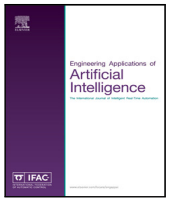
Engineering Applications of Artificial Intelligence 123 (2023) 106299

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Exploring the trade-off between performance and annotation complexity in semantic segmentation

Marta Fernández-Moreno [a,b], Bo Lei [c], Elizabeth A. Holm [c], Pablo Mesejo [a,*], Raúl Moreno [d]

[a] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence, DaSCI, University of Granada, 18071, Granada, Spain
[b] TheNextPangea SL, Residencia la Granda, Gozon, 33440, Asturias, Spain
[c] Materials Science and Engineering, Carnegie Mellon University, Pittsburgh PA, 15213, USA
[d] Department of Computer Science and Automatic Control, National Distance Education University (UNED), Juan del Rosal 16, Madrid 28040, Spain

ARTICLE INFO

Dataset link: https://github.com/martafdezmAM/lessen_supervision

Keywords:
Semantic segmentation
Unsupervised learning
Weakly supervised learning
Deep convolutional neural networks

ABSTRACT

Image semantic segmentation, a fundamental computer vision task, performs the pixel-wise classification of an image seeking to group pixels that share some semantic content. One of the main issues in semantic segmentation is the creation of fully annotated datasets where each image has one label per pixel. These annotations are highly time-consuming and, the more the labelling increases, the higher the percentage of human-entered errors grows. Segmentation methods based on less supervision can reduce both labelling time and noisy labels. However, when dealing with real-world applications, it is far from trivial to establish a method that minimizes labelling time while maximizing performance.

Our main contribution is to present the first comprehensive study of state-of-the-art methods based on different levels of supervision. Image processing baselines, unsupervised, weakly supervised and supervised approaches have been evaluated. We aim to guide anyone approaching a new real-world use case by providing a trade-off between performance and supervision complexity on datasets from different domains, such as street scenes (Camvid), microscopy (MetalDAM), satellite (FloodNet) and medical images (NuCLS). Our experimental results suggest that: (i) unsupervised and weak learning perform well on majority classes, which helps to speed up labelling; (ii) weakly supervised can outperform fully supervised methods on minority classes; (iii) not all weak learning methods are robust to the nature of the dataset, especially those based on image-level annotations; and (iv) among all weakly supervised methods, point-based are the best-performing ones, even competing with fully supervised methods. The code is available at https://github.com/martafdezmAM/lessen_supervision.

1. Introduction

Semantic segmentation (Long et al., 2015; Wang et al., 2018) is one of the most widely used problems in computer vision since it performs both classification and localization tasks within an image at the pixel level in a precise way. Moreover, semantic segmentation allows us to estimate the size of objects in an image by converting each pixel size to units of length measurement, based on information such as image augmentation parameters. This way, we can perform accurate object quantification tasks. During the last few years, methods based on semantic image segmentation have obtained great performance in different fields, such as video surveillance (Muhadi et al., 2020), autonomous driving (Feng et al., 2021), medical prognosis (Wang et al., 2022), or material's characterization (Holm et al., 2020), just to name a few. These techniques are usually model-centric approaches, focusing on improving the quality of the model by, for example, exploring architectures, algorithms or hyperparameters. However, it is not common to find studies related to the data-centric approach, which focuses on exploiting the data to reach the best performance.

The generation of good-quality annotations is a very common challenge involving data-centric approaches due to several issues. While building a semantic segmentation dataset, the main issue is the large amount of time required to generate the necessary labelled data. Complete labelling refers to annotating every pixel per image, often entailing a tremendous amount of labelling effort. For instance, in the case of MS COCO Lin et al. (2014) labelling at the image level took an average of 4.1 s while labelling at the pixel level took an average of 10.1 min. This means that while labelling 1 image at the pixel level, 148 images are labelled at the image level. This is an example of an

* Corresponding author.
E-mail addresses: martafm@correo.ugr.es (M. Fernández-Moreno), blei1@andrew.cmu.edu (B. Lei), eaholm@andrew.cmu.edu (E.A. Holm), pmesejo@decsai.ugr.es (P. Mesejo), raul.moreno.salinas@gmail.com (R. Moreno).

https://doi.org/10.1016/j.engappai.2023.106299
Received 19 July 2022; Received in revised form 4 April 2023; Accepted 5 April 2023
Available online 25 April 2023

image dataset with a 640 × 480 resolution, however, this labelling task becomes even more tedious when facing datasets with images of higher resolution, with a larger number of classes or, when contour outlining is complex and requires a huge effort. For instance, the Camvid dataset (Brostow et al., 2009) reports that labelling each image takes about 60 min. Another good example is Cityscapes (Cordts et al., 2016), which contains higher resolution images and two types of labels, coarse labels consisting of poorly refined polygons and requiring an average of 7 min of labelling per image, and fine labels consisting of well-defined pixel-level labels and taking about 90 min. In the case of MetalDAM (Luengo et al., 2022), with high-resolution images and complex boundaries between classes, the average labelling time per image is 140 min.

A secondary issue, depending on the particular application field, in cases such as analysis of medical image Liu et al. (2021a) or material microstructures (Luengo et al., 2022; Holm et al., 2020), among others, labelling requires expert domain knowledge. Therefore, the labelling work must be carried out, in many cases, by one or a few specialists in those fields, who may not have enough time to perform this kind of labelling task. This increases the need of reducing the amount of labelling.

As for the third issue, we also must consider that a higher image resolution implies an increase in the labelling time, but also the noise on the annotations and the probability of introducing human errors during labelling. Therefore, using fully annotated images is not always the best solution as it is prone to human errors and biases. A recent study (Northcutt et al., 2021) analysed 10 of the most common state-of-the-art datasets and showed that approximately 3.3% of the labels in a dataset are wrong. For instance, in ImageNet (Deng et al., 2009) 6% of labels in the evaluation subset are wrongly annotated. These labelling errors result in inaccurate evaluation metrics, which may mean that some model-centric approaches, rather than improving the model performance, are learning to adapt to these labelling errors.

These challenges can be tackled in several ways by applying methods based on different levels of supervision rather than training fully supervised models on large datasets. On the other hand, reducing the level of supervision may also reduce their performance, since their natural behaviour will be to produce inaccurate segmentations due to the lower amount of supervision (Li et al., 2018). However, this performance depends largely on the training data, as well as the objectives and the tasks to be performed. With all of this in mind, before generating a new model from an unlabelled dataset, the first thing to do is to establish a trade-off between time to be invested in labelling and training algorithms, versus the desired performance level. For instance, if just some hundreds of labelled pixels can provide results that meet the target performance level, training with thousands of annotated pixels will not be the best option as it will require much more annotation effort. There are cases such as detecting people in security environments where there is no need for such high accuracy, on the other hand, cases like medical diagnostics or material characterization require precise results.

Many works have proposed new architectures or methods within each supervision branch: supervised learning (Guo et al., 2018), semi-supervised (Hong et al., 2015), weakly supervised (Chan et al., 2021), few-shot (Dong and Xing, 2018) and unsupervised (Toldo et al., 2020). Some of these works focus on providing new architectures or incremental optimizations for a given data domain or even for a specific dataset. Another part of the literature develops reviews comparing methods from the aforementioned supervision branches. For example, there are studies such as (Lateef and Ruichek, 2019) which include a theoretical and practical comparison of different neural network architectures on different datasets using fully supervised training. We also found reviews of methods based on unlabelled data such as (Li et al., 2018) which compares unsupervised methods with state-of-the-art fully supervised approaches. Regarding weakly supervised methods, Chan et al. (2021) compares a few of them with their fully supervised version.

In the literature, when comparing other supervision approaches with unsupervised methods, classical computer vision techniques are often used, neglecting the potential of more sophisticated unsupervised segmentation methods. However, there are numerous papers proposing novel unsupervised segmentation methods such as Ji et al. (2019), Cho et al. (2021), Hwang et al. (2019), Van Gansbeke et al. (2021), Kanezaki (2018), Kim et al. (2020b) and Hamilton et al. (2022). Despite unsupervised methods obtaining poorer results than others with higher supervision, if the target of the model is relatively easy, they can be a perfect choice. Moreover, unsupervised approaches can also be useful as an aid to labelling, using their output as a pre-annotation, reducing notably the annotation effort. This applies not only to unsupervised but also to weakly supervised approaches so, by testing methods with lower complexity labels, we can generate more complex labels easily. Fig. 1 graphically represents this idea in which three large families of methods are identified: Unsupervised (no labels), weakly supervised (partial labels) and supervised (fully annotated labels).

There are other families such as few-shot learning and domain adaptation. Few-shot learning focuses on reducing the labelling effort by minimizing the number of images. However, they still make use of fully annotated data, which, as discussed in the third issue, leads to the introduction of human errors. Moreover, the process of labelling a single pixel-wise annotation is more tedious for the labellers than annotating diverse images using weak labels. In addition, as these methods use fewer images during the training, they face a hard problem which is the selection of the most representative images. As for domain adaptation techniques, there are papers such as (Dong et al., 2020, 2021; Liu et al., 2021b) that only require a model trained on a similar domain and an unlabelled dataset to obtain very good results. They achieve these results by preserving the source domain knowledge from a pre-trained model via knowledge transfer during model adaptation. However, in this work, we focus on real-world applications where it is not common for the community to share already trained networks or labelled datasets from the same domain.

This paper aims to study the trade-off between performance vs. model generation effort, by exploring diverse supervision approaches. The main goal is to speed up the model generation process and evaluate the necessary amount of annotation for datasets of different natures. Fig. 1 shows the process we follow, which consists in testing methods from the lowest to the highest labelling complexity. In this way, generating a final pixel-wise annotation can be sped up by using the predictions of less complex models as pre-annotations or pseudo-labels.

Our main contribution is to present, up to our knowledge, the first comprehensive study of state-of-the-art methods based on different levels of supervision. We test 12 methods on 4 datasets from different domains: 2 image processing baselines, 3 unsupervised methods, 6 weak learning approaches and 1 fully supervised approach. Our comparative study allows us to draw, among others, the following conclusions:

- Not all methods perform equally well on diverse datasets, being the point-based approach the most robust one.
- The point-based approach performs very similarly to the supervised method. Since the point-based approach requires less than 0.1% of the annotated pixels, they represent an interesting and feasible alternative in real-world use cases.
- Unsupervised methods are generally effective at predicting majority classes, which would help to speed up the labelling process. In this study, unsupervised methods based on Deep learning (DL) techniques have been able to correctly label more than 90% of the pixels.
- Some weakly supervised approaches, such as point-based and scribble-based methods, outperform fully supervised methods on minority classes. This could be due to the negative impact of human errors introduced during the labelling process, as certain less represented classes are sometimes ignored during the labelling.

