Semantic Segmentation of Remote Sensing Images With Sparse Annotations

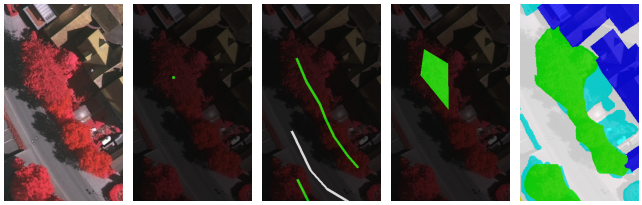
Yuansheng Hua, Diego Marcos, Lichao Mou, Xiao Xiang Zhu, *Fellow, IEEE*, and Devis Tuia, *Senior Member, IEEE*

***Abstract*— Training convolutional neural networks (CNNs) for very high-resolution images requires a large quantity of high-quality pixel-level annotations, which is extremely labor-intensive and time-consuming to produce. Moreover, professional photograph interpreters might have to be involved in guaranteeing the correctness of annotations. To alleviate such a burden, we propose a framework for semantic segmentation of aerial images based on incomplete annotations, where annotators are asked to label a few pixels with easy-to-draw scribbles. To exploit these sparse scribbled annotations, we propose the FEature and Spatial relaTional regulArization (FESTA) method to complement the supervised task with an unsupervised learning signal that accounts for neighborhood structures both in spatial and feature terms. For the evaluation of our framework, we perform experiments on two remote sensing image segmentation data sets involving aerial and satellite imagery, respectively. Experimental results demonstrate that the exploitation of sparse annotations can significantly reduce labeling costs, while the proposed method can help improve the performance of semantic segmentation when training on such annotations. The sparse labels and codes are publicly available for reproducibility purposes.**[1]

***Index Terms*— Aerial image, convolutional neural networks (CNNs), semantic segmentation, semisupervised learning, sparse scribbled annotation.**

I. Introduction

SEMANTIC segmentation of remote sensing imagery aims at identifying the land-cover or land-use category of each pixel in an image. As one of the fundamental visual tasks, semantic segmentation has been attracting wide attention in the remote sensing community and has proven to be beneficial to a variety of applications, such as land cover mapping, traffic monitoring, and urban management. Recently, many studies [1] resort to learning deep convolutional neural networks (CNNs) with full supervision for semantic segmentation and have obtained enormous achievements. However, training a fully supervised segmentation CNN requires a huge volume of dense pixel-level ground truths, which are labor-intensive and time-consuming to generate. Moreover, expert annotators might be needed for correctly identifying pixels located at object boundaries and ambiguous regions (e.g., shadows in Fig. 1).



(a) (b) (c) (d) (e)

Fig. 1. Comparisons of different levels of scribbled annotations. Trees (marked as green) are taken as an example here. Images from left to right are (a) aerial image, (b) point-, (c) line-, (d) polygon-level scribbled annotations, and (e) dense pixelwise labels.

To alleviate the requirement of dense pixelwise annotations, semisupervised learning approaches are proposed to make use of additional information, such as spatial relations (e.g., neighboring pixels are likely to belong to the same class) or feature-level relations (e.g., pixels with similar CNN feature representations are likely to belong to the same class), for semantic segmentation. These methods aim to utilize low-cost annotations, such as points [2], scribbles [3], [4], or image-level labels [5], [6]. As the first attempt, Bearman *et al.* [2] proposed to learn semantic segmentation models with point-level supervision, where only one point is labeled for each instance. In scribble-supervised algorithms, annotations are provided in the form of hand-drawn scribbles. Wu *et al.* [3] propose to learn aerial building footprint segmentation models from scribbles. Maggiolo *et al.* [4] argue that a network directly trained on scribbled ground truths fails to accurately predict object boundaries and propose to employ a fully connected conditional random field (CRF) to refine the shapes of objects. Compared to fully annotated ground truths, scribbled annotations [see Fig. 1(c)] are easier to generate in a user-friendly way. In comparison with point-level annotations [e.g., Fig. 1(b)], scribbles can provide stronger supervisory signals. However, point- and scribble-supervised segmentation methods remain underexplored in the remote sensing community. To this end, we propose a simple yet effective framework for semantic segmentation of remote sensing imagery with low-cost annotations. In this framework, we manually create point- or scribble-level annotations and train networks on them. Besides, we also evaluate polygon-level annotations [see Fig. 1(d)], which can be easily yielded and cover more pixels than the other types of annotations. Since these annotations

