Weakly-supervised Semantic Segmentation in Cityscape via Hyperspectral Image (Supplementary Material)

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

In this supplementary material, we first provide more details on Hyperspectral City Dataset. Then we provide more implementation details of our proposed method (Weakly-supervised HSI Semantic Segmentation Framework). Finally, we provide more experimental results and analysis.

1. Theoretical Background and Dataset

In this section, we provide a supplementary introduction to the dataset and show more details on the task, to show difference between existing tasks and our proposed task.

Traditional datasets based on RGB images mainly contain spatial domain information which limits the development of technology. In recent years, with the rapid development in computational photography theory and semiconductor techniques, spectral video acquisition has become feasible [2]. Since each material has unique signatures in a spectral domain, hyperspectral imaging can provide much more information about the captured scenes. New datasets with high resolution on both space domain and spectral domain are necessary. Hyperspectral City Dataset is an cityscape scenes spectral dataset which utilizes newly developed hyperspectral camera PMVIS [1].

As shown in Figure 1, hyperspectral images in the dataset have 1773 by 1379 spatial resolution and have 129 spectral band ranging from 450 to 950 nm. We plot spectral curves at different pixels in the hyperspectral image. To best exploit the feasibility of modern semantic segmentation, the dataset focuses particularly on complex cityscape scenes as well as complex lighting conditions, shown in Figure 2. Compare with other cityscape scenes datasets (Cityscapes [4] and KITTI [5]), this dataset has diverse scenes, lighting and weather conditions, which can fully illustrate the value of our method and has good research value.

2. Weakly-supervised HSI Semantic Segmentation Framework

In this section, we summarize the proposed algorithm and add more details about the module.

2.1. The Proposed Algorithm

The Algorithm 1 describes our weakly-supervised HSI semantic segmentation framework which contains the operations of "Hyperspectral Semantic Prior Module", "Semantic Fusion Module" and "Finetuning Module". Through this algorithm, we obtain refined labels based on coarse labels and hyperspectral information, and migrate a mature semantic segmentation pre-trained model.

Formally, given a set of M training data $X_h, X_r, Y =$

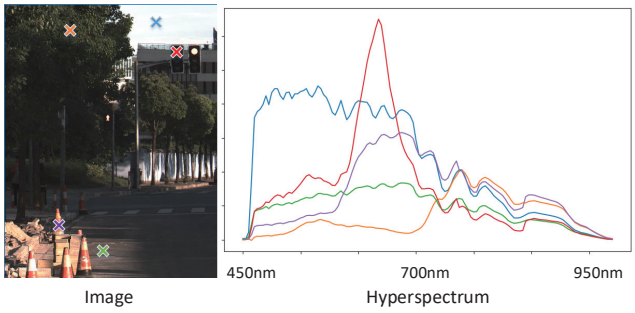


Figure 1: An example of the image and spectral curves of different pixels in Hyperspectral City V1.0 Dataset.

(a) Hyperspectral City V1.0

(b) Cityscapes

(c) KITTI

Figure 2: Comparisons of Hyperspectral City dataset with other datasets: Examples from (a) Hyperspectral City V1.0, (b) Cityscapes [4], (c) KITTI [5].

$\{I_i\}_{i=1}^{M}$, 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 numbers of classes.

2.2. Hyperspectral Semantic Prior Module.

This module first generates HSI cubes $C_{x,y}$ from the hyperspectral images X_h. Then it uses ResNet-50 [6] as the hyperspectral classification network and learn the label under the supervision from the ground-truth $Y_{x,y} = c \in [0, k]$ using the cross-entropy loss. After training, we use the hyperspectral classification network to compute the spectral prior $Z_{x,y}$ for HSI cube $C_{x,y}$. $\phi(C_{x,y}) = Z_{x,y}$ represents the output of ResNet50.

