Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Weakly supervised free-space segmentation by fusing spatial priors and region features for auto-driving

Dongbo Huang
Quanzhou Vocational and Technical University
Hui Wang
224601027@csu.edu.cn
Central South University
Yuqian Zhao
Central South University
Feifei Guo
Quanzhou Vocational and Technical University
Fan Zhang
Central South University
Pei Chen
Central South University
Chunhua Yang
Central South University
Weihua Gui
Central South University

Research Article

Keywords: weak supervision, region-based clustering, free-space segmentation, auto-driving

Posted Date: December 3rd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5487076/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on May 30th, 2025. See the published version at https://doi.org/10.1007/s00530-025-01855-w.

Weakly supervised free-space segmentation by fusing spatial priors and region features for auto-driving

Dongbo Huang [a], Hui Wang [b, *], Yuqian Zhao [b, c], Feifei Guo [a], Fan Zhang [b, c], Pei Chen [b], Chunhua Yang [b, c], Weihua Gui [b, c]

[a] *Quanzhou Vocational and Technical University, Quanzhou 362000, China*

[b] *School of Automation, Central South University, Changsha 410083, China*

[c] *Key Laboratory of Industrial Intelligence and Systems (Central South University), Ministry of Education, Changsha 410083, China*

***Contact:** 224601027@csu.edu.cn

Abstract

Weakly supervised semantic segmentation can significantly alleviate the annotation burden of the pixel-level collection used in full supervision. However, most existing works are based on simple images and only include a few tags, which are not applicable to free-space segmentation under complex driving scenes. In this study, we develop an effective weakly supervised framework with image-level label for free-space segmentation by incorporating the spatial priors with geometric context of road. The proposed method jointly leverages the locating capability of image tags and low-level structure information of superpixels, and fuses spatial weights and superpixel region features by a region-based clustering method. These strategies guarantee the satisfying segmentation results while facing the complex driving scenes. Extensive experimental results on the Cityscapes datasets show that our work outperforms some competing weakly supervised methods in terms of intersection over union (IoU), and is a feasible way to perform free-space segmentation in complex environments for auto-driving application.

Keywords: weak supervision; region-based clustering; free-space segmentation; auto-driving

1. Introduction

Free-space segmentation aims at identifying the available space where safe driving of vehicles is guaranteed on the ground without potential collisions [1], which is a critical component in visual environment perception for auto-driving. This task can be tackled as a semantic segmentation problem by generally classifying each pixel in the image as drivable or undrivable [2, 3]. However, fully supervised segmentation conventionally needs pixel-wise labels, which are expensive and consuming. Meanwhile, the manual labeling deviation

introduced by subjective difference of annotators affects the accuracy of segmentation results, which limits the development and application of autonomous driving systems in new environments.

Weakly supervised semantic segmentation is a solution which can significantly ease the labeling cost. Different from the full supervised semantic segmentation framework requiring the pixel-wise labels, the weakly supervised semantic segmentation has the ability of achieving pixel-level learning with weak tags such as scribbles [4], bounding boxes [5], point [6], and image-level labels [7]. In auto-driving, there is only one drivable space in front of the vehicle and quantity information is not necessary, and therefore the image-level tags with minimum requirements can be utilized for free-space segmentation.

Some existing weakly supervised methods attempt to achieve equivalent performance of fully supervision, but most of them are explored on images under simple scenes. For example, in PASCAL VOC 2012 dataset [8], each image contains only single or a few objects and the same objects may not exist in every image. So, it is relatively easy to learn object's visual representation with weaker labels. While autonomous driving scenes are much more complex, there are a larger number of categories presented in each image, and most objects appear repeatedly in different images, which means the existing researches do not work when it comes to autonomous driving segmentation tasks such as free-space segmentation. Moreover, many salient non-road objects such as cars, persons, and buildings attract more attention during the network learning stage, resulting in the pixels of free-space to be judged as the background category. Therefore, it is a huge challenge to learn distinctive visual representation by weakly supervision for free-space under complex scenes.

In this study, an effective weakly supervised framework based on image-level labels is proposed for free-space segmentation under complex environment by considering the spatial priors and geometric context of free-space. We collect road and non-road images from the Cityscapes [9] and PASCAL VOC 2012 datasets to train a discriminative binary classification network with image tags. During network training process, a bottom-crop strategy is adopted to learn distinctive features of free-space. Then, the spatial weight map is generated by applying the LayerCAM [10] on the trained classification network, informing the spatial significance of any pixel or region in an image. Subsequently, considering that free-space appearing in different images shares similar appearance such as homogeneous texture and color, the superpixel providing inter-pixel consistency and preserving boundary information is employed to form a semantically region feature aggregated from a homogeneous set of pixel features. Next, a region-based clustering method, named Spatial Weight Kmeans (SW-Kmeans), is used

to select superpixels with region features under the guidance of spatial priors for obtaining the pseudo free-space masks. Finally, these generated masks are used as labels to train a fully supervised segmentation network.

The main contributions of this study are as follows:

a) An effective framework is developed for weakly supervised free-space segmentation with only image-level labels by fully exploiting the appearance characteristics of free-space.
b) The proposed method learns discriminative visual representation of free-space under complex environment and presents a fusion mechanism for combing the spatial priors from image tags with geometric context of free-space.
c) Extensive experiments on the Cityscapes dataset demonstrate the superiority of our approach with only image-level annotations.

