All you need are a few pixels: semantic segmentation with PIXELPICK

Gyungin Shin[1] Weidi Xie[1] Samuel Albanie[1,2]
[1] Visual Geometry Group, University of Oxford, UK
[2] Department of Engineering, University of Cambridge, UK
{gyungin, weidi, albanie}@robots.ox.ac.uk
https://www.robots.ox.ac.uk/~vgg/research/pixelpick

Abstract

A central challenge for the task of semantic segmentation is the prohibitive cost of obtaining dense pixel-level annotations to supervise model training. In this work, we show that in order to achieve a good level of segmentation performance, all you need are a few well-chosen pixel labels.

We make the following contributions: (i) We investigate the semantic segmentation setting in which labels are supplied only at sparse pixel locations, and show that deep neural networks can use a handful of such labels to good effect; (ii) We demonstrate how to exploit this phenomenon within an active learning framework, termed PIXELPICK, *to radically reduce labelling cost, and propose an efficient "mouse-free" annotation strategy to implement our approach; (iii) We conduct extensive experiments to study the influence of annotation diversity under a fixed budget, model pretraining, model capacity and the sampling mechanism for picking pixels in this low annotation regime; (iv) We provide comparisons to the existing state of the art in semantic segmentation with active learning, and demonstrate comparable performance with up to two orders of magnitude fewer pixel annotations on the* CAMVID, CITYSCAPES *and* PASCAL VOC 2012 *benchmarks; (v) Finally, we evaluate the efficiency of our annotation pipeline and its sensitivity to annotator error to demonstrate its practicality.*

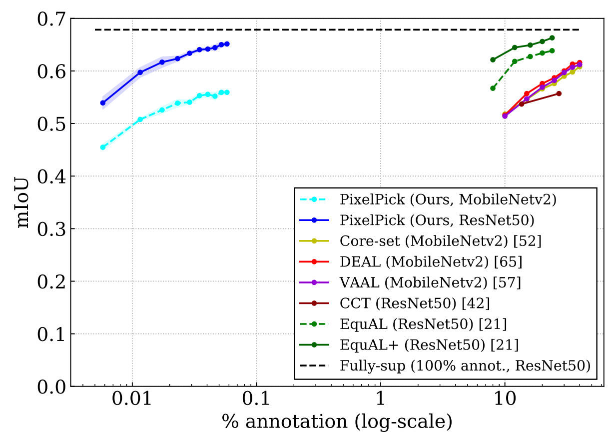


Figure 1: **All you need are a few pixels:** We show that deep neural networks can obtain remarkable performance with just a handful of labelled pixels per image whose spatial coordinates are proposed by the model, rather than the human annotator. We compare our approach, PIXELPICK, with existing active learning and semi-supervised approaches on the CAMVID dataset [7] (see Sec. 4 for further details).

1. Introduction

The coupling of deep neural networks and large-scale labelled datasets has yielded significant progress on a host of core machine perception tasks. A key challenge of training these models is their need for considerable quantities of annotation, which can be prohibitively expensive to collect for applications that require either specialised annotators such as medical image diagnostics [1, 23, 56, 62], or fine-grained labels, such as for detection and segmentation [36].

Semantic segmentation, in particular, has proven valuable for decision making in a variety of applications such as digital pathology [59], remote sensing [64] and autonomous driving [67]. However, its requirement of per-pixel annotations raises significant scalability challenges—on average more than 1.5 hours of annotation and quality control was required for each image in the CITYSCAPES segmentation dataset [13].

The objective of this work is to propose a simple yet effective approach for training a good semantic segmentation model at minimal annotation cost. Our approach is motivated by three observations: (1) Within a given image, pixels exhibit significant spatial mutual information; (2) Deep neural networks possess a strong inductive bias that renders them appropriate for modelling these spatial dependencies [60]; (3) Collecting mask, scribble or click annotations requires annotators to "localise and classify" using a mouse or trackpad. By contrast, assigning a class to a pixel proposal can be "mouse-free", requiring instead only a "classify" task without a localisation component (and which can be performed via a single key-press). The first two factors imply that densely labelling all pixels in images may be highly redundant, while the third suggests the possibility of designing an efficient sparse pixel labelling strategy. Several questions then arise: *how many sparse pixel labels are needed to achieve good performance? how should those pixel locations be selected?* and *how can the selected pixels be annotated efficiently?*

In this paper, we address these questions through the lens of *active learning* [2, 54]. In contrast to passive supervised learning (where the model is tasked to learn the mapping from a fixed set of input-output pairs), active learning considers a dynamic scenario in which a model can interactively request labels for the samples that it believes will be most useful for solving a given task. Our proposed PIXELPICK framework adopts this paradigm, learning a model for semantic segmentation by alternating between training on previously labelled pixels and requesting new labels.

We make the following contributions: (i) We study the problem setting in which labels are supplied at the level of sparse pixels and show that with only a small collection of such labels, modern deep neural networks can achieve good performance; (ii) We show how this phenomenon can be exploited with an efficient and practical "mouse-free" annotation strategy as part of a proposed PIXELPICK active learning framework; (iii) We perform a series of experiments into factors that affect model performance in the low-annotation regime: annotation diversity, architectural choices and the design of the sampling mechanisms for selecting most useful pixels; (iv) We compare with other state of the art active learning approaches on standard segmentation benchmarks: CAMVID, CITYSCAPES and PASCAL VOC 2012, where we demonstrate comparable segmentation performance with significantly lower annotation budget (Fig. 1); (v) Lastly, we assess PIXELPICK from the perspective of practical deployment, assessing its annotation efficiency and robustness.

2. Related work

Our work is related to several themes of research that have sought to minimise labelling costs for semantic segmentation, as discussed next.

