Point-Supervised Semantic Segmentation of Natural Scenes via Hyperspectral Imaging

Tianqi Ren
Nanjing Univerisity
tianqi.ren@smail.nju.edu.cn

Qiu Shen*
Nanjing University
shenqiu@nju.edu.cn

Ying Fu
Beijing Institute of Technology
fuying@bit.edu.cn

Shaodi You
University of Amsterdam
s.you@uva.nl

Abstract

Natural scene semantic segmentation is an important task in computer vision. While training accurate models for semantic segmentation relies heavily on detailed and accurate pixel-level annotations, which are hard and time-consuming to be collected especially for complicated natural scenes. Weakly-supervised methods can reduce labeling cost greatly at the expense of significant performance degradation. In this paper, we explore the possibility of introducing hyperspectral imaging to improve the performance of weakly-supervised semantic segmentation. We take two challenging hyperspectral datasets of outdoor natural scenes as example, and randomly label dozens of points with semantic categories to conduct a point-supervised semantic segmentation benchmark. Then, a spectral and spatial fusion method is proposed to generate detailed pixel-level annotations, which are used to supervise the semantic segmentation models. With multiple experiments we find that hyperspectral information can be greatly helpful to point-supervised semantic segmentation as it is more distinctive than RGB. As a result, our proposed method with only point-supervision can achieve approximate performance of the fully-supervised method in many cases.[1]

1. Introduction

Semantic segmentation is a fundamental task in computer vision, which aims to assign a semantic label to each pixel in an image, and benefits many practical applications, e.g., autonomous driving [7, 13, 45], medical image analysis [4, 25] and remote sensing [15, 27, 44]. However, the task usually needs pixel-level annotations for training, which is expensive, time-consuming and mistakable. The problem is even worse in the case of natural scenes with huge number of objects, large scale difference and multiple occlusions.

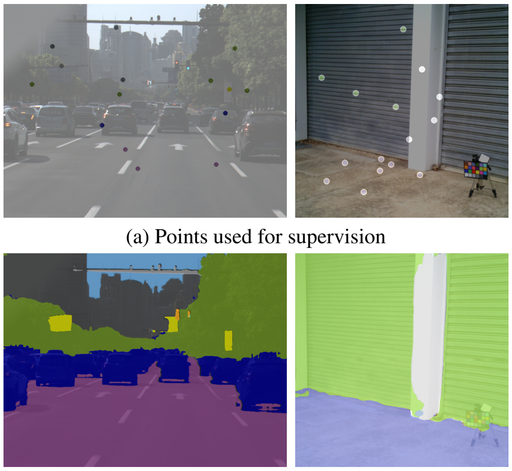
To alleviate this problem, weakly-supervised semantic segmentation methods [46] have been proposed, with image-level supervision [2, 24, 35], box supervision [14, 38], scribble supervision [28] or point supervision [6, 36]. But these methods mainly focus on datasets with simple scenes, such as PASCAL VOC [16] and faces a severe problem of significant performance degradation. For example, point-supervised methods may bring more than 20% performance degradation (measured by mIoU) comparing to full supervision methods [6, 36]. Obviously, it is particularly challenging if only few points are provided as supervision for training a semantic segmentation model. The main reason is that RGB images are in nature not distinctive enough. While same semantics can have huge variety of their RGB values; different semantics can have the same value.

Inspired by the applications in the field of remote sensing [15, 20, 27, 32, 42, 44] using hyperspectral imaging to realize accurate pix-level classification, we imagine that introducing hyperspectral imaging into weakly supervised semantic segmentation of natural scenes can be beneficial. Preliminary researches proved that utilizing hyperspectral information to refine the coarse annotations can improve the semantic segmentation performance [43]. However, their method is still limited to using pixel-level weak annotations, which contain far more information and are yet expensive and time-consuming; beyond that, they train their classification network on the whole dataset, which didn't take different lighting conditions between scenes into consideration.

In this paper, we propose a novel point-supervised semantic segmentation method by utilizing hyperspectral information. As shown in Figure 1, only few points are an-

*Corresponding author: Qiu Shen

[1] This work was supported in part by the National Natural Science Foundation of China under Grant 62071216 and 62231002.



(a) Points used for supervision

(b) Generated pixel-level annotation

Figure 1. **Our method**: Generate pixel-level annotation with few annotated points. Left: image from HSICityV2 dataset; Right: image from LIB-HSI dataset.

notated manually, and other points are classified by training a classification model according to the point-supervision. This process is operated on each image independently to generate accurate pixel-level annotations. Considering that annotated points are sparse and using only hyperspectral information may ignore the spatial relations, we further propose a spectral and spatial feature fusion method to generate refined and semantically correct annotations. Finally, the generated annotations can be used to train any semantic segmentation models in a pixel-level supervision manner. It is worth noting that only RGB images are required as input to train and inference the semantic segmentation model, which means that our proposed method is more practical than directly introducing hyperspectral imaging into the final segmentation procedure. We conduct extensive experiments on LIB-HSI [18] dataset and HSICityV2 [22] dataset. Experiments show that our method can generate pixel-level annotations with more than 90% accuracy comparing to the manually labeled ground truth. Using the generated annotations as supervision can achieve nearing performance of the fully-supervised method (i.e., 62.43% mIoU on LIB-HSI and 55.2% mIoU on HSICityV2), while keeping computational cost at the same level.

