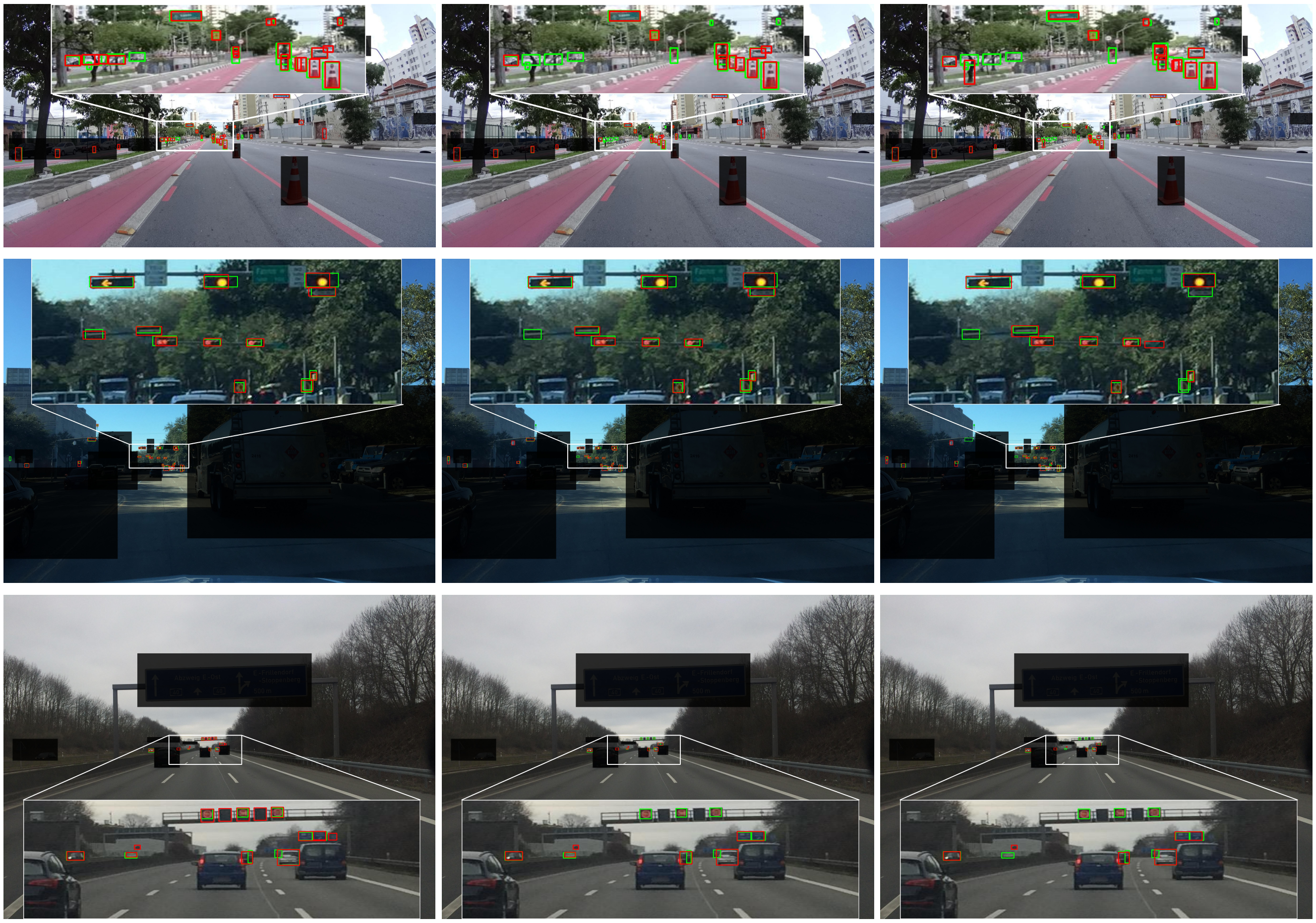
DM-EFS: Dynamically Multiplexed Expanded Features Set Form for Robust and Efficient Small Object Detection (Supplementary Paper)

