

Generating All the Roads to Rome: Road Layout Randomization for Improved Road Marking Segmentation

Tom Bruls, Horia Porav, Lars Kunze, and Paul Newman

Abstract—Road markings provide guidance to traffic participants and enforce safe driving behaviour, understanding their semantic meaning is therefore paramount in (automated) driving. However, producing the vast quantities of road marking labels required for training state-of-the-art deep networks is costly, time-consuming, and simply infeasible for every domain and condition. In addition, training data retrieved from virtual worlds often lack the richness and complexity of the real world and consequently cannot be used directly. In this paper, we provide an alternative approach in which new road marking training pairs are automatically generated. To this end, we apply principles of domain randomization to the road layout and synthesize new images from altered semantic labels. We demonstrate that training on these synthetic pairs improves mIoU of the segmentation of rare road marking classes during real-world deployment in complex urban environments by more than 12 percentage points, while performance for other classes is retained. This framework can easily be scaled to all domains and conditions to generate large-scale road marking datasets, while avoiding manual labelling effort.

I. INTRODUCTION

Safety-critical systems, such as automated vehicles, need interpretable and explainable decision-making for real-world deployment. An important aspect for improving interpretability of such systems is the ability to explain scenes semantically. More specifically, planning the behaviour of an automated vehicle through an urban traffic environment requires understanding of the *road rules*. These are conveyed to the traffic participants by the markings painted on the road.

Although semantic reasoning about road markings is ideally performed at an object and scene level [1], state-of-the-art deep learning methods perform semantic segmentation at the pixel level. This, however, requires thousands of pixel-labelled images for different environments and conditions, which is a problem for several reasons. Firstly, it is impossible to label every pixel of every image for every city in every condition manually. Secondly, simple data augmentation techniques [2] (e.g. flipping, translating, adjusting contrast, etc.) do not deliver the necessary diversity to adapt to all encountered environments and conditions [3].

Even if more efficient hand-labelling techniques become available in the future, we still face the issue of *edge cases* that appear very infrequently in regular driving. In the context of road marking segmentation, data collection during regular driving creates extremely imbalanced datasets. For example, zigzag markings (which indicate a pedestrian crossing, Fig. 1) are encountered rarely, but their detection is

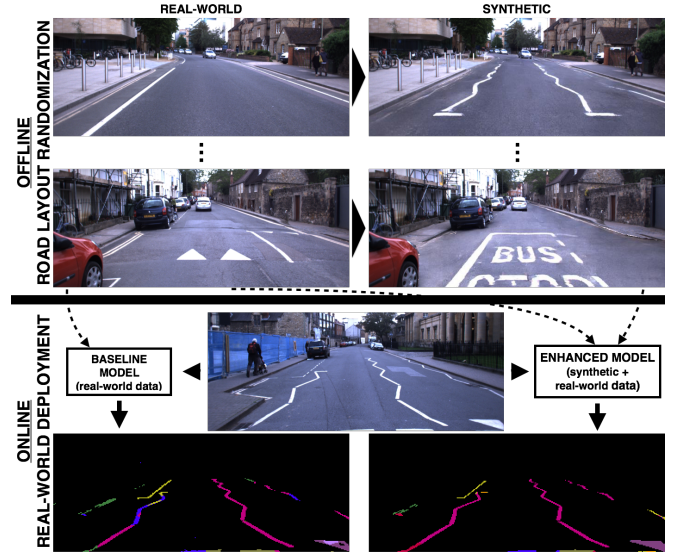


Fig. 1. Road layout randomization for improved road marking segmentation, while avoiding manual labelling. *Offline*: Firstly, new images for training road marking segmentation networks are automatically generated by synthesizing new road surfaces from altered semantic labels. *Online*: Subsequently, mIoU of the segmentation of rare road markings (e.g. zigzags shown in pink) is improved by more than 12 percentage points during *real-world* deployment by the enhanced model trained on a hybrid dataset when compared to a baseline model only trained on the real-world dataset.

critical for safe operation. Resampling or applying a class-weighted loss function are not viable solutions for small, hand-labelled datasets, since these simply contain insufficient examples of rare classes for proper generalization. Retrieving more examples is labour intensive in terms of driving and labelling time. Consequently, trained classifiers show decreased performance on infrequently-occurring classes [4].

The latter problem could be solved by creating a virtual environment (i.e. simulator), in which the desired road markings can be reproduced as many times as necessary. However, this introduces several new challenges. Firstly, even though state-of-the-art simulators can appear realistic to the human-eye, their fidelity lacks the richness and complexity of the real world and consequently there is still an apparent domain gap between simulated environments and their real-world equivalent. As a result, domain adaptation techniques need to be applied for real-world deployment [5], [6]. Secondly, although we might be able to generate simulated environments from real-world data in the future [7], at present their design remains a manual, costly, and time-consuming task. Besides, since urban environments can vary substantially between countries, there is a need for highly-configurable

virtual worlds, which increases the labour cost.

Recently, alternative methods have been developed [8] to synthesize new, photo-realistic scenes for a domain of interest by employing Generative Adversarial Networks (GANs). These approaches require relatively little human effort and can easily extend to all kinds of different conditions [9]. This provides the ability to generate large-scale datasets for semantic scene understanding in a domain of interest at low cost. Most of these frameworks take real-world scenes and augment them by placing or removing objects (e.g. cars, pedestrians, etc.). This can be done randomly [10] or more naturally by learning from real-world examples [11], [12].