Manuscript received September 24, 2020; revised November 29, 2020; accepted December 22, 2020. This work was supported by the German Federal Ministry of Education and Research—AI Future Laboratory "AI4EO" under Grant 01DD20001. *(Corresponding authors: Xiao Xiang Zhu; Devis Tuia.)*

Yuansheng Hua, Lichao Mou, and Xiao Xiang Zhu are with Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany (e-mail: yuansheng.hua@dlr.de; lichao.mou@dlr.de; xiaoxiang.zhu@dlr.de).

Diego Marcos is with the Laboratory of GeoInformation Science and Remote Sensing, Wageningen University, 6708 Wageningen, The Netherlands (e-mail: diego.marcos@wur.nl).

Devis Tuia was with the Laboratory of GeoInformation Science and Remote Sensing, Wageningen University, 6708 Wageningen, The Netherlands. He is now with the Ecole Polytechnique Fédérale de Lausanne, 1950 Sion, Switzerland (e-mail: devis.tuia@epfl.ch).

Color versions of one or more figures in this letter are available at https://doi.org/10.1109/LGRS.2021.3051053.

Digital Object Identifier 10.1109/LGRS.2021.3051053

[1]https://github.com/Hua-YS/Semantic-Segmentation-with-Sparse-Labels

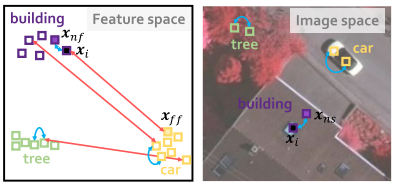


Fig. 2. Illustration of the proposed FESTA. A sample $\boldsymbol{x}_i$ belonging to *building* (filled with black) is taken as an example.

are sparsely distributed across the images, we call them sparse annotations in Sections II–IV. In order to better exploit sparse annotations, we propose a semisupervised learning method that encodes and regularizes the feature and spatial relations. To demonstrate the effectiveness of our learning framework, extensive experiments are conducted on two very high resolution (VHR) data sets: the Vaihingen and Zurich Summer.

II. Methodology

A. Supervision With Sparse Annotations

In contrast to conventional dense annotations, sparse annotations have two characteristics: 1) a very small proportion of pixels are assigned semantic classes and 2) objects do not need to be entirely annotated [see Fig. 1(b)–(d)]. This greatly reduces the effort required from the annotators, as complex boundaries and ambiguous pixels can be avoided.

Here, we consider three levels of sparse annotations: point-level, scribble-level, and polygon-level. Specifically, point-level annotations indicate that, for an annotator interaction, only one single pixel is labeled. Scribble-level annotations, also called line-level annotations, are yielded by drawing a scribble line within an object and assigning all pixels along this line the same class label. Similarly, polygon-level annotations can be generated by drawing a polygon within an object and classifying pixels located in the polygon into the same semantic class. Examples of these three levels of annotations are shown in Fig. 1.

B. Feature and Spatial Relational Regularization

When using sparse annotations, the vast majority of pixels in the training images are left unlabeled. In order to exploit both labeled and unlabeled pixels, we develop a semisupervised methodology, named FEature and Spatial relaTional regulArization (FESTA), to enable a semantic segmentation CNN to learn discriminative features while leveraging the unlabeled image pixels. An assumption shared by many unsupervised learning algorithms [7] is that nearby entities often belong to the same class. Based on this assumption, a recent work [8] achieves success in representation learning by encoding neighborhood relations in the feature space. Inspired by this work, we propose to encode and regularize relations between pixels in both feature and spatial domain, as shown in Fig. 2, so that the learned features become more useful for semantic segmentation.

