Weakly-supervised Semantic Segmentation in Cityscape via Hyperspectral Image

Yuxing Huang
School of Electronic Science and Engineering
Nanjing University
mf1923026@smail.nju.edu.cn

Qiu Shen
School of Electronic Science and Engineering
Nanjing University
shenqiu@nju.edu.cn

Ying Fu
School of Computer Science and Technology
Beijing Institute of Technology
fuying@bit.edu.cn

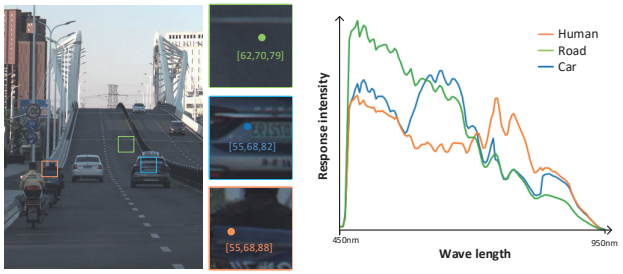
Shaodi You
Computer Vision Research Group
University of Amsterdam
s.you@uva.nl

Abstract

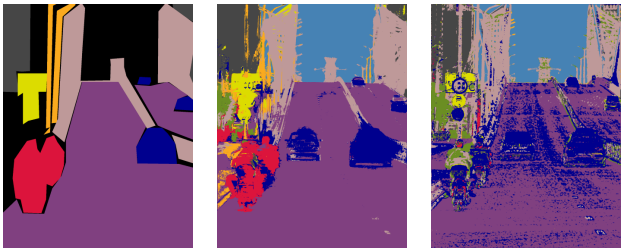
Hyperspectral images (HSIs) contain the response of each pixel in different spectral bands, which can be used to effectively distinguish various objects in complex scenes. While HSI cameras have become low cost, algorithms based on it have not been well exploited. In this paper, we focus on a novel topic, weakly-supervised semantic segmentation in cityscape via HSIs. It is based on the idea that high-resolution HSIs in city scenes contain rich spectral information, which can be easily associated to semantics without manual labeling. Therefore, it enables low cost, highly reliable semantic segmentation in complex scenes. Specifically, in this paper, we theoretically analyze the HSIs and introduce a weakly-supervised HSI semantic segmentation framework, which utilizes spectral information to improve the coarse labels to a finer degree. The experimental results show that our method can obtain highly competitive labels and even have higher edge fineness than artificial fine labels in some classes. At the same time, the results also show that the refined labels can effectively improve the performance of existing semantic segmentation algorithms. The combination of HSIs and semantic segmentation proves that HSIs have great potential in high-level visual tasks for automatic driving.

1. Introduction

Semantic segmentation in cityscape scenes using RGB images has been well exploited (*e.g.*, FCN [28], DeepLabV3 [13] and HRNet [40]). A various of datasets enable such research. (*e.g.*, Cityscapes [15], CamVid [7] , BDD100K [45] and KITTI [21]). We notice that most of the RGB based methods relying on large scale and high quality datasets, large and complex networks and fragile training strategies. This is because, RGB images have inherent limitation on metamerism [10, 9]. As illustrated in Figure 1(a), different objects may have similar RGB value. Metamerism is particularly challenging in cityscape scenes because they contain too many classes, complex lighting and spacial structures. As shown in Figure 1, while RGB based semantic segmentation easily lead to mistakes when



(a) An example of image and metamerism phenomenon.



(b) Coarse label (c) HSI result (d) RGB result

Figure 1: (a) Metamerism: pixels of similar [R,G,B] values may actually have significantly different spectrum. (b) Coarse label. (c, d) An example of spectral classification result based on RGB image and HSI respectively. Based on RGB information, pixels in (a) are all classified as car, but spectrum can distinguish them well.

using shallow network, HSI based methods does not have such problem over the same setting.

In this paper, we propose to introduce HSI into the pipeline of semantic segmentation of cityscape scenes, to break the inherent limitation of RGB images. Firstly, theoretically analysis using t-SNE is executed to show that HSIs are inherently more distinctive than RGB images in cityscape scenes. Based on such advantages, a weakly-supervised semantic segmentation framework is designed to exploiting the advantage of HSI when retain the RGB based semantic segmentation network. Specifically, the coarse label provided by the dataset is refined by learning the prior relationship between hyperspectral information and semantic categories. Then, the refined label is applied to supervise the well-known semantic segmentation network on RGB images (i.e.,HRNet network [40] and DeepLabV3+ [14]). Notice that our proposed framework can generally adopt any RGB segmentation network, and achieve reliable semantic segmentation with coarse label.

Experimental results on Hyperspectral City dataset [44] demonstrated that adopting HSI into cityscape scenes can effectively improve the annotation accuracy. Furthermore, finetuning HRNet [40] with the refined labels can improve 10.10% mIoU over that pre-trained on Cityscapes, and 2.16% mIoU over that fine-tuning with coarse label.

Our main contributions are summarized as follows:

- To the best of our knowledge, our proposed framework is the first paper to apply HSIs into semantic segmentation in cityscape scenes.
- We theoretically analyze the necessity of HSI in cityscape scenes.
- We propose a novel weakly-supervised semantic segmentation framework via HSI only works on coarse labels, which is applicable to any RGB semantic segmentation network.
- We demonstrate the significant performance improved by adopting HSI to label refinement and semantic segmentation in cityscape scenes, which will likely enable a new research area.

