Superpixel clustering with deep features for unsupervised road segmentation

Shunta Saito*
Preferred Networks, Inc.
shunta@preferred.jp

Tommi Kerola*
Preferred Networks, Inc.
tommi@preferred.jp

Satoshi Tsutsui*, †
Indiana University
stsutsui@indiana.edu

Abstract

Vision-based autonomous driving requires classifying each pixel as corresponding to road or not, which can be addressed using semantic segmentation. Semantic segmentation works well when used with a fully supervised model, but in practice, the required work of creating pixel-wise annotations is very expensive. Although weakly supervised segmentation addresses this issue, most methods are not designed for road segmentation. In this paper, we propose a novel approach to road segmentation that eliminates manual annotation and effectively makes use of road-specific cues. Our method has better performance than other weakly supervised methods and achieves 98% of the performance of a fully supervised method, showing the feasibility of road segmentation for autonomous driving without tedious and costly manual annotation.

1. Introduction

Autonomous driving requires various components including vision modules to better understand the environment around the vehicle. At the very minimum, for vision-based driving, a monocular camera facing the front is required to be mounted on the vehicle. Given a road image, the system needs to perceive the objects in the road. For most of objects that appear in traffic, such as vehicles, pedestrians, and traffic signs it is enough for a practical system to localize them using bounding boxes (*i.e*. leveraging object detection). However, bounding boxes are not suitable for detecting the region where a car is able to drive safely. For the road, the system needs to identity the drivable region in a precise pixel-wise manner. For this reason, we focus on the problem of classifying each pixel into road or not.

Semantic segmentation is the problem of classifying each pixel into semantic classes (*e.g*. road, sky, vehicle). Recent advances in semantic segmentation [31, 49, 29, 5, 8] include fully supervised methods that employ convolutional neural networks (CNNs). After a breakthrough in image classification in 2012, CNNs have become fundamental tools for learning-based computer vision [26]. CNNs can be used for segmentation by performing pixel-wise classification. This means that annotations are also required in a pixel-wise manner in order to be able to perform training. Unfortunately, pixel-wise annotations are much more costly than annotations for image classification. For example, while assigning a class to an image only takes a second, masking an object in an image takes approximately 80 seconds [6]. In addition, in the established Cityscapes dataset [11], which contains 2975 pixel-wise annotated images in the training set, it took 1.5 hours per image to annotate all road-related objects in an image including quality control. Being able to train segmentation models without the required annotation work would greatly accelerate scientific progress.

Reducing the expensive annotation cost for semantic segmentation is an active area of research and is often called weakly supervised segmentation [34, 24, 39, 36, 12, 46, 28, 23, 41, 21]. These methods are usually not designed for road segmentation, and mainly focus on foreground classes (*e.g*. humans, cars, and cats) and regard the background objects (*e.g*. sky, building, and road) as a uniform class [35]. For autonomous vehicles, however, it is essential to segment parts of the background, such as the road, which makes these methods unusable in our problem setting.

In this paper, we address the problem of road segmentation from *car-centric* images without using pixel-wise annotations. We define a car-centric image to be an image taken by a monocular front-facing camera mounted on a vehicle. Car-centric images are cheap to collect, and can be easily gathered by driving with a vehicle as usual. We distinguish our work from most other weakly supervised methods in that we focus on road, which is background segmentation. While weakly supervised methods that focus on segmenting foreground objects rely on reformulating segmentation as an image-level classification task [12], this will not work for weak supervision from car-centric images, as all the images contain road. Previous work overcame the issue

*All authors contributed equally. The order was decided by setting the paper ID as random seed and then calling `np.random.permutation`.
†Part of the work was done while the author was an intern at Preferred Networks, Inc.

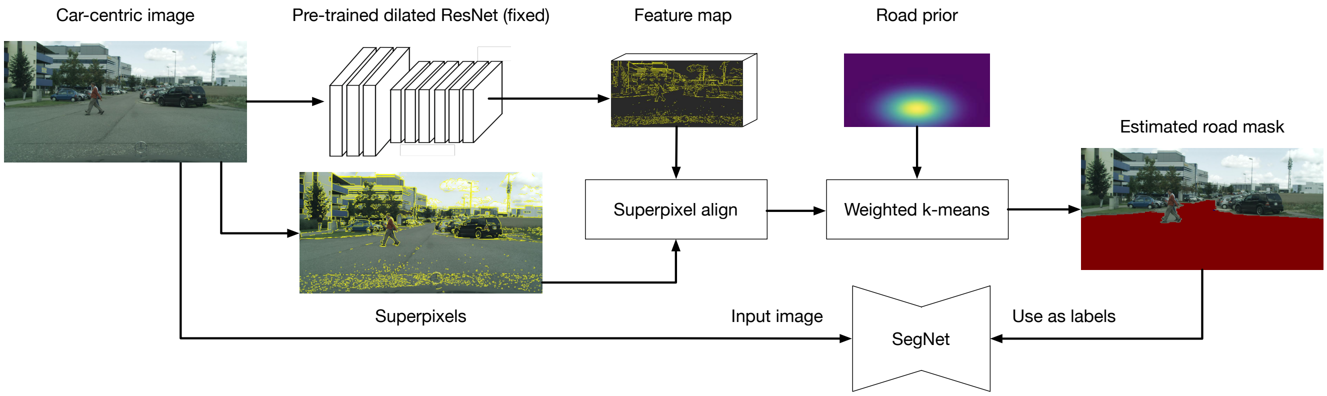

Figure 1. Overview of our method. We extract features from a dilated ResNet pre-trained on ImageNet and leverage a novel superpixel align method and road prior clustering to extract weak labels for training a segmentation CNN. Our method requires no manual annotation of road labels for semantic segmentation. Best viewed in color.

