CEREALS – Cost-Effective REgion-based Active Learning for Semantic Segmentation

Radek Mackowiak[1]
radek.mackowiak@de.bosch.com
Philip Lenz[1]
philip.lenz@de.bosch.com
Omair Ghori[1]
omair.ghori@de.bosch.com
Ferran Diego[1]
ferran.diego@de.bosch.com
Oliver Lange[1]
oliver.lange@de.bosch.com
Carsten Rother[2]
carsten.rother@iwr.uni-heidelberg.de

[1] Robert Bosch GmbH
Corporate Research - Computer Vision
Robert-Bosch-Straße 200
31139 Hildesheim, DE

[2] Heidelberg Collaboratory for Image Processing (HCI)
Berliner Straße 43,
69120 Heidelberg, DE

Abstract

State of the art methods for semantic image segmentation are trained in a supervised fashion using a large corpus of fully labeled training images. However, gathering such a corpus is expensive, due to human annotation effort, in contrast to gathering unlabeled data. We propose an active learning-based strategy, called CEREALS, in which a human only has to hand-label a few, automatically selected, regions within an unlabeled image corpus. This minimizes human annotation effort while maximizing the performance of a semantic image segmentation method. The automatic selection procedure is achieved by: a) using a suitable information measure combined with an estimate about human annotation effort, which is inferred from a learned cost model, and b) exploiting the spatial coherency of an image. The performance of CEREALS is demonstrated on Cityscapes, where we are able to reduce the annotation effort to 17%, while keeping 95% of the mean Intersection over Union (mIoU) of a model that was trained with the fully annotated training set of Cityscapes.

1 Introduction

Deep convolutional neural networks (CNNs) have become the de-facto standard method for solving a large variety of heterogeneous image understanding problems. In the domain of visual scene understanding, semantic segmentation plays an important role due to enabling machines a pixel-wise semantic understanding of their environment. It is therefore a key enabler for applications like autonomous driving or robotic vision. However, one shortcoming of current CNN training algorithms is that they require a large amount of diverse and labeled training data to achieve satisfying results. Furthermore, their performance seems to scale linearly with an exponential increase of training data [48].

(a) Ground Truth (b) Full training set (100% labeling effort) (c) 17% of the effort by random data selection (d) 17% of the effort using our approach.

Figure 1: Qualitative segmentation results using *CEREALS* for data annotation. Our approach reduces labeling effort significantly. We achieve 95% of the performance with only 17% of the labeling effort measured by the number of clicks as compared to annotating the full training set of Cityscapes [5].

While acquiring large amounts of unlabeled data is usually easy, the effort required to manually annotate this data is a costly process due to the requirement of human annotators [5, 24]. Hence, the major bottleneck for rapidly applying CNN models into new domains is the acquisition of large-scale labeled training sets. Active Learning (AL) [4] is an established approach to mitigate the problems associated with data labeling. In essence AL aims to query the data only for annotation, which is more likely to lead to more accurate models when used for training than other data [12, 49, 50]. Consequently, this mitigates the time and monetary cost associated with the labeling effort.

Currently only few AL approaches [12, 40, 41, 53, 56] evaluated on CNNs exist. Furthermore, most of the proposed AL methods in computer vision problems are focusing on image-level classification tasks [12, 25, 40, 41, 52, 53]. In contrast, very few works apply AL on CNNs with spatial and dense output spaces [14, 56].

Regarding semantic segmentation the relationship between annotation time, amount of label noise and a resulting deep model's performance with respect to its capacity has not been investigated to the best of our knowledge. Therefore, we focus our work on how to cost-effectively create *reliable* training data for learning high performing CNNs for semantic segmentation by utilizing AL. We decided to use AL since, as far as we are aware of, it is the only paradigm optimized for cost-effectively generating reliable dense semantic segmentation annotations of real-world imagery data without the need of other input modalities.

In this work, we propose a novel cost-effective active learning framework tailored to multi-class semantic segmentation (*CEREALS*). In particular, we aim to iteratively find the *minimal* set of highly informative data while *minimizing* the annotation effort, in order to achieve a desired high quality performance with minimal costs. We approximate costs by the amount of user interactions measured by the number of clicks performed during the annotation process. The proposed framework reduces the labeling effort by (i) utilizing spatial estimates about annotation costs inferred from a learned cost prediction CNN and (ii) by focusing on image regions promising high information content and low annotation costs in a global context. We demonstrate the performance of *CEREALS* on Cityscapes [5], a very complex dataset consisting of high definition natural urban scene images. A qualitative result of our approach is depicted in Fig.1.

2 Related Work

Densely annotating images with pixel-accurate multi-class semantic segmentation masks requires considerably more time than annotating images with univariate multi-class annota-

tions [54]. To lessen the burden of manual annotation, an array of six different strategies has been explored in the literature.

1) Pre-training is a standard practice whenever the amount of available ground truth data is relatively scarce. Initially training a deep CNN on less complex but large-scale databases such as Imagenet [7] results in better discrimination ability of the final model [13, 17, 34, 48].

