TorchAdapt: Towards Light-Agnostic Real-Time Visual Perception

Khurram Azeem Hashmi[1,2] Karthik Palyakere Suresh[1,2] Didier Stricker[1,2]
Muhammad Zeshan Afzal[1,2]
[1]DFKI [2]RPTU Kaiserslautern-Landau
khurram_azeem.hashmi@dfki.de

Contents

A. Implementation Details

A.1. Object Detection

Following common practice [7] in LL object detection, we adopt YOLOV3 [20] as the baseline detector. Since YOLOV3 operates on Darknet backbone [19], we pre-train the TorchAdapt following settings explained in Sec. 3.3 and Fig. 3 in the main paper, where the visual encoder f is Darknet-53, pre-trained on the ImageNet [8].

To evaluate the performance on the object detection task, for the (Low-Light) LL setting, COCO [14] pre-trained weights are utilized to fine-tune YOLOV3 on the ExDark dataset [16]. Alternatively, we train the model from scratch in the (Light-Agnostic) LA setting, as COCO images are also included in the training data, as summarized in Table 1 in the main paper. Training is conducted using the Stochastic Gradient Descent (SGD) optimizer [21] with an initial learning rate of 1×10^{-3} and batch size of 16 is used. All images are resized to 608×608 pixels to maintain consistency. We train YOLOv3 for 24 epochs, reducing the learning rate by a factor of 10 at epochs 18 and 23. Our implementation is based on the MMDetection toolbox [1]. To evaluate the performance, we follow prior approaches [7, 10] and report mAP scores at mAP@IoU=0.5 and mAP@IoU=0.5:95 using [14].

A.2. Face Detection

For face detection, we choose a different detector and adopt RetinaNet [15] to report results on the DARK FACE dataset [25, 29] in the LL setting and a combination of the DARK FACE and WIDER FACE [28] datasets in the LA setting, as both datasets only contain a single *face* class, making it challenging face detection benchmark with varying illuminations. During pre-training of TorchAdapt, ResNet-50 [13] backbone is utilized as a vision encoder f. For the face detection task, following prior works [10], we resize images to a resolution of 1500×1000 pixels and follow the $1\times$ schedule[1] in MMDetection [1]. We follow the same evaluation protocol to report face detection performance as in § A.1.

A.3. Instance Segmentation

For the instance segmentation task, we adopt RTMDet-Ins-tiny [17] as our baseline model. Since RTMDet employes CSPNext [24] as the light-weight backbone network, we conduct TorchAdapt pre-training with the visual encoder f set to CSPNext-tiny in Fig. 3.

For the LL setting, we adopt the LIS [2] dataset. For the LA setting, we take all images containing common classes between LIS and COCO [14], as explained in Eq. 6 and Table 1 in the main paper. Following Section A.1, we use the COCO pre-trained weights of RTMDet-Ins-tiny for the LL setting and train the model from scratch for the LA setting. We resize images to 640×640 pixels and train the model for 50 and 300 epochs for LL and LA settings, respectively. The rest of the experimental settings follow implementation from

[1] https://github.com/open-mmlab/mmdetection/blob/main/configs/retinanet/retinanet_r50_fpn_amp-1x_coco.py