Aashish Sharma
KLASS Engineering and Solutions, Singapore
sharma.aashish@klasses.com.sg



DM-EFS (Ours) YOLOv7 (Base) CFINet [12]

Figure 1. Results on the SODA-D dataset. We can see that our DM-EFS outperforms both the YOLOv7 base model and recent SOD method CFINet. Note, green and red boxes are ground-truths and predictions respectively, and 'ignore' regions are masked out. All the results are shown without class labels for ease in visualization and performance comparisons.

1. Performance Comparisons

1.0.1. SODA-D Dataset

In the main paper, Fig. 5 shows an example of our qualitative results on the SODA-D [2] dataset (for quantitative comparisons on the SODA-D dataset, please refer to Table 1 in the main paper). We now show additional qualitative results and comparisons with SOD baseline method CFINet [12] and our base model YOLOv7 [10]. The results (without class labels for ease in visualization) are shown in Fig. 1 which clearly highlight the better performance of our DM-EFS method over the baselines.

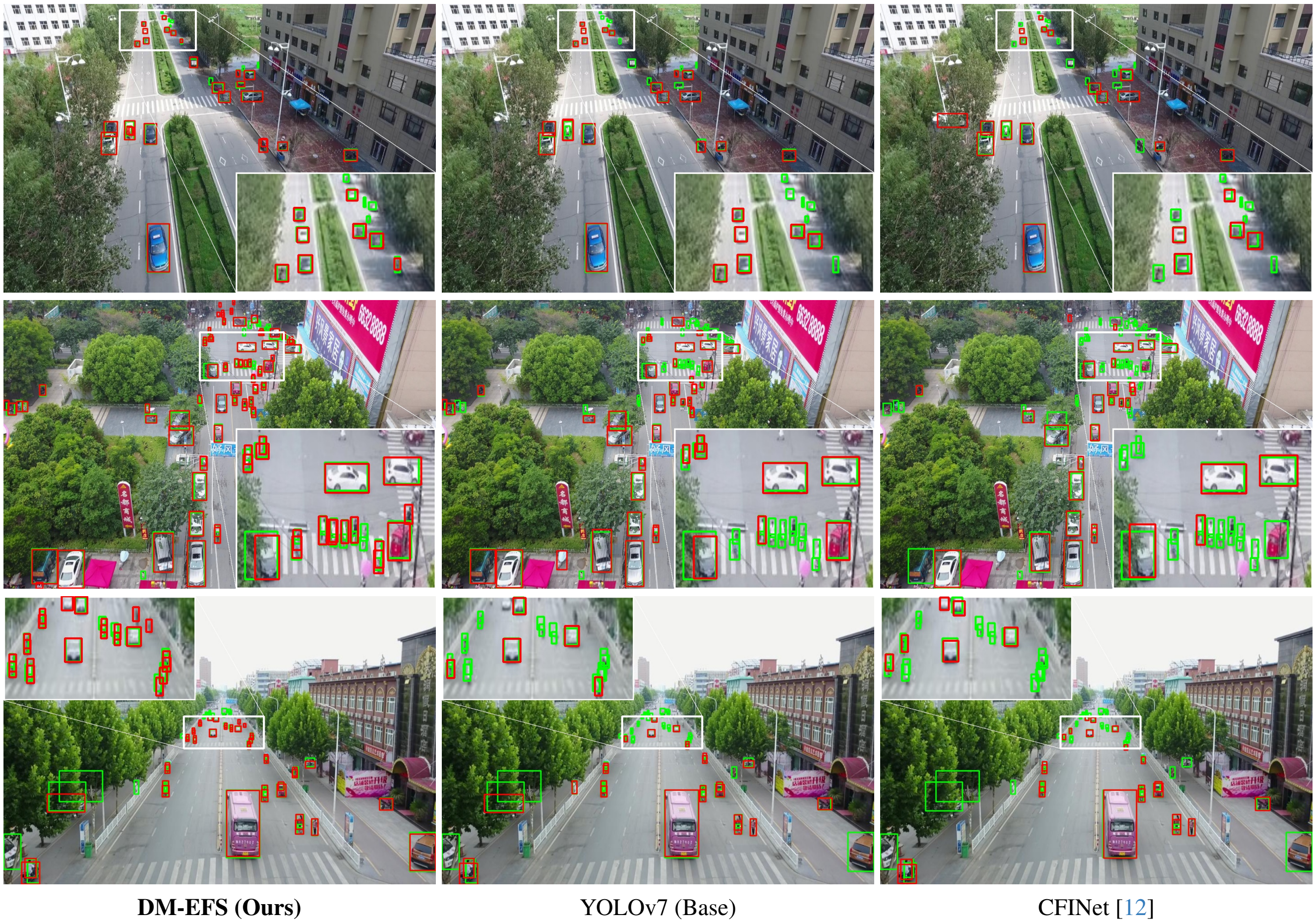


Figure 2. Results on the VisDrone dataset. We can see that our DM-EFS outperforms both the YOLOv7 base model and recent SOD method CFINet in detecting small objects robustly. Note, green and red boxes are ground-truths and predictions respectively. The results are for 640×640 images but are shown on original images without class labels for ease in visualization.

Method	Venue	AP↑	AP_{50}↑	AP_{75}↑	AP_{vt}↑	AP_{t}↑	AP_{s}↑	AP_{m}↑
