Semantic Segmentation with Active Semi-Supervised Learning

Aneesh Rangnekar, Christopher Kanan, Matthew Hoffman
Rochester Institute of Technology
Rochester, NY, USA
aneesh.rangnekar@mail.rit.edu

Abstract

Using deep learning, we now have the ability to create exceptionally good semantic segmentation systems; however, collecting the prerequisite pixel-wise annotations for training images remains expensive and time-consuming. Therefore, it would be ideal to minimize the number of human annotations needed when creating a new dataset. Here, we address this problem by proposing a novel algorithm that combines active learning and semi-supervised learning. Active learning is an approach for identifying the best unlabeled samples to annotate. While there has been work on active learning for segmentation, most methods require annotating all pixel objects in each image, rather than only the most informative regions. We argue that this is inefficient. Instead, our active learning approach aims to minimize the number of annotations per-image. Our method is enriched with semi-supervised learning, where we use pseudo labels generated with a teacher-student framework to identify image regions that help disambiguate confused classes. We also integrate mechanisms that enable better performance on imbalanced label distributions, which have not been studied previously for active learning in semantic segmentation. In experiments on the CamVid and CityScapes datasets, our method obtains over 95% of the network's performance on the full-training set using less than 17% of the training data, whereas the previous state of the art required 40% of the training data.

1. Introduction

Given enough labeled data, incredibly good semantic segmentation systems can be trained using deep learning [66, 4, 45, 42, 12, 91, 63, 78]. However, obtaining pixel-wise labels for semantic segmentation is incredibly time-consuming and expensive. For the COCO dataset, this required over 85,000 annotator hours[1][49]. Thus, minimizing

---
[1]There were 10K hours for determining the categories present in each image, 20K for using point annotations for each object present, and over 55K for creating segmentation masks [49].
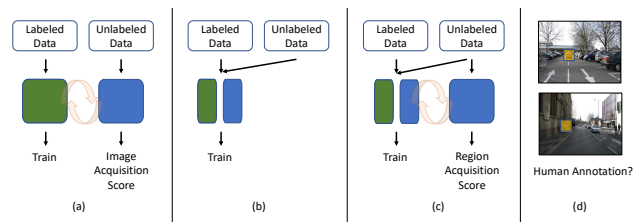

Figure 1: Our active learning approach aims to minimize the number of manual region annotations by using pseudo labels for unlabeled data during training phase. Blending (a) the traditional active learning approach with (b) the traditional semi-supervised approach, (c) our approach used semi-supervised learning on the unlabeled data to generate (d) region acquisition scores that can be queued for human annotations.

the number of annotations is desirable when creating a new semantic segmentation dataset. Active learning (AL) can help achieve this goal. AL is a framework for identifying the most informative samples in an unlabeled pool of samples for annotation. It has been heavily studied in computer vision for classification [27, 6, 71, 44, 85, 52], segmentation [54, 76, 41, 11, 18, 47, 21, 29, 84, 62, 75, 10, 74] and detection [40, 67, 20, 1, 31, 32, 87, 61, 17, 22].

In AL, a machine learning system is iteratively retrained, where for each active learning cycle C it 1) scores each unlabeled sample based on their informativeness, 2) requests annotations for B samples, where B is its annotation "budget," and then 3) re-trains the system with the new samples. This loop continues until the desired performance is achieved. For semantic segmentation, most AL research has aimed to minimize the number of exhaustively annotated images [54, 76, 41, 21, 84, 75]. For segmentation, we argue that this is sub-optimal. Rather than assuming images are exhaustively manually annotated, our approach aims to minimize the total number of manual annotations.

In our framework, only some fully-annotated images are provided manually, as the initial labeled set. The remaining

annotations, for the unlabeled set, are produced automatically by employing semi-supervised learning (SSL) in each active learning cycle. Specifically, we employ pseudo labeling with a teacher-student framework to obtain the missing annotations. While pseudo labeling has been widely used for other problems [79, 77, 60, 25, 50, 16, 2], it has not been used for active learning in semantic segmentation. We also introduce two regularization mechanisms for handling label imbalance.

The key contributions of this paper are the following:

- We pioneer the application of SSL for active learning in semantic segmentation with a teacher-student framework,

- We show that two regularization schemes, Confidence Weighting and Balanced Classmix, improve generalization by mitigating dataset bias,

- Our system achieves performance that rivals or exceeds existing methods on the CamVid [9] and CityScapes [19] datasets.

2. Related Work

Existing methods in active learning can be broadly classified in three categories, stream-learning [92, 69, 82], query-synthesizing [90, 55, 51, 56] and query-computing. Stream-learning methods sequentially receive unlabeled samples and decide on the spot to request a label or discard. At the same time, query-synthesizing approaches often involve the use of generative adversarial networks (GANs) to "synthesize" informative samples from the unlabeled data pool [30]. Query-computing, or pool-based active learning, is the most common approach among the three and involves designing sample acquisition metrics for ranking the most informative samples.

Query-computing methods are further classified into applying Bayesian inference [36, 27, 44, 23, 47, 75, 31], ensemble learning [6, 67, 32, 62, 17], diversity learning [71, 85, 20, 76, 21, 52, 10] and representation learning [54, 40, 1, 41, 11, 18, 28, 29, 84, 87, 74, 61, 22]. Most approaches use least confidence [72], softmax entropy [73], softmax margin [70], mutual information, Core-set [71], Monte-Carlo Dropout [26], or calculating the gradients of the output layer [3]. The premise for all approaches is the hypothesis that only the informative samples will stand out - for example, a low softmax margin indicates high confusion between two classes and hence is a high priority sample within the unlabeled data pool for labeling.

