

Semantic Segmentation by Semantic Proportions

Halil Ibrahim Aysel, Xiaohao Cai, Adam Prugel-Bennett

Abstract—Semantic segmentation is a critical task in computer vision aiming to identify and classify individual pixels in an image, with numerous applications in for example autonomous driving and medical image analysis. However, semantic segmentation can be highly challenging particularly due to the need for large amounts of annotated data. Annotating images is a time-consuming and costly process, often requiring expert knowledge and significant effort; moreover, saving the annotated images could dramatically increase the storage space. In this paper, we propose a novel approach for semantic segmentation, requiring the rough information of individual semantic class proportions, shortened as *semantic proportions*, rather than the necessity of ground-truth segmentation maps. This greatly simplifies the data annotation process and thus will significantly reduce the annotation time, cost and storage space, opening up new possibilities for semantic segmentation tasks where obtaining the full ground-truth segmentation maps may not be feasible or practical. Our proposed method of utilising semantic proportions can (i) further be utilised as a booster in the presence of ground-truth segmentation maps to gain performance without extra data and model complexity, and (ii) also be seen as a parameter-free plug-and-play module, which can be attached to existing deep neural networks designed for semantic segmentation. Extensive experimental results demonstrate the good performance of our method compared to benchmark methods that rely on ground-truth segmentation maps. Utilising semantic proportions suggested in this work offers a promising direction for future semantic segmentation research¹.

Index Terms—Semantic segmentation, semantic proportions, deep neural networks.

I. INTRODUCTION

SEMANTIC segmentation is the task of partitioning an image into different regions depending on their semantic classes/categories. It is widely used in a variety of fields such as autonomous driving [1], medical imaging [2], [3], augmented reality [4] and robotics [5]. Impressive improvements have been shown in those areas with the recent development of deep neural networks (DNNs), benefiting from the availability of extensive annotated segmentation datasets at a large scale [6], [7]. However, creating such datasets can be expensive and time-consuming due to the usual need to annotate pixel-wise labels as it takes between 54 and 79 seconds per object [8], thus requiring a couple of minutes per image with a few objects. Moreover, requiring full supervision is rather impractical in some cases, for example, in medical imaging where expert knowledge is required. Annotating 3D data for semantic segmentation is even more costly and time-consuming due to the additional complexity and dimensionality of the data,

The authors are with the School of Electronics and Computer Science, University of Southampton, UK. (Email: hia1v20@soton.ac.uk; x.cai@soton.ac.uk; and apb@ecs.soton.ac.uk)

¹Code available at https://github.com/Halilbrahimaysel/Semantic_Segmentation_by_Semantic_Proportions

which generally requires voxel (i.e., point in 3D space) annotation. Skilled annotators from outsourcing companies that are dedicated to data annotation may be needed for specific requests to ensure annotation accuracy and consistency, adding further to the cost [9]. In addition, saving the annotated data could also be expensive given the substantial amount of storage space generally needed.

Different approaches have been proposed to reduce the fine-grained level (e.g. pixel-wise) annotation costs. One line of research is to train segmentation models in a weakly supervised manner by requiring image-level labels [10], [11], scribbles [12], eye tracks [13], or point supervision [8], [14] rather than costly segmentation masks of individual semantic classes. In contrast, in this paper we propose to utilise the proportion (i.e., percentage information) of each semantic class present in the image for semantic segmentation. For simplicity, we call this type of annotation *semantic (class) proportions* (SP). To the best of our knowledge, this is the first time of utilising SP for semantic segmentation. This innovative way, different from the existing ways (see e.g. Figure 1), could significantly simplify and reduce the human involvement required for data annotation and storage space in semantic segmentation. Our proposed approach by utilising the SP annotation can achieve comparable and sometimes even better performance in comparison to benchmark methods with full supervision utilising ground-truth segmentation masks. Moreover, we show that our method can sometimes provide free performance improvement in the presence of ground-truth maps as it can be served as a plug-and-play module, which can easily be added on top of existing DNNs trained for segmentation tasks.

Our main contributions are: i) propose a new semantic segmentation methodology and a plug-and-play module, utilising SP annotations; ii) conduct extensive experiments on representative benchmark datasets from distinct fields to demonstrate the effectiveness and robustness of the proposed approach; and iii) draw an insightful discussion for semantic segmentation with weakly annotated data and future directions.

II. RELATED WORK

Supervision levels in semantic segmentation. In recent years, more and more researchers have focused on reducing the annotation cost for semantic segmentation tasks. One way is to use weakly supervised learning techniques that require less precise or less expensive forms of supervision. For instance, the work in [11] proposed to utilise image-level labels, the work in [15], [16] used bounding boxes, and the methods in [12], [17] fed scribbles as labels instead of precise annotations to conduct semantic segmentation. Those approaches can significantly reduce the annotation cost, as they require less manual effort

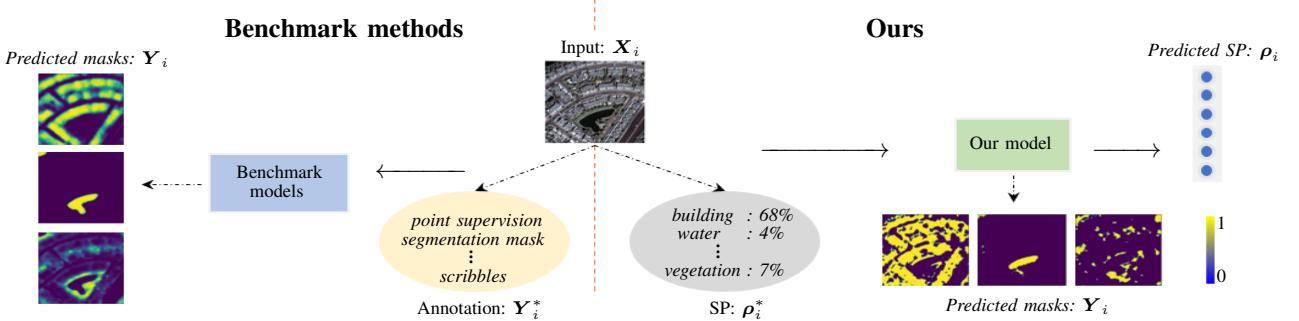


Fig. 1. Difference between the proposed semantic segmentation approach and benchmark methods.

to annotate the data. However, there is always a trade-off between the annotation cost and the model performance, i.e., models trained with higher levels of supervision generally perform better than weakly supervised models. Active learning is an alternative approach to reduce the annotation cost by selecting the most informative samples to annotate based on the current model's uncertainty. With the selected most informative samples, active learning can reduce the amount of data that needs to be labelled, thus reducing the annotation cost [18], [19]. It is worth mentioning that this is actually similar to the way we propose for the SP degraded by clustering presented in Section V-B2. Reducing the annotation cost could also be achieved by generating synthetic data that can be used to augment the real-world data [20]. Synthetic data can be generated using e.g. computer graphics or other techniques to simulate realistic images and labels.