Faster-RCNN [8]	NIPS'15	13.24	23.41	13.51	00.01	00.04	07.44	26.67
RetinaNet [6]	ICCV'17	08.70	15.51	08.90	00.11	00.30	02.61	15.14
Sparse-RCNN [9]	CVPR'21	06.85	12.90	06.11	00.21	01.14	04.34	11.49
CFINet [12]	ICCV'23	14.60	27.80	14.01	00.20	02.55	11.66	25.47
ESOD [7]	TIP'24	23.61	41.89	23.38	-	-	-	-
YOLOv7 (Base)	CVPR'23	27.60	47.81	27.30	01.40	08.11	25.42	43.63
DM-EFS (Ours)	–	**29.71**	**51.80**	**29.30**	**02.94**	**10.88**	**28.54**	**44.49**

Table 1. SOD performance comparisons on the VisDrone dataset using additional metrics, AP_{vt}, AP_{t}, AP_{s}, and AP_{m}, customized for small object detection. The results are generated for 640×640 resolution input images. '↑' means higher is better. '-' means results not available.

1.0.2. VisDrone Dataset

For the VisDrone [14] dataset, Fig. 6 (bottom-row) in the main paper shows an example of qualitative comparisons between our DM-EFS and baselines methods, YOLOv7 base model and recent SOD method CFINet. We now show additional qualitative comparisons in Fig. 2, and the results again show our better performance over the baselines for detecting small objects.

In the main paper, we also show quantitative comparisons in Table. 2 using AP, AP_{50}, and AP_{75} metrics. We now show additional comparisons using additional metrics, namely AP_{vt}, AP_{t}, AP_{s}, AP_{m} which are AP metrics computed for *very tiny* (size $\in [2, 8)$), *tiny* (size $\in [8, 16)$), *small* (size $\in [16, 32)$), and *medium* (size $\in [32, 64)$) sized objects

