-Supplementary Material-
Active Learning and Discovery of Object Categories in the Presence of Unnameable Instances

Christoph Käding, Alexander Freytag, Erik Rodner, Paul Bodesheim, Joachim Denzler
Computer Vision Group, Friedrich Schiller University Jena, Germany
{firstname.lastname}@uni-jena.de | www.inf-cv.uni-jena.de

Abstract

The following document gives additional information with respect to the paper "Active Learning and Discovery of Object Categories in the Presence of Unnameable Instances". *Details for five aspects are presented: (i) an additional comparison of loss functions and probability estimates (Sect. S1), (ii) additional evaluations focusing on labeling time rather than on number of requested labels (Sect. S2), (iii) a visualization of requested samples (Sect. S3), (iv) a visual inspection of the experiments with best and worst results (Sect. S4), and (v) an evaluation of statistical significance of experimental results (Sect. S5). The provided information is not necessary to understand the main paper, but sheds light on interesting aspects not included therein due to the lack of space.*

S1. Comparing loss functions and probability estimates for EMOC

As mentioned in Sect. 3 of the main paper, several choices for loss functions and multi-class classification probabilities are possible when computing the expected model output changes as introduced in Eq. (7). We shortly list two choices for both aspects, and compare the resulting performance within the active discovery scenario of object proposals as tackled in Sect. 6.4.

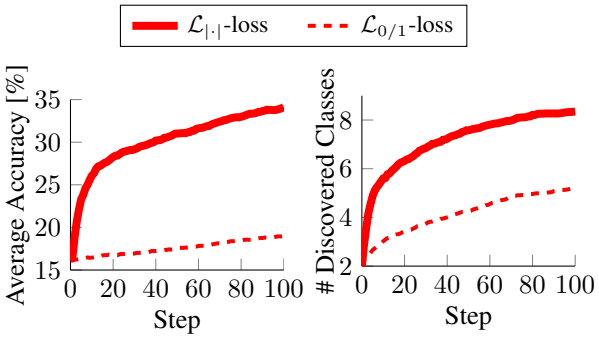
Loss-functions for multi-class scenarios In our paper, we proposed using the L_1-loss on the one-vs-all classification scores to measure the model output change denoted with $\mathcal{L}_{|\cdot|}$:

$$\mathcal{L}_{|\cdot|}\left(f\left(\boldsymbol{x}\right), f'\left(\boldsymbol{x}\right)\right)\right) = \sum_{c=1}^{C} \left| f_c(\boldsymbol{x}) - f'_c(\boldsymbol{x}) \right| \ . \tag{S1}$$

When working in multi-class scenarios, however, an alternative choice would be to directly measure changes in hard classification decisions

$$\mathcal{L}_{0/1}\left(f\left(\boldsymbol{x}\right), f'\left(\boldsymbol{x}\right)\right) = 1 - \delta_{\bar{y}(\boldsymbol{x}),\bar{y}'(\boldsymbol{x})} \tag{S2}$$

where $\delta_{\cdot,\cdot}$ is the Kronecker delta and $\bar{y}(\boldsymbol{x})$ is the hard classification decision as defined in Eq. (10) in the main paper.

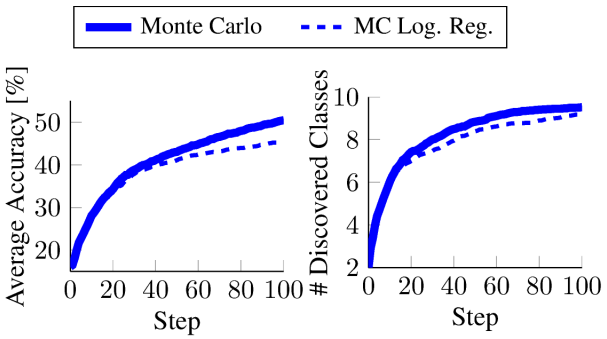


(a) Accuracy (b) Discovery

Figure 1: Comparison of different loss functions when computing EMOC scores according to Eq. (7) in the main paper.