2. Related Work

2.1 Fully supervised Semantic segmentation

Fully supervised semantic segmentation provides pixel-level annotations to train models. Fully convolutional network (FCN) [11, 12] establishes an end-to-end learning framework by replacing fully connected layers with convolutional ones. U-Net [13] adopts a U-shaped structure to concatenate feature maps from encoder and decoder and uses skip connections to preserve the details lost in the consecutive pooling operations of decoder. SegNet [14] develops a new upsampling method by introducing an unpooling layer, which corresponds to a max-pooling one in the downsampling step. The general framework of the above methods is formulated as an encoder network to learn hierarchies of features, followed by a decoder network to produce dense per-pixel labeled outputs. Deeplabv3 [15] applies several parallel dilated convolutions with different rates to capture the contextual information at multiple scales. However, fully supervised methods lead to a cost of manual labor involved in collecting labels. Although these fully supervised methods achieve substantial improvement in segmentation performance, the cost of manual labor involved in collecting pixel-wise labels is high.

2.2 Weakly supervised semantic segmentation

Weakly supervised methods have been extensively studied to address the problem of expensive annotation cost of pixel-level segmentation labels. These methods utilize weak labels such as scribbles [4], bounding boxes [5], point [6], and image-level labels [10, 16-19], which are readily available in a large-scale dataset, to guide the training of network. In this study, we only focus on weak supervision with image tags. Early approaches [16, 17] treat this task as

Multiple Instance Learning (MIL) problem with each image being regarded as a package containing at least one pixel of the known classes, which fails to consider localization information of objects. Recently, Class Activation Map (CAM) [20] has been widely used as the seed to indicate to area of interest within the images.

CAM [10, 21] extracts class-specific heatmaps to localize objects of different classes in a neural network, and researchers have proposed different way to generate the CAM. Zhou et al. [20] first proposed the CAM which can be generated by adopting the parameters of the final Linear layer as the weight. Then, the weight is used to perform weighted summation on the feature map generated by the last convolution layer. However, the method proposed by Zhou et al. required another training for non-classification network. Therefore, Grad-CAM [22] and Grad-CAM++ [21] are proposed to provide better CAM. However, the Grad-CAM and Grad-CAM++ could not provide satisfying result in analyzing shallow layers. To overcome this issue, LayerCAM [10] is proposed by adopting element-wise weight.

In the weakly supervised semantic segmentation, the CAM is usually considered as pixel-level seeds for later propagating to approximate the integral object area. Most state-of-the-art approaches devote to refining the seeds by learning the intra-image information [18, 19, 23, 24] and cross-image relationship [25, 26], but these methods heavily depend on the activation accuracy of the seeds, which may be incorrectly activated in the background. Some researchers also incorporate CAMs with superpixels to estimate object shapes [27, 28], while their methods are not performing well in complex autonomous environment. CAM is employed as attention map in our framework to generate spatial weight for each superpixel region, which can reduce the impact of wrong activation since we don't use CAM as initial seeds.

2.3 Free-space segmentation

Free-space segmentation is a task of estimating the available space where vehicles can drive safely without potential collisions. This task was traditionally solved by geometric modeling [29] or CNN-based learning [2, 3]. Since the expensive cost of obtaining pixel-level annotations, some researchers investigated weakly supervised free-space segmentation by exploiting additional supervision such as stereo disparity [30] and optical flow of videos [31, 32], instead, our method only inputs a collection of still monocular car images which is less supervised. Saito et al. [33] leveraged distant supervision to train saliency map for road segmentation. In their study, the large public image database which is different from the target domain is adopted to train their model for enhancing the ability of extracting saliency. Saleh et al. [31] introduced a two-stream deep architecture that jointly leverages appearance and motion, and the proposed two-stream deep architecture realize the video segmentation method by

treating foreground and background object equally. Laddha et al. [34] adopted OpenStreetMap, vehicle pose estimation sensors, and camera parameters to generate road annotations. With the help of these pseudo annotations, they could train the proposed approach by self-supervised and detection drivable road area in monocular images. Considering the limitation of monocular, Sanberg et al. [35] used stereo-vision disparity system to automatically generate pseudo training labels for their color-based FCN. Similar to [35], Guo at al. [36] also adopted the stereo-vision to realize the drivable road detection and tracking system. In their study, the road detection problem is formulated as a maximum a posteriori problem in markov random field based on stereo with homography. Wang et al. [37] proposed a deep clustering method to find the common appearance of objects including free-space on the road. These methods ignored the appearance characteristics of free-space, whereas we generate region feature for each superpixel, proving to be effective on obtaining consistent and integral free-space area in our experiments. Tsutsui et al. [38] developed a framework for free-space segmentation with a manual adjusted prior which is fixed for all images, while we can automatically generate a distinctive location prior for different inputs.

3. Proposed method

In this section, we introduce a weakly supervised framework for free-space segmentation only using image-level labels. Our method follows four steps: (1) a binary classification network is trained on publicly available image database and the spatial weight map is generated from a binary classification network; (2) high-level features are extracted from a pre-trained CNN and aggregated into region features by applying superpixel pooling to the dense feature map; (3) Spatial Weight Kmeans (SW-Kmeans) is conducted iteratively to generate free-space pseudo masks by adopting spatial weight map generated in step (1) and region features generated in step (2) as input; (4) the free-space pseudo masks generated from step (3) are used as labels to train the free-space segmentation network for higher quality results. We leverage the implicit locating capability of a discriminative classifier for initial object localization, and use the low-level structure information of superpixels to supply inter-pixel cohesiveness in a region and compensate lost boundary information. Fig 1 gives an illustration of the proposed framework and each step is explained in detail in the following part.