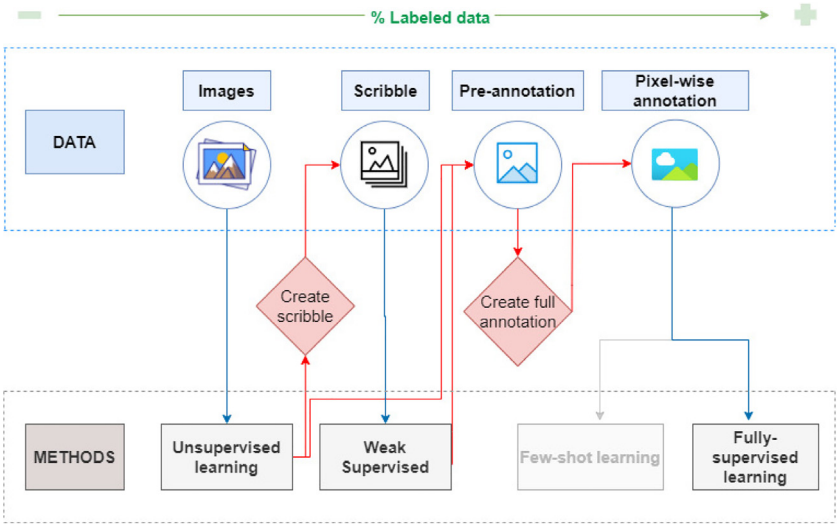


Fig. 1. Trade-off performance vs. model generation effort: The natural process for establishing this trade-off is to run methods from lower to higher supervision rates while speeding up the labelling process by reusing method predictions as pre-annotations.

2. Methods

As we have mentioned so far, in this paper we study the trade-off between performance and model generation effort. Based on this, we have selected state-of-the-art methods trained on different kinds of labels whose training process is simple and as fast as possible.

2.1. Unsupervised learning

As mentioned in the introduction, the main objective of unsupervised approaches is that, despite not obtaining as accurate results as the supervised ones, these models can provide a solution quickly since no labelling is required. There are numerous unsupervised segmentation methods in the state of the art based on various architectures and approaches. We have analysed several methods such as (Ji et al., 2019; Cho et al., 2021; Hwang et al., 2019; Kanezaki, 2018; Caron et al., 2018) and evaluated them on several datasets. However, most of these methods are based on complex architectures whose training time can extend to more than 12 h or even days, depending on the number and size of images in each dataset and, in many cases, hyperparameter optimization requires running the experiments more than 20 or 30 times to get an optimized model. This means that the results highly depend on the hyperparameter values employed and obtaining a reliable model would take a matter of months.

As reasoned above, the advantage of supervised methods is that they provide a fast segmentation without labelling, serving as a pre-annotation or a solution to a simple problem. With this in mind, we have selected some methods whose training takes just a few seconds or minutes depending on the resolution of the input image. The first method is the approach presented in Kanezaki (2018), as it can generate fairly decent segmentations in a matter of minutes. Hereafter, we will refer to this method as Unsupervised Segmentation with Superpixel (USSP) in this article. This method considers three key aspects in image segmentation: feature similarity, spatial continuity of superpixel-based clusters, and the number of unique clusters. Unlike traditional DL methods, USSP does not generate a final model trained on a dataset. Instead, this algorithm is applied individually for each image by training a fairly simple Convolutional Neural Network (CNN) for a few steps before generating the final segmentation. USSP has been successfully tested in specific tasks such as microscopic imaging, solving the characterization of steel microstructures in Kim et al. (2020a).

The other selected method was introduced in Kim et al. (2020b) as a modification of USSP (Kanezaki, 2018) and we will refer to it as Continuity-based Unsupervised Segmentation (USC). The authors claimed as a limitation that the segment boundaries were fixed in Kanezaki (2018) due to superpixel refinement, so they replaced this refinement by adding a continuity term in the loss function. This loss favours cluster labels to be the same as those of neighbouring pixels by considering the L1 norm of the horizontal and vertical differences of the response map as a spatial constraint. This method can be trained in the same way as USSP, by training one model per image, or by doing what the authors call "training by reference", which consists of training the same network on a subset of images. From now on in this paper we will refer to the "training by reference" approach as Reference-based Unsupervised Segmentation (USRef).

2.2. Weakly supervised learning

As mentioned above, we can group weakly supervised methods into different categories according to the type of label they use for training. However, bounding boxes annotations (Hsu et al., 2019) tend to be more common in instance segmentation, as they are more related to providing information about the presence of an object rather than its exact location. In addition, semantic segmentation datasets tend to have highly intermingled classes where bounding boxes would overlap and include parts of other classes. This means that within semantic segmentation, image-level (Zhou et al., 2021, 2022), scribbles (Lin et al., 2016) and points (Papadopoulos et al., 2017) are treated as the most commonly used weak labels. Therefore, for each of these three approaches, a method has been selected based on the same criteria as those established when selecting unsupervised methods, i.e., simplicity of the methods and speed in training and optimizing the model.

The first weak learning method will be an image-level approach as it is the simplest type of weak label. Within this field, there are numerous approaches such as (Du et al., 2022; Zhou et al., 2021, 2022), among others, from which (Zhou et al., 2022) has been selected for its balance between good performance and ease of reproducing the results reported by the authors. This method obtains pseudo-labels by combining semantic contrast and aggregation. These pseudo-labels are then used as labels during subsequent supervised training. We will refer to this paper from now on as RCA.

The approach involving scribbles is a variation of the USC (Kim et al., 2020b) method and we will refer to this one as Semantic Segmentation with Scribbles (SSCR), where the authors introduce another term to the loss function regarding the supervised loss over the scribble information. We have selected this method for the same reasons as

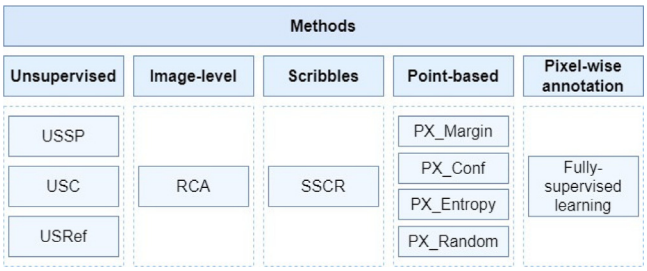


Fig. 2. Summary of the selected methods grouped by the type of label used during training. Unsupervised methods: Unsupervised Segmentation with Superpixel (Kanezaki, 2018) (USSP), Continuity-based Unsupervised Segmentation (Kim et al., 2020b) (USC), Reference-based Unsupervised Segmentation (Kim et al., 2020b) (USRef). Image-level method (Zhou et al., 2022): Regional Contrast Aggregation (RCA) Scribble method (Kim et al., 2020b): Scribble-based Segmentation (SSCR). Point-based methods (Shin et al., 2021): Margin Sampling-based PixelPick (PX_Margin), Confidence-based PixelPick (PX_Conf), Entropy-based PixelPick (PX_Entropy), Random-based PixelPick (PX_Random). Supervised method: DeepLabV3+ (Chen et al., 2018).

USC, since it performs well on different datasets, its training is fast and simple, the model architecture is lightweight and the incorporation of scribbles in the training is simple and efficient in terms of performance.

As for the point-based method, we have selected the approaches in Shin et al. (2021) where the authors present an active learning pipeline inspired by the "extreme clicking" method introduced in Lin et al. (2016). This pipeline consists of N consecutive trainings in which the output proposes M new pixels to be labelled considered as the most relevant ones. The authors compare four different criteria to consider which pixel is more relevant using the measures: least loss confidence, margin sampling, cross-entropy loss value and random selection. We will refer to each of them as PX_conf, PX_margin, PX_entropy and PX_random. These active learning criteria will also be compared with the rest of the methods. This approach has been selected because it allows us not only to simply evaluate the training on a different number of pixels but also to compare different approaches based on active learning. On the other hand, the authors provide very promising results on the Camvid dataset, close to supervised with less than 0.1% of the labelled pixels, so one of our goals will be to check if this behaviour is also true for other datasets.

As a summary, Fig. 2 shows an outline of the methods we have selected for this study since, after an initial analysis of several methods, they are the ones that best suit the objectives of this comparative study. The figure shows each of the methods gathered according to the type of label used during training.

3. Experiments

In this section, we will evaluate the selected methods on several datasets with different characteristics. The objective is to analyse how this trade-off between performance and model generation effort behaves in each dataset by comparing methods trained on labels of different complexities. A common repository has been implemented, where the above methods have been adapted to new datasets. As part of the contribution of this article, this repository has been made public and can be accessed from the link: https://github.com/martafdezmAM/lessen_supervision.

3.1. Datasets

Since not all methods behave in the same way depending on the different characteristics of the datasets, we will evaluate how far we can go in terms of accuracy and performance with different levels of supervision. For this purpose, we must analyse the impact of each method regardless of the type of image used, so we must test these methods on highly different datasets. To achieve this, the following datasets have been selected looking for the greatest possible diversity based on the following criteria: tasks, number of examples, resolution, and number of classes.

- **MetalDAM** (Luengo et al., 2022): MetalDAM consists of grayscale images of steel microstructures taken with different microscopes or similar specialized devices that are capable of obtaining magnified images. In material science, this kind of data is typically used for characterization tasks, and the annotations do not usually contain many classes. However, they are high-resolution images that have a large amount of detail and uncertain boundaries between classes. The areas to be segmented usually involve more abstract and complex criteria than in other kinds of tasks such as the segmentation of well-defined objects such as cars or people.
- **Camvid** (Brostow et al., 2009): Camvid consists of images taken from a car driving on diverse roads. We see a lot of variability of illumination between images due to the different times of the day in which the photo were taken, atmospheric conditions or even the shadows produced by objects as seen from the camera's viewpoint. Camvid has been used as a benchmark in multiple methods, including the point-based method (Shin et al., 2021) selected in this article. Based on this, we have used the same version of the dataset as the authors.
- **FloodNet** (Rahnemoonfar et al., 2021): It is composed of high-resolution satellite images that capture the consequences of some natural disasters. These are aerial images, so the perspective of the images is not highly variable. However, due to their very high resolution, they contain a large amount of information. FloodNet contains thousands of images, but only the labels of the training set are public, so just 398 training images have been used.
- **NuCLS** (Amgad et al., 2022): This dataset contains examples of breast cancer images from The Cancer Genome Atlas (TCGA) programme. The version labelled by non-pathologists and subsequently validated by a pathologist has been selected as it contains a larger volume of better-quality labels. NuCLS has a training subset of 1481 images and a test subset of 263 images.