Active Learning for Semantic Segmentation spans all of the above venues and are addressed at image, region and pixel level. The Variational Adversarial Active Learning (VAAL) approach used adversarial learning to identify

samples that confuse a learned variational auto-encoder and discriminator on whether the latent space indicates labeled or unlabeled data [43, 76]. The Minimax Active Learning (MAL) framework also used a discriminator to classify the most diverse samples as compared to the labeled set and paired it with class prototypes to identify the highest entropy samples [21]. The Difficulty-awarE Active Learning (DEAL) architecture appended a probability attention branch to the standard semantic segmentation framework to learn to attend pixels belonging to the same semantic category before calculating metrics for acquisition [84]. Our approach is encouraged from VAAL and MAL, wherein we improve on the idea of using unlabeled data alongside labeled data, but with pseudo labels instead of an adversarial framework (Sec. 4.3).

On a region level, [54] introduced regional Monte-Carlo Dropout with spatial diversity and cost analysis to select regions within images for labeling while RALIS used reinforcement learning to determine the best blocks to sample [11]. ViewAL used diversity in scene object views in multi-view dataset samples [75], [41] refined labels with Conditional Random Fields, and the paper [10] introduced a class-balanced sampling to select super-pixels regions generated by the SEEDS algorithm [81]. Recently, PixelPick reduced the labeling costs by a significant degree by training networks only with sparse pixel annotations [74]. However, most of these approaches fail to utilize the unlabeled data for training the models to their maximum potential.

A notable exception is EquAL, an active learning approach that incorporates self-consistency on the image and its horizontally flipped version [29]. The authors then use the same constraint as the acquisition metric for queuing regions to be labeled within images of the unlabeled pool. It is worth mentioning that similar ideas have been successful in reducing data labeling costs for classification [28] and object detection [22]. We further develop this line of work to use semi-supervised learning for improving active learning frameworks. More specifically, instead of using consistency in terms of equivariance and data augmentations, we propose to work on pseudo labeling to leverage the unlabeled data pool to its maximum potential.

Semi-Supervised Learning for Semantic Segmentation is typically addressed in three ways, explicit consistency regularization [93, 24, 60, 16, 46], using a teacher-student framework [57, 25, 59], and recently, combining teacher-student with contrastive embeddings [50, 89].

Teacher-student training pipelines use the mean-teacher framework that maintains an exponentially moving averaged copy of the student model to provide smoother pseudo labels [79]. This approach is trained by passing weak image copies to the teacher, acquiring pseudo labels and then training the student on strong perturbed copies of the same set of images [8, 77]. The paper [57] used GAN based learning to

encourage confusions between predictions from the labeled and unlabeled examples. CutMix-Seg showed that applying CutMix boosts the performance of learned representations for semantic segmentation [88, 25], while ClassMix modified the CutMix mechanism to sample masks corresponding to individual classes for mixing [59]. Recently, Regional Contrast (ReCo) and C^3-SemiSeg learned contrastive pixel-embeddings to further strengthen the representations alongside conventional cross-entropy as a separate dedicated branch [50, 89]. For simplicity and to set a baseline, we adopt the mean-teacher framework to generate pseudo labels on the unlabeled data pool and expand on this idea for active learning in the next section.

3. Our Approach

Active learning frameworks have a standard operational cycle - the network is trained on labeled data - for semantic segmentation, this is typically with cross-entropy loss, and then used to score samples within the unlabeled data pool (Fig. 1a). The learned network is used to infer statistics in terms of sample information within the unlabeled data pool, represented as scores for data acquisition. Every active learning cycle C, a portion of samples from the unlabeled pool - "budget" B, is sampled, annotated and added to the labeled set. This loop continues till either the budget is exhausted or the updated labeled data pool achieves acceptable performance.

Active learning for semantic segmentation comes with three significant ordeals. Segmentation datasets have their own distribution bias - there is a set of classes that are heavily under-represented "tail distribution", for example, column pole and sign in CamVid [9], and rider and train in CityScapes [19]. They also have significant variance in the within-scene imagery, for example, illumination, contrast, and viewpoint changes. The most common semantic segmentation models use the ResNet or MobileNet family of backbones, with either Fully Convolutional Networks (FCNs), Dilated Residual Networks (DRNs) or DeepLab family of architectures coupled with additional blocks for inference [35, 68, 37, 53, 86, 13, 14, 15]. A common ingredient in all these networks is the Batch Normalization layer, which coupled with the dataset and scenery bias, makes standard training of neural networks difficult to generalize to new unseen data [38]. This brings us to the third problem - in an active learning scenario, the unlabeled images are ranked in order of a network's response, which may be biased based upon how well it has learned on a relatively smaller training set that has its own distribution of classes.

To mitigate these challenges, we leverage unlabeled data using semi-supervised learning during each active learning cycle C as a means of understanding the unlabeled data distribution (Fig. 1b). We use a teacher-student framework for generating pseudo labels on a per-pixel basis to learn better representations of the unlabeled data pool while assisting the network's learning with labeled data. We also replace sampling entire images with regions instead (similar to [54, 11, 29, 18], Fig. 1c, d).