Weakly-supervised semantic segmentation. Many weak supervisory signals have been explored in the literature as a pragmatic compromise between fully supervised [37] and fully unsupervised approaches to semantic segmentation [26]. These cues include scribbles [35], eye tracking [43], object pointing [3, 45, 12], web-queried samples [27], bounding boxes [15, 28, 58], extreme clicks for objects [44, 39] and image-level labels [70, 63, 17]. Differently from these approaches, we gather labels at sparse pixel locations proposed by the model itself, rather than at locations selected by the annotator, and show that very few such annotations are needed for good performance.

Interactive annotation. There is rich body of computer vision literature considering the related problem of accelerating *interactive* annotation. The seminal work of [6] demonstrated how to exploit scribbles to indicate the foreground/background appearance model and leverage graph-cuts for segmentation [5]. This was later extended to the use of multiple scribbles on both object and background, applied to annotating objects in videos [41]. [48] exploited 2D bounding boxes provided by the annotator and performed pixel-wise foreground/background labelling using EM. Recent work [10] tasks a model with sequentially producing the vertices of a polygon outlining an object, given an appropriate crop. As with the weakly-supervised signals described above, these methods are passive in the sense that the labelling process is driven by the human annotator, rather than the model.

Semi-supervised semantic segmentation. Inspired by classical self-labelling approaches which aim to leverage unlabelled data to improve a classifier [51, 68], a number of semi-supervised approaches have been developed to make use of pseudo-labelling algorithms [30] for semantic segmentation in a low-annotation regime. Consistency-based pseudo-labelling methods have recently demonstrated promising results, highlighting the important role of aggressive data augmentations when only a small number of densely annotated images or regions are available [42, 18].

Our approach differs from theirs in several ways: (i) our model is trained from sparse pixel annotations, rather than a small number of densely labelled images, (ii) we employ active learning (samples are dynamically selected and queried for annotation by the model), which, as we show through experiments, brings additional improvements. We compare our approach quantitatively with theirs in Sec. 4.3.

Active learning for semantic segmentation. At its core, active learning is a set selection problem; the aim being to determine the most informative subset of samples to acquire labels for, given a labelling budget [2, 54, 32, 19, 69]. In this case the maximally informative labelled-pixel subset is the one which yields the lowest generalisation error when used to train a supervised semantic segmentation model. Prior work targeting segmentation has investigated strategies to select superpixels that induce the maximum label change for a CRF on the training set by using weak (image-level category) supervision [61], incorporate geometric constraints [29, 40] and propagate foreground masks to large-scale image collections [25]. For foreground segmentation of medical imagery, FCNs [37] have been coupled with bootstrapping [66], and U-Nets [47] with dropout-based Monte Carlo estimates of uncertainty [22] to drive label acquisition via uncertainty sampling. The strategy of learning an estimator for difficult regions [69] has proven effective as a basis for selecting which images should be densely labelled for semantic segmentation [65]. Adversarial learning has also been explored to query labels that align distributions of labelled and unlabelled images [57].

More closely related to our work, prior studies have considered *region-based* sampling strategies for semantic segmentation, employing reinforcement learning [9], equivariance constraints [21] and learned estimators of labelling cost [38]. In contrast to these lines of research, our work aims to introduce a more efficient paradigm of active learning for segmentation, which is to train models by only querying sparse pixel annotations (removing the localisation component of the annotation task). Concurrent to our work, Cai et al. [8] explore the use of super-pixels with active learning. We compare our work with theirs and with other region-based sampling strategies noted above in Sec. 4.

3. Method

In this section, we describe the problem formulation and introduce our framework for pixel-level active learning semantic segmentation in Sec. 3.1. We then detail our mouse-free annotation tool to efficiently implement the framework in Sec. 3.2.

3.1. PixelPick framework

We seek to train a model for semantic segmentation with *pool-based active learning* [53], in which we alternate between training a model on available annotation and requesting labels for unlabelled samples from an oracle (see Fig. 2).

More formally, let $\mathcal{X} \subset \mathbb{R}^{H \times W \times 3}$ denote the space of colour images and let $\Phi(\cdot; \Theta) : \mathcal{X} \rightarrow \mathcal{Y}^{H \times W}$ represent a ConvNet with parameters Θ that maps a given image to a grid of elements in a C-class semantic label space (here $\mathcal{Y}$ denotes the $(C-1)$-simplex, i.e. $\mathcal{Y} = \{(p_1, \ldots, p_C) \in [0,1]^C : \sum_{p=1}^{C} p_i = 1, p_i \geq 0\}$). We assume access to an initial unlabelled pool of N images, $\mathcal{D}_U$, indexed by the $H \times W \times N$ pixel coordinate lattice, Ω, and an annotation database, $\mathcal{D}_L^0$, initialised to an empty state.

At the k^{th} round of learning, a batch of $B \in \mathbb{N}$ pixel coordinates, $\omega_k \subset \Omega$, are sampled by an acquisition function, $\mathcal{A}$, using the predictions of the model trained in the previous round, $\Phi(\cdot; \Theta_{k-1})$, on the unlabelled pool $\mathcal{D}_U$, i.e. $\mathcal{A}(\mathcal{D}_U, \Phi(\cdot; \Theta_{k-1})) = \omega_k \subset \Omega$. The sampled pixel coordinates ω_k are then sent to an oracle for labelling to produce a corresponding set of one-hot labels $\{y_u \in \mathcal{Y} : u \in \omega_k\}$ that are added to the latest version of the annotation database, $\mathcal{D}_L^{k-1}$. Finally, the model is retrained on this expanded database, $\mathcal{D}_L^k = \cup_{i=1}^k \{(u, y_u) : u \in \omega_i\}$ (comprising all annotations gathered so far), to produce a new model, $\Phi(\cdot; \Theta_k)$, and the process is repeated. We term this framework PixelPick due to its emphasis on selecting appropriate pixels for annotation. The two components of the framework, namely *retraining the segmentation model* and *sampling new pixel coordinates*, are discussed next.

