Supplementary Materials: FeatEnHancer: Enhancing Hierarchical Features for Object Detection and Beyond Under Low-Light Vision

Khurram Azeem Hashmi[1,2] Goutham Kallempudi[2] Didier Stricker[1,2] Muhammamd Zeshan Afzal[1,2]

[1]DFKI - German Research Center for Artificial Intelligence
[2]RPTU Kaiserslautern

Overview: We first provide complete implementation details of our experiments on dark object detection in Appendix 1.1, face detection in Appendix 1.2, semantic segmentation in Appendix 1.3, and video object detection in Appendix 1.4. Then, we present the quantitative and qualitative analysis in Appendix 2.1, 2.2, and 2.3. We compare and visualize the learned features between RetinaNet and Featurized Query R-CNN in Appendix 2.4. We discuss the performance in terms of learned representation and predictions between our intra-scale feature enhancement network and the DCENet [12, 18] in Appendix 2.5. The performance comparison between our FeatEnHancer and unsupervised domain adaptation methods is provided in Appendix 2.6. Finally, in Appendix 2.7, we discuss the performance-efficiency trade-off between our FeatEnHancer, Low-Light Image Enhancement (LLIE) approaches, and task-specific state-of-the-art methods.

1. Implementation Details

1.1. Dark Object Detection

For dark object detection experiments on real-world data, we adopt RetinaNet [19] as a typical detector and Featurized Query R-CNN [44] (FQ R-CNN) as an advanced object detection framework to report results. The implementation of FeatEnHancer with FQ R-CNN is based on detectron2 [36] with ResNet-101 [14] as the backbone network pre-trained with COCO [20] weights. For the training, a batch size of 8 is employed with all the images resized to a maximum scale of 1333 $\times$ 800. Our training follows a $50K$ scheduler using ADAMW [23] optimizer with an initial learning rate set to 0.0000025, which is divided by 10 both at 42000 and 47000 iterations. All experiments are carried out on RTX-2080 Ti GPU.

For RetinaNet, images are resized to 640$\times$640, and we train the network using 1$\times$schedule[1] in mmdetection [2] (12 epochs using SGD optimizer [28] with an initial learning rate of 0.001). For evaluation, along with the common practice of employing mAP@IoU=0.5, we report mAP@IoU=0.5:95 using [20] for completeness. Note that for each object detection framework, we adopt the same settings while reproducing results of our work, baseline, LLIE approaches, and task-specific state-of-the-art methods.

We compare our FeatEnHancer to several state-of-the-art LLIE methods, including KIND [46], RAUS [27], EnGAN [17], MBLLEN [10], Zero-DCE [12], Zero-DCE++ [12], and state-of-the-art dark object detection method, MAET [6]. For LLIE methods, all images are enhanced from their released checkpoints and propagated to the detector. In case of MAET [6], we pre-train the detector using their proposed degrading pipeline and then fine-tune it on both datasets to establish a direct comparison.

1.2. Face Detection

The DARK FACE [41] dataset comprises dark human faces of various sizes in low-illuminated environments. In order to capture the tiny human faces, the images are resized to a maximum of 1333 $\times$ 800 and 1500 $\times$ 1000 for FQ R-CNN [44] and RetinaNet [19], respectively. The other settings and hyperparameters are identical to the Dark Object detection experiments explained in Appendix. 1.1.

1.3. Semantic Segmentation

The semantic segmentation with FeatEnHancer is implemented using MMSegmentation [4], where the images are resized to 2048 $\times$ 1024 for the training. For direct comparison with previous state-of-the-art method [39], we use DeepLabV3+ [3], an encoder-decoder style segmentor

[1] https://github.com/open-mmlab/mmdetection/blob/master/configs/_base_/schedules/schedule_1x.py

with ResNet-50 [14] as the backbone for nighttime semantic segmentation. The backbone is initialized with pre-trained ImageNet [8] weights, and we use a batch size of 4 for the training. The SGD [28] optimizer following a $20K$ scheduler[2] of MMSegmentation [4] is employed with a base learning rate of 0.01 and a weight decay of 0.0005. For the direct comparison with the LLIE methods, we enhance all the images before passing them to the segmentor as done in Appendix. 1.1. The Mean Intersection over Inion (mIoU) is used to report the segmentation results in comparison to the baseline, LLIE approaches and existing state-of-the-art method Xue *et al*. [39].

1.4. Video Object Detection

Besides object detection and semantic segmentation on images, we extend our experiments to the video domain to test the generalization capabilities of the proposed FeatEnHancer. The video object detection under low-light vision is evaluated on the recently emerged DarkVision dataset [42] (see Table 1 in the main paper). Although the dataset is not publicly available yet, we sincerely thank the authors of [42] for providing prompt access to its subset. To evaluate our FeatEnHancer under the low-light setting, we take the low-end camera split on two different illumination levels, i.e., 0.2 and 3.2. For ablation studies, we adopt a 3.2% illumination level split.