On a higher level, we initialize two sets of the network, wherein the student network is trained with cross-entropy and the teacher network is updated in a gradual manner with the student's parameters (Fig. 2a, Eqn. 2). The teacher network is then used to infer pseudo labels on the unlabeled images, which are used to train the student network on the unlabeled data with cross-entropy. At the end of the SSL training for each active learning cycle, we rank the regions within the unlabeled pool of images based on an acquisition metric computed with the teacher's performance and then select the highest ranking set for labeling. We keep retraining the dataset in an SSL fashion for multiple active learning cycle till a satisfactory performance is achieved to end the label acquisitions.

A slow moving update (teacher network) results in more stable cumulative predictions, however, simply adopting this approach is difficult due to the overall class distribution bias in the labeled data. This is undesirable as pseudo labels from the teacher act as ground-truths for the student network, which may quickly learn the false representations, especially for tail distribution classes. Furthermore, when regions from the unlabeled images are scored as candidates for labeling, a poorly trained network combination may provide a significant shift in the scores for regions that otherwise would be necessary for correcting the bias and assist in guiding the next active learning cycle. To circumvent this problem, we introduce two regularizing schemes - Confidence Weighting and Balanced Classmix.

3.1. Teacher-Student Framework

The mean teacher framework, proposed for semi-supervised learning on image classification, consists of two networks, a student (θ) and a teacher (θ') [79]. The student and teacher share the same network architecture, and the teacher is updated by an exponential moving average (EMA) of the student parameters (Eqn. 1):

$$\theta' := m\theta' + (1 - m)\theta, \tag{1}$$

where m is the smoothness coefficient (momentum) and is set to 0.99 [25, 59]. For the labeled images, the student is learned using a supervised loss (cross-entropy) with the ground truth information (Eqn. 2):

$$\mathcal{L}_{sup} = \ell_{ce}(\theta(x_l), y_l), \tag{2}$$

where x_l are the inputs to the network and y_l are the corresponding ground truth labels. The continuously updated teacher model (Eqn. 1) is used to generate pseudo labels on the unlabeled data which are used to train the student
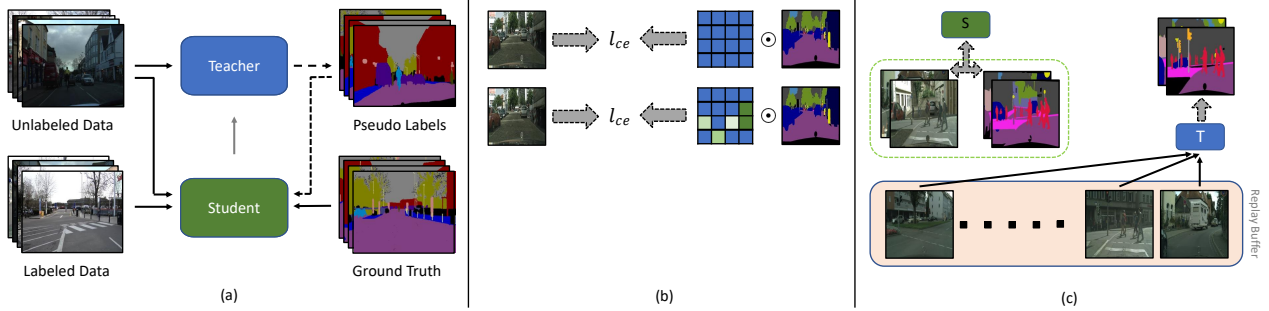
Figure 2: Our SSL setup with (a) teacher-student framework, where (b) shows Confidence Weighting on a per-pixel basis, where the overall cross-entropy loss is weighted by the highest probability for the corresponding pseudo label (blue indicates high, and green indicates low), and (c) shows the Balanced ClassMix, where we maintain a replay buffer of unlabeled images and randomly sample them for ClassMix to further boost the number of samples for tail classes.

network. The teacher receives weakly-augmented version of the images (x_{u-w}) to generate predictions and the student model is trained using the generated pseudo labels as ground truths following (Eqn. 3):

$$\mathcal{L}_{unsup} = \ell_{ce}(\theta(x_{u-s}), [\theta'(x_{u-w})]), \qquad (3)$$

where x_{u-s} is the strong-augmented (perturbed) version as input to the student. The [] indicate the conversion of the logits to one hot vectors, which are used as ground truths for training. We use random flip and random crop operations for x_{u-w}, and random scaling, random flip, color jittering and ClassMix for x_{u-s}.

We treat $[\theta'(x_{u-w})]$ as a one-hot vector indicating the corresponding pseudo label for the pixel under consideration. The final loss for training is then formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \eta \cdot \mathcal{L}_{unsup}, \qquad (4)$$

where η corresponds to the weighting for the unsupervised loss. We calculate η as the ratio of the number of pixels within an image that satisfy $p > 0.97$, (wherein p indicates the max probability for the pseudo label for the pixel) to the total number of pixels within that image, similar to [25, 59, 50].