We tested both loss functions within our active learning and discovery framework and the results on the COCO dataset are visualized in Figure 1. As can be seen, comparing continuous classification scores directly with a simple L_1-loss significantly outperforms the loss working on label changes as given in Eq. (S2). We attribute this behavior to two aspects: on one hand, estimates of continuous scores are likely more reliable compared with hard decisions. Thus, we believe that changes in continuous scores are more meaningful, especially in early stages of learning. In addition, changes in hard decisions do not reward samples that would confirm current class estimates, which however would potentially result in decreased classification uncertainties.

Multi-class classification probabilities As explained in the main paper, we apply a Monte-Carlo sampling strategy



(a) Accuracy (b) Discovery

Figure 2: Comparison of different multi-class probability estimates when computing EMOC scores according to Eq. (7) in the main paper.

to compute multi-class probability estimates given the individual class scores. A common alternative is a multi-class logistic regression model [2]

$$p(y = c|\boldsymbol{x}) \propto \exp(\alpha_c \cdot f_c(\boldsymbol{x}) + \beta_c) \qquad \text{(S3)}$$

with class-specific parameters α_c and β_c estimated from training data. In contrast to [2], we used leave-one-out estimates to learn the parameters in Eq. (S3) [10], which provides a stable estimate compared to the estimation on a validation dataset. We compared the resulting strategy with our sampling approach by directly evaluating the active learning and discovery performance. The results are given in Figure 2 and only show a marginal performance difference slightly in favor of the sampling strategy. Thus, we conclude that the sampling technique leads to appropriate probability estimates in our scenario.

S2. Evaluations with focus on labeling times

In the main paper, we only evaluated the *number of queries* with respect to the performance. However, an aspect of equal importance is the total time required to process a query [12]. In real-world applications, this time reflects the time a human annotator needs for labeling as well as the time the active learning approach requires to automatically select an unlabeled example.

For each of the datasets used in our paper, we additionally present here the total time passed during active learning with respect to our two performance measures (number of discovered classes and recognition rate): USPS (Figure 3), Labeled-Faces-in-the-Wild (Figure 4), and COCO (Figure 5). Several plots are given for different assumptions about the labeling time needed by a human annotator, ranging from 1 second per annotation up to 50 seconds.

First of all, we observe that the expected risk minimization strategy requires significantly more time than all other methods, which is consistent with the observations in [11]. The plots further reveal that our methods are able to outperform established approaches in nearly all cases. Thus, we obtain comparable or higher recognition and discovery rates without dramatically increasing computation times required. We therefore conclude that the time needed by our algorithms to select an unlabeled example pays off, which further underlines the benefit of modeling possible rejections and integrating data density when unnameable samples are to be expected.

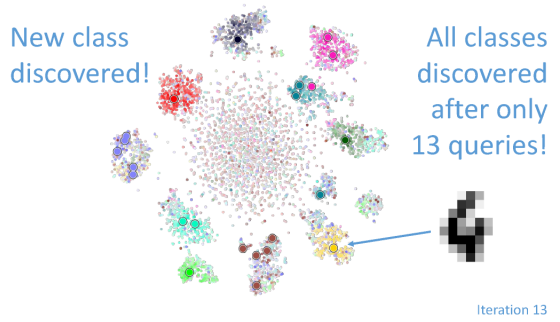


Figure 6: Visualization of query process as shown in the supplementary video.