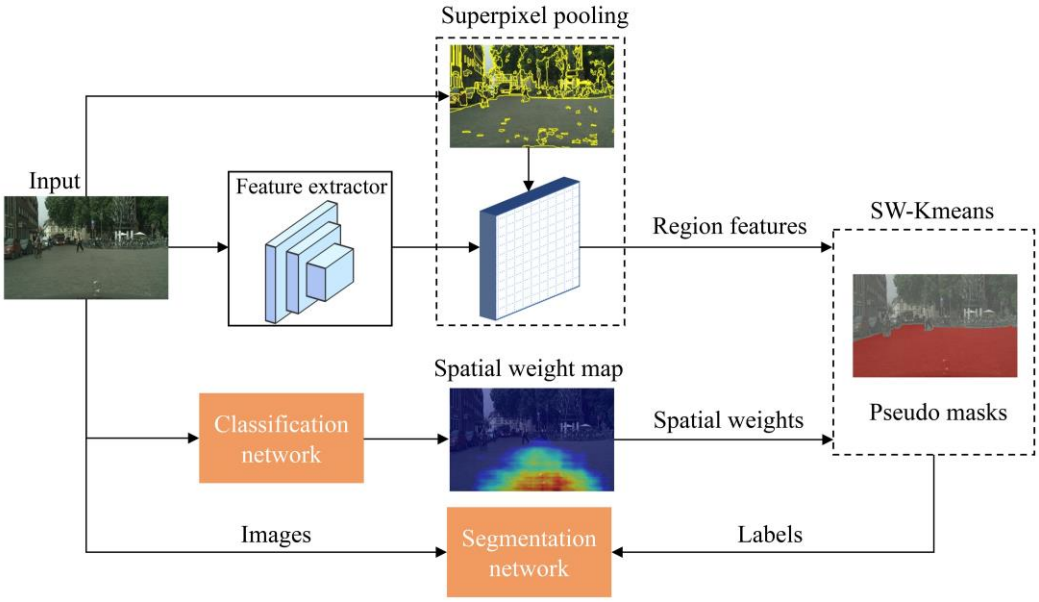


Fig. 1. Overview of the proposed framework.

3.1 Spatial Weight Map

A major challenge in weakly supervised segmentation is that image tags do not directly provide any spatial information of target object. To solve this problem, we train a classification network of free-space for initial localization by associating the tags with discriminative target region in an image.

For simple scenes with only one or a few objects in images, the class labels are highly distinguished to associate with discriminative regions. While the conventional autonomous driving dataset almost contains free-space area in every single image, it is difficult to obtain accurate visual features of target due to the class-imbalance problem. Therefore, we collect road images from the CityScapes dataset, and take advantage of simple scene images from Pascal VOC 2012 for non-road images collection. All collected images have a one-to-one correspondence with class labels.

We aim at learning distinctive visual representation for free-space, however, many salient non-road objects are frequently presented in complex autonomous driving scenes which may confuse the network. Thus, the network tends to pay more attention to these non-road objects and produce biased location guidance. To account for this problem, the bottom of the input image is cropped, and transported to the classification network for training. The bottom-crop image input strategy can largely eliminate the negative effect of the salient non-road objects which are more likely to appear in the center or top of the collected image. The backbone of the classification network is a modified ResNet38 model [39], which is pre-trained on ImageNet [40] and has 38 convolution layers with wide channels. The backbone is transformed

into a fully-convolutional one with dilated convolutions, and the feature maps extracted by the last convolutional layer are input into a global average pooling (GAP) layer to obtain a vector. We add a fully connected layer at the end of the network for classification, enabling the vector obtained by GAP to learn the classification weights in accordance with the image tags, as shown in Fig 2. The classification weights are then utilized by CAM to generate a road-specific spatial weight map.

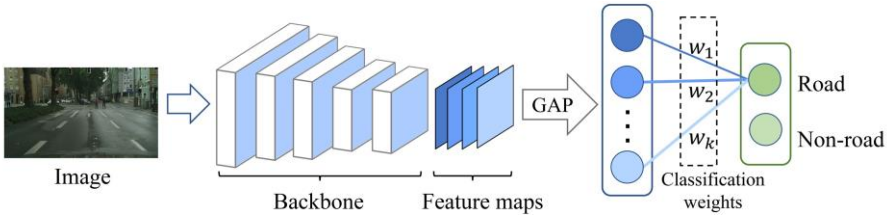


Fig. 2. Free-space classification network architecture.

Given the trained network with GAP followed by a fully connected layer, CAM is computed by applying the LayerCAM on the classification network. The CAM assigns an activation score to individual pixel for each class. We denote the class activation map of class c as M_c, and each spatial element in the map is computed by:

$$M_c(x,y) = relu(\sum_k \hat{f}_k(x,y)) \tag{1}$$

$$\hat{f}_k(x,y) = w_{xy}^{kc} \bullet f_k(x,y) \tag{2}$$

$$w_{xy}^{kc} = relu(\frac{\partial Y^c}{\partial f_k(x,y)}) \tag{3}$$

where $f_k(x,y)$ indicates the activation value of channel k located at (x,y) on the last feature maps, Y^c denotes the predicted results of class c, $relu()$ represents for the linear rectification function. $M_c(x,y)$ is then normalized to be considered as a relevance score for a spatial position (x,y) associated to the free-space class. Finally, we use bilinear interpolation on the normalized map to present the spatial weight map which has the same resolution as the original image. The generated spatial weight map shows the approximate initial location of the object, providing distinctive spatial priors for different images.

3.2 Superpixel region features

Since only a partial region with distinctive visual features will obtain high weight in the spatial weight map, the spatial priors can only locate an initial area of target which is constantly not intact. To meet the pixel-level extraction requirements of semantic segmentation, we adopt superpixel to compensate for the loss of shape details.

