SUPPLEMENTARY FMG-DET: FOUNDATION MODEL GUIDED ROBUST OBJECT DETECTION

Darryl Hannan *Timothy Doster* *Henry Kvinge* *Adam Attarian* *Yijing Watkins*

Pacific Northwest National Laboratory
Seattle, WA

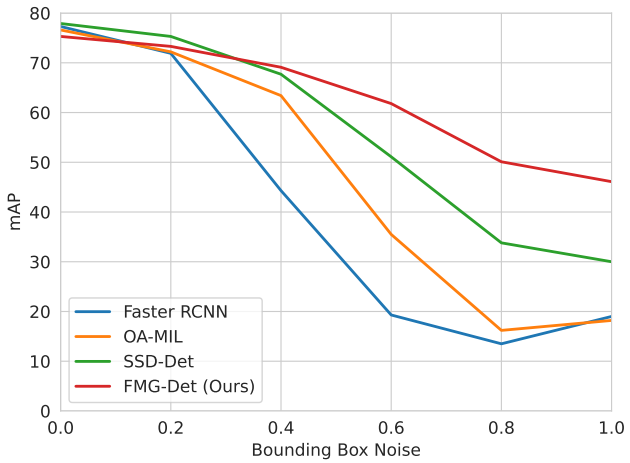

Fig. 1. VOC test mAP demonstrating the considerable impact of bounding box noise on model performance in prior state-of-the-art models, including OA-MIL and SSD-Det, compared to our proposed FMG-Det algorithm.

1. EFFECTIVENESS OF PRIOR APPROACHES IN HIGH NOISE SCENARIOS

Figure 1 presents results on PASCAL VOC at various noise levels, ranging from 0.0 (no noise) to 1.0 (severe noise). Prior work only focused on noise levels up to 0.4; opting to not even run their models at higher noise levels. The Faster RCNN model is just a standard detector without any noise mitigation approaches. OA-MIL [1] and SSD-Det [2] are prior noise mitigation approaches that are added to a Faster RCNN model. As the amount of noise increases beyond 0.4, the mAP deteriorates significantly for each of these approaches. Our proposed method, FMG-DET, addresses this shortcoming, retaining strong performance even under severe noise.

2. INSTANCE INTERPOLATION MODULE

Figure 2 provides an overview of our Instance Interpolation module. The details of this module are available in the main paper.

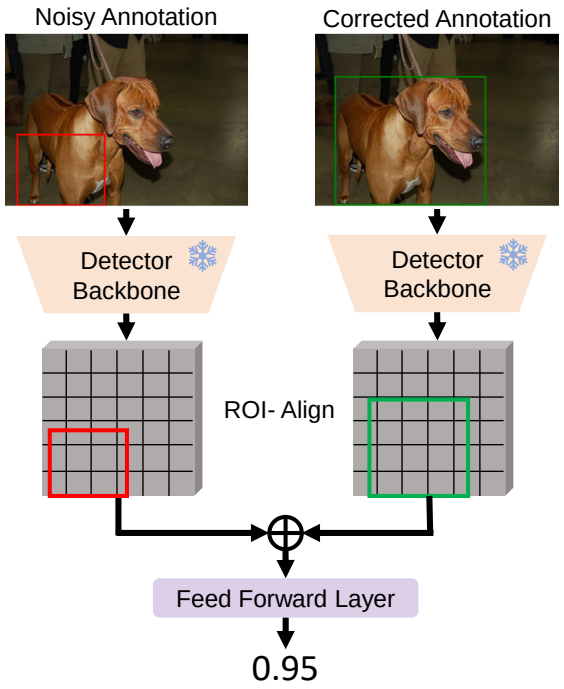

Fig. 2. An overview of our proposed Instance Interpolation module. Both the corrected and noisy bounding boxes are passed to this module. It then extracts features for each bounding box using the backbone that already exists in the detector, and using these features, predicts a value γ that is used to then interpolate between the corrected and noisy boxes.

