Noname manuscript No.
(will be inserted by the editor)

Mask-guided sample selection for Semi-Supervised Instance Segmentation

Miriam Bellver · Amaia Salvador · Jordi Torres · Xavier Giro-i-Nieto

Received: date / Accepted: date

Abstract Image segmentation methods are usually trained with pixel-level annotations, which require significant human effort to collect. Weakly-supervised pipelines are the most common solution to address this constraint because they are trained with lower forms of supervision, such as bounding boxes or scribbles. Semi-supervised methods are another option, that leverage a large amount of unlabeled data and a limited number of strongly-labeled samples. In this second setup, samples to be strongly-annotated can be selected randomly or with an active learning mechanism that chooses the ones that will maximize the model performance. In this work, we propose a sample selection approach to decide which samples to annotate for semi-supervised instance segmentation. Our method consists in first predicting pseudo-masks for the unlabeled pool of samples, together with a score predicting the quality of each mask. This score is an estimate of the Intersection Over Union (IoU) of the segment with the ground truth mask. We study which samples should be annotated based on the quality score, leading to an improved performance for semi-supervised instance segmentation with low annotation budgets.

Keywords Image Segmentation · Semi-Supervised Learning · Active Learning

Miriam Bellver · Jordi Torres
Barcelona Supercomputing Center (BSC)
Jordi Girona street, 29, 31, 08034 Barcelona, Spain
E-mail: miriam.bellver@bsc.es, jordi.torres@bsc.es

Amaia Salvador · Xavier Giro-i-Nieto
Universitat Politècnica de Catalunya (UPC)
Jordi Girona street, 1-3, 08034 Barcelona, Spain
E-mail: amaia.salvador@upc.edu, xavier.giro@upc.edu

1 Introduction

Instance segmentation is a popular task in computer vision in which a mask and a class category are predicted for each target object in a given image. Typically, high-performing models rely on large datasets of annotated data, which are expensive to obtain. This work extends our previous study that presented a semi-supervised scheme, which we refer as BASIS [3] (from **B**udget **A**ware **S**emi-supervised semantic and **I**nstance **S**egmentation). Given a low annotation budget, BASIS outperforms previous works on weakly or semi-supervised semantic and instance segmentation.

The BASIS approach for semi-supervised instance annotation is depicted in Figure 1. The pipeline consists in firstly training an *annotation* network that uses only a few strongly-annotated samples. This *annotation* network is subsequently used to pseudo-annotate unlabeled or weakly-labeled samples. Later, a second *segmentation* network is trained with both the few strongly-annotated samples and the pseudo-annotations. In our previous solution [3], the subset of strongly-annotated samples was chosen randomly. In this work, we propose an alternative selection scheme based on active learning, which leads to an improved performance given a fixed annotation budget.

Our active learning method for sample selection consists in firstly training the *annotation* network with a random subset of very few strongly-annotated images. This model is later used to obtain pseudo-annotation masks, as in BASIS, but in this case a confidence score for the masks is predicted, too. This additional information is leveraged to select more images to be strongly-annotated by humans, allowing a more efficient usage of the available annotation budget.

Our main contribution is the definition of a novel way to estimate the confidence score. Specifically, our model is trained to predict an estimation of the Intersection Over Union (IoU) of the pseudo-labels and their corresponding ground truth masks. IoU prediction has been used in previous works on object detection for filtering object proposals [23]. To the best of the authors' knowledge, our work is the first one to use as a selection criterion for active learning. We name our selection strategy *Mask-guided sample selection*.

The summary of our contributions is as follows: 1) a novel method to estimate the mask confidence score based on IoU score, being the first work to leverage IoU prediction for active learning, 2) a study of the selected images, which concludes that the best images to annotate are those that are neither the easiest nor the most complicated of our dataset. Finally, 3) with the Mask-guided sample selection strategy we reach higher performance compared to our BASIS baseline, leading to state-of-the-art results at low annotation budgets.

The remainder of this manuscript is structured as follows: Section 2 covers the previous works on weakly and semi-supervised segmentation, active learning pipelines and IoU prediction. Next, the benchmark of budget-aware segmentation is explained in Section 3. Following, the semi-supervised architecture that we extend is reviewed in Section 4. In Section 5, the IoU quality prediction pipeline is described. In Section 6 the experimental validation is presented. Finally, Section 7 draws the conclusions of this work.

2 Related Work

Weakly-Supervised Semantic Segmentation. Several works in the literature have proposed to use weak supervision to reduce the annotation cost. For semantic segmentation, one of the most popular forms are image-level labels, as they can be obtained with minimum human intervention. There are approaches that treat image-level labels with multiple instance learning techniques [38][37] [36], but these works achieve an accuracy far from their fully-supervised counterparts. Other works develop Expectation-Maximization methods to learn from weakly-annotated data [35]. More recently, a pool of works have focused on localizing class-specific cues with Class Activation Maps (CAMs) [55] in order to mine regions [49][22][1][50], while others obtain regions with attention mechanisms [53]. An alternative is to leverage noisy labels crawled from the internet [19], and obtaining pseudo-annotations using low-level cues such as saliency and edges.

Weakly-Supervised Instance Segmentation. The task that we address in this work is instance segmentation with low annotation budgets. Few works have addressed weakly-supervised instance segmentation in computer vision. Bounding box labels have been exploited by [24][54][26] to recursively generate and refine pseudo-labels from a weak-labeled set. These methods typically rely on bottom-up segment proposals [39][42]. In contrast with this approach, [41] proposes an adversarial scheme that learns to segment without using any object proposal technique. Although these works tackle weakly-supervised instance segmentation, their weak supervision consists in using bounding boxes, thus their main challenge resides in how to separate the foreground from the background within a bounding box. The first work that uses image-level supervision for weakly-supervised instance segmentation [56] detects peaks of Class Activation Maps (CAMs) [55], producing what they identify as Peak Response Maps (PRMs). With them they generate a query to retrieve the best candidate among a set of pre-computed object proposals (MCG) [39]. Recently, [25] builds on PRMs by using the pseudo-masks to train Mask R-CNN [17] in a fully-supervised way, reaching better performance.

Semi-Supervised Segmentation. Semi-supervised learning has been widely explored for image classification with techniques such as adversarial training [27,5], but to the authors knowledge, only [20][26] have tackled semi-supervised instance segmentation in still images. However, they assume a huge amount of weakly-labeled samples (using MSCOCO [28]). In details, [20] leverage bounding boxes as weak labels, and propose a weight transfer function to predict the mask segmentation network weights from the bounding box prediction network weights. On the other hand, [26] obtain pseudo-annotations from bounding boxes relying on Grabcut [42] and MCG [39], and use these annotations to train their models, which resembles to our pipeline.

In our previous work [3], we focused on low-budget scenarios presenting the first results for semi-supervised instance segmentation for the Pascal VOC benchmark [7] with no extra images from other datasets, obtaining results for very low annotation budgets. In this present work, we work with the same pipeline but extend it to a better selection of samples to be strongly-annotated for the semi-supervised scheme. We want to highlight that in contrast to previous works that have addressed semi-supervised instance segmentation, we exploit image-level labels instead of bounding boxes, which is more challenging because no localization information is provided. On the other hand, our annotations are easier and cheaper to obtain.