Retraining the segmentation model. At round k of the active learning algorithm, we solve for parameters Θ_k by minimising a cross-entropy loss at each labelled pixel coordinate present in the current annotation database $\mathcal{D}_L^k$:

$$\Theta_k = \underset{\Theta}{\operatorname{argmin}}\, \mathcal{L}(\Theta, \mathcal{D}_L^k) \quad \text{where} \tag{1}$$

$$\mathcal{L}(\Theta, \mathcal{D}_L^k) = -\frac{1}{|\mathcal{D}_L^k|} \sum_{(u, y_u) \in \mathcal{D}_L^k}^{n} \sum_{c=1}^{C} y_u(c) \cdot \log(\hat{y}_u(c)). \tag{2}$$

In the expression above, $y_u(c)$ and and $\hat{y}_u(c)$ denote the c^{th} channel of the oracle-provided label and corresponding model prediction at pixel coordinate u, respectively.

Sampling new pixel coordinates for labelling. The objective of the acquisition function, $\mathcal{A}$, is to sample the B pixel locations at round k that maximise improvement in segmentation performance for the current model $\Phi(\cdot; \Theta_{k-1})$. Functionally, it acts by examining the predictions of $\Phi(\cdot; \Theta_{k-1})$ across all candidate pixel coordinates among the unlabelled pool $\mathcal{D}_u$ and sampling B such coordinates according to a specified criterion.

Discussion. The distinction between sampling contiguous spatial patches for annotation (*e.g.* grids of 128x128 pixels or larger as considered in prior work [38, 21, 9]), and sampling *individual pixel coordinates*, as proposed within the PixelPick framework, is a subtle but important one. It has two key benefits. The first, as noted in Sec. 1, is that it allows us to leverage the powerful inductive biases provided by deep neural network architectures that render them well suited to modelling spatial dependencies in natural images [60]. The second is a practical one: by providing annotators with pixel coordinate proposals, the labelling process is transformed from a "localise and classify" task (required when segmenting semantic regions and typically performed with a mouse or trackpad), into simply a "classify" task in which a class label is assigned to a coordinate proposal, and which can often be performed with a single key-press. We validate both claims through experiments in Sec. 4, where we show that (i) deep neural networks achieve strong segmentation performance at extremely sparse annotation levels, (ii) "mouse-free" annotation can be performed very efficiently.

Acquisition functions. The design of the specific criteria employed by the acquisition function has been the subject of considerable attention in the active learning literature (see [53] and [46] for surveys of classical and recent

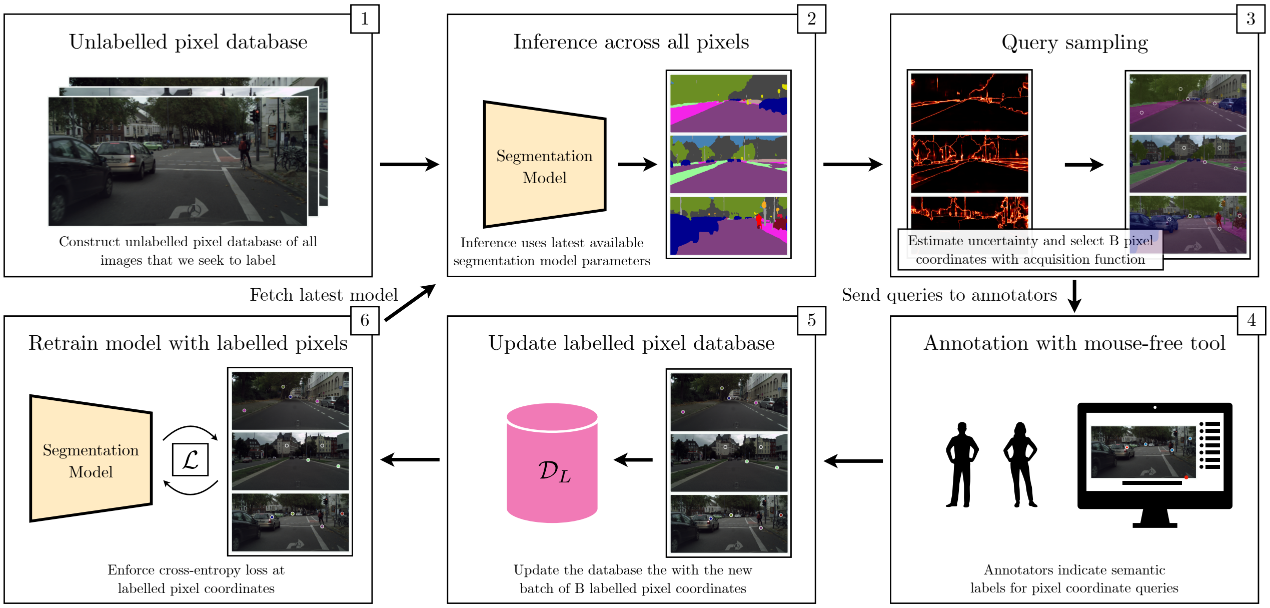


Figure 2: **Overview of the PIXELPICK active learning framework.** Given a database of unlabelled pixels of interest (1) each image is fed to a segmentation model to produce pixel-wise class probabilities (2), which are in turn passed to an acquisition function to estimate per-pixel uncertainties and select a batch of B pixels to be labelled (3). The queries are sent to annotators (4), and the resulting labels are added to the *labelled pixel database*, $\mathcal{D}_L$ (5). Finally, the segmentation model is retrained on the expanded database (6), before the cycle repeats. To bootstrap the process and train the initial segmentation model, we randomly sample B pixels and send them to be annotated. See text in Sec. 3 for further details.

approaches, respectively). Since the focus of our work is not the design of another criterion, but rather on the effectiveness of individual pixels as the base unit for annotation, we consider several existing approaches based on the framework of *uncertainty sampling* [32] that have been noted as effective in the literature, discussed next.