Shots	Model	0.0	0.2	0.4	0.6	0.8	1.0	MAE
1-shot	Faster R-CNN	19.2	15.3	5.1	0.8	0.4	1.7	12.1
	OA-MIL	22.1	14.7	4.9	1.4	0.7	2.1	11.6
	FMC	17.6	15.3	7.8	3.5	2.6	1.3	11.2
	SSD-Det	21.8	18.3	14.7	4.7	1.1	1.6	8.8
	FMG-Det	21.4	19.2	10.8	6.0	3.3	2.4	**8.7**
2-shot	Faster R-CNN	37.7	28.7	12.0	3.6	2.1	3.3	23.1
	OA-MIL	41.2	31.5	11.5	3.6	1.6	3.1	22.3
	FMC	36.5	29.0	17.9	11.1	8.0	5.5	19.7
	SSD-Det	40.1	33.8	28.4	11.9	5.4	4.4	17.0
	FMG-Det	40.4	35.7	22.9	13.7	7.4	6.4	**16.6**
3-shot	Faster R-CNN	52.2	37.8	14.3	5.4	3.3	5.2	32.5
	OA-MIL	52.3	39.1	15.8	5.6	3.3	4.3	32.1
	FMC	50.1	40.7	25.6	15.3	12.2	8.8	26.8
	SSD-Det	51.9	44.8	37.2	15.9	7.4	6.5	24.9
	FMG-Det	51.6	41.9	30.9	17.9	11.2	11.1	**24.7**
5-shot	Faster R-CNN	59.6	44.7	21.2	8.0	5.6	8.3	35.0
	OA-MIL	61.1	49.4	23.4	8.3	5.3	6.3	34.0
	FMC	59.6	49.0	33.4	20.9	17.2	14.2	27.2
	SSD-Det	58.1	49.7	42.8	21.8	10.7	8.7	27.6
	FMG-Det	59.7	50.6	39.5	23.3	16.3	14.0	**25.7**
10-shot	Faster R-CNN	63.0	49.5	22.2	8.9	6.2	10.3	36.3
	OA-MIL	63.6	55.6	28.7	11.3	5.9	7.0	34.3
	FMC	62.2	52.4	40.5	24.6	20.3	18.2	26.6
	SSD-Det	62.4	56.1	50.1	26.5	13.2	13.6	26.0
	FMG-Det	61.8	57.4	44.5	26.9	20.0	17.5	**25.0**

Table 1. Mean average precision for few-shot PASCAL VOC Novel Set 1 dataset.

3. FULL FEW-SHOT DETECTION RESULTS

Table 1 contains the full results for our few-shot experiments in the main paper, highlighting the performance of each approach at all noise levels for each few-shot scenario.

4. COMPUTATIONAL EFFICIENCY OF FMC PIPELINE

While the foundation model correction pipeline involves large foundation models, it can be run entirely offline with the remainder of the FMG-Det architecture being leveraged to prioritize performance, or a more efficient detector leveraged to prioritize training and inference time. For the experiments in this paper, due to compute limitations, we ran our experiments with a batch size of 1 on a single Tesla V100. Computation scales primarily with the number of images but also with the number of bounding boxes in each image. For COCO, this resulted in a rate of 2.475 seconds/image, and for VOC, a rate of 1.350 seconds/image. Note that due to the pipeline being training-free, it is highly conducive to distribution across multiple GPUs, where the dataset can easily be sharded, dramatically increasing the inference speed.

5. FULL ABLATIONS

Table 2 contains ablations for our proposed FMG-Det model, starting from the base detector, Faster RCNN [3], and adding each of our proposed components, along with OA-MIL. Our Foundation Model Correction pipeline makes the largest contribution to the overall performance of our proposed approach, improving MAE from 36.4 to just 17.5. Relative to state-of-the-art approaches, just including this pipeline already achieves state-of-the-art performance on PASCAL VOC. This is critical as it is fully detector agnostic and therefore it is reasonably assumed that virtually any object detector would enjoy similar benefits. Interestingly, adding OA-MIL directly on top of this pipeline slightly decreases performance. However, adding our instance interpolation module to OA-MIL does boost performance further by a substantial margin. We did explore using SSD-Det instead of OA-MIL to see whether performance could be improved further by simply swapping them. Yet, we found that this, similar to adding OA-MIL directly on top of the foundation model correction pipeline, did not result in further performance improvements, suggesting that there is some redundancy between the learned denoising procedure in SSD-Det and our own proposed contributions. We hypothesize that adding the FMC pipeline

Model	0.0	0.2	0.4	0.6	0.8	1.0	MAE
Faster RCNN	77.3	71.9	44.3	19.3	13.5	19.0	36.4
FM Correction	75.0	72.1	66.1	55.2	46.5	44.0	17.5
FM Correction + OA-MIL	75.1	72.2	67.0	55.5	44.4	41.4	18.0
FM Correction + Instance Interpolation	76.6	73.4	67.8	58.2	48.7	44.9	15.7
FMG-Det	75.7	73.2	69.3	62.6	50.2	46.5	**14.4**

Table 2. Ablations for FMG-Det using the Pascal VOC 2007 dataset, demonstrating the performance impact of each component. Starting from Faster RCNN, FMG-Det adds the Foundation Model Correction (FM Correction) pipeline, instance interpolation, and then leverages OA-MIL

.

diminishes the issues of object drift and group prediction, two of the primary issues with OA-MIL that motivated SSD-Det, reducing the effectiveness of the latter's improvements.

