Active Learning for Semantic Segmentation with Area Disagreement

Flavius Cristian Fetean
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
feteanflavius@gmail.com

Razvan Itu
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
razvan.itu@cs.utcluj.ro

***Abstract*—We propose Area Disagreement, a novel uncertainty estimation method for Active Learning in the context of Semantic Segmentation, which capitalizes on the unique characteristics of the task. It relies on the assumption that slight alterations to the learner model's parameters should not produce significant differences in the output, specifically in terms of predicted shapes and objects given the same input image. While a small amount of contradiction is natural, larger inconsistencies suggest that the model's internal representation of the world is perplexed by those images, thus making them valuable for further training. Our uncertainty estimation method prioritizes images with a smaller Dice Coefficient relative to the average prediction, based on multiple Monte-Carlo Dropout inferences. Utilizing this approach, we outperformed baseline methods by a wide margin on the Cityscapes dataset, achieving 95% of the full-scale training performance using only 36% of the dataset and 97.5% of the full-scale training performance using 47% of the dataset.**

I. Introduction

In the past decade, the field of Deep Learning has been supercharged by the abundance of data and the availability of powerful computing hardware such as GPUs. It has managed to revolutionize multiple areas, most notably Computer Vision. However, most approaches require huge amounts of manually annotated data, which has become a significant bottleneck in the development of large-scale Neural Network models, especially for fine-grained tasks like Semantic Segmentation, where pixel-wise labels must be obtained for every ground truth image. The annotation process for Semantic Segmentation typically involves delimiting each object in an image using a polygon, a process that may take up to 1.5 hours for a single image, thereby increasing the need to minimize the number of images requiring manual processing.

Active Learning (AL) is a framework that aims to address this challenge by proposing mechanisms to select only the most promising images for annotation, usually by evaluating each sample from an unlabeled pool based on the model output and a query method. It has been successfully applied in classification [1]–[5], segmentation [6]–[14], object detection [15]–[17] and task-agnostic settings [18], [19], reaching higher performance levels than would otherwise be possible with the same number of randomly chosen images.

979-8-3315-3997-9/24/$31.00 ©2024 IEEE

Most of the recent papers on AL for Semantic Segmentation focus on leveraging Semi-Supervised Learning, selecting only regions of images, or coping with weakly labeled data. While these approaches do decrease the necessary effort for annotation by orders of magnitude, they still heavily rely on uncertainty estimation functions.

The purpose of this work is to introduce a novel uncertainty estimation method tailored for Semantic Segmentation, taking advantage of the particularities of this task, mainly the spatial dependencies between pixels. More specifically, by leveraging the concept of Bayesian Neural Networks implemented using Monte-Carlo Dropout [20], multiple segmentation masks are predicted for each image in the unlabeled pool. The Dice Coefficient is computed between the different predictions and their average, and the images causing the lowest Dice Coefficient are deemed to be the most informative.

The contribution of this paper is the creation of a superior uncertainty estimation function tailored for Semantic Segmentation, which may be used either in isolation, as a query method, or in conjunction with any other method that needs an uncertainty estimation baseline.

II. Related Work

A. Active Learning

The first comprehensive survey on active learning (AL) was conducted by Burr Settles [21]. In his work, he identified the three main scenarios of AL: membership query synthesis, stream-based selective sampling, and pool-based sampling. Among these, the most widely used is the latter, which will be the focus of this paper. The pool-based sampling method treats the AL process as a loop, where at every iteration, a large pool of real, unlabeled data is queried for the most useful samples to be manually annotated and added to a smaller labeled set, which will constitute the training set for the next iteration of the learner model. The process is illustrated in Fig. 1.

Multiple approaches for assessing the usefulness of samples have been studied over the years, with the most prominent being Uncertainty Sampling, where the model's predictions on the unlabeled images are evaluated for uncertainty using various functions, such as Entropy [22] and Margin [23]. Other approaches include Query-By-Committee [24], Expected Model Change [25], and Core-Set [4]. Additionally, Yarin Gal et al.