Table 1 gathers the most relevant information relative to each dataset and Fig. 3 shows an example of an image and its mask for each dataset. From this table and Figure, we can observe the differences between the kind of images and annotations between datasets, as well as the differences of the volume of data and the image resolution.

3.2. Set-up

In summary, the behaviour of MetalDAM, Camvid, FloodNet and NuCLS datasets with different labels will be tested on the methods: USSP (Kanezaki, 2018), USC (Kim et al., 2020b) and USCRef. (Kim et al., 2020b) (unsupervised), RCA (Zhou et al., 2022), SSCR (Kim et al., 2020b), PX_margin, PX_conf, PX_entropy and PX_random (Shin et al., 2021) (weakly supervised approaches) and a CNN trained in a fully supervised fashion (Chen et al., 2018).

Additional methods were tested against the four datasets; however, the previous methods were selected by taking into account key issues such as reduced labelling time, model training and human-introduced error. Extensive experimental work has been carried out since the previous experiments along with those outlined in this paper have been run more than 40 times for each dataset. These experiments have been running for 7 months on 2 DGX servers each with 8 NVIDIA Tesla V100 GPUs and 3 servers equipped with NVIDIA RTX 3090.

To contrast unsupervised methods with more traditional techniques, k-means (Dash et al., 2010) and graph-segmentation (Felzenszwalb and Zabih, 2010) baselines have been included. As for the weak learning methods, the code has been modified to use PyTorch segmentation models library (Yakubovskiy, 2020). This modification gives users the possibility of using new loss functions and a much wider catalogue of network architectures. Each of these architectures was tested, and

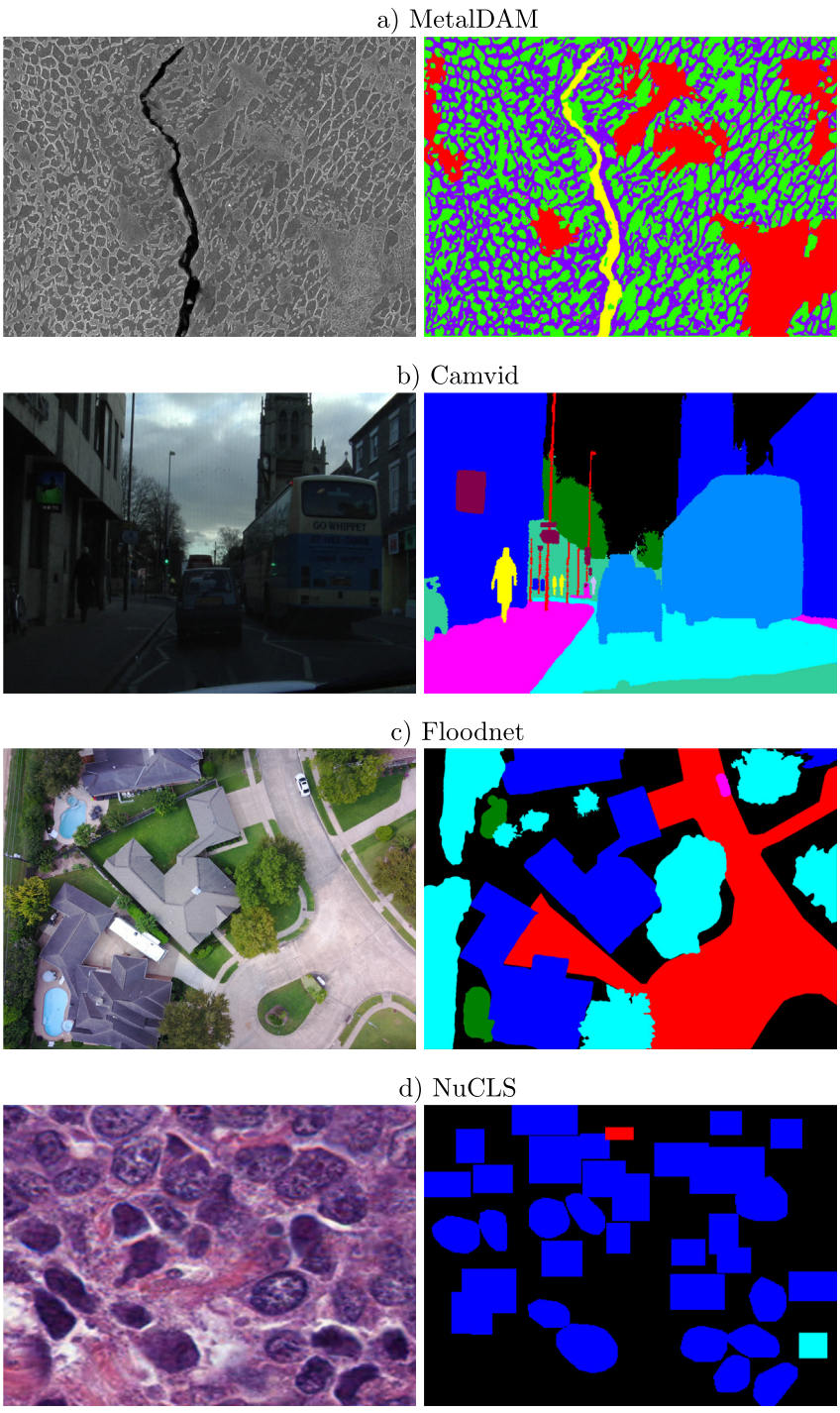


Fig. 3. Comparison of images and their mask: Each subfigure shows an example for each dataset.

Table 1
Summary of the semantic segmentation datasets tested in the experiments.

Dataset	Domain	# Labelled images	# Classes	Resolution
MetalDAM[a]	Microscopy	42	5	1024 × 768
Camvid[b]	Urban Scene	701	12	480 × 384
FloodNet[c]	Satellite	398	10	4000 × 3000
NuCLS[d]	Medical	1744	4	480 × 384

[a] https://dasci.es/transferencia/open-data/metal-dam/.
[b] https://github.com/alexgkendall/SegNet-Tutorial.
[c] https://github.com/BinaLab/FloodNet-Challenge-EARTHVISION2021.
[d] https://sites.google.com/view/nucls/home?authuser=0.

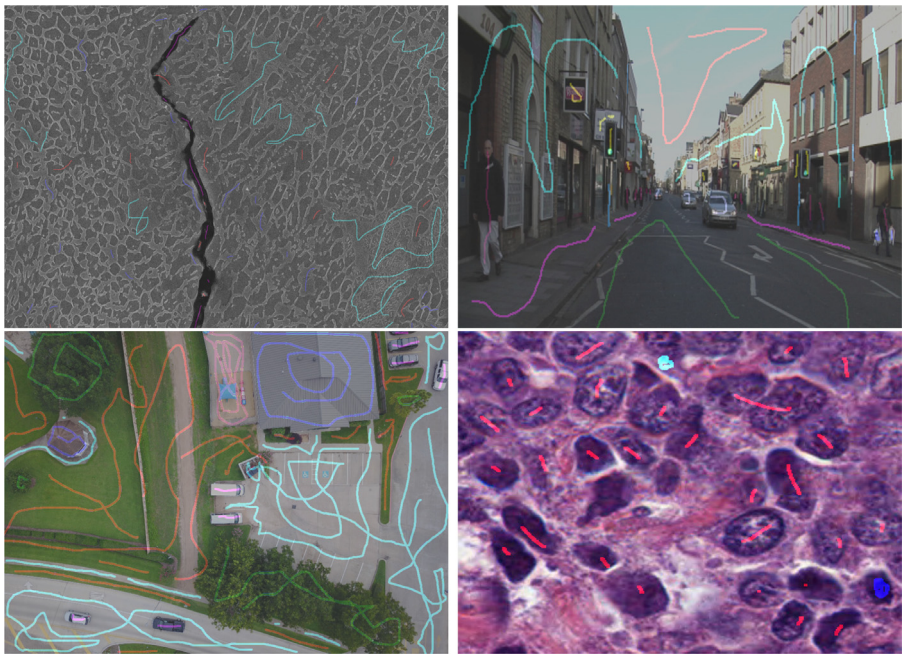


Fig. 4. Comparison of images and their scribble: Each figure represents an example for each dataset,consisting of an image and the scribble that has been manually created.

DeepLabv3+ (Chen et al., 2018) has obtained the best results; therefore, a DeepLabv3+ since they are trained in a fully supervised fashion and also used in RCA, PX_margin, PX_conf, PX_entropy and PX_random methods. Regarding the unsupervised models, as they are trained on each image individually, the 20 most representative images have been selected for each dataset. To identify the most representative images, we applied k-means clustering to the label class distribution of each image and selected the 20 centroids. Regarding the scribble-based method, scribbles have been manually created for each of the 20 images in the four datasets and will be shared along with the implementation of the methods in this article's repository. Fig. 4 shows an example per dataset showing these scribbles on its image.

The optimization of hyperparameters has been performed using Optuna (Akiba et al., 2019), and all unsupervised and SCR methods have been run a minimum of 20 times based on a sequential model optimization using Tree-structured Parzen Estimator (TPE) (Bergstra et al., 2013). As for the PX_margin, PX_conf, PX_entropy and PX_random methods, fixed hyperparameters were selected based on the recommendations of the original paper. Each of the PX methods has been run 10 times, training over N new labelled pixels per image in each run, so that the first training is done on N pixels per image and the last one on $10 \times N$.

Regarding the datasets, the training, validation and test subsets predefined by their authors have been used in each dataset. However, since the labels for the validation and test subsets of FloodNet are not publicly available, the training subset has been divided into three separate subsets, each with the same proportions as the other three datasets. Additionally, the high resolution of the FloodNet images and masks results in longer and more costly training processes. Based on this problem, our results suggest that resizing images to a resolution of 1024×768 does not harm the performance of the models, and all training has been performed on this resolution of images and masks. Concerning the classes in MetalDAM, we have ignored class 3 during the model training and evaluation process, since this class was just partially labelled. Regarding FloodNet and Camvid as well, the image background has also been ignored since this class represents pixels that do not belong to any of the classes to be identified.

3.3. Results and discussion

In this section, we collect the results obtained on each dataset and analyse the impact of each type of label. Since every method has been run 20 times, to show the average result of each method, instead of selecting the best experiment, we report the results of the selected experiment as a function of the median over the mean Intersection Over Union (IOU) of the classes. For the point-based method, the results of the model trained with $10 \times N$ pixels are shown (N being the number of labelled pixels introduced by active learning on each run), since it corresponds to the end of the training pipeline.

First, through Fig. 5 we analyse the mean IOU performance of each method sorted from left to right according to the complexity of the label. Each colour refers to a different type of label:

1. Unsupervised without labels.
2. Weakly supervised with image-level.
3. Weakly supervised with scribbles.
4. Weakly supervised with points.
5. Supervised with fully annotated labels.