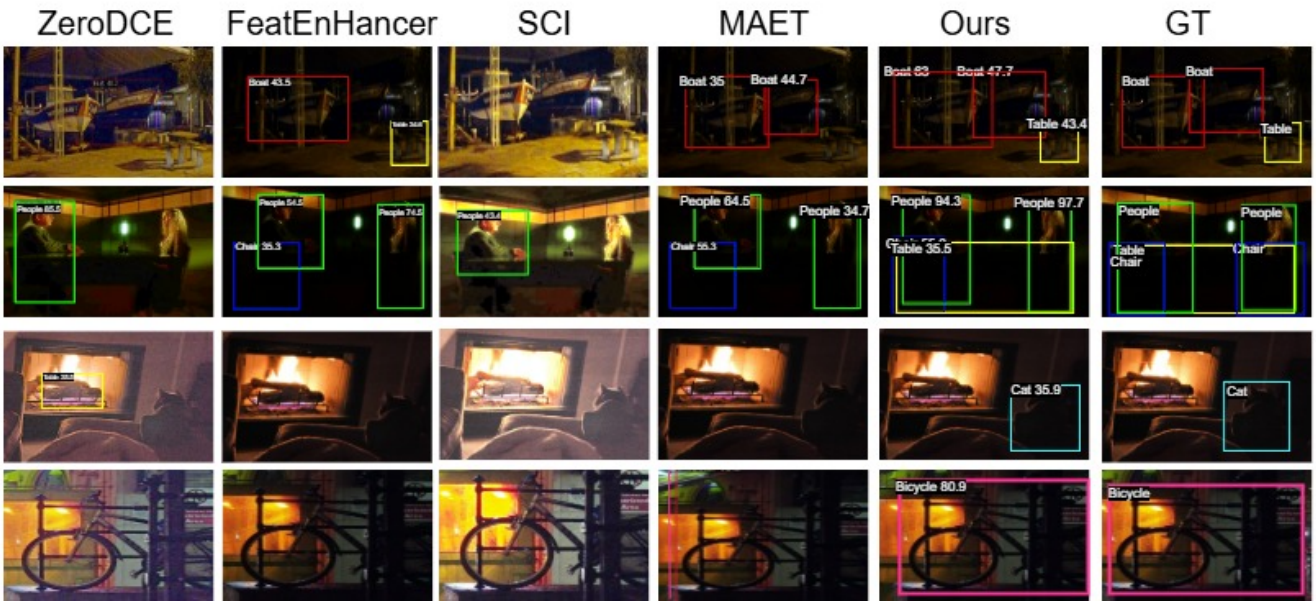

Figure I. **Qualitative comparisons on the ExDark validation set**. All methods are integrated into YOLOV3 [20] and trained end-to-end on the ExDark training set [16]. LLIE methods such as Zero-DCE and SCI improve the visual quality but yield inferior predictions. Nevertheless, our TorchAdapt consistently detects objects more accurately than others, such as ZeroDCE [9], FeatEnhancer [10], SCI [18], and MAET [7], while closely matching the ground truth (GT). These results highlight TorchAdapt's ability to enhance object detection in low-light environments.

Figure II. **Qualitative comparisons on Light-Agnostic Object Detection on the ExDark and COCO validation split.** All methods are integrated into YOLOv3 [20] and trained end-to-end. The figure highlights TorchAdapt's ability to consistently detect objects under both low-light (ExDark) and well-lit (COCO) conditions, outperforming prior task-specific methods, such as FeatEnhancer, Zero-DCE, and SCI, while closely aligning with the ground truth (GT).

Figure III. **Qualitative comparisons on the DARK FACE validation set**. The figure showcases the performance of TorchAdapt (Ours) compared to other LLIE and task-specific methods, including ZeroDCE, SCI, FeatEnhancer, URetinexNet, and MAET, in detecting faces under low-light conditions. TorchAdapt consistently identifies faces more accurately and robustly, closely aligning with the ground truth (GT), demonstrating its effectiveness in challenging low-light scenarios. Best view it on the screen and zoom in.

the MMDetection [1][2]. Following the authors of LIS [2], we employ commonly used mAP and SegAP as the evaluation metrics to report performance on instance segmentation task.

A.4. Semantic Segmentation

Consistent with prior works [10, 27], we adopt DeepLabV3+[3] as our baseline model for the semantic segmentation task. We employ ACDC DAtaset [23] for the LL setting and the combination of ACDC and CityScapes [6] datasets for the LA settings, as summarized in Table 1. Following detection tasks, we fine-tune the pre-trained baseline on the ACDC dataset for low-light evaluation, whereas we train the model from scratch for LA settings.

For training, we resize the images to 2048×1024. To enable a direct comparison with the previous state-of-the-art methods [10, 27], we employ DeepLabV3+ [3] with ResNet-50 [13] as its backbone. For our TorchAdapt, same pre-trained weights used in § A.2 are utilized here. The backbone is pre-initialized with ImageNet [8] weights, and training is performed with a batch size of 1. The optimizer is SGD [21], configured with the $20K$ scheduler[3] from MMSegmentation [4], a base learning rate of 0.001, and a weight decay of 0.0005. Similar to prior related efforts [10, 27], we adopt the commonly used mIoU metrics to evaluate the semantic segmentation performance.