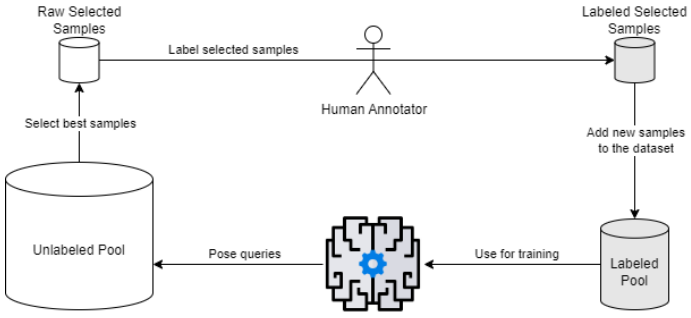


Fig. 1: Pool-Based Active Learning Process.

introduced the concept of using Bayesian Neural Networks to assess uncertainty [3] by approximating Bayesian Inference using the Monte-Carlo (MC) Dropout method [20]. This was confirmed by BALD [5] and its successor BatchBALD [2], which compute the Mutual Information between individual and average predictions on both sample and batch levels, respectively. Furthermore, the Bayesian framework, as discussed in a survey conducted by Di Fiore et al. [26], is particularly well-suited to address some challenges explored by Active Learning. Consequently, a synergy has recently begun to emerge between the two domains.

More modern approaches have emerged, aiming to integrate advances from other fields of Deep Learning, such as Autoencoders [18], [19], Reinforcement Learning [6], [27], GANs [28], [29], and NAS [30]. Unlike classical AL methods, these contemporary approaches typically use additional networks to estimate image usefulness and are generally task-agnostic, making it relatively straightforward to adapt them to any new Deep Learning task, including Semantic Segmentation.

B. Active Learning for Semantic Segmentation

For the Semantic Segmentation task, the most prevalent approaches used Semi-Supervised Learning [10], [12] and Region-Based querying [9] to maximize the utility of unlabeled data and reduce the labeling effort as much as possible by not requiring whole images to be annotated. Taking this effort reduction further, the work of Hwang et al. [8] aimed to reduce the labeling effort to just a few clicks. Their method involved the model requesting all existing classes in a small region of interest, and adapting the learning procedure to support such a multi-class labeling scheme. Here, the annotator's effort is reduced to simply selecting the visible classes from a list. Another approach to minimizing annotator workload is PixelPick [13], where the authors propose a framework where only individual pixels of interest are queried by the model, and annotators only need to pick the class of each pixel.

Two other methods that use approaches similar to ours are EquAL [7] and ViewAL [14]. EquAL suggests that ensuring self-consistency between model predictions on the same input image, with equivariant transformations applied, helps training by adding this enforcement as a loss component. However, the consistency was not applied as a criterion for sample selection. ViewAL is tailored for 3D environments, where there are multiple shots of the same object from different viewpoints. By analyzing predictions from different viewpoints, the authors could identify regions (projected into 3D) causing the most contradictions between viewpoints.

In a manner similar to ours, Dechesne et al. [31] applied the Bayesian framework, specifically the Monte-Carlo Dropout(MCD) method, in order to train a semantic segmentation model end-to-end, achieving promising results. In addition, they utilized the multiple predictions generated by the MCD process to assess uncertainty by computing pixel-wise entropy. They did not use it, however, as an active learning goal.

Although targeted at Semantic Segmentation, existing methods mainly aim to enhance the process by leveraging the unlabeled set or reducing the labeling effort per data point, but do not focus on the acquisition itself and the means to compute the most informative data points. Moreover, most of them select their data points using simple Entropy or slightly modified variants of Entropy. Therefore, we construct and evaluate our method without such additional augmentations, positioning it as an alternative to classical query functions like Entropy and Margin.

III. Method

In order to take advantage of the properties of the Semantic Segmentation's output (e.g.: locality, shape, size), we shall look at the way IoU and Dice Coefficient use them. Instead of averaging pixel-wise difference between the network output and ground truth, these methods compute the ratio of the intersected areas (how much the model got right) over the union of areas (how much the model missed relative to the size of the objects). This formulation pushes the model to learn the shapes of the objects irrespective to their sizes, as the applied penalty is the proportion of the coverage, as opposed to the pixelwise penalty which would be strongly biased towards bigger objects. But in order to compute an IoU or a Dice Coefficient, a Term of Comparison, in this case, the Ground Truth, is needed, and when computing uncertainty, nothing else but the output of the model is available, so such a comparison cannot be done and the uncertainty has to be assessed within the output itself.