The main contributions of this paper are as follows:

- We explore the feasibility and effectiveness of introducing hyperspectral imaging into point-supervised semantic segmentation.
- We propose a framework to extract and fuse spectral and spatial information from hyperspectral images and RGB images, respectively, to generate more accurate annotations for RGB images.
- We prove that point-supervised segmentation can achieve nearing performance of the fully-supervised methods with the assist of hyperspectral information.

2. Related Works

2.1. Semantic Segmentation

Semantic segmentation models based on deep learning usually adopt an encoder-decoder [5] architecture, where the encoder is used to extract features from the input image and the decoder is used to generate the segmentation map. The encoder can be a pre-trained model on the ImageNet dataset, such as ResNet [19], and HRNet [41]. The decoder can be a simple up-sampling layer or a deconvolution layer. Starting from the FCN [30], many methods [8–10, 23, 39, 47] have been proposed to improve the performance of semantic segmentation. For example, DeepLab [8–10] uses astrous convolution to enlarge the receptive field of the network; HRNet [39] uses a high-resolution representation to extract features.

In this paper, we use semantic segmentation models as the final consumer of generated annotation.

2.2. Weakly-supervised Semantic Segmentation

Manually annotating pixel-level labels, which is required for fully-supervised semantic segmentation, is very expensive and time-consuming. It is cheaper to acquire coarse annotations, such as image-level supervision [2, 24, 35], box supervision [14, 38, 40], scribble supervision [3, 28] and point supervision [6, 36]. To the best of our knowledge, there are few works utilizing point supervision currently, especially in the field of natural scene semantic segmentation. In [11], Cheng et al. use point-based annotations along with bounding boxes to avoid the influence of points outside the gt boxes and improve the accuracy of supervision. In [17], Fan et al. use point-based annotations to mark the position of objects and generates pseudo labels. We also find that many unsupervised segmentation metrics, such as Classification Activation Maps (CAM) [48], Superpixels [1], are widely used in weakly-supervised semantic segmentation.

2.3. Hyperspectral Imaging

Hyperspectral imaging capture hundreds of bands in the electromagnetic spectrum. Thus, hyperspectral images can provide more information than RGB images. Several studies in the fields of military [26], agricultural [34] and industrial [37] have shown that hyperspectral images have great potential in many applications.

Currently, there are few works on natural scene semantic segmentation using hyperspectral images. This is due to the lack of large-scale datasets. As hyperspectral cameras become more popular and cheaper, some larger-scale datasets of natural scenes have been released, such as HSIRoad [31], HSICity [22, 43] and LIB-HSI [18].

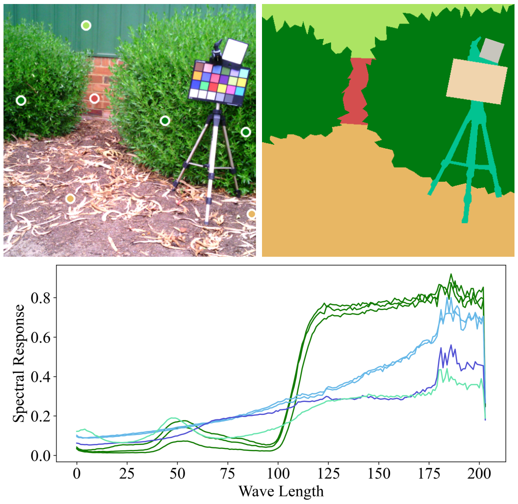


Figure 2. Spectral response of different objects. Object pixel in the scene and its spectral response are assigned with the same color.

It is found that a few accurate hyperspectral information of an object can provide more distinctive features for semantic segmentation since different materials have different spectrum. As a result, hyperspectral information can be treated as additional prior knowledge to improve the performance of weakly-supervised annotations. Huang et. al[21] proposed a semantic segmentation method on HSICityV1 [43], which trains a hyperspectral classification model in advance, and generates more accurate labels for training semantic segmentation model. We adopt the similar idea to generate more accurate pseudo labels and use them to train a segmentation model.

3. Method

3.1. Background

The spectral response of a material is determined by the material's chemical composition and physical structure. Thus, the spectral response of a material is unique. Figure 2 shows that objects belonging to the same category share similar spectral response, while objects belonging to different categories have different spectral response. Compared to RGB images that are composed of three bands (i.e., red, green, and blue), spectral response of a material has a much wider range. As a result, hyperspectral images own a larger feature space and have stronger ability to distinguish different objects than RGB images and can avoid spectral confusion. Figure 3 shows the t-SNE [29] analysis results of the same scene in RGB and hyperspectral images. Pixels with the same label are more clustered in hyperspectral, indicating that objects belonging to the same class share a more general spectral signature.