Method	Venue	'pedestrian'	'car'	'tricycle'	'motor'	'people'	'van'	'awning-tricycle'	'bicycle'	'truck'	'bus'
Faster-RCNN [8]	NIPS'15	06.11	33.51	08.54	07.01	04.39	19.12	05.33	03.49	16.82	27.84
RetinaNet [6]	ICCV'17	04.41	30.19	03.44	04.10	02.65	11.71	02.37	00.90	11.71	15.60
Sparse-RCNN [9]	CVPR'21	03.91	19.93	03.23	04.54	03.77	09.45	02.66	01.54	06.59	10.81
CFINet [12]	ICCV'23	11.61	39.74	08.88	10.40	07.62	19.97	05.11	04.27	14.20	24.45
ESOD [7]	TIP'24	18.81	53.30	15.81	18.70	11.62	31.41	09.39	06.37	26.41	44.72
YOLOv7 (Base)	CVPR'23	20.91	55.12	21.82	22.70	15.91	36.22	12.50	11.72	32.27	46.86
DM-EFS (Ours)	–	**24.60**	**57.33**	**23.21**	**25.84**	**17.95**	**37.84**	**13.33**	**14.19**	**33.04**	**50.64**

Table 2. Class-wise SOD performance comparisons in terms of AP on the VisDrone dataset with ten classes (divided into two tables as shown above with five classes each). The results are generated for 640×640 resolution input images.

Method	Venue	$\text{AP}_{50}\uparrow$	fps$\uparrow$
Deformable-DETR [15]	ICLR'21	48.07	14.51
RT-DETR [13]	CVPR'24	49.79	**42.04**
YOLOv7 (Base)	CVPR'23	47.81	39.71
DM-EFS (Ours)	–	**51.80**	38.24

Table 3. Comparisons with transformer-based detection methods on the VisDrone validation set with 640×640 input images.

respectively. The results are shown in Table 1, which support the qualitative results showing the better performance our DM-EFS over the baselines. We also compute class-wise AP metrics for the VisDrone dataset, shown in Table 2. As the results show, for all the classes, our DM-EFS generates the best SOD results among all the methods. While our DM-EFS is CNN-based using YOLOv7 base detection model (by default), we also compare it with transformer-based general detection models, Deformable-DETR [15] and RT-DETR [13], as transformer-based detection models provide robust and competitive performance. As shown in Table 3, our DM-EFS still provides state-of-the-art results compared to the recent transformer-based methods (while being slightly inferior in terms of fps performance.)

1.0.3. DarkFace Dataset

For our SOD performance comparisons, in the main paper, we also use the DarkFace [11] dataset. It is a suitable candidate for testing diverse SOD conditions as it consists of low-light images for face detection (i.e., only face class) with pre-dominantly tiny faces. This can be observed from the object size histograms for the DarkFace dataset for 640×640 images shown in Fig. 4, which shows an average object size of 11.09px. For reference, the object size histograms for the VisDrone SOD dataset for 640×640 images are shown in Fig. 5, which shows an average object size of 18.97px. This means that the average object size of the DarkFace dataset is even smaller than that of the VisDrone dataset, making it a suitable choice for performance comparisons on a diverse SOD dataset.

In the main paper, Fig. 6 (top-row) and Table 4 show the qualitative and quantitative comparisons for the DarkFace dataset. We now show some additional qualitative results in Fig. 7. As observed from the results in the main paper, the results in Fig. 7 also highlight that our DM-EFS can handle diverse SOD conditions, and is able to detect small faces quite more robustly than compared to the baseline methods.

2. Additional Details

2.0.1. Visualization of Shallow Features

Fig. 3 visualizes the backbone features in our DM-EFS, and we can observe that with shallow features included (which basically is our proposed EFS form), more small objects are captured in the backbone features as shallow features though abstract, are of higher resolution which allows more feature representation capability for small objects.