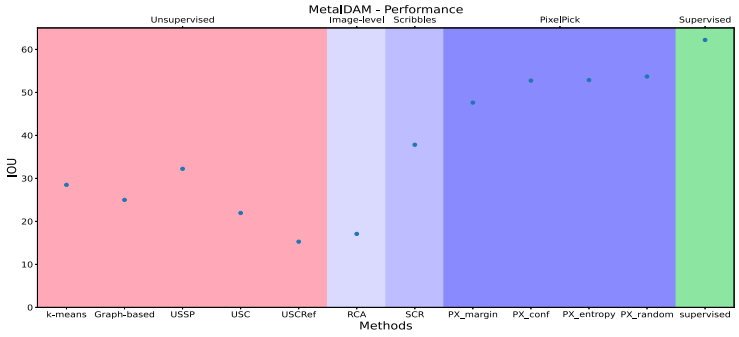
The natural behaviour of the graph, therefore, would be to observe growth on the Y-axis between the different colour bands. Considering the results obtained on the four datasets, we can observe several common behavioural patterns.

An Appendix has been included to show examples of the predictions made by each model compared to the input image and the corresponding label. In this appendix we can see how unsupervised methods obtain segmentations with high granularity, especially in majority classes. As for the image-level method, we can see how its predictions are more oriented towards instance segmentation, since it tends to separate clear objects from the background, as happens in the case of MetalDAM. In the case of scribbles, we see how the SCR method starts to distinguish somewhat more minority classes, although it does not do so with great precision. Finally, the point-based method achieves visually very similar results to the supervised ones.

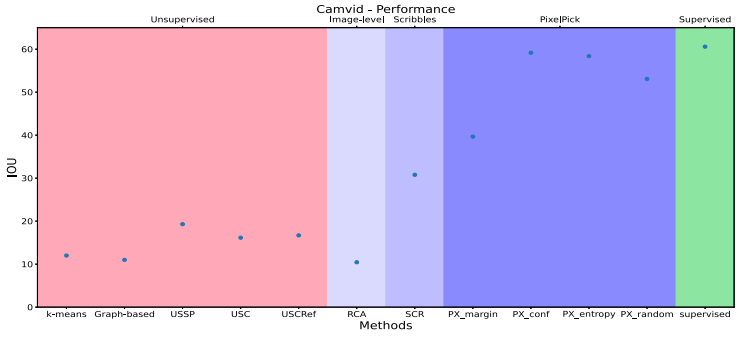
3.3.1. Unsupervised methods

As expected, unsupervised methods generally underperform other methods that incorporate supervision during training, with the exception of the image-level method.

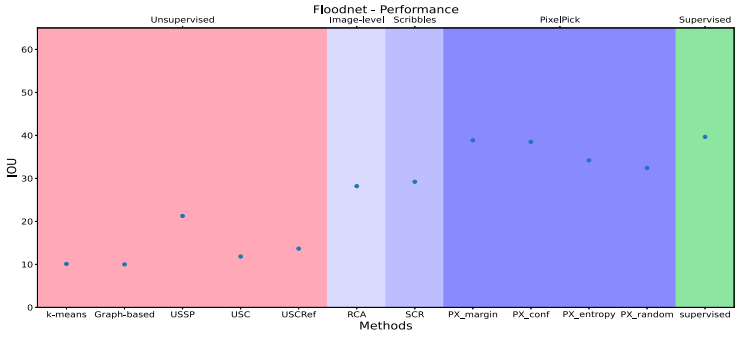
a) MetalDAM results



b) Camvid results



c) Floodnet results



d) NuCLS results

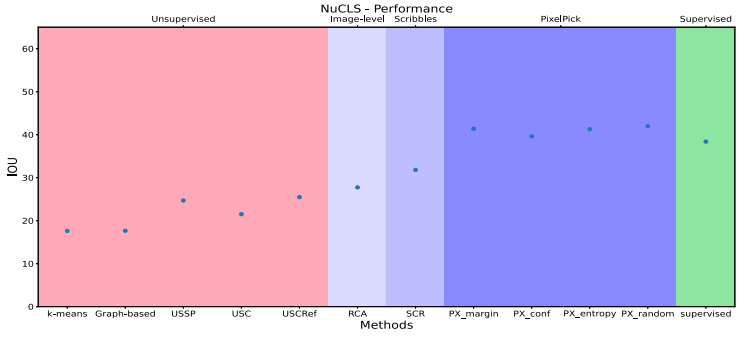


Fig. 5. Supervision effort vs performance trade-off: This figure reports the mean IOU results obtained by each of the selected methods. Each stripe coloured in a different shade collects the methods using a certain kind of label and, as we move between strips from left to right, the annotation effort increases. The subfigures represent the results obtained for each dataset.

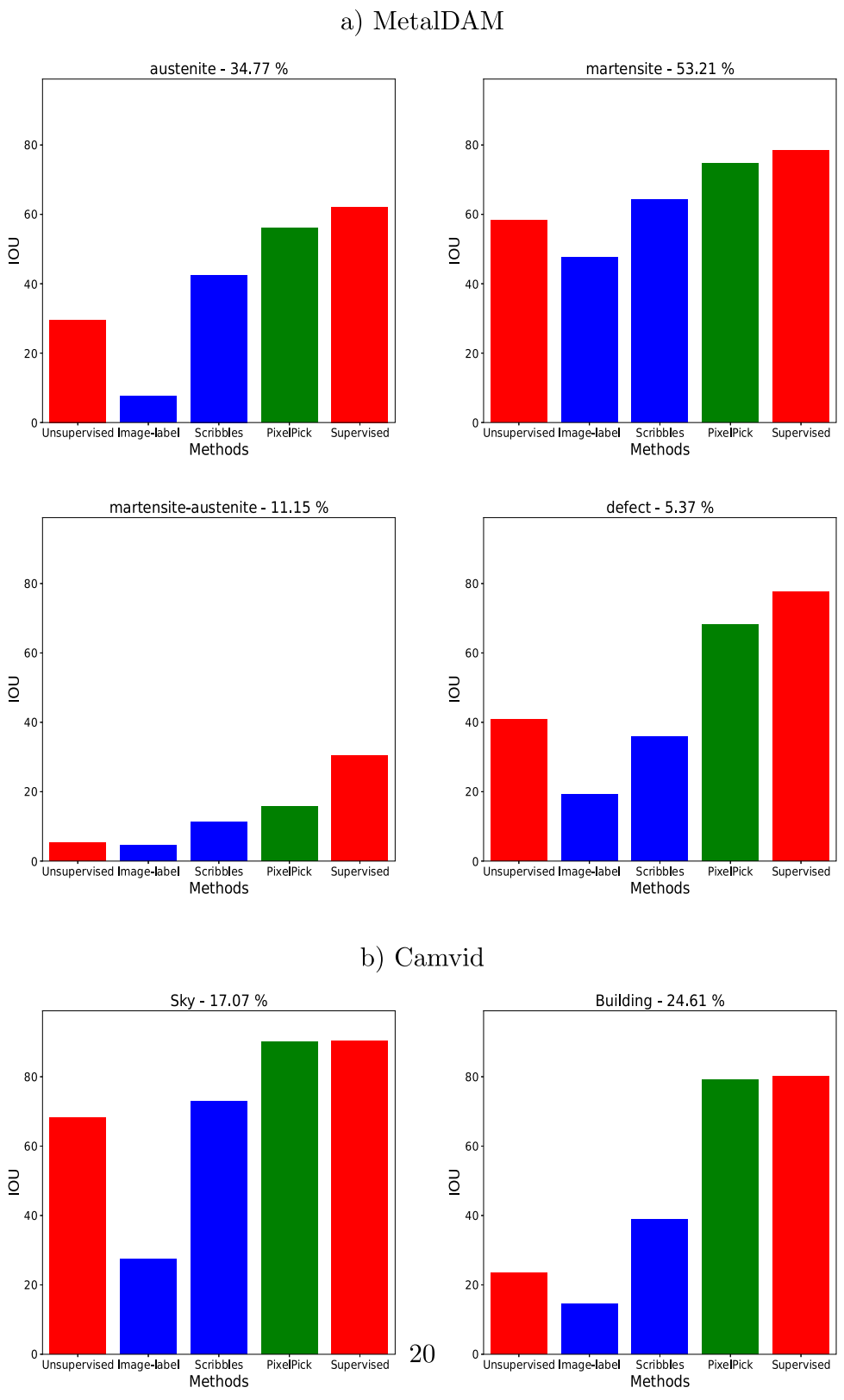


Fig. 6. Comparison of IOU by classes: Each subfigure represents the performance level of the methods based on different types of labels for each of the classes of each dataset. The title of each graph specifies the percentage of pixels in the entire dataset for that class.