Specifically, given a sample $\boldsymbol{x}_i$ (i.e., a CNN feature vector extracted from location i in an image), we first encode its relations to all other samples by measuring the distance in space and feature similarity with respect to all other features in the image. The sample with the smallest similarity is considered as the far-away sample in the feature space, $\boldsymbol{x}_{i_{ff}}$, while that with the highest similarity is defined as the neighboring sample in feature space, $\boldsymbol{x}_{i_{nf}}$. According to the aforementioned proximity assumption, it is highly probable that $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i_{nf}}$ belong to the same class, and thus, the distance between them should be as small as possible. In order to prevent a trivial solution in which all features collapse to the same point, $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i_{ff}}$ are encouraged to further increase their dissimilarity. We apply similar reasoning in the spatial domain since images are smooth in spatial terms. Thus, we take the eight spatial neighbors of $\boldsymbol{x}_i$ into consideration and chose the one most similar in feature space as the spatial neighbor, $\boldsymbol{x}_{i_{ns}}$. This operation is intended to prevent pairing $\boldsymbol{x}_i$ with a spatial neighbor that belongs to the object boundary.

These priors can be incorporated into the learning objectives by using the following loss function:

$$\mathcal{L}_{\text{FESTA}} = \alpha \sum_{i=1}^{N} \mathcal{D}\left(\boldsymbol{x}_i, \boldsymbol{x}_{i_{nf}}\right) + \beta \sum_{i=1}^{N} \mathcal{D}\left(\boldsymbol{x}_i, \boldsymbol{x}_{i_{ns}}\right) + \gamma \sum_{i=1}^{N} \mathcal{S}\left(\boldsymbol{x}_i, \boldsymbol{x}_{i_{ff}}\right) \tag{1}$$

where $\mathcal{D}$ denotes the Euclidean distance and $\mathcal{S}$ represents cosine similarity. α, β, and γ are tradeoff parameters representing the significances of the respective terms, and N represents the number of pixels in a given image. By minimizing $\mathcal{L}_{\text{FESTA}}$, $\boldsymbol{x}_{i_{nf}}$ and $\boldsymbol{x}_{i_{ns}}$ are forced to move closer to $\boldsymbol{x}_i$, while $\boldsymbol{x}_{i_{ff}}$ is pushed far from $\boldsymbol{x}_i$. In order to jointly exploit the sparse scribbled annotations and FESTA for the network training, the final loss is defined as

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{\text{FESTA}} \tag{2}$$

where $\mathcal{L}_{ce}$ indicates the categorical cross-entropy loss calculated from pixels with annotations.

C. CRF for Boundary Refinement

To further refine the predictions of networks trained on scribbled annotations, we integrate a fully connected CRF [9] into our system, and the energy function of CRF model is

$$E = \sum_i \theta_u(x_i) + \sum_{ij} \theta_p\left(x_i, x_j\right) \tag{3}$$

where $\theta_u(x_i)$ is the unary potential and calculated as $\theta_u(x_i) = -\log P(x_i)$. Here, i ranges from 0 to the number of pixels in the image, and $P(x_i)$ is the label probability of pixel i. $\theta_p(x_i, x_j)$ is utilized to measure pairwise potentials between pixel i and j. We tested with two Gaussian kernels

$$k_1 = \exp\left(-\frac{\left\|p_i - p_j\right\|^2}{2\theta_1^2} - \frac{\left\|I_i - I_j\right\|^2}{2\theta_2^2}\right)$$
$$k_2 = \exp\left(-\frac{\left\|p_i - p_j\right\|^2}{2\theta_3^2}\right) \tag{4}$$

where p_i and I_i indicate the position and color intensity of pixel i. θ_1, θ_2, and θ_3 are hyperparameters that control the kernel "scale". In (4), k_1 is known as appearance kernel and tends to classify neighboring pixels with similar appearances [10], i.e., color intensities, into the same classes, while k_2, so-called smoothness kernel, penalizes pixels nearby but assigned diverse labels. This step is expected to make the class map smoother within homogeneous areas.

III. Experimental Results

A. Data Set Description

The Vaihingen data set[2] is a benchmark data set for semantic segmentation provided by the International Society for Photogrammetry and Remote Sensing (ISPRS); 33 aerial images with a spatial resolution of 9 cm were collected over the city of Vaihingen, and each image covers an average area of 1.38 km^2. For each aerial image, three bands are available, near-infrared (NIR), red (R), and green (G). Besides, coregistered digital surface models (DSMs) are provided for all images; 16 images are fully annotated. In total, six land-cover classes are considered: impervious surface, building, low vegetation, tree, car, and clutter/background. In this letter, we follow the train-test split scheme in most existing works [11], [12] and select five images (image IDs: 11, 15, 28, 30, and 34) as the test set. The remaining ones are utilized to train our models.