[2] https://github.com/open-mmlab/mmdetection/blob/main/configs/rtmdet/

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

Figure IV. **Qualitative comparisons on the light-agnostic face detection on the DARK FACE [29] and WIDER FACE [28]**. All methods are incorporated into RetinaNet [15] and trained end-to-end, following light-agnostic experimental setting in Table 1. The figure highlights TorchAdapt's (Ours) superior ability to accurately detect faces in both low-light (DARK FACE) and well-lit (WIDER FACE) conditions, outperforming other methods such as FeatEnhancer, Zero-DCE, and SCI, while closely matching the ground truth (GT).

.

A.5. Video Object Detection

In addition to image-based tasks, we evaluate the generalization capabilities of TorchAdapt in the video domain. Following [10], we employ SELSA [26] as a video object detection baseline framework with a ResNet-50 backbone pre-trained on ImageNet. Similar to § A.2, we use the TorchAdapt with pre-trained weights optimized with ResNet-50 backbone network. In order to obtain results in Table 2, we fine-tune the SELSA baseline on the DarkVision dataset [30] for the LL setting. On the other hand, for the LA setting, we employ a well-lit, commonly used video object detection benchmark ImageNet VID [22]. Since DarkVision has five splits, for convenience of experiments, we choose the 3.2 illumination level split from DarkVision in both LL and LA settings.

In consistent with [10], we follow the implementation details adheres to the $1\times$ schedule[4] in mmtracking [5]. For comparison with low-light image enhancement (LLIE) methods, all video frames are first enhanced using their checkpoints before being processed by the baseline network. For task-specific evaluations, all methods are integrated into the baseline and trained end-to-end, similar to our TorchAdapt. Adopting common practice in video object detection [10–12, 26], we utilize mAP@IoU=0.5 to compute all methods.

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

A.6. Employed Datasets

COCO. The COCO dataset [14] is a large-scale dataset widely used for object detection and instance segmentation tasks. For light-agnostic evaluations, COCO is combined with ExDark [16] for object detection and LIS [2] for instance segmentation. The combined dataset for object detection includes 12 object classes with 92,357 training samples and 5,156 validation samples. For the instance segmentation task, 8 classes are used with 44,321 training samples and 2,550 validation samples.

ExDark. ExDark [16] is a dataset designed specifically for object detection under low-light conditions. It contains 12 object categories, with 5,891 training images and 1,472 validation images. ExDark is also used in combination with COCO for light-agnostic evaluations, providing a comprehensive benchmark for both low-light and mixed illumination settings.

DARK FACE. The DARK FACE dataset [29] focuses on face detection under challenging low-light scenarios. It includes 5,400 training samples and 600 validation samples. For light-agnostic evaluations, it is combined with WIDER FACE [28], forming a dataset with 18,280 training samples

