MetaAT: Active Testing for Label-Efficient Evaluation of Dense Recognition Tasks (Supplementary Material)

Sanbao Su[1⋆], Xin Li[2], Thang Doan[2], Sima Behpour[2], Wenbin He[2], Liang Gou[2], Fei Miao[1], and Liu Ren[2]

[1] University of Connecticut
{sanbao.su,fei.miao}@uconn.edu
[2] Bosch Research North America & Bosch Center for Artificial Intelligence (BCAI)
{xin.li9,thang.doan,sima.behpour,wenbin.he2,liang.gou,liu.ren}@us.bosch.com

1 Experiments Detials

To ensure clarity and thoroughness in presenting our experimental settings, we expand on the description of the hyperparameters and datasets used in our study. The hyperparameters utilized for training our meta-model are detailed in Table 1. This table provides a comprehensive list of all critical parameters employed in the training process.

Parameter	**Ours**
optimizer	SGD
learning rate (LR)	0.03
LR decay type	cosine
LR warmup steps	2000
training epoch	10
batch size	16
pretrained model	ViT-B_16&32
loss function	focal loss
number of bins	50

Table 1: Training details.

Furthermore, to provide comprehensive information on the datasets used in our study, we present the statistics of training, validation, and testing samples for each dataset in Table 2. The selected datasets include Pascal VOC [8], CityScapes [4], COCO [2], ADE20K [21], and COCO Detection [15]. These datasets vary in size and prediction difficulty, thus demonstrating the robustness and versatility of our approach.

Dataset	Pascal VOC [8]	CityScapes [4]	COCO [2]	ADE20K [21]	COCO Detection [15]
Train	1171	2380	7181	16167	94629
Val	293	595	1796	4042	23658
Test	1449	500	1000	2000	5000

Table 2: Datasets statistics.

⋆ Work done while interning at Bosch Research North America.

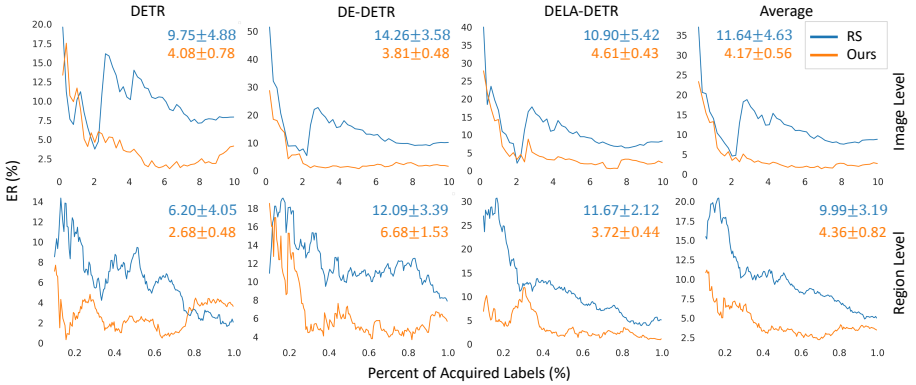

Fig. 1: Performance comparison between MetaAT and RS for evaluating object detection model on the CityScapes dataset [4]. The values in the upper right corner are the mean of ER and its standard deviation. Here, ASE and ATS cannot be directly adapted to object detection as the risk of object detection involves both classification and regression. We can observe that our MetaAT outperforms the RS on both image and region level settings.

2 Additional Experiments for Object Detection

2.1 Results on Extra Dataset

We have extended the application of our approach on the object detection task to one additional dataset CityScapes [4]. As mentioned in the main text, it is nontrivial to adapt ASE and ATS for the object detection task. Consequently, we only compare the performance of our MetaAT with RS, applied to three target models, including DETR [3] DE-DETR [18] and DELA-DETR [18], considering both image-level and region-level annotation settings.

Performance with Respect to Annotation Budget. Figure 1 presents the results of the ER calculated across three target models. We use annotation budget ranges $\leq 10\%$ and $\leq 1\%$ for image and region level experiments, respectively. Overall, our MetaAT consistently outperforms RS in risk estimation under all settings. Notably, it achieves a reduction of up to 98% in ER compared to RS, specifically on DE-DETR at the image level experiment.