2.3. Semantic Fusion Module.

Here, we show the importance of class-wise erosion operation. As shown in Figure 3, the coarse label has some errors in the edge and direct fusion certainly preserves these errors. The erosion preserves the region with the highest confidence of the coarse label, thus maximizing preserving the spatial structure information of the spectral prior in the fusion. The results show that the refined label with erosion fusion is obviously better than that with direct fusion. At the same time, our class-based selection of erosion kernel size further improves performance.

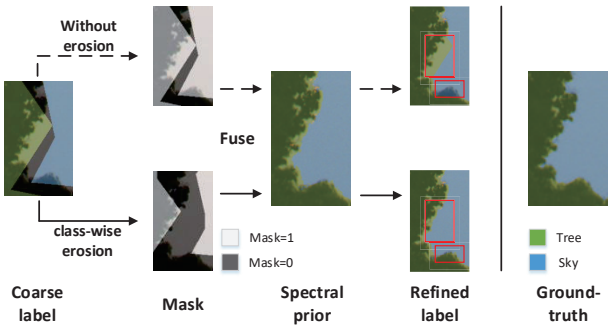


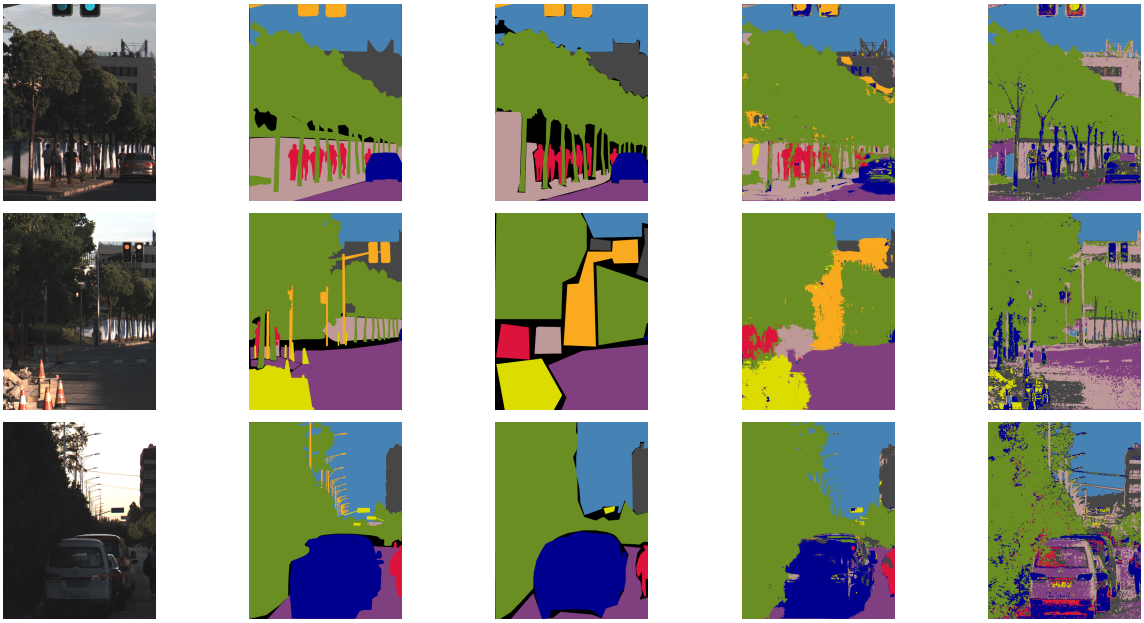
Figure 3: Comparison of the refined label w/ or w/o class-wise erosion.