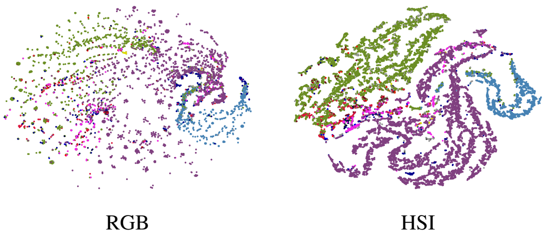


Figure 3. t-SNE [29] analysis results on the same scene in RGB and hyperspectral images. NOTE: The number of points are the same in two sub-figures, while RGB values of many points are the same making the sub-figure of RGB appear more sparse.

Though hyperspectral images have the same spatial resolution as RGB images, it's relatively hard to extract spatial information from them efficiently. On the one hand, some cameras can only capture hyperspectral images with low spatial resolution and reconstruct them to match the spatial resolution of RGB images. On the other hand, hyperspectral images are of great size, which makes it extremely hard to be used to train feature extraction models. By conducting classification and segmentation experiments respectively on hyperspectral images and RGB images, we find that the RGB images contain richer spatial information while hyperspectral images contain more accurate spectral information. Thus in this paper, we propose a novel weakly-supervised semantic segmentation method, that utilize hyperspectral information as distinctive prior to generate pixel-level annotations with sparse point supervision. Furthermore, both RGB and hyperspectral images are used for efficient spatial and spectral feature extraction and fusion.

3.2. Overview

As is illustrated in Figure 4, our method consists of two parts: (a) **Annotation Generation**. We extract spectral and spatial features from hyperspectral images and RGB images respectively, and fuse them to generate more accurate annotations. (b) **Semantic Segmentation**. The semantic segmentation module trains a segmentation model on the generated annotations from (a) using RGB images in a supervised manner.

Formally, given RGB image $X_r \in R^{H \times W \times 3}$, where H and W are the height and width of the image, and a sequence of hyperspectral points $X_h \in R^{(M+N) \times C}$, where C is the number of bands and M and N are the number of annotated and unannotated points. Annotation generation module F_a first trains classification model f_s with point-based annotation $Y \in R^{M \times 1}$ using annotated points from X_h and predicts the onehot labels $Y_p \in \{0,1\}^{(M+N) \times D}$ of X_h, where D is the number of object classes. Then, F_a extracts spatial features from X_r using Superpixels and generates spatial feature map $Y_s \in R^{(M+N) \times 1}$ with X_r. F_a then fuses Y_p and Y_s to generate annotations $Y_a \in R^{(M+N) \times 1}$.

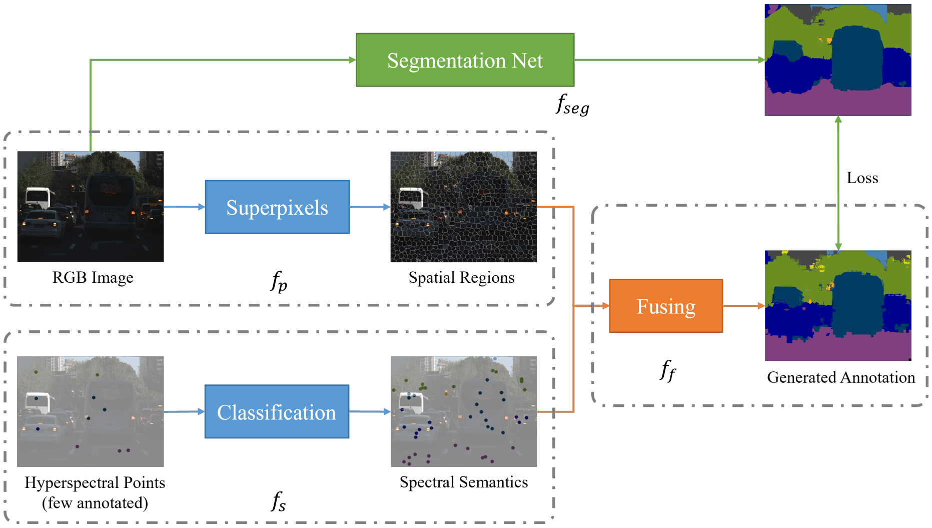


Figure 4. **Overview of our method.** First, RGB image is segmented unsupervisedly into several regions, each of which share similar spatial features. Second, a HSI classification network is trained to extract spectral semantic information. Third, the outputs are fused to generate more accurate annotations. Finally, a semantic segmentation network is trained with generated annotations in a supervised manner.

Finally, X_r and Y_a are fed into semantic segmentation module F_m to train a segmentation model that contains fused features from X_r and X_h.

3.3. Annotation Generation

In this section, we adopt a two-stage method to generate more accurate annotations from RGB images, hyperspectral images and point-based annotations. In the first stage, we use supervised learning method to extract spectral features from hyperspectral images using point supervision and extract spatial features from RGB images in an unsupervised manner. In the second stage, we design a fusion module, which takes both the spectral and spatial features into account to generate annotations with stronger supervision.