DNNs for semantic segmentation. The work in [21] made a breakthrough by proposing fully convolutional networks (FCNs) for semantic segmentation. FCNs utilise convolutional neural network (CNN) to transform input images into a probability map, where each entry of the probability map represents the likelihood of the corresponding image pixel belonging to a particular class. This approach allows the model to learn spatial features and eliminate the need for hand-crafted features. Following FCN, several variants have been proposed to improve the segmentation performance. For example, SegNet [22] is a modification of FCN employing an encoder-decoder architecture to achieve better performance; and DeepLab [23] introduced a novel technique called atrous spatial pyramid pooling to capture multi-scale information from the input image. U-Net [24], one of the architectures used in our proposed methodology, is a type of CNN consisting of a contracting path and an expansive path. The skip connections in U-Net allow the network to retain and reuse high-level feature representations learned in the contracting path, helping to improve segmentation accuracy. The U-Net architecture has been widely used for biomedical image segmentation tasks such as cell segmentation [25], organ segmentation [26] and lesion detection [27], [28], due to its ability to accurately segment objects within images while using relatively few training samples. Furthermore, its modular architecture and efficient training make it adaptable to a wide range of segmentation tasks. Therefore, to demonstrate our methodology utilising SP,

we employ a modified and relatively basic version of the U-Net architecture as the backbone of our models for most of the experiments.

III. METHODOLOGY

Notation. Let \mathcal{X} be a set of images. Without loss of generality, we assume each image in \mathcal{X} contains no more than C semantic classes. $\forall \mathbf{X}_i \in \mathcal{X}, \mathbf{X}_i \in \mathbb{R}^{M \times H}$, where $M \times H$ is the image size. Let $\mathcal{X}_T \subset \mathcal{X}$ and $\mathcal{X}_V \subset \mathcal{X}$ be the training and validation (test) sets, respectively; and let $\Omega_T \subset \mathbb{N}$ be the set containing the indexes of the images in \mathcal{X}_T . $\forall \mathbf{X}_i \in \mathcal{X}_T$, annotations are available. The most general annotation is the ground-truth segmentation maps, say $\{\mathbf{Y}_{ij}^*\}_{j=1}^C$, for \mathbf{X}_i , where each $\mathbf{Y}_{ij}^* \in \mathbb{R}^{M \times H}$ is a binary mask for the semantic class j of \mathbf{X}_i . For simplicity, let \mathbf{Y}_i^* be a tensor formed by $\{\mathbf{Y}_{ij}^*\}_{j=1}^C$, where its j -th channel is \mathbf{Y}_{ij}^* . Note that the ground-truth segmentation maps are not required in our approach for semantic segmentation in this paper unless specifically stated; instead, they are mainly used by benchmark methods for the comparison purpose. Analogously, let \mathbf{Y}_i be the predicted segmentation maps following the same format as \mathbf{Y}_i^* . Let $\rho_i^* = (\rho_{i1}^*, \dots, \rho_{iC}^*)$ be the given SP annotation of image $\mathbf{X}_i \in \mathcal{X}_T$, which will be mainly used to train our approach, where each $\rho_{ij}^* \in [0, 1]$ is the SP of the j -th semantic class of \mathbf{X}_i and $\sum_{j=1}^C \rho_{ij}^* = 1$.

Loss function. Two types of loss functions are introduced in the architectures of our method. One is based on the mean squared error (MSE). MSE is commonly used to evaluate the performance of regression models where there are numerical target values to predict. We employ MSE to measure the discrepancy between the ground-truth SP and the predicted ones. For ease of reference, we call this loss function \mathcal{L}_{sp} throughout the paper, i.e.,

$$\mathcal{L}_{sp} = \frac{1}{|\Omega_T|} \sum_{i \in \Omega_T} \|\rho_i^* - \rho_i\|^2, \quad (1)$$

where ρ_i is the predicted SP for image $\mathbf{X}_i \in \mathcal{X}_T$ and $|\Omega_T|$ is the cardinality of set Ω_T . The other loss function, which will be deferred in Section III-B, is defined based on the binary cross-entropy (BCE). BCE is a commonly used loss function in binary classification problems and measures the discrepancy

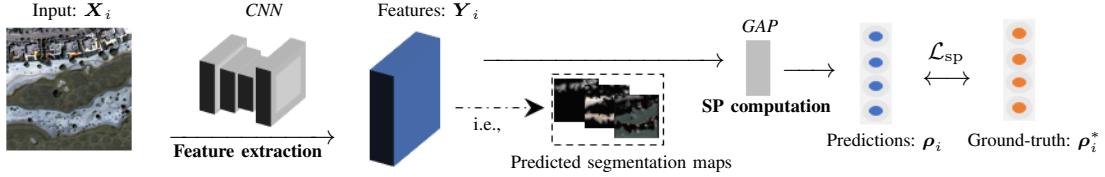


Fig. 2. The SPSS (SP-based semantic segmentation) architecture. In the training stage, features are firstly extracted by a CNN from the input; and then the extracted features are through a GAP layer calculating the SP. After training using the loss function \mathcal{L}_{sp} , the proposed SPSS architecture can force the extracted features to be the prediction of the class-wise segmentation masks.

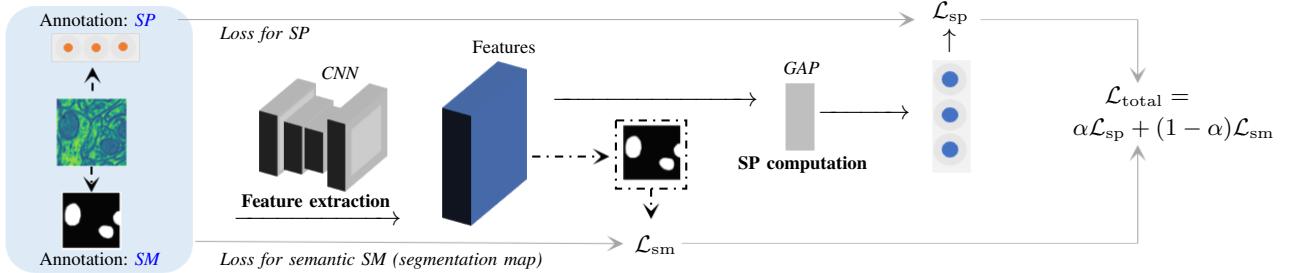


Fig. 3. The SPSS+ architecture (*cf.* the SPSS architecture in Figure 2). In contrast, \mathcal{L}_{total} (see Eq. (3)), a weighted average of \mathcal{L}_{sp} and \mathcal{L}_{sm} , is calculated during training. After training, the SPSS+ architecture can force the extracted features to be the prediction of the class-wise segmentation masks.

between the predicted probabilities and the true binary ones. Below we define the BCE function as

$$\mathcal{L}_{sm} = \frac{1}{|\Omega_T|} \sum_{i \in \Omega_T} \sum_{j=1}^C -(\mathbf{Y}_{ij}^* \log(\mathbf{Y}_{ij}) + (1 - \mathbf{Y}_{ij}^*) \log(1 - \mathbf{Y}_{ij})), \quad (2)$$

where \mathbf{Y}_{ij} is the predicted segmentation map for the j -th semantic class of image $X_i \in \mathcal{X}_T$.