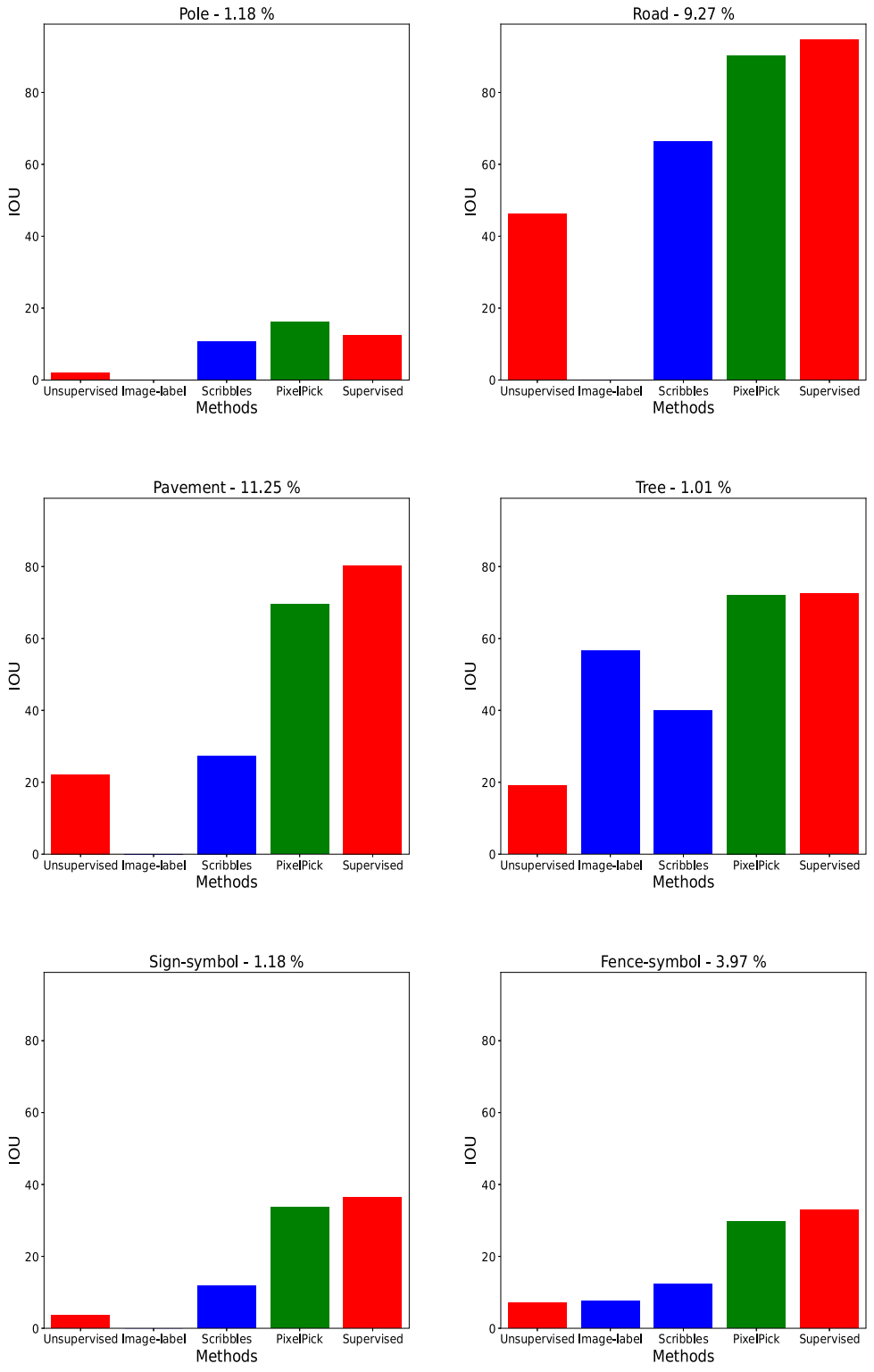
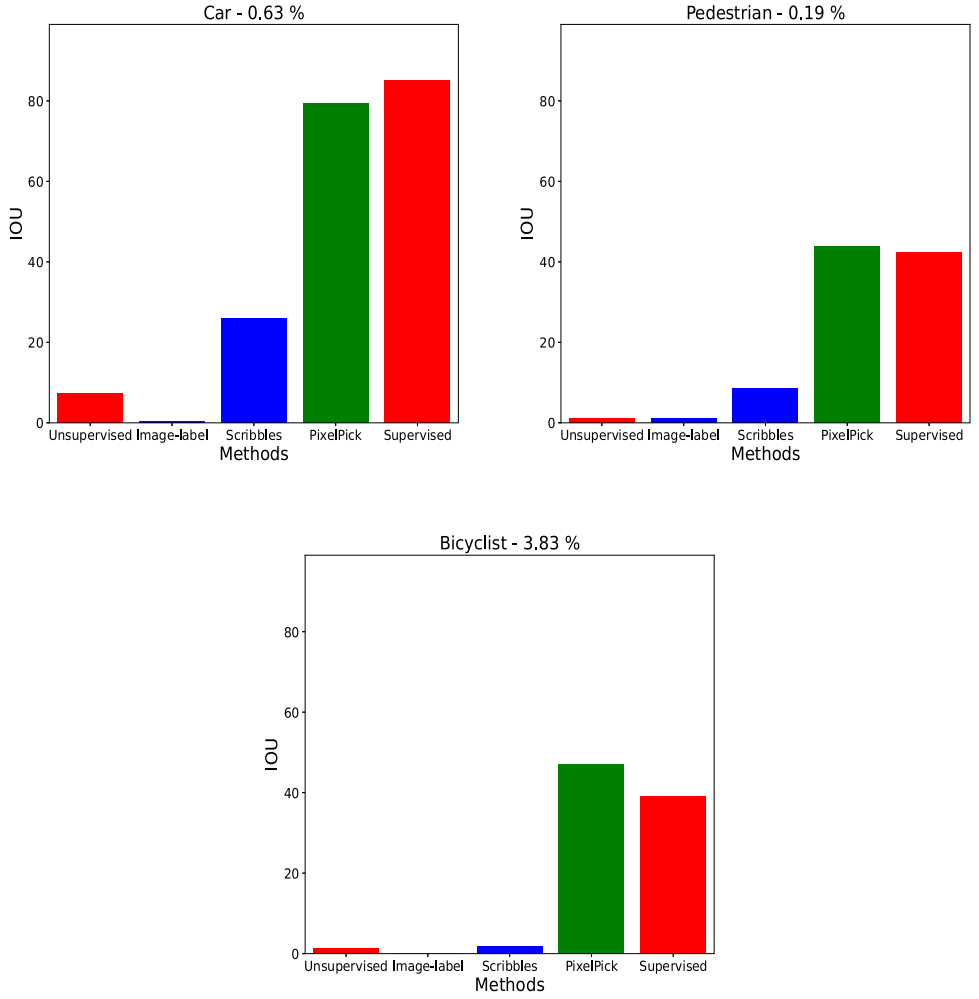


Fig. 6. (*continued*).

However, there are interesting exceptions such as the performance of USSP, which obtains very promising results on all datasets, approaching the scribbles method by 5% on metalDAM. Considering Table 1, one can observe that unsupervised learning achieves a worse performance when the dataset has a larger number of classes. In NuCLS (4 classes) the IOU difference with scribble-based methods is 6%, in MetalDAM (5 classes) 5%, in FloodNet (10 classes) 8% and in Camvid (12 classes) 11%. As can be seen in Fig. 6 unsupervised methods perform well on majority classes, so the more minority classes, the lower the average IOU over all classes.

As a summary, from this analysis of unsupervised methods we can deduce:

- Unsupervised methods obtain highly accurate results in many majority classes, as is the case of Martensite and Austenite in MetalDAM, which make up almost 90% of the labelling, or grass and tree in Floodnet, which make up more than 70% of the



c) FloodNet

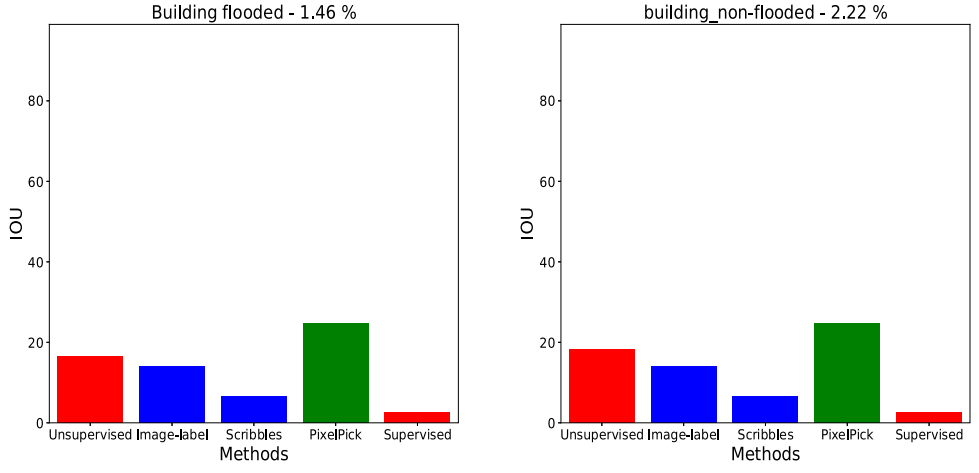


Fig. 6. (*continued*).

labelling. Given this fact, using unsupervised methods as an aid in the labelling process can be a good approach.

- In relation to the previous point, as they obtain good results on majority classes, they perform better on datasets with a smaller number of classes.
- Among the 2 baselines and 3 unsupervised approaches tested, USSP always performs best and compares well with methods such as scribbles and image-level. This suggests that USSP is the optimal unsupervised method for addressing a problem from scratch, in addition to serving as an excellent starting point for comparison.

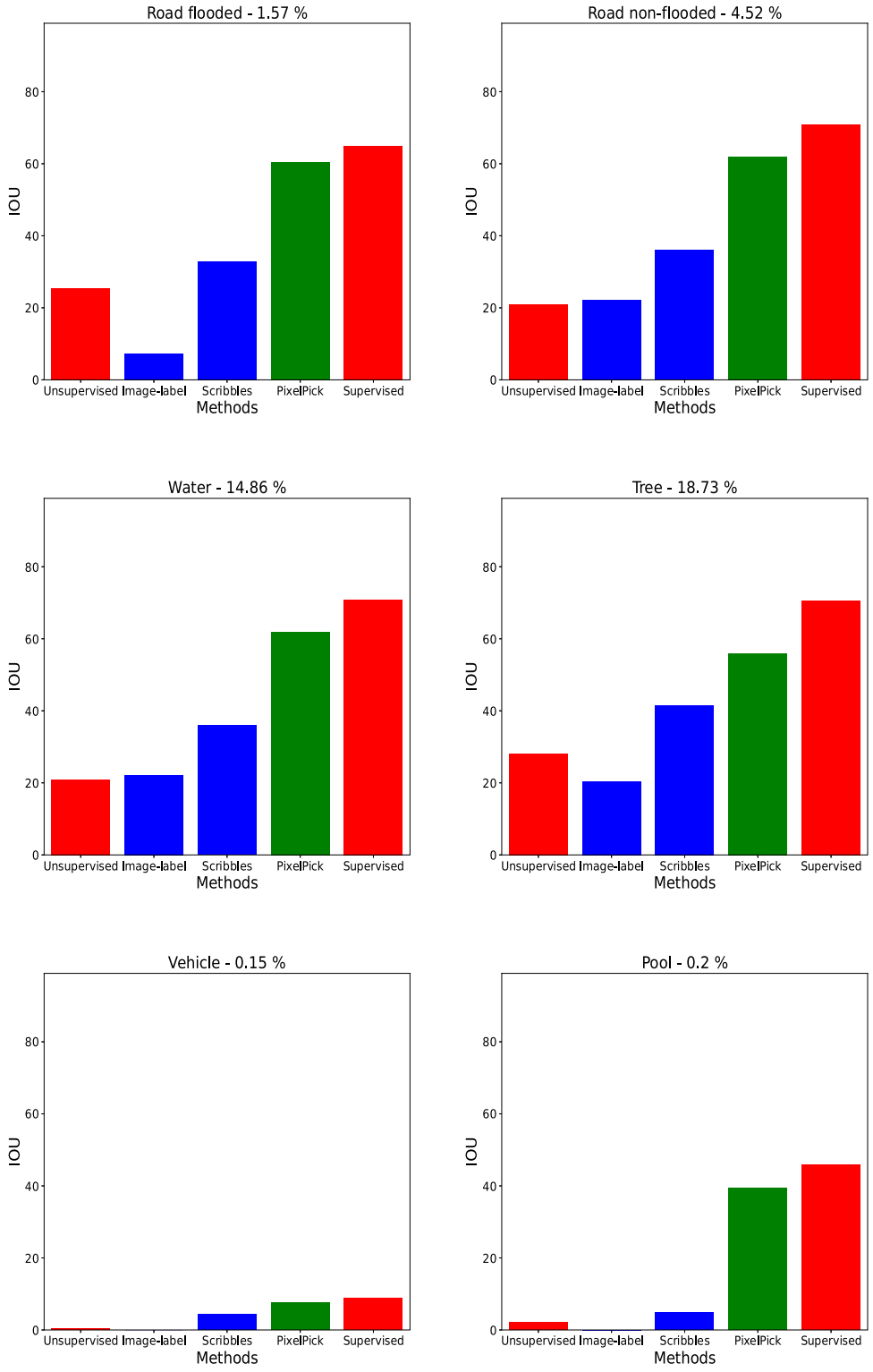
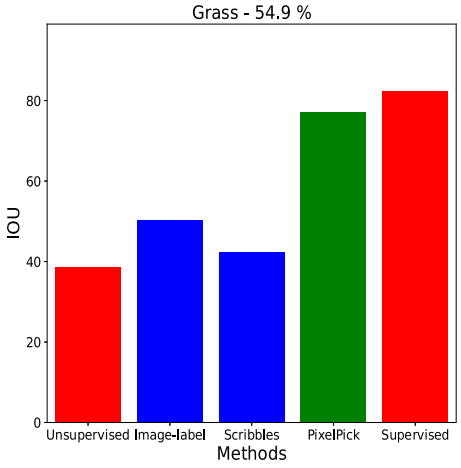