2. Related Work

Hyperspectral Image. Hyperspectral images (HSIs) capture the spectral behavior of every pixel within observed scenes at hundreds of continuous and narrow bands, which provides greater information about the captured scenes and objects. HSIs can overcome adverse environmental conditions (*e.g.*, nighttime, foggy, snowy) and reduce the interference of metamerism phenomenon. Several studies [49, 29, 46, 42] have shown the great potential of spectra in cityscape scenes.

In the past, spectra images are usually acquired by scanning or interferometry in remote sensing (RS)(*e.g.*, Indian Pines [4], Pavia University [1] and Houston [17]). But these approaches can only be applied in practice on static or slow-moving scenes. With the advances in compressive sensing theory, snapshot multispectral cameras (*e.g.*, CTIS [18], PMVIS [10] and SPCS [3]) can measure data in a single exposure on sensor. At present, the acquisition technology has been able to capture high-resolution spectral video [11], which greatly expands the application field of spectral imaging.

Semantic Segmentation in Cityscape Scenes. Semantic segmentation is a task of predicting unique semantic label for each pixel of the input image. It has achieved great progress with the works such as FCN [28], UNet [37], SegNet [5], PSPNet [47], DeepLabv3 [13] and HRNet [40]. Fully-supervised Semantic Segmentation depending on huge datasets with pixel-wise annotation (*e.g.*, Cityscapes [15], KITTI [21] and CamVid [7]) is expensive and labor-consuming.

To solve this problem, numerous papers focus on semi- and weakly-supervised semantic segmentation. The semi-supervised methods, such as video label propagation [12, 35, 8], consistency regularization [19, 33], self-training [30, 51, 26, 25, 50] have made great effect, but still rely on the fine annotations. Weakly-supervised methods usually employ bounding boxes [16], scribbles [27], points [6] and image-level labels [36]. For image-level labels, most of methods [2, 23, 41] refine the class activation map (CAM) [48] generated by the classification network to approximate the segmentation mask. Besides, the network also be trained with bounding boxes [24, 36], scribbles [27], or videos [39]. But comes to complex cityscape scenes, it is difficult to utilize these labels for one image contains almost all classes. For cityscape scenes, annotating coarsely only requires each polygon must only include pixels belonging to a single class, which is a low cost weakly-supervised labels [15]. But due to the sparse supervision, using coarse label directly can not achieve competitive results.

Annotation Refinement. There are rare methods focusing on the refinement of coarse labels. In [43], a fully convolution encoder-decoder network with the dense conditional random field (CRF) is proposed for contour detection in order to refine imperfect annotations. However, the large computational cost and sensitivity to parameter selection restrict its practicability. [31] proposes a coarse-to-fine annotation enrichment strategy which expends coarse annotations to a finer scale. But coding and iterating also make it too complex. Fundamentally, in the absence of the prior between spatial information and semantics in coarse labels, it is difficult to refine coarse labels directly.

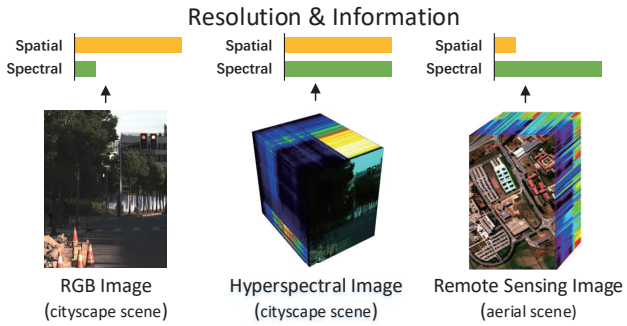
3. Theoretical Analysis

To enable our research in HSIs for low cost and reliable semantic segmentation in cityscape, we adopt a new dataset

(a) Hyperspectral City V1.0

(b) Cityscapes

(c) KITTI



(d) Comparison of RGB and hyperspectral datasets

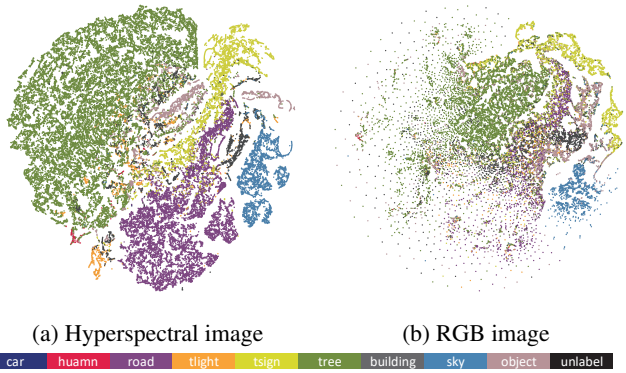
Figure 2: Comparisons of Hyperspectral City dataset with other datasets: Examples from (a) Hyperspectral City [44], (b) Cityscapes [15], (c) KITTI [21]. (d) Comparison of RGB images, HSIs and remote sensing images. The orange and green bars represent roughly the spectral and spatial resolution of the image, as well as the corresponding amount of information.

called the Hyperspectral City Dataset [44]. Based on this dataset, we will introduce the great advantages HSIs present over RGB images in the cityscape scenes.

To best exploit the feasibility of modern semantic segmentation, the dataset focuses particularly on complex cityscape scenes as well as complex lighting conditions. As Figure 2 shows, compared with other cityscape scenes datasets (*e.g.*, Cityscapes [15] and KITTI [21]), the scenes and weather conditions in this dataset are more complicated.