2) Weakly-supervised learning has shown promising results towards solving semantic segmentation tasks [15, 20, 22, 29, 33, 35, 38, 39, 46, 57]. For instance, annotating a dataset just based on the annotator's binary decision if a class is present in an image or not, or annotating bounding boxes is faster than producing dense segmentation masks. However, given a sufficient amount of data, models trained in a supervised way outperform any weakly-supervised method.

3) Semi-supervised learning methods have been shown to increase a model's performance using a mixture of fine-grained labels plus additional unlabeled data to achieve better results than labeled data alone [18, 35, 47].

4) Synthetic data generation methods produce synthetic images together with highly accurate dense annotations [1, 36, 44], but the shortcoming lies in the effort required to generate diverse and realistic sceneries [1], which is a crucial aspect for achieving satisfactory results in real-world scenarios [31, 58].

5) Interactive segmentation is the process of extracting objects of interest by utilizing sparse user input. It directly improves the annotation tools for assisting human annotators by increasing their efficiency [8, 9, 28]. Though it has been recently shown that by utilizing CNNs the segmentation quality can be drastically improved compared to the previous state of the art [55], these methods are still suffering from imprecision [2]. Castrejon *et al*. show how this problem could be treated by allowing the annotator to refine estimated polygons for cost-effectively generating reliable instance annotations [2], however the work doesn't show quantitative results on dense semantic segmentation. Xie *et al*. [54] and Liang-Chieh *et al*. [3] presented methods for effectively generating multitude of ground truth instance segmentations. Both methods are dependent on the utilization of lidar sensors.

6) Active learning is described in the survey from Settles [42]. It offers a high-level overview over the commonly used methods. *Pool-based active learning* [26] exploits the inequality of amount of information in an existing unlabeled pool and aims to find the most valuable sample to be labeled by an oracle able to reveal the ground truth semantics of interest given some data. *AL on Semantic Segmentation* has been investigated in [19, 23, 30, 51]. Both methods [23, 51] rely on a previously processed oversegmentation of an unlabeled image for retrieving its superpixels. Informative regions however are not restricted to the extent of superpixels. Annotating them furthermore does not guarantee a reliable labeling because oversegmentation algorithms often fail to separate semantic regions when the transition from one class to the next is smooth in the underlying input space. The method proposed by Mosinska-Domanska *et al*. [30] operates only on curvilinear structures and [19] is investigating the propagation of segmentations from informative data to unlabeled data. *AL for CNNs on Semantic Segmentation* running on top of unstructured state-of-the-art models, namely fully convolutional neural networks [27], has so far to the best of our knowledge only been investigated in [14, 56]. Both approaches focus on foreground/background segmentation and assume an equal annotation effort for all images, which we later show to be a simplified assumption. *Cost-Effective AL for CNNs* has been recently proposed in [53] for image classification tasks, where highest confidence pseudo-annotated unlabeled samples are added to the training set with no human cost at all. The same idea has been adopted in [14] for medical image segmentation. This method assumes that highly confident predic-

tions are labeled correctly; Hence, selected samples could introduce hard label noise during training. Although this technique shows an improvement regarding general performance, it may lead to unwanted side-effects on corner-cases due to strengthening wrong, but certain predictions. Wherever in this work the authors propose to score the entire images, we are 1.) scoring all possible fixed-size regions across all images in an unlabeled pool and 2.) querying the ones for annotation expected to have the highest positive impact on the model's performance. Most related to our approach regarding cost-effectiveness in active learning are the methods described in [25, 43, 52]. All these works employ a cost prediction model trained on data where target labels are available from the beginning or after previously executed acquisition steps. We adapt this idea to the domain of semantic segmentation by estimating spatial information about costs in order to find highly informative, but cheap regions in an unlabeled pool of images to be annotated, where the quality of annotation depends only on the quality of the executing annotators and their given labeling instructions. Despite the fact that AL could be incorporated alongside all the previously mentioned approaches, in this work we are considering a CNN, specifically a fully convolutional neural network, trained in a strongly supervised manner and a simple polygon-based annotation tool, similar to ones used for constructing training datasets of state-of-the-art benchmarks [5, 32, 59]. Though not being optimized for efficiency, it allows the production of fine-grained annotations [37] for training high quality CNNs on multi-class semantic segmentation tasks.

3 Method

In classical pool-based AL typically only a single sample out of an unlabeled pool is queried to be labeled by an oracle in each step of the iterative algorithm. Since deep CNN training algorithms need a long time to converge on currently available hardware, such a setup is however practically infeasible. Therefore, we consider a pool-based AL scenario running in batch-mode. In such a setting a large unlabeled pool of data exists from which a small, randomly sampled subset, called the seed set, is initially extracted and labeled by an oracle. Using this seed set the algorithm works as follows: First, a model is trained on the currently labeled pool. Secondly, some measure of information on each individual unlabeled sample is being computed. Thirdly, an acquisition function is applied. A subset of a pre-specified amount of elements maximizing the acquisition function is annotated by an oracle. It is then added to the labeled pool. The process is repeated until either a desired performance or labeling budget is reached. Furthermore, the stopping criterion is satisfied whenever the unlabeled pool becomes exhausted which is indicated by no further improvements after several acquisition steps.