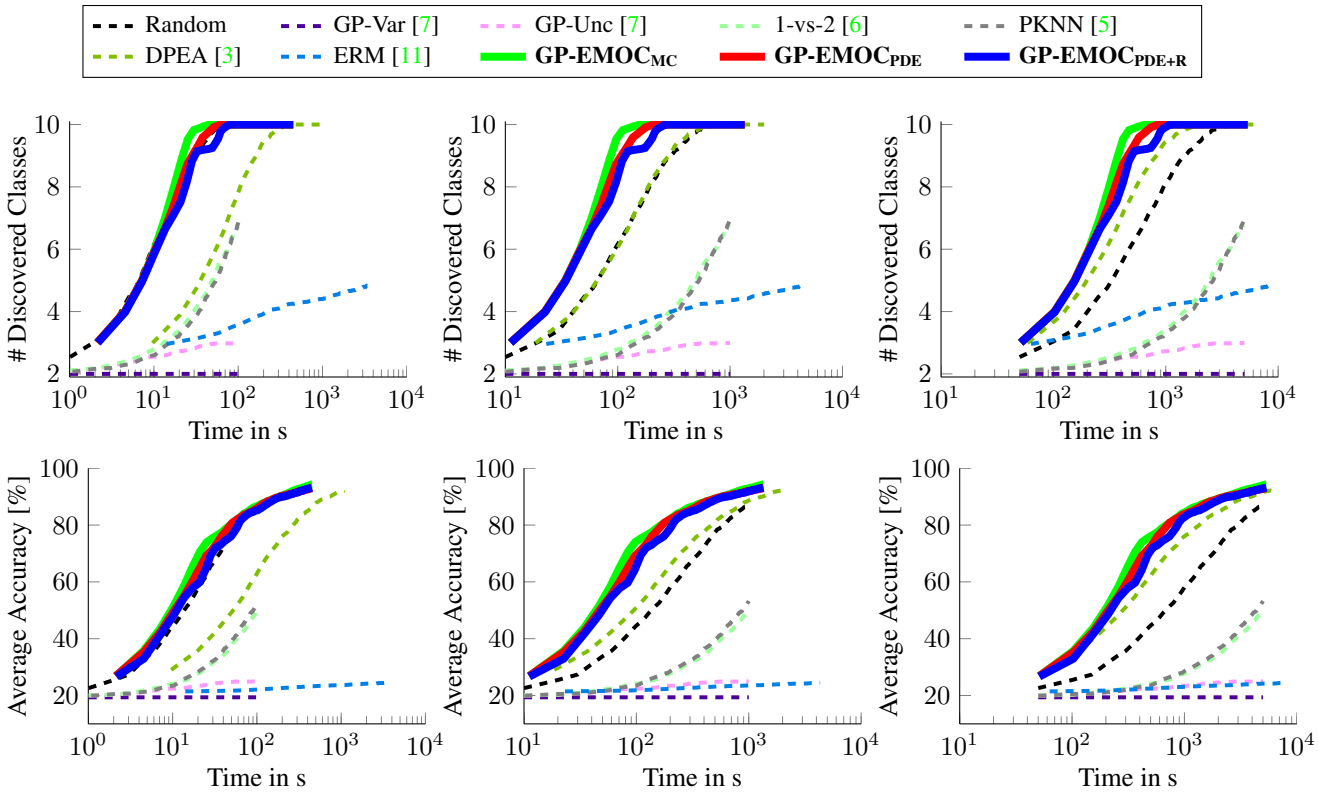
S3. What did we query? Visualizing active learning

Our evaluations are mainly based on quantitative statements, and we limited qualitative results to a few queried segments shown in Figure 5 of the main paper. To further support the understanding of the query process, we created a short video which is part of the supplementary material[1]. After a brief introduction, we visualize the t-SNE feature space for the USPS experiment, we show queries as well as discovered clusters, and we animate the process of active learning and discovery. Given the video, we can again see the benefit of our introduced EMOC method which rapidly discovers new clusters while being unaffected by unnameable instances.

S4. Extreme-case analysis

The experimental evaluation in Sect. 6 of the main paper considers only the mean values of all evaluated test scenarios. In the following, we will provide a more detailed analysis by visualizing the best and worst scenario of $\text{EMOC}_{\text{PDE+R}}$ on the COCO-dataset. Therefore, the individual scenarios are rated according to the relative improvement in accuracy as described in Sect. S5. The obtained scenarios with best and worst improvement are selected for

[1]The video is also available at https://www.youtube.com/watch?v=AEIrYqMHH74.