Numerical Comparison of Performance. The values in the upper right corner of Figure 1 present the average ER across the annotation budget range and its corresponding standard deviation. Our approach always outperforms RS on bias and variance. On average, our MetaAT archives a reduction of 64% and 56% in ER, and 88% and 74% in standard deviation, compared to RS, at the image level and region level.

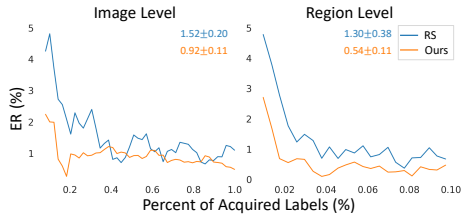

Fig. 2: Performance comparison between MetaAT and RS for YOLOX model.

2.2 Results on Anchor-based Model

Previous target models for the object detection task are mainly transformer-based. Here we extended the application of our approach on the object detection task to the anchor-based target model YOLOX [10] on the COCO Detection dataset [15]. We compare the performance of our MetaAT with RS, considering both image-level and region-level annotation settings. Figure 2 presents the results of the ER calculated on the YOLOX [10] model. Here we use annotation budget ranges $\leq 1\%$ and $\leq 0.1\%$ for image and region level experiments, respectively. The values in the upper right corner of Figure 2 present the average ER across the annotation budget range and its corresponding standard deviation Overall, our MetaAT outperforms RS. On average, our MetaAT archives a reduction of 39% and 58% in ER, and 45% and 71% in standard deviation, compared to RS, at the image level and region level.

3 Discussions About Run Time

Target Model	PSPNet	UNet	SEGNet	FCN
ASE & ATS	67.80h	40.68h	32.88h	62.12h
Our MetaAT	11.15h	11.27h	10.74h	10.98h

Table 3: Comparison of training times (in hours) between our MetaAT and other methods (ASE and ATS) on the VOC dataset.

Training dense recognition models is notably expensive; for example, DETR [3] requires 1152 GPU hours on the V100 (130 TFLOPS), making the current state-of-the-art methods such as ASE and ATS, which necessitate training additional segmentation/detection models, impractical. In contrast, the meta-model used in our approach is a single **regression model** for loss estimation. Utilizing a pre-trained ViT, our approach requires only 10 hours on a single RTX 2080 Ti (14.2 TFLOPS), costing just a few dozen dollars on the cloud (AWS). This is clearly illustrated in Table 3 for the segmentation task. On average, ASE and ATS require 4.6 times more training time compared to MetaAT.

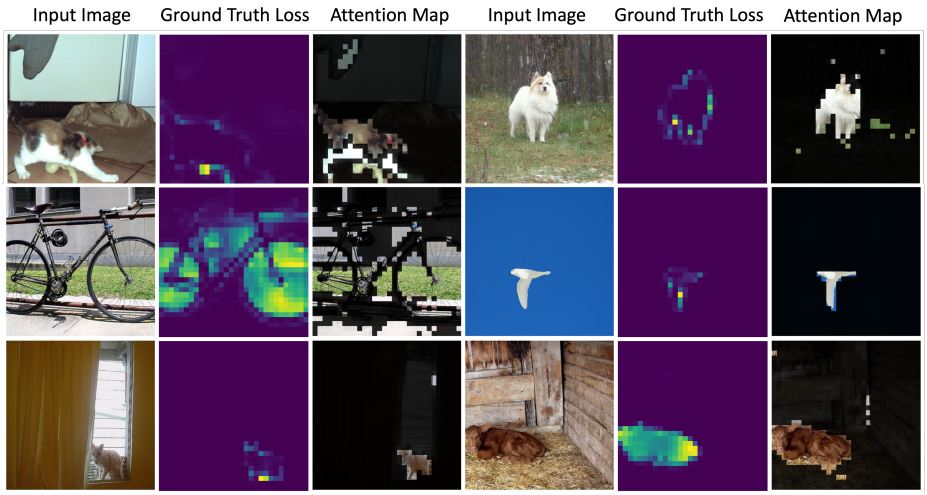

Fig. 3: Examples of the attention maps employed by our MetaAT. The high alignment between the ground-truth loss map and the attention map indicates that our model prioritizes the high-loss areas when performing image-level estimations.