A. Proposed SP-based Semantic Segmentation Architecture

The proposed SP-based semantic segmentation (SPSS) architecture is shown in Figure 2. It contains two main parts. The first part of the SPSS architecture is feature extraction. Employing a CNN is a common approach in current state-of-the-art semantic segmentation methods. In our SPSS, a CNN (or other type of DNNs) is utilised as its backbone to extract high-level image features \mathbf{Y}_i from the input image X_i . The second part of the SPSS architecture is a global average pooling (GAP) layer, which takes the image features \mathbf{Y}_i to generate the SP, ρ_i , for the input image X_i . The SPSS architecture is then trained by using the loss function \mathcal{L}_{sp} defined in Eq. (1). After training the SPSS architecture, the extracted features \mathbf{Y}_i of the trained CNN are, surprisingly, the prediction of the class-wise segmentation masks; that is how the SPSS architecture performs semantic segmentation by just using the SP rather than the ground-truth segmentation maps.

We remark that both parts in the SPSS architecture except for utilising SP are well-known and commonly employed for e.g. computer vision tasks. To the best of our knowledge, it is, for the first time, to combine them for semantic segmentation in reducing the need of labour-intensive (fine-grained) ground-truth segmentation masks to the (coarse-grained) SP level.

B. A Booster: SPSS+

The proposed SPSS architecture in Figure 2 only uses the SP annotation for semantic segmentation, which is quite cheap in terms of annotation generation. Moreover, SPSS is also very flexible. For example, i) the proposed loss function \mathcal{L}_{sp} using SP can be employed as a plug-and-play module in different DNNs; and ii) SPSS can be enhanced directly when additional annotation information is available. Below we give a showcase regarding how to use SP and pixel-level annotations jointly to enhance the SPSS architecture, see Figure 3. For ease of reference, we call the proposed booster in Figure 3 SPSS+.

The total loss for the SPSS+ architecture is

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{sp} + (1 - \alpha) \mathcal{L}_{sm}, \quad (3)$$

where α is an adjustable weight to determine the trade-off between \mathcal{L}_{sp} and \mathcal{L}_{sm} . The SPSS+ architecture uses the loss \mathcal{L}_{total} , which considers the annotations of the SP and segmentation masks for training. Similar to the SPSS architecture (in Figure 2), the extracted features \mathbf{Y}_i of the trained CNN in the SPSS+ architecture are the prediction of the class-wise segmentation masks, i.e., the semantic segmentation results.

Our SPSS can generally achieve comparable performance against benchmark semantic segmentation methods. SPSS+ works as a performance booster and improves the segmentation ability of SPSS without extra training data or model complexity. More details regarding the extensive validation and comparison are given in Section V.

IV. DATA AND SETTINGS

A. Data

The proposed SP-based methodology for semantic segmentation is showcased on four different datasets described below.

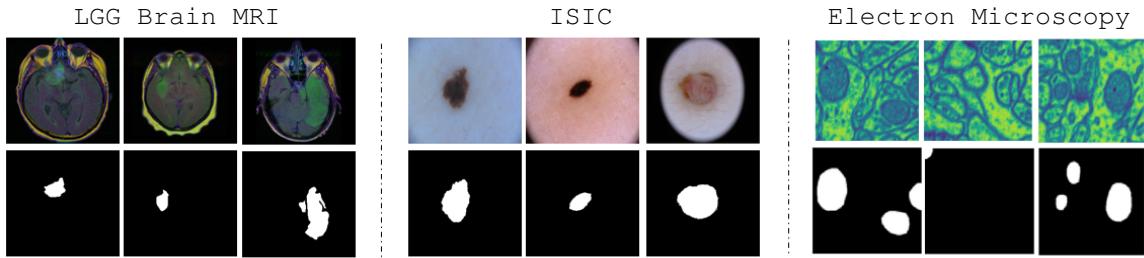


Fig. 4. Example images and ground-truth segmentation masks of the three employed medical imaging datasets.

(i) Satellite images of Dubai, i.e., Aerial Dubai. This is an open-source aerial imagery dataset presented as part of a Kaggle competition². The dataset includes 8 tiles and each tile has 9 images of various sizes and their corresponding ground-truth segmentation masks for 6 classes, i.e., *building, land, road, vegetation, water and unlabeled*.

(ii) Medical imaging dataset ISIC (International Skin Imaging Collaboration). This is a comprehensive collection of dermoscopic images specifically curated for the study and analysis of skin lesions [29], [30]. It contains 2594 training, 100 validation and 1,000 test images with high-resolution capturing various types of skin lesions, including benign and malignant conditions. Each image in the dataset is accompanied by expert annotations including detailed segmentation masks outlining the precise boundaries of the lesions. These annotations are crucial for segmentation methods to accurately delineate the lesion from the surrounding skin. The ISIC dataset is frequently used in research and competitions, such as the ISIC Challenge, to benchmark and advance segmentation algorithms. However, obtaining fine-grained pixel-level segmentation masks is expensive and our SPSS model shows comparable performance despite being trained with dramatically less expensive SP rather than full masks in Section IV in the main paper.

(iii) Medical imaging dataset Electron Microscopy³. It contains 165 slices of microscopy images with the size of 768×1024 . The primary aim of this medical dataset is to identify and classify mitochondria pixels. This dataset is quite challenging since its semantic classes are severely imbalanced, i.e., the size of the mitochondria in most slices is very small (e.g. see the right column of Figure 4 and Figure 7).

(iv) Medical imaging dataset LGG Brain MRI from The Cancer Genome Atlas (TCGA) and The Cancer Imaging Archive (TCIA). We used the version made available by Buda et al. [31] on Kaggle⁴, where the authors selected 120 patients from the TCGA lower-grade glioma collection⁵ which had available preoperative imaging data including at least a fluid-attenuated inversion recovery (FLAIR) sequence. The dataset includes roughly 4000 brain MRI images of 110 patients from 5 institutions. Figure 4 presents some example images for the three medical imaging datasets.

1) Data Preprocessing: The Aerial Dubai and Electron Microscopy datasets contain large images that were preprocessed into smaller patches for analysis. Specifically, each image in the Aerial Dubai dataset was divided into 224×224 pixel patches, resulting in a total of 1,647 images. For the Electron Microscopy dataset, images were divided into 256×256 pixel patches, yielding 1,980 images. The images in the LGG Brain MRI dataset, originally sized at 256×256 pixels, were centre-cropped to 144×144 pixels. Subsequently, images from all datasets including ISIC were then resized to 288×288 pixels. This preprocessing ensures uniformity in image sizes across different datasets, facilitating consistent and effective analysis.