Active learning [45] consists in recursively selecting which samples to annotate in order to train a network. The goal of this approach is the reduction of the annotation cost, by only annotating those samples that will have more impact to the learning of the model. This acquires special relevance in contexts where annotating samples is very expensive, e.g., in image segmentation problems. Common active learning methods select samples according to two main criteria: how *uncertain* and *representative* a sample is. The *uncertainty* is related to how informative a sample is with respect to the learning process.

There are several methods that estimate the *uncertainty*, e.g., dropout has been used to sample from the approximate posterior of a Bayesian CNN to calculate the uncertainty of predictions when varying the model [11]. This quantified metric can be used to request the annotation of subsequent training batches of data [12][15]. More recent methods have also used Bayesian CNNs to calculate the informativeness of images generated by a Generative Adversarial Network (GAN) [31] in order to add these samples to the training set. Another method [6] is based on bootstrapping, and consists in training several networks with different subsets and calculate the variance in predictions across the different networks in order to estimate uncertainty [52].

Some of the aforementioned methods not only base their selection on the *uncertainty* criterion, but also on the *representativeness* of a sample. This criterion is relevant to promote diversity among samples and to avoid redundancy. One strategy used in computer vision is to extract image descriptors with a CNN, and compare images with a cosine similarity metric [52] to avoid picking very similar samples. Maximizing set coverage has also been studied [10]. Other metrics, such as *content distance* have been used to quantify the distance between images to maximize content information [32][33].

Most of the above methods focus on image recognition and region labeling. The first works that handled active learning for large scale object detection [47] used as active learning criterion the *simple margin* selection method for SVMs [46], which seeks points that most reduce the version space. More recently, methods rely on modern object detectors [40][29], but still are based on uncertainty indicators like least confidence or 1-vs-2 [4][43]. Notice that object detection is very close to the instance segmentation task addressed in this work. However, our sample selection criterion is based on the estimated quality of the different masks predicted for each image, instead of using classification scores as the previous approaches. We want to highlight that our method is the first one that proposes active learning for semi-supervised instance segmentation for Pascal VOC benchmark [7], and the first one to explore mask quality prediction as an alternative to classification scores for active learning. Our claim is that classification scores are suitable for object detection pipelines, but do not reflect the quality of the actual pixel-wise annotation used to train instance segmentation models.

IoU prediction IoU prediction has been used in recent works for filtering object proposals in object detection tasks [23]. Precisely, in [23] the IoU between predicted bounding boxes and ground truth bounding boxes is estimated, and the authors argue that this score, in comparison to a class confidence score, considers the localization accuracy. In their work they show how their approach leads to improved performance. Similarly to this work, [21] estimate the IoU between the predicted masks and the ground truth masks, and use this score to better filter object proposals for instance segmentation. In this direction, we propose to also predict the Intersection Over Union of the predicted masks with respect to the ground truth as a measure of the confidence of the prediction.

3 Benchmark for budget-aware segmentation

As in our previous work [3], we propose a unified analysis across different supervision setups and supervision signals for instance segmentation. Our motivation raises from the ultimate goal of weakly and semi-supervised techniques: the reduction of the annotation burden. We adopt the analysis framework from [2] and extend it to any supervision setup to compare to other works considering the total annotation cost.

We estimate the annotation cost of an image from a well-known dataset for semantic and instance segmentation: the Pascal VOC dataset [7]. Our study considers four level of supervision: image-level, image-level labels + object counts, bounding boxes, and full supervision (i.e. pixel-wise masks). The estimated costs are inferred from three statistical figures about the Pascal VOC dataset drawn from [2]: a) on average 1.5 class categories are present in each image, b) on average there are 2.8 objects per image, and c) there is a total of 20 class categories. Hence, the budgets needed for each level of supervision are:

Image-Level (IL): According to [2], the time to verify the presence of a class in an image is of 1 second. The annotation cost per image is determined by the total number of possible class categories (20 in Pascal VOC). Then, the cost is of $t_{IL} = 20\,\text{classes/image} \times 1\text{s/class} = 20\,\text{s/image}$.

Image-Level + Counts (IL+C): IL annotations can be enriched by the amount of instances of each object class. This scheme was proposed in for weakly-supervised object localization [13], in which they estimate that the counting increases the annotation time to 1.48s per class.

Hence, the time to annotate an image with image labels and counts is $t_{IL+C} = t_{IL} + 1.5\,\text{classes/image} \times 1.48\,\text{s/class} = 22.22\,\text{s/image}$.

Table 1 Average annotation cost per image when using different types of supervision.

	IL	**IL+C**	**Full**	**BB**
Cost (s/image)	20	22.22	239.7	38.1

Full supervision (Full): We consider the annotation time reported in [2] for instance segmentation: $t_{Full} = 18.5\,\text{classes/image} \times 1\text{s/class} + 2.8\,\text{mask/image} \times 79\,\text{s/mask} = 239.7\,\text{s/image}$.
Bounding Boxes (BB): Recent techniques have cut the cost of annotating a bounding box to 7.0 s/box by clicking the most extreme points of the objects [34]. Following the same reasoning as for dense predictions, the cost of annotating a Pascal VOC image with bounding boxes is $t_{bb} = 18.5\,\text{classes/image}\times 1\text{s/class} + 2.8\,\text{bb/image} \times 7\,\text{s/bb} = 38.1\,\text{s/image}$.

Table 1 summarizes the average cost of the different supervision signals for a single Pascal VOC image. In this work these annotation costs will be used as reference to compare between different configurations or to other works.

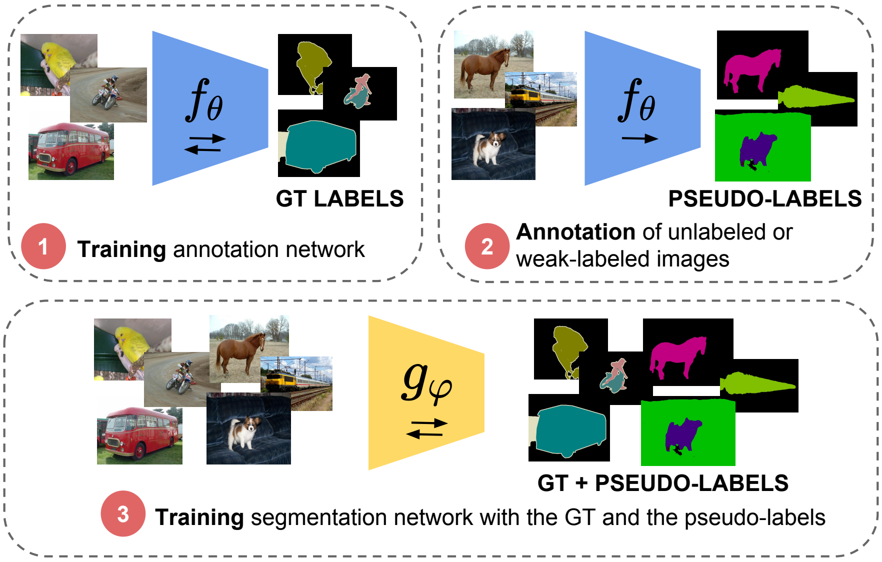
4 BASIS

Our sample selection approach is implemented over the semi-supervised scheme we introduced in [3]. **B**udget-**A**ware **S**emi-Supervised Semantic and **I**nstance **S**egmentation (BASIS) pipeline consists of two different networks. A first fully supervised model f_θ is trained with strong-labeled samples from the ground truth $(X, Y) = \{(x_1, y_1), ..., (x_N, y_N)\}$, being N the total number of strong samples. The network f_θ is an annotation network used to predict pseudo-labels $Y' = \{y'_1, ..., y'_M\}$ for M unlabeled/weakly-labeled samples $X' = \{x'_1, ..., x'_M\}$. A second segmentation network g_φ is trained with $(X, Y) \cup (X', Y')$, as depicted in Figure 1.