4 Image Level Attention Maps

Figure 3 showcases six examples of attention weights employed by our meta model during image-level estimation. Significantly, the model shows a tendency to focus on regions with high losses, which aligns well with our rationale for choosing the ViT as the backbone architecture. Additionally, the model displays versatility in dealing with regions of high losses, regardless of their size. For instance, in small areas with high losses, such as the bird in row two and column four, our model precisely focuses its attention. Conversely, in larger areas with high losses, like the wheels in row two and column one, it effectively identifies and concentrates on these larger regions.

5 Relations Between Entropy and Test Loss

In the main text, we conduct an ablation study to highlight the crucial role of output entropy in our model's inputs for the segmentation task. To further clarify this relationship, Figure 4 provides visual examples of both ground-truth loss and entropy. These examples demonstrate a close alignment between entropy and ground-truth loss, particularly in border regions. For instance, at the top right corner of the figure, we observe high losses around the sheep's margins, which correspond with increased entropy in the same areas. This observation reinforces the strong correlation between entropy and test loss, as discussed in the main text.

However, a detailed analysis of Figure 4 shows that the alignment is not always perfect. For example, at the top left corner, while the entropy effectively

highlights the boundary between two pillows on the sofa, it assigns low entropy to the brown pillow, which is in contrast to the ground-truth loss. This observation explains our rationale for still needing the model output and the original image to further enhance the accuracy of the loss estimation.

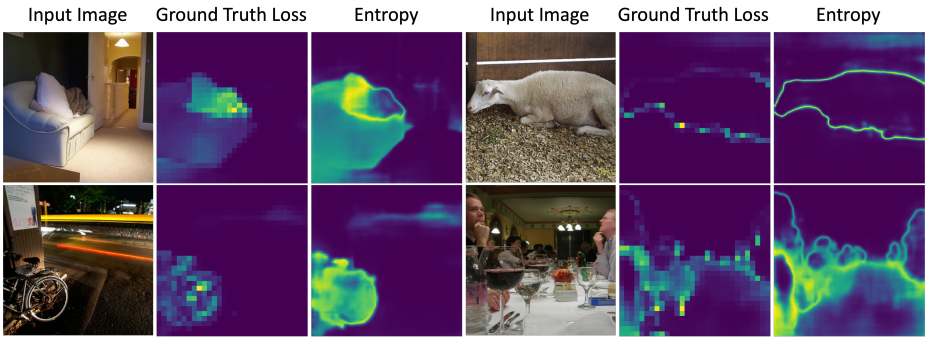

Fig. 4: Examples illustrating the relationship between entropy and ground-truth loss. The entropy map is closely aligned with the ground-truth loss map, providing visual evidence of entropy's crucial role among all inputs for our meta model.

6 Loss Distribution

Figure 5 provides visualizations of the distribution of ground-truth losses and estimated losses from MetaAT, ATS, and ASE at various levels: image, region-32, and region-16.

Firstly, these visualizations at the region-32 and region-16 levels clearly show that the test loss distribution is highly imbalanced. This supports our hypothesis that in segmentation and object detection, only small regions, such as borders and areas of ambiguity, account for most of the test errors. Additionally, this observation explains why we utilize equal-width binning and focal loss in training our meta model. These methods are designed to address the challenges posed by such highly imbalanced data.

Secondly, our approach demonstrates an estimated loss distribution that closely aligns with the ground-truth loss distribution across all sample levels, exhibiting superior performance compared to other methods. For example, at the region-32 level, where 90% of ground-truth losses are zeros, our approach accurately estimates 90% of losses as zeros. In contrast, ASE estimates only 60% of losses as zeros, resulting in a significant overestimation of losses in at least 30% of samples. This discrepancy is due to ASE's reliance on epistemic uncertainty for estimations, which does not consistently correlate with ground-truth losses. The losses estimated by ATS closely resemble a Gaussian distribution, characterized by a low value of loss (< 0.07). It suggests a limited ability in

accurately estimating losses. This might be attributed to the lack of diversity in ATS's deep ensemble models. The behavior of ATS and ASE aligns with the qualitative results shown in the previous subsection.

Finally, it should be noted that, although ASE appears to have a better loss distribution estimation than ATS (Figure 5), it often yields reversed predictions (Figure 6 in the main text), leading to lower performance in active testing compared to ATS.