The Zurich Summer data set [13] is composed of 20 images that are taken over the city of Zurich in August 2002 by the QuickBird satellite. The spatial resolution is 0.62 m, and the average size of images is 1000×1150 pixels. The images consist of four channels: NIR, red (R), green (G), and blue (B). Following previous works [14], [15], we only utilize NIR, R, and G in our experiments and train our model on 15 images; the others (image IDs: 16, 17, 18, 19, and 20) are utilized to test. In total, there are eight urban classes, including road, building, tree, grass, bare soil, water, railway, and swimming pool. Uncategorized pixels are labeled as background.

It is noteworthy that although full pixelwise annotations are provided for all images in the Vaihingen and Zurich Summer data sets, we only use them in the test phase to calculate evaluation metrics. The training of all models is done with scribbled annotations described in the following.

B. Scribbled Annotation Generation

To annotate large-scale images, we employ an online labeling platform, LabelMe,[3] and ask annotators to draw by following these rules: 1) for each class, annotations are supposed to cover diverse appearances (see region (a)–(c) in Fig. 3, where cars of different colors are annotated) and be located in different positions of the image separately and 2) polygon- and line-level annotations are not required to delineate object boundaries precisely [see the annotations of trees in Fig. 1(c) and (d)]. In order to make the time spent on each level of scribbled annotations more equivalent, we ask four annotators (including two nonexperts) to label 7, 5, and 3 objects per class for point-, line- and polygon-level annotations in each aerial image. As a consequence, sparse but accurate annotations can be provided rapidly without effort. Since a point- or line-level annotation is often located in the center area of an object and distant from its boundary, we perform morphological dilation on all point- and line-level annotations with a disk of radius 3. Afterward, pixels involved in dilated annotations are assigned the same class labels as their central points or lines. For polygon-level annotations, pixels within each polygon are assigned the corresponding classes.

Table I shows the average amounts of pixels with sparse and dense annotations in both data sets. It can be seen that sparse annotations are several orders of magnitude fewer than dense annotations. As to the labeling time, it took on average 133, 126, and 161 s per image to produce point-, line- and polygon-level annotations, respectively, for the Vaihingen data set and 177, 162, and 238 s per image for the Zurich Summer data set. In Section III-D, we demonstrate the proposed method allows improving the semantic segmentation results using these sparse annotations. In Section III-D, we discuss the differences observed among the tested annotation types.

TABLE I

Total Numbers of Pixels Labeled With Sparse Point-, Line-, and Polygon-Level Annotations (Middle Three Columns) and Dense Annotations (Right Column) in the Vaihingen and Zurich Summer Data Sets

Dataset Name	Point	Line	Polygon	Dense*
Vaihingen	18,787	480,593	4,591,409	54,373,518
Zurich Summer	29,508	330,767	1,445,270	12,266,287

*Background/Clutter is not considered.

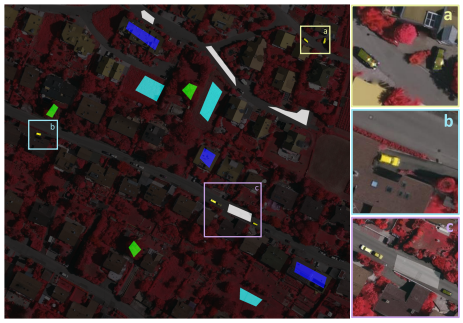


Fig. 3. Example polygon-level annotations of an image (ID: 13) on the Vaihingen data set. Annotations of cars are zoomed in to illustrate that annotations should include variant visual appearances for one class. Legend—white: impervious surfaces; blue: buildings; cyan: low vegetation; green: trees; and yellow: cars.

