Supplementary Materials: EPL-UFLSID: Efficient Pseudo Labels-Driven Underwater Forward-Looking Sonar Images Object Detection

1 Supplementary Materials

1.1 Datasets and Implementation Details

The effectiveness of the proposed method is demonstrated on two main forward-looking sonar object detection datasets, marine-debris-fls-dataset (MDFD) [9] and the underwater acoustic target detection (UATD) [12] dataset.

MDFD is captured by Ocean Systems Lab (Heriot-Watt University) using a ARIS Explorer 3000 forward-looking sonar at 3.0 MHz frequency, which comprises 1868 sonar images across 11 categories of marine debris, including bottles, cans, chains, drink-cartoons, hooks, propellers, shampoo-bottles, standing-bottles, tires and valves. UATD is captured by Tritech 1200ik forward-looking sonar in Weihai, China, which comprises 9200 sonar images of 10 categories, including cubes, balls, cylinders, human bodies, tyres, circle cages, square cages, metal buckets, planes and rovs.

In terms of MDFD dataset, we've allocated 1681 images for training and 187 for testing object detection models. With regard to UATD dataset, we've designated 8280 images for training and 920 for testing object detection models.

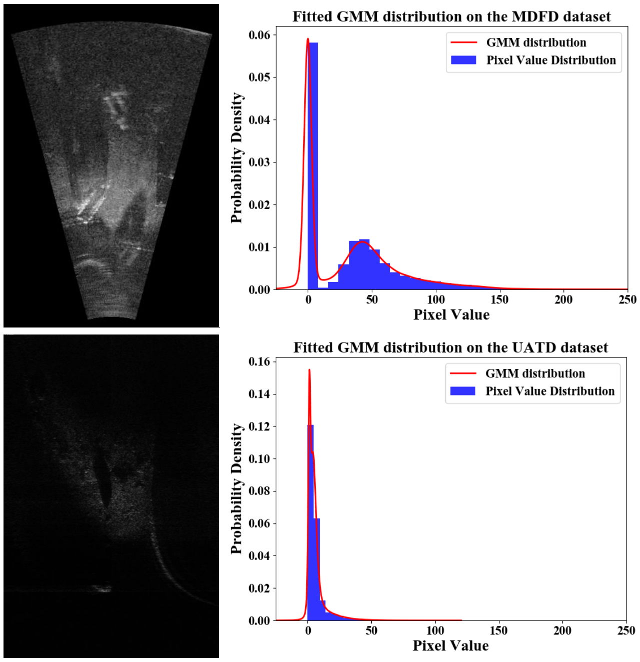
Experiments are conducted in Pytorch 1.2.0 on two NVIDIA GeForce RTX 2080 SUPER GPUs with 8GB memory. GMMDIP is trained via Adam optimizer for 900 iterations per image at an initial learning rate of 1e-1. DFIQA undergoes 200 epochs of training with Adam optimizer, starting at a learning rate of 2e-5, 5e-4 weight decay, and a batch size of 16. EPL-UFLSID is trained for 50 epochs using the Adam optimizer, with an initial learning rate of 1e-4 and a batch size of 4. Subsequently, an additional 50 epochs are conducted with an initial learning rate of 1e-5 and a batch size of 2. The overall training process includes a linear learning rate decay of 0.96.

Table 1: Computational Overhead Comparison

Method	Params(M)	FLOPs(G)
SSD [4]	24.9	121.5
YOLOv3 [6]	61.6	77.7
YOLOv5 [2]	7.1	8.3
YOLOv7 [10]	37.2	52.6
MBSNN [11]	8.2	9.9
Faster R-CNN [7]	136.9	184.9
CenterNet [1]	32.7	49.2
RetinaNet [3]	36.5	74.2
UFIDNet [5]	34.3	71.3
EPL-UFLSID	37.1	76.1

1.2 Statistical distribution on the MDFD and UATD datasets

To further exemplify the performance of Gaussian Mixture Model (GMM) [8] in fitting the distribution of the original sonar image, two images along with their statistical distributions are provided in Figure 1, which shows that GMM basically fits the distribution characteristics of the original sonar image. Therefore, GMM can be reasonably selected as the input of GMMDIP to better denoise the original sonar image.



(a) Original Image **(b) Statistical Distribution Image**

Figure 1: Fitted GMM distribution to the original image of pixel value distribution on the MDFD and UATD datasets.

