig. 1) is critical in such models: their underlying meaning provides rules and guidance to all traffic participants and warns them of potentially dangerous situations. This paper presents a first step towards interpretation of these road rules by presenting a framework for road marking detection in a variety of traffic, lighting, and weather conditions.

In the domain of autonomous vehicles, highly detailed mapping services such as Google Maps, HERE Maps, OpenStreetMap, etc., include road graphs that can support scene understanding. However, relying solely on these can cause problems whenever the traffic situation is updated, or when unmapped places are visited. Even in a future of connected cars, real-time detection and interpretation of road markings will remain an important cue for high-level scene understanding and thereby aid planning, localization [2], and mapping [3].

In this paper, we detect not only separators that mark the different lanes, but the collection of all painted markings on the road surface that dictate the traffic rules for that particular urban setting. Detecting and interpreting these is a more complex problem than lane detection. In general, proposed solutions in that area do not extend easily to the detection of a bigger variety of road markings.

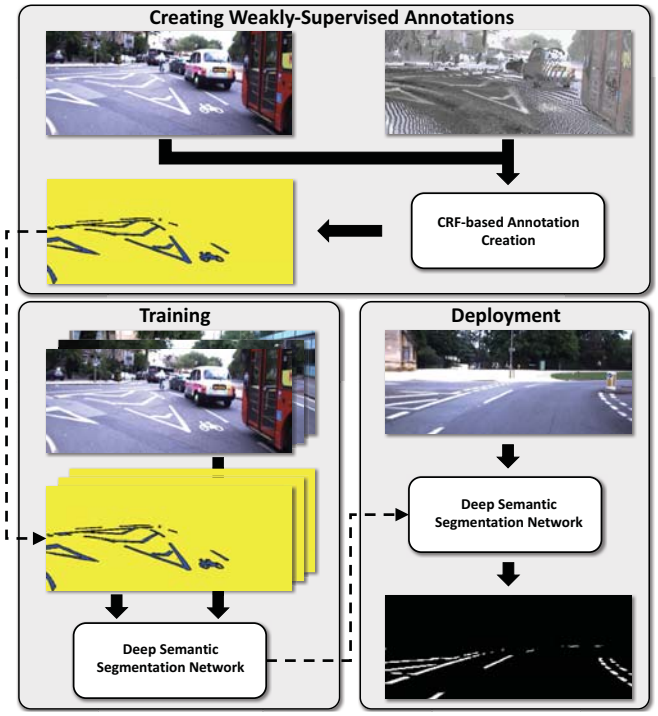

Fig. 1. Road marking detection using weakly-supervised annotations. A LiDAR point cloud of reflectance values is combined with a monocular image to generate road marking annotations in a weakly-supervised way using a conditional random field approach (Section III). A deep semantic segmentation network is then trained using these annotations and the corresponding images (Section IV). During deployment the network performs road marking detection in real time without any additional processing steps using only a monocular camera (Section V).

Road marking detection is a challenging problem for several reasons. Firstly, a proposed method has to cope with occlusions, varying lighting, and changing weather conditions. Secondly, road markings are often degraded and vary in sorts and shapes between countries. Lastly, there are no large datasets available that contain accurate ground-truth labels for road markings. Most datasets for urban scenarios such as KITTI [4], Cityscapes [5], and the Oxford RobotCar dataset [6] do not provide the level of detail that is required for segmenting such small classes.

Road marking detection in images can be posed as a semantic segmentation problem. State-of-the-art methods for these tasks implement deep networks, which are able to learn specific scene context and thereby cope with the challenges stated above, as long as sufficient training data

Authors are from the Oxford Robotics Institute, Dept. Engineering Science, University of Oxford, UK. {tombruls, pnewman}@robots.ox.ac.uk

is available. Although some networks have been trained for road marking recognition [7], [8], their applicability remains limited because of the current lack of ground-truth labels.

Manual generation of these ground-truth labels for semantic segmentation tasks is extremely labour expensive, because of the required pixel-level detail in combination with the aforementioned visual issues. Therefore, we present a method for creating annotations in a weakly-supervised way, by leveraging complementary sensors mounted on the vehicle. We utilize these annotations to train a deep semantic segmentation network (inspired by U-Net [9]) for road marking detection using only a monocular camera. The annotations do not necessarily capture all the road markings in the image perfectly, but are sufficient for training purposes as explained in Section III-C.

We present qualitative results of our approach in a variety of traffic, lighting, and weather conditions on the RobotCar dataset. Furthermore, we show quantitatively that exploiting the weakly-supervised RobotCar annotations significantly reduces the required labelling effort and improves detection performance on the CamVid dataset [10].

We make the following contributions in this paper:

- We present a method for creating road marking annotations in a weakly-supervised way by using complementary sensor modalities. These are used for training a deep semantic segmentation network, thereby avoiding expensive manual labelling.
- We introduce a real-time framework for road marking detection in complex urban settings using a monocular camera without relying on any preprocessing steps or predefined models. This method performs reliably in a wide variety of traffic, lighting, and weather conditions.