Similarly, we place instances of chosen road markings into newly-synthesized, photo-realistic scenes, which are then used to train a road marking segmentation network. In this way, we generate sufficient examples of *rare* road marking classes to achieve the generalization performance required during real-world deployment, as visualized in Fig. 1. However, placing new road markings coherently into the scene is difficult, since there are many dependencies such as the type of road / intersection, traffic lights, parked cars, etc. that need to be taken into account. We avoid solving this hard problem by employing the principles of domain randomization [13]. More concretely, we place road markings at random places on the road surface, not necessarily coherent with other elements in the scene. In this way, we perform *road layout randomization*. Real-world scenes encountered during deployment then appear as samples of the broadened distribution on which the model was trained.

We demonstrate quantitatively that training on these synthetic labels improves mIoU of the segmentation of rare road marking classes, for which it is expensive to attain sufficient real-world examples, during real-world deployment in complex urban environments by more than 12 percentage points. To take full advantage of the synthetic labels we introduce a new class-weighted cross-entropy loss which balances the training. Furthermore, we show qualitatively that the segmentation performance for other classes is retained.

We make the following contributions in this paper:

- We present a method for generating large-scale road marking datasets for a domain of interest by leveraging principles of domain randomization, while avoiding expensive manual effort.
- We introduce a new class-weighted cross-entropy loss to balance the training on synthetic datasets with large class-wise imbalance in terms of their occurrence.
- We demonstrate a real-time framework for improving the segmentation of (rare) road marking classes in *real-world*, complex urban environments.

II. RELATED WORK

Road Marking Segmentation: Deep networks are increasingly used to perform lane detection in highway scenarios [14]–[16]. However, the urban environments and road markings targeted in this paper are substantially different and more complex, and thus require a different approach. This problem has seen significantly fewer deep learning solutions,

due to a lack of large-scale datasets containing road markings. The first large-scale semantic road marking dataset was recently introduced in [17], however it is extremely expensive to manually expand this to all environments and conditions.

Road marking segmentation as demonstrated in [4] is closest to the application of this paper. The authors train a network for semantic road marking segmentation and improve their results by predicting the vanishing point simultaneously. In contrast to this paper, they require thousands of hand-labelled images, which is very labour expensive. Alternatively, the authors of [18] hand-label road markings such as arrows and bicycle signs and train an object detection network to predict bounding boxes instead of pixel segmentations. In previous work [19] (includes more extensive review), we have introduced a weakly-supervised approach for binary road marking segmentation, which is used here to acquire road marking labels for real-world scenes.

Synthetic Training for Automated Driving Tasks: To prevent costly and time-consuming manual labelling of training data, many approaches leverage synthetic datasets. Early works trained on purely virtual data to perform object detection [20], [21] or semantic segmentation [5], [6].

However, virtual data lacks the richness and complexity of the real world. A possible alternative is to augment real-world data. For the task of semantic segmentation this means either generating new, photo-realistic images from semantic labels [8], [22], [23] or enriching semantic labels with virtually-generated information [24]. Both of these principles are applied in this paper. For object detection tasks, the main difficulty is to place the (dynamic) objects coherently into the scene. The simplest solution is random object placement (i.e. domain randomization) [10]. Alternatively, the authors of [25], [26] place photo-realistic, synthetic cars into real-world images by taking into account the geometry of the scene. The most recent approaches [11], [12], [27], [28] learn context-aware object placement from real-world examples. However, placing dynamic objects, such as pedestrians, seems less complex than road markings, because the space of realistic solutions is less restrictive. Therefore, we place road markings randomly onto the road surface in this paper.

Scene Manipulation: Recently, several approaches have been introduced for more complex scene manipulation, beyond simple augmentation. Additional sensor modalities are used in [29] to offer the flexibility (e.g. different view points) of a virtual simulator, while generating data with the fidelity and richness of real-world images. The authors of [30] introduce a probabilistic programming language to synthesize complex scenarios from existing domain knowledge. Another system [31] offers similar levels of control, while the camera sensor is modelled accurately at the same time. These frameworks potentially offer a way to generate improved training data for our approach.

III. GENERATING SYNTHETIC TRAINING PAIRS

In this section, we explain in detail how to generate synthetic training pairs for road marking segmentation networks to improve performance during real-time deployment, as

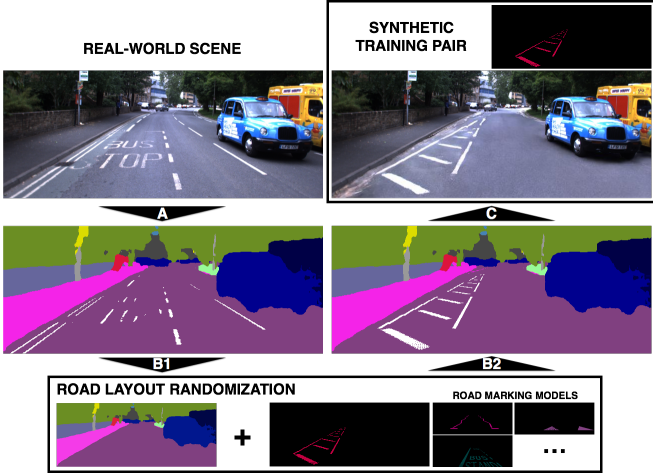


Fig. 2. Road layout randomization: generating synthetic training data based on real-world scenes. The process has the following steps (as described in the respective subsections of Section III): (A) semantic segmentation of the real-world scene is acquired, (B1) the road markings are removed and replaced with road surface, (B2) instances of chosen road markings (modelled according to the UK Highway Code) are placed randomly on the road surface, and finally (C) the road surface of the original image is replaced with a GAN-synthesized, photo-realistic alternative based on the altered semantic label. The composite image is then paired with the generated road marking label.

shown in Fig. 2. We demonstrate that this framework can be employed on any driving dataset even when no ground-truth semantic or road marking labels are available.