The Bayesian Framework enables predicting multiple slightly modified variants by the same model, by tweaking the parameters. In classifier based AL, an entropy would be measured on these inferences afterwards, or disagreement by computing mutual information. For semantic segmentation, it means that for a single image we could have multiple masks for the same scene, but with the objects within them having slight differences in properties (e.g.: slightly dislocated, marginally smaller or bigger) compared to the homologous objects in the other predictions. This enables us to observe exactly which objects and properties the model tends to confuse, by what degree, even if the true composition of the image is not known yet. Therefore, we borrow the concept of "Disagreement", as we want to compute how much the

predictions contradict with each other, and we apply it to "Areas" resulted by the Semantic Segmentation process, though not exactly in the same way described in the original BALD method, yet still similar. We aim to measure the spread in terms of Dice Coefficient of the Bayesian inference given by the model, as the Dice Coefficient is able to capture the rich information contained in the surfaces described by the segmentation outputs. This implies computing the average prediction of the MC Dropout ensemble, and averaging the Dice Coefficients between each of the individual predictions against the average one, essentially capturing the variance of the predictions with respect to their mean. While the exact methodology for each step will be detailed further in the section, the intuition behind the Area Disagreement concept may be visualized in Fig. 2.

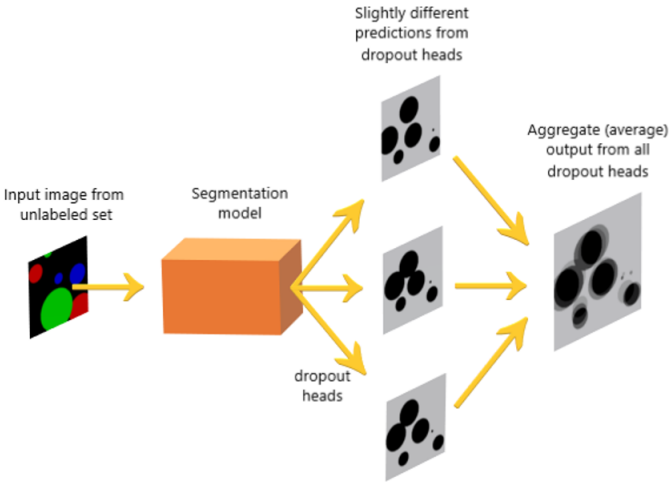


Fig. 2: Illustration of the Disagreement using 2 classes (foreground, background) for better visibility. Each iteration of the MC Dropout process inferences on the same input image and outputs a similar segmentation mask. When these masks are overlaid, a shady output mask results which is the actual prediction of the overall network.

A. Uncertainty Estimation

As presented in Fig 1, the distinctive part of the Pool-Based Active Learning (AL) process is the querying step, where an acquisition function scores the unlabeled data samples by estimating the model's uncertainty for each one. The focus of this paper is to describe the methodology behind Area Disagreement, a superior uncertainty estimation method for the AL process in semantic segmentation.

Let $f_i(x)$ be the prediction of the model in the i-th MC Dropout iteration for a given sample x. As mentioned earlier in the section, to apply the Dice Coefficient, a one-hot encoded tensor is needed to act as the ground truth. In our case, we will construct $f_{agg}(x)$ as the one-hot encoded final prediction of the network by averaging all the MC Dropout predictions together and one-hot encoding the result.

Having the final aggregate prediction of the network, the computation of the Dice Coefficient for each MC iteration is straightforward:

$$Dice(f_i(x), f_{agg}(x)) = 2 \times \frac{\sum(f_i(x) \cdot f_{agg}(x))}{\sum f_i(x) + \sum f_{agg}(x)} \quad (1)$$

Finally, the Area Disagreement of the sample will be the average of the Dice Coefficients between every MC iteration's prediction and the aggregate prediction:

$$AD(x) = -\frac{1}{N}\sum_{i=1}^{N} Dice(f_i(x), f_{agg}(x)) \quad (2)$$

where $AD(x)$ is the Area Disagreement of the model for input sample x, and N the number of MC iterations. However, the Dice Coefficient value is smaller for images where the disagreement is higher, meaning the actual uncertainty of the model for a given sample is inversely proportional to its average Dice Coefficients. Therefore, a minus sign is needed at the beginning to make AD proportional to the true uncertainty for the sample.