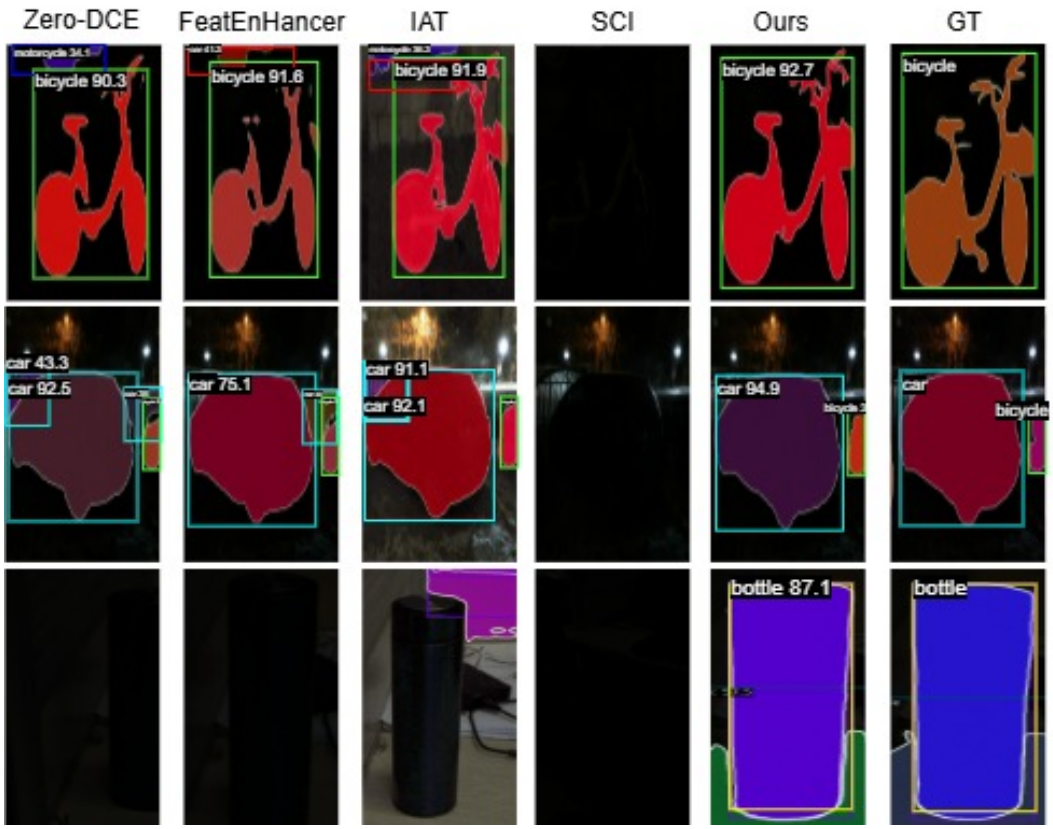

Figure V. **Qualitative comparisons on the LIS validation set under low-light setting.** TorchAdapt not only produces accurate instance segmentations but also avoids false positives, such as misclassifications of motorcycles and cars, which are evident in prior state-of-the-art methods like Zero-DCE and FeatEnhancer.

and 3,822 validation samples.

WIDER FACE. WIDER FACE [28] is a comprehensive face detection dataset featuring diverse challenges, such as variations in scale, occlusion, pose, and lighting conditions. It contains a total of 32,203 images and 393,703 labeled faces, making it one of the most diverse face detection benchmarks. For light-agnostic evaluations, WIDER FACE is combined with DARK FACE [29] to create a balanced benchmark covering both well-lit and low-light scenarios. This combined dataset includes 18,280 training samples and 3,822 validation samples, enabling robust evaluation of face detection models across varying illumination conditions.

LIS. The LIS dataset [2] is tailored for instance segmentation in low-light environments. It contains 8 object classes with 1,561 training images and 669 validation images. For light-agnostic evaluations, LIS is merged with COCO, forming a dataset with 44,321 training samples and 2,550 validation samples.

ACDC. The ACDC dataset [23] is designed for semantic segmentation under adverse weather and lighting conditions, including night time scenarios. It consists of 19 semantic classes, with 400 training images and 106 validation images. For light-agnostic evaluations, ACDC is combined with CityScapes [6], yielding a dataset with 3,375 training samples and 606 validation samples.

CityScapes. CityScapes [6] is a widely used semantic segmentation dataset designed to evaluate scene understanding in urban environments. It includes high-resolution images with pixel-level annotations covering 19 semantic classes, such as roads, buildings, vehicles, and pedestrians, captured under diverse weather and lighting conditions. For light-agnostic evaluations, CityScapes is combined with ACDC [23] to ensure comprehensive coverage of both well-lit and low-light urban scenes.

DarkVision. DarkVision [30] is a recently introduced dataset for video object detection under low-light conditions. It includes 4 classes with 26 video samples for training and 6 video samples for validation in low-light settings, covering 5 different illumination levels. For light-agnostic evaluations, it is merged with ImageNet VID [22], resulting in a dataset with 401 training video samples and 51 validation video samples.

ImageNet VID. ImageNet VID [22] is a large-scale benchmark dataset specifically designed for video object detection.