The main attention in pool-based AL research has been given to information measures being computed on the posterior probability distribution of a classifier given some input data. An acquisition function for batch-mode AL scenarios is often extended with density weighting approaches aiming to select samples maximizing not only information content but also diversity. Since the problem of semantic segmentation does not only allow the exploration of novel information measures and/or diversity maximization methods, our approach focuses on the acquisition process itself. In a typical AL scenario for image classification problems, a subset of promising images from an unlabeled pool is sampled. Such an acquisition function has been adopted for semantic segmentation in [14] for retrieving images, but based on their accumulated per-pixel information content projecting all the information extracted from an image onto a single value. We are proposing to design acquisition to explicitly focus

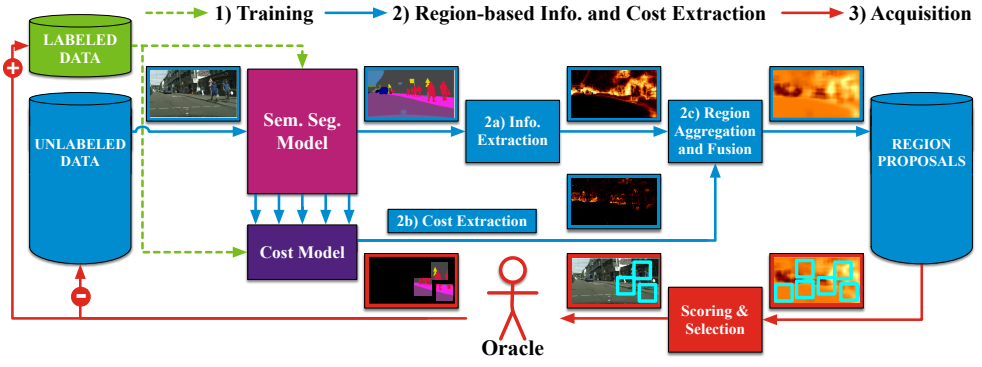


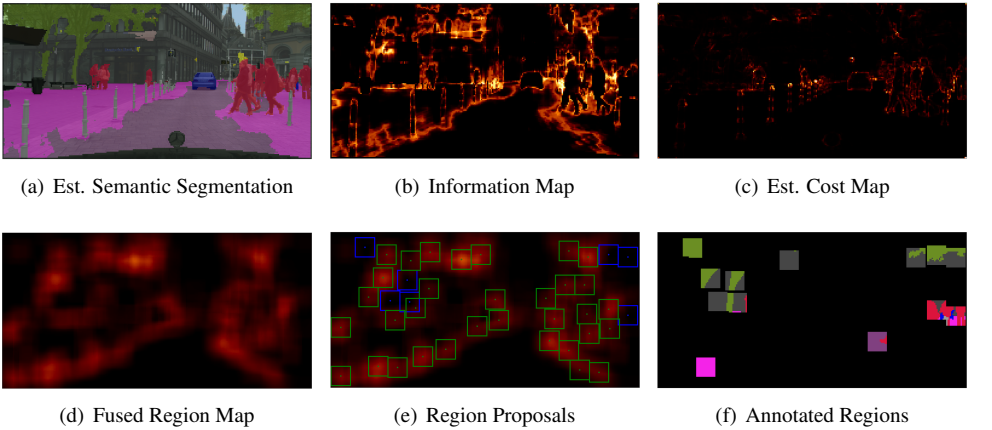
Figure 2: Diagram of the proposed framework for cost effective region-based active learning for semantic segmentation. We fuse spatial information about information content and cost estimates in order to query the most promising regions for annotation maximizing an information/cost trade-off. Our approach consequently aims to boost the performance of a CNN as cheaply as possible.

on image regions inside of the entire unlabeled pool of images and further to not only consider information during region selection but also annotation costs. The proposed method is depicted in Fig.2 and works as follows:

1) Training For constructing the seed set we uniformly sample n images to be fully labeled by an oracle. We then learn two deep convolutional neural networks. One being the *semantic segmentation model* based on the *FCN8s* architecture [27], whose training is initialized using imagenet-pretrained weights. For faster training we apply the width multiplier introduced in [16] with its value set to 0.25. Furthermore, we discarded the $8\times$ upsampling and instead scale down the spatial annotations during training, since we observe it to have only marginal impact on the model's final performance. For validation however, we add a $8\times$ bilinear upsampling layer. The other model, which is trained directly after the *semantic segmentation model*, is a *cost model* based on [21]. It utilizes the semantic segmentation networks learned knowledge as prior information to estimate the clicks an annotator would have needed to execute for densely annotating an image. All further implementation details are reported in A.1.

2a) Information Extraction In this work we raise awareness towards costs. We therefore compare two classical heterogeneous information measures only. We are computing both information measures for each pixel location individually given the *a-posteriori* probability distributions retrieved from the activations produced by the employed semantic segmentation CNN's softmax layer. In the following formulas $P^{(u,v)} \hat{=} P^{(u,v)}(f_\theta(x))$ is the probability class distribution at a specific pixel position (u,v) retrieved from a model f parameterized by θ given some image x. A specific class out of a set of considered classes is denoted by c. The resulting *information map* (Fig.3(b)) contains the information content for each pixel of an image at a current acquisition step.