by collecting additional images [44, 35], which brings the cost of additional data collection. In contrast, we make use of the fact that images are car-centric images, and assume prior knowledge about the location of the road. Given these assumptions, we can eliminate the need for any annotation and only need features from a CNN pre-trained on ImageNet (§3.1), which is reported to have built-in priors for semantic segmentation [36]. We then propose *superpixel align*, a novel aligning method for boundary cohesive superpixels that uses bilinear interpolation and random sampling to create a feature for each superpixel (§3.2). We further propose a custom superpixel clustering method that makes use of the road prior knowledge and automatically selects the cluster corresponding to the road (§3.3). The road clusters are then regarded as labels used to train a CNN for segmentation (§3.4). An overview of our method can be seen in Fig. 1. Our experiments on the Cityscapes dataset [11] show that our approach is superior to other weakly supervised methods for road segmentation and achieves 98% of the performance of a fully-supervised approach, while requiring no annotation work at all (§4). The implementation of our method will be made publicly available.

In summary, we make the following contributions:

- We propose an approach for road segmentation for autonomous driving that does not require any annotation other than having car-centric images of road, and demonstrate that CNN built-in priors are enough for extracting the road region.
- We propose a novel algorithm for combining CNN feature maps and superpixels, and a clustering method that incorporates the prior knowledge about the road.
- We empirically show the performance of our approach and compare with fully supervised and other weakly supervised methods. Our approach uses no supervision but has better performance than other weakly supervised methods.

2. Related Work

Fully supervised segmentation Recent advances in semantic segmentation are built upon fully convolutional networks (FCNs) [31], which extended CNNs designed for image classification by posing semantic segmentation as a dense pixel-wise classification problem.

A problem caused by CNN pooling layers is that they discard spatial information that is critical for image segmentation. To address this problem, an approach is to use dilated convolutions [47], which allows receptive field expansion without pooling layers. Dilated convolutions (or atrous convolutions) are incorporated into recent frameworks such as DeepLab [8] and PSPNet [49]. Although this paper does not focus on engineering CNN architectures, this direction inspired our choice of a CNN for image feature extraction, since we want to obtain a high resolution feature map. We use dilated ResNet [48] trained on the ImageNet classification task, which yields a higher resolution feature map than the normal ResNet [20].

In this paper, we use the established SegNet CNN architecture [5] as our segmentation model, although this particular choice is not crucial and other modern FCNs could be leveraged as well.

Weakly supervised segmentation Annotations for segmentation are very costly so weakly supervised segmentation utilizes less costly annotations such as image tags [34, 24, 39, 36, 12, 46], scribbles [28], bounding boxes [23], or videos capturing objects [41, 21]. Although most previous studies only focus on foreground objects, our work can be regarded as a tag-based weakly supervised method in that we assume all images having a tag of road. However, previous work extract segmentation masks from CNNs for tag-based classification, which is not directly applicable as we cannot train a tag classifier without collecting additional images with a different tag than road. Technically, some methods propose a new CNN architecture [12] or a better loss function [24], but others focus on automatically generating segmentation masks for training available CNNs [23]. Our focus is also generating segmentation masks for CNNs. We also do not use the approach of gradually refining the segmentation mask [39], because we believe it is important to output a trained CNN that can be deployed on an autonomous car in practice.

Road segmentation Road segmentation (often called free space estimation) is the task of estimating the space where a vehicle can drive safely without collision. This task is critical for autonomous driving and has traditionally addressed by geometric modeling [25, 2, 4, 45], handcrafted features [3, 17] or even a patch-based CNN [1]. We leverage FCNs in this paper, which Oliveira *et al*. [33] demonstrated to be efficient for road segmentation. However, as pixel-wise annotations are expensive, weakly supervised methods have been developed also in the context of road segmentation.

There are several studies about weakly supervised road segmentation. While an early study [16] trains a probabilistic model, other studies train FCNs [44, 35, 27, 37]. Saleh *et al*. [35] develop a video segmentation algorithm for general background segmentation including road. Tsutsui *et al*. [44] propose distantly supervised monocular image segmentation for road. However, both methods require additional images to train a saliency or attention extractor. Laddha *et al*. [27] use external maps of the road using the corresponding vehicle position from GPS. Sanberg *et al*. [37] and Guo *et al*. [16] use stereo information for automatically generating road masks. We distinguish our work from the previous studies in that we only use a collection of monocular car-centric images, which makes our approach the least supervised method among others.