The *Least Confidence* acquisition strategy [33, 14] draws, at each iteration, the pixel coordinate for which the model has *least confidence* in its *most likely* class label. The *Margin Sampling* strategy [50] looks for samples that exhibit the smallest difference (i.e. lowest "margin") between the first and second most probable labels. Finally, the *Entropy Sampling* strategy aims to select the pixel coordinate with the greatest conditional entropy [55] under the current model. Formal mathematical descriptions of each method are given in the supplementary.

As noted in prior work [4, 69], these strategies can suffer from a lack of diversity if applied naively, but can be readily adapted to minimise this effect by first sub-sampling the unlabelled pool and then employing the acquisition function to choose only from this restricted subset. A variation of this diversity heuristic worked well on our task: We first rank all pixels using the acquisition function, then uniformly sampling B/N pixel coordinates from the top $M\%$ ranked locations in each image, where M is a hyperparameter and N denotes the number of images we distribute our budget B amongst. We note that while more sophisticated strategies (e.g. [52]) could also be considered within our framework, a simple *Margin Sampling* strategy coupled with the modification described above proved effective (shown through experiments in Sec. 4), and thus we adopt it in this work.

Sampling batches. The number of pixel coordinates sampled in each round, B, is set as a hyperparameter. A larger value of B corresponds to fewer rounds of annotation (and therefore a potentially faster deployment cycle), at some cost in performance. A detailed study of the effects of B is provided in the supplementary.

3.2. PIXELPICK Annotation tool

To demonstrate the practical utility of the PIXELPICK framework, we created an annotation tool to support the labelling process (Fig. 3). The tool is simple: for each image, the annotator is presented with a few pixels that were selected by the PIXELPICK acquisition function (described in Sec. 3.1). They are also shown a mapping from keyboard keys to semantic labels (Fig. 3, right hand side). The tool iterates over the pixel locations, highlighting the current pixel in red and the annotator simply presses the appropriate key to classify it. The tool then moves on to the next pixel proposal, and the procedure repeats until all proposals in the image are exhausted, when a new image is shown.

We note that an important difference between this annotation technique and those considered in prior work (e.g.

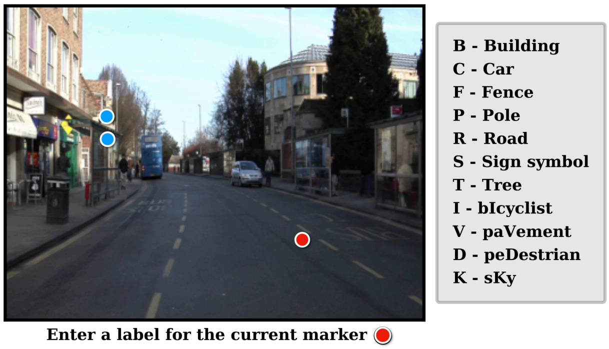


Figure 3: **PIXELPICK mouse-free annotation tool.** The annotator classifies the highlighted point (in red) by pressing the keyboard character of the corresponding class for the dataset. The tool then highlights the next pixel proposal and the process repeats. Note that the task requires the annnotator to perform classification, but not localisation.

scribbles [35], object pointing [3, 45], extreme clicks [44, 39] etc.) is that it is "mouse-free"—requiring only key presses from the user—but avoids the complexity of specialised approaches such as eye tracking [43]. In Sec. 4, we conduct experiments to validate the efficiency of the proposed annotation tool.

4. Experiments

In this section, we first describe datasets and implementation details for our experiments in Sec. 4.1. In Sec. 4.2, we conduct extensive ablation studies, and we then compare with existing state of the art approaches in Sec. 4.3. Finally, in Sec. 4.4, we demonstrate the practical feasibility and robustness of PIXELPICK by reporting annotation times and investigating its sensitivity to annotator errors.

4.1. Experimental setup

Datasets. We conduct experiments on the widely used CAMVID [7], CITYSCAPES [13] and PASCAL VOC 2012 [16] (abbreviated to VOC12) datasets. Detailed descriptions of each dataset are provided in the supplementary.

Training setting. During each round of active learning, we enforce the cross-entropy loss only on the labelled pixels (i.e. those in $\mathcal{D}_L^k$ for round k), as described in Sec. 3.1. Unless otherwise stated, M, the hyperparameter defining the $\%$ of top ranked pixel coordinates used as a basis for uniform sampling is set to 5, while B, the pixel labelling budget per round is set to $10N$ for CAMVID and CITYSCAPES and $5N$ for VOC12, where N is the number of images in the dataset. At the beginning of each round, we reinitialise the model and train from scratch with the updated labelled pixels. For details about optimisation and data augmentation, please refer to the supplementary.