C. Training Details

We segment the images with a standard fully convolutional network (FCN) (i.e., FCN-16s [17]) and initialize convolutional layers with Glorot uniform [18] initializers. Specifically, VGG-16 is taken as the backbone, and the outputs of the last two convolutional blocks are upsampled to the original resolution and fused with an elementwise addition. The fused feature maps are finally fed into a convolutional layer, where the number of filters is equivalent to the number of classes. In the training phase, all weights are trainable and updated with Nesterov Adam [19], using $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1\text{e}{-}08$ as recommended. We initialize the learning rate as $2\text{e}{-}04$ and let it decay by a factor of 10 when the validation loss is saturated. To train the network, we define the loss as (2), and λ is set experimentally to 0.1 and 0.01 for the Vaihingen and Zurich Summer data sets, respectively. Tradeoff parameters, α, β, and γ, are set as 0.5, 1.5, and 1, to ensure that: 1) the regularizers governing feature and spatial relations are balanced and 2) neighboring pixels in the image space receive more attention. The network, as well as FESTA, is implemented on TensorFlow and trained on one NVIDIA Tesla P100 16-GB GPU for 100k iterations. The size of the minibatch is set as five during the training procedure. In the training phase, we use a sliding window to crop training images into 256×256 patches, and its stride is set to 64 pixels. Besides, no class-dependent configurations are considered. In the test phase, we employ dense CRF to refine predictions before calculating metrics. We tuned the

[2]http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html

[3]http://labelme.csail.mit.edu/Release3.0/

TABLE II

NUMERICAL RESULTS ON THE VAIHINGEN DATA SET (%): WE SHOW THE PER-CLASS F_1 SCORE, MEAN F_1 SCORE, AND OA ON THE TEST SET. MEAN AND STANDARD DEVIATION OF EACH METRIC ARE CALCULATED FROM RESULTS ON SPARSE ANNOTATIONS PRODUCED BY FOUR ANNOTATORS. RESULTS ON DENSE ANNOTATIONS ARE PROVIDED AS REFERENCE

Scribble	Model	Imp. surf.	Build.	Low veg.	Tree	Car	mean F_1	OA
Point	FCN-WL [16]	69.81 ± 1.52	75.02 ± 2.32	60.25 ± 3.40	76.17 ± 1.42	12.29 ± 3.60	58.71 ± 0.33	67.11 ± 0.97
	FCN+dCRF [4]	75.37 ± 0.93	**81.37** ± 3.10	**61.93** ± 5.54	**78.50** ± 1.69	17.51 ± 6.70	62.94 ± 0.44	72.53 ± 0.42
	FCN-FESTA	74.65 ± 2.73	78.64 ± 4.74	60.24 ± 3.33	76.15 ± 2.07	23.65 ± 4.24	62.66 ± 2.54	71.43 ± 2.93
	FCN-FESTA+dCRF	**77.62** ± 1.93	80.08 ± 5.27	60.78 ± 4.00	76.70 ± 2.00	**31.40** ± 5.24	**65.32** ± 2.56	**73.65** ± 2.52
Line	FCN-WL [16]	78.44 ± 3.24	83.45 ± 1.58	64.02 ± 2.34	79.32 ± 0.54	29.01 ± 2.96	66.85 ± 1.81	76.12 ± 1.52
	FCN+dCRF [4]	**81.32** ± 2.45	**84.88** ± 1.88	63.71 ± 3.92	79.88 ± 1.33	38.95 ± 4.50	69.75 ± 2.23	**78.03** ± 1.82
	FCN-FESTA	78.12 ± 3.92	83.76 ± 2.00	**65.78** ± 1.88	**80.49** ± 0.93	38.24 ± 10.31	69.28 ± 3.66	77.24 ± 2.27
	FCN-FESTA+dCRF	80.06 ± 3.32	84.47 ± 2.23	64.35 ± 2.38	80.32 ± 0.92	**43.72** ± 9.62	**70.58** ± 3.42	77.99 ± 2.14
Polygon	FCN-WL [16]	76.71 ± 3.63	80.03 ± 1.42	59.40 ± 6.09	78.50 ± 2.86	26.28 ± 11.06	64.19 ± 4.40	74.18 ± 2.97
	FCN+dCRF [4]	78.37 ± 3.08	80.85 ± 1.13	57.92 ± 7.67	78.67 ± 2.87	29.13 ± 8.15	64.99 ± 3.99	75.15 ± 2.94
	FCN-FESTA	78.98 ± 3.82	83.10 ± 2.62	**62.59** ± 4.89	**79.91** ± 3.31	33.04 ± 7.71	67.52 ± 4.07	76.65 ± 3.39
	FCN-FESTA+dCRF	**80.62** ± 3.22	**83.62** ± 2.29	60.79 ± 5.04	79.81 ± 2.52	**40.27** ± 8.30	**69.02** ± 4.01	**77.32** ± 2.92
Dense	FCN [17]	88.67	92.83	76.32	74.21	86.67	83.74	86.51