3.2. Confidence Weighting

Eqn. 4 weighs the contribution of pseudo labeled pixels from the teacher network based on the amount of corresponding probabilities p which pass a preset confidence threshold [25, 59, 50]. However, the weight is applied on an image-level for a mini-batch setting, which implies that all pixels within an image are trained with the same importance irrespective of their prediction confidence. We hypothesize that the student network can quickly overfit to the distribution of classes present in the limited training set and as the

teacher is a slow update (Eqn. 1), this can eventually introduce bias with respect to classes that have limited pixel annotations.

Most semi-supervised learners function on a one-time run agenda. In an active learning setting, however, a class distribution bias can prove more harmful if the networks become very confident of their predictions and they have false confidence on their predictions when computing an acquisition metric. In a long run, this is undesirable as the entire goal of active learning is to continuously query informative samples, minimize labeling costs and yet maintain good performance.

There are multiple approaches in semi-supervised image classification which strive to solve this class imbalance problem, which are broadly based on re-sampling or re-weighting [7, 39, 83]. Most of the approaches involve keeping a running statistic of the number of images falling under a certain class label, which can be daunting when directly applied to semantic segmentation on a pixel basis as the head classes (sky, road, building) heavily overshadow the tail classes (bicycle, pedestrian, fence) . Following the fundamentals of these approaches, we propose a simple modification to Eqn. 3 as follows:

$$\mathcal{L}_{unsup} = \ell_{ce}(\theta(x_{u-s}), p \cdot [\theta'(x_{u-w})]), \qquad (5)$$

where p is the corresponding max probability for every pseudo label from the teacher network (Fig. 2b). This approach ensures that high confidence pseudo labels and annotated labels within an active learning cycles get more importance as compared low confidence pseudo labels.

3.3. Balanced ClassMix

Our second regularization scheme focuses on tail classes imbalance from an oversampling perspective with data augmentation. For this task, we build on ideas from continual learning and ClassMix [64, 5, 33, 34, 59]. Replay buffers

are widely used to mitigate catastrophic forgetting by storing previous sets of samples and mixing them with new samples for training neural networks [58, 80, 33, 34]. These buffers are usually restricted in size due to memory constraints and there are multiple schemes for replay sampling, but, for our task, we rely on uniform sampling from the buffer.

Specifically, we initialize a replay buffer with limit M. At every iteration, we add images into the buffer for sampling later. For the current iteration, we randomly sample images of the same batch size from the replay buffer. These set of images are then passed through the same pipeline for calculating Eqn. 5, except we modify the probability distribution for ClassMix data augmentation. ClassMix uniformly samples the mask from the entire set of classes in its original setting. For our setting, we bias the sampling rate to focus on more samples from the tail classes as compared to the head classes, which is assessed based on the distribution within the labeled data. We accomplish this by sampling classes for ClassMix by two separate distributions, head and tail, instead of a single combined distribution. This assists in the active learning loops as we expect the regions queued by the acquisition metric for labeling belong to the tail classes, due to more challenges posed during teacher-student training with our ClassMix variant (Fig. 2c).

Thus, for the current mini-batch, we get two different x_{u_s}, where the first set corresponds to within mini-batch augmentations (L_{unsup1}), and the second set corresponds to replay augmentations (L_{unsup2}). For computation efficiency, we only maintain the ClassMix-d versions equal to the batch size. It is entirely plausible to use standard ClassMix with the replay images, however, our formulation ensures that more pixels from the tail classes are seen during training, especially post a few cycles of label acquisitions on the unlabeled data pool. Our total training loss from Eqn. 4 becomes,

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \eta_1 \cdot \mathcal{L}_{unsup1} + \eta_2 \cdot \mathcal{L}_{unsup2}, \quad (6)$$

where the values of η in Eqn. 6 are still calculated in the same manner as mentioned before.

3.4. Sampling Strategy

Our entire framework, semi-supervised semantic segmentation for active learning (S4AL) is a two-step process where we follow the standard protocols of active learning and iterate over multiple active learning cycles C, but at the same time, use pseudo labeling to leverage the unlabeled data pool in a much efficient manner. A natural choice for sampling shifts from image-level to region-level as there are multiple sub-regions whose predictions become more confident over cycles, thus lowering the overall image scores

and increasing the possibility of missing out on key annotations that may belong to tail classes. For our acquisition metric, we adopt four strategies: random sampling, least confidence, softmax entropy, and softmax margin. We refer the readers to [75] for an in-depth explanation for all sampling strategies in active learning. From our initial set of experiments, we found softmax entropy to be the best suited acquisition metric for all our datasets (details are enclosed in the supplemental).

4. Experimental Setup and Discussions

4.1. Datasets

We evaluate our proposed approach on CamVid and CityScapes datasets for semantic segmentation [9, 19]. We follow the widely adopted protocol for both datasets: we sample 10% of the data as labeled data from the train set as our labeled data pool [76, 84, 21]. For simplicity and reproducibility, we sample images for the labeled set uniformly from the ground truth train set and consider all others from the train set as the unlabeled image pool. To ensure fairness in a comparison between all methods, we treat the ignore index for both datasets as part of the labeling information so that those areas are also potential regions to be acquisitioned for labeling.

CityScapes is a relatively larger dataset for semantic analysis of urban driving scenes at 1024×2048 resolution with 30 classes. It contains a total of 2975, 500, and 1525 images for training, validation and test respectively. We use the standard split of 2675, 300, and 500 images for training, validation, and test by replacing the validation set as test set and randomly sampling 300 images as validation. We downsample the images to 688×688 resolution and use 19 classes for training and evaluations, similar to [76, 84, 21].