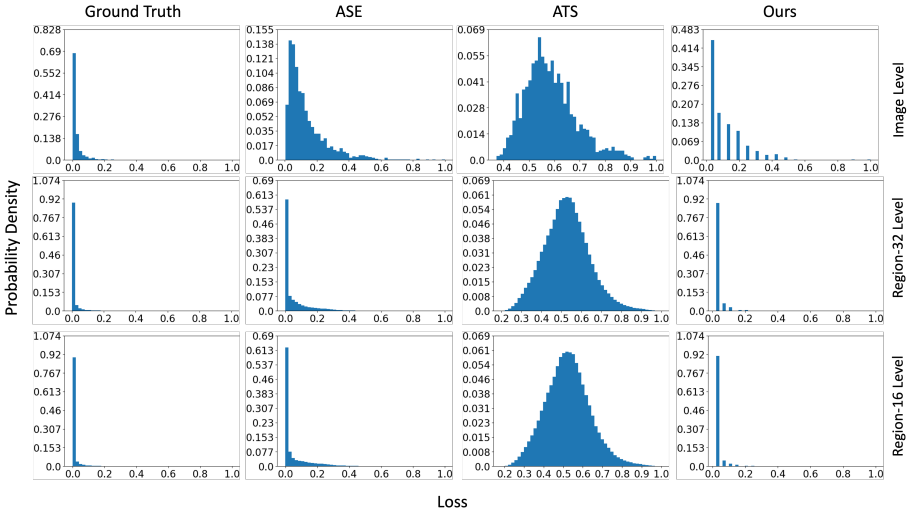

Fig. 5: Examples of loss distribution of the ground-truth losses and estimated losses from MetaAT, ATS and ASE, at the image level, region-32 level and region-16 level. Our MetaAT outperforms other approaches on loss estimation at all sample levels.

7 Extended Related Work

7.1 Active Testing

The work in [16] was the first to utilize importance sampling to estimate active risk; however, it does not adhere to the pool-based setting. In [20], active testing based on Poisson sampling was proposed. While these methods outperform simple baselines, they rely on non-adaptive acquisitions that cannot adjust to the test data. To address this challenge, Active Testing Surrogate (ATS) [14] and Active Surrogate Estimator (ASE) [13] were introduced, as discussed in Section 2. Our MetaAT not only retains the adaptive acquisition policy but also enhances active testing performance in dense recognition tasks.

7.2 Model Evaluation without Labels

Testing with a few test labels or even without test labels can also make model evaluation more efficient. The works in [5–7,11] estimate accuracy using multiple other datasets. Specifically, [5] estimates the performance gap between training and test datasets through persistent topology measures, while [7] trains a regression model on a meta-dataset to estimate classification accuracy on test datasets. However, these methods require access to multiple related datasets with available labels, thereby increasing labeling costs. Another approach, which mainly focuses on segmentation task, involves utilizing few-shot segmentation models [9,17,19] to generate pseudo-labels and compute the risk as the loss between the pseudo-labels and the target model's outputs. These few-shot segmentation models are trained on the training dataset and a few test samples. Nonetheless, they often fail to provide accurate labels in challenging areas, which are crucial for risk evaluation, resulting in ineffective estimates.

8 Additional Ablation Study

8.1 Model Architecture

In Table 2, we show that our meta-model outperforms ResNet [12] by around 20% at the image level. Here, we present new results at the region-16 level: our model achieves 4.27 ± 1.4 compared to ResNet's 7.99 ± 3.25, indicating a significant error reduction. For the object detection task, the meta-model takes a list of query features as input, which represent objects in images. Unlike image inputs for CNNs, these query features do not have a fixed spatial relationship, making it non-trivial to modify a CNN model to process them.

8.2 Few-shot Training

In the few-shot setting, our MetaAT faces challenges due to fewer labels. However, our method's region design, which divides images into 225/900 patches, and the joint training of the image and region heads, help address this problem. Table 4 compares the performance of our MetaAT on 100% and 1% training samples of VOC [8] with SEGNet [1]. For VOC, 1% training samples mean 11 samples, which is around 1 shot. The results show minor performance drops. Especially for the region-16 level, there is an 18.58% reduction. These results indicate that the impacts, while present, are not statistically significant (P-values of 0.24 and 0.38).