We consider SELSA [35] as our baseline due to its simple and effective design and impressive performance on video object detection benchmarks [29, 7]. As the backbone network, we use ResNet-50 [14] pre-trained on ImageNet [8]. For the detection, we apply Region Proposal Network (RPN) [26] on the output of *conv4* to generate candidate proposals on each frame. In RPN, a total of 12 anchors with four scales $\{4, 8, 16, 32\}$ and three aspect ratios $\{0.5, 1.0, 2.0\}$ are used. The final 300 proposals are selected from each frame. In summary, we follow identical experimental settings by following the config of $1\times$ schedule[3] in mmtracking [5].

To establish a direct comparison, we enhance all video frames first through LLIE methods and feed these frames to the baseline, as done in Appendix. 1.1. For our method, we integrate FeatEnHancer in the baseline in an end-to-end fashion (see Fig. 2 in the main paper). Following standard practice in video object detection [11, 13, 35], the mAP@IoU=0.5 is utilized as an evaluation metric to report results. *Note that the goal of this experiment is not to surpass prior state-of-the-art results on dark video object detection [42]. Instead, the target is to demonstrate the effectiveness and generalization capabilities of the proposed FeatEnHancer in the video domain. Furthermore, we believe that far better results can be attained by incorporating our FeatEnHancer with better baselines and optimal experimental configurations.*

[2] https://github.com/open-mmlab/mmsegmentation/blob/master/configs/_base_/schedules/schedule_20k.py

[3] https://github.com/open-mmlab/mmtracking/blob/master/configs/vid/selsa/selsa_faster_rcnn_r50_dc5_1x_imagenetvid.py

2. Results and Discussion

2.1. Detailed Results on ExDark

Table I summarizes the exhaustive quantitative analysis, comparing our FeatEnHancer with several LLIE approaches, including RAUS [27], KIND [46], EnGAN [17], MBLLEN [10], Zero-DCE [12], Zero-DCE++ [18], and state-of-the-art dark object detection method MAET [6] on the ExDark dataset [22]. It is evident that our FeatEnHancer yields impressive improvements using both object detection frameworks. Furthermore, we exhibit a comprehensive qualitative analysis in Fig. I. Despite the visually unappealing images, the detector equipped with our FeatEnHancer produces accurate detections compared to other LLIE and existing state-of-the-art methods. By looking at the first row of Fig. I, note that while all methods detect the bigger boat, they all miss the smaller boat. However, our FeatEnHancer, enriched with hierarchical multi-scale features, conveniently detects both of the boats. Similarly, bigger and smaller cars are accurately detected by our method in the same figure. These architectural innovations contribute to remarkable improvements in the baseline and deliver new state-of-the-art mAP_{50} of 86.3 on the ExDark dataset with Featurized Query R-CNN.

2.2. Detailed Results on DARK FACE

We demonstrate a qualitative comparison of our FeatEnHancer with existing LLIE methods and existing state-of-the-art dark object detection method on the DARK FACE dataset in Fig. II. Albeit the darkness and tiny faces, our FeatEnHancer provides strong visual cues to the detector and brings maximum gains in the baseline compared to other methods.

2.3. Detailed Results on ACDC

In Table II, we present a detailed quantitative analysis, comparing our FeatEnHancer with several LLIE approaches, including RetinexNet [34], DRBN [38], FIDE [40], KIND [46], EnGAN [17], Zero-DCE [12], SSIENet [45], and prior state-of-the-art nighttime semantic segmentation method Xue *et al*. [39] on the ACDC dataset [30]. The results show that our FeatEnHancer generates powerful semantic representations, producing significant boosts in the baseline performance and leading to a new state-of-the-art mIoU of 54.9. Moreover, we illustrate a detailed visual comparison in Fig. III. Note that our method produces accurate segmentations in both cases of small objects, such as traffic signs