IV. Experimental Setup

A. Dataset

Cityscapes [32] is a comprehensive large-scale dataset used for semantic urban scene understanding. It comprises 5,000 high-quality annotated images (1024x2048 pixels) captured in 50 different cities under varying weather conditions and seasons, providing diverse urban street scenes. Each image is finely annotated with pixel-level labels for 30 different classes, encompassing common urban objects such as pedestrians, vehicles, and road markings. In our experiments 1500 of the images were randomly held out for ensuring fair testing among the different acquisition methods and 3500 were used for training. Furthermore, the number of classes was reduced to 8 by mapping similar classes to a single centroid.

B. Comparison Baseline

The overwhelming majority of the recent AL studies for Semantic Segmentation experiment with augmenting the classic AL framework and use either Entropy or Margin for uncertainty estimation when selecting images. For this reason, the baseline chosen for our experiments consists in Random Acquisition, Entropy Acquisition and Margin Acquisition. As it is the case for the majority of the related papers, our implementation consists in averaging the pixel-wise Entropy or Margin values for the Segmentation output of a given image.

C. Implementation Details

For each experiment, we randomly select 10% (350 images) of the dataset as the initial labeled set. In each iteration, we hold out 20% of the available labeled images for validation. We then train a reduced U-Net model (as described in [33]) from scratch on the remaining labeled images. The model consists of four downsampling blocks, starting with 32 filters and doubling the number of filters at each subsequent block (64, 128, 256). This is followed by a bottleneck layer with 512

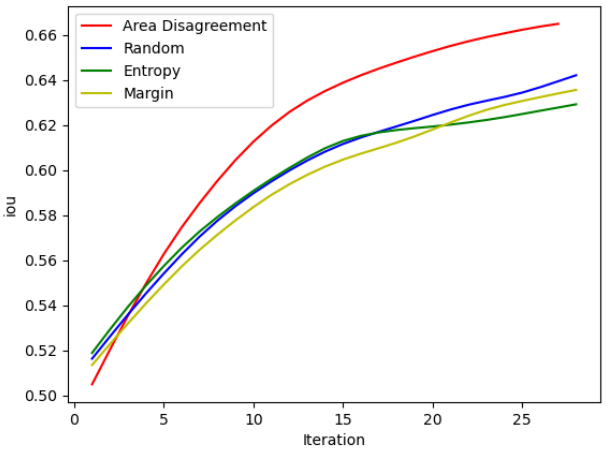


Fig. 3: Comparative performance analysis. The plot compares the four active learning strategies—Area Disagreement, Random, Entropy, and Margin—in terms of Mean Intersection over Union (IoU) performance across 30 iterations. The x-axis represents the iteration number of the active learning process, and the y-axis corresponds to the test performance (IoU).

filters, and a symmetric upsampling path. At acquisition time, we introduce an additional dropout layer with 20% dropout rate after the last parameterized layer of the network. We use this model to predict the uncertainty of each unlabeled image and select the 50 images with the highest predicted uncertainty computed on the basis of 10 MCD predictions. These images, along with their ground truth annotations, are added to the labeled set for the next iteration. We perform 27 iterations, adding 50 images in each iteration until no more than 50% of the total dataset (1750 images) is labeled. We repeat each experiment 10 times to ensure statistical stability.

During training, we employ data augmentation (random cropping, flipping, and brightness adjustment) to mitigate overfitting due to the limited number of images. We train with a batch size of 4 for a maximum of 50 epochs but utilize early stopping, halting training if the validation loss does not improve for 5 epochs, to further prevent overfitting.

V. Results

A. Performance Evaluation

In Fig. 3 we analyze the comparison between the performance in mIoU of Area Disagreement and the baseline methods at each AL step. It may be seen that after 26 iterations, the Area Disagreement method leads to a performance of the model of 0.67 IoU. This marks an important increase of 0.023 over the Random Selection and over 0.032 over the Entropy and Margin methods, or a 3.4% increase in performance. Furthermore, considering that a full training yields 0.685 IoU [1], it means that the introduced method only lags behind the full use of the dataset by 0.015 with only half of it. In comparative terms, as opposed to the Random Acquisition, which trails the full training by 0.038, Area Disagreement closes 60% of the performance gap to full-scale training. However, the presented figure also suggests a decrease of the advantage in performance which seems to follow after half of the dataset has already been acquired, as the gain naturally starts to plateau, possibly because the most informative samples have already been selected and the remaining ones do not provide much new knowledge to the model.