A. Retrieving Semantic Labels for Real-World Scenes

In order to generate synthetic training pairs for road marking segmentation, the road layout of semantic labels of real-world scenes is altered and from these new, photo-realistic images are synthesized. Ground-truth semantic labels are not required for the domain of interest, since semantic segmentation of reasonable (i.e. sufficient) quality can be acquired from a model pretrained on the Cityscapes dataset¹. In this way, we retrieve semantic labels of real-world scenes from the Oxford RobotCar dataset [32], as shown in Fig. 3.

Unfortunately, the available model is not trained to segment road markings (Cityscapes does not contain road marking masks). However, semantic labels including road markings and their corresponding real-world images are necessary to train the GAN described in Section III-C. We prevent manual labelling of road markings by employing the techniques of [19] to generate large quantities of road marking annotations automatically. Because these annotations are generated automatically, they are not equivalent to the ground-truth, however they have proven to be sufficient for training purposes if regularization techniques are applied. The road markings are added to the semantic labels acquired from the Cityscapes model, as visualized in Fig. 3.

B. Road Layout Randomization

To form new road marking training pairs, we alter the road layout (i.e. road markings) of the retrieved semantic labels

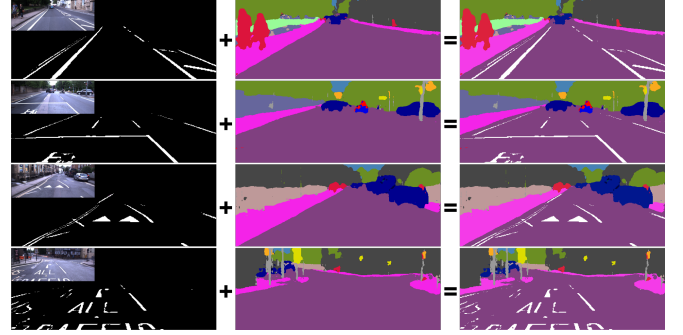


Fig. 3. We augmented the Oxford Robotcar Dataset with semantic labels including road markings to train the CGAN described in Section III-C. The semantic segmentation label is retrieved from inference with a pretrained Cityscapes model and combined with automatically generated road marking annotations from [19]. The resulting labels are not perfect ground-truth, but they are sufficient for the task and can be acquired at low cost.

and subsequently synthesize a new corresponding image. In order to rebalance datasets collected during regular driving, we create new semantic labels with road markings which occur relatively infrequently in the real world (e.g. pedestrian crossing, arrows, etc.). By training the road marking segmentation network on the rebalanced dataset, the goal is to improve the performance for these respective *rare* classes, while at the same time retaining the overall performance.

As mentioned before, the type and placement of road markings is dependent on many factors of the scene such as the type of road, traffic lights, and even the traffic participants. Altering all of these coherently according to the real world is difficult and seems similar in terms of complexity to designing a simulator. Therefore, we choose to leverage domain randomization principles [10]. We vary position (and scale accordingly), rotation, quantity, and partial occlusion of the road markings that are placed into the environment and in that way perform *road layout randomization* to create vast quantities of new training pairs automatically. For accurate placement, we use the camera sensor calibration of the vehicle and assume that the road surface is planar and horizontal. Training the network on many randomly-generated pairs improves generalization in newly-encountered, real-world scenes, which then appear as variations of the distribution on which the network was trained.

Concretely, we start by erasing the original road markings from the real-world semantic labels and subsequently place a new road marking instance onto the cleared road surface. The classes are realistically modelled according to the UK Highway Code so that their shape, size, colour and configuration (e.g. zigzags appear in dual or triple configurations) resemble the real world. Some examples for different classes of rare road markings are given in Fig. 4.

C. Synthesizing Photo-Realistic Images

In order to create a synthetic training pair, we train a Conditional Generative Adversarial Network (CGAN), as introduced in [8], to synthesize a photo-realistic RGB image for the altered semantic label (from Section III-B). In this

¹https://github.com/tensorflow/models/blob/master/research/deeplab/g3doc/model_zoo.md



Fig. 4. Examples of newly-synthesized training images for several rare road marking classes (i.e. zigzag, diagonal stripes, bus stop, and small warning triangles) by employing road layout randomization. The top two rows show images of real-world scenes of the Oxford RobotCar dataset together with the corresponding (partial) semantic labels (Section III-A). The third row visualizes the altered semantic labels in which instances of chosen road markings are placed randomly on the road surface (Section III-B). The last row presents the newly-synthesized road surfaces substituted into the real-world images. The GAN is able to generate road surfaces with photo-realistic textures, lighting, and even degradation as exemplified on the letters of the bus stops.

framework the generator G aims to synthesize the RGB images, while the discriminator D tries to distinguish synthesized from real-world images. The CGAN is trained in a supervised setting using real-world images and corresponding semantic labels retrieved in Section III-A. After the training is completed a photo-realistic image can be synthesized by the generator from the altered semantic labels generated in Section III-B, as shown in Fig. 4.