Spectral feature extraction of hyperspectral images. Given a set of HSI sequences $X_h \in R^{(M+N)\times C}$ and point-based annotation $Y \in R^{M\times 1}$, the spectral feature extraction module f_s is trained to map X_h to onehot labels $Y_p \in \{0,1\}^{(M+N)\times D}$, where $D = \max(Y)$. As a result of the stable spectral response of different objects, only a few samples of X_h need to be annotated for f_s to generate reliable annotations for the rest of X_h. The spectral feature extraction module f_s is trained by minimizing the following loss function:

$$\begin{aligned}\mathcal{L}_s &= \mathcal{L}_{ce}(f_s(X_h), Y_h) \\ &= -\sum_{i=1}^{M} Y_{h,i} \log(f_s(X_{h,i}))\end{aligned} \quad (1)$$

Different from most HSI classification methods, which usually requires HSI cube with both spatial and spectral information, our method only requires HSI sequences with spectral information. As a result, traditional methods for HSI classification, such as SVM, Logistics regression, can be adopted as f_s. Though these methods are not as powerful as deep learning methods, they are of great efficiency and satisfying performance when X_h is of small size. When X_h is of large size, we can use modified classification models, such as TwoCNN, HybridSN, etc. to train f_s. The main difference between original and modified models is the remove of spatial branch, since spatial information are not included in X_h. We propose a modified version of TwoCNN by removing the spatial branch of TwoCNN, and replaced it with an average pooling branch to extract lightness feature of HSI sequences. The modified network is of less parameters and stronger ability to extract spectral features.

Spatial feature extraction of RGB images. Given an RGB image $X_r \in R^{H\times W\times 3}$, the spatial feature extraction module f_p is trained to map X_r to segmented labels $Y_s \in R^{H\times W\times 1}$ in an unsupervised manner. The pixels are clustered into multiple clusters using Simple Linear Iterative Clustering (SLIC)[1] algorithm. Given a target region count k, SLIC treats pixel as a 5D vector, which consists of pixel's RGB values and spatial coordinates (r, g, b, x, y), and clusters pixels with similar values into k categories. The pixels in the same cluster are considered as belonging to the same object and are assigned the same label. Y_s is composed of the the result category c of each pixel in X_r, where $c \in [0, k)$. Since the result generated by SLIC contains no semantic information, Y_s consists of k regions of pixels with spatial feature extracted from X_r, but lack of semantic information.

Feature fusion. Given Y_p and Y_s, the fusing algorithm f_f utilizes Y_p and Y_s to annotations $Y_a \in R^{H \times W \times 1}$. For each pixel cluster in Y_s, we find the corresponding pixel in Y_p with the same label and assign the label to all pixels in the cluster. If there are several pixels in the cluster with different labels, we assign the label with the highest probability to all pixels in the cluster; if there's no pixel in the cluster with a label, the pixels in this cluster are assigned with a special label C_i, which is ignored in loss computation. For superpixel category $c \in [0, k)$, the fusing algorithm can be formulated as:

$$S_c = (x, y) \text{ s.t. } Y_s(x, y) = c$$
$$Y_a(x, y) = \begin{cases} \arg\max \sum\limits_{(x,y)\in S} Y_p((x, y)) & S_c \neq \varnothing \\ C_i & S_c = \varnothing \end{cases} \tag{2}$$

3.4. Semantic Segmentation

In this section, we train a semantic segmentation module f_{seg} using generated annotations Y_a and RGB images X_r in a supervised manner. The semantic segmentation module F_s utilizes a semantic segmentation network f_{seg} trained by minimizing the following loss function:

$$\mathcal{L}_{seg} = \mathcal{L}_{ce}(f_{seg}(X_r), Y_a) \tag{3}$$

Since we use existing semantic segmentation networks, such as FCN, PSPNet, DeepLab, etc., to train f_{seg}, the semantic segmentation module can be easily adapted to different tasks.

After training, the semantic segmentation module f_{seg} can be used to segmenting new RGB images. Hyperspectral images are not required during inference, which reduces the computation cost in real-world applications, making our method more practical.

4. Experiments

4.1. Datasets

As is mentioned in Section 3.1, we focus on the semantic segmentation of natural scenes with complicated objects and backgrounds. Thus, we choose LIB-HSI and HSICityV2 datasets which contains large-scale urban scenes with both hyperspectral and RGB images for experiments.

LIB-HSI. LIB-HSI is a facade dataset for semantic segmentation, which consists of 513 scenes with both HS and RGB images. The objects are divided into 9 categories and 44 sub-categories, the first of which contains stuffs for hyperspectral data validation such as whiteboard, palette, etc. Each hyperspectral image is 512×512 and the spectral resolution is 204 bands from 400nm to 1000nm. The spatial resolution of RGB images is 512×512.

HSICityV2. HSICityV2 is an urban street view dataset for semantic segmentation, which consists of 1306 scenes with both hyperspectral and RGB images. Scenes are annotated according to Cityscapes dataset, which contains 19 categories, with an additional whiteboard category for hyperspectral data validation. Each hyperspectral image is 1889×1422 and the spectral resolution is 128 bands from 450nm to 950nm. The spatial resolution of RGB images is 1889×1422.