[1]The reported figure was obtained by the described U-net model in our training and augmentation setting. It is worth noting that the state-of-the-art at the time of writing is 0.864, achieved by VLTSeg [34]

TABLE I: Number of iterations needed (percentage of dataset) for each method to pass a certain IoU threshold, given selection sizes of 50 images per iteration.

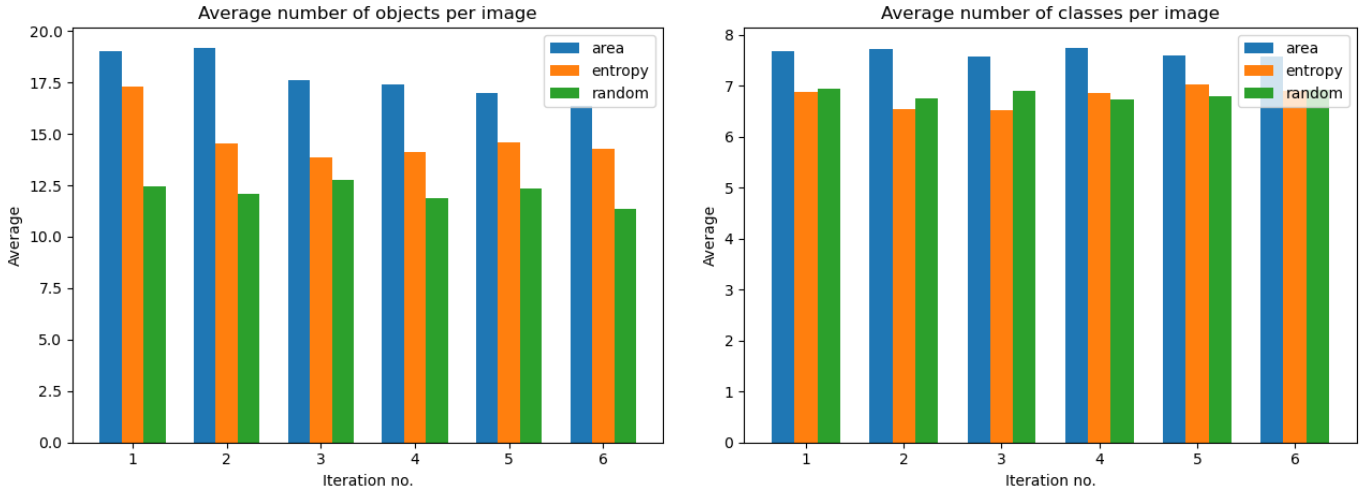
IoU	Area Disagreement	Random	Entropy	Margin
0.6	**9 (23%)**	12 (27%)	11 (26%)	11 (26%)
0.625	**12 (27%)**	20 (37%)	15 (31%)	22 (40%)
0.65	**19 (36%)**	-	-	-
0.67	**26 (47%)**	-	-	-

Another benefit of the Area Disagreement approach is the faster growth in performance, which may also be noticed in Fig. 3. As it is able to more effectively select the most relevant samples, especially in the beginning, it manages to hit milestones in mIoU much faster than the other methods, translating in less images used and fewer AL iterations needed. This behavior is detailed in the Table I. Area Disagreement manages to hit an mIoU of 0.6 with 100-150 less images, 0.625 with 150-500 less images and manages to hit 0.65 (95% of full-scale training) within 36% of the dataset and 0.67 (97.5% of full-scale training) within 47% of the dataset while the other methods do not within the first half.

An additional observation is the underperformance of the Entropy and Margin approaches. As noted in [7], training on a small number of images can lead to weak generalization, contributing to a poor choice of new samples. Thus, the margin and entropy methods likely picked more redundant samples initially due to overfitting, causing each iteration to have weaker generalization than the previous one and generating a compounding underperformance. The relative superiority of the random process, on the other hand, may be because the choice of new samples does not depend on the quality of already chosen samples, thus having no reason to degrade over time. Nevertheless, even in this context, the Area Disagreement method managed to pick more relevant samples and avoid being affected by overfitting, likely due to the high diversity and difficulty of the chosen samples.