Entropy [45] is the most widely used information measure seen in active learning literature. Here, the data with the highest positive impact on the model's performance is estimated to be the one where it's posterior probability distribution produces the highest entropy. Entropy is, inter alia, used as a measure of uncertainty, since its value is maximized when the



(a) Est. Semantic Segmentation (b) Information Map (c) Est. Cost Map

(d) Fused Region Map (e) Region Proposals (f) Annotated Regions

Figure 3: Visualization of *CEREALS* query selection behaviour during acquisition step two. Blue boxes in (e) represent regions annotated at the end of the previous acquisition step one. Green boxes represent regions within the region proposal pool of the current acquisition step. *CEREALS* queries the best regions out of the region proposal pool for annotation (f).

model assigns each considered class the same probability and very small if the model is sure about its decision. We compute entropy for retrieving per-pixel information as follows.

$$H^{(u,v)} := -\sum_c P_c^{(u,v)} \cdot log(P_c^{(u,v)}) \tag{1}$$

The Vote Entropy [6] information measure entails first constructing a committee E of N_E different classifiers that ideally are all consistent with the labeled pool. Each committee member e places a vote on vector $P_e^{(u,v)}$. Then a disagreement factor among the members is calculated. We utilize vote entropy which we adapt for the semantic segmentation case as follows.

$$V^{(u,v)} := -\sum_c \frac{\sum_e D(P_e^{(u,v)},c)}{N_E} \cdot log \frac{\sum_e D(P_e^{(u,v)},c)}{N_E} \quad where \quad D(a,c) = \begin{cases} 1, & \text{if } argmax(a) = c \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Instead of training N_E different classifiers on the same training data, we leverage the stochasticity provided by the dropout layers of our employed *semantic segmentation model* and construct a Monte-Carlo dropout ensemble as in [11]. The most informative data points are the ones having the highest disagreement factor among the committee members. The aim of such *Query by Committee* [10] approaches is to sample data expected to reduce the version space of given committee members.

2b) Cost Extraction Our work is based on the assumption that some samples in an unlabeled pool are more costly to label by a human oracle than others and further that this also applies to regions within images. To the best of our knowledge no published dataset addressing semantic segmentation provides information on annotation costs. As in [9, 55], we are approximating costs by the number of clicks necessary to annotate an image. Cityscapes is the only dataset providing information about where and how often a user has clicked to label an image. Obviously, this information is unknown for unlabeled data. For this reason we train the *cost model* on all the click data which was produced by human annotators at previous acquisition steps. During actual cost extraction we perform a forward pass for each

individual image within the current unlabeled pool through the *cost model* for retrieving an estimate about clicks. We denote the result given an image as *cost map* (Fig.3(c)).

2c) Region Aggregation and Fusion We argue that not all regions in an image boost a CNN's performance equally. Thus, some regions may have not only different labeling costs but also may provide supervisory signals of different impact on the model's performance than other regions. Theoretically, some regions could be very costly to label while having only little positive impact on the models performance and vice-versa. Based on this assumption, we leverage the varying information content and cost of regions within an unlabeled pool of images in order to query the highest density samples to be passed to a human oracle for labeling. We aim to maximize a trade-off, such that for the minimal cost we achieve maximal performance. Various definitions of regions exist. In this work, we only consider regions of quadratic shape and investigate the impact on the employed model's performance regarding their varying sizes.

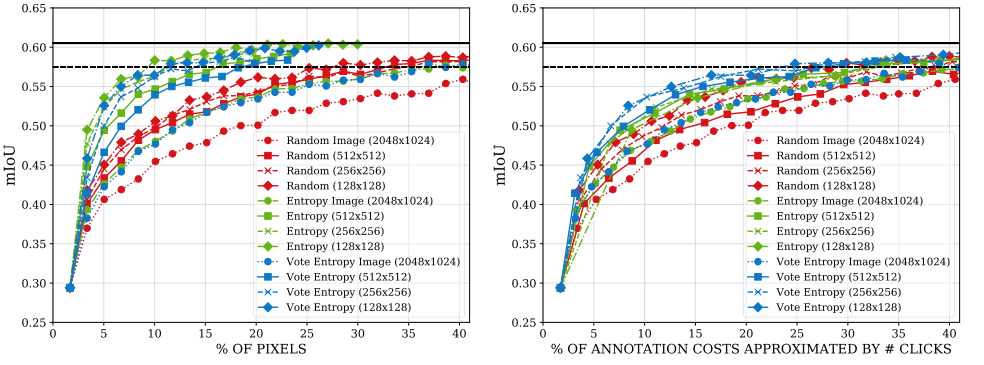
We utilize a sliding-window approach for selecting the most informative regions from within the acquired *information maps* computed for each individual image of the unlabeled pool. For a pre-specified window size we proceed as follows: At each sliding-window location (u,v), we accumulate all the values of our *information map* encompassing the dimensions of the window and store this density in a matrix denoted as *region information map* having the same spatial dimensionality as the considered image. We proceed similarly to generate *region cost maps* given the estimated *cost maps*.