(a) $1s$ for labeling (b) $10s$ for labeling (c) $50s$ for labeling

Figure 3: Evaluating active class discovery (*top*) and improving recognition accuracy with active learning (*bottom*). Results are obtained on the *USPS dataset* [1] with labeling times ranging from $1s$ (*left*) over $10s$ (*middle*) to $50s$ per query. Baselines are indicated with dotted lines, whereas our techniques are plotted solidly. See main paper for details on the experimental setup. Best viewed in color.

further analysis. In Figure 8, we show for both scenarios exemplary images which either resulted in a drastic change of the recognition performance or which lead to no performance gain at all. A complete overview of all training examples as well as all queried samples in both scenarios is additionally given in Figure 7. Besides their visual beauty, however, we are not able to see any specific characteristics for these images that might give reason for the performance difference of the best and the worst run.

S5. Significance of results

The evaluations in Sect. 6 of the main paper are based on averaging results over 100 individual runs per experiment. We already concluded that our techniques lead to improved learning curves, however, the statistical significance is still unanswered. Therefore, we applied a paired students t-test to evaluate the significance of differences in performance. Evaluations are applied to areas under learning curves corrected by the corresponding initial accuracy, thus, we compare accuracy *improvements* of different techniques. Results given in Table 1 are conducted for the scenario of learning with object proposals from the COCO dataset as introduced in Sect. 6.4 using a significance level of $\alpha = 5\%$. We observe that the resulting improvements in accuracy are statistically significant with p-values smaller than 10^{-2}.