Our setup consists in working with heterogeneous annotations: strongly-annotated samples (with pixel-level annotations) and weakly-annotated samples with image-level plus counts (IL+C). This type of weak annotation consists in indicating the class labels in each image, and the counts of how many times each category appears.

The architecture for the *segmentation* network g_φ that we use is the recurrent architecture for instance segmentation RSIS [44]. We use the open-source code released by the authors. RSIS [44] consists in an encoder-decoder architecture. The encoder is a ResNet-101 [18], and the decoder is composed of a set of stacked ConvLSTM's [51]. At each time step, a binary mask and a class category for each object of the image is predicted by the decoder. The architecture also includes a stop branch that indicates if all objects have been covered. The main feature of this architecture is that its output does not need any post-processing as in object proposal-based methods, where proposals need to be filtered a posteriori. This way, the pseudo-annotations are directly the output of the network itself.

Fig. 1 BASIS pipeline from [3] consists of two networks, an annotation network trained with strong supervision, and a segmentation network trained with the union of pseudo-labels and strong-labeled samples (GT).

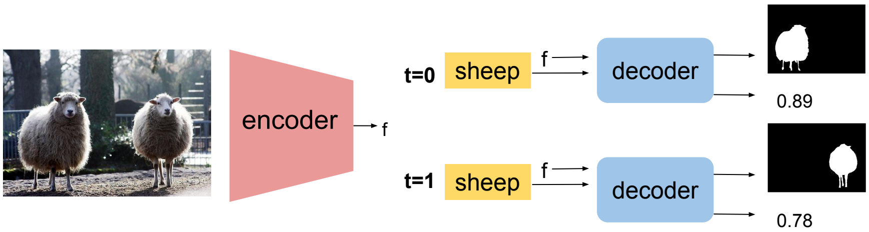

f_θ
GT LABELS
1 **Training** annotation network
f_θ
PSEUDO-LABELS
2 **Annotation** of unlabeled or weak-labeled images
g_φ
GT + PSEUDO-LABELS
3 **Training** segmentation network with the GT and the pseudo-labels

Regarding the *annotation* network f_θ, we modify the RSIS architecture to adapt it to the weakly-supervised setup. The main difference is that, besides the features extracted by the encoder, the decoder receives at each time step a one-hot encoding of a class category representing each of the annotated instances of the objects in the image. If there are several instances belonging to the same class, a one-hot encoding of that class will be given as input at several time steps, as many as the counts of instances of each depicted class. As we did in our previous work [3], we call this architecture *W-RSIS*, where W- refers to the weakly supervised approach.

5 IoU quality prediction

The main contribution of this work is proposing an additional output to the *W-RSIS annotation* network that predicts the quality of each predicted mask. This confidence score can guide an active learning algorithm in choosing which images should be strongly-annotated given a limited budget. We propose to predict the Intersection over Union (IoU) of the predicted masks over a hypothetical ground truth as the guiding signal. As ground truth masks are available for the training data, the model can be trained and the confidence score estimated. The pipeline can be seen in Figure 2. We call this new architecture *IoU-W-RSIS*. The IoU measures the intersection between two regions divided by its union, and it is a common metric to assess segmentation performance (Equation 1).

Fig. 2 *IoU-W-RSIS* model with the IoU branch. The model consists of an encoder for the image, that is a ResNet-101, and a recurrent decoder that receives at each time step a class category label, in this example, it receives at the first time step the label *sheep*, and in the second time step it receives the same label, as in the image there are two instances of the *sheep* category. The decoder also receives the features obtained by the encoder, and at each time step produces a binary mask with the segmented instance, and a prediction of the IoU of the produced mask.



$$IoU(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

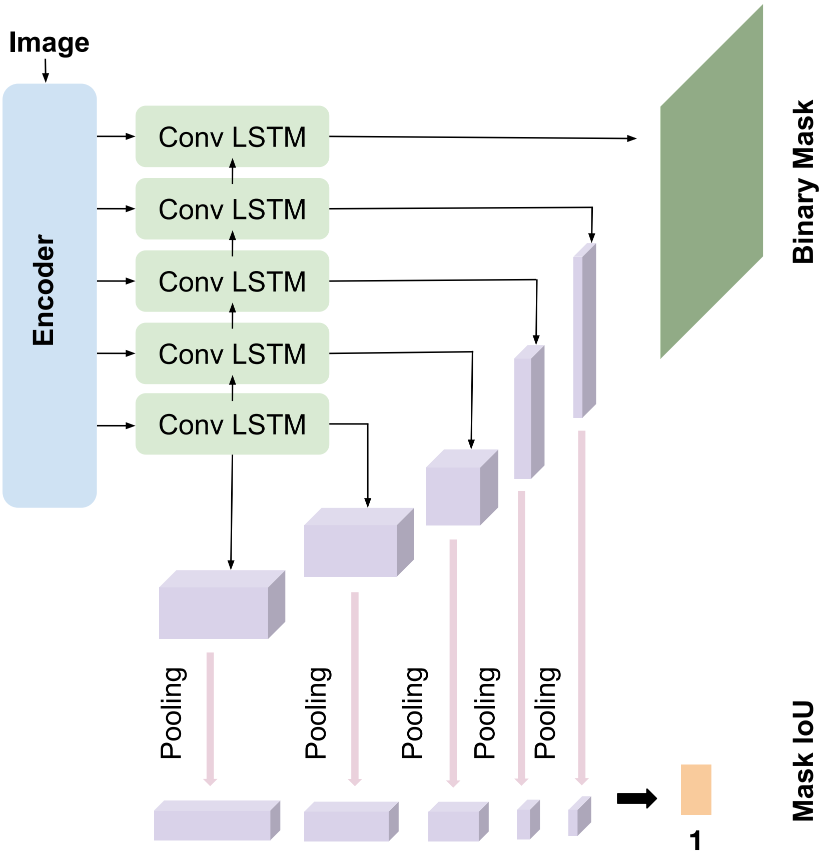
IoU-W-RSIS will segment an object mask of the category fed in the input and predict a confidence score of the segmentation quality at each time step.

The architecture that predicts the IoU is depicted in Figure 3. A branch for IoU prediction is added to the decoder of the network, indicated in the Figure as *Mask IoU*. This branch aggregates features of the decoder at different spatial resolutions, concatenates them, and computes global average pooling. Afterwards, we add a fully connected layer that predicts the IoU using an L1 regression loss. This loss term is introduced once the segmentation loss has already converged. At that point, the network weights are frozen and only the additional IoU branch is trained for a few epochs. To give more relevance to the predictions of low IoUs, we predict the squared IoU, as suggested in other scenarios in which small values have important relevance, as bounding box offset regression for object detection [40].