The combination of these contributions (see Fig. 1) provides a first step towards road marking classification in datasets without ground-truth labels to support high-level scene understanding, mapping, and planning.

II. Related Work

Our work on road marking segmentation based on weakly-supervised learning from multimodal data is mainly related to work in the area of road marking detection — which we discuss first. We further discuss related work in the areas of lane detection, semantic segmentation, and automatic label generation.

1) Road Marking Detection: Work on road marking detection can generally be distinguished by the used sensor modalities (e.g. camera or LiDAR) and whether learning algorithms are applied (unsupervised or supervised).

Unsupervised camera-based road marking detection systems often follow a four stage pipeline: preprocessing, filtering/binarization, feature extraction, and (rule-based) classification. An early evaluation of several techniques is given in [11]. While effective in moderate environments, these approaches fail in the presence of extreme lighting conditions and shadows. Other disadvantages include hand-crafted features used for template matching [12] and shape-based classification, which both perform badly in the presence of occlusions.

Supervised approaches often use a similar pipeline with the exception that the last step is replaced by a supervised classification algorithm. Popular classifiers include random forests [13], SVMs [14], shallow neural nets [15], and OCR for text recognition [16]. Computed features include HOG and Hu spatial moments, which are rotation and scale invariant and thus perform better under challenging conditions and occlusions. Other approaches [17], [18], do not classify detected road markings independently, but take the spatial configuration of the entire scene into account. This is preferable because road markings are often found in the same spatial configuration.

More recently, deep networks have been successfully introduced for road marking recognition [7], [19] or purely for classification [8]. However, these approaches either implement additional preprocessing algorithms or require detected road markings as an input, because of the current lack of ground-truth road marking labels in large-scale urban datasets. We resolve this issue by creating annotations in a weakly-supervised way.

Lately, the use of LiDAR reflectance values has become more popular as an indication for road markings, since they are not affected by varying lighting. Most solutions generate an interpolated 2D reflectance image [20], so that well-known image processing techniques can be applied. In contrast, the latest approaches work directly on the point cloud [21]. However, because LiDARs are still relatively expensive, these approaches are mainly applied for mapping purposes and not for real-time road marking detection. Therefore, we make use of LiDAR sensors only during the offline annotation creation, and rely solely on a monocular camera during deployment.

2) Lane Detection: Most lane detection systems consist of detection, model fitting, and tracking stages, as summarized in [22]. More recently, deep networks [23], [24] have been proposed, because they perform better under challenging conditions. However, the extracted information does not extend beyond detecting driving lanes.

3) Semantic Segmentation: Semantic segmentation solves a structured pixel-wise labelling problem over meaningful objects in the scene. In early research, maximum-a-posteriori inference in a conditional random field (CRF) [25] was used to compute the labelling layout. More recently, researchers started exploiting deep networks for modelling and extracting these latent feature hierarchies with Fully Convolutional Networks (FCNs) [26]. To improve the output resolution, which suffers from the down- and upsampling in the encoder and decoder path, several solutions have been proposed such as skip connections [9], dilated convolutions [27], and end-to-end integration of a CRF [28].

4) Automatic Label Generation: To fully exploit scene context, the aforementioned networks require large-scale semantic datasets [5]. To reduce the labeling effort for such datasets, several automatic annotation solutions have been proposed. In [29] a single 3D scene annotation is projected

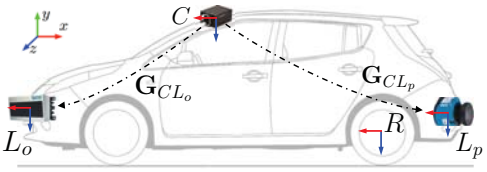

Fig. 2. The vehicle's reference frame R is located at the middle of the rear axle. The approximate sensor locations are shown for the monocular camera C, pushbroom LiDAR L_p, and object detection LiDAR L_o.

into multiple 2D images. The methods proposed in [30], [31] create weakly-supervised annotations for training networks for applications which require less detail and are sometimes supported by a small, manually annotated dataset as in [32]. In this work we automatically create road marking annotations from multimodal data.

III. Weakly-Supervised Annotations from Multimodal Data

We present a method for creating road marking annotations in a weakly-supervised way by leveraging complementary sensor modalities. After the network is trained using these annotations, it requires only a monocular camera at run time. The annotations are computed offline and thus do not require real-time generation.

We exploit the property that road markings are highly reflective and must lie on the road surface. We utilize a LiDAR to capture a point cloud of the environment, with a range and reflectance value associated with each point. The latter is not prone to varying lighting conditions, and thus provides benefits over using (only) camera images. The road surface is extracted from the point cloud using a surface normal region-growing approach and projected into the image to decrease the search area for road markings. A dense CRF is then employed to identify the road marking image pixels by corresponding them with the high-reflectance laser points.