The superpixel merges pixels with similar low-level features into a region, providing boundary information with internal consistency, which is in line with the appearance characteristics of free-space. The global weights related to free-space category can be reflected in the spatial weight map by simply averaging the weights of pixels in each superpixel region. We can directly find and aggregate superpixels with high confidence that are very likely to be free-space, but this only considers the low-level appearance of the superpixel and ignores high-level semantic information. Thus, it is necessary to create appearance feature to semantically distinguish free-space for each superpixel.

We first extract the high-level features from a network trained on ImageNet dataset. Such features have been found to capture latent features with rich semantic information, even though the ImageNet dataset does not include free-space as one of its annotated classes [41]. We adopt a dilated ResNet [42] with higher receptive field as the feature extractor, and use the features of different layers in order to obtain multi-scale semantic information. Then, we aggregate the pixel-level features in the superpixels into global features by a pooling operation on the final dense feature map, as shown in Fig 3 In this work, we employ a classic graph-based segmentation algorithm [43] to create the superpixels.

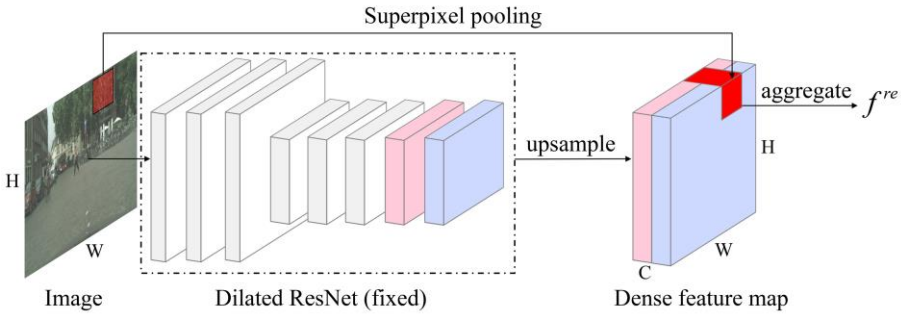


Fig. 3. Overall architecture for computing superpixel region features. The image is inputted to the fixed dilated ResNet to generate a dense feature map. The output feature f^{re} is obtained by aggregating multi-scale features of homogeneous set of pixels within a superpixel.

The superpixel pooling can be looked as the average pooling of local area, which outputs a matrix with size of $N \times C$, where N is the number of superpixels, C the number of feature channels. Specifically, the feature maps of the last two layers are concatenated and upsampled to the resolution of original image for generating the final dense feature map $F \in \mathbb{R}^{W \times H \times C}$. The region features f_i^{re} of superpixel P_i can be expressed as:

$$f_i^{re} = \text{average}\{F(x, y)\}, (x, y) \in P_i \tag{4}$$

where $(x, y) \in P_i$ represents the spatial location (x, y) belonging to superpixel P_i, $\text{average}\{\cdot\}$ denotes the average pooling. That is, for each superpixel, we obtain a semantically

global feature to characterize the region appearance similarity and preserve boundary information.

3.3 Spatial Weight K-means

The remaining challenge is to select candidate free-space superpixels by incorporating the spatial priors with semantic features. We thus propose a simple clustering algorithm to select superpixels with robust region features. Under the guidance of the global spatial weight for each superpixel, we can effectively identify the cluster corresponding to free-space with higher confidence. The cluster center is set to be a weighted combination of region features and spatial weights. In practice, the cluster center of road is calculated as the spatially weighted average of region features assigned to free-space class. For non-road clusters, the center computation method remains the same, but the features assigned to them will have a repellent weight to keep the spatial meaning of other cluster completely away from that of the free-space class. The clustering process is performed on multiple images, and iteratively conducted to generate finer masks. For better illustration, the process of the region-based clustering method, SW-Kmeans, can be formed as Algorithm 1.

Algorithm 1 Process of SW-Kmeans

Input: Spatial weight map W_{map}, superpixel P_i and its corresponding region feature f_i^{re}, number of clusters K

1: $\forall i : w_i = \frac{1}{|P_i|}\sum_{(x,y)\in P_i} W_{map}(x,y)$

2: $\forall i : Class_i = \begin{cases} 0 & \text{if } w_i > \text{median}(w_i) \\ \text{unif}(1, K-1) & \text{otherwise} \end{cases}$

3: **while** not converged **do**

4: $c_0 = \sum_{Class_i=0} w_i f_i^{re} \Big/ \sum_{Class_i=0} w_i$ # cluster center of the zero class

5: $\forall k > 0 : c_k = \sum_{Class_i=k} (1-w_i) f_i^{re} \Big/ \sum_{Class_i=k} (1-w_i)$ # cluster center of the k class

6: $\forall i : Class_i = \arg\min_k \left\| f_i^{re} - c_k \right\|_2^2$

7: **end while**

Output: $Class_i$ for each superpixel P_i

3.4 Semantic segmentation using pseudo masks

The free-space pseudo masks of the training image are generated by the SW-Kmeans method, and are used as labels to train the free-space segmentation network for obtaining higher

quality results. In our framework, the classical segmentation model DeepLabv3 is used as the basic network.

4. Experiments and discussion

4.1 Implementation Details

The training images for free-space classification are collected from the CityScapes and PASCAL VOC 2012 datasets. CityScapes contains 5000 urban scene images of 50 cities with resolution of 2048×1024, and each image has high quality pixel-level annotations for 19 types of traffic objects. This dataset is divided into 2975 training images, 500 validation images, and 1525 testing images. The keyword 'road' is searched as the tags of free-space class, yielding a total of 3418 images as positive examples for classification.