HSI Acquisition. While HSI have been exploited in remote sensing, it cannot be directly used in cityscapes. Remote sensing images are acquired based on scanning or interferometry methods [20, 34], which limit the use of spectra. The Hyperspectral City Dataset was captured by PMVIS [10], which can capture both high-resolution spectral and RGB videos in real time. Specifically, as shown in Figure 2, the PMVIS camera works well in highly dynamic and complex scenes such as cityscape.

HSI in natural scenes and difference from remote sensing. As shown in Figure 2 (d), We compare three kinds of data. The middle cube represents the HSI in Hyperspectral City, which has high spatial and spectral resolution. The left cube represents the RGB image, which is the result of the integration of the spectral image over the spectral channel. RGB image mainly contains spatial information. The right cube represents the aerial remote sensing image, which has high spectral resolution and low spatial resolution. The different imaging scenes and acquisition methods make a great difference between cityscape HSI and remote sensing image.



(a) Hyperspectral image (b) RGB image

Figure 3: Visualizing data using t-SNE. We performed t-SNE visualization of RGB images and HSIs respectively under the same experimental conditions. Compared with RGB images, HSIs have stronger category prior.

Hyperspectral information as semantic feature. Hyperspectral images have higher spectral resolution, so they have better discrimination of semantic features. We use t-SNE [32] to visualize HSI and RGB image. We select the HSI and RGB image in the same frame, which have the same fine label. Due to computational limitations, we use the nearest interpolation to sample the HSI, the RGB image and the fine label to a same low resolution, which keeps each pixel corresponding. We use HSI and RGB image respectively to create a t-SNE visualization. As shown in Figure 3, the result of HSI has a continuous distribution and a clear boundary between each category. On the contrary, for RGB image, the confusion of spectra is more serious. Figure 1(a) plots the spectral curves of different substances with the similar color. HSI can reduce the interference of metamerism phenomenon and provide more powerful information support for cityscape scene analysis.

4. Weakly-supervised HSI Semantic Segmentation Framework

4.1. Overview

Although HSI has inherit advantages, applying it in our task in not trivial. A hyperspectral image requires more than 1G memory. It is problematic that the naive extension of existing network structure by just increasing the channels with research in memory overflow. From the perspective of label, our method avoids the limitation of memory, effectively

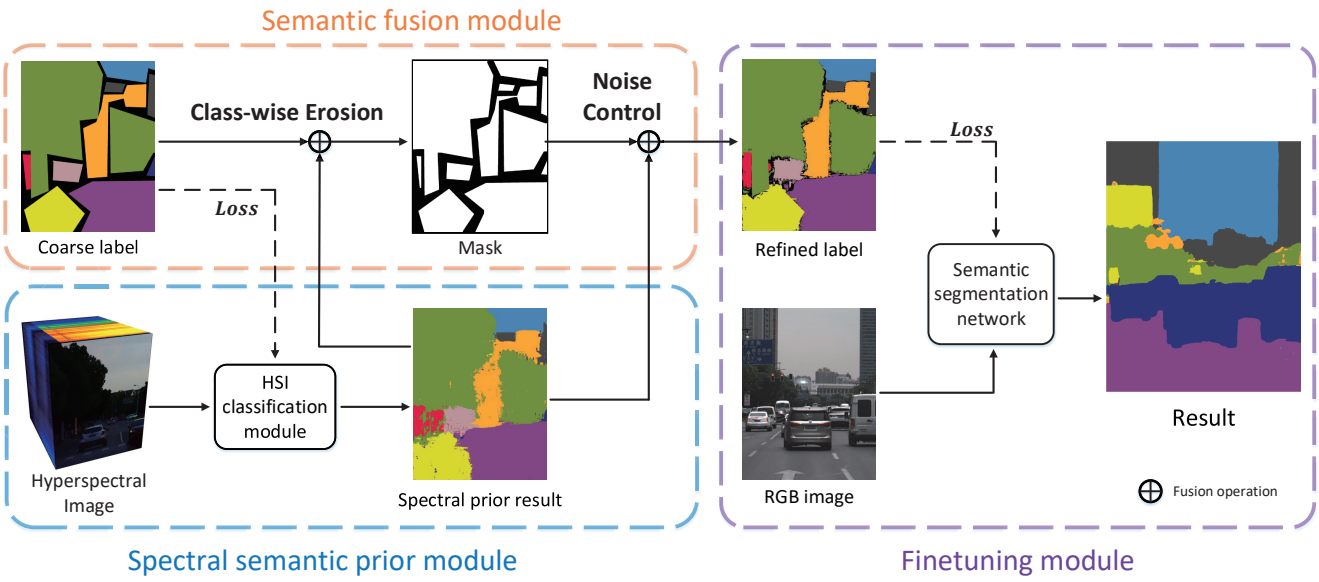


Figure 4: The proposed framework. In the blue dashed box, we train HSI classification module with coarse label as supervision, and then input HSI to generate spectral prior result. In the orange dashed box, we combine the coarse label and the spectral prior through the mask to generate the refined label. In the purple dashed box, we fine-tune the HRNet pre-trained model with the refined label as the supervision.