A. Extracting the Road Surface

As road markings only occur on the road itself, coarse segmentation of the road surface can decrease the search domain. This speeds up the algorithm and makes it less prone to false detections (i.e. high-reflectance objects such as white vehicles).

A training route is segmented in $25\,\mathrm{m}$ chunks of laser and image data. The normal of every laser point is calculated using a local neighborhood (empirical evaluation showed that a radius of $0.35\,\mathrm{m}$ achieved good results). From these, the surface normal for the selected point is calculated using principal component analysis (PCA). We employ a per scan-line based region-growing approach (we build our point clouds with a LiDAR mounted in push broom configuration, see Fig. 2) starting at the position of the vehicle and going outwards. The boundary of the road surface is found whenever the surface normal is not parallel to the z-axis of the vehicle anymore. The road surface point cloud is then projected into the camera image using the extrinsic transform $\mathbf{G}_{CL_p}$ to extract the pixels belonging to the road.

Fig. 3. Four examples of extracted road surfaces generated by the surface normal region-growing approach and object detection mask. Highly accurate results are not necessary as this step is only used to restrict the search domain for later steps.

Since LiDAR L_p is mounted in a push broom configuration (see Fig. 2), at any given time, the fields-of-view of LiDAR L_p and camera C do not overlap. Sensor covisibility is simulated by integrating vehicle egomotion estimates. Thus, and since urban scenes are dynamic, the extracted road surface points can project onto dynamic objects such as cars, cyclists, etc. in the image. Hence, we use an additional horizontal LiDAR L_o on the front of the vehicle to capture static and dynamic objects in the scene, and implement the "stixels"-inspired approach of [30] to remove objects from the extracted road region. In Fig. 3 four examples of extracted road surfaces are shown.

B. Classifying the Road Marking Pixels

After the road surface image pixels are extracted, each pixel should be classified as either *road marking* or *non-road marking*. This is a difficult classification problem, since the non-road marking class has a diverse color and texture domain. We use a CRF to associate image pixels with the high-reflectance points of the sparse laser point cloud, because this is a state-of-the-art method for contextual coherent image segmentation in the presence of prior knowledge (i.e. reflectance values).

A CRF models pixel labels as random variables in an undirected graphical model given some observations (i.e. the image). The labelling task is then posed as an energy minimization problem. The framework of [25] is utilized, in which each pixel is represented by a vertex of the graph, and all vertices are connected to each other by Gaussian edge potentials. These pairwise potentials take into account long-range interactions between pixels. At the same time, they ensure that the mean field approximation of the CRF can be computed in a highly efficient manner, so that optimization over a dense pixel-wise model can be performed within seconds.

Let $X_i \in \mathbf{X}$ be the random variable, which represents the label assigned to pixel $i = \{1, \ldots, N\}$, where N is the number of pixels. Each pixel takes a value in the label space $\mathcal{L} = \{l_r, l_n\}$, where l_r denotes the class *road marking* and l_n denotes the class *non-road marking*. Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be the undirected graph, whose vertices X_i are contained in $\mathcal{V}$ and whose edges are contained in $\mathcal{E}$. Given the graph, the combination of the observed image pixels $\mathbf{I}$ and the label

configuration $\mathbf{X}$ can be modelled as a CRF characterized by the Gibbs distribution

$$p(\mathbf{X} = \mathbf{x}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp\left(-E\left(\mathbf{x}|\mathbf{I}\right)\right), \tag{1}$$

where $Z(\mathbf{I})$ is the normalization constant and $E\left(\mathbf{x}|\mathbf{I}\right)$ is the Gibbs energy function defined as

$$E\left(\mathbf{x}|\mathbf{I}\right) = \sum_{c \in \mathcal{C}_\mathcal{G}} \Phi_c(\mathbf{x}_c|\mathbf{I}). \tag{2}$$

In (2), $\mathcal{C}_\mathcal{G}$ denotes the set of cliques associated with $\mathcal{G}$, in which each clique c induces a potential Φ_c. The most probable label assignment given the observed image data is thus found by minimizing the energy: $\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{L}^N} p(\mathbf{X} = \mathbf{x}|\mathbf{I})$. Omitting the conditioning on $\mathbf{I}$ for notational convenience, we use the energy function

$$E\left(\mathbf{x}\right) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j), \tag{3}$$

where $\psi_i(x_i) : \mathcal{L} \rightarrow \mathbb{R}$ are the unary potentials that denote the cost of pixel i taking label x_i, and $\psi_{ij}(x_i, x_j) : \mathcal{L} \times \mathcal{L} \rightarrow \mathbb{R}$ are the pairwise potentials that denote the cost of assigning the labels x_i and x_j to pixel i and j simultaneously. The unary potential can thus be seen as an independent, discriminative pixel classifier, whereas the pairwise potentials are smoothing terms that encourage similar labels for pixels with similar features.