Annotation Level	100% Training Sample	1% Training Samples
Image Level	17.52 ± 2.09	21.64 ± 4.36
Region-16 Level	9.13 ± 4.78	10.83 ± 3.04

Table 4: Ablation study for the size of training samples on SEGNet model and VOC.

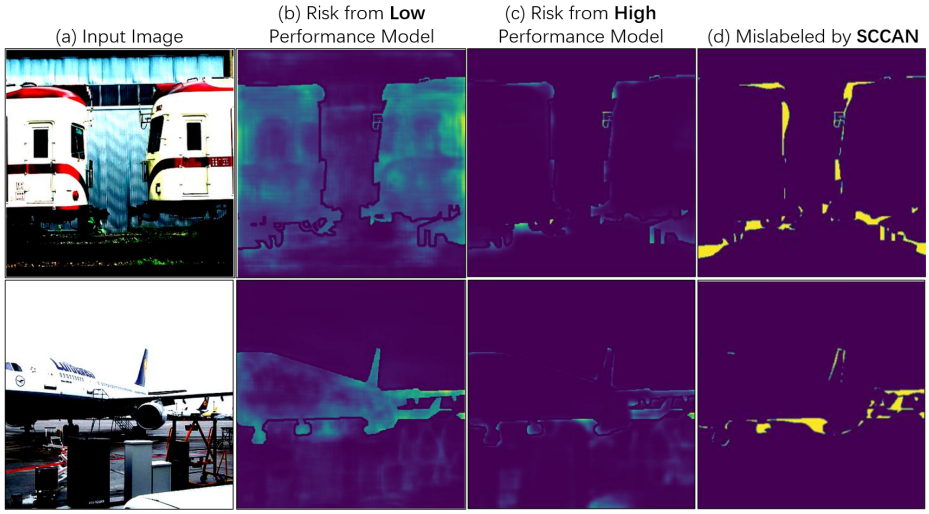

Fig. 6: Qualitative examples from SCCAN [19].

8.3 Comparison with Few-shot Segmentation

Another active testing pipeline for the segmentation task utilizes the few-shot segmentation model to predict the pseudo labels. The trained few-shot segmentation models, such as SCAAN [19] and GFS-Seg [17], are retrained on the original training dataset and the selected annotated test dataset. Then the predictions of the few-shot segmentation model are served as pseudo labels to estimate the risk of the target segmentation model on the whole test dataset. Here we compare the performance of this pipeline, which uses SCAAN [19] as the few-shot segmentation model, with our MetaAT on VOC [8], as shown in Table 5. All approaches utilize the whole training dataset and 10% test labels. We also provide qualitative examples of SCCAN [19] in Figure 6. While SCCAN achieves a high IOU (91.9) after training with the training dataset and 10% of testing labels, it underperforms compared to MetaAT. SCCAN struggles to provide correct labels in challenging areas, as shown in Figure 6d, which are crucial for risk evaluation in high-performance segmentation models (Figure 1c), leading to ineffective estimates. However, in scenarios with weaker models like UNet, where risk is more widespread as shown in Figure, SCCAN's correct labeling in those areas can match MetaAT's performance at the image level.

Methods	PSPNet	UNet	SEGNet	FCN
SCCAN [19]	35.15	3.11	33.31	20.97
Ours (Image)	5.93	2.71	9.79	4.10
Ours (Region-16)	0.64	0.42	0.44	0.46

Table 5: Results of SCCAN as baseline (ER on **10%** test labels).

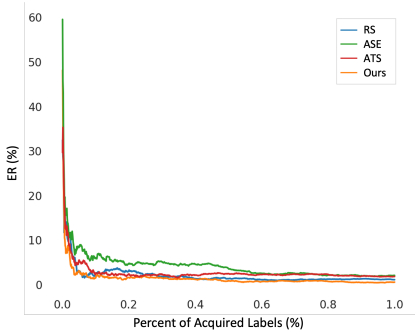

Fig. 7: Performance comparison between our MetaAT and three competitors with respect to larger annotation budgets for the region-16 level annotation and the semantic segmentation task on the ADE20K dataset. The ER is very close to zero under 1% acquired labels with our MetaAT.