TABLE III

NUMERICAL RESULTS ON THE ZURICH SUMMER DATA SET (%): WE SHOW THE PER-CLASS F_1 SCORE, MEAN F_1 SCORE, AND OA ON THE TEST SET. MEAN AND STANDARD DEVIATION OF EACH METRIC ARE CALCULATED FROM RESULTS ON SPARSE ANNOTATIONS PRODUCED BY FOUR ANNOTATORS. RESULTS ON DENSE ANNOTATIONS ARE PROVIDED AS REFERENCE

Scribble	Model	Road	Build.	Tree	Grass	Soil	Water	Rail.	Pool	mean F_1	OA
Point	FCN-WL [16]	69.74±3.98	78.94±3.01	82.33±2.55	82.20±2.40	53.37±7.03	87.87±1.40	0.81±1.42	48.89±9.42	63.02±2.14	77.38±2.73
	FCN+dCRF [4]	**72.13**±4.99	**80.71**±1.84	82.87±2.08	83.55±2.07	63.92±8.90	92.71±1.26	**2.09**±4.17	59.96±14.60	67.24±1.93	**80.03**±2.26
	FCN-FESTA	70.64±3.44	77.34±4.13	**82.91**±2.48	83.73±2.34	56.67±5.64	89.67±2.25	0.94±1.89	73.62±4.06	66.94±2.56	78.17±3.00
	FCN-FESTA+dCRF	71.23±2.61	77.71±3.17	82.81±1.99	**84.18**±1.96	**66.34**±3.69	**93.40**±1.81	0.00±0.00	**77.38**±8.87	**69.05**±1.15	79.11±2.14
Line	FCN-WL [16]	73.00±4.60	**81.17**±3.77	**82.82**±2.78	81.88±1.41	67.02±8.77	90.98±1.79	1.19±1.60	58.77±7.82	67.10±2.02	**79.75**±2.25
	FCN+dCRF [4]	71.71±4.83	79.22±4.01	81.22±3.06	80.43±2.10	**71.72**±9.20	84.65±14.90	**2.35**±4.71	67.58±17.39	68.39±3.10	78.84±2.15
	FCN-FESTA	**73.34**±3.88	79.08±3.60	82.71±2.10	**84.27**±1.41	60.67±13.36	92.37±1.44	1.02±0.83	74.27±8.24	68.47±2.45	79.52±2.86
	FCN-FESTA+dCRF	71.74±2.78	75.81±4.18	81.20±1.60	83.44±1.51	66.49±15.57	**94.68**±0.52	0.00±0.00	**82.06**±6.80	**69.43**±2.57	78.51±2.21
Polygon	FCN-WL [16]	64.18±6.14	72.17±6.01	79.64±4.25	77.10±3.92	49.17±16.96	89.26±3.52	1.31±1.09	76.90±6.33	63.72±4.35	73.09±4.49
	FCN+dCRF [4]	62.63±5.77	70.35±4.88	78.30±3.53	75.94±4.42	52.11±14.06	91.03±4.39	0.84±1.69	**85.13**±2.72	64.54±4.08	72.37±3.89
	FCN-FESTA	**66.53**±5.07	**74.06**±3.06	**80.05**±3.66	**79.42**±3.56	57.83±11.38	90.80±2.42	5.87±4.86	65.68±16.06	65.03±1.98	**75.00**±3.17
	FCN-FESTA+dCRF	65.10±4.42	71.96±2.76	79.44±3.26	78.87±4.58	**61.86**±9.72	**92.50**±2.96	**6.37**±6.63	77.21±6.63	**66.66**±2.41	74.41±2.86
Dense	FCN [17]	88.34	93.27	92.40	89.48	67.96	96.87	2.98	88.10	77.42	90.51

parameters of dense CRF [θ_1, θ_2, and θ_3 in (4)] on validation images and find that satisfactory results can be achieved for both FCN and FCN-FESTA when setting them to 30, 10, and 10, respectively. In the case of large homogeneous areas of an image belonging to the same class, α should be set to a small value, which encourages the network to focus more on geographically nearby samples. Besides, large batch size and sliding window can also help alleviate the influence of such a scenario.