1) Unary Potentials: Ideally, the measured reflectance value provides a good feature for pixel-wise road marking classification, because we have ensured that the search domain only contains the road surface. Unfortunately, this simple classifier will not give satisfactory results for two reasons.

Firstly, the measured reflectance value is a function of the material, the viewing angle, and the distance of the object. We perform a two-step procedure on a per-beam basis to make the reflectance values of a scene comparable: 1) subtract the per-beam median reflectance value, calculated over the entire dataset, from that beam (since in most cases it will not hit a road marking), 2) normalize the values of that beam by dividing them by the per-beam variance calculated over the entire dataset.

Secondly, a point cloud is significantly sparser than an image. In order to compute a unary potential for every vertex (i.e. pixel), the reflectance values of the point cloud are interpolated linearly. This results in a smooth synthetic laser image (see Fig. 4), which cannot be used for creating pixel-accurate unary potentials. Under the assumption that there exists a correlation between the reflectance and brightness of road marking pixels, a simple solution is to multiply the grayscale pixel intensities g_i with the reflectance values of the synthetic image r_i

$$\psi_i(x_i) = g_i \cdot r_i(x_i). \tag{4}$$

In this way, color and reflectance form a joint, discriminative feature for road marking pixels given the road surface, so that only bright *and* highly reflective pixels are assigned an increased potential.

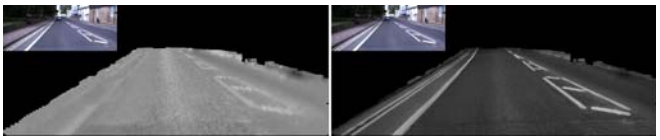

Fig. 4. Generating the unary potentials for the CRF. Interpolating the laser reflectance values results in a smooth synthetic image (*left*) not sufficient for the task. The potentials can be improved by multiplying them with the grayscale intensities of the original image (*right*).

2) Pairwise Potentials: In order to ensure efficient optimization as in [25], define the Gaussian edge potentials as

$$\psi_{ij}(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{M} k_m\left(\mathbf{f}_i, \mathbf{f}_j\right) = \mu(x_i, x_j) K(\mathbf{f}_i, \mathbf{f}_j), \tag{5}$$

where each k_m is a Gaussian kernel which takes a feature vector $\mathbf{f}$ from the respective pixel. We take the compatibility function $\mu(x_i, x_j) = [x_i \neq x_j]$. In contrast to [25], we do not weigh the Gaussian kernels, because learning these weights requires ground-truth labels. However, the same two-kernel potentials are used where the feature vectors $\mathbf{f}$ include the pixel RGB values I at the pixel position p

$$K(\mathbf{f}_i, \mathbf{f}_j) = \exp\left(-\frac{\|p_i - p_j\|^2}{2\theta_\alpha} - \frac{\|I_i - I_j\|^2}{2\theta_\beta}\right) + \exp\left(-\frac{\|p_i - p_j\|^2}{2\theta_\gamma}\right). \tag{6}$$

The first exponential function forces nearby pixels with similar features to have the same label, while the second smoothens the results by removing small, isolated regions. The θ parameters control the amount of influence between pixel i and j; increasing θ will increase long-range interactions. We empirically choose $\theta_\alpha = 43$, $\theta_\beta = 9$, and $\theta_\gamma = 3$. This choice was inspired by [29].

C. Annotation Results

Qualitative evaluation of the created annotations demonstrates that high-quality results are achieved, as illustrated in Fig. 5. The current approach does not classify all the road marking pixels in every image perfectly. This happens due to the fact that the method is unsupervised and the dataset contains images with varying lighting conditions and reflectance range. Learning weights for the kernels to adjust to specific images is challenging due to the lack of ground-truth labels. The results might be improved if weights are learned from a relatively small set of manually labelled images.

However, as shown later in Section V, the generated annotations are sufficient for detecting road markings in urban settings under varying conditions. The most likely reason for this is that several regularization techniques incorporated in the network such as dropout and batch normalization prevent overfitting to the imperfect annotations. The best generalized binary segmentation that the network is able to achieve, groups the road marking pixels in one class, since

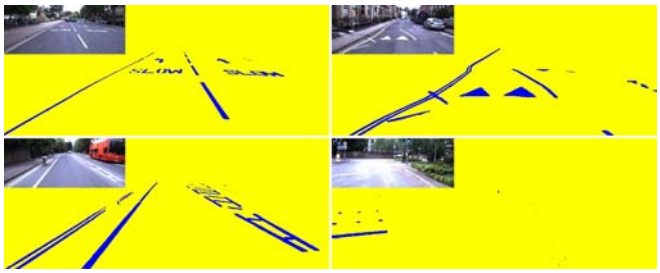

Fig. 5. High-quality annotation results achieved by the CRF approach. Although not all road markings are captured perfectly, these results are sufficient for training. In the case of over-exposure (*bottom right*), annotations are conservatively estimated.

their appearance is very similar to the correctly labelled road marking pixels.