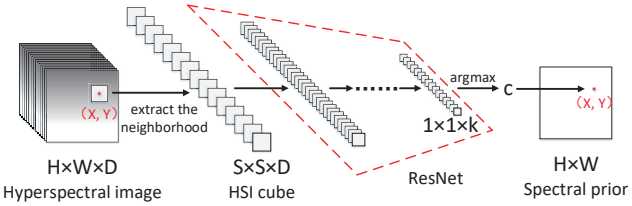


Figure 5: The illustration of HSI classification module. ResNet50 is used to classify the HSI cubes and obtain the label of the pixel. For each pixel of the HSI is classified and finally generate a spectral prior result.

reduces the cost of labeling and improves the performance of segmentation.

As illustrated in Figure 4, there are 3 modules of our method. First, coarse label is used to supervise HSI classification to generate a spectral prior result. This step gets a prior relationship between HSI and coarse label to generate high edge precision result. Second, coarse label and spectral prior are fused to generate more detailed and accurate label, which is called refined label. Third, the refined label is utilized as a supervision to improve the migration effect of the mature semantic segmentation pre-trained model.

Formally, given a set of M training data $X_h, X_r, Y = \{I_i\}_{i=1}^{M}$, let $\mathbf{X}_h \in R^{H\times W\times D_h}$ and $\mathbf{X}_r \in R^{H\times W\times D_r}$ denote a pair of hyperspectral image data and RGB image data, where H and W are the spatial dimensions of the input tensor, height and width, and D_h, D_r is the number of spectral channels. Every X_h, X_r has a pixel at location x, y contains a same one-hot label $\mathbf{Y}_{x,y} = (y_1, y_2, ..., y_k) \in R^{1\times 1\times k}$ where k represents the number of classes.

4.2. Hyperspectral Semantic Prior Module

In this section, we use HSI classification module to generate semantic prior. We hope to use the prior relationship between coarse label and hyperspectral information to obtain label with high fineness, and at the same time, to prevent the influence of coarse spatial information of coarse label.

As shown in Figure 5, we first generate HSI cube $C \in R^{S\times S\times D_h}$ with the size $S \times S$ from X_h, whose center at the space position is (x, y) where $x \in [(S-1)/2+1, H-(S-1)/2]$, $y \in [(S-1)/2+1, W-(S-1)/2]$. Thus, a HSI cube at the position (x, y) is denoted by $C_{x,y}$. The HSI cube covers the height from $x-(S-1)/2$ to $x+(S-1)/2$, the width from $y-(S-1)/2$ to $y+(S-1)/2$ and the whole spectral dim D_h. The number of HSI cubes generated from X_h is $(H-S+1)\times(W-S+1)$. The label of the $C_{x,y}$ is the one-hot label $Y_{x,y}$ of the pixel at the position (x, y).

After we generate HSI cubes from the dataset, we use ResNet-50 [22] as the hyperspectral classification network. We change the input dims of ResNet50 from D_r to D_h to adapt HSI cube. During training, we learn the label under the supervision from the ground-truth $Y_{x,y}(c) = 1c \in [0, k]$ using the cross-entropy loss and $f(C_{x,y}) = Z_{x,y}$ the output of ResNet50, as the following equation shows:

$$\begin{aligned} L(Z_{x,y}, c) &= -\log\left(\frac{\exp(Z_{x,y}[c])}{\sum_k \exp(Z_{x,y}[k])}\right) \\ &= -Z_{x,y}[c] + \log\left(\sum_k \exp(Z_{x,y}[k])\right). \end{aligned} \tag{1}$$

After training, we use the hyperspectral classification network to compute the result for each pixel of HSI X_h to generate spectral prior Z. Sparse random selection of pixels and the use of shallow network prevent overfitting in experiments.

4.3. Semantic Fusion module

In this section, we combine coarse labels and spectral prior to generate refined labels. Although spectral prior Z have higher edge fineness, manual coarse labels have higher confidence in the central region. So a label fusion algorithm is proposed to fuse the advantages of two kinds of labels. First, we remove the low confidence pixels in the spectral prior. Then, we use a class-based erosion strategy to combine spectral prior Z and coarse label Y to generate refined label.

Noise control. For each pixel $Z_{x,y}$ in spectral prior, We use the softmax function to calculate the confidence, where the softmax function is defined as:

$$f_{softmax}(Z_{x,y}) = \frac{\exp(Z_{x,y}[k'])}{\sum_k \exp(Z_{x,y}[k])} \tag{2}$$

where $k' = argmax(Z_{x,y}) \in [1,\, k]$. For pixel $Z_{x,y}$, if the confidence is below the threshold α, the pixel will be set to the label of 'background'; otherwise it is assigned the label class-k'. It plays an important role in controlling spectral prior quality.

Class-wise erosion fusion. Due to the edge of manual coarse label will have some errors beyond the boundary of the classes. At the same time, in the internal area of some classes (*e.g.*, car, building), spectral prior has misclassification. So we propose a class-wise erosion kernel size selection method to obtain the optimal mask, and then fuse two labels.

Coarse label $Y \in R^{1\times1\times k}$ is a one-hot label. For each class ks, the coarse label Y_k is eroded by each category. We choose a square E with size $l \times l$ as the kernel of erosion. For class $i \in [1,\, k]$, the erosion operation as the following equation shows:

$$f_{erode}(Y_i(x,\, y)) = \min_{x',\, y' \in (-l,\, l),\, l \neq 0} Y_i(x{+}x',\, y{+}y')\,. \tag{3}$$

After eroding each class, the regions eroded by each class are added together to form a mask. The mask Y_{mask} after the erosion operation retains the area near the center of each class. Then we use a mask Y_{mask} to fuse the spectral prior Z with the coarse label Y to generate refined label $Y_{refined}$, as the following equation shows:

$$Y_{refined} = Y \times \sum_{i=1}^{k} f_{erode}(Y_i) + Z \times (1 - \sum_{i=1}^{k} f_{erode}(Y_i))\,. \tag{4}$$

Next we use class-wise intersection over union (IoU) as the evaluation index to calculate the IoU scores of each class under different erosion kernel sizes $l \in (1,\, n)$. We select the erosion kernel size l_i with the highest IoU score for each class. By this method, we obtained the final optimal erosion kernel size l_i for each class $i \in [1,\, k]$. Finally, we generate mask with kernel size l.