More specifically, the framework incorporates several advancements over previous works which make it possible to generate higher-resolution images. Firstly, the generator architecture follows a traditional downsample-bottleneck-upsample model, but splits into a global generator and a local enhancer, where the local component is forced to learn high-resolution details for the stabilized features of the global component. Secondly, to overcome discriminator capacity limitations which arise from training with high-resolution images, the framework incorporates three similar discriminators that work on different scales. The discriminators with bigger receptive field enforce more globally consistent image generation, while the smaller receptive fields steer the generator towards more realistic, fine-level details. Lastly, the traditional GAN loss is augmented to include a feature matching loss based on the discriminator. Formally, following the architecture described in [8], given $K = 3$ discriminators D_k , each operating on a different scale, along with the input and label images $I_{\text{input}}^{\text{SEG}}$ and $I_{\text{label}}^{\text{RGB}}$, respectively, the final objective to be minimized is:

$$\mathcal{L}_{\text{tot}} = \min_G \left(\max_{D_1, D_2, D_3} \sum_{k=1,2,3} \mathcal{L}_{\text{GAN}}(G, D_k) \right) + \lambda_{\text{FM}} \sum_{k=1,2,3} \mathcal{L}_{\text{FM}}(G, D_k) + \lambda_{\text{VGG}} \mathcal{L}_{\text{VGG}}(G). \quad (1)$$

Here, $\mathcal{L}_{\text{GAN}}(G, D_k)$ represents the usual GAN loss (see [8]) defined over K scales, $\mathcal{L}_{\text{FM}}(G, D_k)$ is the discriminator

feature loss defined over K scales:

$$\mathcal{L}_{\text{FM}}(G, D_k) = \sum_{i=1}^{l_D} \frac{1}{w_i} \|D_k(I_{\text{label}}^{\text{RGB}})_i - D_k(G(I_{\text{input}}^{\text{SEG}}))_i\|_1, \quad (2)$$

with l_D defining the number of layers from the discriminator used in the discriminator feature loss and $\mathcal{L}_{\text{VGG}}(G)$ being the perceptual loss:

$$\mathcal{L}_{\text{VGG}}(G) = \sum_{i=1}^{l_P} \frac{1}{w_i} \|\text{VGG}(I_{\text{label}}^{\text{RGB}})_i - \text{VGG}(G(I_{\text{input}}^{\text{SEG}}))_i\|_1, \quad (3)$$

with l_P defining the number of layers from an ImageNet-trained network (in this case VGG16) used in computing the perceptual loss. The factors $w_i = 2^{l-i}$ are utilized to scale the weight of each network layer used in computing the losses. We train the model on 3351 overcast training pairs while using the settings as specified in [8] to generate images with a resolution of 256×640 .

Unfortunately, the RobotCar dataset does not contain any boundary or instance labels (as used in [8]) necessary to generate sharp, high-quality images. Consequently, the generated images can be smudgy around object boundaries (e.g. rows of parked cars are merged because of the image perspective, as exemplified in [8]) and contain unnatural artifacts. Therefore, we choose to substitute only the newly-generated road surface and keep the rest of the original image intact. The RobotCar dataset contains sufficient real-world images so that no background duplicates have to exist in the new road marking dataset. In this way, we are able to generate a large-scale urban datasets for road marking segmentation, while avoiding expensive manual labelling.

The above-described framework can easily be extended to different (weather and lighting) conditions by training condition-specific models. If it is not possible to retrieve semantic labels of sufficient quality under difficult conditions, a state-of-the-art invertible generator, that can transform the images into the desired appearance similar to [9], [33], can be employed. In this way the semantic label acquired from the

overcast image can be paired with an image which resembles a different weather or lighting condition.

IV. TRAINING FOR ROAD MARKING SEGMENTATION

In this section, the network trained for road marking segmentation is described in detail, along with some important considerations that have to be taken into account when rebalancing datasets.

A. Network Architecture

Deep networks for road marking segmentation have several advantages over traditional heuristic or shallow-learning pipelines. Firstly, they are more robust to spatial deformations, degradation, and partial occlusion. Secondly, the scene context can be leveraged to improve semantic segmentation and thereby understand the road rules. For instance, similarly-shaped road markings (e.g. lane separators and separators that mark a parking spot) can be classified differently based on their place in the scene and relationship with other objects, whereas this is difficult to accomplish with traditional rule-based systems.

We train a U-Net model [34], but include batch normalization and dropout as regularization techniques. These are paramount in our framework, since we train on partial labels that are generated automatically. Dropout allows the network to extend its prediction towards road marking pixels that were wrongly assigned to the background in the partial labels, because they share more similarities with the road marking class than the background class. The architecture and training settings used are similar to our previous work [19], with the major exception that the output now predicts multiple classes of road markings instead of a binary segmentation. More specifically, the output of the network is computed by applying a channel-wise softmax activation over the final feature maps and assigning a class to each respective pixel by taking the channel-wise $\arg\max$ over the output channels, yielding a one-channel discrete class activation map.

At run time, the Tensorflow implementation of the network performs inference on an input image in real-time (~ 62.5 Hz) on an NVIDIA TITAN Xp GPU.

B. Balancing of the Classes

As mentioned before, datasets collected during regular driving are extremely imbalanced in terms of the occurrences of particular road marking classes. For instance, zigzag markings are only found in $\sim 7\%$ of the images, whereas lane separators occur in $\sim 70\%$. Solutions such as resampling the dataset or applying a class-weighted loss function are not viable for small, hand-labelled datasets, because they simply contain an insufficient number of examples of the rare classes to generalize well to unseen cases during deployment.