4.2. Implementation Details

Pre-processing. As a consequence of light conditions, the spectral reflectance of HSIs is usually very low and may differ a lot between different scenes. Thus, a whiteboard correction and a normalization process is needed to make the spectral response of hyperspectral images more consistent. For the whiteboard correction, we calculate the average reflectance vector of the whiteboard by averaging the spectral response of all pixels in the whiteboard area, which are labeled as whiteboard in the annotation. Then we divide the spectral response of each pixel in the HSIs by the average reflectance vector of the whiteboard. After the whiteboard correction, HSIs are normalized by dividing the max value of the spectral response of each pixel.

Annotation. Both LIB-HSI and HSICityV2 datasets contain hyperspectral and RGB images. However, the contained hyperspectral images are in the form of spectral cubes instead of points and annotations are at pixel-level. Thus, we need to collect hyperspectral point sets and point-based annotations for experiments manually. For this purpose, we divide the full image into N regions spatially, and then in every region, we randomly sample k points. The number of regions N and the number of points k are hyper-parameters. Then, the collected $N \times M$ points are divided into training set and testing set according to the ratio of p. The points in the training set should be annotated manually, while in this paper the annotations are directly obtained from the two existing datasets with available annotations. The points in the testing set will be classified by the network training with training set.

Annotation generation. We obtain SVM and modified TwoCNN models to classify the hyperspectral points. The SVM model is implemented using scikit-learn[33] and the TwoCNN model is implemented using PyTorch. The TwoCNN model is trained with SGD optimizer with learning rate 0.01 and batch size 32, and training process is stopped after 40k iterations. For RGB spatial feature extraction, we adopt the SLIC algorithm from OpenCV library and set the region size to 50. After that, pixel-level annotations are generated by fusion the results from the two branches.

Semantic segmentation. We adopt FCN, Deeplab-v3p and HRNet as semantic segmentation models for training with generated annotations. Every model is trained with default training schedules from MMSegmentation [12] on each dataset. The batch size is set to 4 and training process is stopped after 80k iterations. After the models are trained, we use images from LIB-HSI and HSICityV2 test sets for inference. We also select Huang et al.'s method [21] for comparison. Since their method is based on pixel-level annotations, we pad pointed-based annotations to original size and use them as pixel-level annotations for training.

Reproducibility. All the experiments above are conducted on a server with NVIDIA RTX3090 GPUs. Code implementation, training configurations and generated hyperspectral images with point-based annotations are available at `https://github.com/iori2333/pointseg-hss`.

4.3. Results

4.3.1 Annotation generation

Since the final segmentation model is trained using generated annotations, the performance of the segmentation model highly depends on the quality of the generated annotations. Table 1 shows the quantitative results of the generated annotations on LIB-HSI and HSICityV2 train sets with different classification methods (i.e., SVM and TwoCNN). The accuracy and mIoU is computed over the manual pixel-level fine annotations, which are provided by the datasets. It is obvious that, the generated annotations using our method can achieve comparable results with the ground truth with extremely high accuracy and mIoU. It demonstrates that hyperspectral information is distinctive enough for generating accurate pixel-level annotations from sparse labeled points.

To better study the results qualitatively, we compare the generated annotations using proposed method with the ground truth (i.e., pixel-level labels annotated manually) on LIB-HSI and HSICityV2 in Figure 5. Benefiting from the additional information from hyperspectral information, generated annotations can achieve almost consistent with the ground truth. The inconsistency appears almost at ambiguous boundaries. It is interesting that,the generated annotations has more detailed boundaries (*e.g.*, the border between leaves) than manually labeled ones. This is because the there may be inaccurate annotations in the ground truth due to human errors, while our method generates annotations based on the spectral information, which depends on the physical properties of the object surface and is more accurate.

	LIB-HSI		HSICityV2	
	Acc	mIoU	Acc	mIoU
SVM (RGB)	66.22	51.04	35.12	18.41
SVM (HS)	**90.50**	**85.28**	**87.13**	**72.99**
TwoCNN (RGB)	83.21	75.91	60.64	54.76
TwoCNN (HS)	**91.44**	**86.64**	**92.55**	**89.25**

Table 1. Quantitative performance of generated annotations.

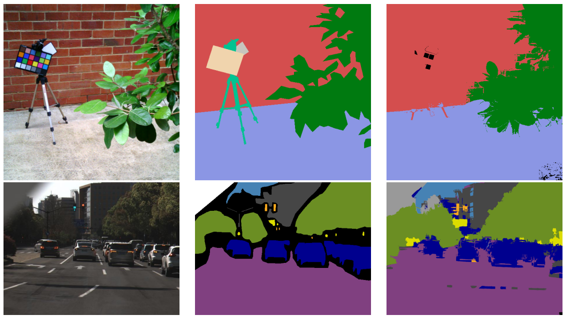

(a) Image (b) Manual (c) Ours (TwoCNN)

Figure 5. Different annotations of LIB-HSI and HSICityV2 train set. (a) Original Image in RGB format; (b) Manual annotation at pixel-level; (c) and (d) Generated annotations using sampled point-based annotations with our method. The whiteboard and palette categories are omitted as only used for HSI data validation.