Algorithm 1 The proposed weakly-supervised HSI semantic segmentation framework

Input: X_h: hyperspectral image; X_r: RGB image; Y: coarse label; ϕ: HSI classification network (ResNet50); $f(X;\theta)$: segmentation pretrained model; n: maximum erosion kernel size; α: noise control threshold;

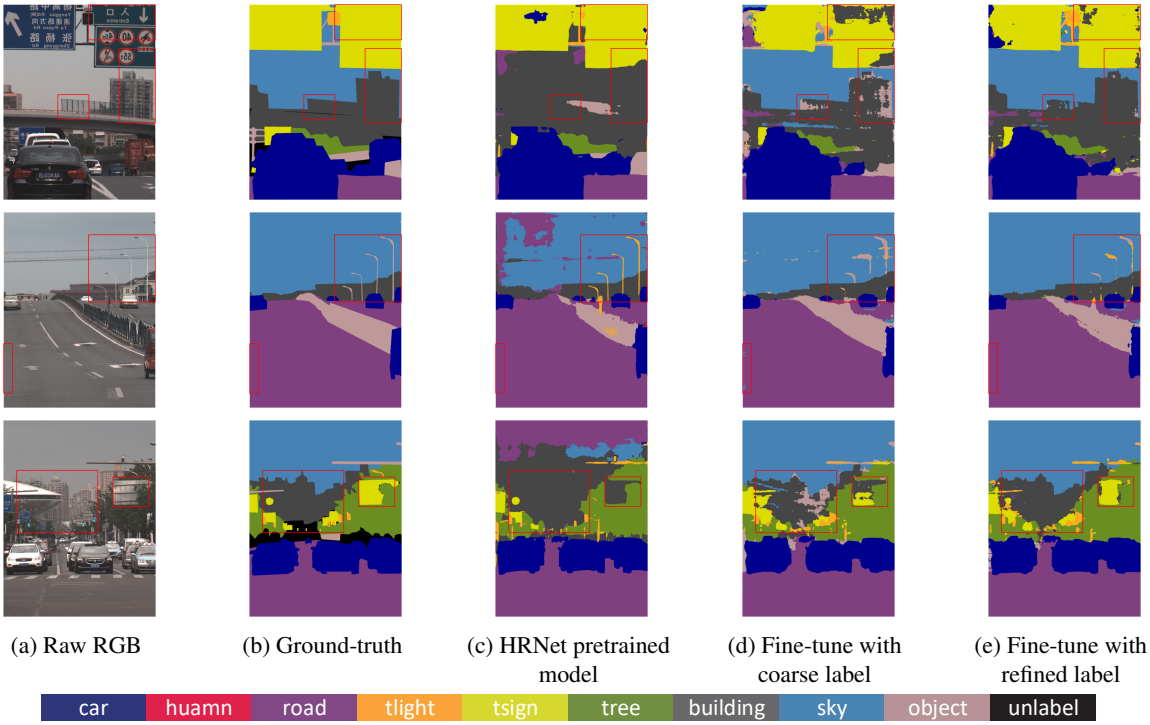
Output: segmentation result on testing set R

generate HSI cube $C_{x,y}$ from X_h;
train HSI classification network ϕ with the cross entropy loss $L(\phi(C_{x,y})), Y_{x,y})$ on training set;
compute spectral prior $Z(x, y) = \phi(C_{x,y})$ with noise control α;
on validation set:
for each kernel size $l \in [1, n]$ **do**
    compute the mask Y_{mask} with the erosion kernel size l;
    compute the refined label $Y_{refined} = Y \times Y_{mask} + Z \times (1 - Y_{mask})$;
    compute the IoU for each class $i \in [1, k]$;
end for;
for each class $i \in [1, k]$ **do**
    select the kernel size l_i with the highest IoU for class i;
end for;
on training set:
compute the optimal mask Y_{mask} with the selected erosion kernel size l_i ($i \in [1, k]$);
compute the refined label $Y_{refined}$ on training set;
finetune HRNet pretrained model $f(X;\theta)$ with the cross entropy loss $L(f(X_r;\theta), Y_{refined})$ with refined label;
compute the segmentation results on testing set $R = f(X_r;\theta)$;
return R;



(a) Raw RGB (b) Ground-truth (c) Coarse label (d) Spectral prior (HSI) (e) Spectral prior (RGB image)

Figure 4: Comparisons of spectral prior result of hyperspectral semantic prior module based on RGB image and HSI respectively with coarse label and fine label on Hyperspectral City validation set.