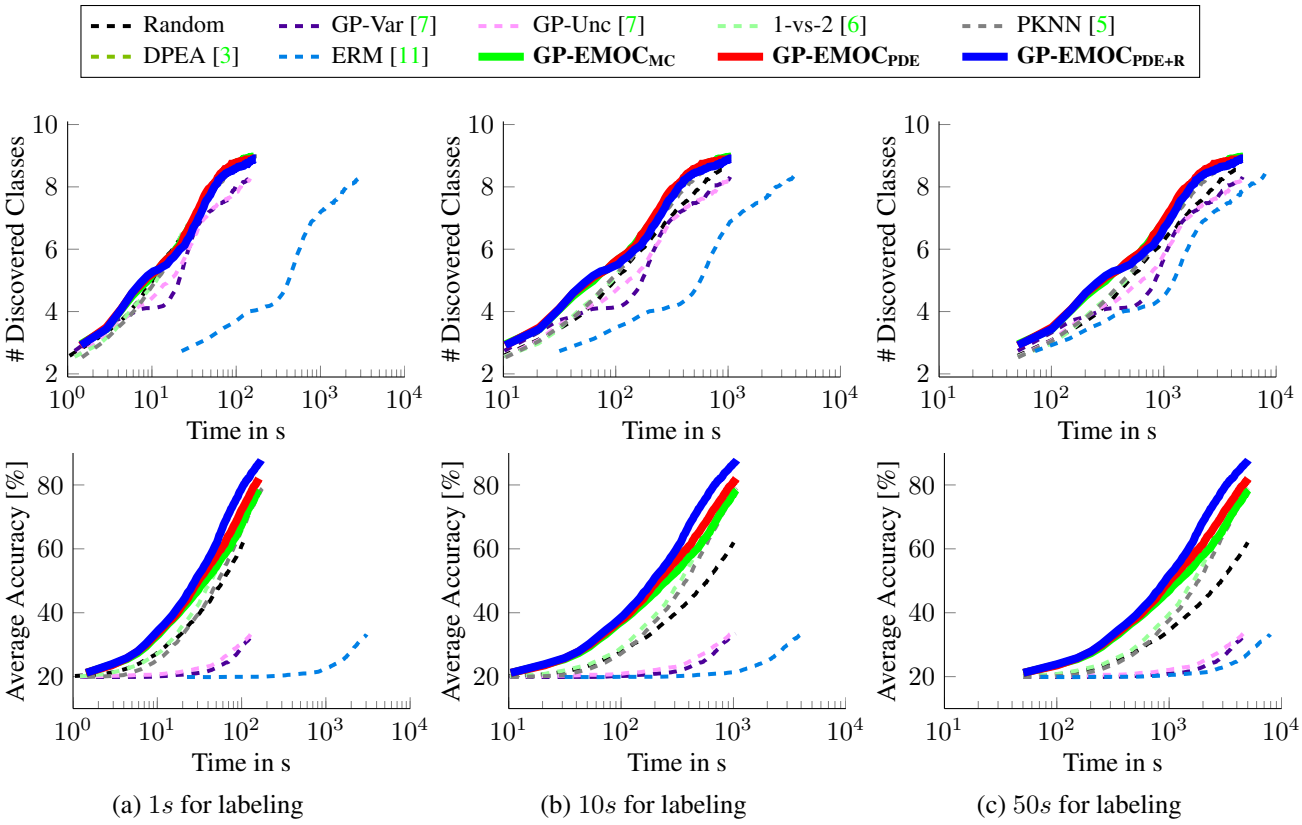
References

[1] K. Bache and M. Lichman. UCI machine learning repository, 2013. 3

[2] K. Duan, S. S. Keerthi, W. Chu, S. K. Shevade, and A. N. Poo. Multi-category classification by soft-max combination of binary classifiers. In *Multiple Classifier Systems (MCS)*, pages 125–134, 2003. 2

[3] T. M. Hospedales, S. Gong, and T. Xiang. A unifying theory of active discovery and learning. In *European Conference on Computer Vision (ECCV)*, pages 453–466, 2012. 3, 4, 5

[4] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 4



(a) $1s$ for labeling (b) $10s$ for labeling (c) $50s$ for labeling

Figure 4: Evaluating active class discovery (*top*) and improving recognition accuracy with active learning (*bottom*). Results are obtained on the *LFW dataset* [4] with labeling times ranging from $1s$ (*left*) over $10s$ (*middle*) to $50s$ per query. Baselines are indicated with dotted lines, whereas our techniques are plotted solidly. See main paper for details on the experimental setup. Best viewed in color.

							GP-EMOC	
	Random	GP-Var [7]	GP-Unc [7]	1-vs-2 [6]	PKNN [5]	ERM [11]	**MC**	**PDE**
GP-EMOC$_{\text{PDE+R}}$	1.4e-4	3.8e-3	2.0e-4	5.2e-6	2.3e-3	3.5e-3	4.8e-21	1.9e-4

Table 1: Evaluating statistical significance of differences in learning curves obtained on the COCO dataset. A paired student t-test validates statistical significance. Numbers shows probabilities for pairwise equality on a significance level of $\alpha = 5\%$.

[5] P. Jain and A. Kapoor. Active learning for large multi-class problems. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 762 –769, 2009. 3, 4, 5

[6] A. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2372 –2379, 2009. 3, 4, 5

[7] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Gaussian processes for object categorization. *International Journal of Computer Vision (IJCV)*, 88:169–188, 2010. 3, 4, 5