Fig. 6. (*continued*).

3.3.2. *Weakly supervised methods*

Continuing with the weakly supervised methods, in Fig. 5 we can observe that for all four datasets every point-based method outperforms the other weakly supervised methods. However, the difference in performance between these methods varies depending on the dataset.

To better understand these differences, Table 2 provides, for each dataset, information related to the amount of labelled information used by each method. This table shows the percentage of labelled pixels in each image and the number of total pixels used taking into account the number of training images. It is important to take into account both



d) NuCLS

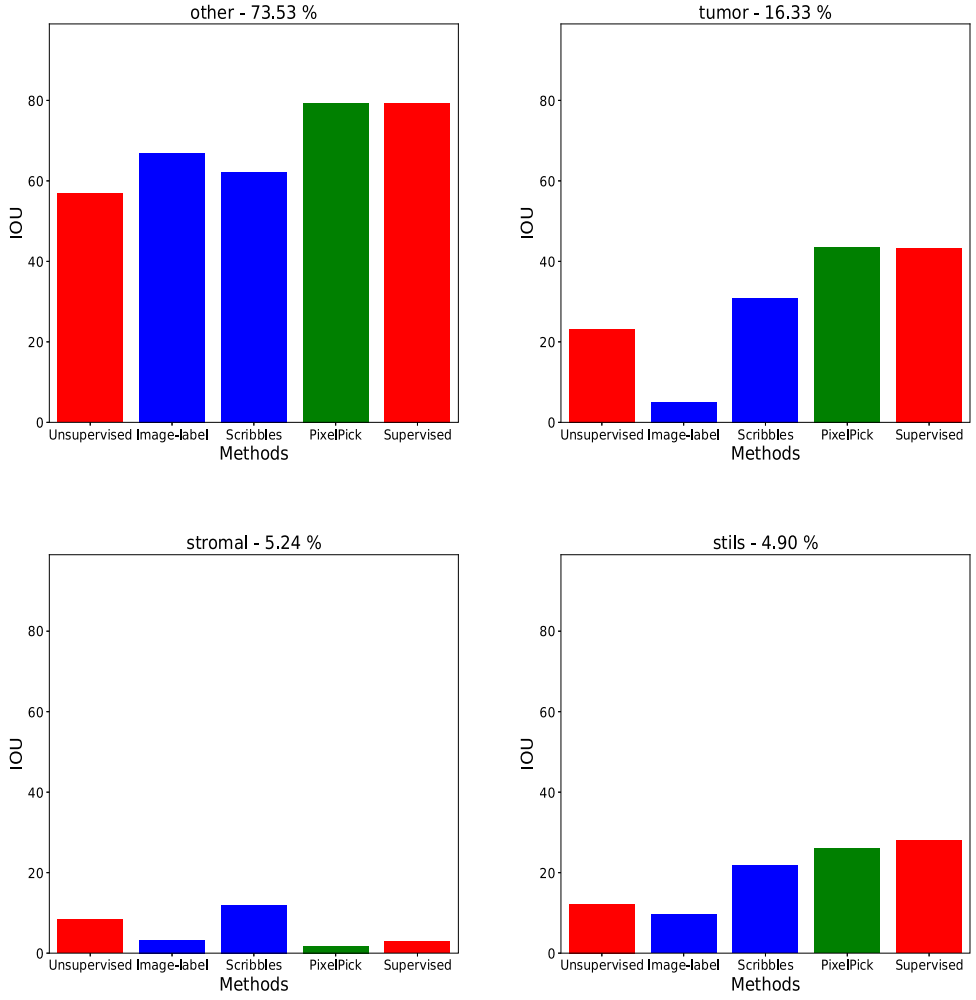


Fig. 6. (*continued*).

measures since not all methods train with the same number of images. We will now analyse the three methods one by one.

First, in Fig. 5 we observe how RCA, the image-level approach, obtains worse results than the rest of the weakly supervised methods. This makes sense since, unlike the other two approaches, RCA knows which classes are present in the image, but has no information regarding their location and extent.

Secondly, the scribble-based method (SCR) always contains more annotated pixels per image than the point-based. However, SCR trains on a single image, while the point-based approach uses all images in the training subset. Therefore, as the number of images increases, the number of labelled pixels remains the same for scribbles and grows for point-based methods. For instance, for MetalDAM, having 29 images, the percentage and number of labelled pixels are higher in scribbles

Table 2
Comparison between amount of information provided vs performance. The table shows the number of images used to train the model, the percentage of labelled pixels used per image, the number of labelled pixels used during the training and the resulting *Mean IOU* value.

MetalDAM				
Approach	# training images	% labelled pixels	# labelled pixels	Mean IOU
Unsupervised	1	0%	0	32.24%
Image-level	29	0%	0	14.78%
Scribbles	1	1.16%	9113	37.83%
Point-based	29	0.01%	2900	53.70%
Supervised	29	100%	22806528	62.20%
Camvid				
Approach	# training images	% labelled pixels	# labelled pixels	Mean IOU
Unsupervised	1	0%	0	19.30%
Image-level	367	0%	0	10.44%
Scribbles	1	3.129%	5748	30.78%
Point-based	367	0.054%	36700	59.17%
Supervised	367	100%	67645440	60.58%
FloodNet				
Approach	# training images	% labelled pixels	# labelled pixels	Mean IOU
Unsupervised	1	0%	0	21.26%
Image-level	318	0%	0	28.20%
Scribbles	1	12.41%	97601	29.23%
Point-based	318	0.054%	31800	38.87%
Supervised	318	100%	312999936	39.65%
NuCLS				
Approach	# training images	% labelled pixels	# labelled pixels	Mean IOU
Unsupervised	1	0%	0	25.47%
Image-level	1481	0%	0	25.56%
Scribbles	1	18.06%	33288	31.82%
Point-based	1481	0.01%	148100	38.20%
Supervised	1481	100%	272977920	38.38%

but, for Camvid, having 367 images, the number of annotated pixels in SCR is lower.

As for the results of the point-based methods, the ones obtained on Camvid reproduce the results of the original authors and confirms that the method obtains similar results for the other three datasets of a very different nature. Comparing the point-based approach with the supervised method, it can be seen that in MetalDAM, which has a reduced number of images, the point-based method is almost 10% behind the supervised method. However, in FloodNet, NuCLS and Camvid, while training with more images, the results are practically equal to the supervised ones, only about 1% worse. This may also be due to the aspect demonstrated by Northcutt et al. (2021) on the problem of model training and evaluation on noisy data sets. MetalDAM contains very tightly intertwined classes, which causes the introduction of many human errors when defining the boundary between classes. In addition, according to the authors, the labelling of certain classes is very ambiguous, leading to cases in which several experts in the area of materials science do not agree on their labelling.

- Despite the fact that scribbles have more labelled information, point-based methods are always superior to scribbles, indicating that better results are obtained with the selection of a smaller volume of sparse pixels based on active learning criteria.
- Based on these results we can see how, apparently, increasing the percentage of supervision has less impact than increasing the number of images with a lower level of supervision.
- Among the 3 weakly supervised approaches tested, the point-based method obtains the best results on very diverse datasets. Its results are very close to the supervised approach, which makes it a highly recommended solution for real-world problems.

4. Conclusions

Our study explores one of the biggest challenges when dealing with semantic segmentation techniques: annotating a dataset with the lowest possible error rate in an efficient manner. We had conducted the first comprehensive study of state-of-the-art methods with different levels of supervision. We had tested 12 methods on 4 datasets from different domains, including 2 image processing baselines, 3 unsupervised methods, 6 weak learning approaches, and 1 fully supervised approach. The results obtained had shown that spending a lot of time on pixel-level annotations is generally unnecessary. We had also provided guidance on the performance offered by different kinds of labels, such as mean IOU vs labelling complexity.

We have found that unsupervised methods, such as USSP (Kanezaki, 2018), perform very well on large classes. For instance, USSP is able to segment the two majority classes that represent the 90% of the MetalDAM labelling. By using these pre-annotations, the time spent on labelling MetalDAM could be reduced from more than 4 days to less than 10 h.

Recent studies, such as the one referenced in Northcutt et al. (2021), have emphasized the importance of incorporating a high level of noise in labels. These studies showed that models perform better when adjustments are made to account for such noise. We have found that the point-based method tested in this paper outperforms the supervised approach on certain minority classes. This is because minor classes are left out during labelling, which causes models trained from these human annotations to overlook them in their predictions.