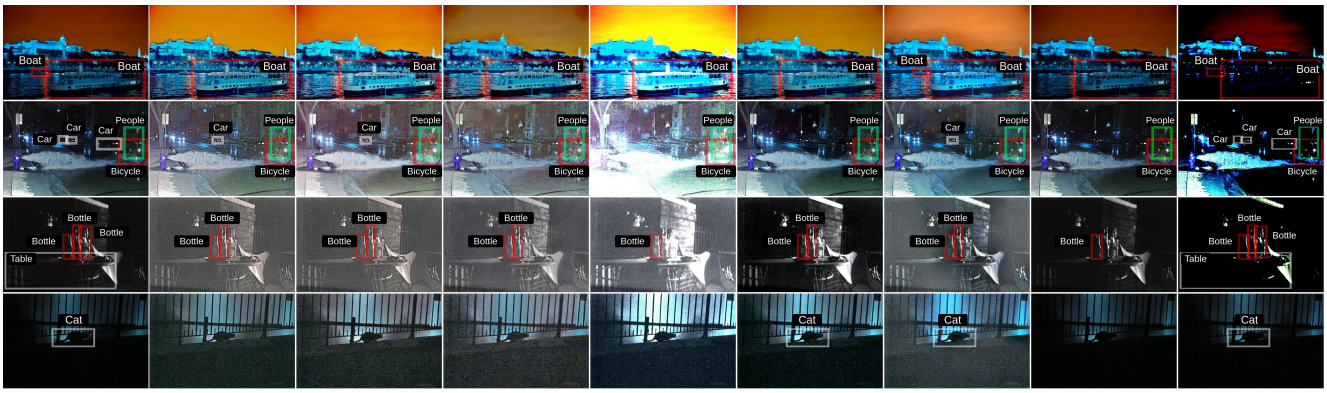

Figure I: Visual comparison of FeatEnHancer with several LLIE approaches and a previous state-of-art dark object detection method on the ExDark dataset. Zoom in for the best view.

Figure II: Visual comparison of FeatEnHancer with several LLIE approaches and an existing state-of-the-art dark object detection method on the DARK FACE dataset. Zoom in for the best view.

(second row) and larger objects, such as trains and terrains (first and fourth row). These extra gains demonstrate the effectiveness of the hierarchical multi-scale feature learning and scale-aware attentional feature aggregation schemes in the proposed method.

2.4. Analysing Features from RetinaNet and Featurized Query R-CNN

During experiments with RetinaNet and Featurized Query R-CNN on ExDark and DARK FACE datasets, we observe that our FeatEnHancer brings inferior improvements with RetinaNet (+0.5 mAP_{50}), compared to Featurized Query R-CNN (+11.8 mAP_{50}) (see Table I). Therefore, we visualize both the learned hierarchical representations of our FeatEnHancer and feature activations from the backbone network (*Res4 block*) in Fig. IV. Note that an identical backbone network of ResNet-50 is employed with both object detectors. From the figure, one can see that with RetinaNet, weaker representations are produced by our method, leading to suboptimal feature extraction, causing inferior gains. On the other hand, with an improved detector like Featurized Query R-CNN, our FeatEnHancer produces more meaningful representations, enabling subsequent backbone networks to extract suitable features. This behaviour is aligned with our network design which is optimized using a task-related loss function. Therefore, our FeatEnHancer can be integrated with advanced downstream vision methods to achieve substantial gains.

Method	Bicycle		Boat		Bottle		Bus		Car		Cat		Chair		Cup		Dog		Motorbike		People		Table		**AP50**		**AP75**		**mAP**	
	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ	Ret	FQ
Baseline	50.4	57.1	40.9	42.0	34.2	45.8	73.1	73.9	57.4	56.4	45.2	41.2	42.0	39.4	46.2	46.5	50.6	52.2	40.2	36.7	41.6	42.8	33.5	30.5	72.1	74.5	51.4	44.1	46.3	47.0
RAUS [27]	49.4	53.9	37.9	43.8	33.6	42.5	68.3	69.5	53.6	52.9	41.5	42.6	40.9	47.2	41.0	47.7	48.5	48.1	37.4	39.7	39.8	43.9	33.3	44.3	64.7	77.0	44.1	49.0	44.0	48.1
KIND [46]	49.4	55.6	38.8	46.0	34.5	46.1	72.4	71.7	56.5	57.8	41.6	48.0	40.1	48.4	44.6	56.0	50.8	51.8	39.0	44.3	41.0	45.1	32.3	47.4	70.7	80.5	49.6	57.2	45.1	51.5
Zero-DCE++ [18]	50.0	53.4	39.8	45.0	34.1	44.4	72.2	71.4	56.6	55.1	41.7	46.9	41.0	45.0	44.2	47.6	50.2	50.4	40.3	44.7	40.4	41.7	32.2	44.4	70.3	79.5	50.1	49.2	45.2	49.2
EnGAN [17]	49.5	55.1	39.9	47.2	33.7	43.3	72.6	74.5	56.1	57.7	42.0	46.9	40.3	49.2	43.1	55.4	50.1	53.1	38.8	45.0	40.7	45.8	31.6	49.1	70.4	80.0	49.7	58.7	44.9	51.9
MBLLEN [10]	50.1	55.4	38.0	45.0	33.8	47.2	72.6	72.6	57.3	59.6	41.7	46.5	41.4	46.6	43.5	52.6	49.8	51.9	40.6	45.7	41.0	46.1	32.5	47.7	70.6	80.0	49.0	58.3	45.1	51.0
Zero-DCE [12]	50.8	55.8	39.6	47.0	34.9	45.2	73.5	73.0	56.7	59.0	40.2	46.8	41.0	48.1	44.1	53.9	50.0	52.9	39.5	47.4	40.8	46.5	32.3	48.1	71.0	80.6	49.8	56.7	45.2	52.0
MAET [12]	50.8	56.2	39.8	47.8	35.7	45.3	74.5	73.3	56.9	59.4	40.9	46.9	41.7	48.7	44.5	54.3	50.5	53.9	39.7	47.7	40.9	46.7	32.6	48.4	71.8	81.6	49.8	56.7	45.7	52.4
FeatEnHancer	**51.0**	**60.3**	**41.6**	**49.1**	**47.6**	**53.7**	**69.8**	**78.4**	**54.2**	**62.1**	**47.5**	**52.1**	**41.8**	**51.9**	**41.0**	**57.7**	**34.5**	**61.2**	**42.7**	**50.8**	**45.8**	**54.4**	**40.3**	**46.9**	**72.6**	**86.3**	**51.4**	**63.6**	**46.4**	**56.5**