1.3 Computational Efficiency

Table 1 shows the computational overhead on a (3*600*600) image of the proposed EPL-UFLSID and other 9 object detection methods. EPL-UFLSID achieves the best performance with high computing efficiency and moderte memory consumption. It needs to be mentioned that although EPL-UFLSID training involves the introduction of pseudo labels selected by DFIQA as additional supervision information, the inference stage only requires the detection backbone (ResNet 50), adding no extra computational burden. Therefore, EPL-UFLSID increases Params and FLOPs by only 0.6M and 1.9G compared to the baseline (RetinaNet), which further demonstrates that the performance advantage of EPL-UFLSID is not due to the increase in computational complexity, but rather because we select the pseudo-labels that are most detection-friendly to constrain the detection network to extract clean features.

1.4 Reconstructed Module of EPL-UFLSID

In order to make the detection backbone network optimized by pseudo labels, we consider reconstructing the backbone features into images through the reconstructed module, which is depicted in Figure 2. Since the first three layers of the backbone network have rich spatial and semantic information, which helps us reconstruct images more easily, we choose to fuse the features of these three layers through convolutional layers and average pooling layers, and then upsample them to obtain a reconstructed image of the same size as the pseudo-label, so as to achieve the effect of additional constrained optimization of the detection network.

In other words, the reconstructed module is served as a medium to allow the pseudo label to optimize the entire detection network with the MSE loss and detection loss, enabling the backbone network to extract features of the original input image that are submerged by noises.

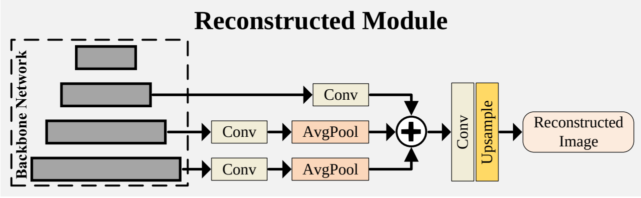


Figure 2: The structure of Reconstructed Module.

1.5 Visualization of pseudo labels generated by GMMDIP with different inputs at different iterations on the MDFD and UATD datasets

To further illustrate the superior performance of GMMDIP, Figure 3 shows denoised sonar images generated by GMMDIP with different inputs at different iterations on the MDFD and UATD datasets. It can be found that the denoising effect of the denoised image with GMM input is significantly better than the denoised image with uniform noise input, especially at 200,300,400 iterations, which is due to the fact that fitting a Gaussian mixture model of the original image allows GMMDIP to obtain the general characteristics of the distribution and structure of the original image, achieving bettter quality of details. Consequently, DFIQA can more effectively select the most detection-friendly pseudo labels generated by GMMDIP for EPL-UFLSID, resulting in better detection performance.

1.6 Qualitative comparison of denoised images on the MDFD and UATD datasets

Qualitative comparison of the selected denoised images by DFIQA and the denoising method in UFIDNet [5] is shown in Figure 4. It needs to be noted that UFIDNet denoises the sonar images by characterizing the noise of sonar images as multiplicative speckle noise, and sets them as pseudo labels to enhance the detection performance. They simply model the sonar image noises like speckle noise with a known and simple distribution. That is to say that the denoising effect of the denoising method in UFIDNet is simplex, not as diverse as the denoising images selected by DFIQA at different iterations, which are more tailored for detection. At the same time, we can find that the pseudo labels selected by DFIQA are not as visually good as those of UFIDNet, as shown in the first row of column (b) in Figure 4. This further illustrates that the pseudo labels selected by DFIQA are targeted at machine vision and are therefore more friendly to object detection tasks.

References

[1] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6569–6578.

[2] Glenn Jocher. 2022. *ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation.* https://doi.org/10.5281/zenodo.7347926

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 21–37.

[5] Hui Long, Liquan Shen, Zhengyong Wang, and Jinbo Chen. 2023. Underwater Forward-Looking Sonar Images Target Detection via Speckle Reduction and Scene Prior. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–13.

[6] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[8] Douglas A Reynolds et al. 2009. Gaussian mixture models. *Encyclopedia of biometrics* 741, 659-663 (2009).

[9] Deepak Singh and Matias Valdenegro-Toro. 2021. The marine debris dataset for forward-looking sonar semantic segmentation. In *Proceedings of the ieee/cvf international conference on computer vision*. 3741–3749.

[10] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7464–7475.