Finally, we observed that image-level methods within the weakly supervised methods obtained significantly different results depending on the dataset. As they do not provide information about which class every pixel belongs to, their ability to distinguish between classes

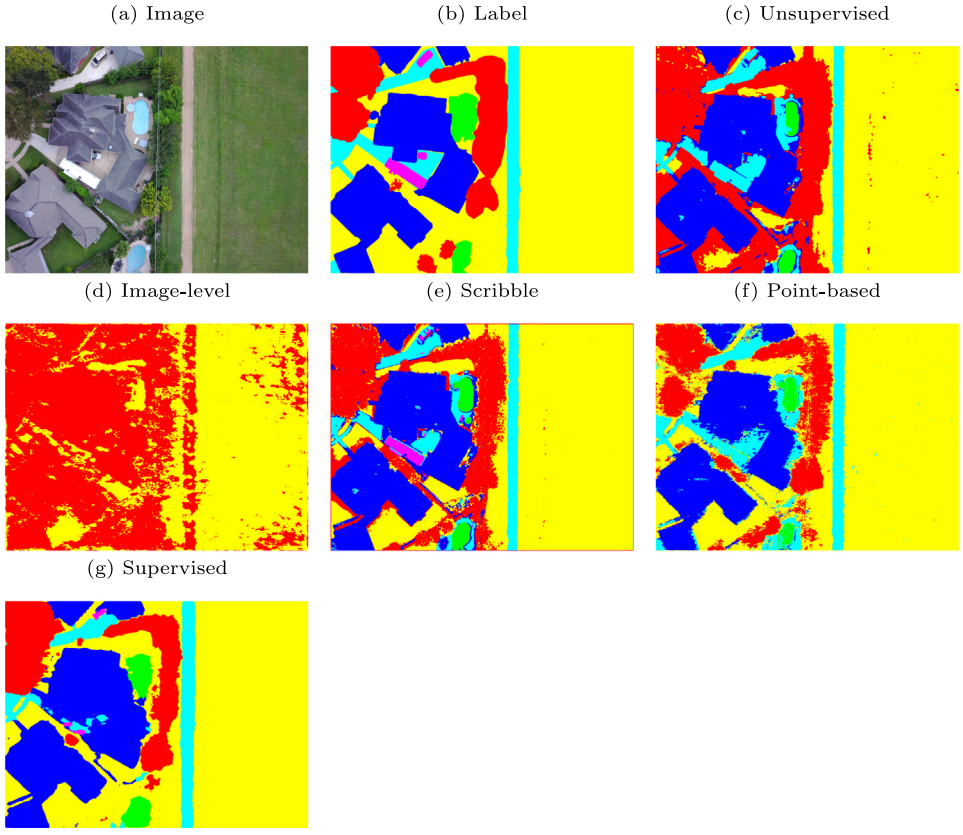


Fig. A.7. FloodNet: Image, label and predictions of each learning approach.

is considerably lower. Although scribbles contain more labelled pixels than point-based methods, they consistently perform worse. This leads us to reflect on the impact of using sparse labels where the labelled information is scattered throughout the image, rather than annotating several pixels close to each other. In our experiments, point-based approaches outperformed scribble-based methods. However, an in-depth analysis of the efforts required for scribble labelling and point-based labelling could be very interesting, but it is beyond the scope of our paper. Future work will include the study of the performance of weak learning annotations using various types of labels with different amounts of labelled information.

CRediT authorship contribution statement

Marta Fernández-Moreno: Conceptualization, Methodology, Research. **Bo Lei:** Research, Reviewing, Editing. **Elizabeth A. Holm:** Supervision, Reviewing. **Pablo Mesejo:** Conceptualization, Methodology, Research, Supervision, Reviewing. **Raúl Moreno:** Conceptualization, Methodology, Research, Supervision, Reviewing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Marta Fernandez Moreno reports a relationship with ArcelorMital Global Research and Development Center - Library that includes: employment. Raul Moreno Salinas reports a relationship with ArcelorMital Global Research and Development Center - Library that includes: employment. Pablo Mesejo Santago reports a relationship with University of Granada Department of Computer Science and Artificial Intelligence that includes: employment. Elizabeth A. Holm reports a relationship with Carnegie Mellon University that includes: employment.

Data availability

https://github.com/martafdezmAM/lessen_supervision

Acknowledgements

Funding: This work was supported by the National Science Foundation, United States [grant numbers CMMI, 1826218]; and the Air Force D3OM2S Center of Excellence [grant number FA8650-19-2-5209]. This work was also supported by the Spanish Ministry of Science and Innovation, the Andalusian Government, and European Regional Development Funds (ERDF) under grants CONFIA (PID2021-122916NB-I00) and FORAGE (B-TIC-456-UGR20)

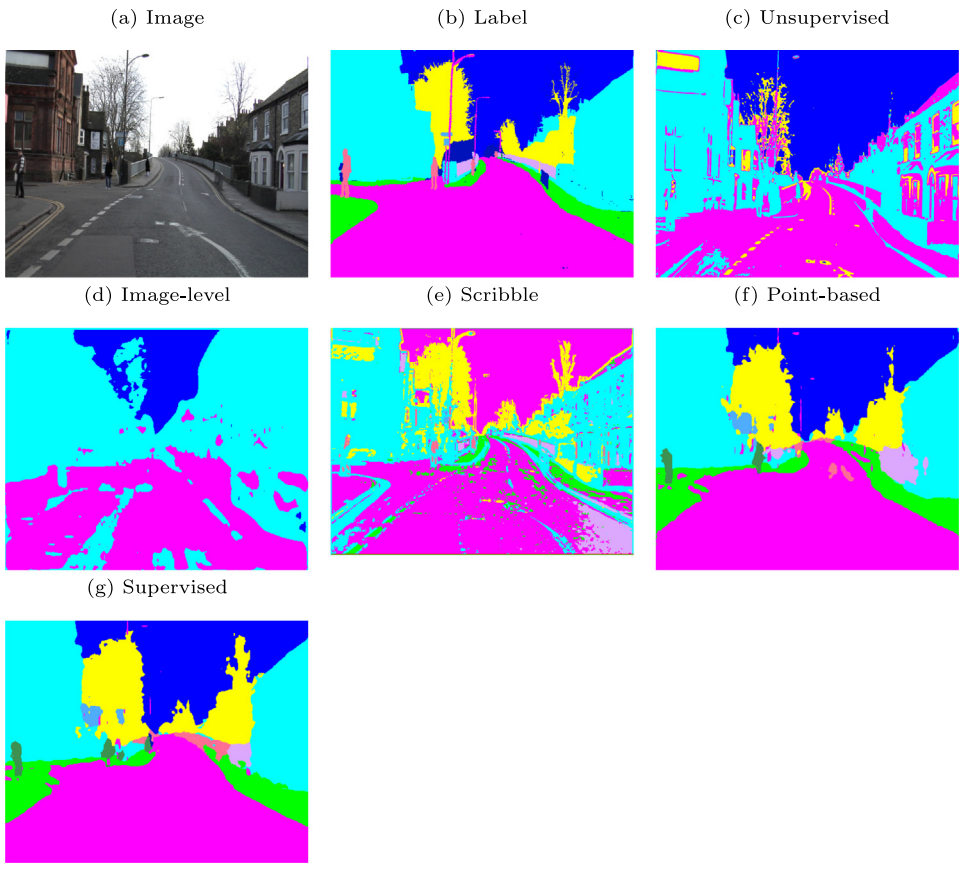


Fig. A.8. Camvid: Image, label and predictions of each learning approach.

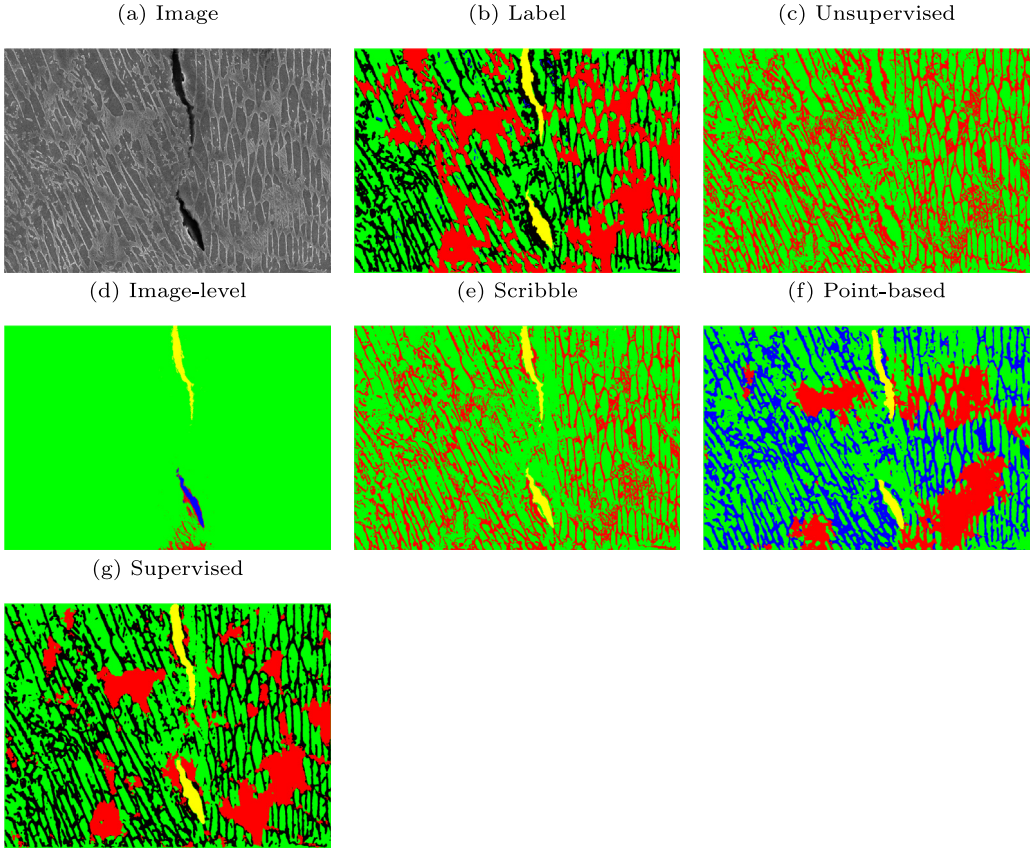


Fig. A.9. Metaldam: Image, label and predictions of each learning approach.