PASCAL VOC 2012 is a standard dataset for weakly supervised semantic segmentation with 21 class annotations, i.e., 20 foreground objects in real scenes and one background. We take the images from train set and val set as non-road category, and 2913 samples are collected as negative examples for classification.

During the training process of classification network, the cross entropy loss function is minimized by using the stochastic gradient descent (SGD) [44] with learning rate of 10-4. A 224×224 crop is randomly sampled from the input training image or from its horizontal flip for simple data augmentation. The model implemented by PyTorch is trained on RTX2080Ti GPU with batch size 16 for 15 epochs. And in the clustering step, the number of clusters is set as 4 by default.

The proposed weakly supervised segmentation method is evaluated on the CityScapes. We only use the road class and adopt the intersection over union (IoU) between predicted segmentation and ground truth as performance metric to evaluate the free-space segmentation results.

4.2 Ablation studies

To verify the effectiveness of the proposed method, we first evaluate the accuracy of the generated pixel-level pseudo masks of free-space, and conduct ablation study to figure out the contribution of each module adopted in the framework. All results are evaluated on the CityScapes val set.

We compare the performance of the adopted bottom-crop image input scheme with that of the whole image input scheme when they are solely used in training process. Fig 4 shows some visualization examples of the road-specific spatial weight map generated by these two schemes. It can be seen that with the former, the generated spatial weight map focuses more details on the road and locate discriminative regions of the free-space, while with the latter, the

map is disturbed by some salient non-road objects like cars, buildings, etc., and produces biased location guidance. The accuracies of these two methods in terms of IoU are listed in Table 1, from which we can find that the IoU reaches 47.9% with the whole image input scheme, while 79.4% with the adopted one. The quantitative and qualitative results demonstrate that the bottom-crop strategy can effectively prevent the interference of non-road objects, and the classification network can learn more distinctive visual features related to the free-space.

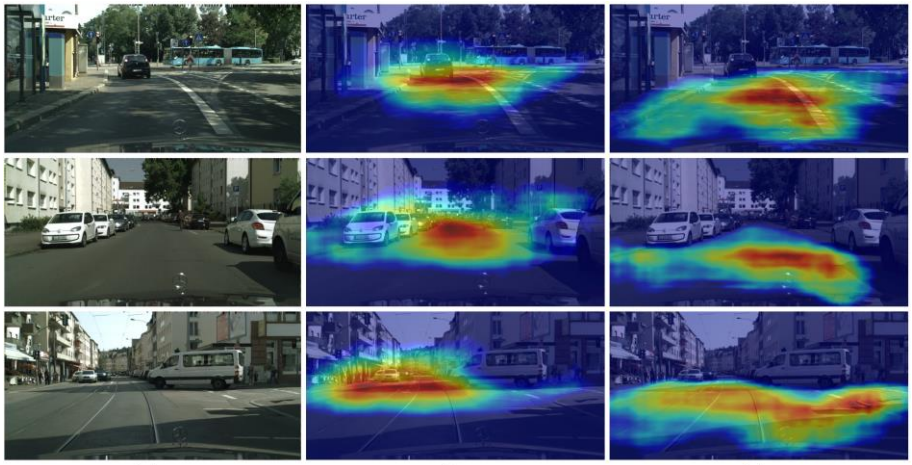


Fig. 4. Comparison of produced road spatial weight maps with different image input schemes during training. (a) Initial image, (b) results of whole image input scheme, and (c) results of bottom-crop image input scheme.

Table 1 Experimental results on Cityscapes val set with different image input schemes during network training.

Crop size	No Crop	Crop50	Crop70	Crop90	Crop110	Crop130	Crop150
IoU	47.9%	76.4%	77.2%	77.7%	**79.4%**	72.6%	60.5%

Note: "No Crop" in this table denotes the whole image input scheme. "Crop50" denotes that pixel difference between the top and bottom of the crop region is 50, and the bottom edge of the crop region is the same as that of the image.

To find an optimal cropped size in the adopted scheme, we conduct experiments on 6 bottom-crop cases with different cropped size. The results of the generated masks in each case are summarized in Table 1. It can be seen that the values of IoU are gradually increasing from 47.9% to 79.4% with the cropped size enlarging from 0 to 110, while decreasing with the cropped size enlarging from 110 to 150. The reason is that the bottom-crop image input scheme with larger cropped size is difficult to eliminate interference factors of non-road objects in classification network, while the scheme with smaller one may lead to lack of the discriminative

features of free-space. The relatively better IoU score, say 79.4%, can be obtained by selecting about half size of the input image.

To evaluate the rationality of using spatial weight map and superpixel region features in the proposed framework, the ablation studies on four strategies are conducted, as summarized in Table 2. **Initial mask** uses the naive seeds generated by CAM to estimate a rough pseudo mask, which is a baseline commonly used in weakly supervised methods. **Spatial aggregation** directly clusters the high-level features of pixels with spatial weights to estimate the free-space region. **Superpixel aggregation** selects superpixels with high global weights to estimate the free-space region. **Manual prior** represents the location prior as an average of Gaussians function by manually adjusting the prior parameters empirically [38].

Table 2 Ablation studies on Cityscapes val set.