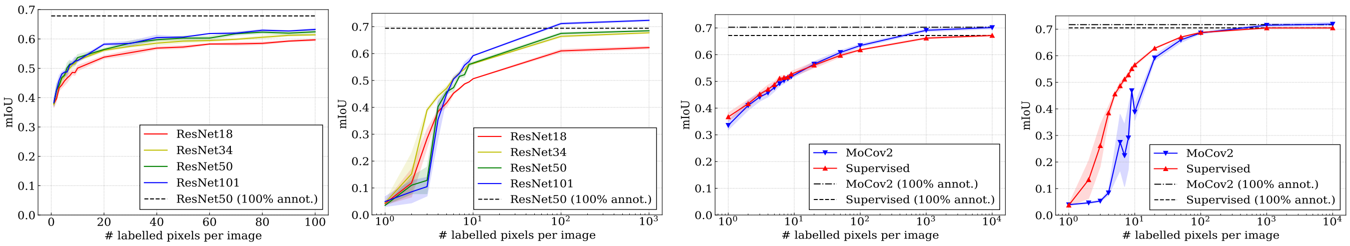
Evaluation metrics. Following standard practice [21, 65, 42, 38], we compute mean intersection over union (mIoU), report our results on the test set for CAMVID, and on the validation set for CITYSCAPES and VOC12 datasets. To provide a measure of variance in our low data regime, we report the average of 3 different runs (i.e., different seeds) on PASCAL VOC 2012 and 5 runs on CAMVID and CITYSCAPES for all experiments. We plot their standard deviations as shaded regions (± 1 std. dev.).

4.2. Ablation studies

In this section, we explore the effect of four factors that affect performance in the PIXELPICK framework, with a particular focus on *annotation diversity* (with the goal of finding the most effective way to spend an annotation budget); *encoder depth* (varying the capacity of the encoder); *encoder initialisation* (self-supervised vs supervised pretraining); and *acquisition function* (determining the best way to select pixels). Note that, while investigating the first three factors, all pixels are selected via simple uniform random sampling, with the goal of validating the effectiveness of inductive bias in modern ConvNets. We simulate the active learning process, following standard practice [21, 65, 9], *i.e.* to label the queried pixels, we simply reveal labels by querying the ground truth annotations at their spatial coordinates.

Annotation diversity. Given a fixed pixel labelling budget (B pixels), a natural question arises: *is it better to label a small number of images densely or a large number of images sparsely*? To address this question we design a simple experiment, where a fixed annotation budget of n pixels is to be distributed over a dataset of N_{total} images. We define the *annotation diversity ratio*, $\eta = \frac{N_{\text{img}}}{N_{\text{total}}}$, where N_{img} refers to the number of images that have had at least one pixel labelled (for simplicity, we assume the labelling budget is evenly distributed over the selected set of images). Therefore, $\eta \to 1$ refers to a budget uniformly distributed over the full dataset (thereby forming a sparse, but diverse, label set), $\eta \to 0$ denotes the case where the budget is only spent on a few images (yielding a densely annotated subset of images). We then train DeepLabv3+ models on CAMVID and CITYSCAPES, fixing B so as to end up with 10 pixel labels per image when $\eta = 1$, and experiment with 5 different diversity ratios η from 0.01 to 1.0. In Fig. 5(a), we observe that mean IoU increases monotonically with η. This indicates that, given a fixed budget, it is better to sparsely annotate as many images as possible, rather than a smaller number more densely, motivating our sparse PIXELPICK approach. In the remaining experiments, we likewise spend our annotation budget evenly across all images within a dataset (as described in Sec. 3.1), with each image being only sparsely labelled.

Encoder depth. We next investigate the effect of encoder capacity in the low annotation regime. Specifically, we



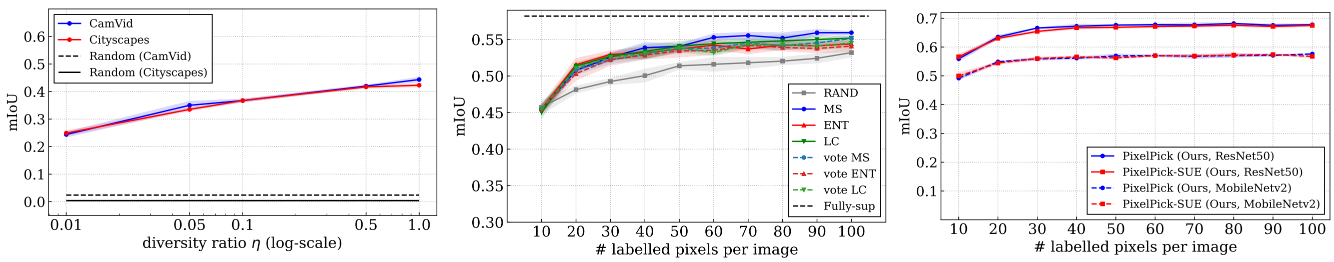
(a) Effect of depth (CAMVID) (b) Effect of depth (VOC12) (c) Effect of pretraining (CAMVID) (d) Effect of pretraining (VOC12)

Figure 4: **Ablation studies.** In (a) and (b), we investigate the effect of segmentation encoder *depth* on CAMVID and VOC12, respectively. We observe that greater depth consistently helps performance above a threshold of 10 pixel labels per image. In (c) and (d), we compare fully-supervised ImageNet classification pretraining with self-supervised ImageNet (MoCov2 [11]) pretraining for the encoder on CAMVID and VOC12, respectively, where we see that for lower numbers of pixel labels per image that fully-supervised pretraining is a better choice, but the situation reverses as more annotations become available.

experiment with a ResNet-based FPN by changing the number of layers in the encoder from 18 to 101 layers. All encoders are initialised with a model pretrained for classification on ImageNet [49]. We conduct experiments both on CAMVID (training each model with 1 to 100 randomly labelled pixel coordinates per image) and VOC12 (training each model with 1 to 1000 randomly sampled labelled pixel coordinates per image), reporting results in Fig. 4(a) and Fig. 4(b), respectively. We observe that deeper networks yield higher performance above a minimum number of labelled pixels (approximately 10) per image. This implies that, at the cost of greater computational complexity, the use of a deeper network may be a viable way to reduce annotation requirements in low annotation regimes (above some minimum labelling threshold).