2.0.2. DarkFace Training

Following [3], before training on the DarkFace dataset, we first pre-train our model on synthetic low-light COCO [5] dataset, which is obtained using the low-light transformation process proposed in [1] to render low-light COCO images from normal well-lit COCO images. Note that, during

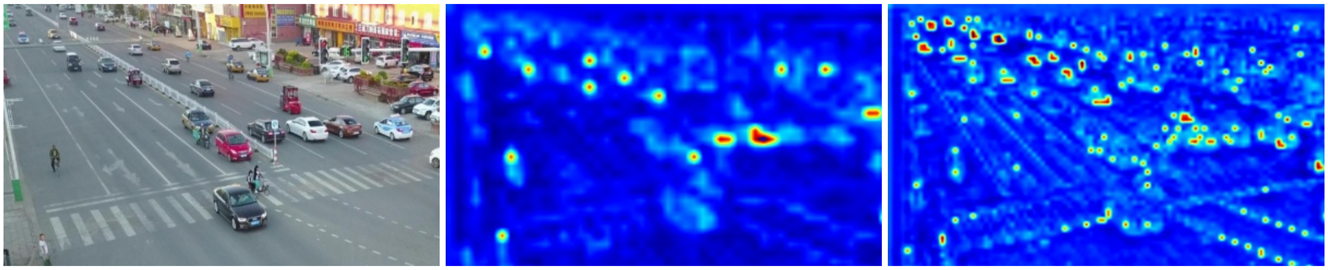


Figure 3. For the input image (left) taken from the VisDrone dataset, backbone features in our DM-EFS without (middle) and with (right) shallow features included (which essentially means our EFS form).

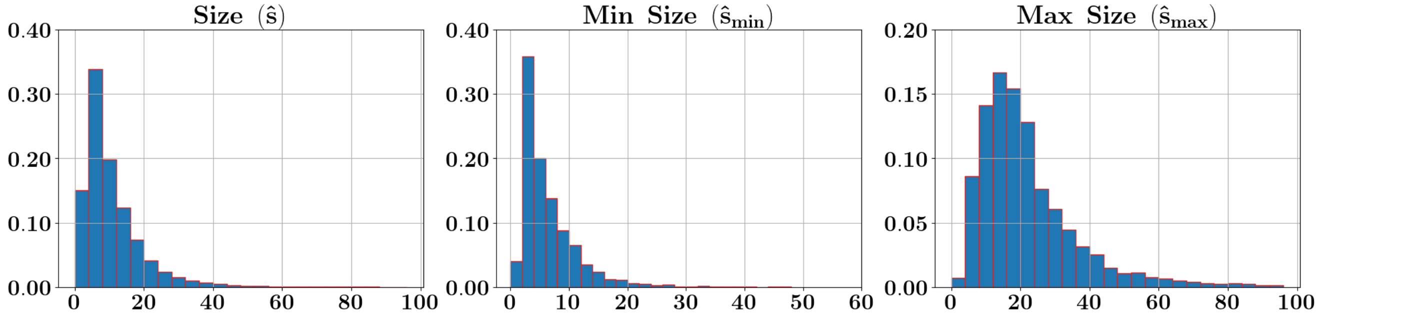


Figure 4. From left-to-right, normalized histograms showing distributions of true sizes $\hat{s}$, true min sizes $\hat{s}_{\text{min}}$, and true max sizes $\hat{s}_{\text{max}}$ for the DarkFace training data. Note that, the histograms are obtained for the input resolution of 640×640.