B. Experimental Settings

Benchmark methods with different CNN backbones (e.g., U-Net [24] or Feature Pyramid Network (FPN) [32] with VGG16 [33] and ResNet34 [34]) are trained end-to-end for semantic segmentation using the ground-truth segmentation masks, comparing to ours using the SP. For fair comparison, the same training images are used to train all the models.

1) Deep Neural Architecture Details:

- We employed U-Net [24] and FPN [32] architectures with pre-trained weights from VGG16 [33] and ResNet34 [34] on the Aerial Dubai dataset. For the medical imaging datasets and all the ablation experiments presented in Section V, we consistently utilized a U-Net with VGG16 weights.
- To adapt U-Net and FPN for predicting SP rather than fine-grained masks, a 1×1 convolutional layer with n filters is employed to match the C number of the semantic classes. Thus n is set to 6 and 1 to output feature maps of the size $288 \times 288 \times 6$ and $288 \times 288 \times 1$ respectively for the Aerial Dubai and medical imaging datasets. Note that there is no need to set n to 2 for the binary segmentation problem with medical imaging datasets. Finally, a global average pooling (GAP) layer added on top to get n float to be used as the predicted SP values.
- To obtain segmentation maps during the test stage, we extract the feature maps prior to the GAP layer and visualise them per semantic class (*cf.* Figures 2 and 3).

2) Training Setup: For all experiments, an 80/20 split for the training/test, Adam optimizer with a learning rate of 10^{-3} , and a batch size of 16 were chosen. The number of epochs was set to 100 with early stopping applied with patience

²<https://www.kaggle.com/datasets/humansintheloop/semantic-segmentation-of-aerial-imagery>

³<https://www.epfl.ch/labs/cvlab/data/data-em/>

⁴<https://www.kaggle.com/datasets/mateusbuda/lgg-mri-segmentation>

⁵<https://cancergenome.nih.gov/cancersselected/lowergradeglioma>

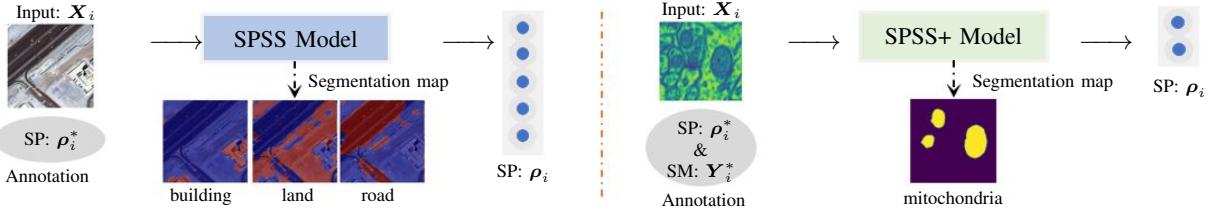


Fig. 5. Diagrams of the proposed models SPSS and SPSS+ on the datasets Aerial Dubai (left) and Electronic Microscopy (right; significant class imbalance), respectively.

TABLE I
QUANTITATIVE SEMANTIC SEGMENTATION RESULTS (MEAN IOU AND F1 SCORES) ON THE AERIAL DUBAI DATASET.

Model	U-Net				FPN			
	VGG16		ResNet34		VGG16		ResNet34	
Backbone	Mean IoU	F1						
Benchmark	71.3 ± 1.2	88.3 ± 0.7	69.2 ± 0.8	86.1 ± 1.2	68.5 ± 0.5	82.1 ± 0.3	67.2 ± 0.8	81.3 ± 0.8
SPSS	64.2 ± 0.6	83.7 ± 0.4	64.4 ± 0.4	80.6 ± 0.8	60.5 ± 0.2	77.2 ± 0.4	61.7 ± 0.6	77.5 ± 1.1
SPSS+	71.6 ± 0.6	88.7 ± 0.6	70.4 ± 0.5	86.4 ± 0.3	67.7 ± 1.2	80.5 ± 0.5	69.2 ± 1.0	82.5 ± 0.7

set to 10 based on the validation loss. All the experiments were implemented on a personal laptop with the following specifications: i7-8750H CPU, GeForce GTX 1060 GPU and 16GB RAM. Training of SPSS and SPSS+ takes around 30 minutes and 40 minutes, respectively.

V. EXPERIMENTS

We highlight that the main aim here is to show that semantic segmentation can be achieved with significantly weaker annotations, i.e., the SP annotation, rather than segmentation accuracy enhancement only. Recall that the difference between SPSS and SPSS+ is just the way of using the annotations for their training, i.e., SPSS+ addresses scenarios that ground-truth segmentation maps are available. Figure 5 illustrates the difference by utilising the SPSS and SPSS+ architectures on the datasets Aerial Dubai and Electronic Microscopy, respectively. To demonstrate the effectiveness of our semantic segmentation approach, we evaluate performance using mean Intersection over Union (IoU) and F1 scores.

A. Segmentation Performance Comparison

Quantitative comparison. Tables I and II give the quantitative results of our method and the benchmark methods for the Aerial Dubai and the three medical imaging datasets, respectively. Well-known evaluation metrics, i.e., mean intersection over union (Mean IoU) and F1 scores are employed. Estimated errors in the mean are obtained by training the models three times with randomly initialised weights. Tables I and II show that SPSS performs comparably to the benchmark methods for all tasks, demonstrating the utility of the SP annotation for semantic segmentation that our methodology introduces. Moreover, SPSS+, i.e., using both ground-truth maps and SP, outperforms the benchmark methods for all the cases except for using the FPN with VGG16 backbone, indicating the usefulness of involving the SP annotation. Note again that SPSS+ does not require any additional data

collection or increase in model complexity, hence offering performance improvements for semantic segmentation tasks nearly for free. Without loss of generality, U-Net with VGG16 is adopted in our method for the rest of the experiments.

TABLE II
QUANTITATIVE SEMANTIC SEGMENTATION RESULTS (MEAN IOU SCORES) ON THE MEDICAL IMAGING DATASETS USING U-NET WITH VGG16 BACKBONE.

Method \ Data	ISIC	Mitocondria	Brain MRI
Benchmark	78.4 ± 0.3	83.7 ± 0.6	72.3 ± 0.2
SPSS	73.2 ± 0.5	76.5 ± 0.2	69.5 ± 0.6
SPSS+	79.1 ± 0.1	84.3 ± 0.5	72.8 ± 0.4

Qualitative comparison. Figure 6 shows the qualitative results of our method and the benchmark method for the Aerial Dubai dataset. Surprisingly, the class-wise segmentation maps that our method achieves (middle of Figure 6) are visually significantly better than that of the benchmark method (right of Figure 6) in terms of the binarisation ability, indicating the effectiveness of the loss \mathcal{L}_{sp} (defined in Eq. (1)) using the SP annotation we introduce. For the significant class imbalance dataset Electronic Microscopy, Figure 7 shows the qualitative results of our method and the benchmark method for some challenging cases. Again, our method exhibits superior performance against the benchmark method. For example, our method can accurately segment the mitochondria on the top-left corner of the second image despite employing much less annotation, but the benchmark method completely misses it despite being trained using the ground-truth segmentation masks. This again validates the effectiveness of the SP annotation for semantic segmentation. Moreover, due to the great binarisation ability of the loss \mathcal{L}_{sp} using SP, it may serve as an auxiliary loss functioning as a plug-and-play module even in scenarios where ground-truth segmentation masks are available to enhance the segmentation performance of many existing methods as SPSS+ does.