While our approach requires a pretrained CNN as a generic feature extractor, the ImageNet creation cost has been ignored in previous work when comparing annotation cost [6] as it is a standard practice to use it as a generic off-the-shelf feature extractor. Moreover, we note that the ImageNet classification task does not include the road class, so we do not require the pre-trained CNNs to be trained for the task of road classification; our assumption is that the CNN features are generic enough for being able to capture the semantic meaning of the road compared to other parts of the image.

In terms of evaluation, two major datasets KITTI [15] and CamVid [7] have been used before, but these are not large enough to leverage the power of CNNs. Recently, a larger dataset called Cityscapes [11] was proposed for the object segmentation task in autonomous driving. Reportedly [11], CNNs trained on Cityscapes perform better on the test set of KITTI and CamVid than each corresponding state-of-the-art method. Hence, we will conduct our experiments in this paper on Cityscapes.

3. Method

This section explains how we automatically obtain weak labels for training a CNN for road segmentation. Since we do not have any annotations, we turn to unsupervised learning, and use features extracted from CNNs trained for the ImageNet Visual Recognition Challenge to represent the input image (§3.1). Even though the ImageNet challenge does not include road classification explicitly, we believe these features are generic enough. In fact, these features have been effectively applied for the domains outside ImageNet such as document image analysis [18] or medical image analysis [13]. Moreover, ImageNet pretrained CNNs are reported to have built-in priors for semantic segmentation [36].

Raw CNN features are, however, still too coarse to be able to localize the road precisely. To handle this issue, we develop a novel superpixel align method that creates aggregated CNN features for a set of superpixels (§ 3.2). In order to extract the road information, we use a simple unsupervised learning technique: clustering. In other words, we perform clustering on the resulting superpixel features. A challenge is how to select the cluster corresponding to the road, but it can be automatically selected assuming prior knowledge about the road (§3.3). In this way, we can automatically obtain a weak mask covering the road region, which is used as training data for CNNs for segmentation (§3.4).

3.1. CNN Features

Our inspiration of using features from a pretrained CNN for capturing semantics is inspired by Saleh *et al*. [36]. As is well-known, earlier layers in convolutional neural networks tend to learn lower level features (*e.g*. edges), while later layers capture an increasing amount of semantic information until finally resulting in features suitable for the explicit classification problem (*e.g*. object types like cars) that the network was trained to solve.

Similar to previous work [36], we visually inspected some features from a dilated ResNet [48] with 26 layers trained for the task of ImageNet classification. Unlike using the VGG [40] architecture as in their work, we choose to leverage a dilated network architecture in order to be able to get higher resolution feature maps, which is important for being able to localize the road more finely. We found that earlier layers indeed convey information like edges, and later layers tend to highlight the semantic objects in the scene, and therefore propose to extract the features from the last convolutional layer in the network to convey information about the road.

3.2. Superpixel Align

While the dilated ResNet features capture semantic information for localizing the road, it does not cover the ma-

function ROADPRIORKMEANS(k, $\{S_i\}_{\forall i}$, μ_{prior}, σ_{prior})
  $\forall i, \forall S_{xy} \in S_i : p_{xy} = \text{spatial\_coord}(S_{xy})$
  $\forall i, \forall S_i : w_i = \frac{1}{|S_i|} \sum_{S_{xy} \in S_i} e^{-\|p_{xy} - \mu_{\text{prior}}\|^2 / 2\sigma^2_{\text{prior}}}$
  $\forall i : m_i \leftarrow \begin{cases} 0 & \text{if } w_i > \text{median}_j \, w_j \\ \text{unif}(1, k-1) & \text{otherwise} \end{cases}$
  while not converged **do**
    $c_0 \leftarrow \sum_{m_i = 0} w_i S_i / Z_0$
    $\forall q > 0 : c_q \leftarrow \sum_{m_i = q} (1 - w_i) S_i / Z_q$
    $\forall i : m_i \leftarrow \text{argmin}_q \|S_i - c_q\|_2^2$
  end while
  return $\{c_0, \ldots, c_{k-1}\}$ ▷ Cluster centers
end function

Figure 2. Road prior k-means. Here, $Z_q = \sum_{m_i = q} w_i$.

Figure 3. Example of using road prior k-means on raw dilated ResNet (CNN) or superpixel align features to find the location of the road. Note that due to the design of our algorithm, the road will typically be located in the first cluster (red), which means that we do not need any sophisticated post-processing in order to find the road cluster. Best viewed in color.

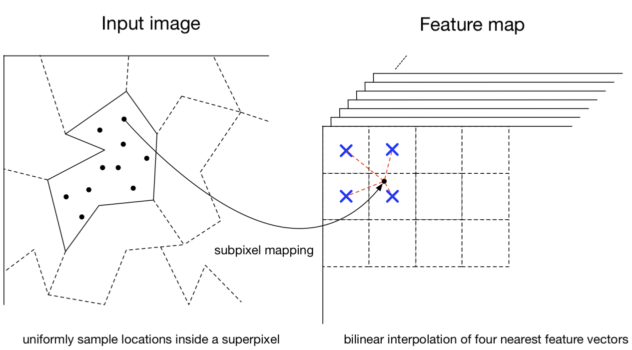

Figure 4. Superpixel align. The feature vector for each superpixel is defined by random sampling of 10 points inside the superpixel, and taking the average of their bilinearly interpolated feature maps.

jor part of the actual road area, due to certain parts of the road being more semantically important for localizing it than other locations. Therefore, we turn to an edge-based superpixel method [14] in order to create superpixels that better match the visual shape of the road. We note that using superpixels is a common practice in other weakly supervised methods [44, 28, 41, 21] and road segmentation methods [9].