(a) Raw RGB (b) Ground-truth (c) HRNet pretrained model (d) Fine-tune with coarse label (e) Fine-tune with refined label

Figure 5: Semantic segmentation results of fine-tuning HRNet pretrained model under different supervision on Hyperspectral City testing set.

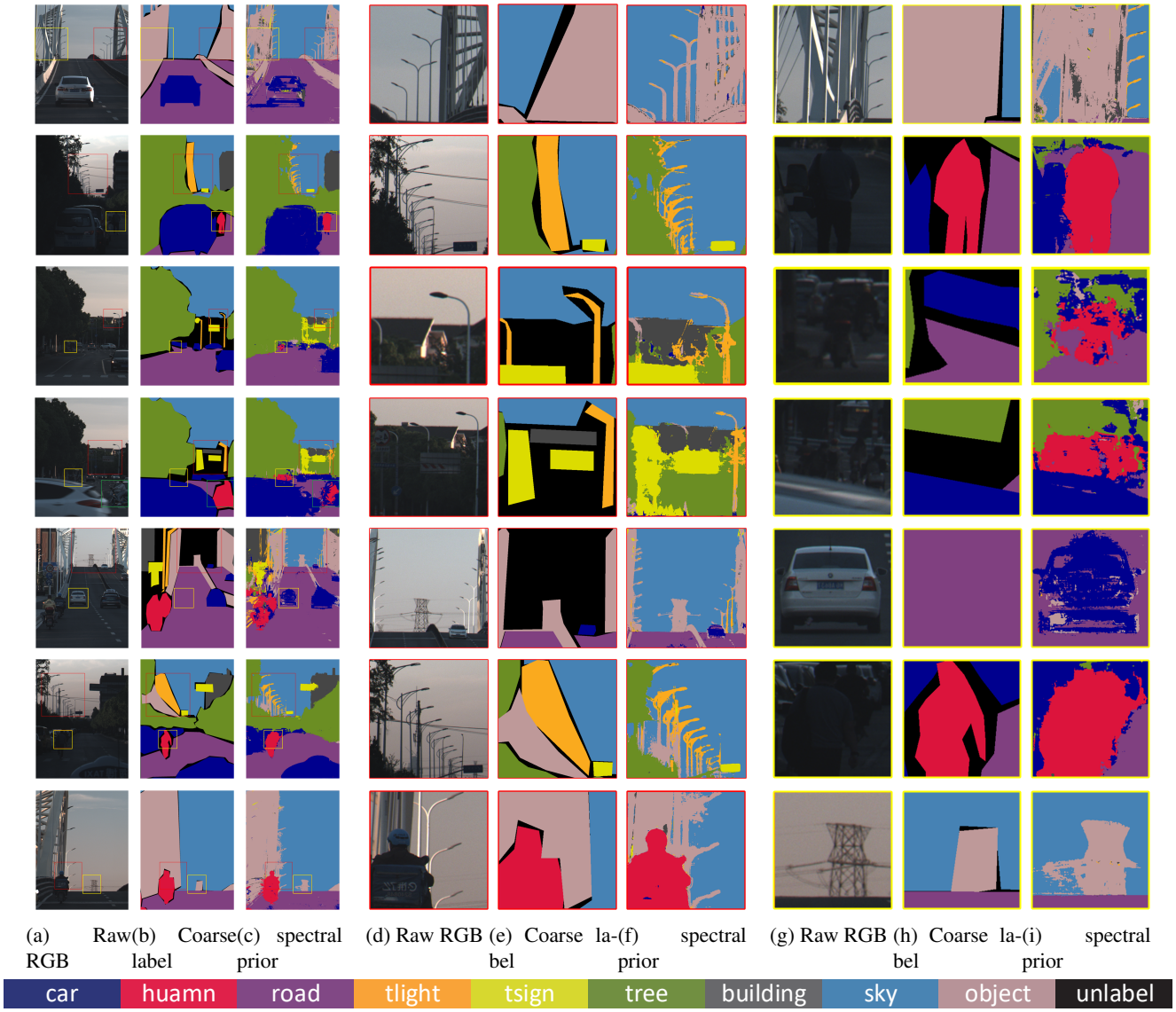


Figure 6: Comparisons of spectral prior based on HSI with coarse label on Hyperspectral City training set. The first 3 columns are the full image and label, and the last 6 columns are area zooms. Refined labels improve accuracy and correct errors compared to coarse labels.