Note that, although the CRF approach achieves good results, it is not suited for real-time applications with the current inference algorithm, because processing of a single high-resolution image takes several seconds. Furthermore, we do not claim that the feature and parameter choices are optimal (see Section III-D), but they generate annotations that are sufficient for training the network, which is the end goal.

D. Alternative Features for the CRF Potentials

We have experimented with different features for the unary and pairwise potentials in order to improve the annotations. Below we briefly share our findings. However, a more extensive analysis is necessary to determine the best overall feature type for this specific application.

For the unary potentials, we found that the Nguyen feature [33] tends to work well in certain settings as a substitute for the grayscale intensities. This is likely because the Nguyen feature emphasizes elongated structures such as lane markings, and is thus less prone to regions of over-exposure.

Intuitively, the RGB values in the pairwise potentials do not seem to be the best feature to discriminate the road surface from road markings, especially not in over-exposed images. Therefore, we have experimented by adding the interpolated reflectance value for every pixel to the feature vector. However, this gave unsatisfactory smoothed results, even when the respective θ value was decreased. Furthermore, we have experimented with different color spaces such as CIELUV and HSV, but empirically achieved the best results across the entire dataset using the RGB values.

IV. Deep Semantic Segmentation Network

Deep neural networks are the state-of-the-art solution for semantic segmentation. We argue that these methods (with adequate training data) will also improve road marking detection and classification, since they are able to leverage the global scene context and are robust to spatial deformations, degradation, and partial occlusion. Besides that, classification is not limited to shapes, but the difference in underlying meaning of similarly shaped road markings (e.g. lane separators and separators that mark a parking spot) can be retrieved based on their place and context in the scene.

A. Network Architecture

We train a U-Net inspired architecture shown in Fig. 6. Like most deep semantic segmentation networks, it consists of an encoding and a decoding path, and a way to provide fine-grained input information to the decoder.

The size of the image is repeatedly reduced by a factor of 2 in the encoder path to increase the receptive field of the filters. Consequently, they become invariant to tiny deformations of the road markings and are able to take contextual information and long-range interactions into account. The decoding path is identical to the encoding path except that the feature maps are now repeatedly upsampled to generate an output image of the same resolution as the input. The upsampling is performed with trainable filters. Skip connections concatenate high-resolution features from the encoding path to the decoding path, so that fast convergence is ensured and a fine-grained segmentation output can be achieved. We modified the original U-Net to include batch normalization after every convolutional filter, and added zero-padding to the sides so that the output resolution is equivalent to the input resolution.

The output of the network is computed by a pixel-wise softmax over the final feature maps

$$p_{i,k} = \frac{\exp(a_{i,k})}{\sum_{m=1}^{M} \exp(a_{i,m})}, \tag{7}$$

where $a_{i,k}$ denotes the activation in feature map k at pixel i, and M is the number of classes. Then, pixel i is assigned a label by $l_i = \arg\max_k p_{i,k}$. Since the number of road marking pixels is much lower than non-road marking pixels, a weighted cross entropy loss is implemented to cope with the class imbalance

$$E = -\sum_{i=1}^{N} w_{l_\star} \log(p_{i,l_\star}), \tag{8}$$

where $l_\star$ is the ground-truth class for that pixel and $w_{l_\star}$ is the weight associated with the ground-truth class of that pixel.

Weights for the two classes are calculated by median frequency balancing $w_m = \tilde{f}/f_m$ [34], where f_m is the total number of pixels of class m divided by the total number of pixels in images where class m is present. The scalar $\tilde{f}$ denotes the median of f_m.

B. Network Training

The parameters that were used during training against the created RobotCar annotations are shown in Table I. We use dropout as a supplementary regularization tool besides batch normalization to prevent overfitting. Training is done from scratch with weight initialization as described in [35].

For the quantitative results, we split up the CamVid dataset into 490 train, 105 validation, and 105 test frames. We select the epoch for testing in which the accuracy is highest among the evaluations on the validation set.

At run time, the TensorFlow implementation of our network in Python performs inference on an input image in 16 ms (=62.5 Hz) using an NVIDIA TITAN Xp GPU.

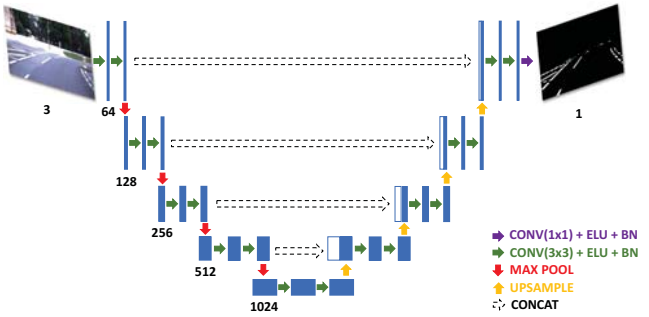

Fig. 6. The U-Net [9] architecture consisting of an encoder and a decoder path, which compresses the feature maps to increase the receptive field of the filters before expanding to a full resolution per-pixel class prediction.