Inspired by RoIAlign [19], to improve the localization of the dilated ResNet features, we propose to apply bilinear interpolation of the raw feature maps to a random subset of the pixels inside each superpixel. While the raw dilated ResNet features capture important semantic information, they are not well localized, and it can thus be expected that aligning these to the spatially coherent superpixels will lead to a better representation of the scene. We do bilinear interpolation [22] of dilated ResNet features F^c_{mn} at spatial location (m, n) and channel c as

$$S^c_{xy} = \sum_{(m,n) \in \mathcal{N}_{xy}} F^c_{nm} \max(0, \hat{\Delta}_{xm}) \max(0, \hat{\Delta}_{yn}) \, , \quad (1)$$

where $\hat{\Delta}_{xm} = 1 - |x - m|$ and $\mathcal{N}_{xy}$ is the set of neighbors for spatial location (x, y) in superpixel S. We sample 10 locations uniformly at random inside the superpixel, and then use the four nearest neighbors of each selected pixel for the bilinear interpolation. Finally, we aggregate the features inside each superpixel using average pooling. Note that unlike RoIAlign, we assume each superpixel to consist of a homogeneous set of pixels and therefore avoid computing the bilinear interpolation densely for all pixels by using a small randomly sampled set of pixels to describe the superpixel, which we have found works well in practice. To improve the spatial cohesiveness of the feature, we propose to also append the center of mass spatial coordinates of the superpixel to the pooled feature vector. This gives us one image feature for each superpixel. The procedure for superpixel align is shown in Fig. 4.

3.3. Feature Clustering

Automatic cluster selection by road prior k-means When applying a clustering algorithm to the extracted image features, we are faced with the problem of determining which cluster corresponds to the road. One simple way would be to select the largest cluster at the bottom half of the image, although this would fail in crowded scenes with a vast amount of foreground objects on the road.

Instead, inspired by previous work [43] in bioinformatics, we wish to utilize prior information about the spatial location of the road in order to find a good cluster that represents the road. For this, we adapt Lloyd's algorithm [30] for solving a weighted variant of the k-means clustering problem. We assume that we have a prior for the location of the road, that ought to be true for most scenes based on human intuition, such as the pixels right above the visible hull of the ego-car. We assume that this prior follows a mixture of Gaussians distribution, that is, each superpixel S_i has a prior weight w_i that is parameterized by μ_{prior} and σ_{prior} and the spatial coordinates p_{xy} of each pixel inside the superpixel.

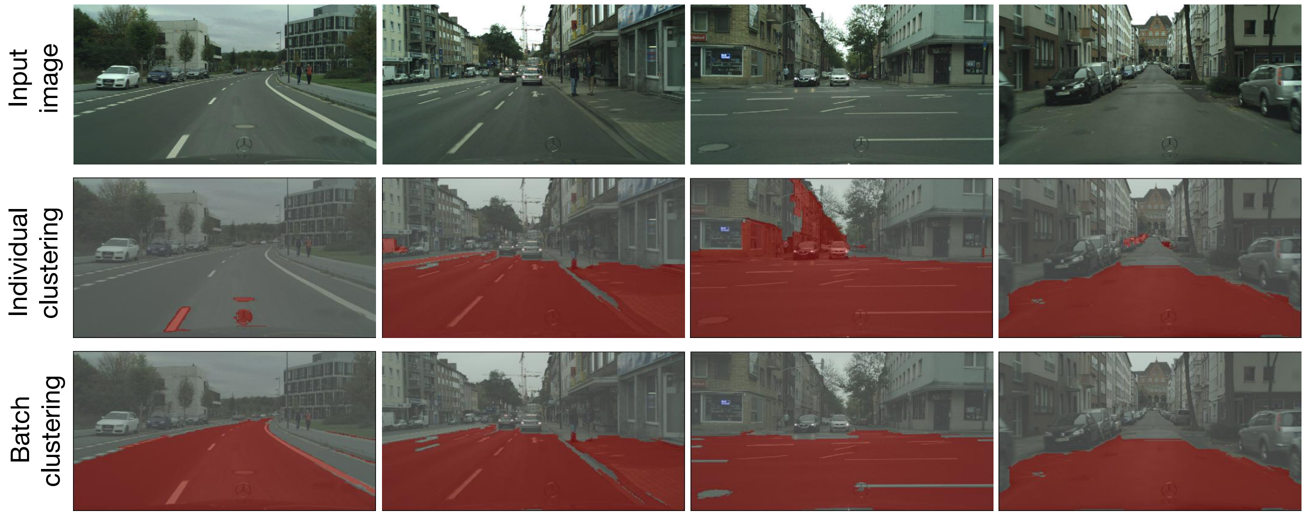

Figure 5. Batch clustering. (top) Input images. (middle) Results of clustering each image individually. (bottom) Results of clustering the four images together in a batch manner. The left image has a spot around the center of the road prior so only the spot is recognized as road when clustered individually, but batch clustering can avoid this mistake.