8.4 Large Annotation Budget

To evaluate the performance of active testing with an increased annotation budget, we assessed our MetaAT model and three competitors using up to a 1% region-16 level annotation budget on the ADE20K dataset. This evaluation served as a sanity check to verify that a sufficiently large annotation budget leads to error convergence towards zero. Figure 7 shows the average ER calculated across all four target models under this larger annotation budget for the region-16 level. As the percentage of acquired labels increases, the ER converges to zero, with MetaAT consistently achieving the best performance. Specifically, Table 6 presents the ER of all approaches under a 1% region-16 level annotation budget. The ER of our MetaAT is very close to zero. Moreover, MetaAT achieves up to a 72% ER reduction compared to competitors under this large annotation budget, demonstrating the robustness of our approach.

Model	RS	ASE	ATS	Ours
PSPNet	1.79 ± 1.72	3.56 ± 2.05	2.67 ± 1.38	**0.40 ± 0.34**
UNet	**0.23 ± 0.11**	1.06 ± 0.55	1.96 ± 0.27	0.75 ± 0.59
SEGNet	1.63 ± 1.07	1.60 ± 0.33	0.89 ± 0.39	**0.42 ± 0.33**
FCN	0.90 ± 0.59	2.17 ± 1.40	1.88 ± 1.33	**0.74 ± 0.54**
Average	1.14 ± 0.87	2.10 ± 1.08	1.85 ± 0.84	**0.58 ± 0.45**

Table 6: The mean of ER and standard deviation of segmentation model evaluation for the region-16 level annotation under 1% test labels of ADE20K dataset. The best results are highlighted in **bold**.

References

1. Badrinarayanan, V., Kendall, A., SegNet, R.C.: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561 **5** (2015)
2. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset. In: CVPR Workshop on the Future of Datasets in Vision. vol. 2. sn (2015)
5. Corneanu, C.A., Escalera, S., Martinez, A.M.: Computing the testing error without a testing set. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2677–2685 (2020)
6. Deng, W., Gould, S., Zheng, L.: What does rotation prediction tell us about classifier accuracy under varying testing environments? In: International Conference on Machine Learning. pp. 2579–2589. PMLR (2021)
7. Deng, W., Zheng, L.: Are labels always necessary for classifier accuracy evaluation? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15069–15078 (2021)
8. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**, 98–136 (2015)
9. Fan, Q., Pei, W., Tai, Y.W., Tang, C.K.: Self-support few-shot semantic segmentation. In: European Conference on Computer Vision. pp. 701–719. Springer (2022)
10. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
11. Guerra, S.B., Prudêncio, R.B., Ludermir, T.B.: Predicting the performance of learning algorithms using support vector machines as meta-regressors. In: Artificial Neural Networks-ICANN 2008: 18th International Conference, Prague, Czech Republic, September 3-6, 2008, Proceedings, Part I 18. pp. 523–532. Springer (2008)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Kossen, J., Farquhar, S., Gal, Y., Rainforth, T.: Active surrogate estimators: An active learning approach to label-efficient model evaluation. NIPS **35**, 24557–24570 (2022)
14. Kossen, J., Farquhar, S., Gal, Y., Rainforth, T.: Active testing: Sample-efficient model evaluation. In: ICML. pp. 5753–5763. PMLR (2021)
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: 2014 European Conference on Computer Vision. pp. 740–755. Springer (2014)
16. Sawade, C., Landwehr, N., Bickel, S., Scheffer, T.: Active risk estimation. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). pp. 951–958 (2010)
17. Tian, Z., Lai, X., Jiang, L., Liu, S., Shu, M., Zhao, H., Jia, J.: Generalized few-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11563–11572 (2022)

18. Wang, W., Zhang, J., Cao, Y., Shen, Y., Tao, D.: Towards data-efficient detection transformers. In: European conference on computer vision. pp. 88–105. Springer (2022)
19. Xu, Q., Zhao, W., Lin, G., Long, C.: Self-calibrated cross attention network for few-shot segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 655–665 (2023)
20. Yilmaz, E., Hayes, P., Habib, R., Burgess, J., Barber, D.: Sample efficient model evaluation. arXiv preprint arXiv:2109.12043 (2021)
21. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)