Obviously, refined label combines the high internal confidence of coarse label and the high edge fineness of spectral prior. Using hyperspectral information, we get high quality label only based on coarse label.

4.4. Finetuning Module

The network structure of semantic segmentation has been relatively mature. To make use of the existing semantic segmentation mature network and prove that our method is useful to semantic segmentation, we fine-tune the HRNet and DeeplabV3+ pre-trained model with our refined labels. More details can be found in supplementary material.

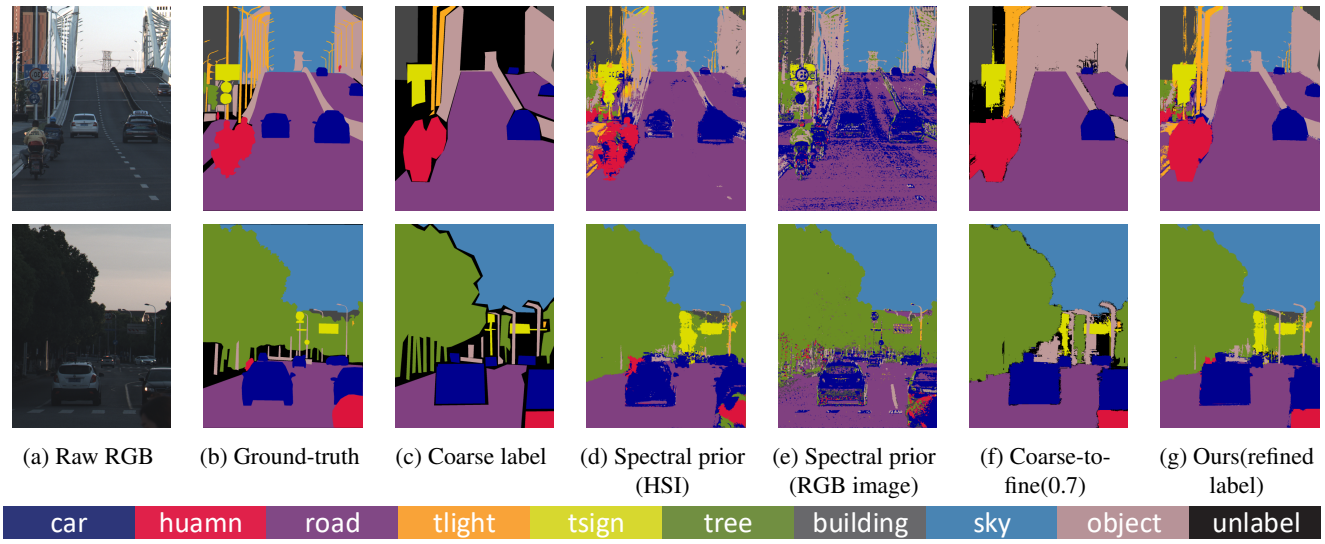
5. Experiments

5.1. Implementation detail

Hyperspectral City dataset. The Hyperspectral City dataset [44] has 367 frames with coarse labels and 55 frames with fine labels. Coarsely and finely annotated images are used for training and testing respectively. There are 6 images have both fine and coarse annotations, which are used for validating. Spectral camera (PMVIS [10]) can capture RGB and spectral images of the same spatial region at the same time. Therefore, each frame captures both the RGB image and the HSI, which have the same spatial resolution 1379 by 1773 and same label. The HSI has 129 spectral channels. The spectrum range is 450 to 950 nm (visible and near-infrared bands) and spectral resolution is 4nm.

Spectral prior. ResNet50 is adopted as the hyperspectral image classification network. Verified by experiment, we set the initial learning rate as 0.01, weight decay as 0.0005 and epoch as 30.

Because HSIs consume a lot of memory, we prepare the data in two steps to balance memory and network training. First, we choose images from training dataset with the batch size 6. Second, we randomly select 10,000 pixels from one HSI, excluding whose corresponding coarse label is "0"(background). Each pixel generates a HSI cube, totally 60,000. Then we randomly choice cubes from these HSI cubes for training with the batch size 256. This method allows us to maximize the use of memory and prevents overfitting at one HSI. The number of HSI cubes used in each image is a small fraction of the total. The spectral information of the same substance has high similarity, which can ensure that sufficient prior information can be learned by using few HSI cubes. We set the spatial resolution of the HSI cube is 11×11.



(a) Raw RGB (b) Ground-truth (c) Coarse label (d) Spectral prior (HSI) (e) Spectral prior (RGB image) (f) Coarse-to-fine(0.7) (g) Ours(refined label)

Figure 6: Visualization results on Hyperspectral City validation set. First we show the results of spectral semantic prior module based on RGB image and HSI respectively. Then we compare the results of semantic fusion module with coarse-to-fine annotation enrichment method.