[11] Jianjun Wang, Chen Feng, Lingyu Wang, Guangliang Li, and Bo He. 2022. Detection of weak and small targets in forward-looking sonar image using multi-branch shuttle neural network. *IEEE Sensors Journal* 22, 7 (2022), 6772–6783.

[12] Kaibing Xie, Jian Yang, and Kang Qiu. 2022. A dataset with multibeam forward-looking sonar for underwater object detection. *Scientific Data* 9, 1 (2022), 739.

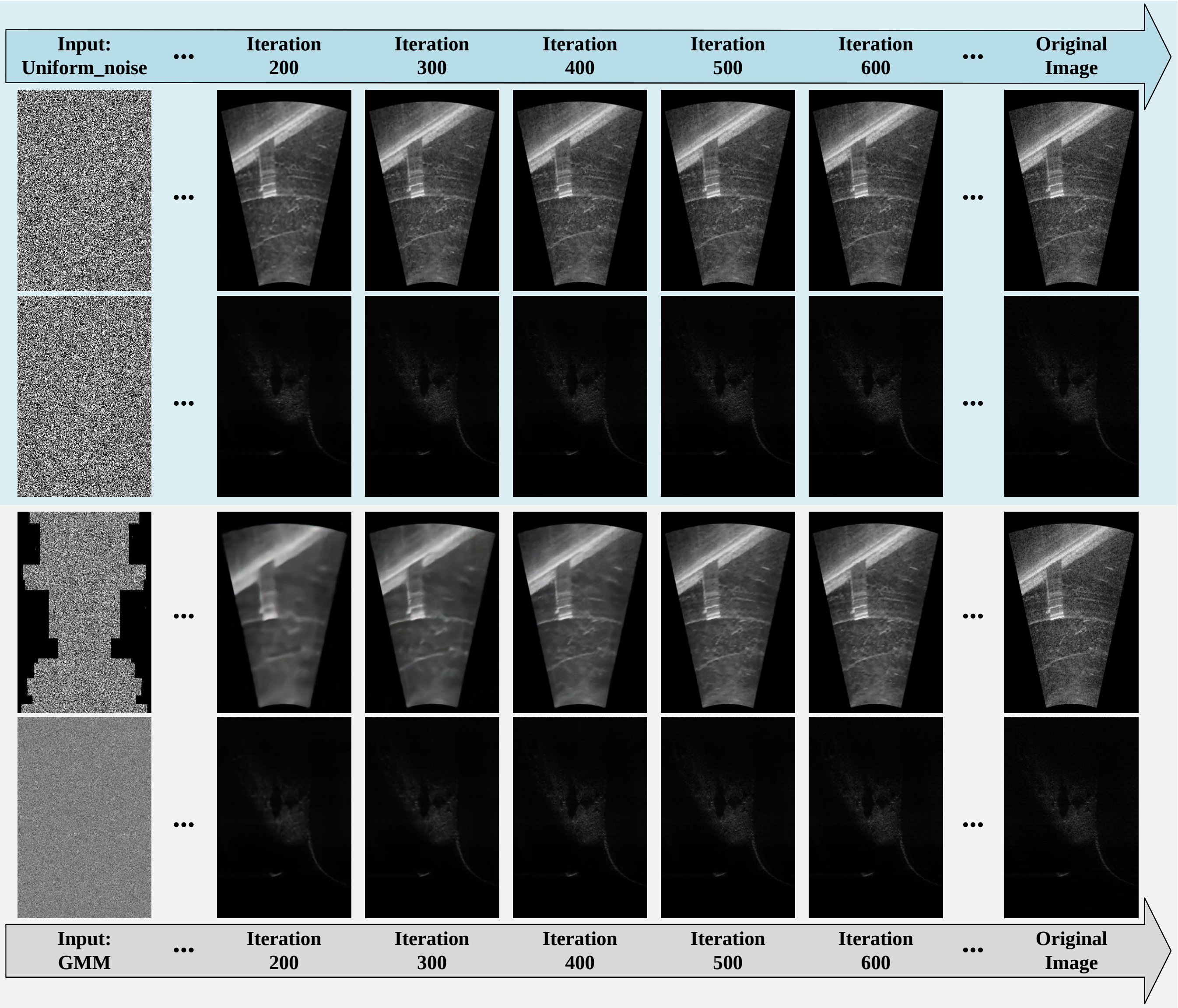


Figure 3: : Pseudo labels generated by GMMDIP with different inputs at different iterations on the MDFD and UATD datasets. The image is divided into two parts, the top half with blue background is the denoised image generated by GMMDIP when the input is uniform noise, while the bottom half with gray background is the denoised image generated by GMMDIP when the input is GMM.