In this paper, we opt for a different approach in which we synthesize new training pairs for rare classes automatically and add them to an existing dataset. This ensures that there are enough examples of these classes for the network to learn from. However, it is not obvious how to produce a rebalanced dataset including synthetic training pairs that is optimal for

training. To counteract the fact that we might add too many synthesized training pairs, we experiment with three types of class-weighted cross-entropy losses:

- 1) Equal weighting (EQ) of all classes irrespective of their occurrence in the dataset.
- 2) Median frequency balancing (FB) [35], in which each pixel is weighted by

$$w_c = \frac{\text{median}(F)}{f_c}, \quad (4)$$

where $F = \{f_1, \dots, f_C\}$ with f_c denoting the total number of pixels of class c divided by the total number of pixels in labels where c is present and C the total number of classes.

- 3) Median total balancing (TB), in which each pixel is weighted by

$$w_c = \frac{\text{median}(G)}{f_c + n_c}, \quad (5)$$

where $G = \{f_1 + n_1, \dots, f_C + n_C\}$ with f_c equivalent to 2) and n_c denoting the number of labels in which class c is present divided by the total number of training pairs.

It is important to note that median frequency balancing only corrects for the fact that some classes naturally occupy less pixels in the images. For instance, dotted lines indicating a pedestrian crossing are smaller in accumulated area than an alternative zebra crossing. However, median frequency balancing does not account for imbalance in occurrences across the dataset; whether $\sim 7\%$ of the images contain zigzag markings or $\sim 70\%$, the weight remains the same as long as their pixel size remains equivalent. This is not ideal, since we artificially create an imbalance in the number of occurrences by adding labels of specific classes. The third weighting function, introduced in this paper, is designed to take this into account, balancing the average pixel area as well as the imbalance in occurrences across the dataset.

V. EXPERIMENTAL RESULTS

In this section we describe the experimental setup and the datasets that we have created, before we present the quantitative and qualitative results.

A. Experimental Setup

We have selected four types of rare road markings for evaluation: bus stops, diagonal stripes (must not enter), small warning triangles, and zigzag markings. These classes function as a proof of concept, but the framework can be applied to any class (i.e. model) of road markings. For quantitative pixel-wise evaluation, we have hand-labelled 102, 102, 96, and 102 *real-world* images containing bus stops, diagonal structures, small warning triangles, and zigzag markings, respectively. Note that in these images only these respective classes were labelled and all other classes present were ignored (see Fig. 7). While we train all models to predict the *full* set of 20 different road markings and show these results qualitatively, we only evaluate the four selected classes quantitatively. We define the pixel-wise metrics $\text{PRE} = \frac{\text{TP}}{\text{TP} + \text{FP}}$, $\text{REC} = \frac{\text{TP}}{\text{TP} + \text{FN}}$, $\text{F}_1 = 2 * \frac{\text{PRE} * \text{REC}}{\text{PRE} + \text{REC}}$, and

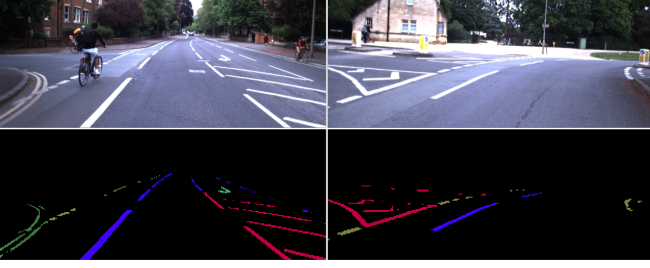


Fig. 5. Examples of the partial labels created by semantically classifying the binary annotations of [19]. Although not perfect ground-truth, these labels can be used to train a baseline model to predict the full set of road markings.

$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$ with TP, FP, and FN denoting the true positive, false positive, and false negative pixels, respectively. In contrast to binary classification, all metrics are evaluated at the operating point defined by taking the channel-wise $\arg \max$ over the multi-class output on a per image basis and averaged over the test set, without any further fine tuning of the operating characteristics. Furthermore, we have hand-labelled 25 real-world images for each respective class for validation. We train until convergence and select the epoch for testing in which the mIoU is highest among the evaluations on the validation set. It should be noted that road marking segmentation is arguably a harder task than scene segmentation, because road marking elements are fairly small in general, often degraded, and the different types share many visual and geometric similarities. State-of-the-art approaches achieve a mIoU of around 40%, however a benchmark has only been established recently [17].

As a reasonable baseline, 1000 partial, binary labels generated by [19] collected during regular driving were hand-labelled class-wise. Although not equivalent to the ground-truth, we have proven in [19] and will demonstrate again in Section V-C that these labels are sufficient to achieve full segmentation, when regularization techniques are applied. A few examples are given in Fig. 5. The labels contain the 20 different types of road markings, so that the network functions as a full road marking segmentation system. However, many classes occur too infrequently to achieve state-of-the-art performance, because the network fails to generalize to new scenarios during deployment. For instance, the baseline dataset only contains 63, 109, 39, and 74 images with bus stops, diagonal stripes, small warning triangles, and zigzag markings, respectively. For the other experiments, we add synthetic training pairs of the four classes to the baseline dataset. In this way, the network still predicts all 20 classes, but is given a sufficient number of labels of the rare classes to improve generalization during real-world deployment.