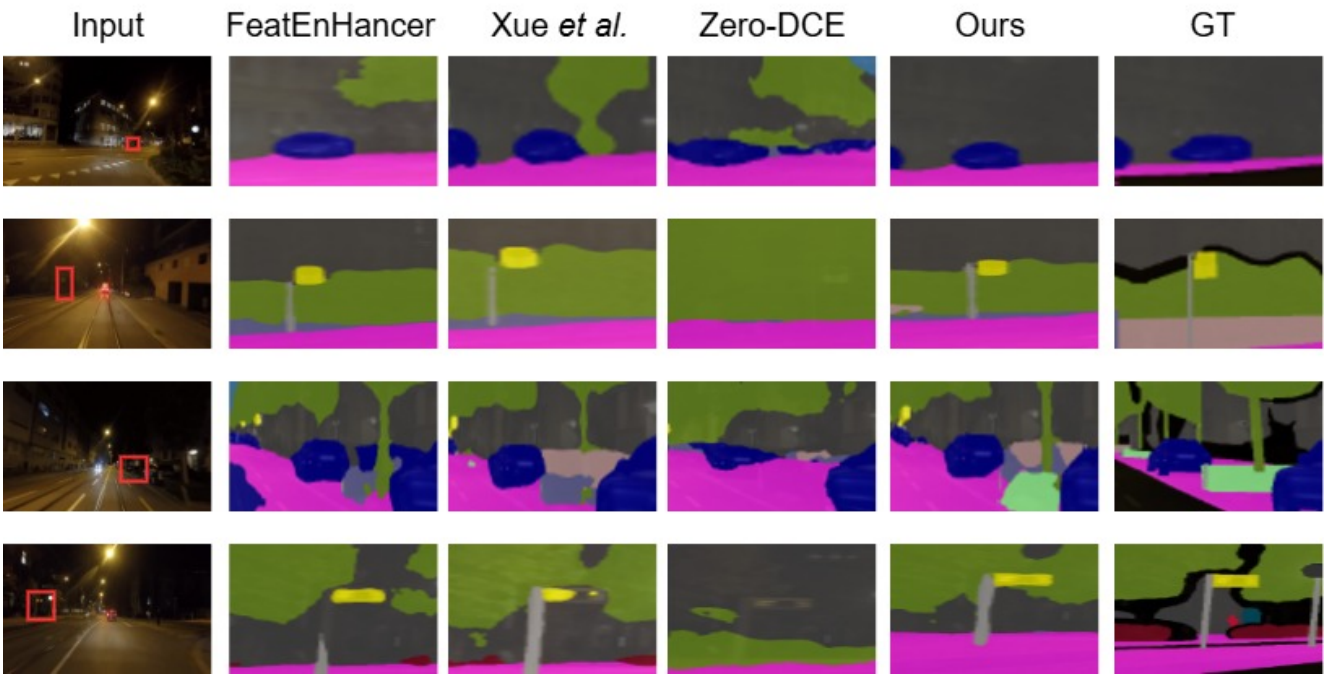

Figure VI. **Qualitative analysis of night time semantic segmentation task on the ACDC validation set.** The figure highlights TorchAdapt's (Ours) ability to generate accurate segmentations under low-light conditions. In contrast, prior methods such as FeatEnhancer, Xue et al., and Zero-DCE exhibit noticeable segmentation errors, particularly in critical regions like road boundaries, Poles, and Fences. While TorchAdapt demonstrates impressive performance, certain discrepancies remain in challenging areas, such as fine-grained structures visible in the GT masks, indicating room for further improvement.

It includes 30 object categories and provides a total of 3862 training and 555 validation videos captured under diverse real-world conditions, including varying object motions, occlusions, and viewpoints. This dataset serves as a critical benchmark for evaluating the temporal and spatial reasoning capabilities of detection models. For light-agnostic evaluations, ImageNet VID is combined with DarkVision [30], a low-light video object detection dataset. This combination ensures a comprehensive benchmark to test the robustness and generalization of models across both well-lit and low-light video sequences, covering diverse lighting conditions and challenging scenarios.

B. Results and Discussion

B.1. Object Detection

We provide the quantitative analysis for the object detection task under both Low-light (LL) and Light-Agnostic (LA) settings in Table 2 in the main paper. Here, we provide qualitative analysis with LL settings in Fig. I and with LA settings in Fig. II. Experimental settings explained in § A.1 are adopted to reproduce these figures.

B.2. Face Detection

We provide the quantitative analysis for the face detection task under both LL and LA settings in Table 2 in the main paper. Here, we provide qualitative analysis with LL settings in Fig. III and with LA settings in Fig. IV. Experimental settings explained in § A.2 are adopted to reproduce these figures.