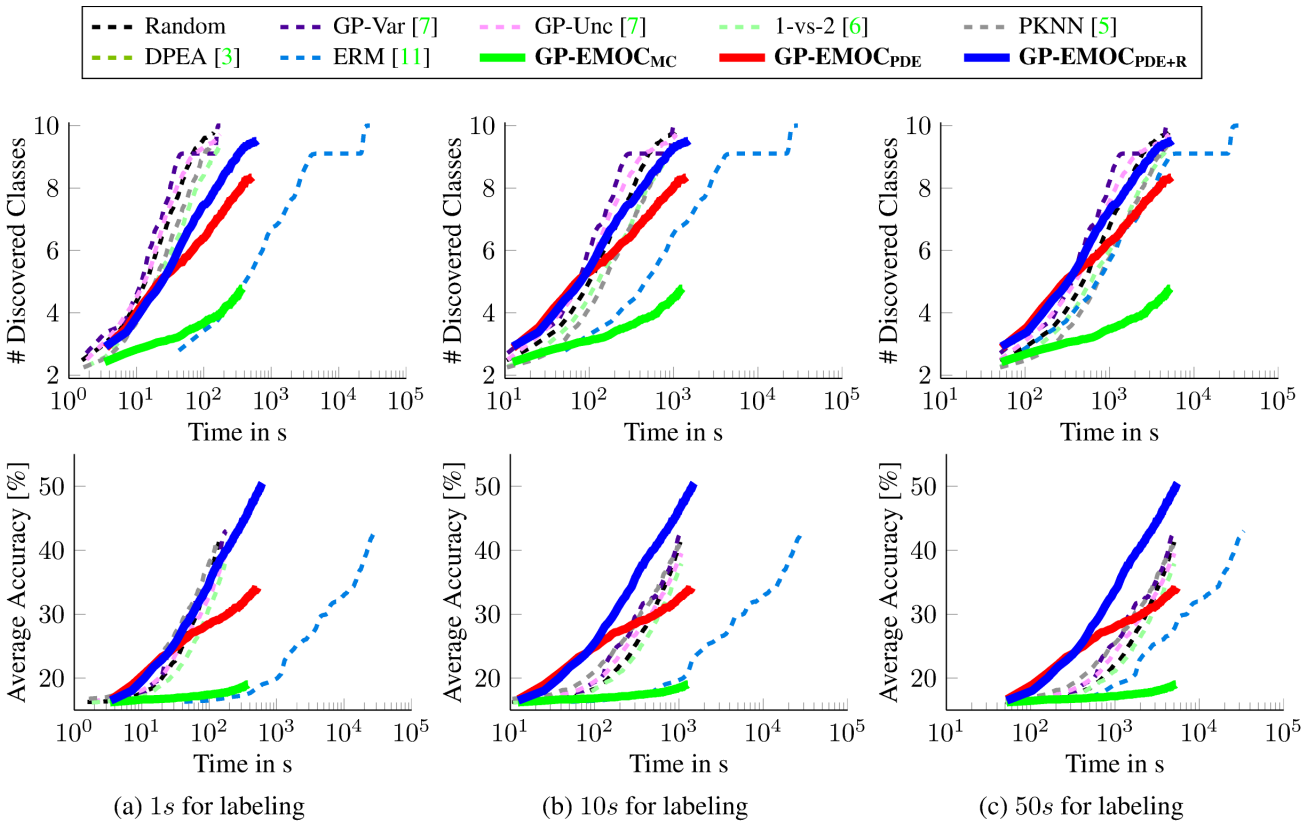
[8] P. Krähenbühl and V. Koltun. Geodesic object proposals. In *European Conference on Computer Vision (ECCV)*, pages 725–739, 2014. 5, 6

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 5

[10] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, MA, USA, 2006. 2

[11] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *International Conference on Machine Learning (ICML)*, pages 441–448, 2001. 2, 3, 4, 5

[12] S. Vijayanarasimhan, P. Jain, and K. Grauman. Hashing hyperplane queries to near points with applications to large-scale active learning. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(2):276–288, 2014. 2



(a) 1s for labeling (b) 10s for labeling (c) 50s for labeling

Figure 5: Evaluating active class discovery (*top*) and improving recognition accuracy with active learning (*bottom*). Results are obtained on the *COCO dataset* [9] with labeling times ranging from 1s (*left*) over 10s (*middle*) to 50s per query. Baselines are indicated with dotted lines, whereas our techniques are plotted solidly. See main paper for details on the experimental setup. Best viewed in color.

Figure 7: A second visual inspection of the scenarios with best and worst results on COCO. We displayed all training images (*top row*) and all queried samples (*bottom blocks, row-wise*) for both, the best (*left*) and worst (*right*) evaluated scenario of EMOC$_{PDE+R}$ on COCO. Again, unsupervised object proposals obtained with [8] as well as the corresponding bounding box for feature extraction are overlayed in red and green. Figure is best viewed in color and by zooming in.