With the proposed architecture, an *IoU Score* for each mask is predicted. In our methodology we use an overall IoU per image instead of individual IoU scores per object. This means that a human annotator will be asked to annotate all object instances from the selected images. Therefore, to compute the *IoU Score* for an image with M objects, we simply average the scores predicted per each object, as seen in Equation 2.

$$IoU\ Score = \frac{1}{M} \sum_{i \epsilon M} IoU_i \tag{2}$$

Fig. 3 *IoU-W-RSIS* model with the IoU branch for a single time step. The class label is omitted in this figure for clarity. The input image is fed into the encoder and the features at different resolutions are fed at different levels of the decoder. Each level of the decoder has a convolutional LSTM (Conv LSTM) layer that receives a hidden state from the previous time step, and produces features for the current time step. The features of the different levels are pooled and concatenated, and are the input of a fully connected layer that predicts the mask IoU. On the other hand, the features of higher resolution are the ones that produce the binary mask that corresponds to the segment.



6 Experiments

The *IoU-W-RSIS annotation* network presented in Section 5 is tested considering one active learning iteration for the task of instance segmentation. Our experiments aimed at measuring the gain of a IoU-guided selection of the images to strongly-annotate, compared with a baseline of random selection as in [3], and with baseline techniques for active learning based on Dropout [11]. We

Table 2 Mean Absolute Error (MAE) of IoU prediction. Each column indicates the number of samples used to train the IoU prediction branch, and each row is a different configuration that we test. The one that yields best performance is when we freeze the segmentation network and when the prediction of the model is the squared root of the IoU.

	100	**200**	**400**	**800**	**1464**
Baseline	31.1	39.8	49.3	47.7	51.0
+ Freeze Seg. Network	24.8	**16.7**	19.0	17.1	**16.6**
+ Sqrt Loss	**23.6**	19.5	**18.0**	**17.0**	**16.6**

present experiments for the instance segmentation task for the Pascal VOC 2012 benchmark [7].

The standard semi-supervised setup adopted for this benchmark consists in using the Pascal VOC 2012 train images (1464 images) as strong-labeled images, and an additional set (9118 images) from [16] as unlabeled/weak-labeled. In our work, we select which samples to strongly-annotate from the Pascal VOC 2012 train images. The additional set of Pascal is used to obtain pseudo-annotations for the semi-supervised pipeline.

This section is divided in two subsections, first we focus on the IoU prediction task (Section 6.1), and then we study how to use this score for tackling sample selection (Section 6.2).

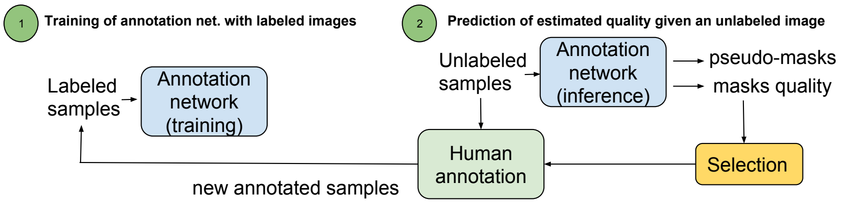
6.1 IoU Prediction

In this first set of experiments we try several configurations to train the IoU branch of the *IoU-W-RSIS* architecture. We train our proposed annotation network *IoU-W-RSIS* with $N \in \{100, 200, 400, 800, 1464\}$, where N is the amount of strongly-annotated samples. These N samples are randomly selected from the Pascal VOC 2012 train set (that has a total of 1464 images). Table 2 contains the Mean Absolute Error (MAE) computed as the mean of the MAE of *IoU Scores* (Eq. 2) of the dataset for the different configurations. The *Baseline* configuration consists in training the IoU branch at the same time as the segmentation branch. In the next row, we freeze the weights of the segmentation network after 150 epochs and only train the IoU branch (for 250 epochs). Finally, we optimize the squared root of the IoU, as small values are specially relevant for this task, and this option leads to the best results. As expected, the MAE tends to decrease from left to right in the table, which corresponds to considering more strongly annotated images.

6.2 Mask-guided sample selection

The second set of experiments exploit the estimated IoU to select which images should be manually annotated and used as supervision to train the *annotation* network in the BASIS pipeline.

Fig. 4 Active Learning pipeline to select next samples to be labeled by a human annotator. The first step consists in training the annotation network with few strongly-labeled samples. The second step consists in using the annotation network to obtain pseudo-masks for the unlabeled samples, together with the masks quality score. Based on this score, some samples are selected to be manually annotated by a human, and added to the pool of labeled samples to re-train the annotation network.


1
Training of annotation net. with labeled images
2
Prediction of estimated quality given an unlabeled image
Labeled samples
Annotation network (training)
Unlabeled samples
Annotation network (inference)
pseudo-masks
masks quality
Human annotation
Selection
new annotated samples

Considering a fixed set of 1464 images from Pascal VOC 2012, our proposal firstly trains an *IoU-W-RSIS annotation* network with a few randomly selected samples (100), and later pseudo-annotates the remaining samples (1364) with it. Together with the pseudo-annotations, the *IoU-W-RSIS annotation* network predicts the IoU for each pseudo-label. With the estimated IoU for each predicted mask we explore different approaches to select which subset of samples should be manually annotated. Finally, the manually annotated samples are added to the training of the *annotation* network. This pipeline is depicted in Figure 4. We follow the classic active learning setup, in which the samples to be annotated are iteratively selected. In our case, we experiment with a single iteration, but it could be easily extended to a looped pipeline.

6.2.1 Criterion for sample selection based on IoU:

The experiment in this section explores a criterion for selecting which images should be strongly annotated by a human given their estimated *IoU Scores*. As we want our analysis to focus on the selection criterion only, in this section we will not use the IoU value predicted by our model but the real ground truth value (*oracle*).

Our experiments start with an *IoU-W-RSIS annotation* network trained with only 100 samples, which obtains a performance of 19.0 Average Precision (threshold=0.5). After that, we select another *N'* samples to be manually annotated, being $N' \in \{100, 300, 700\}$ to make a total of $N \in \{200, 400, 800\}$ strongly-annotated samples. The criterion used to select these *N'* samples consists in first defining a set of *IoU Scores* (from 0 to 1.0 in steps of 0.1), that we name β, and select the *N'* images (being $N' \in \{100, 300, 700\}$) whose *IoU Scores* are closest to these β values. Finally, the samples used to train the *annotation* networks are the 100 initial random images plus these *N'* selected images. The performance obtained with these different subsets is presented in Table 3, which reports the Average Precision (threshold=0.5). All configurations have been trained 5 times, and the reported results are the average with

Table 3 Oracle: mean Average Precision (th=0.5) for different selection criteria (5 runs for each configuration). Each column indicates the number of images used to train the segmentation models. The first row shows the results obtained with random selection of samples, the second and third rows show a baseline sample selection method, whereas the following rows show different selection criteria with our method. If $\beta = 0.0$ and the number of training samples is 200, means that the first 100 samples are randomly-selected and the next 100 are the ones that have a IoU closer to 0.0.