Annotations	FCN	DeepLab-v3p	HRNet
Point-based	24.24	20.48	28.81
TwoCNN+SLIC	60.04	56.65	**62.43**
SVM+SLIC	55.60	54.32	57.91
GT	61.22	59.98	62.22

Table 2. Quantitative performance (mIoU) of semantic segmentation using generated annotations on LIB-HSI test set

Annotations	FCN	DeepLab-v3p	HRNet
Point-based	25.75	21.39	27.30
TwoCNN+SLIC	49.44	51.39	**55.20**
SVM+SLIC	43.46	47.14	50.01
GT	56.22	60.55	60.29

Table 3. Quantitative performance (mIoU) of semantic segmentation using generated annotations on HSICityV2 test set

4.3.2 Semantic segmentation

LIB-HSI Table 2 shows the quantitative results on LIB-HSI of semantic segmentation using generated annotations. We can see that the semantic segmentation results using the generated annotations are close to the fine labels, which means our method can generate annotations with similar semantic segmentation results, especially for the categories with more training samples. The final result is convincing, outperforming the original point-based annotations by a large margin of 33.62% mIoU.

HSICityV2 Table 3 shows the comparison results on HSICityV2. The results are similar to LIB-HSI, however,

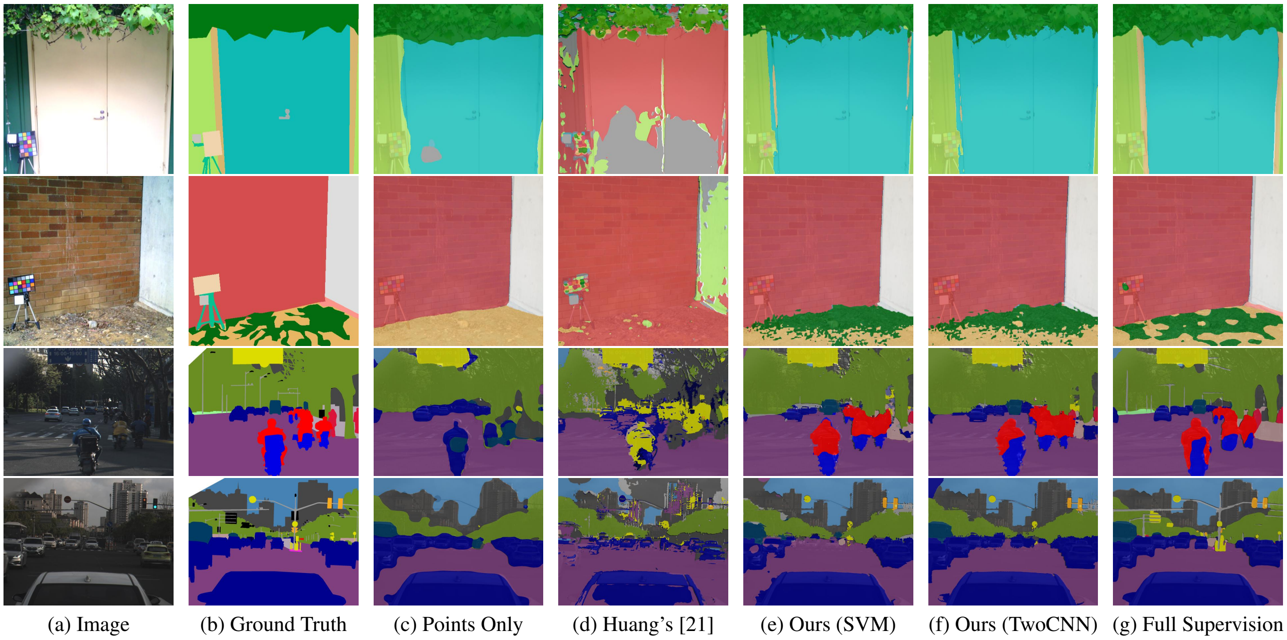


Figure 6. Semantic segmentation results on LIB-HSI and HSICityV2 test set. All models adopts HRNet-w48 as backend, and are trained using different annotations including (c) point-based annotations sampled from gt, (d) generated annotations using our method with SVM, (e) generated annotations using our method with TwoCNN, (f) Huang's method [21] and (g) pixel-level annotations with full supervision.

the performance is not as good as LIB-HSI. The reason is that the scenes in HSICityV2 contain much more small objects (e.g. traffic lights, pole), which may be missed by the point annotations and also missed by the generated annotations. Besides, the manual annotation of the dataset is labeled based on the semantic category instead of the material of the object surface, making the semantic supervision less accurate. Despite these disadvantages, the final results are still convincing, outperforming the original point-based annotations by a large margin of 27.9% mIoU.

Figure 6 shows the results of semantic segmentation using different annotations. We can observe that using generated annotations for training can achieve better performance than using point-based annotations, while some categories with few samples (e.g. poles in HSICityV2) are less accurate. This is because the performance of semantic segmentation highly depends on the generated annotations, which is less accurate on few-shot categories. But in general, our method can achieve comparable performance with the fully-supervised method on categories with sufficient training samples, which is promising.