Table 1: Comparisons of the results of hyperspectral semantic prior module and semantic fusion module with coarse-to-fine annotation enrichment [31] on Hyperspectral City validation set $w.r.t$ mIoU, mean Acc. and IoU of each class.

	mIoU	Acc.	car	human	road	light	sign	tree	building	sky	object
Coarse label	66.3	77.3	80.17	**58.47**	86.52	**26.05**	67.98	80.41	80.53	73.11	43.77
Coarse-to-fine(0.3)	63.6	80.8	80.17	46.91	87.72	23.45	62.34	82.05	81.27	75.03	33.56
Coarse-to-fine(0.7)	64.4	80.3	80.85	48.60	87.57	24.13	66.27	81.96	81.43	74.66	34.16
Coarse-to-fine + HSI	65.0	80.5	80.64	49.94	87.43	23.94	64.58	82.23	81.92	74.67	39.59
Spectral prior (RGB)	28.2	43.3	18.80	2.31	40.45	4.05	18.03	62.98	19.47	75.51	11.93
Spectral prior (HSI)	54.2	77.0	47.93	46.62	81.07	14.74	47.99	72.83	58.48	**90.96**	27.33
Refined label	**69.4**	**82.1**	**81.90**	56.50	**88.33**	24.83	**70.97**	**83.81**	**82.63**	88.75	**46.85**

Refined label. After generating spectral prior. We use the method at section 4.3 to generate fusion label. We generate the mask by eroding each class of coarse label. Ablation experiments are performed on selecting the optimal erosion kernel size and noise control threshold. Then we generate the refined labels on training dataset for finetuning module.

Finetune network. After getting refined label, we use refined label to fine-tune segmentation pre-trained model. For fine-tuning network, we fix the parameters of feature extraction layers, and only fine-tune the last two 1×1 convolution layers. We set the initial learning rate as 0.001, weight decay as 0.0005, crop size as 1773×1379, epoch as 200 and batch size as 3 on four GPUs (GTX 1080Ti). We perform the polynomial learning rate policy with factor $1 - (\frac{iter}{iter_{max}})^{0.9}$. We use $InPlace - ABN^{sync}$ [38] to synchronize the mean and standard-deviation of BN across multiple GPUs. For the data augmentation, we perform random flipping horizontally and random brightness. For evaluation, we use class-wise intersection over union (IoU) and pixel-wise accuracy $(Acc.)$ metrics.

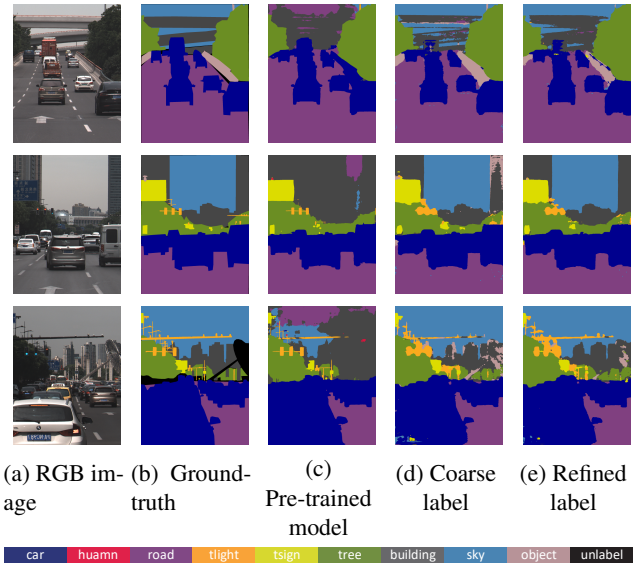
5.2. Quantitative Results

Spectral prior. In Table 1, we compare the spectral prior with coarse label at validation set. First, the spectral prior based on the HSIs is much better than that based on the RGB images. HSIs have stronger semantic prior than RGB images. Second, because the spectral prior has misclassification within some classes, spectral prior is almost equal to coarse labels in Acc. and lower in mIoU. Third, Figure 6 and Figure 7 are the comparisons of some spectral prior and coarse labels on the validation set and training set, which show that spectral prior can effectively improve the edge fineness and correct the wrong or missing annotation in the coarse labels.

Refined label. We compare the best refined labels with the original coarse labels, spectral prior and the result of coarse-to-fine annotation enrichment method [31]. IoU and Acc. are applied to measure the percentage of correctly labeled pixels. Noise control and class-wise erosion fusion are applied to generate refined labels. The parameters selection will be described in the ablation study.

(a) Raw RGB (b) Coarse label (c) Spectral prior (d) Raw RGB (e) Coarse label (f) Spectral prior

Figure 7: Comparisons of spectral prior based on HSI on Hyperspectral City training set. The first 3 columns are the full image and label, and the last 3 columns are area zooms. Spectral prior improves accuracy and corrects errors compared to coarse labels.



(a) RGB image (b) Ground-truth (c) Pre-trained model (d) Coarse label (e) Refined label

Figure 8: Semantic segmentation results of fine-tuning HRNet pre-trained model under different supervision on Hyperspectral City testing set.

As Table 1 shows, the refined label has the highest Acc.(82.1%) and mIoU (69.4%) which brings 4.8% and 3.1% improvement than coarse label. Compare with coarse-to-fine annotation enrichment method [31], our refined label achieves 5.0% mIoU and 1.8% Acc. absolute gains. As Figure 6 shows, due to the coarse-to-fine method relies heavily on the coarse labeled region and does not fit well on large background region, the mIoU score is even lower than coarse label. To further compare with coarse-to-fine method, we add spectral information to the coarse-to-fine method. As Table 1 shows, HSIs can bring 0.6% mIoU improvement. But the results are still weaker than ours. Compared with existing methods, our method has lower requirements on the quality of coarse labels and achieves better results.