	200	**400**	**800**
Random subset	22.7 ± 1.8	27.1 ± 0.8	34.5 ± 2.0
Dropout Baseline (highest)	21.4 ± 1.4	23.9 ± 1.3	28.1 ± 1.9
Dropout Baseline (lowest)	20.0 ± 0.9	24.8 ± 1.5	32.2 ± 1.4
$\beta = 0.0$	20.9 ± 1.5	24.1 ± 0.7	29.1 ± 1.3
$\beta = 0.1$	22.3 ± 1.5	23.8 ± 0.6	28.6 ± 0.7
$\beta = 0.2$	23.3 ± 0.8	24.4 ± 0.3	31.6 ± 1.1
$\beta = 0.3$	**23.9 ± 0.8**	26.5 ± 2.6	32.9 ± 1.4
$\beta = 0.4$	23.4 ± 2.7	**29.0 ± 1.3**	35.0 ± 0.6
$\beta = 0.5$	22.2 ± 1.1	28.9 ± 0.7	35.1 ± 0.9
$\beta = 0.6$	22.2 ± 2.4	28.6 ± 1.3	**35.4 ± 2.4**
$\beta = 0.7$	22.3 ± 1.2	26.7 ± 1.3	**35.4 ± 1.4**
$\beta = 0.8$	21.9 ± 2.0	25.3 ± 1.2	33.4 ± 3.1
$\beta = 0.9$	20.4 ± 1.1	25.9 ± 1.1	34.8 ± 1.9
$\beta = 1.0$	20.3 ± 1.1	25.2 ± 2.3	34.5 ± 1.3

the standard deviation of the performance of these different models. Notice that we compare our approach to a random selection and to two baseline selection criteria. These baselines consist in adding a dropout layer at the end of the encoder of our model, with 50% of probability of dropping out the neurons. Following, we test each of the trained models 5 different times with the dropout, and obtain the predicted masks for each run. We compute the standard deviation of the pixels from the masks predicted, to see which samples vary significantly between different runs when different neurons are dropped, as a way to estimate the uncertainty of the predictions. Finally, we select the images related to the lowest standard deviation values (Dropout Baseline lowest) or the highest values (Dropout Baseline highest), similarly to previous works [12][15].

The results in Table 3 show that there are multiple subsets that outperform the random and the baseline selections. This means that our selection strategy based on IoU is effective to reach better performance. We also notice that the optimal predefined *IoU Score* is not fully consistent across different subsets sizes (at $N = 800$ the optimal score is 0.6, whereas at $N = 200$ the optimal score is 0.3). Interestingly, these optimal *IoU Score* values suggest that the best options are the ones that select images that are neither the most challenging of the dataset nor the easiest ones. We also observe that the dropout baselines do not surpass our method. As our results with *IoU Score* indicate, the best options for the dropout baselines may be related to neither the highest nor lowest standard deviations. However, choosing a mid-range option for standard deviation is not as intuitive as it is with *IoU Score*. In the latter case, we

Table 4 Predicted IoU: mean Average Precision (th=0.5) for different selection criteria (5 runs for each configuration). Each column indicates the number of images used to train the segmentation models. The first row shows the results obtained with random selection of samples, the second and third rows show a baseline sample selection method, whereas the following rows show different selection criteria with our method. If $\beta = 0.0$ and the number of training samples is 200, means that the first 100 samples are randomly-selected and the next 100 are the ones that have a predicted IoU closer to 0.0.

	200	**400**	**800**
Random subset	22.7 ± 1.8	27.1 ± 0.8	34.5 ± 2.0
Dropout Baseline (highest)	21.4 ± 1.4	23.9 ± 1.3	28.1 ± 1.9
Dropout Baseline (lowest)	20.0 ± 0.9	24.8 ± 1.5	32.2 ± 1.4
$\beta = 0.0$	21.5 ± 1.1	23.7 ± 0.6	30.1 ± 1.7
$\beta = 0.1$	21.8 ± 1.6	23.7 ± 0.7	30.3 ± 1.7
$\beta = 0.2$	22.6 ± 0.9	25.0 ± 0.8	29,9 ± 2.2
$\beta = 0.3$	**24.0 ± 1.3**	26.9 ± 3.2	33,5 ± 3.1
$\beta = 0.4$	23.2 ± 0.4	24.8 ± 2.2	35.3 ± 0.9
$\beta = 0.5$	20.9 ± 3.1	25.0 ± 0.9	**37.0 ± 2.0**
$\beta = 0.6$	20.6 ± 1.2	**27.5 ± 2.7**	34.8 ± 3.0
$\beta = 0.7$	20.3 ± 1.0	26.2 ± 3.1	36.3 ± 1.1
$\beta = 0.8$	20.7 ± 2.1	26.9 ± 1.6	35,9 ± 2.5
$\beta = 0.9$	20.8 ± 0.8	26.1 ± 1.2	35,5 ± 1.1
$\beta = 1.0$	21.1 ± 1.5	24.8 ± 1.5	34.6 ± 2.1

only need to select an IoU which directly reflects the quality of the predicted masks. Moreover, our method does not need to run the model several times to select the samples, as it happens with the dropout baselines, thus it is computationally more efficient.

6.2.2 Predicted IoU-selection:

The experiments in Section 6.2.1 with the real ground truth IoU (the *oracle* experiment) showed that choosing samples based on the IoU quality metric leads to better results than performing a random selection or a baseline active learning selection.

In this section, we address the realistic case in which the IoU is predicted by the same annotation network, instead of using the ground truth value as in Section 6.2.1. Table 4 shows that for the three set sizes (N = 200, 400, 800) better results are also obtained by selecting with the IoU criterion instead of performing a random selection or using the dropout baseline defined previously. The optimal *IoU scores* are between 0.3 and 0.6. In fact, we observe a tendency that for smaller subsets, a lower *IoU score* is optimal, whereas for larger subsets, a higher *IoU score* works better. We also observe there is no significant difference between the results obtained with the *oracle* and the predicted IoU configuration.

6.2.3 Sets analysis:

In this section we will analyse the properties of the *N'* samples selected based on the sample selection criterion when considering different *IoU Scores* predefined values. We compare the subsets obtained from the *oracle* and the predicted IoU configurations. In Figure 5 we depict an histogram of the average number of objects per image and the mean size of objects per image for each of the subsets, depending on the predefined *IoU Scores*. The plot has two different columns, the first one belongs to the *oracle* configuration and the second one to the predicted IoU configuration. For both the *oracle* and the predicted IoU configurations, we observe that lower *IoU scores* are related to images with more objects per image and smaller objects. These two scenarios correspond to very challenging cases in object detection, as pointed out by previous works [14]. Finally, we can observe that the subsets created by the predicted IoU follow a similar distribution to the *oracle* one.

As we observed in Section 6.2.2, the optimal *IoU Scores* are between 0.3 and 0.6. In Figure 5 we can see how images associated to these values tend to have a close to the average number of objects per image (2.8 objects/image). Regarding object size, we observe that objects tend to be neither the largest ones nor the smallest.