CamVid is a driving scene understanding dataset consisting of images captured at 720×960 resolution with 32 classes. It contains a total of 701 images, with a split of 367, 101 and 233 images for training, validation, and test respectively. We use the widely adopted down sampled scenes at a resolution of 360×480 for our training and evaluation with 11 classes [4, 29, 84].

4.2. Experimental Configuration

We run our evaluations with MobileNetv2 as the backbone on the DeepLabv3+ semantic segmentation architecture [15, 68], a widely adopted standard for semantic segmentation in active learning. We adjust the MobileNetv2 backbone to have a higher stride of 16, similar to [84]. We also benchmark our results on the CityScapes dataset using DRNs [86] to compare against [76, 21], as they follow the same foundation approach in terms of initial labeled sets and image resolutions. We train all our networks with a batch size of four for 100 epochs and 200 epochs on the

Table 1: **IoU:** Class-wise and mean on CityScapes using MobileNetv2 - while all approaches use 40% data to achiever their goal, we achieve our goal with only 16% data.

Method	Road	Side walk	Building	Wall	Fence	Pole	Traffic Light	Traffic Sign	Vegetation	Terrain
Supervised	97.58	80.55	88.43	51.22	47.61	35.19	42.19	56.79	89.41	60.22
Random	96.03	72.36	86.79	43.56	44.22	36.99	35.28	53.87	86.91	54.58
Entropy	96.28	73.31	87.13	43.82	43.87	38.10	37.74	55.39	87.52	53.68
Core-Set [71]	96.12	72.76	87.03	44.86	45.86	35.84	34.81	53.07	87.18	53.49
DEAL [84]	95.89	71.69	87.09	45.61	44.94	38.29	36.51	55.47	87.53	56.90
S4AL	97.73	81.76	88.63	51.42	47.40	36.00	43.91	58.27	89.72	62.01
	Sky	Pedes-trian	Rider	Car	Truck	Bus	Train	Motor-Cycle	Bicycle	mIoU
Supervised	92.69	65.12	37.32	90.67	66.24	71.84	63.84	42.35	61.84	65.30
Random	91.47	62.74	37.51	88.05	56.64	61.00	43.69	30.58	55.67	59.00
Entropy	92.05	63.96	34.44	88.38	59.38	64.64	50.80	36.13	57.10	61.46
Core-Set [71]	91.89	62.48	36.28	87.63	57.25	67.02	56.59	29.34	53.56	60.69
DEAL [84]	91.78	64.25	39.77	88.11	56.87	64.46	50.39	38.92	56.59	61.64
S4AL	92.81	65.62	39.71	90.52	66.07	65.31	46.03	46.88	61.77	64.80

Table 2: **Discussions:** of various studies conducted on CamVid and CityScapes dataset. We use MobileNetv2 for all our experiments, except in Table 2c.

(a) **Region vs Image:** the entire dataset, with top three classes with the largest difference in IoU and (Recall).

	IoU	
	Region	Image
CamVid	61.78	60.31
· Sign	40.27 (73.75)	34.66 (75.22)
· Pedestrian	49.96 (70.30)	43.99 (68.29)
· Bicyclist	51.31 (77.14)	49.76 (72.08)

(b) **Region vs Image:** the entire dataset, with top three classes with the largest difference in IoU and (Recall).

	IoU	
	Region	Image
CityScapes	64.70	64.73
· Truck	66.07 (85.17)	60.12 (74.77)
· Bus	65.31 (89.05)	67.13 (81.91)
· Train	46.03 (51.07)	57.30 (66.67)

(c) **IoU:** using DRN on VAAL [76] and MAL [21]. IoU_s indicates the score at starting with 10% of the data, IoU_f indicates the score at end of active learning, IoU_{sup} are the scores with fully supervised data, and the last row indicates the amount of data utilized.

	VAAL [76]	MAL [21]	**S4AL**
IoU_s	46.2	48.9	57.9
IoU_f	56.5 (+10.3)	58.4 (+9.5)	65.7 (+7.8)
IoU_{sup}	62.95 ± 0.70	61.9 ± 0.70	67.68 ± 1.2
% data	40	30	16

final stage of the active learning cycle. We start with an initial learning rate of 1×10^{-2} and use the "poly" learning schedule to gradually decrease the learning rate [15, 84]. We set the value for the replay buffer size M to 50 and 500 for CamVid and CityScapes, respectively. We sample four regions per image of 30×30 and 43×43 for two and five active learning cycles on CamVid and CityScapes, respectively. We refer to the supplemental for exact training details.

Comparative Algorithms We compare our methods against four other active selection methods as shown in Table 1. These methods are designed for active learning on an image level basis - namely random selection, Entropy [73], Core-Set [71], and DEAL [84]. Attempts to compare with region level algorithms were partially unsuccessful because they use random sampling to initialize the labeled training set, which makes it challenging to replicate results. We refer the readers to the supplemental section for more details.

4.3. Main Results

On the CamVid dataset, we achieve 97% of the performance (Fig. 3a) as compared to utilizing the full dataset with only 13.8% of the labeled pixel data. The previous state-of-the-art approach achieved 94% performance using 40% of the data [84].