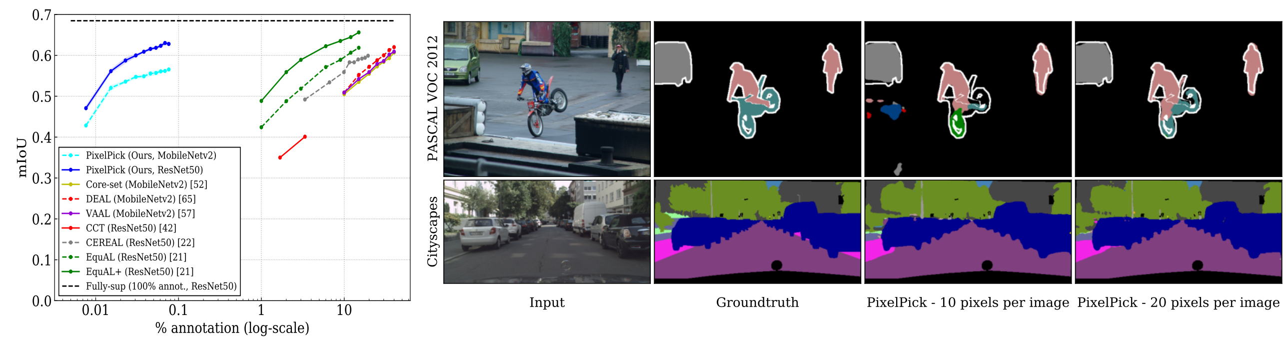
Encoder initialisation. Next, we investigate whether supervised pretraining is necessary for good segmentation performance in a low annotation regime. Concretely, we compare the performance of an FPN-based architecture with a ResNet50 encoder that is initialised using either supervised (ImageNet classification) or self-supervised (MoCov2 [11]) pretraining. To study how performance differs with the number of labelled pixels, we vary the annotation budget from 1 to 10^4 randomly sampled labelled pixels per image on CAMVID (Fig. 4(c)) and VOC12 (Fig. 4(d)). On CAMVID, we observe an interesting biphasic phenomenon: when the number of labelled pixels per image is fewer than 10, the model initialised with supervised ResNet50 shows superior performance. However, as the number of pixel labels increases, self-supervised pretraining gradually outperforms its supervised counterpart. This phenomenon is also observed on the VOC12 dataset, with a cross-over occurring at approximately 10^2 labelled pixels per image. Thus, supervised pretraining may be an appropriate choice for low annotation budgets, when suitable pretraining annotations are readily available, but its advantage wanes the annotation budget grows. Given its superiority at low annotation levels, we adopt supervised pretraining for the remaining experiments.

Acquisition function. Thus far, we have only labelled pixels selected via simple uniform random sampling, showing that modern CNNs—with their strong inductive biases—can be trained for semantic segmentation with just a handful of pixel annotations per image. Here, we go one step further, investigating whether a better choice of acquisition function can further improve learning efficiency. To this end, we experiment on CAMVID with three popular uncertainty sampling methods (described in Sec. 3.1): *Least Confidence* (LC), *Margin Sampling* (MS) and *Entropy Sampling* (ENT). In addition, we also experiment with a Query-By-Committee (QBC) [54] approach that queries labels using model ensembles [53]. We implement this with dropout after each convolutional layer, repeating inference 20 times to obtain a Monte Carlo estimate following [20]. Due to the large number of models to be trained (i.e. different acquisition functions, each trained five times to estimate variance), we employ the lightweight MobileNetv2-based DeepLabv3+ model. We initialise training with 10 uniform randomly selected labelled pixels per image. Once training converges, we query 10 additional pixel labels with the given acquisition function. As described in Sec. 3.1, we first take top M% ranked pixels (here, $M = 5$) per image under the uncertainty estimation ranking and uniformly sample 10 pixels from these pixels. Fig. 5(b) shows the results. We see that all uncertainty-based methods outperform the random baseline in every round. Interestingly, dropout-based voting variants of LC, MS and ENT each show worse performance than their counterparts voting—a similar observation was also made in [9]. We note that in our problem setting, *Margin Sampling* (MS) outperforms other strategies, reaching about 96% of the performance of the fully supervised baseline with only 100 pixels per image (0.06% of the annotations). Therefore, we use MS as our sampling method for PIXELPICK to compare against



(a) Effect of annotation diversity (b) Acquisition functions (CAMVID) (c) Sensitivity to annotator error (VOC12)

Figure 5: **Ablation studies.** In (a), we observe that sparsely annotating a larger number of images (higher η value) outperforms denser labelling of fewer images, with consistent trends on the CAMVID and CITYSCAPES datasets. In (b), we compare acquisition functions on CAMVID and find that *Margin Sampling* performs best. In (c), we investigate the sensitivity of the PIXELPICK framework to annotator errors by simulating a pixel classification user error (SUE) rate of 10%. We observe that performance is only marginally affected, indicating the practical robustness of the PIXELPICK framework.



(a) Comparison to prior work on CITYSCAPES. (b) Qualitative results for models trained with PIXELPICK on VOC12 (top) and CITYSCAPES (bottom).

Figure 6: **Comparison to state-of-the-art and qualitative results**. In (a) we observe that PIXELPICK performs favourably against existing state-of-the-art approaches for active learning and semi-supervised learning on CITYSCAPES. In (b) we show qualitative results. With only 10 labelled pixels per image, segmentation models trained with PIXELPICK achieve promising visual quality, which further improves to capture fine details (e.g. the cleanly segmented thin lamppost in the bottom right image) as further labelled pixels are used.

previous work in the following section.