B. Quantitative Evaluation

In order to understand how the number of added synthetic images influences the performance, we have added different numbers of synthetic zigzag pairs to the baseline dataset, while keeping the other classes constant. The results for the three different cross-entropy losses are presented in Fig. 6.

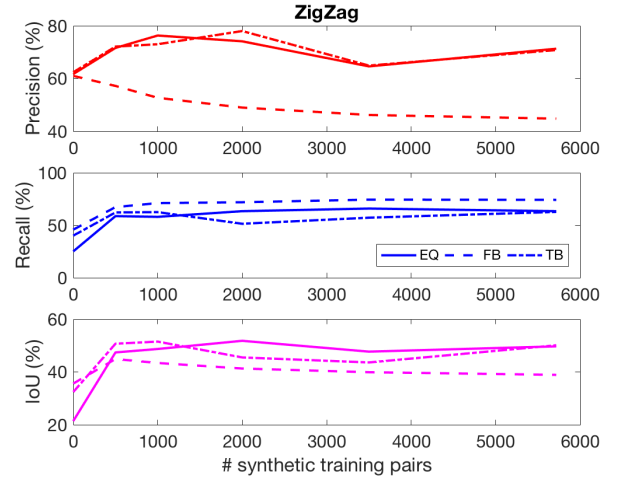


Fig. 6. Pixel-wise performance of zigzag segmentation when training with different cross-entropy losses for a variable number of synthetic images added to the baseline dataset.

The following key observations can be made:

- Adding as little as 500 synthetic training pairs already makes a substantial difference in terms of overall performance.
- Adding more than 2000 synthetic training pairs does not provide extra benefits in general. Further performance increase beyond this level might require higher-quality, more diverse, more coherent synthetic images.
- FB struggles to balance training as more synthetic pairs are added, due to the fact that it does not account for occurrence imbalance across the dataset. The precision drops significantly as the network learns from an abundance of zigzag markings and starts classifying other classes incorrectly as zigzag.
- TB alleviates the precision drop of FB, but does not outperform EQ consistently among all metrics.

Assuming that these observations hold similarly for the other classes, 1000 synthetic training pairs of each respective class were added to the baseline dataset as a proof of concept to train enhanced networks (with the different loss functions). From the results, as presented in Table I, the following key observations can be made:

- By adding synthetic training pairs, IoU performance similar to the state-of-the-art can be achieved when only very few real-world examples are available. mIoU is increased by 12.4% (comparing the best baseline and enhanced models) without using any manual labelling effort.
- The enhanced networks always achieve better overall performance (i.e IoU) by a substantial margin for the equivalent cost function. Segmentation performance can thus be boosted cheaply by the presented framework.
- TB outperforms FB in terms of F_1 and IoU in general, because it accounts for the class imbalance across the dataset that was artificially created by adding synthetic pairs. TB offers a good trade-off between high precision achieved by EQ and high recall achieved by FB.

TABLE I
PIXEL-WISE PERFORMANCE FOR RARE CLASSES FOR THE BASELINE (B) AND ENHANCED (E) MODELS

Model	Loss	BUS STOP				DIAGONAL				TRIANGLE				ZIGZAG				MEAN			
		PRE	REC	F ₁	IoU	PRE	REC	F ₁	IoU	PRE	REC	F ₁	IoU	PRE	REC	F ₁	IoU	PRE	REC	F ₁	mIoU
B	EQ	61.6	17.8	27.6	16.1	59.1	24.6	34.7	21.8	60.7	41.1	49.0	34.4	65.9	22.5	33.6	20.1	61.8	26.5	36.2	23.1
B	FB	64.7	26.3	37.3	22.8	58.8	31.0	40.6	26.4	60.1	47.7	53.2	36.9	64.7	34.8	45.3	29.5	62.1	35.0	44.1	28.9
B	TB	62.2	19.1	29.2	17.1	58.4	33.3	42.4	27.7	59.9	51.6	53.4	39.4	62.3	31.3	41.7	26.5	60.7	33.8	42.2	27.7
E	EQ	74.8	28.5	41.2	26.3	73.0	40.9	52.4	35.8	61.4	46.9	53.2	38.4	69.4	49.8	58.0	40.7	69.7	41.5	51.2	35.3
E	FB	54.8	62.9	58.6	40.1	45.8	67.4	54.6	39.1	46.9	73.8	57.3	37.9	51.0	66.6	57.8	40.3	49.6	67.7	57.1	39.4
E	TB	58.5	55.3	56.8	39.2	51.4	59.1	55.0	40.2	50.8	75.1	60.6	43.4	61.4	57.3	59.3	42.5	55.5	61.7	57.9	41.3