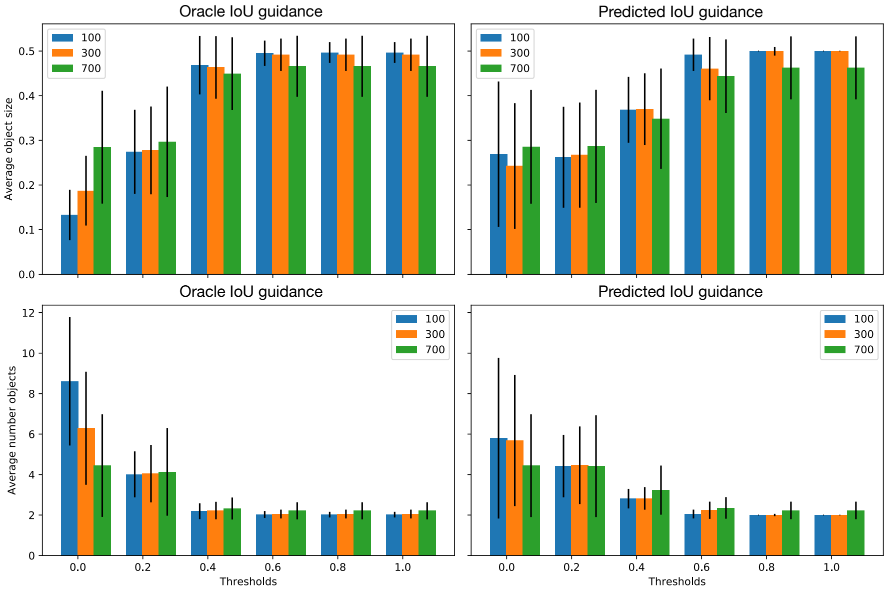
Figure 6 shows some of the selected images when different *IoU Scores* are considered. We observe that at high *IoU Scores* values (0.8 or 1.0), images selected are easy, with only one or two large objects in the image. On the other hand, at low *IoU Scores* (0.0 or 0.2) images have multiple, rather small, instances. As our results indicate, the optimal selected samples to be strongly annotated are those in the middle of the range. These are images that have multiple instances but that are not too complicated to segment. We hypothesize that training with very difficult images can be inefficient if the model is not capable to learn from them, while easy cases do not add much value to the learning process.

6.2.4 Training of segmentation network:

In this section we focus on the final goal of the pipeline: training the *segmentation* network. As a first step, an *annotation* network of $N = 200$ is trained with 100 random samples and 100 selected samples (the ones that are closest to the *IoU score* of 0.3, which is the optimal for this set size). The same procedure applies for $N = 400$ and $N = 800$, with thresholds 0.6 and 0.5 respectively. Once the *annotation* network has been trained with the optimal selection of samples given our mask-guided criterion, we use the network to pseudo-annotate the additional Pascal set from [16], a total of 9118 images. Finally, we train the *segmentation* network with the obtained pseudo-annotations and the available strongly-labeled samples.

Table 5 compares the random selection of samples with the mask-guided selection in terms of mean Average Precision (threshold 0.5). We observe that,

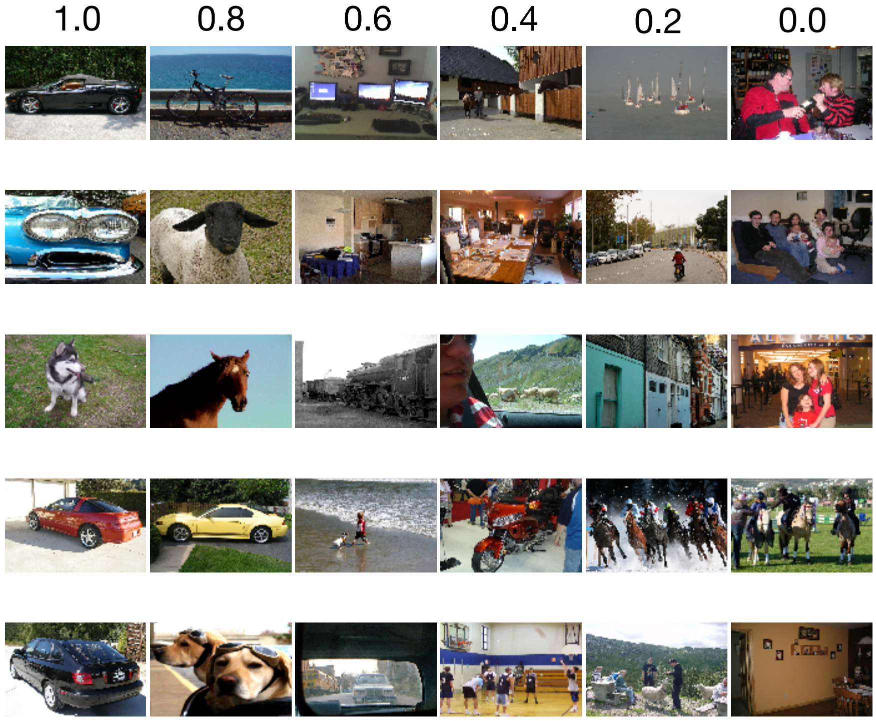
Fig. 5 Analysis of the mean object size (first row) and number of objects (second row) of the selected images when considering the oracle and the predicted IoU. We consider three different scenarios, when 100, 300 or 700 samples are added to the initial 100 randomly-selected set of samples.



for both the *annotation* and the *segmentation* networks, the mask-guided selection reaches better results. Notice that the results obtained with the *annotation* network correspond to a case without semi-supervision, when only strong labels are used to train *W-RSIS* or *IoU-W-RSIS* architectures. On the other hand, the semi-supervised solution would correspond to the *segmentation* network, with the *RSIS* architecture, whose results are significantly better. The fully-supervised setup, when approximately 10k strongly-labeled images are used for training, corresponds to an Average Precision at threshold 0.5 of 57.0, as reported in [44].

The annotation budgets for each configuration in Table 5 are reported in Table 6. We observe that the mask-guided selection options have a slight higher budget compared to the random ones. This is because the *IoU-W-RSIS annotation* network takes as input the image-level labels plus counts. We first predict the *IoU score* for all samples from the Pascal VOC 2012 training set (1464 samples), and then use those scores to perform the mask-guided selection. Therefore, we need to add the cost for the image-level plus counts labels for *1464-N* samples (as *N* will be strongly-annotated and already considered in the annotation budget). Figure 7 provides a qualitative comparison between the models obtained from the *annotation* and *segmentation* networks.

Fig. 6 Examples of images of each subset. Each column are images related to different predicted IoUs. For instance, the first column belongs to the images that have a mean IoU closer to 1, and we can see that effectively these images look easy, with a single and big instance appearing.



We observe that the more strongly-annotated samples (N), the better quality for the obtained masks. We also observe that the results for the *segmentation* networks are higher than those from the *annotation* networks, proving that the pseudo-annotations are beneficial.

In Table 7 we report the mean Average Precision at threshold 0.5 when only the 50% of the additional set of Pascal VOC [16] is weakly-annotated, and therefore the associated budget (Table 8) is lower. In this case we also observe how the configuration with mask-guided selection outperforms the random one. We lead this experimentation to show that at lower annotation budgets, this configuration still works better.

The final results obtained by the *segmentation* network are presented in Tables 9 and 10 with three complementary metrics: Average Precision and Recall at different thresholds, the F measure at threshold 0.5, which corresponds to $F = 2 * (precision * recall)/(precision + recall)$, and the Structural Similarity Index (SSI), as an effort to have a metric that considers the overall

Table 5 Comparison of *annotation* and *segmentation* networks mean Average Precision (th=0.5) on Pascal VOC depending on the selection strategy and the number of strongly-annotated samples used to train the *annotation* network. Notice that the *segmentation* network is trained with both the strongly-annotated samples and the pseudo-labels obtained with the *annotation* network. The pseudo-labeled samples are the complete additional set of Pascal (9118 images).