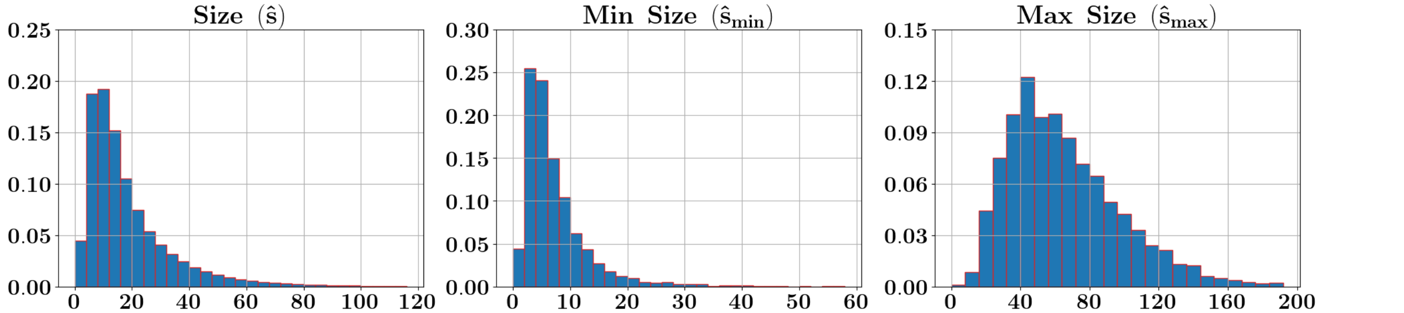


Figure 5. From left-to-right, normalized histograms showing distributions of true sizes $\hat{s}$, true min sizes $\hat{s}_{\text{min}}$, and true max sizes $\hat{s}_{\text{max}}$ for the VisDrone training data. Note that, the histograms are obtained for the input resolution of 640×640.

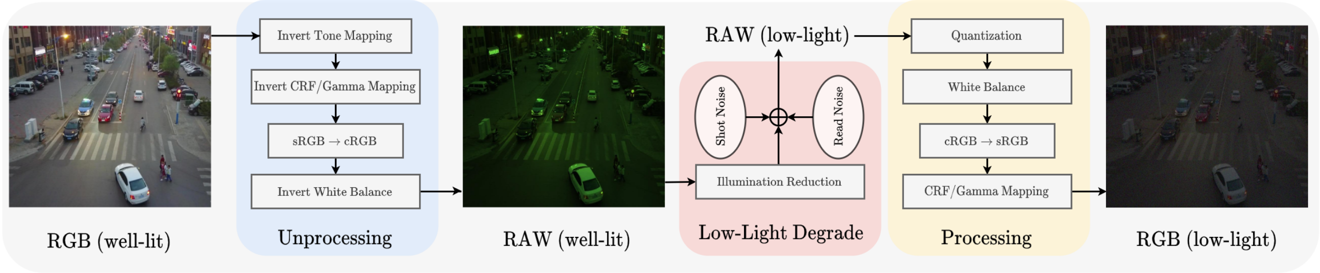


Figure 6. Illustration of the low-light transformation process [1] employed in generating synthetic low-light COCO dataset that is used in the pre-training involved in the DarkFace dataset training scheme [3].

the pre-training step, for ease in the training process, we disable our DFM design and set all the control signals for the neck and head multiplexers to 1. We also disable learning min and max object sizes by setting λ_{szp} to 0 which controls the corresponding loss term. After pre-training, in the main training step, we then follow our proposed DM-EFS