Table I: Detailed comparison of FeatEnHancer with LLIE approaches and existing state-of-the-art dark object detection method on the ExDark dataset. Here, Ret and FQ represent RetinaNet and Featurized Query R-CNN, respectively. Results obtained on the commonly used evaluation metrics are highlighted. Our FeatEnHancer consistently boosts the baseline performance and achieves new state-of-the-art results with Featurized Query R-CNN.

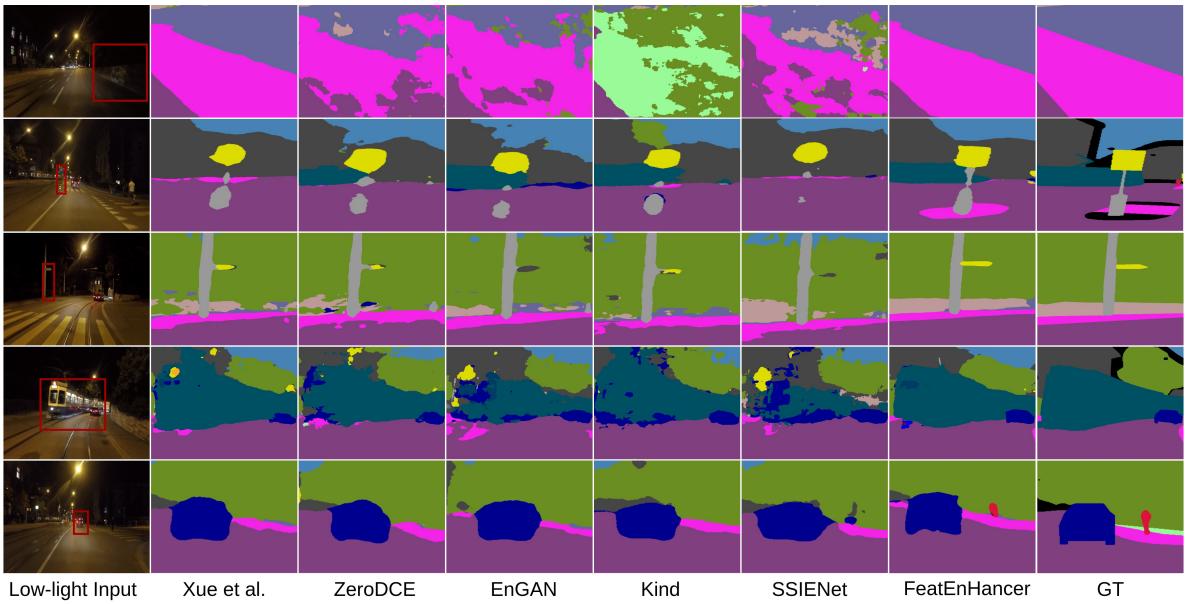

Figure III: Visual comparison of FeatEnHancer with several LLIE approaches and a previous state-of-art nighttime semantic segmentation method on the ACDC dataset. Zoom in for the best view.