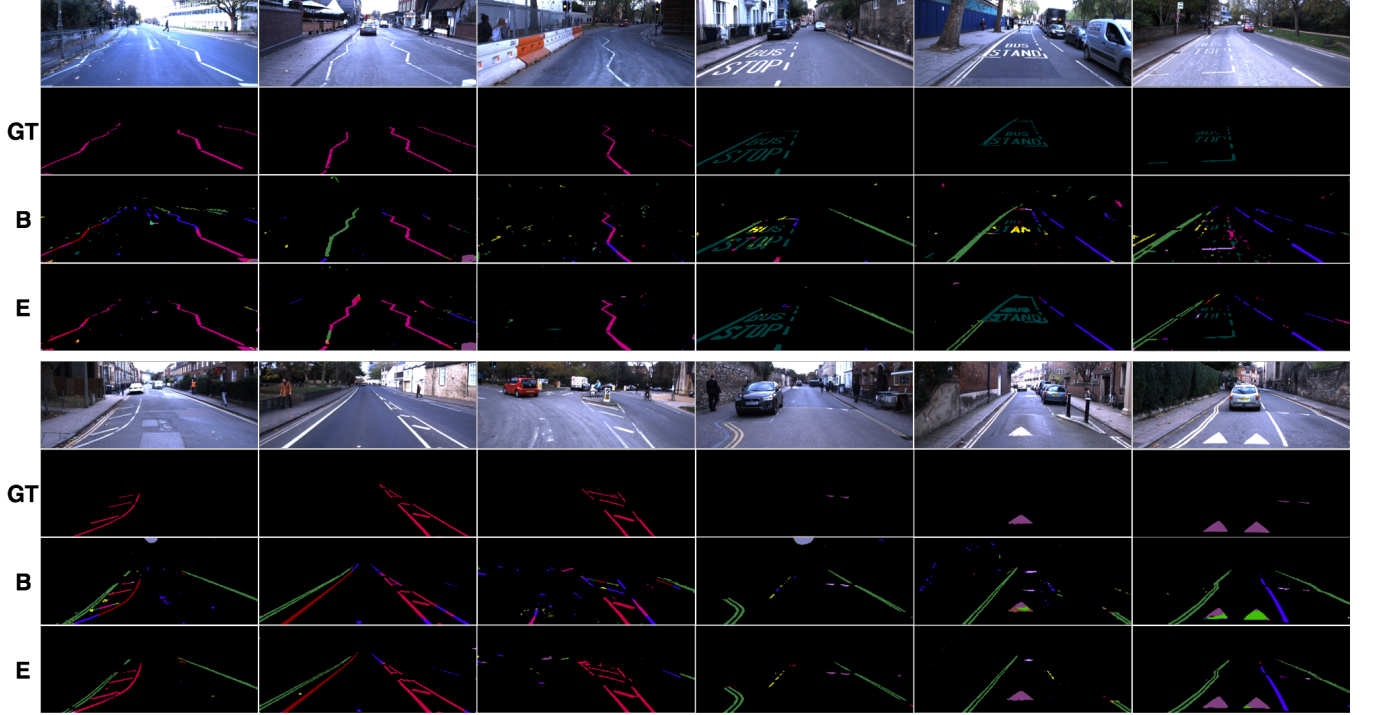


Fig. 7. Road marking segmentation (full set of classes) in traffic environments with rare classes. The *top two* rows of each scene show the input image together with the corresponding ground-truth (GT) label of the rare class, which is used for quantitative evaluation. The *bottom two* rows of each scene depict the segmentation results for the best performing baseline (B) and enhanced model (E), respectively. The enhanced model provides more consistent and correct segmentation of the rare classes, while retaining reasonable and sometimes achieving improved performance for other classes (e.g. *green* double boundaries, *blue* separators, *yellow* parking spot separators, etc.).

C. Qualitative Evaluation

In Fig. 7, the best baseline and enhanced models are compared qualitatively for different traffic scenes. All networks are trained to predict the full set of 20 different road marking classes, however scenes with the respective rare classes are selected for visualization.

It is clear that adding synthetic images to the training set results in more consistent and correct segmentation of the rare classes, while retaining reasonable and sometimes achieving improved performance for other classes. The latter could be caused by the general increase of the number of training examples and/or better balancing of the cost function. The enhanced model trained with TB offers more satisfying (i.e. less noisy) visual results than the baseline model trained with FB. Furthermore, it is clear that full segmentation of the road marking elements is possible from partial labels when regularization techniques are applied correctly. Thus, this framework offers an effective and ef-

ficient step towards a road marking classification system for automated driving pipelines.

VI. CONCLUSION

We have presented a weakly-supervised approach for improving road marking segmentation in complex urban environments. To this end, we alter semantic labels of real-world scenes with instances of chosen road markings using domain randomization principles and synthesized corresponding, photo-realistic images to generate vast quantities of synthetic training pairs, thereby avoiding the need for expensive manual labelling. During deployment, we predict 20 classes of road markings in real time and we have demonstrated quantitatively that this framework improves mIoU of rare classes by more than 12 percentage points and thus reaches state-of-the-art performance with very few real-world labels. This is achieved by introducing a new class-weighted cross-entropy loss to balance the training of

synthetic datasets. Furthermore, we have shown qualitatively that the segmentation performance for other classes is retained. The presented framework can easily be extended to include other classes or work under different conditions and results can be expected to improve as more advanced synthesizing networks will emerge in the future. Hence, road layout randomization is an effective and efficient technique to enhance road marking classification systems in automated driving pipelines.

ACKNOWLEDGMENT

The work has been supported by the EPSRC/UK Research and Innovation Programme Grant EP/M019918/1 (Mobile Autonomy: Enabling a Pervasive Technology of the Future). We acknowledge the support of NVIDIA Corporation with the donation of Titan Xp and Titan V GPUs.