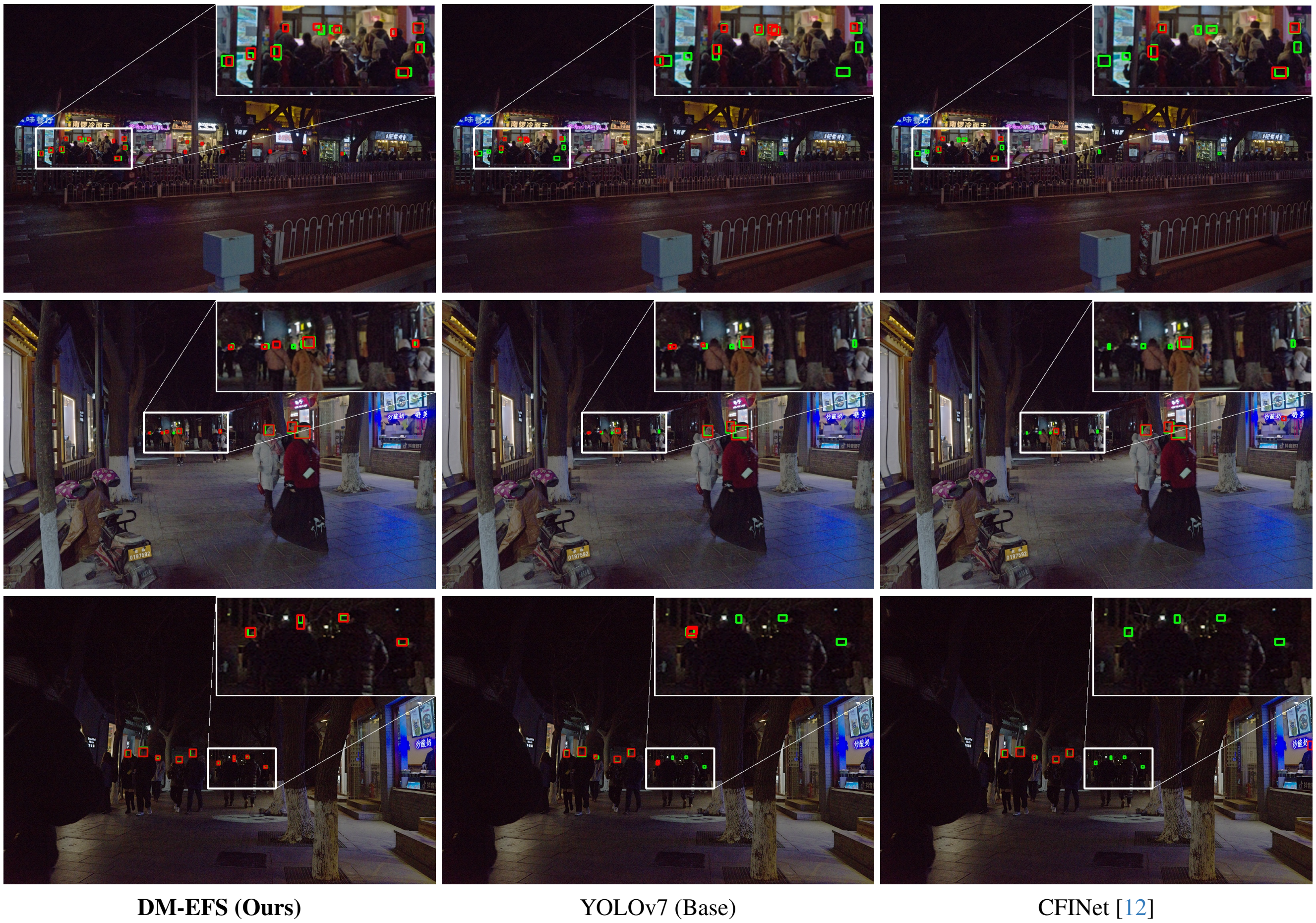


Figure 7. Results on the DarkFace dataset. We can see that our DM-EFS outperforms both the YOLOv7 base model and recent SOD method CFINet in detecting small faces robustly. Note, green and red boxes are ground-truths and predictions respectively. The results are for 640×640 images but are shown on original images for ease in visualization.

training procedure normally.

The process is also illustrated in Fig. 6 for reference. There are primarily three steps: (1) 'Unprocessing' the normal well-lit RGB images to their corresponding RAW form, (2) Low-light degradation of the RAW images (obtained in the previous step) via illumination reduction and non-uniform rrandom noise addition, and (3) 'Processing' the degraded low-light RAW images back to RGB form (see [1, 3] for further details).

2.0.3. Regression based Min-Max Sizes Prediction

As described in the main paper, to learn min and max object sizes prediction, we cast them as single label classification problems, and predict them with a two-branch classification network in the control module Φ_C (see Fig. 3 in the main paper showing our model architecture).