2.5. Comparing FEN in FeatEnHancer with DCENet in Zero-DCE [12, 18]

Our intra-scale feature enhancement network (FEN) is inspired by the enhancement network DCENet employed in [12] and [18]. However, we incorporate several modifications, such as learning feature enhancement at multiple scales and scale-aware attentional feature aggregation schemes. For direct comparison, we exhibit the learned enhanced representations from both Zero-DCE, Zero-DCE++ and our FeatEnHancer in Fig. V. All three visualizations are achieved from identical experimental settings on the ExDark validation set. It is obvious that the proposed modification in our FEN produces more meaningful semantic representations compared to DCENet employed in [12, 18]. Both Zero-DCE and Zero-DCE++ produce false negatives, such as missing bicycles and people in the first and third rows, respectively. However, the scale-aware enhancement in our FeatEnHancer leads to accurate detections.

2.6. Comparison with Domain Adaptation Methods

Recently some methods [15, 21, 24, 43] have exploited the unsupervised domain adaption scheme to improve object detection in challenging environments. These methods are mainly designed to tackle object detection in harsh weather conditions such as foggy weather. Furthermore, for training, they require pre-training on large synthetic datasets (source dataset) labelled with the same classes that match the classes of the target dataset. Therefore, we refrain from including these works [15, 21, 24, 43] when comparing performance on dark object detection in Table I. Nevertheless, for direct comparison with results reported in [24], we incorporate our FeatEnHancer in the identical experimental settings[4] (YOLOv3 [25] as a baseline detector) and use the same

[4] https://github.com/NIvykk/DENet

Method	RO	SI	BU	WA	FE	PO	TL	TS	VE	TE	SK	PE	RI	CA	TR	TA	BI	**mIoU**
Baseline [37]	90.0	61.4	74.2	32.8	34.4	45.7	49.8	31.2	68.8	14.6	80.4	27.1	12.6	62.1	0.0	76.3	14.4	45.7
RetinexNet [34]	89.4	61.0	70.6	30.1	28.1	42.4	47.6	25.7	65.8	8.6	77.3	21.5	13.8	54.8	0.0	67.4	8.2	41.9
DRBN [38]	90.5	61.5	72.8	31.9	32.5	44.5	47.3	27.2	65.7	10.2	76.5	24.2	13.2	55.4	0.0	71.1	11.9	43.3
FIDE [40]	90.0	60.7	72.8	32.4	34.1	43.3	47.9	26.1	67.0	13.7	78.0	26.5	5.8	57.1	0.0	71.0	12.4	43.4
KIND [46]	90.0	61.0	73.2	31.9	32.8	43.5	42.7	27.7	65.5	13.3	77.4	22.8	8.1	55.1	0.0	74.5	11.5	43.0
EnGAN [17]	89.7	58.9	73.7	32.8	31.8	44.7	49.2	26.2	67.3	14.2	77.8	25.0	10.6	59.0	0.0	71.2	7.8	43.8
Zero-DCE [12]	90.6	59.9	73.9	32.6	31.7	44.3	46.2	25.8	67.2	14.6	79.1	24.7	7.7	59.4	0.0	66.8	13.9	43.4
SSIENet [45]	89.6	59.3	72.5	29.9	31.7	45.4	43.9	24.5	66.7	10.6	78.3	22.8	0.2	52.6	0.0	71.1	5.4	41.4
Xue *et al*. [39]	93.2	72.6	78.4	43.8	46.5	48.1	51.1	38.8	68.6	14.9	79.1	21.9	2.2	61.6	**5.2**	85.2	36.1	49.8
FeatEnHancer	**94.0**	**75.1**	**78.6**	**44.9**	**41.6**	**53.9**	**66.0**	**49.9**	**71.2**	**15.1**	**82.7**	**45.3**	**10.2**	**72.5**	0.0	**89.5**	**43.0**	**54.9**

Table II: Comparing FeatEnHancer with LLIE approaches and existing state-of-the-art nighttime segmentation method on the ACDC dataset. For brevity, we represent classes {road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, train, bicycle} with {RO, SI, BU, WA, FE, PO, TL, TS, VE, TE, SK, PE, RI, CA, TR, TA, BI}, respectively. Our FeatEnHancer yields remarkable improvements in the baseline, producing new state-of-the-art results.

10 categories that match with Pascal VOC dataset [9]. We present the results of this experiment in Table III. Note that our FeatEnHancer surpasses the previous best method (DE-YOLO [24] that leverages Laplacian Pyramid [1] to generate multi-scale features) in this specific experimental setting and reaches the mAP_{50} of 53.70.

Methods	**mAP_{50}**
Baseline [25]	43.02
MS-DAYOLO [15]	44.25
DAYOLO [43]	44.62
DSNet [16]	45.31
MAET [6]	47.10
IA-YOLO [21]	49.53
DE-YOLO [24]	51.51
FeatEnHancer	**53.70**

Table III: Quantitative results of our FeatEnHancer and existing UDA methods on the ExDark dataset. For direct comparison, an identical framework of YOLOv3 as the baseline is adopted. Only 10 classes that match with the Pascal VOC dataset [9] are used.

