1. Implementation Details.

We adopt a ResNet-50 [1] as a backbone and a feature pyramid network (FPN) [2]. For meta-training, we follow the standard training schedule of RetinaNet [3]. We train the model using a stochastic gradient descent (SGD) for 90k iterations with a batch size of 16 and a learning rate of 0.01 decayed twice at 50k and 80k iterations by a factor of 10. Moreover, we use a weight decay of 0.0001 and momentum of 0.9. We meta-train in a 5-way-5-shot manner, setting \( N = 5 \) in the proposed MWST algorithm, resulting in a total of 25 shots per task. Horizontal image flipping for the query image is utilized as the only data augmentation technique during meta-training. As for meta-testing, we set the number of iterations to 6000 with a learning rate of 0.005 decayed by a factor of 10 at iteration 4000. Moreover, to exploit the proposed MWST, we set \( N = 15 \). All experiments are conducted using 4 Nvidia Tesla V100 GPUs.

2. Ablation study

2.1. Effect of Focal Loss Parameters

In Table 1, we analyze the effect of the focal loss hyperparameters during the meta-training phase. We show the results for 4 different settings of the \( \alpha \) and \( \gamma \) hyperparameters. Using the default parameters of RetinaNet leads to divergence during the training. Higher values for \( \alpha \) and \( \gamma \) are needed because there are fewer positive anchors in FSOD than the general OD task due to the smaller number of bounding boxes in the query image. This problem has been partially alleviated by the MWST but is found to be further alleviated by tuning the focal loss.

2.2. Effect of the Number of Anchors

In Table 2, the performance of our meta-detector for different number of anchors trained only using \( L_{loc} \) and \( L_{cls} \) (without \( L_{margin} \)) is shown. The results highlight the importance of the anchor density for meta-detectors. Although the number and size of anchors have always played an important role in dense object detectors [3], their effect becomes more pronounced in FSOD. In our design, we find that by increasing the number of the anchors, the base and novel performance improve. We argue that a greater number of anchors provides a stronger learning signal to the post-fusion network, refining the instance-level features. We notice that the training becomes unstable when increasing the number of anchors beyond 15, arguably because the model capacity becomes smaller for the task.

3. Qualitative Results

In Figure 1 we present different success and failure cases of our proposed FSRN model in a 10-shot setting on MS-COCO dataset.

---

*Both authors have contributed equally to this work

karim.guirguis@de.bosch.com
Figure 1: Qualitative results for the proposed FSRN model in 10-shot setting on MS-COCO dataset.
References