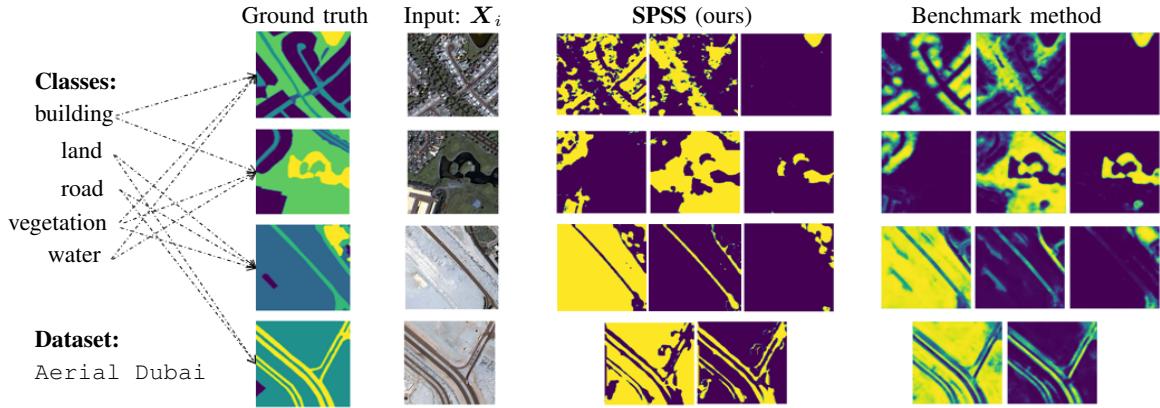


Fig. 6. Qualitative semantic segmentation comparison between our SPSS method (*middle*) and the benchmark method (*right*).

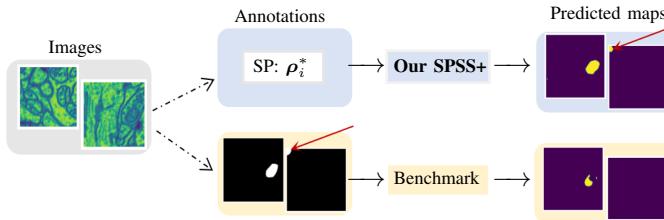


Fig. 7. Comparison between our SPSS+ method (*upper*) and the benchmark method (*lower*) on some images from the Electronic Microscopy dataset.

B. Sensitivity Analysis

Obtaining precise SP annotations may be challenging and, as a result, annotators may provide rough estimates instead. We showcase that rough estimated SP is quite sufficient for our model to achieve good performance (further results are deferred in Section V-C). Below we first investigate the robustness of our models corresponding to the quality of the SP. Two extreme ways degrading the SP are examined: one is adding noises to the SP directly and the other is assigning images in individual clusters the same SP.

1) SP degraded by different noise: We firstly conduct sensitivity analysis of our method SPSS by systematically adding Gaussian noise to the SP for the Aerial Dubai dataset. Let $\mathcal{N}(0, \sigma)$ be the normal distribution with 0 mean and standard deviation σ . For the given SP $\rho_i^* = (\rho_{i1}^*, \dots, \rho_{iC}^*)$, where $\forall \mathbf{X}_i \in \mathcal{X}_T$, let $\tilde{\rho}_i^* = (\tilde{\rho}_{i1}^*, \dots, \tilde{\rho}_{iC}^*)$, where

$$\tilde{\rho}_{ij}^* = \rho_{ij}^* + \mathcal{N}(0, \sigma), \quad j = 1, \dots, C. \quad (4)$$

The above steps are also summarized in Algorithm 1 in Appendix. Then the softmax operator is used to normalise $\tilde{\rho}_i^*$, and the normalised $\tilde{\rho}_i^*$ is used as the new SP to train our model. Here the standard deviation σ controls the level of the Gaussian noise being added to the SP; e.g., $\sigma = 0.1$ represents 10% Gaussian noise. Table III showcases the robustness of our methodology, as it continues performing well even with the SP degraded by quite high levels of noise. E.g., the Mean IoU our method suffers a drops in performance of $\sim 4\%$ for 10% Gaussian noise being added to the SP. Our method still works significantly above random guessing even with the SP

which is degraded by 50% Gaussian noise. This shows that our method is quite robust corresponding to the SP, which means the annotators could in practice spend much less effort for providing rough SP rather than the precise SP.

For medical imaging datasets, the SP of the positive class region, i.e., ρ_{i1}^* , is degraded by a different noise generation process to present diverse noise injection scenarios. Noise is added in a controlled manner utilising the uniform distribution $\mathcal{U}(a, b)$ bounded by a and b , ensuring that the degraded SP remains within a specified range, i.e.,

$$\tilde{\rho}_{i1}^* = \rho_{i1}^* + \lambda \mathcal{U}(a, b) \rho_{i1}^*, \quad (5)$$

where λ is a parameter with value -1 or 1 selected randomly. The above way ensures that the degraded SP is relative to the size of the original SP controlled by bounds a and b . The above steps are also summarized in Algorithm 1 in Appendix. The results presented in Table IV again show that our method SPSS is robust against high level of noise imposed on the SP.

TABLE III
PERFORMANCE OF OUR MODEL IN TERMS OF MEAN IOU TRAINED BY USING THE SP DEGRADED BY GAUSSIAN NOISE.

	Dataset Aerial Dubai							
Noise (%)	0	5	10	15	20	30	40	50
Mean IoU	64.2	62.4	60.1	57.8	52.2	48.3	43.4	38.3

TABLE IV
PERFORMANCE OF OUR MODEL IN TERMS OF MEAN IOU TRAINED BY USING THE DEGRADED SP FOR MEDICAL IMAGING DATASETS.