We linearly scale *region information maps* and *region cost maps* w.r.t. the whole dataset, such that all values are in $[0,1]$. We then fuse corresponding region maps using one of the following fusion functions. The three simple fusion functions we have evaluated are denoted in (3), (4), and (5) with the *region information map* I and the *region cost map* C. The parameter α in (5) allows to set a trade-off for linearly interpolating between both region maps. An example of a resulting *fused region map* is depicted in Fig.3(d).

$$g_1 = \frac{I}{1+C} \quad (3) \qquad g_2 = (1-C)\cdot I \quad (4) \qquad g_3 = I\cdot\alpha + (1-C)\cdot(1-\alpha) \quad (5)$$

After fusing the region information and the region cost map pairs for all images in the current unlabeled pool, we perform non-maximum-suppression to retrieve fixed-size region candidates for each individual image. Regarding non-maximum-suppression we always favor higher scoring regions regarding its computed information/cost trade-off while not allowing any overlap until maximum coverage. We store the region candidates (Fig.3(e)) for each individual image of the unlabeled pool within a region proposal pool. Note that the region candidates are not allowed to overlap in between a round, since in an asynchronous annotation mode we do not want to assign the same pixels to be labeled to different annotators.

3) Acquisition From the region proposal pool we extract as many top scoring regions as would correspond to extracting m images out of a pool of equally sized images regarding their amount of pixels for a fair comparison to the image-based acquisition of labels (Fig.3(f)). Instead of employing a real annotator for evaluating our method we utilize a robot user as our oracle. Whenever annotations are being requested, the robot user uses the ground truth annotation of the considered training set. We then update the labeled and unlabeled pool and learn our *semantic segmentation model* and *cost model* from scratch.



(a) (b)

Figure 4: AL curves showing the relationship between pixels and annotation costs approximated by the number of clicks regarding different acquisition functions. The solid black line shows the mIoU achieved by training the model on the whole training set of Cityscapes. The dashed black line marks 95% of the performance achieved by this model. a) Resulting mIoU as function over the amount of labeled pixels queried from an annotator. b) Same obtained results but plotted as a function over the annotation effort measured by the number of clicks.

4 Results

All processed experiments presented in this work are repeated five times and we report the average *mean Intersection over Union* (mIoU) calculated on the validation dataset of Cityscapes after training convergences. We claim convergence whenever a model's mIoU on the validation dataset does not increase within ten epochs. The seed set is initialized to $n = 50$ fully annotated images randomly selected from the unlabeled pool.

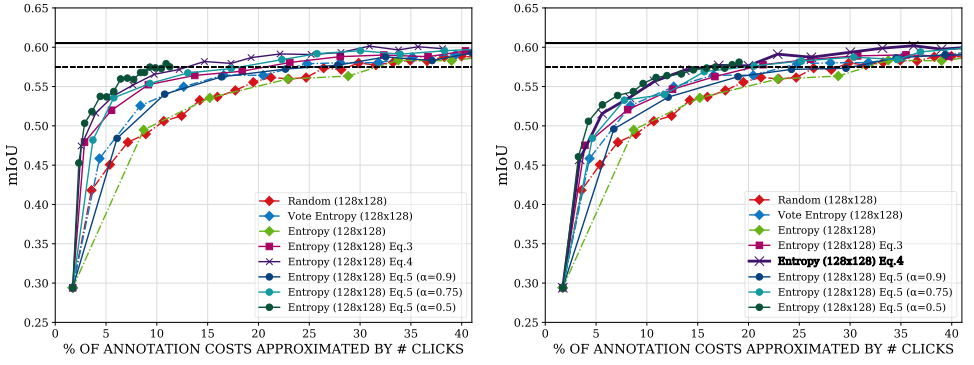
Our experiments are structured as follows: First, we explore the impact of varying region sizes on the models performance w.r.t. the number of queried pixels. Secondly, we show how the results relate w.r.t the labeling costs approximated by the number of clicks. Thirdly, we provide evidence that knowledge about costs can be utilized to further reduce labeling efforts.

Finally, we demonstrate that knowledge about costs can be inferred from a learned CNN, regressing spatial information about costs.

In our first experiment, we are querying the $m = 50$ top scoring images maximizing the considered information measures only, exactly as suggested in [14]. We do not perform any special treatment of semantic boundaries.

In Fig.4(a) we plot the obtained results regarding the percentage of labeled pixels relative to all labels present in the training dataset of Cityscapes against the achieved mIoU of the trained CNN. All considered acquisition functions show better results than random sampling. After 21 acquisition steps corresponding to 35.29% of queried labels by using entropy sampling, we achieve an mIoU of 0.575 which corresponds to 95% of the performance as compared to the obtained result of 0.605, when training on the full training set of Cityscapes. We will refer to the former performance measure as $p95$ and to the latter as $p100$.