B. Selection Interpretation

To continue the investigation into the effectiveness of *Area Disagreement*, it makes sense to also take a deeper look into the properties of the selected images, to understand the ways in which they stand out and enforce the superiority of the method by basing it on observable facts. The composition of images was studied in terms of the number of objects they contain, and the number of classes represented in them. A comparative



(a) Average number of objects

(b) Average number of classes

Fig. 4: Diversity of selected images for Random, Entropy and Area Disagreement methods: Each graph depicts the statistics for the batch chosen in the specified iteration (i.e.: for the 50 images selected in that step), for the first 6 iterations.

plot is illustrated in Figure 4 which measures for the first 6 selection iterations the average number of distinct objects and the average number of distinct classes present throughout the batch of 50 images' masks selected by each method. What is considered a distinct object is an area in the mask containing connected pixels of the same class (i.e.: any two pixels in that area may be connected by an uninterrupted sequence of neighboring pixels), as the annotations of the Cityscapes dataset do not have instance id for different objects. This definition causes the inclusion of object fragments also (the same object disconnected in space will be double counted), or the merging of objects which are overlapping from the point of view of the camera (multiple overlapping objects will be counted as one).

The Area Disagreement method consistently selects more complex scenes, as demonstrated by the increased object and class diversity in the chosen images. Figure 4a shows that selected images contain 50% more objects than those chosen by Random Selection and 10-20% more than those chosen by Entropy Selection. This exposes the model to more instances of the same class in each batch, allowing for more effective feature learning with fewer images. Furthermore, Figure 4b reveals that our method tends to select images with more classes, averaging nearly 8 classes per image in most iterations. In contrast, the other two methods typically choose images with one fewer class. This difference provides the model with greater exposure to various classes, facilitating better differentiation over time.

VI. Conclusions

In this paper, we introduced Area Disagreement, a novel uncertainty estimation method for Active Learning in Semantic Segmentation. By leveraging the unique characteristics of the task and the concept of Bayesian Neural Networks with Monte Carlo Dropout, our method effectively identifies informative images for annotation. Our experiments on the Cityscapes dataset demonstrate that Area Disagreement significantly outperforms baseline methods, achieving 95% of full-scale training performance with only 36% of the dataset and 97.5% with 47%. These results highlight the potential of our approach to reduce annotation efforts while maintaining high performance in semantic segmentation tasks. Furthermore, due to its strong performance and general applicability, the method is well-suited to be used as a base acquisition function for more complex methods which augment the AL process. Future work should explore the integration of Area Disagreement with other Active Learning strategies, its application to different datasets and segmentation models, and its potential adaptation for object detection tasks.

References

[1] X. Du, D. Zhong, and H. Shao, "Building an active palmprint recognition system," in *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*. IEEE, 2019, pp. 1685–1689.

[2] A. Kirsch, J. van Amersfoort, and Y. Gal, "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 7024–7035.

[3] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with bernoulli approximate variational inference," *CoRR*, vol. abs/1506.02158, 2015. [Online]. Available: http://arxiv.org/abs/1506.02158

[4] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=H1aIuk-RW

[5] N. Houlsby, F. Huszar, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," *CoRR*, vol. abs/1112.5745, 2011. [Online]. Available: http://arxiv.org/abs/1112.5745

[6] A. Casanova, P. O. Pinheiro, N. Rostamzadeh, and C. J. Pal, "Reinforced active learning for image segmentation," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SkgC6TNFvr

[7] S. A. Golestaneh and K. Kitani, "Importance of self-consistency in active learning for semantic segmentation," in *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. [Online]. Available: https://www.bmvc2020-conference.com/assets/papers/0010.pdf

[8] S. Hwang, S. Lee, H. Kim, M. Oh, J. Ok, and S. Kwak, "Active learning for semantic segmentation with multi-class label query," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[9] R. Mackowiak, P. Lenz, O. Ghori, F. Diego, O. Lange, and C. Rother, "CEREALS - cost-effective region-based active learning for semantic segmentation," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 121. [Online]. Available: http://bmvc2018.org/contents/papers/0437.pdf

[10] S. Mittal, M. Tatarchenko, Ö. Çiçek, and T. Brox, "Parting with illusions about deep active learning," *CoRR*, vol. abs/1912.05361, 2019. [Online]. Available: http://arxiv.org/abs/1912.05361

[11] S. Mittal, J. Niemeijer, J. P. Schäfer, and T. Brox, "Best practices in active learning for semantic segmentation," in *Pattern Recognition - 45th DAGM German Conference, DAGM GCPR 2023, Heidelberg, Germany, September 19-22, 2023, Proceedings*, ser. Lecture Notes in Computer Science, U. Köthe and C. Rother, Eds., vol. 14264. Springer, 2023, pp. 427–442.