Methods	IoU
Initial mask	50.3%
Spatial aggregation	55.7%
Superpixel aggregation	61.2%
Manual prior	76.1%
Ours	**79.3%**

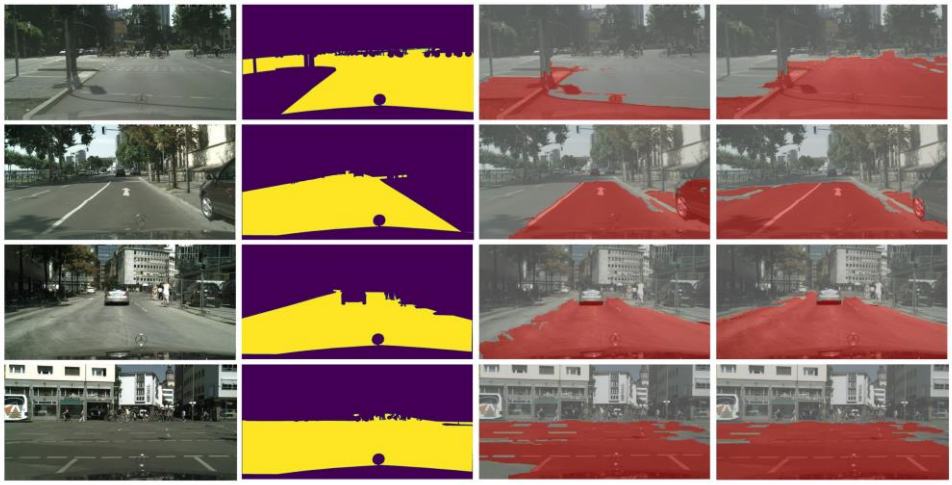


Fig. 5. Free-space masks comparison generated by our method and manual prior. (a) Input image, (b) ground truth, (c) visualization of pseudo mask generated by manual prior and (d) by our method.

As shown in Table 2, the IoU of the spatial aggregation is improved by 5.4% compared to that of the baseline, indicating that the provided spatial priors can effectively guide the feature clustering. In addition, the superpixel aggregation achieves significant elevation by 5.5% on IoU since the appearance characteristics of the free-space are fully utilized without losing shape details. The generated masks of our method further achieve 76.1% on IoU due to the integration of richer semantic information. Our method also achieves higher IoU than the manual prior, since the latter cannot generate road-specific location priors for different images, which may provide wrong guidance when the road is spatially far away from the preset prior. The free-space masks generated by our method and the manual prior are shown in Fig 5, from which it is easy to observe that the proposed method can provide more accurate guidance and achieve better segmentation results while the manual prior does not perform well.

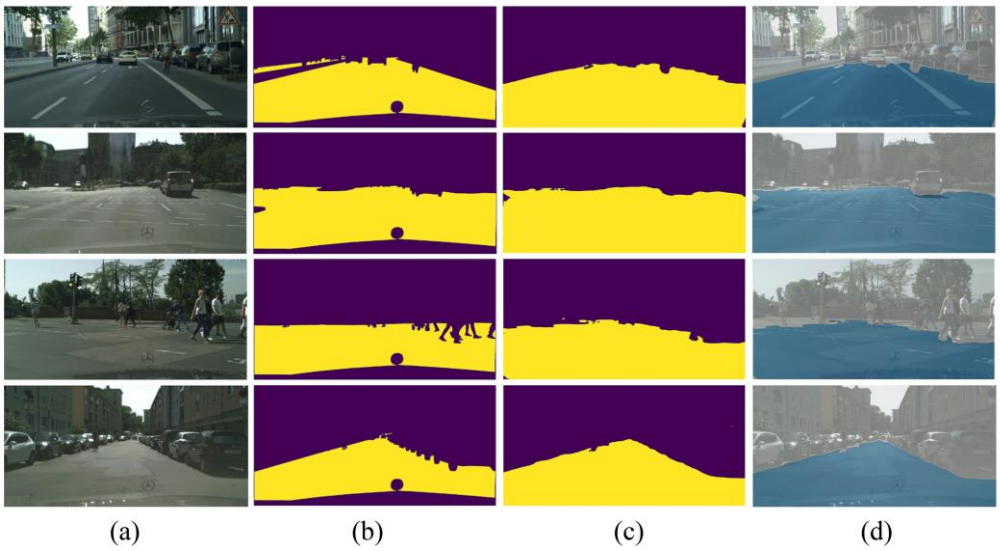


Fig. 6. Qualitative segmentation results of our method on Cityscapes val set. (a) Input images, (b) ground truths, (c) segmentation results, and (d) overlayed results.

4.3 Comparisons to Previous Work

Our ultimate goal is to automatically generate pixel-level annotations for a fully supervised free-space segmentation model. To further evaluate the quality of the generated pseudo masks, we train the classical segmentation model DeepLabv3 on these pseudo masks with full supervision to achieve final results. In Table 3, we compare our framework with other weakly supervised segmentation methods in terms of IoU in autonomous driving scenes. MMS [38] introduce a manually adjusted location prior to select free-space region, followed by Stochastic Co-Teaching U-Net(SCTU)[45] and Augment and Recursive U-Net(ARU)[46]

which also adopted the manual prior as the supervision. Other methods such as BBF [31], DTS [33], and DCWS [37] also make use of saliency maps or heatmaps for target localization. Our method yields the best results of 84.7% on IoU, as shown in Table 3. Some qualitative segmentation results on the Cityscapes val set are shown in Fig 6, from which we can find that the proposed method achieves high-quality segmentation results on free-space even in complex road scenes with pedestrians, vehicles, lane signs, etc.

Table 3 Segmentation result comparison on Cityscapes val set with different methods.

Methods	Supervision	IoU
DCWS [37]	image tags	57.1%
BBF [31]	image tags & videos	75.9%
DTS [33]	image tags	80.0%
MMS [38]	manual prior	83.5%
SCTU [45]	manual prior	82.6%
ARU [46]	manual prior	83.8%
Ours	image tags	**84.7%**

Table 4 Segmentation result comparison on Cityscapes test set with different methods.