Selection	Annotation Network			Segmentation Network		
Strong labels	200	400	800	200	400	800
Pseudo-labels	-	-	-	9118	9118	9118
Random (W-RSIS)	22.7	27.1	34.5	33.3	36.8	43.8
Mask-guided (IoU-W-RSIS)	24.0	27.5	37.0	34.4	41.8	47.1

Table 6 Comparison of *annotation* and *segmentation* networks annotation budget in days depending on the selection strategy and the number of strongly-annotated samples used to train the *annotation* network. Notice that the *segmentation* network is trained with both the strongly-annotated samples and the pseudo-labels obtained with the *annotation* network. The samples pseudo-labeled are the complete additional set of Pascal (9118 images).

Selection	Annotation Network			Segmentation Network		
Strong labels	200	400	800	200	400	800
Pseudo-labels	-	-	-	9118	9118	9118
Random (W-RSIS)	0.55	1.11	2.22	2.90	3.45	4.56
Mask-guided (IoU-W-RSIS)	0.90	1.38	2.39	3.25	3.73	4.73

structure of the mask instead of pixel-wise errors. Observing these metrics, we see that for $N = 400$ and $N = 800$, using more pseudo-labels (9118 vs. 4559) leads to better performance, while for $N = 200$ it is not the case. This may be produced by the different ratio of pseudo-labeled samples vs. strongly-labeled samples, which is significantly larger for $N = 200$. Having a large ratio of noisy labels compared to reliable ones, could damage the training of the model. We also observe that when varying the number of pseudo-labels (9118 vs. 4559), the SSIM does not change as much as we see for the other metrics. Interestingly, the SSIM for $N = 200$ is higher than for $N = 400$. The reason could be that with $N = 200$ the blobs obtained in the masks are coarser and this could favour this metric because it is based on structural similarity. Nevertheless, the difference is not very significant between both configurations.

7 Conclusion and Future Work

We have proposed a novel method to select which samples to strongly-annotate in the semi-supervised instance segmentation setup. Our method, based on IoU prediction, outperforms the baseline random selection and a solution based on neural dropout to estimate pixel-wise uncertainty. We guided a thorough analysis of which samples are best to annotate given the confidence score of the predictions, and we observe that the best samples are those that fall in the mid-range of the IoU scores. With our pipeline, we present a very simple

Table 7 Comparison of *annotation* and *segmentation* networks mean Average Precision (th=0.5) on Pascal VOC depending on the selection strategy and the number of strongly-annotated samples used to train the *annotation* network. Notice that the *segmentation* network is trained with both the strongly-annotated samples and the pseudo-labels obtained with the *annotation* network. The samples pseudo-labeled are 50% of the additional set of Pascal (4559 images).

Selection	Annotation Network			Segmentation Network		
Strong labels	200	400	800	200	400	800
Pseudo-labels	-	-	-	4559	4559	4559
Random (W-RSIS)	22.7	27.1	34.5	33.3	36.8	43.8
Mask-guided (IoU-W-RSIS)	24.0	27.5	37.0	34.6	38.8	46.2

Table 8 Comparison of *annotation* and *segmentation* networks annotation budget in days depending on the selection strategy and the number of strongly-annotated samples used to train the *annotation* network. Notice that the *segmentation* network is trained with both the strongly-annotated samples and the pseudo-labels obtained with the *annotation* network. The samples pseudo-labeled are 50% of the additional set of Pascal (4559 images).

Selection	Annotation Network			Segmentation Network		
Strong labels	200	400	800	200	400	800
Pseudo-labels	-	-	-	4559	4559	4559
Random (W-RSIS)	0.55	1.11	2.22	1.73	2.28	3.39
Mask-guided (IoU-W-RSIS)	0.90	1.38	2.39	2.08	2.55	3.56

Table 9 Average Precision and Average Recall at different thresholds, F measure and SSIM for the *segmentation* networks with the Mask-guided criterion. The samples pseudo-labeled are the complete additional set of Pascal (9118 images).

	AP @[0.5:0.95]	AP @[0.5]	AP @[0.7]	AR @[0.5:0.95]	AR @[0.5]	AR @[0.7]	F@[0.5]	SSIM
200	18.7	34.4	20.6	26.1	41.6	28.7	37.7	84.0
400	24.8	41.8	28.2	33.6	50.2	37.5	45.6	83.6
800	29.2	47.1	32.7	38.7	55.4	42.9	50.9	85.8

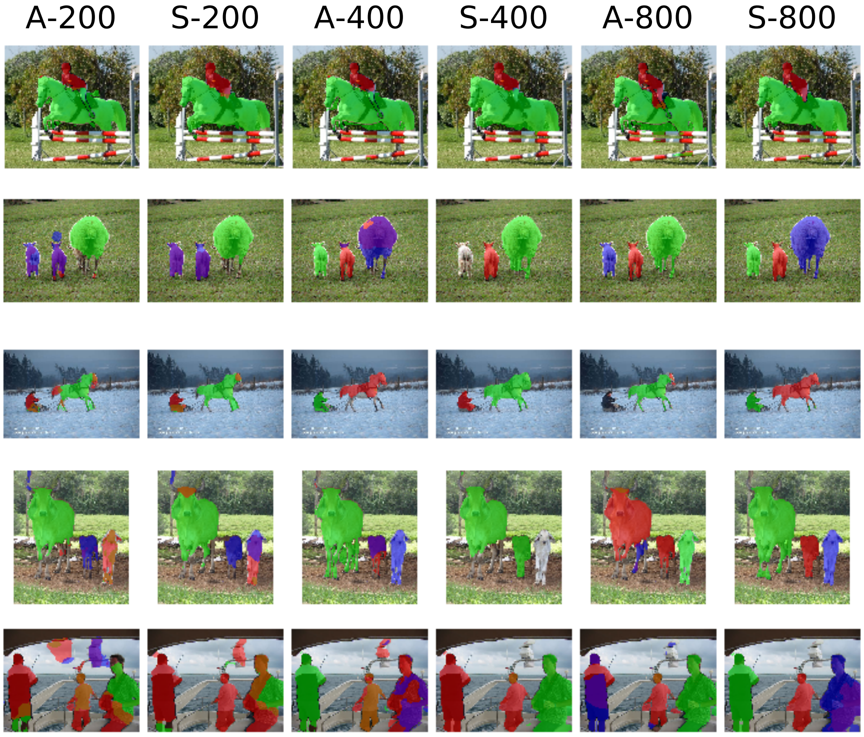
Table 10 Average Precision and Average Recall at different thresholds, F measure and SSIM for the *segmentation* networks with the Mask-guided criterion. The samples pseudo-labeled are 50% of the additional set of Pascal (4559 images).