Noise ($[a, b]$)	Data	ISIC	Mitochondria	Brain MRI
Noise free	73.2	76.5	69.5	
$[0, 0.5]$	70.1	70.5	62.5	
$[0, 1]$	67.3	66.2	60.1	
$[0.5, 1]$	69.3	64.2	63.1	

2) SP degraded by clustering: We now conduct the sensitivity analysis of our method by degrading the SP of the training images by clustering. The degradation procedures are: i) clustering the set of the given SP, i.e., $\{\rho_i^*\}_{i \in \Omega_T}$, into K clusters by K -means; ii) clustering the training set \mathcal{X}_T into the same K clusters, say \mathcal{X}_T^k , $k = 1, \dots, K$, corresponding to the

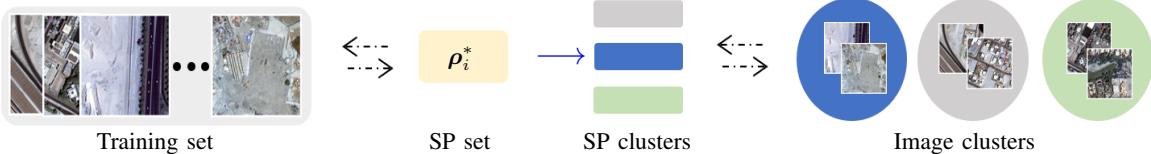


Fig. 8. Diagram of the SP annotation degraded by clustering. Images are clustered corresponding to the SP clusters which are achieved by applying K -means on the SP set. An SP annotation for one image in each image cluster is then randomly selected from that cluster and is assigned to all the images in that image cluster.

TABLE V
PERFORMANCE OF OUR MODEL IN TERMS OF MEAN IOU TRAINED BY
USING THE SP DEGRADED BY CLUSTERING.

	Dataset Aerial Dubai					
# Clusters K	100	50	30	20	10	5
Mean IoU	61.7	59.4	56.5	51.2	47.4	38.3

TABLE VI
COMPARISON BETWEEN THE ANNOTATION STYLES OF OBTAINING THE
SEGMENTATION MASKS AND THE SP IN TERMS OF TIME AND MEMORY.
THE AERIAL DUBAI DATASET IS USED.

Annotation style	Average time per image	Memory per image	
	Original	Compressed	
Segmentation masks	~ 330s	~ 148 kB	~ 4 kB
SP (via annotators)	~ 20s		~ 0.02 kB

SP clusters; and iii) assigning all the training images in cluster \mathcal{X}_T^k the same SP which is randomly selected from the SP of one image in this cluster; see also Figure 8 for illustration. Obviously, implementing this way of degrading the SP, all the images' SP in the training set \mathcal{X}_T are changed except for K (i.e., the number of clusters) images if every training image has different SP annotation in the original SP set. The smaller the number K , the severer the SP degradation.

The performance of our method regarding the SP degraded by clustering is shown in Table V, indicating again the robustness of our methodology corresponding to the SP. For example, after just using $K = 100$ images' SP for the whole training set \mathcal{X}_T , the Mean IoU of our method only drops by $\sim 2.5\%$; and just using $K = 5$ images' SP for the whole training set, our method can still work to some extent (i.e., the Mean IoU just drops less than half). This again shows that our method is indeed quite robust corresponding to the SP. This suggests one possible strategy to reduce effort is to cluster images (for example from patients with a similar level of disease) and then estimate SP on representative images in the cluster.

C. Further Comparison and Analysis

For demonstration purpose, the SP information used in the previous experiments is simply obtained from the given annotated ground-truth segmentation masks. Certainly, in practice, we need the estimated SP information directly from annotators rather than from the ground-truth segmentation masks and thus to significantly simplify the data annotation process. Below we showcase that rough estimated SP directly from annotators can indeed be obtained efficiently and cheaply and is quite sufficient for our models to achieve good performance.

To directly obtain the SP annotations (in the absence of ground-truth masks), 52 images were randomly picked from the Aerial Dubai dataset, and then three annotators were asked to estimate the SP for the provided images. The estimated SP scores were then averaged. Afterwards, data augmentation techniques such as flipping and rotation were applied to obtain 416 images for training. Further details of the annotation process are given in Appendix. Table VI highlights the time and memory cost to produce the SP annotations compared to producing the ground-truth segmentation masks. Pixel annotation for a single image with 5 objects takes roughly 330 seconds which is around 16 times more than the time required for SP annotation⁶. Regarding memory, a mask with the size of 224×224 takes up around 148 kB. With compression, this value can drop to as low as 4 kB, which is still roughly 200 times larger than the SP which consists of only 5 numbers. This huge efficiency brought by our proposed SP strategy is quite significant particularly for big datasets which are required for semantic segmentation.

We now further compare the semantic segmentation performance between the benchmark model with ground-truth segmentation maps and our SPSS with the SP simply obtained from the ground-truth segmentation maps and the rough SP produced by the annotators (the details of the annotation process are given in Appendix), separately. Table VII presents the results on the same test set used in Table I. The results are quite impressive as SPSS with the rough SP estimations surpasses not only the way of using the SP obtained by the ground-truth maps but also the benchmark model trained using the costly ground-truth maps.

VI. DISCUSSION AND LIMITATION

SP (semantic proportions) for each training image is required as annotation/label information for the presented semantic segmentation model. In this work, we obtained these proportions from both the segmentation maps available for the chosen datasets and three annotators directly to demonstrate the effectiveness and robustness of our proposed SP-based methodology. We would like to stress that the reason why we benefited from the existing segmentation maps, which seems controversial to our main aim at first glance, is to show that the proposed methodology is feasible in the presence of SP. Arguably, reasonable proportions can be simply extracted from the ground-truth segmentation maps if they are annotated properly. Therefore, obtaining SP from the readily available maps to achieve our aim is sensible. Clearly, our goal is to

⁶Average time taken for per-pixel annotation is estimated based on [8].

TABLE VII
QUANTITATIVE COMPARISON ON THE AERIAL DUBAI DATASET WITH ROUGH ESTIMATED SP ANNOTATIONS.

Model	Mean IoU	Per-class F1 score					Mean accuracy
		Building	Land	Road	Vegetation	Water	
Segmentation masks	39.5 ± 1.3	52.7 ± 1.2	84.8 ± 0.6	2.4 ± 0.6	43.2 ± 1.3	75.4 ± 0.5	67.9 ± 1.1
SP (via seg. masks)	37.9 ± 0.8	39.8 ± 1.3	84.6 ± 0.3	4.5 ± 0.2	41.3 ± 0.8	77.2 ± 0.9	67.4 ± 0.3
SP (via annotators)	41.6 ± 1.3	46.2 ± 0.7	85.7 ± 1.3	26.6 ± 2.1	44.3 ± 0.8	75.6 ± 0.3	68.7 ± 0.4

train our proposed model when the segmentation maps are unavailable. It is evident from our experiments that obtaining SP annotation could be much cheaper than obtaining the precise segmentation maps particularly for data volumes in high dimensions. There are obviously various ways to obtain SP readily in the absence of the segmentation maps, such as by employing mechanical turks. There may exist applications such as estimating the density of housing in a particular area where information may be extracted from other studies or even obtained from pre-trained large language models, e.g., ChatGPT [35].