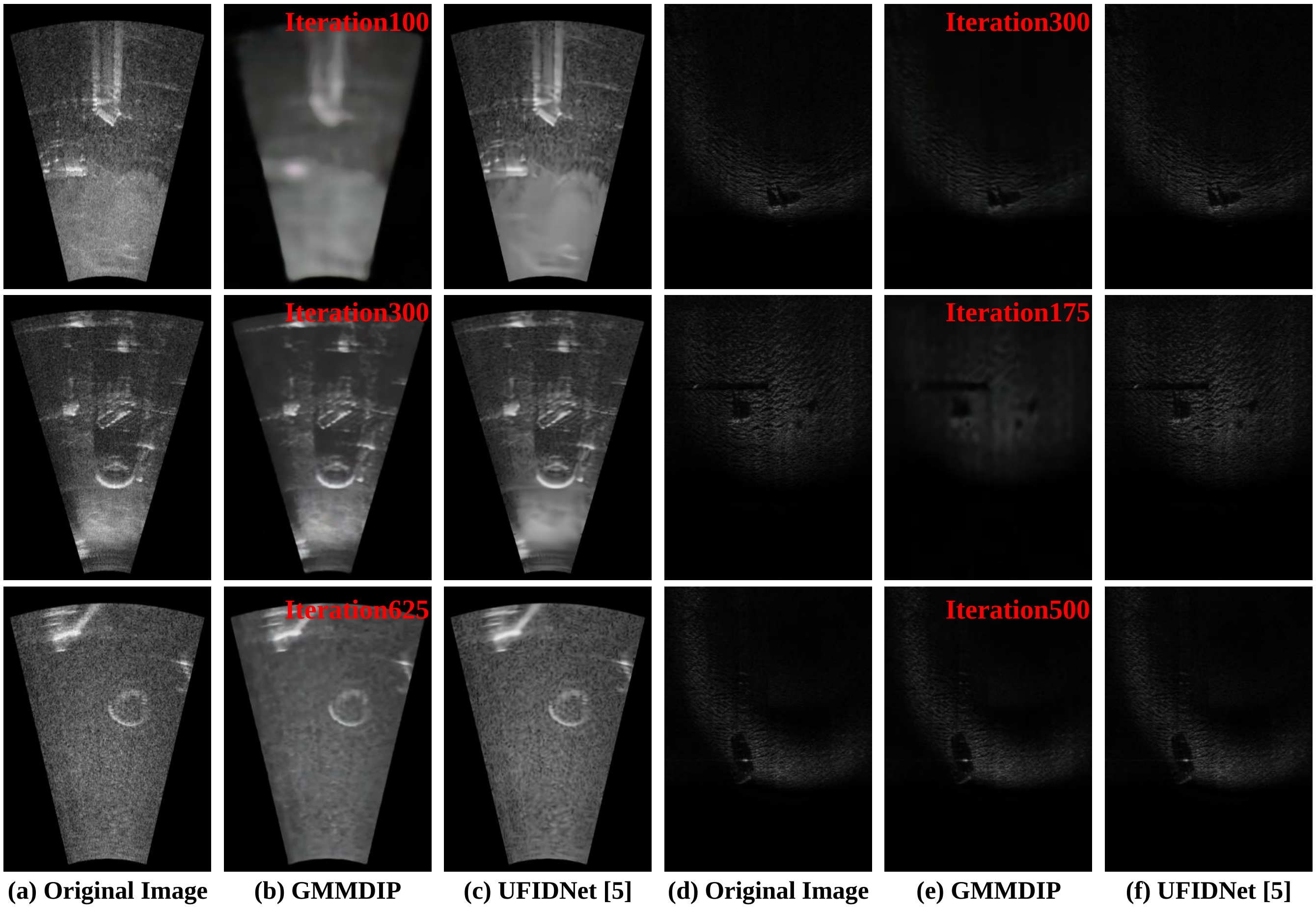


Figure 4: Qualitative comparison of denoised images generated by GMMDIP with GMM input and the denoising method in UFIDNet. Columns (a), (b), (c) represent the images on the MDFD dataset. Columns (d), (e), (f) represent the images on the UATD dataset. The denoised images from columns (b) and (d) have IterationX displayed in the upper right corner, representing the iteration values of the efficient pseudo labels selected by DFIQA from iteration 0 to iteration 900.