Subsequently, we initialize half of the feature vectors to the road cluster (which we assume is the first cluster) based on these weights. The first cluster is then encouraged to consist of features corresponding to road by setting its cluster center to be the spatially weighted average of features assigned to it. The other clusters have a repellent weight assigned to their members to encourage them to spatially spread away from the road prior location. Cluster memberships are updated in the same manner as the standard k-means algorithm without taking the weights into account. Our algorithm is summarized in Fig. 2. An example of the output of the algorithm is shown in Fig. 3. Note in the figure how the superpixel align features are much more spatially cohesive than the raw CNN features.

Although our cluster update breaks the convergence criterion of standard k-means clustering [10], we have found that our algorithm converges to a stable solution in practice. We note that similar prior information could also be incorporated into other types of clustering algorithms, such as spectral clustering [38], although we do not explore this direction in this paper for brevity.

Batch image clustering While clustering could be performed on each image individually, we avoid this and conduct joint clustering on a batch of images. The reason is because while we expect that our road prior corresponds to an area with road in most images, for some images inevitably this assumption will not hold true. For example, there could be a vehicle or pedestrian located on the center of the road prior, which would then wrongly assign the first cluster to consist of features corresponding to non-road locations. In Fig. 5, we show an example where only a single spot at the center of the road prior is recognized as a road, but clustering with three other images prevents this mistake. As our experiments will show, batch clustering is effective for increasing the quality of the resulting weak labels.

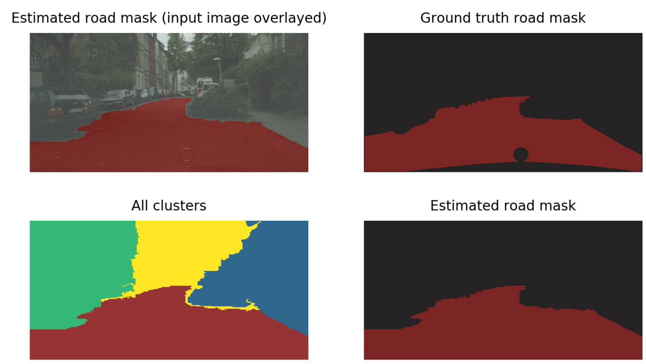

Figure 6. Example of superpixel align compared to ground truth.

3.4. CNN Training from Weak Labels

At this point we have a weak label for road gotten by road prior clustering of superpixel align features (an example can be seen in Fig. 6). We use this automatically generated road mask to train a road segmentation CNN.

4. Experiments

Dataset For evaluating our proposed method, we conduct a series of experiments on the established Cityscapes [11] dataset. This dataset is designed for the evaluation of segmentation algorithms for autonomous driving applications, and includes a set of fine-grained pixel-wise annotations for 19 types of objects typically occurring in traffic scenes, including the road. We report the intersection over union

Technique	IoU
raw CNN features road prior clustering	0.410
+ batch clustering	0.463
+ superpixel overlap [44]	0.633
superpixel align road prior clustering	0.788
+ batch clustering	0.805

Table 1. Ablative results for automatic road mask generation on the Cityscapes training set (300 random samples).

(IoU) metric for the road class, while ignoring regions in the ground truth defined as void.

4.1. Automatic Road Mask Generation

In this section, we evaluate the quality of the automatically generated road masks. We conduct ablative experiments to see how each technique contributes to better segmentation masks and also investigate the sensitivity of the parameter selection involved in the each phase. For these experiments, we randomly sampled 300 images from the Cityscapes training set, and report the road IoU.

Ablation study We compute the road IoU after the steps introduced in Sec. 3, and show how much each technique contributes to the performance. The results are summarized in Table 1. We compare our proposed superpixel align road prior clustering method with clustering the raw CNN features from the dilated ResNet. As can be seen, superpixel align is able to get a much higher IoU than the raw CNN features. We can also see that batch clustering improves both methods, and helps superpixel align to achieve higher IoU. For the sake of comparison, we also compare with previous work that proposes to use superpixels that have sufficient overlap with saliency as features [44]. We treat the road cluster from the raw CNN features as saliency, and use these for selecting superpixels as in previous work [44]. As can be seen, this technique improves the performance of the raw CNN features, but it is still unable to beat superpixel align.

Parameter sensitivity As our method does not use any annotations, we need to manually select some parameters. A question is how sensitive our method is to the choice of the parameters. We change each of the key parameters and compare the final road IoU. The key parameters are listed below with their default values.

- number of clusters: 4
- batch size: 30
- superpixel granularity scale: 300

Fig. 7 shows the sensitivity for the number of clusters. We can see that having too few clusters makes it difficult

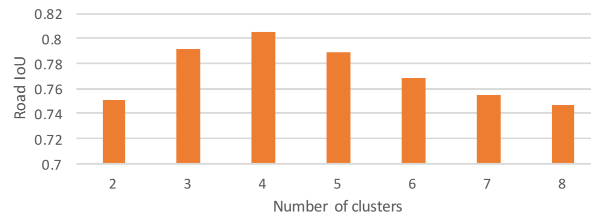

Figure 7. Parameter sensitivity for the number of clusters.