In our second experiment, we are evaluating the region-based acquisition of labels and sample 512×512-sized most promising regions out of the entire unlabeled pool. Despite the very large region size we observe a significant improvement for all evaluated information



(a) (b)

Figure 5: In this plot we show the results achieved by our region-based acquisition function when optimizing towards both; minimal costs and high information using entropy sampling. a) Selecting regions by using the ground truth clicks, in order to show how our method would perform if the utilized cost prediction network would always perform perfectly b) Selection of regions by entropy and minimization of estimated costs.

measures just caused by the spatial exploitation of the unlabeled set. We are achieving $p95 = 16.74\%$ by using entropy sampling. The region-based random sampling approach also shows better results compared to sampling whole images randomly, which we argue must be due to the increase in data variability introduced by sampling regions instead of entire images. We then proceed to investigate smaller region sizes, concretely 256×256 and 128×128 and observe the performance to increase. We argue, that this is because smaller region sizes allow querying higher density regions.

An even more interesting result is found when one compares mIoU vs. the effort measured by the number of clicks relative to the total number of clicks which were executed to annotate the whole training dataset of Cityscapes. Similarly to $p95$, $c95$ will denote the performance index for achieving 95% of the performance related to the amount of clicks relative to $c100$ standing for the performance which was achieved with all polygon base points of the training set. Note that $p100 = c100$ when training on whole images. During evaluation we also count every additional click that might occur on the region borders for a fair comparison, such that theoretically the value of $c100$ could get bigger than $p100$ when sampling regions. The results indicate that highly informative data is also costly to label (Fig.4(b)). For sampling whole images based on the entropy information measure, where $p95 = 35.29\%$ we see that this corresponds to $c95 = 39.2\%$. Furthermore the gain in using a sophisticated information measure compared to random sampling regarding the effort is much smaller when sampling regions than initially indicated when comparing against the amount of pixels. This can be clearly seen by comparing Fig.4(a) with Fig.4(b). For example, for 128×128 regions maximizing the acquisition function based on the entropy information measure we achieved $p95 = 10.01\%$ compared to $c95 = 33.76\%$. The queried 10.01% of labels thus require an annotation effort of 33.76%. We report all results in A.2. In terms of the best performing 128×128 setting the results show worse numbers for highly informative region sampling using the entropy information measure compared to the random selection of non-overlapping regions. We also observe entropy sampling to prefer more costly regions than vote entropy

sampling, which is slightly better then random.

In order to further reduce labeling effort while looking for highly informative regions we now leave the region size fixed to the best performing setting regarding our previous experiments (128×128). Equations (3), (4) and (5) are evaluated with different $\alpha \in \{0.5, 0.75, 0.9\}$ in order to empirically determine a good information/cost trade-off. We first establish an upper bound for the utilized information measures fused with optimal cost estimates by using the actual ground truth data about clicks. We observe that entropy sampling fused with the information about true costs to achieve better results than the more powerful vote entropy information measure alone (Fig.5(a)). We can see our method to achieve $c95 = 14.68\%$ using fast to compute entropy sampling assuming that the cost prediction network would always decide correctly.

We now utilize the *cost model* which is trained on the oracle feedback acquired at previous acquisition steps. With our multiplicative fusion approach we reach a performance of $c95 = 17.07\%$ (Fig.5(b)).

5 Conclusion

We have proposed a novel method for cost effective active learning for semantic segmentation tailored to fully convolutional neural networks. We have demonstrated our framework's performance on Cityscapes, a highly diverse high definition dataset consisting of images of urban scenes captured in the wild. We show that combining information content and cost estimates is a powerful approach for cost-effectively building new training datasets from scratch. With only 17% of the effort measured by the amount of clicks which were executed for annotating the Cityscapes training set, we are able to achieve 95% of the full training set's performance.

We leave the question on how the performance of *CEREALS* scales to other network architectures with varying representational power for further research, due to the high computational demands of such an evaluation. Furthermore, we want to encourage the community to provide ground truth information about human annotation costs of upcoming and, if available, already existing manually labeled computer vision datasets. This will help further research in cost-effectively learning high performance models in data-hungry deep learning era.

6 Acknowledgements

We thank the Center for Information Services and High Performance Computing (ZIH) at TU Dresden for generous allocations of computer time. Furthermore we thank our colleagues, Lucas Rego Drumond[1], Thomas Wenzel[1], Dimitrios Bariamis[1], Uwe Brosch[1], Masato Takami[1], Alexander Lengsfeld[1], Jens Mehnert and Volker Fischer for helpful discussions. We also thank the anonymous reviewers for their valuable comments.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars M. Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision : Efficient data generation for urban driving scenes. *CoRR*, abs/1708.01566, 2017. URL http://arxiv.org/abs/1708.01566.

[2] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4485–4493, 2017. doi: 10.1109/CVPR.2017.477. URL https://doi.org/10.1109/CVPR.2017.477.

[3] Liang-Chieh Chen, Sanja Fidler, and Raquel Urtasun. Beat the mturkers: Automatic image labeling from weak 3d supervision. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3198–3205, 2014. doi: 10.1109/CVPR.2014.409. URL https://doi.org/10.1109/CVPR.2014.409.

[4] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. In *Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]*, pages 705–712, 1994. URL http://papers.nips.cc/paper/1011-active-learning-with-statistical-models.

[5] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223, 2016. doi: 10.1109/CVPR.2016.350. URL https://doi.org/10.1109/CVPR.2016.350.