Discussion. To summarise, we can draw the following conclusions from the ablation studies: First, given a fixed pixel annotation budget, it is best to spread it over as many images as possible; Second, the inductive bias in modern ConvNets makes them well-suited to capture local correlations within an image, evidenced by the first three experiments, where models trained with randomly sampled pixels still perform well; Third, although it might be thought that deeper networks with greater capacity would suffer significantly from over-fitting in the low-annotation regime, we found that for many budget choices, deeper networks are the preferred option. Fourth, in terms of acquisition functions, active learning outperforms random sampling, and in particular, *Margin Sampling* performs best in our setting.

4.3. Comparison to state of the art methods

We next validate our framework by comparing against prior work in active/semi-supervised learning on CAMVID (Fig. 1) and CITYSCAPES (Fig. 6(a)) and in weakly-supervised learning on PASCAL VOC 2012 (Tab. 1). To strike a balance between computation complexity and performance, we adopt the FPN model with a ResNet50 backbone, and query additional samples each round with *Margin Sampling*, as suggested by the ablation study. We train for 10 query rounds, with each round adding 10 labelled pixels per image for CAMVID and CITYSCAPES and 5 pixels for VOC12.

Comparison to region-based active learning and semi-supervised methods. In Fig. 1 and Fig. 6(a), we observe that our approach performs favourably to existing semi-supervised and region-based active learning approaches in terms of label efficiency on both CAMVID

Method	Backbone	Train set (anno. type)	mIoU
Weakly-supervised methods			
GAIN [34]	VGG16	10.5K imgs (classes)	55.3
MDC [63]	VGG16	10.5K imgs (classes)	60.4
DSRG [24]	ResNet101	10.5K imgs (classes)	61.4
FickleNet [31]	ResNet101	10.5K imgs (classes)	64.9
BoxSup [15]	VGG16	10.5K imgs (boxes)	62.0
ScribbleSup [35]	VGG16	10.5K imgs (scribbles)	63.1
Interactive weak supervision			
PIXELPICK (Ours)	ResNet50	1.5K imgs (sparse pixels)	**65.6**

Table 1: **Comparison to existing weakly-supervised methods on VOC12 validation set.** PIXELPICK is competitive against existing methods, using a budget of 20 pixel annotations per image when trained on a much smaller number of images.

and CITYSCAPES, highlighting that the strong inductive bias of CNNs can make good use of pixel-level supervision.

Comparison to previous weakly-supervised learning work. In Table 1, we compare our results on VOC12 to weakly-supervised models which leverage different forms of ground truth annotation. We observe that PIXELPICK compares favourably to existing methods (each trained on 10.5K images) when PIXELPICK uses a budget of 20 pixel annotations per image across only 1.5K training images (since active learning involves retraining the model multiple times, we use a limited size of training set). For reference, note that other methods such as BoxSup [15] and ScribbleSup [35] use multiple annotations per image, depending on the number of objects present.

Comparison to concurrent work of active learning for semantic segmentation In Fig. 7, we compare with the concurrent work of [8], which employs sparse superpixel annotation. To account for the differences between the architectures and training image resolutions used in each method, we compare the methods according to the % of fully supervised performance attained by the model. On the x-axis, we indicate the annotation cost (measured as the number of keyboard inputs for querying pixel labels, *i.e.* the "class clicks" referred to in [8]). We observe that our proposed PIXELPICK approach is competitive with the concurrent work. In particular, we note that PIXELPICK is able to reach 90% performance with fewer annotation cost on both CITYSCAPES and VOC12 benchmarks.

4.4. Practical deployment

Thus far, we have largely followed the common practice in previous active learning segmentation work, mimicking the labelling process by simply disclosing the corresponding labels from the fully-annotated dataset. In this section, we evaluate the efficiency of PIXELPICK (Fig. 3) and its sensitivity to annotator noise during model training.

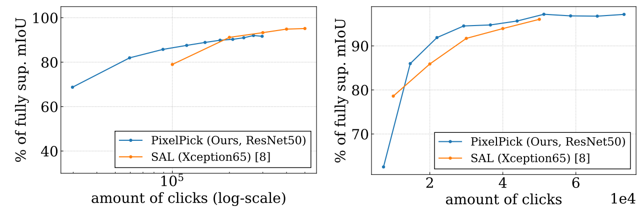


Figure 7: **A comparison based on estimated annotation cost.** We compare PIXELPICK against the concurrent work of [8] in terms of mIoU ratio vs annotation cost approximated by the amount of clicks on (left) CITYSCAPES and (right) VOC12 datasets.

In detail, we ask eleven annotators to label 100 images from VOC12 dataset, with 10 pixels per image, we measure the average time and accuracy (between annotator and the groundtruth from original dataset). As a result, with our simple unoptimised annotation tool, it takes 1.42 seconds on average to label the queried pixel (14s per image), with 87.7% average accuracy. To our knowledge, this annotation speed is significantly faster than drawing bounding boxes or scribbles [15, 35], and approximately twice as fast as picking extreme points according to times reported by [44]. Additionally, given the observed annotation error rate, we conduct an experiment to assess the influence of these noisy annotations, that is, we artificially jitter 10% of the groundtruth annotations to simulate errors during the annotation process and train a model on pixels containing this label noise. As shown in Fig. 5(c), the performance gap incurred from annotation noise is negligible, indicating that our framework is not only efficient with respect to annotation time but also robust to potential errors caused by annotators. This is an encouraging sign for the practical potential of the PIXELPICK framework for real-world deployment.

5. Conclusion

In this work we proposed PIXELPICK, a framework for semantic segmentation that employs a small number of sparsely annotated pixels to train effective segmentation models. We showed that PIXELPICK requires considerably fewer annotations than existing state-of-the-art to achieve comparable performance. Finally, we showed how annotation for pixel-level active learning can be obtained efficiently with a mouse-free labelling tool, facilitating real-world deployment. We hope that our work encourages further research into the promising use of sparse pixel-level annotation for image understanding.