2.7. Performance-Efficiency Tradeoff

Table IV compares the performance-efficiency tradeoff between the proposed FeatEnHancer and LLIE methods and task-specific previous state-of-the-art approaches. While Xue [39] and Zero-DCE [12] contain fewer parameters, they bring sub-optimal gains to the baseline method. On the contrary, with a slightly more number of parameters, our FeatEnHancer demonstrates generalizability and robustness in 4 different downstream vision tasks under low-light conditions. *Nevertheless, it is worth mentioning that most of the parameters come from the proposed scale-aware attentional feature aggregation module in our FeatEnHancer. We believe that instead of traditional attention mechanism [32], incorporating an optimized attentional scheme such as [33] will further reduce the parameters in our FeatEnHancer. We leave this for future works to explore.*

Methods	**#Params**	**mAP_{50}**	**mIoU**
MBLLEN [10]	450K	80.0	-
KIND [46]	8M	80.5	43.0
EnlightenGAN [17]	9M	80.0	43.8
Zero-DCE [12]	79K	80.6	43.4
Xue et. al [39]	**14K**	-	49.8
MAET [6]	40M	81.6	-
FeatEnHancer	138K	**86.3**	**54.9**

Table IV: Comparing the number of parameters in the FeatEnHancer against LLIE approaches and task-specific state-of-the-art methods. #Params is the number of parameters, K and M denote thousands and millions, respectively. mAP_{50} is computed on the ExDark dataset, and mIoU values are taken from the ACDC dataset. We take top competitors of the ExDark and ACDC datasets for comparison. For ExDark, results from Featurized Query R-CNN are adopted.