We now discuss an alternate way to cast them as regression problems. For this, we simply change the output size of the last FC layer of the two-branch classification network to 1 (with sigmoid activation), and re-formulate the min-max object size prediction loss $\mathcal{L}_{\text{szp}}$ to perform regression by:

$$\mathcal{L}_{\text{szp}} = \lambda_{\text{szp}} \Big[\|s^*_{\text{min}} - \hat{s}^*_{\text{min}}\|^2 + \|s^*_{\text{max}} - \hat{s}^*_{\text{max}}\|^2 \Big], \quad (1)$$

where $\hat{s}^*_{\text{min}}$ and $\hat{s}^*_{\text{max}}$ are normalized values of true min size and true max size, $\hat{s}_{\text{min}}$ and $\hat{s}_{\text{max}}$, obtained by $\hat{s}^*_{\text{min}} = \hat{s}_{\text{min}}/\bar{s}$ and $\hat{s}^*_{\text{max}} = \hat{s}_{\text{max}}/\bar{s}$ respectively, where $\bar{s}$ is the max object size parameter. Therefore, both $\hat{s}^*_{\text{min}}$ and $\hat{s}^*_{\text{max}}$ are in $[0, 1]$, which means that with the loss re-formulation described above, we also predict normalized min and max object sizes s^*_{min} and s^*_{max} in the range of $[0, 1]$. Hence, we obtain the final predicted min and max object sizes, s_{min} and s_{max}, by simply de-normalizing them, i.e., $s_{\text{min}} = s^*_{\text{min}} \cdot \bar{s}$ and $s_{\text{max}} = s^*_{\text{max}} \cdot \bar{s}$.

With this above regression based design, the overall computational complexity is lesser than compared to the classification based design (which potentially can provide a higher inference fps speed to our DM-EFS). However, in all our trails, we found the above regression based design to be less effective and accurate than the proposed classification based design. This is because for SOD datasets, majority of the objects are of very small size say 10-20px (as can be seen in Figs. 4 and 5), which when normalized, become extremely small values (close to 0). Since the regression based loss described above is affected by the magnitude of object sizes, this creates a mode collapse (local minima) like situation when majority of the min-max sizes predictions made by our model are close to 0. Using a classification based design bypasses this issue since it is not affected by the magnitude of object sizes. Furthermore, using Focal Loss [6] based classification helps in addressing the class-imbalance problem created by long-tail distributions of true min and max object sizes (see Figs. 4 and 5).

Hence, to sum up, for learning min and max object sizes prediction, we propose to use a classification based design − which even though has a higher computational cost than the regression based design − is more effective in learning min and max object sizes correctly.

2.0.4. Limitations and Future Wok

While our DM-EFS has achieved SOTA performance on various SOD benchmarks, there still remains a few issues. For e.g., handling more difficult small objects such as of very small size (≤8px) or with extreme inter-class similarities [4] remains a challenge, since they may not be adequately represented in the shallow features that contributes to our SOD performance.

Our DM-EFS also employs a control module that has its own inference overhead. For e.g., on the VisDrone dataset, we obtain 38.24 inference fps or 26.15 ms/image runtime, in which the control module takes 1.79 ms/image (and, the rest is taken by the detection modules). While the added overhead from the control module is relatively small, improving our overall inference fps is another area of improvement, and addressing this will be part of our future work.

References

[1] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4, 5

[2] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[3] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2553–2562, 2021. 3, 4, 5

[4] Yecheng Huang, Jiaxin Chen, Di Huang, and NA NA. Ufpmp-det: Toward accurate and efficient object detection on drone imagery. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1026–1033, 2022. 6

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 3, 6

[7] Kai Liu, Zhihang Fu, Sheng Jin, Ze Chen, Fan Zhou, Rongxin Jiang, Yaowu Chen, and Jieping Ye. Esod: Efficient small object detection on high-resolution images. *IEEE Transactions on Image Processing*, 2024. 2, 3

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2, 3

[9] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021. 2, 3

[10] Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao, and NA NA. 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*, pages 7464–7475, 2023. 1

[11] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, and et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020. 3

[12] Xiang Yuan, Gong Cheng, Kebing Yan, Qinghua Zeng, and Junwei Han. Small object detection via coarse-to-fine proposal generation and imitation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6317–6327, 2023. 1, 2, 3, 5

[13] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024. 3

[14] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021. 2

[15] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3