2.4. 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 HSIs is useful to semantic segmentation in cityscape scenes, we finetune the segmentation pretrained model.

In detail, we use HRNetV2-W48 [7] and DeeplabV3+ [3] pretrained on Cityscapes as baseline. HRNet first uses two strided 3×3 convolutions to decrease the resolution to $1/4$. Then it contains 4 stages that are formed by repeating modularized multi-resolution blocks. The last layer mix the output of the four resolution channels by a 1×1 convolution and the output are upsampled (4 times) to the input size by bilinear upsampling. HRNet augment the high-resolution representation by aggregating the (upsampled) representations from all the parallel convolutions. This leads to stronger representation and gets superior results. And in order to prevent overfitting, we fix the parameters of the first two strided 3×3 convolutional layers and the 4 stages of HRNet. Only the parameters of the last two 1×1 convolution layers are trained. Similarly, the DeeplabV3+ also fixes the last two convolution layers. The finetuning module can show the advantage of the spatial fineness of the refined label.

Table 1: Comparisons of refined label with erosion kernel size l from 1 to n and noise control on Hyperspectral City validation set $w.r.t$ mIoU. ($n = 1$ means without erosion operation.)

Noise control	kernel size l	mIoU(%)	car	human	road	light	sign	tree	building	sky	object
$\alpha = 0$	n=1	68.35	**79.57**	**53.31**	88.38	**24.02**	**66.98**	83.40	83.05	89.84	46.58
	3	**68.37**	79.50	53.13	88.46	23.84	66.86	**83.41**	**83.09**	90.11	46.90
	5	68.34	79.36	52.86	88.53	23.62	66.72	83.38	83.07	90.32	47.12
	7	68.25	79.18	52.47	88.57	23.36	66.56	83.33	82.97	90.49	47.24
	9	68.12	78.98	51.97	88.61	23.07	66.39	83.26	82.85	90.62	47.30
	11	67.99	78.78	51.43	88.64	22.84	66.21	83.19	82.73	90.74	**47.32**
	13	67.85	78.58	50.85	88.66	22.65	66.00	83.11	82.61	90.83	47.30
	15	67.70	78.35	50.25	**88.67**	22.46	65.78	83.04	82.49	**90.91**	47.27
$\alpha = 0.7$	n=1	69.20	**81.87**	**56.61**	88.21	**24.84**	70.28	83.75	82.58	88.51	46.17
	3	**69.26**	81.82	56.58	88.27	24.59	70.43	83.79	**82.59**	88.69	46.51
	5	69.24	81.72	56.4	88.31	24.27	70.54	**83.79**	82.54	88.82	46.75
	7	69.17	81.57	56.05	**88.33**	23.90	70.63	83.76	82.43	88.90	46.91
	9	69.06	81.40	55.58	**88.33**	23.51	70.69	83.71	82.28	88.95	47.02
	11	68.92	81.22	55.02	88.32	23.17	70.70	83.64	82.13	88.98	47.07
	13	68.77	81.01	54.42	88.32	22.87	**70.71**	83.55	81.98	**88.99**	**47.09**
	15	68.60	80.78	53.77	88.29	22.57	70.60	83.47	81.83	88.98	47.08

Table 2: Comparisons of refined label with class-wise erosion and noise control on Hyperspectral City validation set $w.r.t$ mIoU. For each class, we select the kernel size l_i with the highest mIoU from 1 to n as shown below. ($n = 1$ means without erosion operation.)