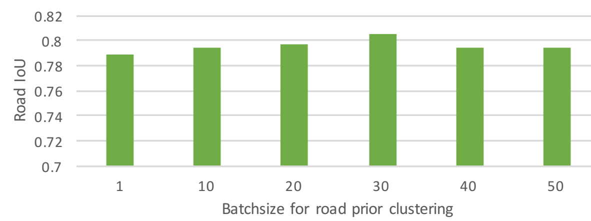

Figure 8. Batchsize vs. road IoU

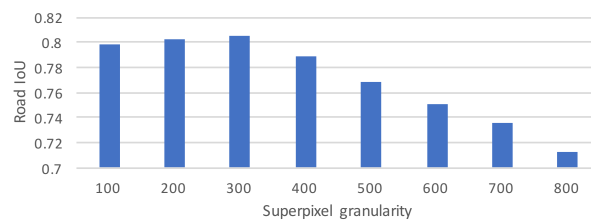

Figure 9. Parameter sensitivity for the superpixel granularity.

to separate road from other parts of the image, while having too many also has some diminishing returns. Nevertheless, the IoU is still fairly high regardless of the number of clusters. The effect of varying the batch size can be seen in Fig. 8. As can be seen, increasing the batch size improves the IoU, but there are diminishing returns after some point. Fig. 9 shows IoU when changing the granularity of the superpixels, which is specified as a parameter called *scale*. Smaller scale tend to be slightly better which seems to be due to large scale causing heavy undersegmentation of the image, which makes it prone to false positives due to *e.g.* merging the road with building walls. In summary, we did not find any highly sensitive parameters. The difference in the final IoU is only a few percent if we change a parameter into neighboring value, which suggests that our method is robust to the choice of the parameters.

4.2. Training a CNN from the Generated Mask

This section describes the road segmentation results obtained by training a CNN on our generated road masks. We use SegNet [5] as our choice of CNN, but we note that our method is not dependent on the CNN, so any CNN could be used.

Experiment setup We train SegNet with our generated road masks as labels for the Cityscapes training images using the Chainer and ChainerCV frameworks [42, 32]. We evaluate our method with several baselines on the valida-

Method	Annotation work	IoU
largest superpixel	none	0.659
bottom half	none	0.720
ground truth as road cluster	-	0.824
distant supervision [44]	additional images	0.800
video segmentation [35][1]	additional images	0.759
fully supervised (from [44])	pixel-wise	0.853
ours (weak labels)	none	0.761
ours (CNN training)	none	0.835

Table 2. Results evaluated on the Cityscapes validation set.

tion set as the test set annotations are not published. After the comparison to previous work on the validation set, we also submitted our model to the Cityscapes test set evaluation server.

Baselines We compare with the following methods:

1. Use the largest superpixel as the road mask.
2. Set the bottom half of the image as the road mask.
3. Use ground truth as clustering results.
4. Distant supervision [44].
5. Video segmentation [35].
6. Fully supervised model.

The first two serve as trivial baselines, which indicates the difficulty of the task. In the third baseline, we use the ground truth mask as the clustering results and combine with the superpixels in a similar manner as in previous work [44]. Although using superpixels is a common approach in weakly supervised segmentation [28, 41, 21, 9], we are interested in knowing its limitations, which is the purpose of this baseline. Therefore, an IoU of 0.824 should be the upper bound of our automatically generated road masks before training a CNN. The fourth one shares a similar motivation with us and uses external images (what they call a distant supervisor) to perform road segmentation in a weakly supervised manner. We do not use any distant supervisor in this work, which makes our method cheaper to use in terms of annotation cost. The fifth one not only uses external images but also videos, which makes the task easier than ours due to the motion cues. The last baseline is using a SegNet model trained from the ground truth annotations [44], which results in an IoU of 0.853. We include this baseline to see how our weak labels compare to the fully supervised performance. The performance of these baselines are summarized in the top of Table 2.

Evaluation We compute the IoU on the evaluation images and obtain the road IoU of 0.835. This is much higher than the trivial baselines, and higher than the video segmentation and distant supervision baseline. Note that while the relative high IoU of the bottom half baseline might make this task seem easy, we emphasize that our method has a much lower false positive rate, which is crucial for employing the method in a practical system. Using the bottom half of the image as a road mask gives precision 0.754, while our weak labels and trained CNN have a precision of 0.867 and 0.898, respectively, thus showing that our method is less prone to fatal false positives, such as a car being mistaken for road. Compared to the fully supervised baseline, we are able to obtain 98% of the IoU of the fully supervised model. This indicates that our proposed method is able to perform proper road segmentation while using no manual annotations for training the CNN.

[1] We contacted the authors and they kindly provided their results on the Cityscapes validation set.

Method	Annotation work	IoU
video segmentation [35]	additional images	0.785
ours (CNN training)	none	0.857

Table 3. Results evaluated on the Cityscapes test set.

Some results on the validation set can be seen in Fig. 10. We can see that while our method typically follows the shape of the true road and avoids labelling cars as road, it has some trouble with *e.g.* pedestrian legs and some parts of the sidewalk being labeled as road. As future work, it would be interesting to see if using a more powerful (although more computationally heavy) CNN architecture might help mitigate these problems [49].