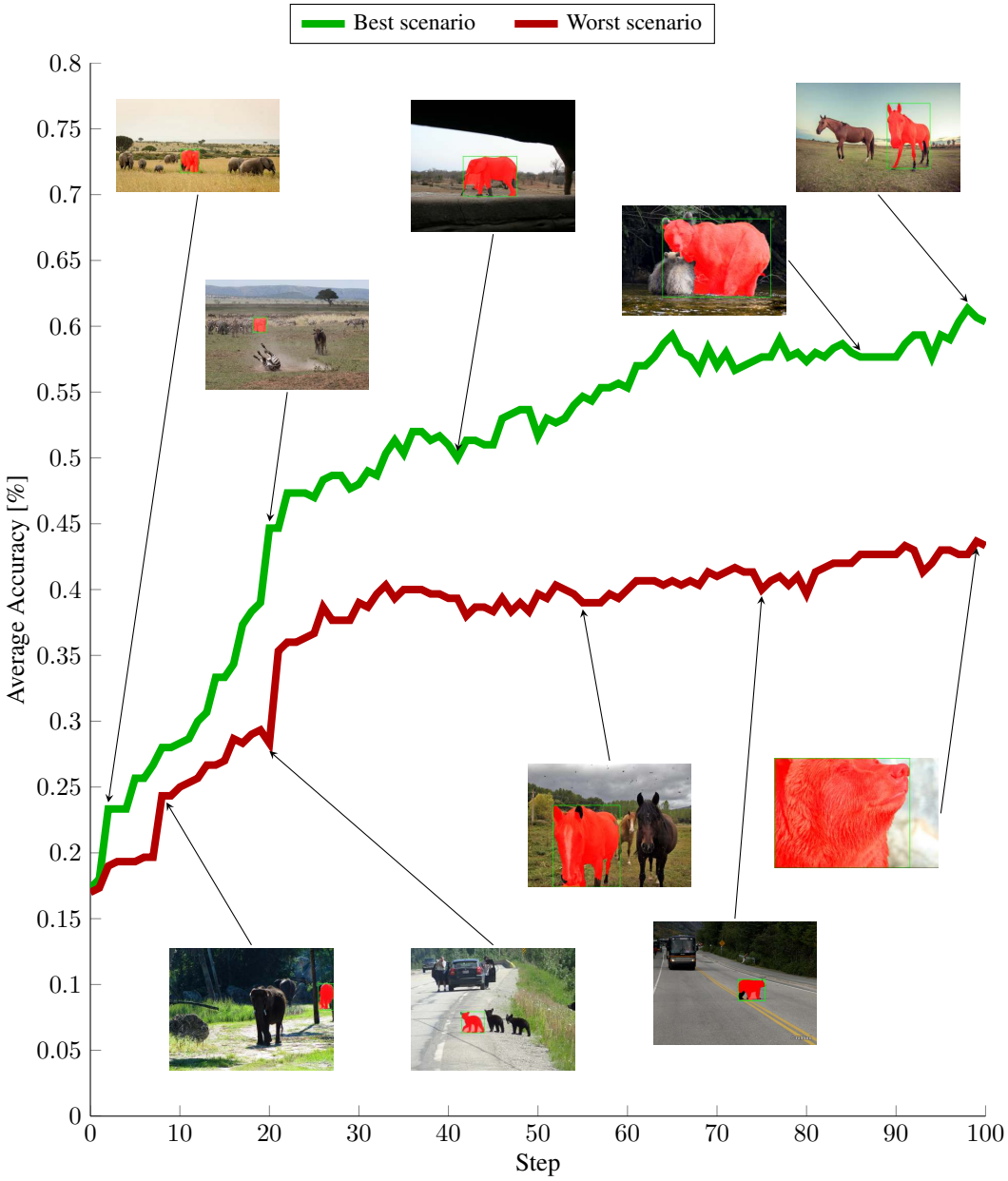


Figure 8: A visual analysis of active learning and discovery on COCO. We displayed learning curves for the best and worst scenario of $\text{EMOC}_{\text{PDE+R}}$. Samples which either resulted in a drastic change of the recognition performance or which lead to no performance gain at all are additionally shown. Unsupervised object proposals obtained with [8] as well as the corresponding bounding box for feature extraction are overlayed in red and green.