TABLE I
NETWORK & TRAINING PARAMETERS

Network	Value	Training	Value
loss function	weighted cross entropy	batch size	10
activation function	ELU	epochs	100
number of layers	5	optimizer	Adam
filter size	3×3	learning rate	0.0001
max pool size	2	dropout	0.5
stride	1		
image resolution	128 × 320		

V. EXPERIMENTAL RESULTS

Due to the absence of a readily available dataset that contains LiDAR data and pixel-wise ground-truth labels of road markings, we employ the following approach to test our system. We train the network using the weakly-supervised annotations created on the RobotCar (RC) dataset, and then fine-tune with manually created labels on the CamVid (CV) dataset to adapt to the different domain. This process allows for pixel-wise evaluation of our approach against the CamVid labels, which will be used as ground truth. Additionally, we show qualitative performance of the network when trained using only the annotations created on the RobotCar dataset, in a variety of traffic, lighting, and weather conditions.

A. Quantitative Evaluation

We performed five experiments, all tested against the 105 selected ground-truth CamVid labels (see Table II).

The first two experiments depict baseline results on CamVid by training against a small set of, and all available ground-truth labels, respectively. For the remaining experiments, the network was trained using 24238 weakly-supervised RobotCar annotations. Herein, the third experiment was tested directly against the CamVid labels, whereas for the fourth and fifth experiments, the network was fine-tuned on a varying number of CamVid labels. Evaluating the results, the following three key observations can be made:

1) The third experiment clearly illustrates that fine-tuning is necessary. Interestingly, the result demonstrates also that training against a large dataset of another domain outperforms training against a small dataset of the actual test domain. This likely occurs because the network is trained on a bigger variety of traffic and lighting conditions, which improves generalization.

2) The fourth experiment shows that training using the weakly-supervised annotations, while fine-tuning using only 25 manually created CamVid labels, achieves comparable performance (in terms of IoU and F_1) to the baseline result trained on 490 manual labels. This significantly reduces the required labelling effort.

3) The last experiment shows that we outperform the baseline result, when we fine-tune using all available ground-truth labels. This is not trivial, since adding more data from a different domain potentially alters the data distribution. The result indicates that more training data of another domain (which requires no additional labelling effort in our case) improves performance.

Note that the RobotCar annotations were uniquely generated without the use of data augmentation techniques. The results further show that pre-training on the annotations increases the precision but decreases the recall. This is expected, since the annotations are created conservatively (see Fig. 5).

Although the manually created CamVid road marking labels are of high quality, there are instances where the labels do not accurately represent the ground truth. As shown in Fig. 7, the predicted output can then correspond better to the actual ground truth than the label itself. Besides, it is important to keep in mind that object-level performance is more relevant than pixel-wise performance, when road marking detection is performed for planning purposes (which is our future goal).

Fig. 7. The CamVid label (*middle*) and the predicted output (*right*) for a test image. The predicted output reflects the ground truth better at several places in the images such as the lower part of the bounding box around the bicycle and the bicycle itself.

TABLE II
QUANTITATIVE PIXEL-WISE RESULTS ON ROBOTCAR (RC) AND CAMVID (CV) DATASETS

Train Dataset	ACC	PRE	REC	IoU	F_1
25 CV	96.82	46.17	**87.64**	42.03	58.33
490 CV	98.22	63.96	86.33	57.17	71.10
24238 RC	97.92	62.92	65.25	46.17	62.52
24238 RC + 25 CV	98.20	66.39	78.39	54.27	69.54
24238 RC + 490 CV	**98.60**	**72.64**	81.63	**61.20**	**75.04**

B. Qualitative Evaluation

Fig. 8 shows qualitative results on a RobotCar test dataset. The results demonstrate that the network segments the road markings from the image without any preprocessing steps when trained using the weakly-supervised annotations. Even