Our results on the CityScapes dataset are given in Table 1. Our approach outperforms existing methods on the CityScapes dataset using just 16% of the labeled data as compared to 40%. Except for the class Train, we improve the IoU scores on multiple tail classes (traffic light, traffic sign, truck, motorcycle and bicycle) by a significant amount.
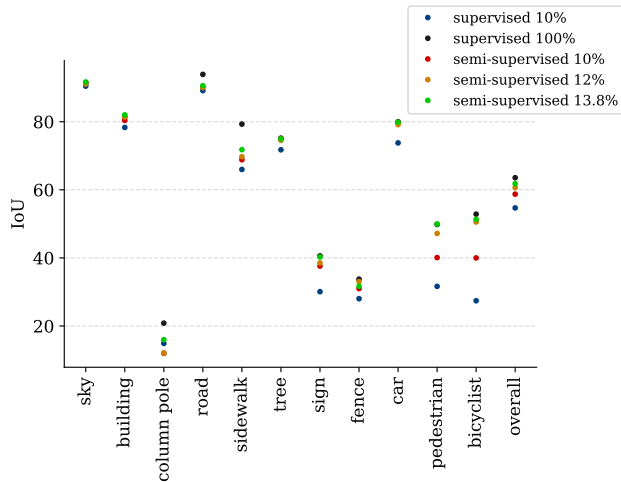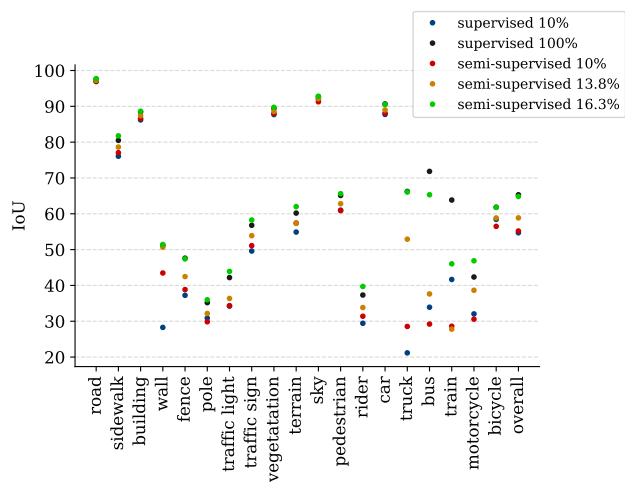
Figure 3: **IoU:** Performance curves of using pseudo labels as a function of the total data used for training on (a) CamVid and (b) CityScapes. We reduce the labeling effort significantly, by requiring only ~4% and 6% additional data on both datasets respectively. On the same initial data, the best state-of-the-art method achieves comparative performance with an additional 30% data.

Figs. 3a and 3b show the gradual increase in IoU per class and overall as a function of data being utilized. Using semi-supervised learning, we improve the scores on both datasets by an initial 1-2 % with the same amount of data. While this is lower than the gains seen in some other semi-supervised algorithms [25, 60, 16, 50], we believe that the limited number of parameters of MobileNetv2 act as a constraint as opposed to the much larger networks used in those studies. Irrespective of the initial boost, our approach iteratively increases the IoU on both datasets, while minimizing the overall labeling effort.

In addition to MobileNet, we also experiment with a comparatively lighter DRN-D-22 [2] network for comparisons with VAAL and MAL approaches for active learning [86, 76, 21] and report our results in Table 2c. Direct comparisons are impossible due to the unknown labeled-unlabeled data split, and hence we report the results on a comparative basis with respect to the IoUs achieved with 10%, maximum data sampled via active learning and fully supervised data pool. We make two key observations: 1) Training with semi-supervised learning on the initial 10% data leads to an up-rise in the starting IoU, and 2) Our approach is able to achieve 97% of the performance while using only 16% of the data, as compared to 40% and 30% on VAAL and MAL respectively.

Finally, we compare our results to other region-based selection methods, namely EquAL and RALIS [29, 11].

Again, direct comparison is not possible due to the unknown labeled-unlabeled split, so instead we compare performance using the same labeled fraction of the data. On CamVid, using 12% of the data and a MobileNetv2 backbone, we achieved 95% of the performance from the fully supervised training regime versus the 94% that EquAL achieved using a relatively heavier ResNet-50 backbone with 12% data and the 96% that RALIS attained also using a ResNet-50 pretrianed on GTA and more (20%) data. Our approach achieved an mIOU of 65.3 on CamVid, as compared to 63.4 from [29], when starting with 8% labeled data and with a budget of 12% labeled data, using a ResNet-50 backbone with DeepLabv3+.

4.4. Additional Results

We discuss the ablations with respect to our approach in this section. In all approach-specific experiments, we sample additional combinations of labeled-unlabeled combinations using random sampling over 5 runs and report our results as the mean and standard deviation over all experiments.

Region vs Image: Instead of using regions, we sample entire images based on the acquisition metric and follow with the semi-supervised learning approach where the newly acquired data is added to the labeled set of images. We repeat the active learning cycle 2 times, sampling 5% of the queried image pool every time, resulting in a 20% data usage. We observe that using regions benefit across both datasets (Tables 2a, 2b), often resulting in higher re-

---

[2]We reached out to the corresponding paper's author for exact DRN version used in their experiments

Table 3: **Discussions:** of additional studies conducted on CamVid and CityScapes dataset. We use MobileNetv2 for all our experiments.