Table 2: Results of fine-tuning HRNet pre-trained on Cityscapes with different supervision for semantic segmentation on Hyperspectral City testing set.

	Supervision	mIoU(%)	Acc(%)
HRNet	Pre-trained model	59.30	78.57
	Coarse label	67.24	87.21
	Refined label	**69.40**	**89.12**
DeepLabV3+	Pre-trained model	55.66	79.04
	Coarse label	61.90	85.23
	Refined label	**63.36**	**85.71**

Finetune network analysis. As shown in Table 2, first we use HRNet (HRNetV2-W48) pertrained on Cityscapes as the baseline, whose mIoU is 81.1% on Cityscapes. We divided 19 categories of Cityscapes into 10 categories in the Hyperspectral City by class affiliation. HRNet per-trained model achieves 59.30% mIoU. Although the pre-trained model based on the Cityscapes dataset with fine annotation has high segmentation accuracy, the results of the direct migration pretrained model are poor. We use coarse label to train the network. Coarse label gives network some se-

Table 3: Comparisons of refined label with erosion kernel size l on Hyperspectral City validation set $w.r.t$ mIoU. The threshold of noise control is 0.7. For each class, we select the kernel size l_i with the highest mIoU from 1 to n.

kernel size l	mIoU(%)	car	human	road	light	sign	tree	building	sky	object
Baseline(n=1)	69.20	1	1	1	1	1	1	1	1	1
n=5	69.33	1	1	5	1	5	5	3	5	5
n=9	69.37	1	1	7	1	9	5	3	9	9
n=11	**69.41**	1	1	7	1	11	5	3	11	11
n=15	69.39	1	1	7	1	13	5	3	13	13

Table 4: Comparison of spectral prior generated from different HSI cube sizes on Hyperspectral City validation set.

HSI cube size	mIoU(%)	Acc(%)
5×5	42.04	75.43
11×11	**54.21**	**76.95**
25×25	42.30	74.95

mantic supervision, but it will weaken the precision of pretrained model. As shown in Table 2, we use refined label to fine-tune HRNet and achieve the best mIoU (69.40%) and Acc. (89.12%). We also use DeepLabV3+ (MobileNetV2 as backbone) for fine-tuning module in Table 2. Although our fine-tuning fixed most of the parameters of network, the results also show that our method can bring great segmentation performance improvement.

5.3. Further Ablation Study

Spectral prior. Smaller hyperspectral cubes contain too little spectral information, while larger hyperspectral cubes contain too much coarse spatial information, which all will affect the classification accuracy. Two kinds of information should be taken into account in the selection of hyperspectral cube size. As shown in Table 4, three spectral cube sizes (5, 11 and 25) are compared. The HSI cubes of size 11 achieve the highest mIoU and Acc. on the validation set.

Noise control and Class-wise erosion. In this ablation study, we give comparisons of the noise control and class-wise erosion fusion. Since noise control will introduce few unlabeled area, mIoU is more suitable as an evaluation index of label quality.

First, we test the threshold α of noise control. The spectral prior is directly fused with the coarse label after the noise control operation. After generating the refined label, the mIoU scores under different thresholds are tested on the ground truth of the validation set. As shown in Figure 9, we test α from 0.1 to 0.9, and find that refined label with $\alpha = 0.7$ achieves the best result. Then, we test the erosion kernel size l of class-wise erosion. Firstly The spectral prior for fusion is subjected to noise control with $\alpha = 0.7$. Then we use the class-wise erosion fusion mentioned in section 4.3 to fuse spectral prior and coarse label. As shown in Table 3, the refined label achieves highest mIoU (69.41%) with n=11. Class-wise erosion selects the

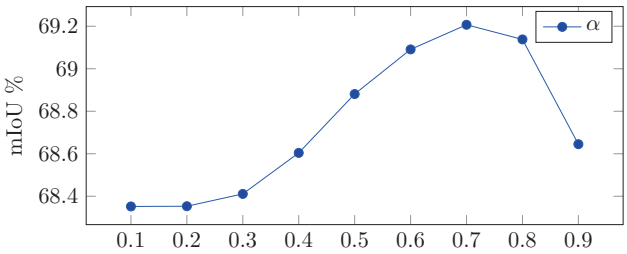


Figure 9: Ablation study of threshold α in noise control.

Table 5: Comparisons of semantic fusion module methods (CE: class-wise erosion(n=11), NS: noise control(α = 0.7)) with fine label of Hyperspectral City validation set $w.r.t$ mIoU.

	CE	NS	mIoU(%)
Coarse label			66.3
Refined label	✗	✗	68.3
	✓	✗	68.4
	✗	✓	69.2
	✓	✓	**69.4**

most appropriate annotation area for different class, which better combines the advantages of the two kinds of labels. The results show that class-wise erosion fusion can further improve the performance of refined label. Finally, we give the qualitative comparisons on Table 5, which demonstrates that operations in semantic fusion module can improve the performance of the refined label respectively.