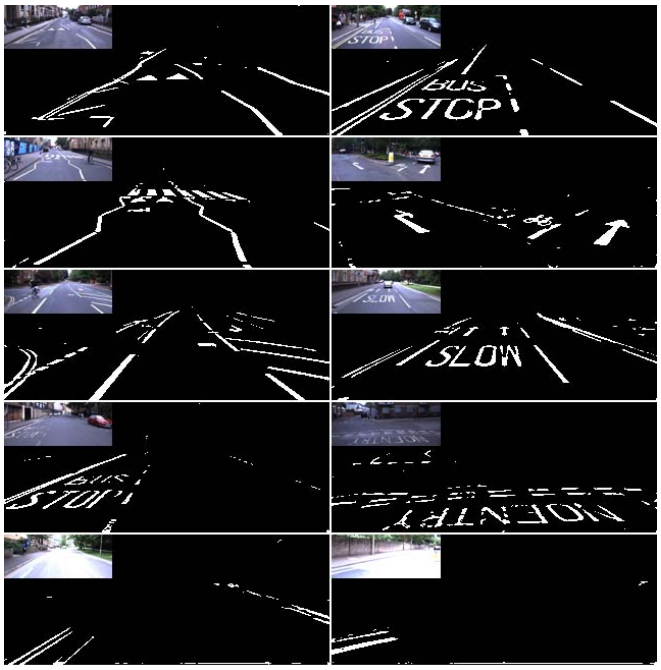

Fig. 8. Network output on RobotCar images when trained against weakly-supervised annotations. The network accurately detects all road markings without limitations to shape, even when the road markings start to degrade (*fourth row*). In case of over-exposure (*fifth row*), a conservative segmentation is achieved.

in case of degradation, the network is able to sufficiently segment the road markings. The network achieves a conservative segmentation in cases of over-exposure, where intensity-based approaches most likely fail.

Additionally, we trained a network using annotations generated under different lighting and weather conditions. Fig. 9 shows the network output under these conditions at the same location. Although the method performs best in overcast conditions, the results under more difficult conditions appear satisfactory considering the image quality.

C. Limitations

Under some conditions the quality of the annotation is poor, as illustrated in Fig. 10. Bright parts of the pavement can be mistaken for road marking, when the extraction algorithm has difficulties finding the correct road surface border. These failure cases could be addressed by more complex road extraction algorithms, or a more discriminative (supervised) feature set for the CRF potentials. These annotations were not included in the training set.

Furthermore, the network output can be spurious at times in the presence of parked cars or stark shadow lines, as shown in Fig. 11. False detections occur, because object edges introduce high-intensity gradients at the same place in the image where road markings normally appear. This can likely be resolved when the network is given annotations with road marking types, so that it can learn improved spatial and contextual coherence.

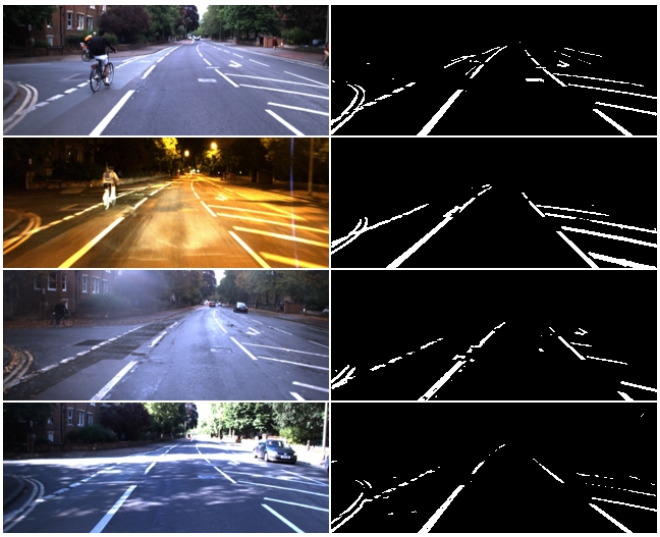

Fig. 9. Road marking detection under different conditions (overcast, night, rain, and sun) at the same location. Despite significant changes in appearance, the method achieves satisfactory results.

Fig. 10. Poor quality annotation due to insufficient road extraction, because the pavement is approximately at road height. The result can be improved by more accurate road extraction algorithms or a more discriminative feature set for the CRF potentials.

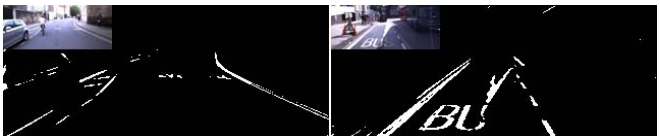

Fig. 11. Examples of spurious network output in the presence of parked cars and stark shadow lines, because edges introduce high-intensity gradients at the same place in the image where road markings normally appear.

VI. Conclusion

We have presented a weakly-supervised system for real-time road marking detection using images of complex urban environments obtained from a monocular camera. At run time, the road markings in the scene are detected using a deep segmentation network without relying on any pre-processing step or predefined models. Crucially, by leveraging LiDAR reflectance values in a CRF approach, we generated vast quantities of annotated road marking images for training purposes in a weakly-supervised way, thereby avoiding the need for expensive manual labelling. We have demonstrated reliable qualitative performance under varying traffic, lighting, and weather conditions on the Oxford RobotCar dataset. Furthermore, we showed quantitatively on the CamVid dataset that weakly-supervised annotations of another domain significantly reduce the required labelling effort and improve performance.

In future work we will extend the current framework to

include semantic classification of the road markings in the scene to retrieve the rules of the road. This information will be exploited to aid high-level scene understanding, mapping, and planning in complex urban environments.