4.3.3 Comparison with related works

Table 4 shows the comparison results between our method and several related works. Results are evaluated on test sets of LIB-HSI and HSICityV2 with different backbones and different inputs. Note that $\mathcal{P}$ indicates point-based annotation, $\mathcal{F}$ indicates full mask annotations, and $\mathcal{M}$ indicates weak mask annotations with only sampled points from $\mathcal{F}$. Compared with vanilla semantic segmentation, generated annotations can provide richer supervision than point-based annotations and thus improving the performance significantly. It demonstrates that the usage of hyperspectral information can provide promising supervision for learning. Compared with vanilla hyperspectral classification models, our method take spatial context of the scenes into consideration, which improves the feature exploitation throughout different scenes. Compared with Huang's weakly-supervised hyperspectral semantic segmentation method, which uses full size hyperspectral image along with RGB image, our method can better exploit the spatial features and achieve much better performance. Compared with fully-supervised learning methods, our method can also achieve competitive performance without the need of accurate annotations.

To emphasize the necessity of using hyperspectral information as additional supervision, we compare the performance of our method with that using only RGB images. To adopt only RGB inputs, we treat RGB pixel values as spectral image with 3 bands only and adapt the same learning tragedies. We can see that our method using both hyperspectral and RGB inputs can achieve 20% and 15% higher mIoU respectively on LIB-HSI and HSICityV2 datasets, denoting that hyperspectral information is necessary in our method. The result is also included in Table 4.

Methods	Backbone	Inputs	Supervision	LIB-HSI		HSICityV2	
				mIoU	Acc	mIoU	Acc
HRNet	HRNet-48	RGB	$\mathcal{M}$	28.81	38.24	27.30	32.89
FCN	FCN-50	RGB	$\mathcal{M}$	24.24	31.43	25.75	30.53
RSSAN	RSSAN-w17	HS	$\mathcal{P}$	6.58	12.18	23.86	31.41
JigSawHSI	JigSawHSI-w17	HS	$\mathcal{P}$	5.09	8.69	18.17	24.11
Huang et al. (2019)	ResNet-50 + HRNet-48	HS+RGB	$\mathcal{M}$	3.46	8.15	13.17	21.14
FCN (sup)	FCN-50	RGB	$\mathcal{F}$	61.22	71.27	56.22	74.57
HRNet (sup)	HRNet-48	RGB	$\mathcal{F}$	**62.22**	**71.24**	**60.29**	**77.74**
Ours	TwoCNN + HRNet-48	HS+RGB	$\mathcal{P}$	**62.43**	**69.52**	**55.20**	**66.18**
Ours	SVM + HRNet-48	HS+RGB	$\mathcal{P}$	57.91	65.57	50.01	60.94
Ours	TwoCNN + HRNet-48	RGB	$\mathcal{P}$	48.18	60.21	32.63	37.58
Ours	SVM + HRNet-48	RGB	$\mathcal{P}$	38.07	53.39	16.31	25.43

Table 4. Comparison with related works on LIB-HSI and HSICityV2 datasets. NOTE: HS is abbreviation for Hyperspectral

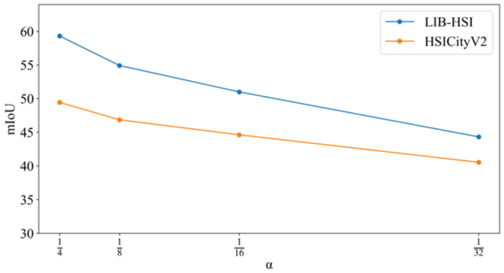
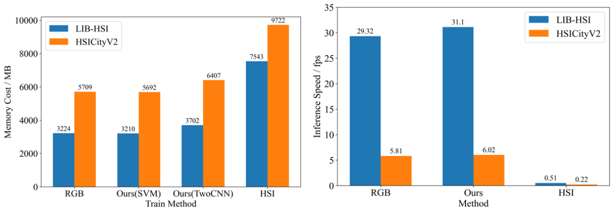


Figure 7. **Comparison of our method using different α.** All results uses TwoCNN + SLIC to generate annotations and HRNet-w48 to perform semantic segmentation.



(a) Memory Cost

(b) Inference Time

Figure 8. **Training performance of our method.** We compare the performance at two aspects: (a) max GPU memory allocation during training progress; (b) inference time on the whole test set. All experiments are conducted using SVM/TwoCNN + SLIC to generate annotations and HRNet-48 to perform semantic segmentation.

4.4. Ablation Study

Annotated Ratio. The spectral information extraction highly depends on the number of annotated points. With more supervised points, the classification model can extract more accurate spectral information, but it will certainly increase the workload of annotation. In experiments above, the ratio of training samples to validation samples α is set to 1:4. In this experiment, we test the performance of α from 1:4 to 1:32 (Figure 7). We can see that the performance of semantic segmentation increases with the ratio of training samples to validation samples. It's important to find a balance between the performance and workload of annotation.