5.1. Experimental Details

Our model was built off of the OA-MIL [1] repository, which is in turn built on top of MMDetection [4]. We use Faster R-CNN [3] with a ResNet-50 [5] backbone as our object detector architecture due to its simplicity and the fact that we are not prioritizing overall model performance, rather we are focused on improving model robustness. However, note that FMG-Det is directly compatible with any 2-stage detector and the Foundation Model Correction pipeline is compatible with virtually any detector or alternative denoising technique. We use many of the defaults provided by MMDetection for the Faster R-CNN model. For our experiments on the full VOC and COCO datasets, we use a batch size of 8 for all of our experiments. All models are trained using the standard 1x learning schedule, which is run over 12 epochs and consists of SGD with a learning rate of 0.02, momentum of 0.9, a weight decay of 0.0001, a warmup linear scheduler that executes over 500 iterations with a warmup ratio of 0.001, and a multistep scheduler with milestones at 8 and 11 epochs with a gamma value of 0.1. For the OA-MIL [1] and SSD-Det [2] baselines, we used the defaults provided in the authors' repositories. For our foundation model correction pipeline, we used an α of 0.5 to mix the scores from SAM and CLIP and a λ of 0.05 as our IoU threshold for accepting a correction. We selected these values by empirically running on subsets of the data and adapting them to minimize the number of dramatically shifted bounding boxes, e.g., egregious mistakes such as placing the bounding box around the background rather than the target.

We use the same training procedure outlined in [6] to train our few-shot models, where a base model is trained, the weights are frozen except the head, then the model is fine-tuned on a mixed base+novel few-shot set. We train the base model using each of our proposed techniques and baselines, using the SAM hyperparameters that are outlined above. We do leverage slightly different hyperparameters for fine-tuning on the novel set. Namely, we fine-tune most of the models with a learning rate of 0.1 and a fixed learning schedule. The only exception is SSD-Det [2], which we found works best with the same learning scheduler as above, just with a lower learning rate of 0.01.

6. LIMITATIONS

Our proposed approach is reliant upon the performance of SAM on the target dataset. While SAM has demonstrated exceptionally strong performance across a wide variety of domains [7], in highly specialized domains that involve distinct image modalities, such as medical imaging, our proposed approach might not be effective if the quality of the extracted masks is low. However, this issue can be mitigated by leveraging a SAM variant that is better suited for the target dataset, whether it is a model like MedSAM [8], which has been specifically trained for medical imagery, or simply a SAM model that has been finetuned on images that are more closely aligned with the detection task. Another limitation of our proposed approach is that it struggles with bounding boxes that have no overlap with the groundtruth object. In these cases, SAM is likely to segment part of the background or an adjacent object. This can result in severe failure scenarios, where the box becomes more inaccurate than the noisy groundtruth. We attempt to mitigate such scenarios by discarding corrected boxes with no overlap, defaulting to the noisy groundtruth. However, this solution negates the benefits of the foundation model correction pre-processing step. Lastly, we were also unable to test our approach on noisy, publicly available datasets that could be reported in our paper. This dataset would need to have naturally noisy training labels but clean, high-quality testing labels for validation. However, as discussed in the main paper, we do believe that our synthetic setting is more challenging than most real world settings. Real world noise is likely to follow a standard pattern, e.g., perhaps a target is frequently occluded and a common mistake is to place the box over the entire target rather than just the portion that is visible. This standard pattern would be much easier to learn and correct than the stocastic pattern that is applied in our experiments.

7. REFERENCES

[1] Chengxin Liu, Kewei Wang, Hao Lu, Zhiguo Cao, and Ziming Zhang, "Robust object detection with inaccurate bounding boxes," in *ECCV*. Springer, 2022, pp. 53–69.

[2] Di Wu, Pengfei Chen, Xuehui Yu, Guorong Li, Zhenjun Han, and Jianbin Jiao, "Spatial self-distillation for object detection with inaccurate bounding boxes," in *ICCV*, 2023, pp. 6855–6865.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, vol. 28, 2015.

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint*, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[6] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu, "Frustratingly simple few-shot object detection," in *ICML*, 2020, pp. 9919–9928.

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick, "Segment anything," *arXiv preprint*, 2023.

[8] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, pp. 1–9, 2024.