REFERENCES

[1] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2174–2182.

[2] M. Schreiber, C. Knöppel, and U. Franke, "Laneloc: Lane marking based localization using highly accurate maps," in *Intelligent Vehicles Symposium (IV), 2013 IEEE*. IEEE, 2013, pp. 449–454.

[3] M. Schreiber, F. Poggenhans, and C. Stiller, "Detecting symbols on road surface for mapping and localization using ocr," in *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*. IEEE, 2014, pp. 597–602.

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[6] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017.

[7] O. Bailo, S. Lee, F. Rameau, J. S. Yoon, and I. S. Kweon, "Robust road marking detection and recognition using density-based grouping and machine learning techniques," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 760–768.

[8] T. Ahmad, D. Ilstrup, E. Emami, and G. Bebis, "Symbolic road marking recognition using convolutional neural networks," in *Intelligent Vehicles Symposium (IV), 2017 IEEE*. IEEE, 2017, pp. 1428–1433.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[10] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[11] T. Veit, J.-P. Tarel, P. Nicolle, and P. Charbonnier, "Evaluation of road marking feature extraction," in *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*. IEEE, 2008, pp. 174–181.

[12] T. Wu and A. Ranganathan, "A practical system for road marking detection and recognition," in *Intelligent Vehicles Symposium (IV), 2012 IEEE*. IEEE, 2012, pp. 25–30.

[13] D. Hyeon, S. Lee, S. Jung, S.-W. Kim, and S.-W. Seo, "Robust road marking detection using convex grouping method in around-view monitoring system," in *Intelligent Vehicles Symposium (IV), 2016 IEEE*. IEEE, 2016, pp. 1004–1009.

[14] B. Qin, W. Liu, X. Shen, Z. J. Chong, T. Bandyopadhyay, M. Ang, E. Frazzoli, and D. Rus, "A general framework for road marking detection and analysis," in *Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on*. IEEE, 2013, pp. 619–625.

[15] F. Poggenhans, M. Schreiber, and C. Stiller, "A universal approach to detect and classify road surface markings," in *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*. IEEE, 2015, pp. 1915–1921.

[16] J. Greenhalgh and M. Mirmehdi, "Automatic detection and recognition of symbols and text on the road surface," in *International Conference on Pattern Recognition Applications and Methods*. Springer, Cham, 2015, pp. 124–140.

[17] B. Mathibela, P. Newman, and I. Posner, "Reading the road: road marking classification and interpretation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2072–2081, 2015.

[18] T. Woudsma, L. Hazelhoff, P. H. de With, and I. Creusen, "Automated generation of road marking maps from street-level panoramic images," in *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*. IEEE, 2015, pp. 925–930.

[19] T. Chen, Z. Chen, Q. Shi, and X. Huang, "Road marking detection and classification using machine learning algorithms," in *Intelligent Vehicles Symposium (IV), 2015 IEEE*. IEEE, 2015, pp. 617–621.

[20] M. Cheng, H. Zhang, C. Wang, and J. Li, "Extraction and classification of road markings using mobile laser scanning point clouds," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 3, pp. 1182–1196, 2017.

[21] Y. Yu, J. Li, H. Guan, F. Jia, and C. Wang, "Learning hierarchical features for automated extraction of road markings from 3-d mobile lidar point clouds," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 2, pp. 709–726, 2015.

[22] A. Bar Hillel, R. Lerner, D. Levi, and G. Raz, "Recent progress in road and lane detection: A survey," *Mach. Vision Appl.*, vol. 25, no. 3, pp. 727–745, Apr. 2014.

[23] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, F. Mujica, A. Coates, and A. Y. Ng, "An empirical evaluation of deep learning on highway driving," *CoRR*, vol. abs/1504.01716, 2015.

[24] B. He, R. Ai, Y. Yan, and X. Lang, "Lane marking detection based on convolution neural network from point clouds," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Nov 2016, pp. 2475–2480.

[25] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 109–117.

[26] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017.

[27] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[28] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[29] J. Xie, M. Kiefel, M.-T. Sun, and A. Geiger, "Semantic instance annotation of street scenes by 3d to 2d label transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3688–3697.

[30] D. Barnes, W. Maddern, and I. Posner, "Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 203–210.

[31] J. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez, "Road scene segmentation from a single image," in *European Conference on Computer Vision*. Springer, 2012, pp. 376–389.

[32] W. Wang, N. Wang, X. Wu, S. You, and U. Neumann, "Self-paced cross-modality transfer learning for efficient road segmentation," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1394–1401.

[33] U. T. Nguyen, A. Bhuiyan, L. A. Park, and K. Ramamohanarao, "An effective retinal blood vessel segmentation method using multi-scale line detection," *Pattern recognition*, vol. 46, no. 3, pp. 703–715, 2013.

[34] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.