[6] Ido Dagan and Sean P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*, pages 150–157, 1995. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.30.6148.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009. doi: 10.1109/CVPRW.2009.5206848. URL https://doi.org/10.1109/CVPRW.2009.5206848.

[8] Alireza Fathi, Maria-Florina Balcan, Xiaofeng Ren, and James M. Rehg. Combining self training and active learning for video segmentation. In *British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings*, pages 1–11, 2011. doi: 10.5244/C.25.78. URL https://doi.org/10.5244/C.25.78.

[9] Jie Feng, Brian L. Price, Scott Cohen, and Shih-Fu Chang. Interactive segmentation on RGBD images via cue selection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 156–164, 2016. doi: 10.1109/CVPR.2016.24. URL https://doi.org/10.1109/CVPR.2016.24.

[10] Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Information, prediction, and query by committee. In *Advances in Neural Information Processing Systems 5, [NIPS Conference, Denver, Colorado, USA, November 30 - December 3, 1992]*, pages 483–490, 1992. URL http://papers.nips.cc/paper/622-information-prediction-and-query-by-committee.

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1050–1059, 2016. URL http://jmlr.org/proceedings/papers/v48/gal16.html.

[12] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1183–1192, 2017. URL http://proceedings.mlr.press/v70/gal17a.html.

[13] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587, 2014. doi: 10.1109/CVPR.2014.81. URL https://doi.org/10.1109/CVPR.2014.81.

[14] Marc Gorriz, Axel Carlier, Emmanuel Faure, and Xavier Giró i Nieto. Cost-effective active learning for melanoma segmentation. *CoRR*, abs/1711.09168, 2017. URL http://arxiv.org/abs/1711.09168.

[15] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. Weakly supervised semantic segmentation using web-crawled videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2224–2232, 2017. doi: 10.1109/CVPR.2017.239. URL https://doi.org/10.1109/CVPR.2017.239.

[16] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. URL http://arxiv.org/abs/1704.04861.

[17] Mi-Young Huh, Pulkit Agrawal, and Alexei A. Efros. What makes imagenet good for transfer learning? *CoRR*, abs/1608.08614, 2016. URL http://arxiv.org/abs/1608.08614.

[18] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *CoRR*, abs/1802.07934, 2018. URL http://arxiv.org/abs/1802.07934.

[19] Suyog Dutt Jain and Kristen Grauman. Active image segmentation propagation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2864–2873, 2016. doi: 10.1109/CVPR.2016.313. URL https://doi.org/10.1109/CVPR.2016.313.

[20] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1665–1674, 2017. doi: 10.1109/CVPR.2017.181. URL https://doi.org/10.1109/CVPR.2017.181.

[21] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: From edges to instances with multicut. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7322–7331, 2017. doi: 10.1109/CVPR.2017.774. URL https://doi.org/10.1109/CVPR.2017.774.

[22] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 695–711, 2016. doi: 10.1007/978-3-319-46493-0_42. URL https://doi.org/10.1007/978-3-319-46493-0_42.

[23] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Introducing geometry in active learning for image segmentation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2974–2982, 2015. doi: 10.1109/ICCV.2015.340. URL https://doi.org/10.1109/ICCV.2015.340.

[24] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. Crowdsourcing in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 10(3):177–243, 2016. doi: 10.1561/0600000071. URL https://doi.org/10.1561/0600000071.

[25] Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1915–1924, 2017. URL http://proceedings.mlr.press/v70/krishnamurthy17a.html.

[26] David D. Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. *SIGIR Forum*, 29(2):13–19, 1995. doi: 10.1145/219587.219592. URL http://doi.acm.org/10.1145/219587.219592.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965. URL https://doi.org/10.1109/CVPR.2015.7298965.

[28] Yao Lu, Xue Bai, Linda G. Shapiro, and Jue Wang. Coherent parametric contours for interactive video object segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 642–650, 2016. doi: 10.1109/CVPR.2016.76. URL https://doi.org/10.1109/CVPR.2016.76.

[29] Zhiwu Lu, Zhenyong Fu, Tao Xiang, Peng Han, Liwei Wang, and Xin Gao. Learning from weak and noisy labels for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(3):486–500, 2017. doi: 10.1109/TPAMI.2016.2552172. URL https://doi.org/10.1109/TPAMI.2016.2552172.

[30] Agata Mosinska-Domanska, Raphael Sznitman, Przemyslaw Glowacki, and Pascal Fua. Active learning for delineation of curvilinear structures. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5231–5239, 2016. doi: 10.1109/CVPR.2016.565. URL https://doi.org/10.1109/CVPR.2016.565.

[31] Yair Movshovitz-Attias, Takeo Kanade, and Yaser Sheikh. How useful is photo-realistic rendering for visual learning? In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, pages 202–217, 2016. doi: 10.1007/978-3-319-49409-8_18. URL https://doi.org/10.1007/978-3-319-49409-8_18.

[32] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5000–5009, 2017. doi: 10.1109/ICCV.2017.534. URL https://doi.org/10.1109/ICCV.2017.534.

[33] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5038–5047, 2017. doi: 10.1109/CVPR.2017.535. URL https://doi.org/10.1109/CVPR.2017.535.