Methods	Supervision	Scene	IoU
SC-CAM [23]	image tags	S	20.9%
AffinityNet [18]	image tags	S	31.7%
SEC [19]	image tags	S	43.2%
CCNN [47]	image tags	S	44.3%
OAA [24]	image tags	S	57.5%
BBF [31]	image tags & videos	C	78.5%
MMS [38]	manual prior	C	85.7%
Ours	image tags	C	**86.0%**

The applicable scenarios: S-simple scene, C-complex scene.

We also evaluate the proposed method on the CityScapes test set, and compare it with several state-of-the-art image-labeling weakly-supervised methods such as AffinityNet [18], SEC [19], SC-CAM [23], OAA [24], and CCNN [47], as depicted in Table 4. It should be noted that most of original models provided by these methods are trained on Pascal VOC 2012, so the final results come from our reproduction retrained with CityScapes or from the related work

[48]. The SOTA methods have achieved satisfactory results on simple scenes where only one or a few categories exist in each image. However, they obviously do not perform well in the Cityscapes due to the difficulty in achieving discriminative feature learning via image tags under complex scene. In contrast, our method produces accurate spatial priors and is customized for free-space segmentation by considering the geometric context of road. Therefore, our method achieves IoU of 86.0% on the CityScapes test set, outperforming the SOTA methods. In addition, Table 4 also presents that the segmentation results of our method surpass those of the BBF and MMS methods applied in complex environment, which are consistent with the Cityscapes val set results, indicating the rationality and effectiveness of our method.

5. Conclusion

We introduce an effective image-level labeling weakly supervised framework designed for free-space segmentation in complex environments by considering the spatial priors and geometric context of free-space. The proposed method capitalizes on the locating ability of image tags to generate road-specific spatial priors for different images, and builds discriminative representations of free-space after adopting a bottom-crop strategy during network training. The superpixel region features are extracted to combine the low-level structure information and high-level semantic information to fully exploit the appearance characteristics of free-space. The spatial priors and region features are incorporated by a region-based clustering method to generate the pixel-level labels. The segmentation network retrained by our synthesized labels achieves IoU scores of 84.7% and 86.0% on the Cityscapes val set and test set, respectively, which outperforms some competing methods. The superior performance of the fusion mechanism shows that our work provides a feasible way to perform free-space segmentation in complex environments without tedious manual annotations for auto-driving application.

Acknowledgements

This work is partially supported by the Key R&D Plan Program of Hunan Province (Grant No. 2023GK2021), the Scientific Research Project of Hunan Provincial Education Department (23A0011), the Opening Foundation of Quanzhou Vocational and Technical University in 2024 (LERIS24-04), and the 111 Project (Grant No. B18059). We are grateful for resources from the High Performance Computing Center of Central South University.

References

[1] Janai J., Güney F., Behl A. and Geiger A. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12, 1–3 (2020), 1-308.
[2] Fan R., Wang H., Cai P. and Liu M. *Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection*. Springer, City, 2020.
[3] Tang J., Li S. and Liu P. A review of lane detection methods based on deep learning. *Pattern Recognition*, 111 (2021/03/01/ 2021), 107623.
[4] Liu X., Yuan Q., Gao Y., He K., Wang S., Tang X., Tang J. and Shen D. Weakly Supervised Segmentation of COVID19 Infection with Scribble Annotation on CT Images. *Pattern Recognition*, 122 (2022/02/01/ 2022), 108341.
[5] Hsu C.-C., Hsu K.-J., Tsai C.-C., Lin Y.-Y. and Chuang Y.-Y. Weakly supervised instance segmentation using the bounding box tightness prior. *Advances in neural information processing systems*, 32 (2019), 6586–6597.
[6] Lejeune L. and Sznitman R. A positive/unlabeled approach for the segmentation of medical sequences using point-wise supervision. *Medical Image Analysis*, 73 (2021/10/01/ 2021), 102185.
[7] Huang Z., Wang X., Wang J., Liu W. and Wang J. *Weakly-supervised semantic segmentation network with deep seeded region growing*. City, 2018.
[8] Everingham M., Van Gool L., Williams C. K., Winn J. and Zisserman A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88 (2010), 303-338.
[9] Cordts M., Omran M., Ramos S., Rehfeld T., Enzweiler M., Benenson R., Franke U., Roth S. and Schiele B. *The Cityscapes Dataset for Semantic Urban Scene Understanding*. City, 2016.
[10] Jiang P.-T., Zhang C.-B., Hou Q., Cheng M.-M. and Wei Y. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30 (2021), 5875-5888.
[11] Zhang H., Zhang S., Xing L., Wang Q. and Fan R. Expressive feature representation pyramid network for pulmonary nodule detection. *Multimedia Systems*, 30, 6 (2024), 1-18.
[12] Chen Q., Wang J., Yin J. and Yang Z. CFFANet: category feature fusion and attention mechanism network for retinal vessel segmentation. *Multimedia Systems*, 30, 6 (2024), 1-15.
[13] Yu Z., Yu L., Zheng W. and Wang S. EIU-Net: Enhanced feature extraction and improved skip connections in U-Net for skin lesion segmentation. *Computers in Biology and Medicine* (2023), 107081.
[14] Badrinarayanan V., Kendall A. and Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 12 (2017), 2481-2495.
[15] Liu Y., Bai X., Wang J., Li G., Li J. and Lv Z. Image semantic segmentation approach based on DeepLabV3 plus network with an attention mechanism. *Engineering Applications of Artificial Intelligence*, 127 (2024), 107260.
[16] Carbonneau M.-A., Cheplygina V., Granger E. and Gagnon G. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77 (2018/05/01/ 2018), 329-353.
[17] Roy A. and Todorovic S. *Combining Bottom-Up, Top-Down, and Smoothness Cues for Weakly Supervised Image Segmentation*. City, 2017.
[18] Ahn J. and Kwak S. *Learning Pixel-Level Semantic Affinity with Image-Level Supervision for Weakly Supervised Semantic Segmentation*. City, 2018.
[19] Kolesnikov A. and Lampert C. H. *Seed, expand and constrain: Three principles for weakly-supervised image segmentation*. Springer, City, 2016.