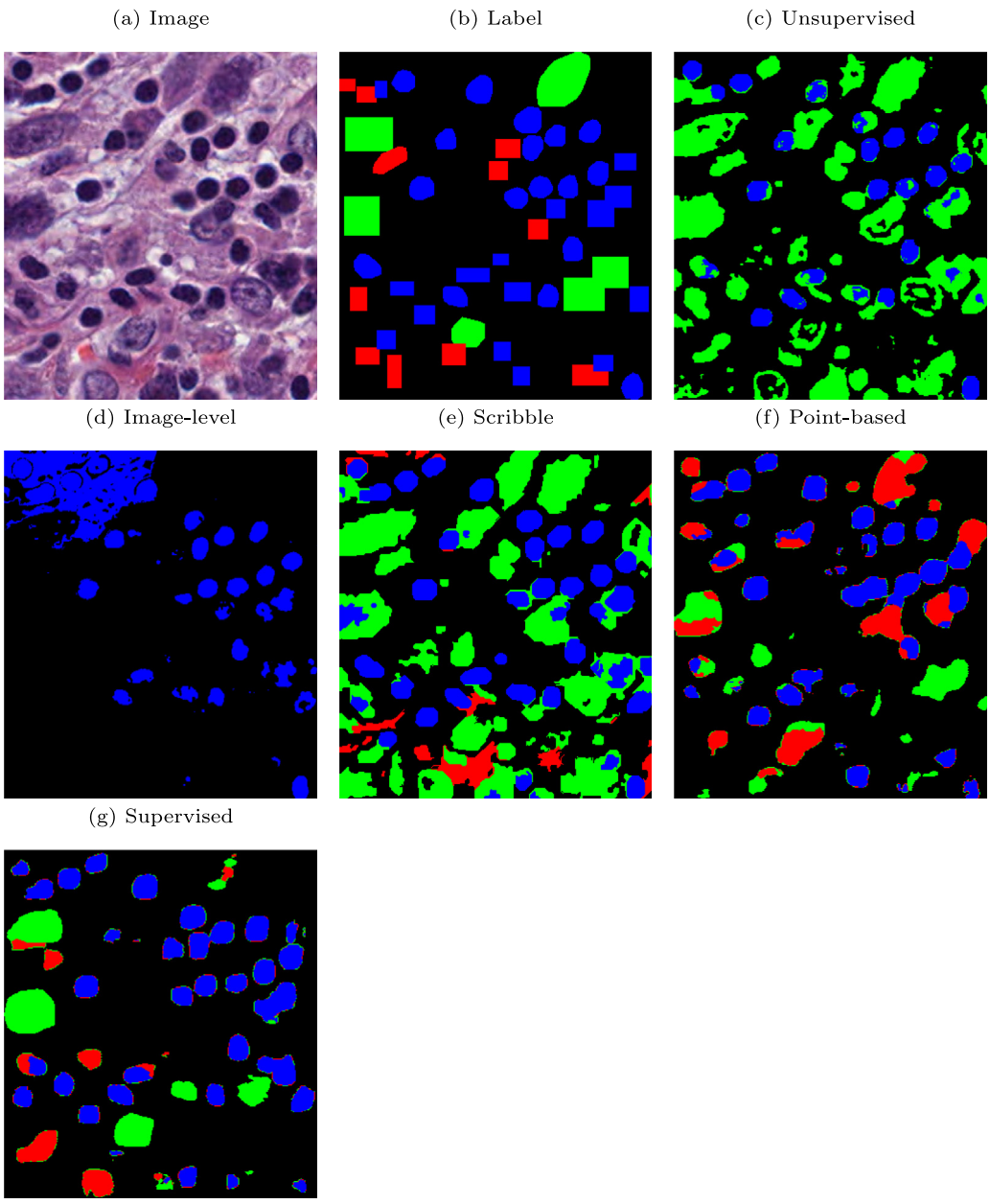


Fig. A.10. NuCLS: Image, label and predictions of each learning approach.

Appendix

Figs. A.7–A.10 show a comparison between a random image and its label from each dataset with the output prediction of the model from each learning approach.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework. In: International Conference on Knowledge Discovery & Data Mining KDD. pp. 2623–2631.

Amgad, M., Atteya, L.A., Hussein, H., Mohammed, K.H., Hafiz, E., Elsebaie, M.A., Alhusseiny, A.M., AlMoslemany, M.A., Elmatboly, A.M., Pappalardo, P.A., et al., 2022. NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. GigaScience 11.

Bergstra, J., Yamins, D., Cox, D.D., et al., 2013. Hyperopt: A Python library for optimizing the hyperparameters of machine learning algorithms. In: Annual Scientific Computing with Python Conference, Vol. 13. SciPy, p. 20.

Brostow, G.J., Fauqueur, J., Cipolla, R., 2009. Semantic object classes in video: A high-definition ground truth database. Pattern Recognit. Lett. 30 (2), 88–97.

Caron, M., Bojanowski, P., Joulin, A., Douze, M., 2018. Deep clustering for unsupervised learning of visual features. In: IEEE European Conference on Computer Vision. pp. 132–149.

Chan, L., Hosseini, M., Plataniotis, K., 2021. A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. Int. J. Comput. Vis. 129 (2), 361–384.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: IEEE European Conference on Computer Vision. pp. 801–818.

Cho, J.H., Mall, U., Bala, K., Hariharan, B., 2021. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 16794–16804.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3213–3223.

Dash, B., Mishra, D., Rath, A., Acharya, M., 2010. A hybridized K-means clustering approach for high dimensional dataset. Int. J. Eng. Sci. Technol. 2 (2), 59–66.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255.

Dong, J., Cong, Y., Sun, G., Fang, Z., Ding, Z., 2021. Where and how to transfer: Knowledge aggregation-induced transferability perception for unsupervised domain adaptation. IEEE Trans. Pattern Anal. Mach. Intell..

Dong, J., Cong, Y., Sun, G., Zhong, B., Xu, X., 2020. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In: Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition. pp. 4023–4032.

Dong, N., Xing, E.P., 2018. Few-shot semantic segmentation with prototype learning. In: British Machine Vision Conference, Vol. 3, no. 4.

Du, Y., Fu, Z., Liu, Q., Wang, Y., 2022. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4320–4329.

Felzenszwalb, P.F., Zabih, R., 2010. Dynamic programming and graph algorithms in computer vision. IEEE Trans. Pattern Anal. Mach. Intell. 33 (4), 721–740.

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Gläser, C., Timm, F., Wiesbeck, W., Dietmayer, K., 2021. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Trans. Intell. Transp. Syst. 22 (3), 1341–1360.

Guo, Y., Liu, Y., Georgiou, T., Lew, M.S., 2018. A review of semantic segmentation using deep neural networks. Int. J. Multimed. Inf. Retrieval 7 (2), 87–93.

Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T., 2022. Unsupervised semantic segmentation by distilling feature correspondences. arXiv preprint arXiv:2203.08414.

Holm, E.A., Cohn, R., Gao, N., Kitahara, A.R., Matson, T.P., Lei, B., Yarasi, S.R., 2020. Overview: Computer vision and machine learning for microstructural characterization and analysis. Metall. Mater. Trans. A 51, 5985–5999.

Hong, S., Noh, H., Han, B., 2015. Decoupled deep neural network for semi-supervised semantic segmentation. In: Advances in Neural Information Processing Systems, Vol. 28.

Hsu, C.-C., Hsu, K.-J., Tsai, C.-C., Lin, Y.-Y., Chuang, Y.-Y., 2019. Weakly supervised instance segmentation using the bounding box tightness prior. Adv. Neural Inf. Process. Syst. 32.

Hwang, J.-J., Yu, S.X., Shi, J., Collins, M.D., Yang, T.-J., Zhang, X., Chen, L.-C., 2019. SegSort: Segmentation by discriminative sorting of segments. In: IEEE International Conference on Computer Vision. pp. 7333–7343.

Ji, X., Henriques, J.F., Vedaldi, A., 2019. Invariant information clustering for unsupervised image classification and segmentation. In: IEEE International Conference on Computer Vision. pp. 9865–9874.

Kanezaki, A., 2018. Unsupervised image segmentation by backpropagation. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1543–1547.

Kim, H., Inoue, J., Kasuya, T., 2020a. Unsupervised microstructure segmentation by mimicking metallurgists' approach to pattern recognition. Sci. Rep. 10 (1), 1–11.

Kim, W., Kanezaki, A., Tanaka, M., 2020b. Unsupervised learning of image segmentation based on differentiable feature clustering. IEEE Trans. Image Process. 29, 8055–8068.

Lateef, F., Ruichek, Y., 2019. Survey on semantic segmentation using deep learning techniques. Neurocomputing 338, 321–348.

Li, B., Shi, Y., Qi, Z., Chen, Z., 2018. A survey on semantic segmentation. In: IEEE International Conference on Data Mining Workshops. pp. 1233–1240.

Lin, D., Dai, J., Jia, J., He, K., Sun, J., 2016. ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3159–3167.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: IEEE European Conference on Computer Vision. pp. 740–755.

Liu, X., Song, L., Liu, S., Zhang, Y., 2021a. A review of deep-learning-based medical image segmentation methods. Sustainability 13 (3), 1224–1253.

Liu, Y., Zhang, W., Wang, J., 2021b. Source-free domain adaptation for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1215–1224.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Luengo, J., Moreno, R., Sevillano, I., Charte, D., Peláez-Vegas, A., Fernández-Moreno, M., Mesejo, P., Herrera, F., 2022. A tutorial on the segmentation of metallographic images: Taxonomy, new MetalDAM dataset, deep learning-based ensemble model, experimental analysis and challenges. Inf. Fusion 78, 232–253.

Muhadi, N.A., Abdullah, A.F., Bejo, S.K., Mahadi, M.R., Mijic, A., 2020. Image segmentation methods for flood monitoring system. Water 12 (6), 1825.

Northcutt, C.G., Athalye, A., Mueller, J., 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749.

Papadopoulos, D.P., Uijlings, J.R.R., Keller, F., Ferrari, V., 2017. Extreme clicking for efficient object annotation. In: IEEE International Conference on Computer Vision. pp. 4940–4949.

Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., Murphy, R.R., 2021. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. IEEE Access 9, 89644–89654.

Shin, G., Xie, W., Albanie, S., 2021. All you need are a few pixels: Semantic segmentation with PixelPick. In: IEEE International Conference on Computer Vision Workshops). pp. 1687–1697.

Toldo, M., Maracani, A., Michieli, U., Zanuttigh, P., 2020. Unsupervised domain adaptation in semantic segmentation: A review. Technologies 8 (2), 35.

Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Van Gool, L., 2021. Unsupervised semantic segmentation by contrasting object mask proposals. In: IEEE International Conference on Computer Vision. pp. 10052–10062.

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G., 2018. Understanding convolution for semantic segmentation. In: IEEE Winter Conference on Applications of Computer Vision. pp. 1451–1460.

Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K., 2022. Medical image segmentation using deep learning: A survey. IET Image Process. 16 (5), 1243–1267.

Yakubovskiy, P., 2020. Segmentation models pytorch. GitHub Repository, GitHub, https://github.com/qubvel/segmentation_models.pytorch.

Zhou, T., Li, L., Li, X., Feng, C.-M., Li, J., Shao, L., 2021. Group-wise learning for weakly supervised semantic segmentation. IEEE Trans. Image Process. 31, 799–811.

Zhou, T., Zhang, M., Zhao, F., Li, J., 2022. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4299–4309.