[34] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1717–1724, 2014. doi: 10.1109/CVPR.2014.222. URL https://doi.org/10.1109/CVPR.2014.222.

[35] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1742–1750, 2015. doi: 10.1109/ICCV.2015.203. URL https://doi.org/10.1109/ICCV.2015.203.

[36] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, pages 102–118, 2016. doi: 10.1007/978-3-319-46475-6_7. URL https://doi.org/10.1007/978-3-319-46475-6_7.

[37] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008. doi: 10.1007/s11263-007-0090-8. URL https://doi.org/10.1007/s11263-007-0090-8.

[38] Fatemehsadat Saleh, Mohammad Sadegh Ali Akbarian, Mathieu Salzmann, Lars Petersson, Stephen Gould, and Jose M. Alvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 413–432, 2016. doi: 10.1007/978-3-319-46484-8_25. URL https://doi.org/10.1007/978-3-319-46484-8_25.

[39] Fatemehsadat Saleh, Mohammad Sadegh Ali Akbarian, Mathieu Salzmann, Lars Petersson, and Jose M. Alvarez. Bringing background into the foreground: Making all classes equal in weakly-supervised video semantic segmentation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2125–2135, 2017. doi: 10.1109/ICCV.2017.232. URL https://doi.org/10.1109/ICCV.2017.232.

[40] Ozan Sener and Silvio Savarese. A geometric approach to active learning for convolutional neural networks. *CoRR*, abs/1708.00489, 2017. URL http://arxiv.org/abs/1708.00489.

[41] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

[42] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. URL http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.

[43] Burr Settles, Mark Craven, and Lewis Friedl. Active learning with real annotation costs. In *In Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, pages 1–10, 2008. URL http://burrsettles.com/pub/settles.nips08ws.pdf.

[44] Alireza Shafaei, James J. Little, and Mark Schmidt. Play and learn: Using video games to train computer vision models. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016. URL http://www.bmva.org/bmvc/2016/papers/paper026/index.html.

[45] Claude E. Shannon. A mathematical theory of communication. *Mobile Computing and Communications Review*, 5(1):3–55, 2001. doi: 10.1145/584091.584093. URL http://doi.acm.org/10.1145/584091.584093.

[46] Wataru Shimoda and Keiji Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 218–234, 2016. doi: 10.1007/978-3-319-46493-0_14. URL https://doi.org/10.1007/978-3-319-46493-0_14.

[47] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5689–5697, 2017. doi: 10.1109/ICCV.2017.606. URL https://doi.org/10.1109/ICCV.2017.606.

[48] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 843–852, 2017. doi: 10.1109/ICCV.2017.97. URL https://doi.org/10.1109/ICCV.2017.97.

[49] Simon Tong and Daphne Koller. Active learning for parameter estimation in bayesian networks. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 647–653, 2000. URL http://papers.nips.cc/paper/1795-active-learning-for-parameter-estimation-in-bayesian-networks.

[50] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2001. URL http://www.ai.mit.edu/projects/jmlr/papers/volume2/tong01a/abstract.html.

[51] Alexander Vezhnevets, Joachim M. Buhmann, and Vittorio Ferrari. Active learning for semantic segmentation with expected change. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3162–3169, 2012. doi: 10.1109/CVPR.2012.6248050. URL https://doi.org/10.1109/CVPR.2012.6248050.

[52] Sudheendra Vijayanarasimhan and Kristen Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 2262–2269, 2009. doi: 10.1109/CVPRW.2009.5206705. URL https://doi.org/10.1109/CVPRW.2009.5206705.

[53] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Trans. Circuits Syst. Video Techn.*, 27(12): 2591–2600, 2017. doi: 10.1109/TCSVT.2016.2589879. URL https://doi.org/10.1109/TCSVT.2016.2589879.

[54] Jun Xie, Martin Kiefel, Ming-Ting Sun, and Andreas Geiger. Semantic instance annotation of street scenes by 3d to 2d label transfer. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3688–3697, 2016. doi: 10.1109/CVPR.2016.401. URL https://doi.org/10.1109/CVPR.2016.401.

[55] Ning Xu, Brian L. Price, Scott Cohen, Jimei Yang, and Thomas S. Huang. Deep interactive object selection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 373–381, 2016. doi: 10.1109/CVPR.2016.47. URL https://doi.org/10.1109/CVPR.2016.47.

[56] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 - 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III*, pages 399–407, 2017. doi: 10.1007/978-3-319-66179-7_46. URL https://doi.org/10.1007/978-3-319-66179-7_46.

[57] Wei Zhang, Sheng Zeng, Dequan Wang, and Xiangyang Xue. Weakly supervised semantic segmentation for social images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2718–2726, 2015. doi: 10.1109/CVPR.2015.7298888. URL https://doi.org/10.1109/CVPR.2015.7298888.

[58] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas A. Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5057–5065, 2017. doi: 10.1109/CVPR.2017.537. URL https://doi.org/10.1109/CVPR.2017.537.

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5122–5130, 2017. doi: 10.1109/CVPR.2017.544. URL https://doi.org/10.1109/CVPR.2017.544.