D. Comparing With Existing Methods

We compare an FCN [17] learned using the proposed FESTA (FCN-FESTA) against an FCN learned with weighted loss function (FCN-WL) [16] on sparse annotations. We also report segmentation results of the baseline FCN trained on dense labels. In addition, we study the influence of the fully connected CRF by comparing FCN-FESTA+dCRF and FCN+dCRF [4]. Each model is trained and validated on sparse annotations independently. Per-class F_1 scores, mean F_1 scores, and overall accuracy (OA) are calculated on test images with dense annotations. Considering that each model is learned on labels from four annotators, respectively, we average metrics obtained by each annotator and report them in the form of mean ± standard deviation.

Table II exhibits numerical results on the Vaihingen data set. FCN-FESTA+dCRF achieves the highest mean F_1 scores in training with all kinds of scribbled annotations, which demonstrates its effectiveness. To be more specific, with the point- and polygon-level supervision, FCN-FESTA improves the mean F_1 score by 3.95% and 3.33% compared to FCN-WL, respectively. By refining predictions with dense CRF, FCN-FESTA + dCRF achieves improvements of 2.38% and 4.03% in comparison with FCN + dCRF. It is interesting to observe that line-level scribbles improve the segmentation performance the most, and FCN-FAST + dCRF learned with such annotations obtains the highest mean F_1 score, 70.58%. Moreover, we note that FESTA can enhance the network capability of recognizing small objects, i.e., car, in high-resolution aerial images. Example segmentation results of networks trained online annotations are visualized in Fig. 4.

Numerical results on the Zurich Summer data set are shown in Table III. As can be seen, FESTA contributes to increments of 3.92%, 1.37%, and 1.31% in the mean F_1 score when training with point-, line-, and polygon-level annotations. By utilizing line annotations and dense CRF, FCN-FESTA + dCRF obtains the highest mean F_1 score, 69.43%. Besides, we note that the exploitation of dense CRF plays a significant role in improving the results of networks trained on point-level scribbles. Example visual results of networks trained online annotations are shown in Fig. 5. In our experiments, we also train networks with multiclass dice loss and find that results are comparative to those learned with cross-entropy loss.

E. Discussion on Annotation Type

To further study the influence of annotations, we also train baseline FCNs on dense annotations and report numerical results in Tables II and III. As shown in Tables II and III, line-level annotations lead to the best performance on both data sets, even though the number of labeled pixels is an order of magnitude smaller than polygon annotations (see Table I). Although it was expected that line annotations would outperform point annotations, due to their ability to capture within-object variations, we were surprised to see that they also outperformed polygon annotations. We suspect that this

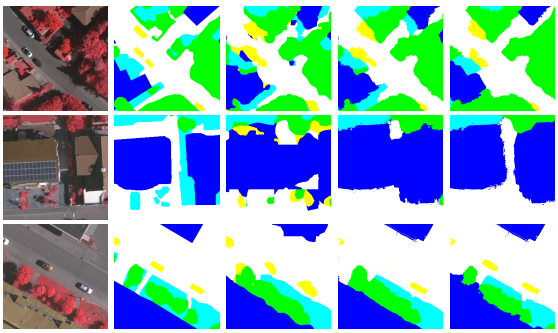


Fig. 4. Examples of segmentation results on the Vaihingen data set. All models are trained on line annotations. The legend is the same as that in Fig. 3. (a) Image. (b) Dense GT. (c) FCN-WL. (d) FCN+dCRF. (e) Ours.