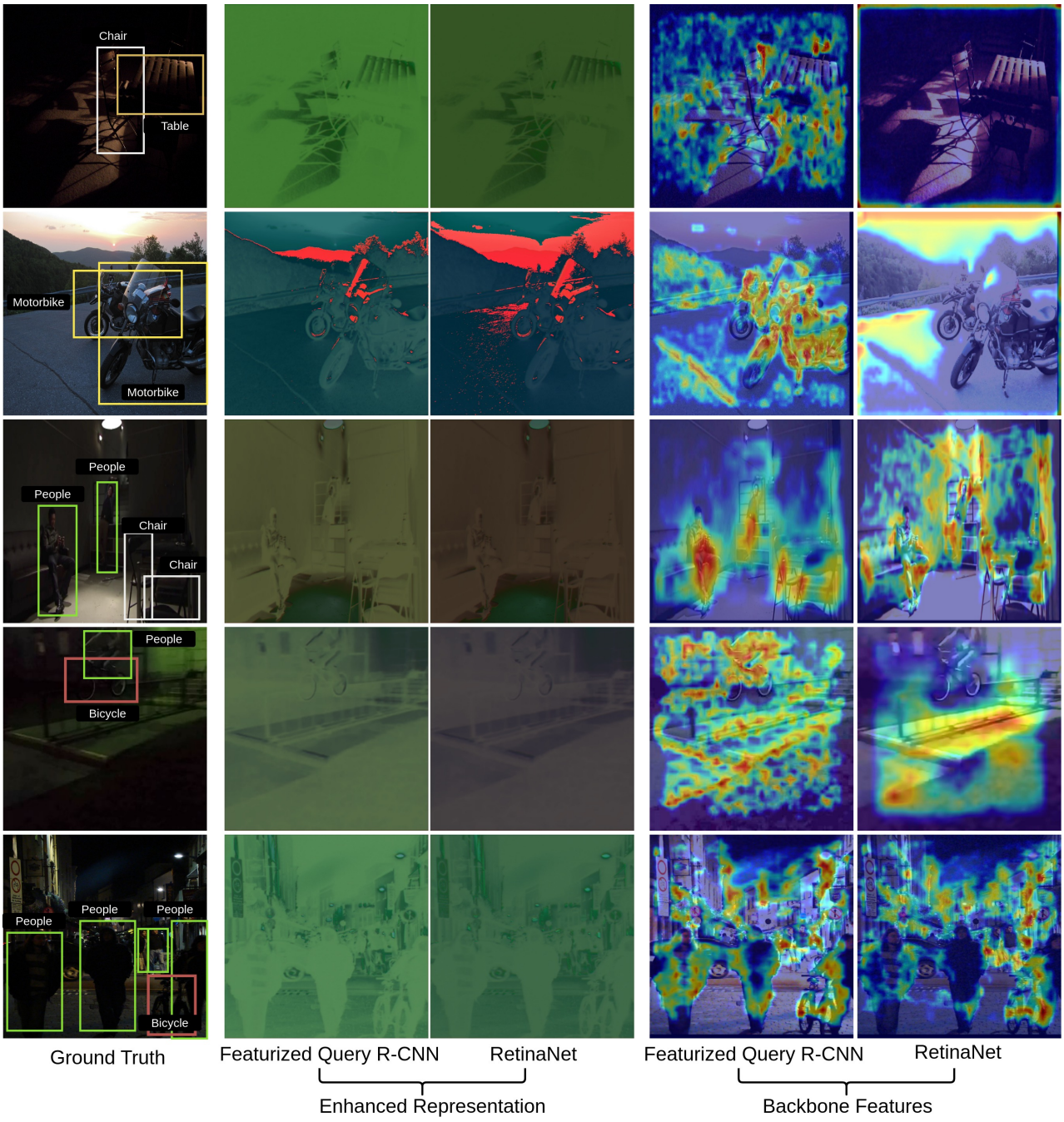

Figure IV: Visualizing enhanced hierarchical representation and backbone features from RetinaNet and Featurized Query R-CNN. We visualize the output of the final enhanced representation achieved after aggregating multi-scale hierarchical features. For backbone features, we employ gradcam [31] and illustrate learned features from the last *Res4* block in the ResNet-50 backbone network.

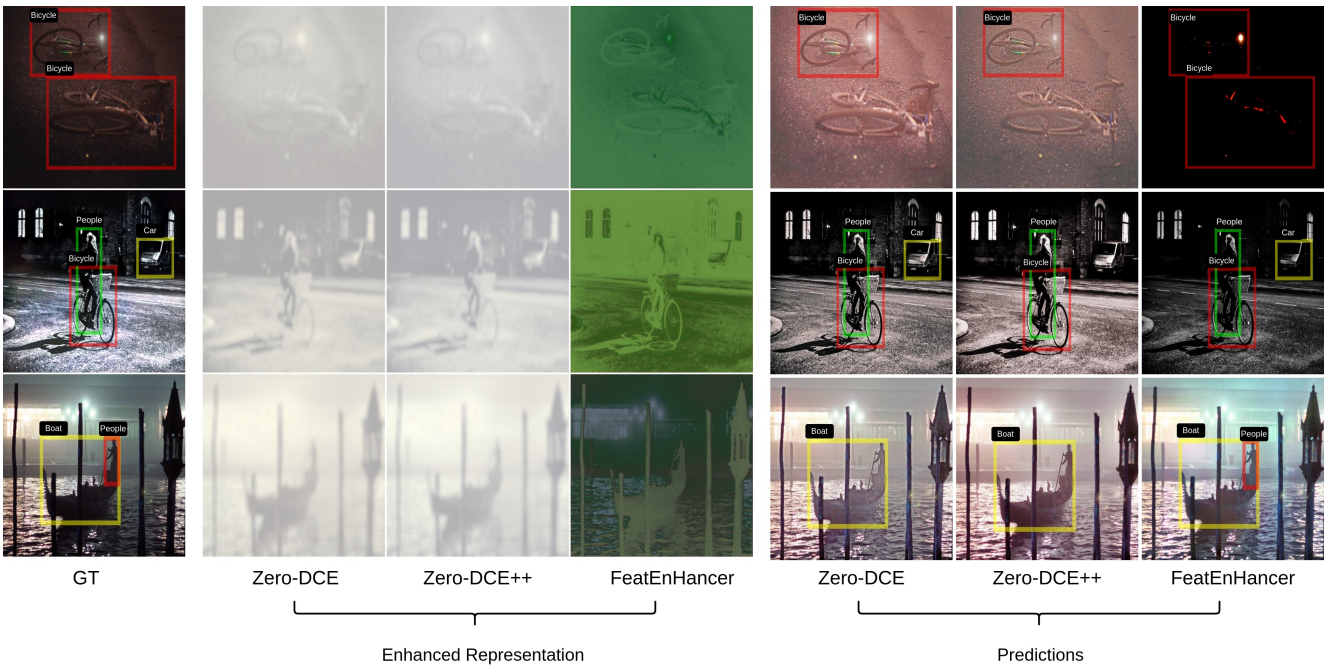

Figure V: Visual comparison between our feature enhancement network and DCENet proposed in [12, 18]. For comparison, we illustrate the learned mean of all eight curves from DCENet in Zero-DCE and Zero-DCE++. For our FeatEnHancer, we visualize the learned aggregated hierarchical feature representation. All methods are incorporated with Featurized Query R-CNN and trained on the ExDark dataset.