6. Conclusion

In this paper, we present a weakly-supervised semantic segmentation framework via HSI based on hyperspectral cityscape scenes. Specifically, first, we introduce a new hyperspectral dataset. The comparisons between hyperspectral images (HSIs) and RGB images prove that richer spectral information of HSIs is important to semantic prior. Second, we use the character that spectral information is independent of fine annotation to optimize the semantic segmentation coarse annotation. The label with higher precision is obtained in the case of lower annotation cost. Third, we use the refined label to finetune the semantic segmentation pretrained model, which significantly improves the segmentation accuracy. In future, we hope to continue to explore the application value of spectra in more scenes.

References

[1] Aviris salinas valley and rosis pavia university hyperspectral datasets. http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes. 2

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 2

[3] Yitzhak August, Chaim Vachman, Yair Rivenson, and Adrian Stern. Compressive hyperspectral imaging by random separable projections in both the spatial and the spectral domains. *Applied optics*, 52(10):D46–D54, 2013. 2

[4] NW Aviris. Indiana's indian pines 1992 data set. http://cobweb.ecn.purdue.edu/biehl/MultiSpec/documentation.html. 2

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 2

[6] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 2

[7] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 1, 2

[8] Ignas Budvytis, Patrick Sauer, Thomas Roddick, Kesar Breen, and Roberto Cipolla. Large scale labelled video data augmentation for semantic segmentation in driving scenarios. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 230–237, 2017. 2

[9] Xun Cao, Hao Du, Xin Tong, Qionghai Dai, and Stephen Lin. A prism-mask system for multispectral video acquisition. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2423–2435, 2011. 1

[10] Xun Cao, Xin Tong, Qionghai Dai, and Stephen Lin. High resolution multispectral video capture with a hybrid camera system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 297–304. IEEE, 2011. 1, 2, 3, 5

[11] Linsen Chen, Tao Yue, Xun Cao, Zhan Ma, and David J Brady. High-resolution spectral video acquisition. *Journal of Zhejiang University Science C*, 18(9):1250–1260, 2017. 2

[12] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. *arXiv preprint arXiv:2005.10266*, 2020. 2

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 2

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2, 3

[16] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015. 2

[17] Christian Debes, Andreas Merentitis, Roel Heremans, Jürgen Hahn, Nikolaos Frangiadakis, Tim van Kasteren, Wenzhi Liao, Rik Bellens, Aleksandra Pižurica, Sidharta Gautama, et al. Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2405–2418, 2014. 2

[18] Michael Descour and Eustace Dereniak. Computed-tomography imaging spectrometer: experimental calibration and reconstruction results. *Applied optics*, 34(22):4817–4826, 1995. 2

[19] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham D Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. *arXiv: Computer Vision and Pattern Recognition*, 2019. 2

[20] Nahum Gat. Imaging spectroscopy using tunable filters: a review. In *Wavelet Applications VII*, volume 4056, pages 50–64. International Society for Optics and Photonics, 2000. 3

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1, 2, 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[23] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. *arXiv preprint arXiv:1810.09821*, 2018. 2

[24] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 2

[25] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 2

[26] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. *arXiv: Computer Vision and Pattern Recognition*, 2019. 2

[27] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 2

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[29] Jiarou Lu, Huafeng Liu, Yazhou Yao, Shuyin Tao, Zhenming Tang, and Jianfeng Lu. Hsi road: a hyper spectral image dataset for road segmentation. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 2

[30] Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. Predicting deeper into the future of semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 648–657, 2017. 2

[31] Yadan Luo, Ziwei Wang, Zi Huang, Yang Yang, and Cong Zhao. Coarse-to-fine annotation enrichment for semantic segmentation learning. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 237–246, 2018. 2, 6, 7

[32] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 3

[33] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high- and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 2

[34] Hannah R Morris, Clifford C Hoyt, and Patrick J Treado. Imaging spectrometers for fluorescence and raman microscopy: acousto-optic and liquid crystal tunable filters. *Applied spectroscopy*, 48(7):857–866, 1994. 3

[35] Siva Karthik Mustikovela, Michael Ying Yang, and Carsten Rother. Can ground truth label propagation from video help semantic segmentation? In *European Conference on Computer Vision*, pages 804–820. Springer, 2016. 2

[36] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[38] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5639–5647, 2018. 6

[39] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M Alvarez. Bringing background into the foreground: Making all classes equal in weakly-supervised video semantic segmentation. In *2017 IEEE international conference on computer vision (ICCV)*, pages 2125–2135. IEEE, 2017. 2

[40] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *arXiv: Computer Vision and Pattern Recognition*, 2019. 1, 2

[41] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017. 2

[42] Haitao Xiong, Wenjie Cai, and Qiong Liu. Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene. *Infrared Physics & Technology*, 113:103628, 2021. 2

[43] Jimei Yang, Brian Price, Scott Cohen, Honglak Lee, and Ming-Hsuan Yang. Object contour detection with a fully convolutional encoder-decoder network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 193–202, 2016. 2

[44] Shaodi You, Erqi Huang, Shuaizhe Liang, Yongrong Zheng, Yunxiang Li, Fan Wang, Sen Lin, Qiu Shen, Xun Cao, Diming Zhang, et al. Hyperspectral city v1. 0 dataset and benchmark. *arXiv preprint arXiv:1907.10270*, 2019. 2, 3, 5

[45] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018. 1

[46] Yifei Zhang, Olivier Morel, Ralph Seulin, Fabrice Mériaudeau, and Désiré Sidibé. A central multimodal fusion framework for outdoor scene image segmentation. *Multimedia Tools and Applications*, pages 1–14, 2021. 2

[47] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2

[49] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. *arXiv preprint arXiv:2008.03043*, 2020. 2

[50] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019. 2

[51] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2