(a) **IoU:** with respect to different block sampling ratios on CamVid and CityScapes datasets.

CamVid	mIoU	CityScapes	mIoU
$30 \times 30 \times 2$	60.4 ± 1.4	$43 \times 43 \times 2$	61.8 ± 0.8
$30 \times 30 \times 4$	61.4 ± 0.6	$43 \times 43 \times 4$	62.6 ± 2.2
$60 \times 60 \times 1$	60.8 ± 2.4	$86 \times 86 \times 1$	61.4 ± 1.6

(b) **IoU:** with respect to different sampling schemes on CamVid and CityScapes datasets.

	mIoU	
	CamVid	CityScapes
Random	59.1 ± 1.8	59.8 ± 2.5
LS	60.5 ± 0.5	60.3 ± 1.4
Ent	61.2 ± 0.5	62.5 ± 1.8
Margin	60.8 ± 0.6	61.8 ± 0.5

call across tail distribution classes except for the 'Train' class for CityScapes, which gets confused with 'Bus'. We believe stronger regularizers can help prevent pseudo label confusion and mitigate this performance gap.

Confidence Weighting: benefits both datasets, most noticeable in the very first active learning cycle (CamVid: $57.2 \longrightarrow 58.5$, CityScapes: $56.15 \longrightarrow 56.6$). This is important as having the correct knowledge and not falling to the dataset bias is crucial towards obtaining the perfect set of samples. More specifically, for CityScapes, we further observe an average difference of nearly 5% for tail classes - traffic light, traffic sign, pedestrian, rider, and motorcycle - which implies a lot of potential false label associations were averted by imposing a simple constraint on the pseudo labels.

Balanced Classmix: also benefits both datasets, and this is most noticable in the final active learning cycle (CamVid: $60.4 \longrightarrow 61.5$, CityScapes: $61.3 \longrightarrow 63.15$). Active learning performs significantly better with using Balanced Classmix - depending on the initial data pool, we observe more inconsistencies in the final results if Balanced Classmix is not used as it compensates for lack of enough contributions from the tail classes. This also comes as a little deterrent, specially for the 'Train' category, as this same regularizer works negatively to over-sample areas that confuse the category.

Region Size: We experiment with different region sizes for both datasets in Table 3a. Statistically, there are multiple

combinations possible, and we achieve the best consistent results across our final size of choosing.

Sampling Strategy: We conclude our study of hyperparameters in the experiment with the metric used for sampling the regions within an image for labeling (Table 3b). The values indicate that softmax entropy works the best for our approach, with softmax margin performing nearly well. This can also be explained by the Confidence Weighting that enforces a tougher constraint and thus makes identifying key areas possible.

5. Conclusions

We visit active learning for semantic segmentation with the goal of incorporating semi-supervision based pseudo labeling into each training cycle to reduce labeling costs. We propose two regularizers to prevent issues with dataset bias in terms of head and tail classes, Confidence Weighting and Balanced Classmix. Our approach achieves comparable performance to full supervised data with a significant reduction in the amount of labels used. A current limitation of our work is lack of understanding with respect to classes that have not been seen during training (open-set or heavily under-represented), however, we believe that our work can set a precedent for further research in the area.

Acknowledgements

References

[1] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López. Active learning for deep detection neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3672–3680, 2019.

[2] Iñigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a classwise memory bank. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[3] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[5] Yogesh Balaji, Mehrdad Farajtabar, Dong Yin, Alex Mott, and Ang Li. The effectiveness of memory replay in large scale continual learning. *arXiv preprint arXiv:2010.02418*, 2020.

[6] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018.

[7] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.

[8] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[9] 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.

[10] Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10988–10997, June 2021.

[11] Arantxa Casanova, Pedro O. Pinheiro, Negar Rostamzadeh, and Christopher J. Pal. Reinforced active learning for image segmentation. In *International Conference on Learning Representations*, 2020.

[12] Aayush K Chaudhary, Rakshit Kothari, Manoj Acharya, Shusil Dangi, Nitinraj Nair, Reynold Bailey, Christopher Kanan, Gabriel Diaz, and Jeff B Pelz. Ritnet: Real-time semantic segmentation of the eye for gaze tracking. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3698–3702. IEEE, 2019.

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:834–848, 2018.

[14] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017.

[15] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[16] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2613–2622, June 2021.

[17] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M. Alvarez. Active learning for deep object detection via probabilistic modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10264–10273, October 2021.

[18] Pascal Colling, Lutz Roese-Koerner, Hanno Gottschalk, and Matthias Rottmann. Metabox+: A new region based active learning method for semantic segmentation using priority maps. *arXiv preprint arXiv:2010.01884*, 2020.

[19] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[20] Sai Vikas Desai, Akshay L Chandra, Wei Guo, Seishi Ninomiya, and Vineeth N Balasubramanian. An adaptive supervision framework for active learning in object detection. *arXiv preprint arXiv:1908.02454*, 2019.

[21] Sayna Ebrahimi, William Gan, Dian Chen, Giscard Biamby, Kamyar Salahi, Michael Laielli, Shizhan Zhu, and Trevor Darrell. Minimax active learning. *arXiv preprint arXiv:2012.10467*, 2020.

[22] Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixe, and Jose M. Alvarez. Towards reducing labeling cost in deep object detection, 2021.

[23] Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep active learning for efficient training of a lidar 3d object detector. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 667–674, 2019.