Finally, we also evaluated our method on the Cityscapes test set evaluation server. Results are shown in Table 3, where we can see that consistent with the validation set results, our method is able to gain better performance than the previous work by Saleh *et al*.

5. Conclusion

In this paper, we proposed a method for unsupervised road segmentation from car-centric images. Our method extracts the road region by performing clustering of superpixel features, which are created by a novel superpixel align method that bases features on the last layer in an ImageNet-pretrained CNN. We use road prior knowledge to select the cluster corresponding to the road then perform training of a road segmentation CNN. Unlike previous work, our method is fully unsupervised, and experimental results showed that our method is able to gain superior performance compared to other weakly supervised methods without requiring manual annotations. Compared to a fully supervised model, our method achieved 98% of performance without tedious and costly manual annotations. As future work, it would be interesting to make the road prior image-dependent, to better handle road segmentation of distant road areas, which the current models struggles with.

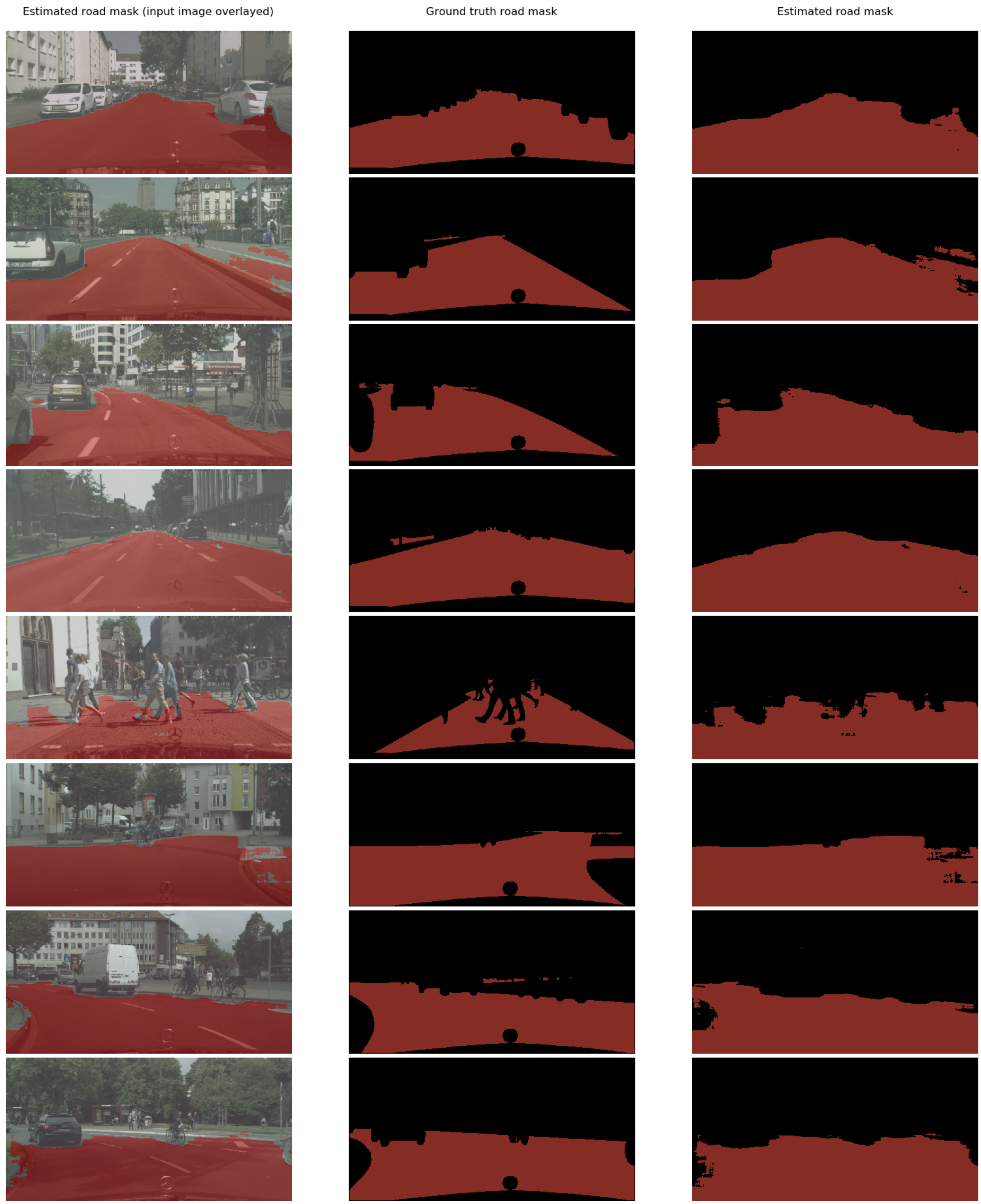

Figure 10. Sample results on the Cityscapes validation set, comparing our method with ground truth.

Acknowledgments We would like to thank the authors of the video segmentation paper [35] for sharing their results on the validation set. We would also like to thank Masaki Saito and Richard Calland for helpful discussions.