The results that we present in Section V are promising and one may wonder if the exact proportions are a must, which would make the proposed setting as expensive as the traditional one. To demonstrate that it is not the case and that our methodology only needs rough SP, we presented sensitivity analysis regarding SP, where we added various amounts of noise to the extracted SP and demonstrated that the model performs satisfactorily well when trained with noisy SP. We also presented sensitivity analysis through investigating degraded SP by clustering to further support the robustness of our methodology when the precise SP is unavailable. The analysis suggests that our methodology not only works well with rough SP, but also with rough SP for only some representative images from the whole training set, indicating its need of significantly less annotation effort.

Additional annotations. In many scenarios, different types of annotations may exist. This raises the question that whether it is feasible for semantic segmentation methods to use the combination of different types of annotations to boost their performance. In this regard, our proposed semantic segmentation methodology based on SP delivers quite promising results.

For datasets where the ground-truth segmentation maps are available, the SP annotation can be calculated directly. In these cases an additional loss function using the SP scores can be used as demonstrated by the SPSS+ model we have proposed. The results shown in Tables I and II demonstrated the good performance of SPSS+. The enhanced performance of our method by utilising both annotation types may benefit from our introduced loss function $\mathcal{L}_{\text{total}}$ in Eq. (3). It contains the \mathcal{L}_{sp} loss defined in Eq. (1), which measures the MSE between the predicted SP and the given SP. The visualisation results in Figure 6 showed that our \mathcal{L}_{sp} loss may produce better segmentation than the loss directly measuring the segmentation maps (that the benchmark method uses) in terms of the binarisation ability. Therefore, combining the \mathcal{L}_{sp} loss with the \mathcal{L}_{sm} loss and then forming the $\mathcal{L}_{\text{total}}$ loss could boost the semantic segmentation performance, e.g. see the visualisation given in Figure 7.

Limitations. SP provides much less information than standard segmentation annotations. In some scenarios, for example, with large number of classes or where some classes represent only a tiny proportion of any image, the semantic proportions might not provide enough information for the network to infer the classes. Thus the utility of SP will be problem dependent. In many ways the surprising observation for us was to discover how powerful SP is on a range of problems given how little information we are providing to the network. Although SP will not be a solution for all segmentation problems, we believe that its relative cheapness means that it may be the method of choice in a number of applications where semantic segmentation is required, but the resources to hand annotation images is limited.

In this work, we proposed a new semantic segmentation methodology by introducing the SP annotation. In the scenario of quite limited annotation, using SP for semantic segmentation can already achieve competitive results. If additional annotations are available, our method can easily utilise them for performance boost. Moreover, for existing segmentation methods that use different types of annotations, we also suggest involving SP in these methods; e.g., our proposed \mathcal{L}_{sp} loss could be served as a type of regularisation given its effectiveness in binarisation.

VII. CONCLUSION

Semantic segmentation methodologies generally require costly annotations such as the ground-truth segmentation masks in order to achieve satisfying performance. Motivated by reducing the annotation time and cost for semantic segmentation, we in this paper presented a new methodology SPSS, relying on the SP annotation instead of the costly ground-truth segmentation maps. Extensive experiments validated the great potential of the proposed methodology in reducing the time and cost required for annotation, making it more feasible for large-scale applications. Furthermore, this innovative design opens up new opportunities for semantic segmentation tasks where obtaining the ground-truth segmentation maps may not be feasible or practical. We believe that the use of the SP annotation suggested in this paper offers a new and promising avenue for future research in the field of semantic segmentation, with evident and wide real-world applications.

Acknowledgements. Halil Ibrahim Aysel is thankful for the support from the Republic of Turkiye Ministry of National Education.

Reference Images Average SP Estimations by Annotators

Classes
B: building
L: land
R: road
V: vegetation
W: water

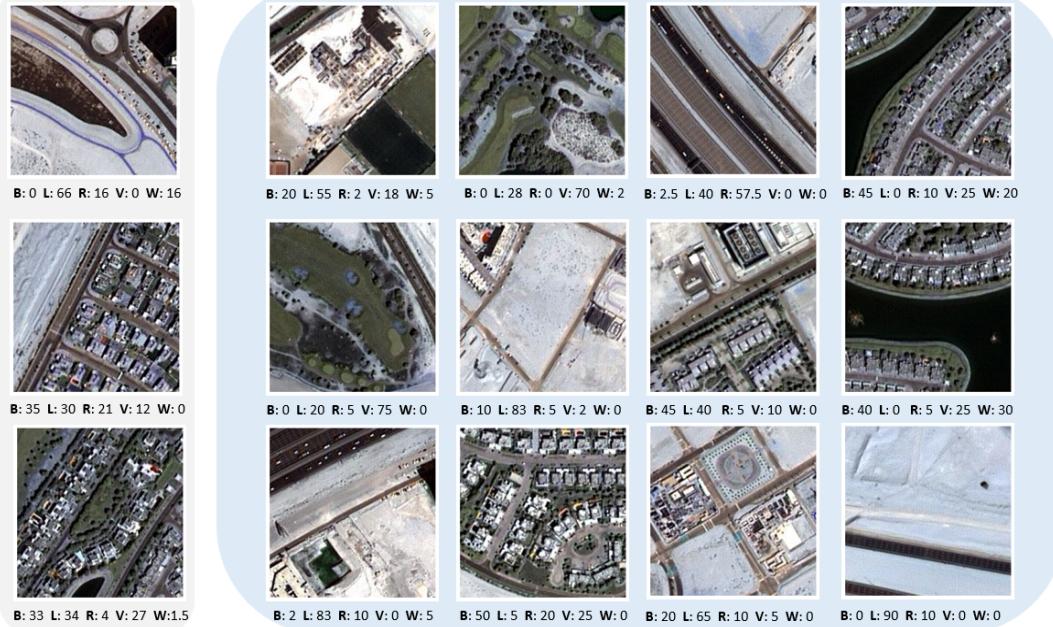


Fig. 9. Showcase of the SP annotation process by annotators directly. Three annotators were asked to annotate a batch with 52 images for training. Left: reference images whose SP information is calculated from the pixel-wise annotated ground-truth segmentation maps. Right: some randomly selected images with their average SP estimations by the three annotators.

APPENDIX

A. Semantic Proportions Annotation

For the experiments presented in Section V-C, three annotators were asked to annotate a small batch containing 52 images from the Aerial Dubai dataset each with the size of 288×288 to show the efficiency of the SP annotation process compared to the pixel-wise annotation, as well as the excellent semantic segmentation ability of the proposed SPSS model compared to the benchmark model (with the ground-truth segmentation maps).

- The annotators were provided with three reference images whose SP information is simply obtained via the pixel-wise segmentation maps, see the left of Figure 9 above. The reference images could be helpful for annotators to adjust their estimations; for instance, for the last image in the first row of Figure 9 regarding the SP estimations, it is clear that the water area is a little larger than that in the first reference image, which helps the annotators to estimate a proportion with a larger value than that for the water area in the reference image (i.e., 20% vs. 16%). The average estimation of the three annotators for the water area in the mentioned image is around 20%, which is quite close to the value obtained by its ground-truth map, i.e., 21.3%, showing the efficiency of the SP annotation directly by annotators in this manner. Moreover, our sensitivity experiments showed that obtaining precise SP information for training is not a must for our SPSS model to perform well, making the SP annotation process even more efficient and relaxing given its tolerance of rough deviation in the SP estimations.