[24] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Dmt: Dynamic mutual training for semi-supervised learning. *arXiv preprint arXiv:2004.08514*, 2020.

[25] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations, 2020.

[26] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[27] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

[28] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pages 510–526. Springer, 2020.

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

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[31] Ali Harakeh, Michael Smart, and Steven L Waslander. Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 87–93. IEEE, 2020.

[32] Elmar Haussmann, Michele Fenzi, Kashyap Chitta, Jan Ivanecky, Hanson Xu, Donna Roy, Akshita Mittel, Nicolas Koumchatzky, Clement Farabet, and Jose M Alvarez. Scalable active learning for object detection. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1430–1435. IEEE.

[33] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *European Conference on Computer Vision*, pages 466–483. Springer, 2020.

[34] Tyler L Hayes, Giri P Krishnan, Maxim Bazhenov, Hava T Siegelmann, Terrence J Sejnowski, and Christopher Kanan. Replay in deep learning: Current approaches and missing biological elements. *arXiv preprint arXiv:2104.04132*, 2021.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[36] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[37] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[38] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[39] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020.

[40] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *Asian Conference on Computer Vision*, pages 506–522. Springer, 2018.

[41] Tejaswi Kasarla, Gattigorla Nagendar, Guruprasad M Hegde, Vineeth Balasubramanian, and CV Jawahar. Region-based active learning for efficient labeling in semantic segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1109–1117. IEEE, 2019.

[42] Ronald Kemker, Carl Salvaggio, and Christopher Kanan. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS journal of photogrammetry and remote sensing*, 145:60–77, 2018.

[43] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[44] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32:7026–7037, 2019.

[45] Simon AA Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus H Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *arXiv preprint arXiv:1806.05034*, 2018.

[46] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1205–1214, June 2021.

[47] Bo Li and Tommy Sonne Alstrøm. On uncertainty estimation in active learning for image segmentation. *arXiv preprint arXiv:2007.06364*, 2020.

[48] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[49] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[50] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew J Davison. Bootstrapping semantic segmentation with regional contrast. *arXiv preprint arXiv:2104.04465*, 2021.

[51] Yezheng Liu, Zhe Li, Chong Zhou, Yuanchun Jiang, Jianshan Sun, Meng Wang, and Xiangnan He. Generative adversarial active learning for unsupervised outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1517–1528, 2019.

[52] Zhuoming Liu, Hao Ding, Huaping Zhong, Weijia Li, Jifeng Dai, and Conghui He. Influence selection for active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9274–9283, October 2021.

[53] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[54] Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. Cereals-cost-effective region-based active learning for semantic segmentation. *arXiv preprint arXiv:1810.09726*, 2018.

[55] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 580–588. Springer, 2018.

[56] Christoph Mayer and Radu Timofte. Adversarial sampling for active learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3071–3079, 2020.

[57] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[58] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017.

[59] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1369–1378, January 2021.

[60] Yassine Ouali, Celine Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[61] Fengchao Peng, Chao Wang, Jianzhuang Liu, and Zhen Yang. Active learning for lane detection: A knowledge distillation approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15152–15161, October 2021.

[62] Aneesh Rangnekar, Emmett Ientilucci, Christopher Kanan, and Matthew J. Hoffman. Uncertainty estimation for semantic segmentation of hyperspectral imagery. In Frederica Darema, Erik Blasch, Sai Ravela, and Alex Aved, editors, *Dynamic Data Driven Applications Systems*, pages 163–170, Cham, 2020. Springer International Publishing.

[63] Aneesh Rangnekar, Nilay Mokashi, Emmett J Ientilucci, Christopher Kanan, and Matthew J Hoffman. Aerorit: A new scene for hyperspectral image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):8116–8124, 2020.

[64] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[65] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[67] Soumya Roy, Asim Unmesh, and Vinay P Namboodiri. Deep active learning for object detection. In *BMVC*, volume 362, page 91, 2018.

[68] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[69] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016.

[70] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

[71] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[72] Burr Settles. Active learning literature survey. 2009.

[73] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[74] Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: Semantic segmentation with pixelpick. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1687–1697, October 2021.

[75] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9433–9443, 2020.

[76] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.

[77] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[78] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[79] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1195–1204, Red Hook, NY, USA, 2017. Curran Associates Inc.

[80] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

[81] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *European conference on computer vision*, pages 13–26. Springer, 2012.

[82] Sarah Wassermann, Thibaut Cuvelier, and Pedro Casas. Ral-improving stream-based active learning by reinforcement learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in*

Databases (ECML-PKDD) Workshop on Interactive Adaptive Learning (IAL), 2019.

[83] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10857–10866, 2021.

[84] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[85] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93–102, 2019.

[86] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[87] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5330–5339, June 2021.

[88] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[89] Yanning Zhou, Hang Xu, Wei Zhang, Bin Gao, and Pheng-Ann Heng. C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7036–7045, 2021.

[90] Jia-Jie Zhu and José Bento. Generative adversarial active learning, 2017.

[91] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8865, 2019.

[92] Indrė Žliobaitė, Albert Bifet, Bernhard Pfahringer, and Geoffrey Holmes. Active learning with drifting streaming data. *IEEE transactions on neural networks and learning systems*, 25(1):27–39, 2013.

[93] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *International Conference on Learning Representations (ICLR)*, 2021.