Acknowledgements. GS is supported by AI Factory, Inc. in Korea. WX and SA are supported by Visual AI (EP/T028572/1). The authors would like to thank Tom Gunter for suggestions. SA would also like to thank Z. Novak and S. Carlson for support.

References

[1] Michael David Abramoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology & Visual Science*, 2016. 1

[2] Les E Atlas, David A Cohn, and Richard E Ladner. Training connectionist networks with queries and selective sampling. In *NeurIPS*, 1989. 2

[3] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016. 2, 5

[4] William H. Beluch, Tim Genewein, A. Nürnberger, and J. Köhler. The power of ensembles for active learning in image classification. In *Proc. CVPR*, 2018. 4

[5] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *TPAMI*, 2004. 2

[6] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proc. ICCV*, 2001. 2

[7] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 1, 5

[8] Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *Proc. CVPR*, 2021. 3, 8

[9] Arantxa Casanova, Pedro O Pinheiro, Negar Rostamzadeh, and Christopher J Pal. Reinforced active learning for image segmentation. *ICLR*, 2020. 3, 5, 6

[10] Lluıs Castrejón, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *Proc. CVPR*, 2017. 2

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv: 2003.04297*, 2020. 6

[12] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. *arXiv*, 2021. 2

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016. 1, 5

[14] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, 2005. 4

[15] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proc. ICCV*, 2015. 2, 8

[16] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 5

[17] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R. Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[18] Geoffrey French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Consistency regularization and cutmix for semi-supervised semantic segmentation. In *BMVC*, 2019. 2

[19] Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *ECCV*, 2014. 2

[20] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. ICML*, 2016. 6

[21] S Alireza Golestaneh and Kris M Kitani. Importance of self-consistency in active learning for semantic segmentation. In *BMVC*, 2020. 3, 5

[22] Marc Gorriz, Axel Carlier, Emmanuel Faure, and Xavier Giro-i Nieto. Cost-effective active learning for melanoma segmentation. *arXiv preprint arXiv:1711.09168*, 2017. 3

[23] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, and et al. Jorge Cuadros. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 2016. 1

[24] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proc. CVPR*, 2018. 8

[25] Suyog Dutt Jain and Kristen Grauman. Active image segmentation propagation. In *Proc. CVPR*, 2016. 2

[26] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proc. ICCV*, 2019. 2

[27] Bin Jin, Maria V Ortiz Segovia, and Sabine Susstrunk. Webly supervised semantic segmentation. In *Proc. CVPR*, 2017. 2

[28] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proc. CVPR*, 2017. 2

[29] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Introducing geometry in active learning for image segmentation. In *Proc. ICCV*, 2015. 2

[30] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 2, 2013. 2

[31] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proc. CVPR*, 2019. 8

[32] D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *SIGIR '94*, 1994. 2, 4

[33] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning Proceedings*. 1994. 4

[34] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proc. CVPR*, 2018. 8

[35] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proc. CVPR*, 2016. 2, 5, 8

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[37] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. CVPR*, 2015. 2

[38] Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. Cereals-cost-effective region-based active learning for semantic segmentation. In *BMVC*, 2018. 3, 5

[39] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proc. CVPR*, 2018. 2, 5

[40] Agata Mosinska-Domanska, Raphael Sznitman, Przemyslaw Glowacki, and Pascal Fua. Active learning for delineation of curvilinear structures. In *Proc. CVPR*, 2016. 2

[41] Naveen Shankar Nagaraja, Frank R. Schmidt, and Thomas Brox. Video segmentation with just a few strokes. In *Proc. ICCV*, 2015. 2

[42] Yassine Ouali, Celine Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proc. CVPR*, 2020. 2, 5

[43] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *ECCV*, 2014. 2, 5

[44] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *Proc. ICCV*, 2017. 2, 5, 8

[45] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. Weakly supervised scene parsing with point-based distance metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8843–8850, 2019. 2, 5

[46] Pengzhen Ren, Y. Xiao, Xiao jun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and X. Wang. A survey of deep active learning. *ArXiv*, abs/2009.00236, 2020. 3

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[48] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 2

[49] Olga Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, Michael S. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 6

[50] T. Scheffer, Christian Decomain, and S. Wrobel. Active hidden markov models for information extraction. In *IDA*, 2001. 4

[51] H Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965. 2

[52] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv: Machine Learning*, 2018. 4

[53] Burr Settles. Active learning literature survey. 2009. 3, 6

[54] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT '92*, 1992. 2, 6

[55] C. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423, 1948. 4

[56] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, and et al. Maya Galperin-Aizenberg. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 2019. 1

[57] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[58] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proc. CVPR*, 2019. 2

[59] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. In *Proc. CVPR*, 2019. 1

[60] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proc. CVPR*, pages 9446–9454, 2018. 1, 3

[61] Alexander Vezhnevets, Joachim M Buhmann, and Vittorio Ferrari. Active learning for semantic segmentation with expected change. In *Proc. CVPR*, 2012. 2

[62] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549, Nov 2020. 1

[63] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *Proc. CVPR*, 2018. 2, 8

[64] Michael Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *Isprs Journal of Photogrammetry and Remote Sensing*, 150:59–69, 2019. 1

[65] Shuai Xie, Zunlei Feng, Y. Chen, Songtao Sun, Chao Ma, and Ming-Li Song. Deal: Difficulty-aware active learning for semantic segmentation. In *ACCV*, 2020. 3, 5

[66] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *MICCAI*, 2017. 2

[67] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proc. CVPR*, 2018. 1

[68] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995. 2

[69] Donggeun Yoo and I. Kweon. Learning loss for active learning. *Proc. CVPR*, 2019. 2, 3, 4

[70] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proc. CVPR*, 2016. 2