- After each annotator completed their SP annotation, the average SP annotation of the three annotators is obtained for the 52 images.
- Finally, two types of augmentation strategies were carried out to increase the training dataset size. Each image was flipped horizontally and rotated by 90, 180 and 270 degrees clockwise. The rotations were also applied to every flipped image. Therefore, 8 images were obtained for every image, and a training dataset consisting of 416 images in total is formed. Note that, since the SP information is irrelevant to the position of the content in an image, the estimated SP for one image is also applied to all of its 7 augmented versions.

B. Algorithm

Algorithm 1 shows the noise injection processes for the experiments presented in Section V-B1.

REFERENCES

- [1] M. Siam, M. Gamal, and et al., “A Comparative Study of Real-time Semantic Segmentation for Autonomous Driving,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 587–597.
- [2] S. Asgari Taghanaki, K. Abhishek, and et al., “Deep semantic segmentation of natural and medical Images: a review,” *Artificial Intelligence Review*, vol. 54, pp. 137–178, 2021.
- [3] R. Yang and Y. Yu, “Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medical Imaging Analysis,” *Frontiers in oncology*, vol. 11, p. 638182, 2021.
- [4] H. Zhang, B. Han, and et al., “Slimmer: Accelerating 3D Semantic Segmentation for Mobile Augmented Reality,” in *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 2020, pp. 603–612.

Algorithm 1 Noisy SP $\tilde{\rho}_i^*$ Generation

```

1: Input: Ground-truth SP  $\rho_i^*$  of image  $X_i$ , standard deviation  $\sigma$ , lower bound  $a$ , and upper bound  $b$ .
2: Output: Noisy SP  $\tilde{\rho}_i^*$ 
3: if  $\text{length}(\rho_i^*) == 1$  then ▷ E.g., medical imaging datasets
4:   Randomly select  $\lambda$  from  $\{-1, 1\}$ ;
5:    $\tilde{\rho}_{i1}^* = \rho_{i1}^* + \lambda \mathcal{U}(a, b)\rho_{i1}^*$ ;
6: else ▷ E.g., Aerial Dubai dataset
7:   for  $j = 1$  to  $\text{length}(\rho_i^*)$  do
8:      $\tilde{\rho}_{ij}^* = \rho_{ij}^* + \mathcal{N}(0, \sigma)$ ;
9:   end for
10: end if
11: return  $\tilde{\rho}_i^*$ 

```

- [5] A. Milioto, P. Lottes, and C. Stachniss, “Real-time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs,” in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 2229–2235.
- [6] A. Garcia-Garcia, S. Orts-Escolano, and et al., “A Review on Deep Learning Techniques Applied to Semantic Segmentation,” *arXiv preprint arXiv:1704.06857*, 2017.
- [7] S. Hao, Y. Zhou, and Y. Guo, “A Brief Survey on Semantic Segmentation with Deep Learning,” *Neurocomputing*, vol. 406, pp. 302–321, 2020.
- [8] A. Bearman, O. Russakovsky, and et al., “What’s the Point: Semantic Segmentation with Point Supervision,” in *European conference on computer vision*. Springer, 2016, pp. 549–565.
- [9] K. Genova, X. Yin, and et al., “Learning 3D Semantic Segmentation with Only 2D Image Supervision,” in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 361–372.
- [10] P. O. Pinheiro and R. Collobert, “From Image-level to Pixel-level Labeling with Convolutional Networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1713–1721.
- [11] Y. Wei, X. Liang, and et al., “Learning to Segment with Image-level Annotations,” *Pattern Recognition*, vol. 59, pp. 234–244, 2016.
- [12] D. Lin, J. Dai, and et al., “Scribblesup: Scribble-supervised Convolutional Networks for Semantic Segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3159–3167.
- [13] D. P. Papadopoulos, A. D. Clarke, and et al., “Training Object Class Detectors from Eye Tracking Data,” in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 361–376.
- [14] R. A. McEver and B. Manjunath, “Pcams: Weakly Supervised Semantic Segmentation Using Point Supervision,” *arXiv preprint arXiv:2007.05615*, 2020.
- [15] G. Papandreou, L.-C. Chen, and et al., “Weakly-and Semi-supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1742–1750.
- [16] J. Dai, K. He, and J. Sun, “Boxsup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1635–1643.
- [17] H. Lee and W.-K. Jeong, “Scribble2label: Scribble-supervised Cell Segmentation via Self-generating Pseudo-labels with Consistency,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, 2020, pp. 14–23.
- [18] S. Mittal, J. Niemeijer, and et al., “Revisiting Deep Active Learning for Semantic Segmentation,” *arXiv preprint arXiv:2302.04075*, 2023.
- [19] S. Xie, Z. Feng, and et al., “Deal: Difficulty-aware Active Learning for Semantic Segmentation,” in *Proceedings of the Asian conference on computer vision*, 2020.
- [20] Y. Chen, W. Li, and et al., “Learning Semantic Segmentation from Synthetic data: A Geometrically Guided Input-output Adaptation Approach,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1841–1850.
- [21] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [22] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A Deep Convolutional Encoder-decoder Architecture for Image Segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [23] L.-C. Chen, G. Papandreou, and et al., “Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [24] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [25] H. Hu, Y. Zheng, and et al., “MC-Unet: Multi-scale Convolution Unet for Bladder Cancer Cell Segmentation in Phase-contrast Microscopy Images,” in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 1197–1199.
- [26] H. Chen, W. Zhang, and et al., “Multi-organ Segmentation Based on 2.5D Semi-supervised Learning,” in *Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation: MICCAI 2022 Challenge, FLARE 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*. Springer, 2023, pp. 74–86.
- [27] M. Dildar, S. Akram, and et al., “Skin Cancer Detection: a Review Using Deep Learning Techniques,” *International journal of environmental research and public health*, vol. 18, no. 10, p. 5479, 2021.
- [28] Y. Cao, A. Vassantachart, and et al., “Automatic detection and segmentation of multiple brain metastases on magnetic resonance image using asymmetric UNet architecture,” *Physics in Medicine & Biology*, vol. 66, no. 1, p. 015003, 2021.
- [29] N. Codella, V. Rotemberg, and et al., “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic),” *arXiv preprint arXiv:1902.03368*, 2019.
- [30] P. Tschandl, C. Rosendahl, and H. Kittler, “The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.
- [31] M. Buda, A. Saha, and M. A. Mazurowski, “Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm,” *Computers in biology and medicine*, vol. 109, pp. 218–225, 2019.
- [32] T.-Y. Lin, P. Dollár, and et al., “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
- [33] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [34] K. He, X. Zhang, and et al., “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [35] OpenAI, “Introducing ChatGPT,” <https://openai.com/blog/chatgpt>, 2022.