REFERENCES

- [1] L. Kunze, T. Bruls, T. Suleymanov, and P. Newman, "Reading between the lanes: Road layout reconstruction from partially segmented scenes," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Nov 2018, pp. 401–408.
- [2] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.
- [3] R. Krajewski, T. Moers, and L. Eckstein, "VeGAN: Using GANs for augmentation in latent space to improve the semantic segmentation of vehicles in images from an aerial perspective," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Jan 2019, pp. 1440–1448.
- [4] S. Lee, J. Kim, J. S. Yoon, S. Shin, O. Bailo, N. Kim, T. Lee, H. S. Hong, S. Han, and I. S. Kweon, "VPGNet: Vanishing point guided network for lane and road marking detection and recognition," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 1965–1973.
- [5] Y. Chen, W. Li, X. Chen, and L. Van Gool, "Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach," *arXiv preprint arXiv:1812.05040*, 2018.
- [6] A. Dundar, M.-Y. Liu, T.-C. Wang, J. Zedlewski, and J. Kautz, "Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation," *arXiv preprint arXiv:1807.09384*, 2018.
- [7] R. Cura, J. Perret, and N. Paparoditis, "Streetgen: In base city scale procedural generation of streets: road network, road surface and street objects," *arXiv preprint arXiv:1801.05741*, 2018.
- [8] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 8798–8807.
- [9] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," *arXiv preprint arXiv:1903.07291*, 2019.
- [10] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Bochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2018, pp. 1082–1088.
- [11] D. Lee, S. Liu, J. Gu, M.-Y. Liu, M.-H. Yang, and J. Kautz, "Context-aware synthesis and placement of object instances," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 414–10 424.
- [12] A. Prakash, S. Bochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapira, and S. Birchfield, "Structured domain randomization: Bridging the reality gap by context-aware synthetic data," *arXiv preprint arXiv:1810.10093*, 2018.
- [13] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2017, pp. 23–30.
- [14] B. De Brabandere, W. Van Gansbeke, D. Neven, M. Proesmans, and L. Van Gool, "End-to-end lane detection through differentiable least-squares fitting," *arXiv preprint arXiv:1902.00293*, 2019.
- [15] N. Garnett, R. Cohen, T. Pe'er, R. Lahav, and D. Levi, "3D-LaneNet: end-to-end 3D multiple lane detection," *arXiv preprint arXiv:1811.10203*, 2018.
- [16] M. Ghafoorian, C. Nugteren, N. Baka, O. Booi, and M. Hofmann, "EL-GAN: Embedding loss driven generative adversarial networks for lane detection," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 256–272.
- [17] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The ApolloScape dataset for autonomous driving," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2018, pp. 1067–10676.
- [18] T. M. Hoang, P. H. Nguyen, N. Q. Truong, Y. W. Lee, and K. R. Park, "Deep retinanet-based detection and classification of road markings by visible light camera sensors," *Sensors*, vol. 19, no. 2, 2019.
- [19] T. Bruls, W. Maddern, A. A. Morye, and P. Newman, "Mark yourself: Road marking segmentation via weakly-supervised annotations from multimodal data," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 1863–1870.
- [20] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtualworlds as proxy for multi-object tracking analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4340–4349.
- [21] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 746–753.
- [22] S. Liu, J. Zhang, Y. Chen, Y. Liu, Z. Qin, and T. Wan, "Pixel level data augmentation for semantic image segmentation using generative adversarial networks," *arXiv preprint arXiv:1811.00174*, 2018.
- [23] K. Li, T. Zhang, and J. Malik, "Diverse image synthesis from semantic layouts via conditional IMLE," *arXiv preprint arXiv:1811.12373*, 2018.
- [24] Q. Geng, F. Lu, X. Huang, S. Wang, X. Cheng, Z. Zhou, and R. Yang, "Part-level car parsing and reconstruction from single street view," *arXiv preprint arXiv:1811.10837*, 2018.
- [25] H. A. Alhajja, S. K. Mustikovele, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 961–972, 2018.
- [26] H. A. Alhajja, S. K. Mustikovele, A. Geiger, and C. Rother, "Geometric image synthesis," *arXiv preprint arXiv:1809.04696*, 2018.
- [27] R. Khrodgar, D. Yoo, and K. M. Kitani, "VADRA: Visual adversarial domain randomization and augmentation," *arXiv preprint arXiv:1812.00491*, 2018.
- [28] J. Fang, F. Yan, T. Zhao, F. Zhang, D. Zhou, R. Yang, Y. Ma, and L. Wang, "Simulating LiDAR point cloud for autonomous driving using real-world scenes and traffic flows," *arXiv preprint arXiv:1811.07112*, 2018.
- [29] W. Li, C. Pan, R. Zhang, J. Ren, Y. Ma, J. Fang, F. Yan, Q. Geng, X. Huang, H. Gong *et al.*, "AADS: Augmented autonomous driving simulation using data-driven algorithms," *arXiv preprint arXiv:1901.07849*, 2019.
- [30] D. J. Fremont, X. Yue, T. Dreossi, S. Ghosh, A. L. Sangiovanni-Vincentelli, and S. A. Seshia, "Scenic: Language-based scene generation," *CoRR*, vol. abs/1809.09310, 2018.
- [31] Z. Liu, M. Shen, J. Zhang, S. Liu, H. Blasinski, T. Lian, and B. Wandell, "A system for generating complex physically accurate sensor images for automotive applications," *arXiv e-prints*, p. arXiv:1902.04258, Feb 2019.
- [32] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford Robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.
- [33] H. Porav, W. Maddern, and P. Newman, "Adversarial training for adverse conditions: Robust metric localisation using appearance transfer," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 1011–1018.
- [34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.
- [35] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 2650–2658.