Computational Complexity. We compared the training memory cost and inference time of our method and others. As shown in figure 8a, the training memory cost of our method is a bit higher than training using only RGB images, but much lower than training using full hyperspectral images. The reason is that our method only uses a small number of spectral points, which reduces the memory cost of training. Figure 8b shows that the inference time of our method is almost same as inference using only RGB images, but much lower than inference using full hyperspectral images. The reason is that reading full hyperspectral images from hard disk and unzipping them is very time-consuming, while our method only uses a small number of spectral bands, which reduces the inference time.

5. Conclusions

In this paper, we propose a novel and simple framework of weakly-supervised semantic segmentation for natural scenes. We adopt point-based annotations to reduce cost, and use hyperspectral information which are easier to acquire as additional information to improve the performance of point-supervised semantic segmentation. Our approach provides a new perspective for using hyperspectral information as prior knowledge in semantic segmentation, and improves the performance without increasing the cost of annotation. In the future, we will explore more effective methods to extract semantic information from HS sequences, and try to use hyperspectral information to improve the performance of other weakly-supervised and semi-supervised tasks.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[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.

[3] Zaid Al-Huda, Donghai Zhai, Yan Yang, and Riyadh Nazar Ali Algburi. Optimal scale of hierarchical image segmentation with scribbles guidance for weakly supervised semantic segmentation. *International Journal of Pattern Recognition and Artificial Intelligence*, 35(10):2154026, 2021.

[4] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54:137–178, 2021.

[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.

[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.

[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.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

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

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

[11] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation, 2022.

[12] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

[13] 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.

[14] 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.

[15] Lei Ding, Hao Tang, and Lorenzo Bruzzone. Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1): 426–435, 2020.

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[17] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Pointly-supervised panoptic segmentation. In *European Conference on Computer Vision*, pages 319–336. Springer, 2022.

[18] Nariman Habili, Ernest Kwan, Weihao Li, Christfried Webers, Jeremy Oorloff, Mohammad Ali Armin, and Lars Petersson. A hyperspectral and rgb dataset for building façade segmentation. In *European Conference on Computer Vision*, pages 258–267. Springer, 2022.

[19] 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.

[20] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[21] Lingbo Huang, Yushi Chen, and Xin He. Weakly supervised classification of hyperspectral image based on complementary learning. *Remote Sensing*, 13(24): 5009, 2021.

[22] Yuxing Huang, Tianqi Ren, Qiu Shen, Ying Fu, and Shaodi You. HSICityV2: Urban Scene Understanding via Hyperspectral Images, 2021.

[23] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation.

In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.

[24] Longlong Jing, Yucheng Chen, and Yingli Tian. Coarse-to-fine semantic segmentation from image-level labels. *IEEE transactions on image processing*, 29:225–236, 2019.

[25] Mithun Kumar Kar, Malaya Kumar Nath, and Debanga Raj Neog. A review on progress in semantic image segmentation and its application to medical images. *SN computer science*, 2(5):397, 2021.

[26] Chen Ke. Military object detection using multiple information extracted from hyperspectral imagery. In *2017 International Conference on Progress in Informatics and Computing (PIC)*, pages 124–128. IEEE, 2017.

[27] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M Atkinson. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021.

[28] 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.

[29] George C. Linderman and Stefan Steinerberger. Clustering with t-sne, provably, 2017.

[30] 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*, pages 3431–3440, 2015.

[31] 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.

[32] Jaime Moraga and H Sebnem Duzgun. Jigsawhsi: A network for hyperspectral image classification. *arXiv preprint arXiv:2206.02327*, 2022.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[34] Robert Pike, Guolan Lu, Dongsheng Wang, Zhuo Georgia Chen, and Baowei Fei. A minimum spanning forest-based method for noninvasive cancer detection with hyperspectral imaging. *IEEE Transactions on Biomedical Engineering*, 63(3): 653–663, 2015.

[35] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 90–105. Springer, 2016.

[36] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. Weakly supervised scene parsing with point-based distance metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8843–8850, 2019.

[37] Raúl Siche, Ricardo Vejarano, Victor Aredo, Lia Velasquez, Erick Saldana, and Roberto Quevedo. Evaluation of food quality and safety with hyperspectral imaging (hsi). *Food Engineering Reviews*, 8:306–322, 2016.

[38] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3136–3145, 2019.

[39] 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.

[40] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7):1562–1573, 2018.

[41] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition, 2020.

[42] Jiaqi Yang, Bo Du, Di Wang, and Liangpei Zhang. Iter: Image-to-pixel representation for weakly supervised hsi classification. *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, 2023.

[43] 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.

[44] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169:114417, 2021.

[45] Hongyi Zhang, Andreas Geiger, and Raquel Urtasun. Understanding high-level semantics by modeling traffic patterns. In *Proceedings of the IEEE international conference on computer vision*, pages 3056–3063, 2013.

[46] Man Zhang, Yong Zhou, Jiaqi Zhao, Yiyun Man, Bing Liu, and Rui Yao. A survey of semi-and weakly supervised semantic segmentation of images. *Artificial Intelligence Review*, 53:4259–4288, 2020.

[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.

[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.