[20] Zhou B., Khosla A., Lapedriza A., Oliva A. and Torralba A. *Learning deep features for discriminative localization*. City, 2016.
[21] Chattopadhay A., Sarkar A., Howlader P. and Balasubramanian V. N. *Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks*. IEEE, City, 2018.
[22] Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D. and Batra D. *Grad-cam: Visual explanations from deep networks via gradient-based localization*. City, 2017.
[23] Chang Y. T., Wang Q., Hung W. C., Piramuthu R., Tsai Y. H. and Yang M. H. *Weakly-Supervised Semantic Segmentation via Sub-Category Exploration*. City, 2020.
[24] Jiang P. T., Hou Q., Cao Y., Cheng M. M., Wei Y. and Xiong H. *Integral Object Mining via Online Attention Accumulation*. City, 2019.
[25] Fan J., Zhang Z., Tan T., Song C. and Xiao J. *Cian: Cross-image affinity net for weakly supervised semantic segmentation*. City, 2020.
[26] Sun G., Wang W., Dai J. and Van Gool L. *Mining cross-image semantics for weakly supervised semantic segmentation*. Springer, City, 2020.
[27] Chen Z., Wang T., Wu X., Hua X.-S., Zhang H. and Sun Q. *Class re-activation maps for weakly-supervised semantic segmentation*. City, 2022.
[28] Sun K., Shi H., Zhang Z. and Huang Y. *Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps*. City, 2021.
[29] Wedel A., Badino H., Rabe C., Loose H., Franke U. and Cremers D. B-Spline Modeling of Road Surfaces With an Application to Free-Space Estimation. *IEEE Transactions on Intelligent Transportation Systems*, 10, 4 (2009), 572-583.
[30] Xie Q., Liu R., Sun Z., Pei S. and Cui F. A flexible free-space detection system based on stereo vision. *Neurocomputing*, 485 (2022), 252-262.
[31] Saleh F. S., Aliakbarian M. S., Salzmann M., Petersson L. and Alvarez J. M. *Bringing Background into the Foreground: Making All Classes Equal in Weakly-Supervised Video Semantic Segmentation*. City, 2017.
[32] Liu P., He Z., Yan X., Jiang Y., Xia S.-T., Zheng F. and Maowei H. *Weclick: Weakly-supervised video semantic segmentation with click annotations*. City, 2021.
[33] Tsutsui S., Saito S. and Kerola T. *Distantly Supervised Road Segmentation*. City, 2017.
[34] Laddha A., Kocamaz M. K., Navarro-Serment L. E. and Hebert M. *Map-supervised road detection*. IEEE, City, 2016.
[35] Sanberg W. P., Dubbelman G. and de With P. H. Free-space detection with self-supervised and online trained fully convolutional networks. *arXiv preprint arXiv:1604.02316* (2016).
[36] Guo C., Mita S. and McAllester D. Robust road detection and tracking in challenging scenarios based on Markov random fields with unsupervised learning. *IEEE Transactions on intelligent transportation systems*, 13, 3 (2012), 1338-1354.
[37] Wang X., Ma H. and You S. Deep clustering for weakly-supervised semantic segmentation in autonomous driving scenes. *Neurocomputing*, 381 (2020/03/14/ 2020), 20-28.
[38] Tsutsui S., Kerola T., Saito S. and Crandall D. J. *Minimizing Supervision for Free-Space Segmentation*. City, 2018.
[39] Wu Z., Shen C. and van den Hengel A. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. *Pattern Recognition*, 90 (2019/06/01/ 2019), 119-133.
[40] Deng J., Dong W., Socher R., Li L. J., Kai L. and Li F.-F. *ImageNet: A large-scale hierarchical image database*. City, 2009.

[41] Saleh F. S., Aliakbarian M. S., Salzmann M., Petersson L., Alvarez J. M. and Gould S. Incorporating Network Built-in Priors in Weakly-Supervised Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40, 6 (2018), 1382-1396.

[42] Yu F., Koltun V. and Funkhouser T. *Dilated Residual Networks*. City, 2017.

[43] Felzenszwalb P. F. and Huttenlocher D. P. Efficient graph-based image segmentation. *International journal of computer vision*, 59 (2004), 167-181.

[44] Bottou L. *Stochastic gradient descent tricks*. Springer, City, 2012.

[45] Robinet F., Parera C., Hundt C. and Frank R. *Weakly-supervised free space estimation through stochastic co-teaching*. City, 2022.

[46] Robinet F. and Frank R. *Refining weakly-supervised free space estimation through data augmentation and recursive training*. Springer, City, 2021.

[47] Pathak D., Krähenbühl P. and Darrell T. *Constrained Convolutional Neural Networks for Weakly Supervised Segmentation*. City, 2015.

[48] Li X., Ma H., Yi S., Chen Y. and Ma H. Single annotated pixel based weakly supervised semantic segmentation under driving scenes. *Pattern Recognition*, 116 (2021/08/01/ 2021), 107979.