References

[1] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987.

[2] 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 arXiv:1906.07155*, 2019.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[4] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[5] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking, 2020.

[6] 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 (ICCV)*, pages 2553–2562, October 2021.

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[9] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015.

[10] Yu Li Feifan Lv and Feng Lu. Attention-guided low-light image enhancement. *arXiv preprint arXiv:1908.00682*, 2019.

[11] Tao Gong, Kai Chen, Xinjiang Wang, Qi Chu, Feng Zhu, Dahua Lin, Nenghai Yu, and Huamin Feng. Temporal roi align for video object recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1442–1450, 2021.

[12] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. *CoRR*, abs/2001.06826, 2020.

[13] Khurram Azeem Hashmi, Didier Stricker, and Muhammamd Zeshan Afzal. Spatio-temporal learnable proposals for end-to-end video object detection, 2022.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[15] Mazin Hnewa and Hayder Radha. Multiscale domain adaptive yolo for cross-domain object detection. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3323–3327, 2021.

[16] Shih-Chia Huang, Trung-Hieu Le, and Da-Wei Jaw. Dsnet: Joint semantic learning for object detection in inclement weather conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2623–2633, 2021.

[17] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021.

[18] Chongyi Li, Chunle Guo Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[19] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[21] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):1792–1800, June. 2022.

[22] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *CoRR*, abs/1805.11227, 2018.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[24] Qingpao Qin, Kan Chang, Mengyuan Huang, and Guiqing Li. Denet: Detection-driven enhancement network for object detection under adverse weather conditions. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 2813–2829, December 2022.

[25] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[26] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017.

[27] Liu Risheng, Ma Long, Zhang Jiaao, Fan Xin, and Luo Zhongxuan. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[28] Sebastian Ruder. An overview of gradient descent optimization algorithms, 2016.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[30] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10765–10775, October 2021.

[31] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[33] Keqi Wang, Ziteng Cui, Ge Wu, Yin Zhuang, and Yuhua Qian. Linear array network for low-light image enhancement. *CoRR*, abs/2201.08996, 2022.

[34] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[35] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[36] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[37] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[38] Ke Xu, Xin Yang, Baocai Yin, and Rynson W.H. Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[39] Xinwei Xue, Jia He, Long Ma, Yi Wang, Xin Fan, and Risheng Liu. Best of both worlds: See and understand clearly in the dark. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 2154–2162, New York, NY, USA, 2022. Association for Computing Machinery.

[40] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[41] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, Yuqiang Zheng, Yanyun Qu, Yuhong Xie, Liang Chen, Zhonghao Li, Chen Hong, Hao Jiang, Siyuan Yang, Yan Liu, Xiaochao Qu, Pengfei Wan, Shuai Zheng, Minhui Zhong, Taiyi Su, Lingzhi He, Yandong Guo, Yao Zhao, Zhenfeng Zhu, Jinxiu Liang, Jingwen Wang, Tianyi Chen, Yuhui Quan, Yong Xu, Bo Liu, Xin Liu, Qi Sun, Tingyu Lin, Xiaochuan Li, Feng Lu, Lin Gu, Shengdi Zhou, Cong Cao, Shifeng Zhang, Cheng Chi, Chubing Zhuang, Zhen Lei, Stan Z. Li, Shizheng Wang, Ruizhe Liu, Dong Yi, Zheming Zuo, Jianning Chi, Huan Wang, Kai Wang, Yixiu Liu, Xingyu Gao, Zhenyu Chen, Chang Guo, Yongzhou Li, Huicai Zhong, Jing Huang, Heng Guo, Jianfei Yang, Wenjuan Liao, Jiangang Yang, Liguo Zhou, Mingyue Feng, and Likun Qin. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.

[42] Bo Zhang, Yuchen Guo, Runzhao Yang, Zhihong Zhang, Jiayi Xie, Jinli Suo, and Qionghai Dai. Darkvision: A benchmark for low-light image/video perception, 2023.

[43] Shizhao Zhang, Hongya Tuo, Jian Hu, and Zhongliang Jing. Domain adaptive yolo for one-stage cross-domain detection. In Vineeth N. Balasubramanian and Ivor Tsang, editors, *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pages 785–797. PMLR, 17–19 Nov 2021.

[44] Wenqiang Zhang, Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Qian Zhang, and Wenyu Liu. Featurized query r-cnn, 2022.

[45] Yu Zhang, Xiaoguang Di, Bin Zhang, and Chunhui Wang. Self-supervised image enhancement network: Training with low light images only. *arXiv preprint arXiv:2002.11300*, 2020.

[46] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. *CoRR*, abs/1905.04161, 2019.