References

[1] J. Alvarez, T. Gevers, Y. LeCun, and A. Lopez. Road scene segmentation from a single image. In *ECCV*, 2012. 3

[2] J. M. Alvarez, T. Gevers, and A. M. Lopez. 3d Scene Priors for Road Detection. In *CVPR*, 2010. 3

[3] J. M. Á. Alvarez and A. M. Lopez. Road detection based on illuminant invariance. *IEEE Trans. Intell. Transp. Syst.*, 2011. 3

[4] H. Badino, U. Franke, and R. Mester. Free space computation using stochastic occupancy grids and dynamic programming. In *ICCV Workshop*, 2007. 3

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. 1, 2, 6

[6] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016. 1, 3

[7] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 3

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016. 1, 2

[9] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. F. Wang, and M. Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, 2017. 4, 7

[10] D. M. Christopher, R. Prabhakar, and S. Hinrich. Introduction to information retrieval. *An Introduction To Information Retrieval*, 151:177, 2008. 5

[11] 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 *CVPR*, 2016. 1, 2, 3, 5

[12] T. Durand, T. Mordan, N. Thome, and M. Cord. WILDCAT : Weakly Supervised Learning of Deep ConvNets for Image Classification , Pointwise Localization and Segmentation. In *CVPR*, 2017. 1, 2

[13] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. 3

[14] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision (IJCV)*, 59(2):167–181, 2004. 4

[15] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *CVPR*, 2012. 3

[16] C. Guo, S. Mita, and D. McAllester. Robust road detection and tracking in challenging scenarios based on markov random fields with unsupervised learning. *IEEE Trans. Intell. Transp. Syst.*, 13(3):1338–1354, 2012. 3

[17] S. Hänisch, R. H. Evangelio, H. H. Tadjine, and M. Pätzold. Free-Space Detection with Fish-Eye Cameras. In *IEEE Intelligent Vehicles Symposium (IV)*, 2017. 3

[18] A. W. Harley, A. Ufkes, and K. G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *ICDAR*, 2015. 3

[19] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In *ICCV*, 2017. 4

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[21] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han. Weakly supervised semantic segmentation using web-crawled videos. In *CVPR*, 2017. 1, 2, 4, 7

[22] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 4

[23] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple Does It: Weakly Supervised Instance and Semantic Segmentation. In *CVPR*, 2017. 1, 2, 3

[24] A. Kolesnikov and C. H. Lampert. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation. In *ECCV*, 2016. 1, 2, 3

[25] H. Kong, J.-Y. Audibert, and J. Ponce. Vanishing point detection for road detection. In *CVPR*, 2009. 3

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[27] A. Laddha, M. K. Kocamaz, L. E. Navarro-Serment, and M. Hebert. Map-supervised road detection. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 118–123. IEEE, 2016. 3

[28] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation. In *CVPR*, 2016. 1, 2, 4, 7

[29] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In *CVPR*, 2017. 1

[30] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theory*, 28(2):129–137, 1982. 4

[31] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 2

[32] Y. Niitani, T. Ogawa, S. Saito, and M. Saito. Chainercv: a library for deep learning in computer vision. In *ACM Multimedia*, 2017. 6

[33] G. L. Oliveira, W. Burgard, and T. Brox. Efficient deep models for monocular road segmentation. In *IROS*, 2016. 3

[34] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015. 1, 2

[35] F. Sadat Saleh, M. Sadegh Aliakbarian, M. Salzmann, L. Petersson, and J. M. Alvarez. Bringing background into the foreground: Making all classes equal in weakly-supervised video semantic segmentation. In *ICCV*, 2017. 1, 2, 3, 7

[36] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, J. M. Alvarez, and S. Gould. Incorporating network built-in priors in weakly-supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. 1, 2, 3

[37] W. P. Sanberg, G. Dubbelman, and P. H. de With. Free-space detection with self-supervised and online trained fully convolutional networks. In *IS&T Electronic Imaging - Autonomous Vehicles and Machines (EI-AVM)*, pages 54–61, 2017. 3

[38] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000. 5

[39] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016. 1, 2, 3

[40] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2014. 3

[41] P. Tokmakov, K. Alahari, and C. Schmid. Weakly-supervised semantic segmentation using motion cues. In *ECCV*, 2016. 1, 2, 4, 7

[42] S. Tokui, K. Oono, S. Hido, and J. Clayton. Chainer: a next-generation open source framework for deep learning. In *NIPS workshop*, 2015. 6

[43] G. C. Tseng. Penalized and weighted k-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics*, 23(17):2247–2255, 2007. 4

[44] S. Tsutsui, T. Kerola, and S. Saito. Distantly supervised road segmentation. In *ICCV Workshop*, 2017. 2, 3, 4, 6, 7

[45] A. Wedel, H. Badino, C. Rabe, H. Loose, U. Franke, and D. Cremers. B-spline modeling of road surfaces with an application to free-space estimation. *IEEE Trans. Intell. Transp. Syst.*, 2009. 3

[46] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, 2017. 1, 2

[47] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2015. 2

[48] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *CVPR*, 2017. 2, 3

[49] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid Scene Parsing Network. In *CVPR*, 2017. 1, 2, 7