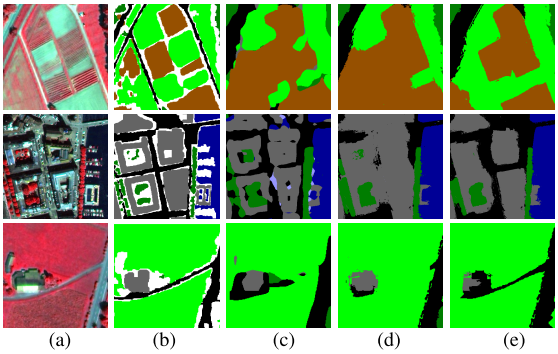


Fig. 5. Examples of segmentation results on the Zurich Summer data set. All models are trained on line annotations. Legend—black: road; brown: soil; green: grass; dark green: tree; gray: buildingl; and white: background. (a) Image. (b) Dense GT. (c) FCN-WL. (d) FCN+dCRF. (e) Ours.

is linked to the fact that the number of pixels per object grows quadratically for polygons and linearly for lines. This would lead to a more balanced weighing of differently sized objects in the case of line annotations and an underweighting of smaller objects in the case of polygon annotations, which could harm the model's performance. Another reason could be that, since drawing a line is faster than drawing a polygon, annotators for the line features provided more scribbles in the same time budget.

In spite of the mean F_1 performance boost provided by FESTA, there is still a large gap with respect to the FCN model trained with dense ground truths of 13% in Vaihingen and 8% in Zurich. This gap is, however, not evenly distributed across the classes. The gap is smaller or nonexistent in classes, such as water, tree, grass, or soil, which are often homogeneous in terms of materials. On the contrary, it is larger for classes with more diverse materials (and therefore, observed spectral values), such as building and car (in the Vaihingen data set). It is noteworthy to mention that the class railway, in the Zurich data set, is systematically missed in all cases, including the densely supervised FCN.

IV. Conclusion

In this letter, we propose a simple yet efficient framework for semantic aerial image segmentation using sparse annotations and a semisupervised learning objective. In order to validate the effectiveness of our approach, we conduct experiments on the Vaihingen and Zurich Summer data sets. Numerical and visual results suggest that the proposed method contributes to the improvement of semantic segmentation results using several kinds of sparse annotations. Although models learned on sparse annotations achieve relatively lower accuracies than those using dense annotations, we show that using a semi-supervised deep learning approach can help to close this performance gap while leveraging sparse annotations that can significantly reduce the costs of label generation. As future work, the proposed framework can be further improved by introducing graph-based models and prior knowledge learned from label semantics.

Acknowledgment

The authors would like to thank Yingya Xu, Li Hua, and Yanping Tang for contributing to this work with annotation.

References

[1] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[2] A. Bearman, O. Russakovsky, V. Ferrari, and F. Li, "What's the point: Semantic segmentation with point supervision," in *Proc. ECCV*, 2016, pp. 549–565.

[3] W. Wu, H. Qi, Z. Rong, L. Liu, and H. Su, "Scribble-supervised segmentation of aerial building footprints using adversarial learning," *IEEE Access*, vol. 6, pp. 58898–58911, 2018.

[4] L. Maggiolo, D. Marcos, G. Moser, and D. Tuia, "Improving maps from CNNs trained with sparse, scribbled ground truths using fully connected CRFs," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 2099–2102.

[5] A. Nivaggioli and H. Randrianarivo, "Weakly supervised semantic segmentation of satellite images," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, May 2019, pp. 1–4.

[6] R. Zhu, L. Yan, N. Mo, and Y. Liu, "Semi-supervised center-based discriminative adversarial learning for cross-domain scene-level land-cover classification of aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 155, pp. 72–89, Sep. 2019.

[7] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[8] M. Sabokrou, M. Khalooei, and E. Adeli, "Self-supervised representation learning via neighborhood-relational encoding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8010–8019.

[9] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. NeurIPS*, 2011, pp. 109–117.

[10] K. Schindler, "An overview and comparison of smooth labeling methods for land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4534–4545, Nov. 2012.

[11] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7092–7103, Dec. 2017.

[12] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*. [Online]. Available: http://arxiv.org/abs/1606.02585

[13] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 1–9.

[14] D. Tuia, M. Volpi, and G. Moser, "Decision fusion with multiple spatial supports by conditional random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3277–3289, Jun. 2018.

[15] C. Wendl, D. Marcos, and D. Tuia, "Novelty detection in very high resolution urban scenes with density forests," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, May 2019, pp. 1–4.

[16] Ö. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*, 2016, pp. 424–432.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[18] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.

[19] T. Dozat, "Incorporating Nesterov momentum into Adam," in *Proc. Int. Conf. Learn. Represent. Workshop*, May 2016, p. 4.