Noise control	Kernel size l	mIoU(%)	car	human	road	light	sign	tree	building	sky	object
$\alpha = 0$	Baseline(n=1)	68.35	1	1	1	1	1	1	1	1	1
	n=5	68.41	1	1	5	1	1	3	3	5	5
	n=9	68.44	1	1	9	1	1	3	3	9	9
	n=11	**68.45**	1	1	11	1	1	3	3	11	11
	n=15	68.43	1	1	15	1	1	3	3	15	15
$\alpha = 0.7$	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

3. More Experimental Results

We provide more results, namely, quantitative results and ablation study of hyperspectral semantic prior module, semantic fusion module and finetuning module.

3.1. Quantitative Results

Spectral Prior. First, we use HSI classification method to generate spectral prior based on HSI and RGB image respectively. As Figure 4 shows, the spectral prior learned from HSI is close to the ground-truth and far better than that learned from RGB images. Meanwhile, it can correct coarse labels and improve fineness. This experiment illustrates that in the acquisition stage of RGB image, a lot of spectral information has been lost. We can obtain detailed results based on hyperspectral semantic prior module.

Second, Figure 6 shows the comparison of some spectral prior and coarse labels on the training set. For example, in row 1, 2, 4, 6, 7, the spectral prior significantly improves the labeling accuracy compared with the coarse label. Especially in row 1, for object class, our method achieves the precision close to manual fine label only by using the coarse label and hyperspectral information. In row 4, 6, our method also has good robustness for the scene with large lighting variation. In row 3, 5, for some unlabeled classes or mislabeled areas, spectral prior can also be a good supplement and correction. Especially in row 3, 4, for pedestrians, our method can identify humans that are not labeled in the coarse labels. In row 5, for object and sky, in the case of complete absence of labels, we can also calculate the precise labels, which is achieved by using rich hyperspectral semantic prior.

The spectral prior based on hyperspectral semantic prior module can effectively improve the edge fineness of all classes and correct the wrong or missing labels in the coarse labels.

Finetune Network. As shown in Figure 5, we add some results of the finetuning network. First, the direct migration of the pretrained model of HRNet still contains many errors. Second, finetuning with coarse labels can improve the

migration results but at a great loss of fineness and accuracy. Finally, the precision and accuracy are both obtained by using the refined label for finetuning.

Although the pretrained model based on the Cityscapes dataset with fine annotation has high segmentation accuracy, the results of the direct migration pretrained model are poor. Experimental results show that our refined labels can effectively improve the quality and classification accuracy of labels with the help of hyperspectral information.

3.2. Further Ablation Study

In this section, we provide complete noise control and class-wise erosion experimental results to illustrate the necessity and effectiveness of these methods.

In the paper, experimental results have proved that noise control achieves the best effect when $\alpha = 0.7$. Firstly, we directly erode coarse label and fuse the coarse label and spectral prior without noise control ($\alpha = 0$). As shown in Table 1, kernel size $l = 3$ achieves the highest mIoU, but only brings 0.02% improvement. Eroding with the larger kernel size, road and sky can effectively improve the accuracy through the refinement of the edge region. But in coarse labels, it is still important to preserve high confidence internal regions for most classes. Therefore, as shown in Table 2, our class-wise erosion can more effectively utilize the respective advantages of spectral prior and coarse label according to the characteristics of different classes.

The experimental results with noise control ($\alpha = 0.7$) further illustrate the function of semantic fusion module. As shown in Table 1 and Table 2, the results are similar to those without noise control. Direct erosion of coarse label will reduce the quality of the refined labels. However, our class-wise erosion makes good use of spectral prior and coarse label respectively. Meanwhile, in Table 2, comparison of the 'sign' class w/ and w/o noise control can illustrate that noise control provides a better basis for class-wise erosion fusion. The two modules are combined to achieve the best refined label.

References

[1] 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] 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. 1

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4

[4] 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

[5] 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

[6] 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. 2

[7] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 4