	AP @[0.5:0.95]	AP @[0.5]	AP @[0.7]	AR @[0.5:0.95]	AR @[0.5]	AR @[0.7]	F@[0.5]	SSIM
200	19.5	34.6	21.7	27.3	42.8	30.1	38.3	84.4
400	22.8	38.8	26.4	32.5	48.2	37.0	43.0	83.7
800	29.0	46.2	33.0	37.9	54.2	42.4	49.9	85.9

but effective manner to perform sample selection to improve performance at a negligible annotation cost.

To conclude, instance segmentation is a very challenging task in the field of scene understanding, and our experimental validation proves that the task can be addressed with low annotation budgets, and that exploiting few but interesting samples can lead to better results. As an improvement for our

Fig. 7 Visualization of Pascal VOC test set for the annotation f_{θ_N} *(A-)* and segmentation networks g_{φ_N} *(S-)*, depending on the number of strong labels used $N \in \{200, 400, 800\}$.



method, in the future we would like to add samples to the training pipeline in an iterative way, similar to traditional active learning pipelines.

As future work we think that IoU prediction for sample selection can be exploited in other tasks, such as semantic segmentation and object detection. In addition, the impact of our semi-supervised methodology can be particularly relevant for video-related tasks, such as video object segmentation, as videos require significant effort to be annotated pixel-wise due to the large amount of data. Previous works on video object segmentation with low annotation budgets have focused on unsupervised approaches [30] leveraging attention and salient cues, widely used for video [48]. We believe that a semi-supervised pipeline can reach significant better performance with a low annotation cost. Another task worth exploring is salient object detection, as our sample selection strategy of predicting the segmentation quality can also be applied to salient segments [48][9][8].

8 Acknowledgments

This work was partially supported by the Spanish Ministry of Economy and Competitivity under contracts TIN2012-34557 by the BSC-CNS Severo Ochoa program (SEV-2011-00067), and contracts TEC2013-43935-R and TEC2016-75976-R. It has also been supported by grants 2014-SGR-1051 and 2014-SGR-1421 by the Government of Catalonia, and the European Regional Development Fund (ERDF). We would also like to acknowledge the valuable discussions with Victor Campos.

References

1. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
2. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
3. Bellver, M., Salvador, A., Torrres, J., Giro-i Nieto, X.: Budget-aware semi-supervised semantic and instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 93–102 (2019)
4. Brust, C.A., Käding, C., Denzler, J.: Active learning for deep object detection. arXiv preprint arXiv:1809.09875 (2018)
5. Dong, J., Lin, T.: Margingan: Adversarial training in semi-supervised learning. In: Advances in Neural Information Processing Systems, pp. 10440–10449 (2019)
6. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. CRC press (1994)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision (2010)
8. Fan, D.P., Cheng, M.M., Liu, J.J., Gao, S.H., Hou, Q., Borji, A.: Salient objects in clutter: Bringing salient object detection to the foreground. In: Proceedings of the European conference on computer vision (ECCV), pp. 186–202 (2018)
9. Fan, D.P., Wang, W., Cheng, M.M., Shen, J.: Shifting more attention to video salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8554–8564 (2019)
10. Feige, U.: A threshold of ln n for approximating set cover. Journal of the ACM (JACM) **45**(4), 634–652 (1998)
11. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning, pp. 1050–1059 (2016)
12. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1183–1192. JMLR. org (2017)
13. Gao, M., Li, A., Yu, R., Morariu, V.I., Davis, L.S.: C-wsl: Count-guided weakly supervised localization. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
14. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448 (2015)
15. Gorriz, M., Carlier, A., Faure, E., Giro-i Nieto, X.: Cost-effective active learning for melanoma segmentation. arXiv preprint arXiv:1711.09168 (2017)
16. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Proceedings of the IEEE international conference on computer vision (ICCV) (2011)
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2016)
19. Hou, Q., Cheng, M.M., Liu, J., Torr, P.H.: Webseg: Learning semantic segmentation from web searches. arXiv preprint arXiv:1803.09859 (2018)
20. Hu, R., Dollár, P., He, K., Darrell, T., Girshick, R.: Learning to segment every thing (2018)
21. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6409–6418 (2019)
22. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
23. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 784–799 (2018)
24. Khoreva, A., Benenson, R., Hosang, J.H., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
25. Laradji, I.H., Vazquez, D., Schmidt, M.: Where are the masks: Instance segmentation with image-level supervision. arXiv preprint arXiv:1907.01430 (2019)
26. Li, Q., Arnab, A., Torr, P.H.: Weakly-and semi-supervised panoptic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
27. Li, W., Wang, Z., Yue, Y., Li, J., Speier, W., Zhou, M., Arnold, C.W.: Semi-supervised learning using adversarial training with good and bad samples. arXiv preprint arXiv:1910.08540 (2019)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV) (2014)
29. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision, pp. 21–37. Springer (2016)
30. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3623–3632 (2019)
31. Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Reyes, M.: 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, pp. 580–588. Springer (2018)
32. Ozdemir, F., Fuernstahl, P., Goksel, O.: Learn the new, keep the old: Extending pretrained models with new anatomy and images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 361–369. Springer (2018)
33. Ozdemir, F., Peng, Z., Tanner, C., Fuernstahl, P., Goksel, O.: Active learning for segmentation by optimizing content information for maximal entropy. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 183–191. Springer (2018)
34. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Extreme clicking for efficient object annotation. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
35. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision (ICCV) (2015)
36. Pathak, D., Krahenbuhl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE international conference on computer vision (ICCV) (2015)
37. Pathak, D., Shelhamer, E., Long, J., Darrell, T.: Fully convolutional multi-class multiple instance learning. arXiv preprint arXiv:1412.7144 (2014)

38. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
39. Pont-Tuset, J., Arbelaez, P., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE transactions on pattern analysis and machine intelligence (2017)
40. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788 (2016)
41. Remez, T., Huang, J., Brown, M.: Learning to segment via cut-and-paste. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
42. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM transactions on graphics (TOG) (2004)
43. Roy, S., Unmesh, A., Namboodiri, V.P.: Deep active learning for object detection. In: British Machine Vision Conference, pp. 3–6 (2018)
44. Salvador, A., Bellver, M., Campos, V., Baradad, M., Marqués, F., Torres, J., Giro-i Nieto, X.: Recurrent neural networks for semantic instance segmentation. arXiv preprint arXiv:1712.00617 (2017)
45. Settles, B.: Active learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (2009)
46. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of machine learning research **2**(Nov), 45–66 (2001)
47. Vijayanarasimhan, S., Grauman, K.: Large-scale live active learning: Training object detectors with crawled data and crowds. International Journal of Computer Vision **108**(1-2), 97–114 (2014)
48. Wang, W., Shen, J., Yang, R., Porikli, F.: Saliency-aware video object segmentation. IEEE transactions on pattern analysis and machine intelligence **40**(1), 20–33 (2017)
49. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
50. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
51. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems (2015)
52. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, pp. 399–407. Springer (2017)
53. Zhang, T., Lin, G., Cai, J., Shen, T., Shen, C., Kot, A.C.: Decoupled spatial neural attention for weakly supervised semantic segmentation. arXiv preprint arXiv:1803.02563 (2018)
54. Zhao, X., Liang, S., Wei, Y.: Pseudo mask augmented object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
55. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
56. Zhou, Y., Zhu, Y., Ye, Q., Qiu, Q., Jiao, J.: Weakly supervised instance segmentation using class peak response. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)