[12] A. Rangnekar, C. Kanan, and M. Hoffman, "Semantic segmentation with active semi-supervised learning." Waikoloa, HI, USA: IEEE, 2023, pp. 5955–5966.

[13] G. Shin, W. Xie, and S. Albanie, "All you need are a few pixels: semantic segmentation with pixelpick," in *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*. IEEE, 2021, pp. 1687–1697.

[14] Y. Siddiqui, J. Valentin, and M. Nießner, "Viewal: Active learning with viewpoint entropy for semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 9430–9440.

[15] H. H. Aghdam, A. Gonzalez-Garcia, A. M. López, and J. van de Weijer, "Active learning for deep detection neural networks," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 3671–3679.

[16] S. V. Desai, A. C. Lagandula, W. Guo, S. Ninomiya, and V. N. Balasubramanian, "An adaptive supervision framework for active learning in object detection," in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*. BMVA Press, 2019, p. 230. [Online]. Available: https://bmvc2019.org/wp-content/uploads/papers/0921-paper.pdf

[17] A. Harakeh, M. Smart, and S. L. Waslander, "Bayesod: A bayesian approach for uncertainty estimation in deep object detectors," in *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*. IEEE, 2020, pp. 87–93.

[18] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 5971–5980.

[19] K. Kim, D. Park, K. I. Kim, and S. Y. Chun, "Task-aware variational adversarial active learning." Nashville, TN, USA: IEEE, 2021, pp. 8162–8171.

[20] Y. Gal and Z. 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*, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 1050–1059. [Online]. Available: http://proceedings.mlr.press/v48/gal16.html

[21] B. Settles, "Active learning literature survey," 2009.

[22] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[23] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Advances in Intelligent Data Analysis, 4th International Conference, IDA 2001, Cascais, Portugal, September 13-15, 2001, Proceedings*, ser. Lecture Notes in Computer Science, F. Hoffmann, D. J. Hand, N. M. Adams, D. H. Fisher, and G. Guimarães, Eds., vol. 2189. Springer, 2001, pp. 309–318.

[24] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992, Pittsburgh, PA, USA, July 27-29, 1992*, D. Haussler, Ed. ACM, 1992, pp. 287–294.

[25] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2007, pp. 1289–1296.

[26] F. Di Fiore, M. Nardelli, and L. Mainini, "Active learning and bayesian optimization: A unified perspective to learn with a goal," *Archives of Computational Methods in Engineering*, vol. 31, no. 5, p. 2985–3013, Apr. 2024. [Online]. Available: http://dx.doi.org/10.1007/s11831-024-10064-z

[27] M. Fang, Y. Li, and T. Cohn, "Learning how to active learn: A deep reinforcement learning approach," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, M. Palmer, R. Hwa, and S. Riedel, Eds. Association for Computational Linguistics, 2017, pp. 595–605.

[28] T. Tran, T. Do, I. D. Reid, and G. Carneiro, "Bayesian generative active deep learning," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 6295–6304. [Online]. Available: http://proceedings.mlr.press/v97/tran19a.html

[29] J. Zhu and J. Bento, "Generative adversarial active learning," *CoRR*, vol. abs/1702.07956, 2017. [Online]. Available: http://arxiv.org/abs/1702.07956

[30] Y. Geifman and R. El-Yaniv, "Deep active learning with a neural architecture search," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 5974–5984.

[31] C. Dechesne, P. Lassalle, and S. Lefèvre, "Bayesian deep learning with monte carlo dropout for qualification of semantic segmentation," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 2536–2539.

[32] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. 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*. IEEE Computer Society, 2016, pp. 3213–3223.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., vol. 9351. Springer, 2015, pp. 234–241.

[34] C. Hümmer, M. Schwonberg, L. Zhou, H. Cao, A. Knoll, and H. Gottschalk, "Vltseg: Simple transfer of clip-based vision-language representations for domain generalized semantic segmentation," 2023. [Online]. Available: https://arxiv.org/abs/2312.02021