B.3. Instance Segmentation

We provide qualitative analysis of Instance segmentation, visualizing predictions from the top 5 performing methods in Fig. V. Experimental settings explained in § A.3 are adopted to reproduce this figure.

B.4. Semantic Segmentation

We provide qualitative analysis of Semantic segmentation, visualizing predictions from the top 4 performing methods in Fig. VI. Experimental settings explained in § A.4 are adopted to reproduce this figure.

B.5. Enhancement As Emergence

In Sec. 5 of the main paper, we explore the enhancement as an emergent property of our TorchAdapt. Here, we visualize more examples in Fig. VII. For each input image I, we

illustrate the learned modulated enhancement from the Adapt module and the final enhanced image from the complete TorchAdapt module.

C. Additional Experiments and Ablations

C.1. Performance Comparison on Well-lit Datasets

Table I evaluates the performance of TorchAdapt on well-lit datasets, comparing it with prior SOTA low-light image enhancement and task-specific methods. The experiments are conducted on COCO [14] for object detection and CityScapes [6] for semantic segmentation. TorchAdapt achieves an mAP of **34.7** and mAP_{50} of **57.6** in object detection, improving by **+0.9** and **+0.8**, respectively, over the baseline. For semantic segmentation, TorchAdapt attains an mIoU of **80.0**, surpassing the baseline by **+0.4**. These results highlight TorchAdapt's ability to preserve or even enhance performance in well-lit conditions, leveraging its illumination-invariant learning. Notably, TorchAdapt outperforms prior methods such as Zero-DCE, SCI, and Feat-EnHancer, demonstrating its effectiveness and adaptability across diverse illumination settings.

C.2. LLIE methods as a *Torch* in TorchAdapt

Table II demonstrates the improved performance of LLIE methods, Zero-DCE [9] and SCI [18], when integrated as the *Torch* module in the TorchAdapt framework. For Zero-DCE, incorporating it as a *Torch* results in significant gains, achieving mAP improvements of **+0.1** in LL and **+0.3** in LA for object detection, and SegAP improvements of **+0.2** in LL and **+0.4** in LA for instance segmentation. Similarly, SCI as *Torch* improves mAP by **+0.1** in both LL and LA for object detection, and SegAP by **+1.1** in LL and **+0.7** in LA for instance segmentation. These results highlight the ability of TorchAdapt to make existing LLIE methods more effective and light-agnostic by leveraging its modular framework, significantly enhancing their performance across diverse illumination conditions.

C.3. Can TorchAdapt work without Light-Agnostic Training?

Table III highlights the impact of light-agnostic training on TorchAdapt for object detection and instance segmentation under low-light (LL) and light-agnostic (LA) settings. It consistently improves performance, achieving mAP gains of **+0.2** in LL and **+1.1** in LA for object detection, as well as SegAP gains of **+0.7** in LL and **+0.8** in LA for instance segmentation. These results demonstrate that light-agnostic training enhances TorchAdapt's ability to learn illumination-invariant features, leading to improved task performance across both settings. Notably, even without light-agnostic training, TorchAdapt delivers competitive results compared to prior LLIE and task-specific methods.

Figure VII. **Illustrating low-light image Enhancement as an emergent property of TorchAdapt on the DARK FACE validation set.** This figure is an extension of Fig. 6 in the main paper. Left side of the figure is the input image, the middle part is the adaptive enhancement learned by the Adapt module, and the rightmost is the enhanced image from the TorchAdapt. Although TorchAdapt is designed to produce illumination-invariant features for high-level vision tasks without employing any explicit enhancement loss functions during training, it inherently enhances low-light images, specifically the region of interest. For instance, human faces in the DARK FACE dataset.

Method	Object Det.		Semantic Seg.
	mAP	mAP_{50}	mIoU
Baseline	33.8	56.8	79.6
Zero-DCE [9]	33.1	55.9	79.2
SCI [18]	32.2	49.8	77.0
FeatEnHancer [10]	33.6	56.5	79.5
TorchAdapt	**34.7**+0.9	**57.6**+0.8	**80.0**+0.4

Table I. **Comparison of TorchAdapt with prior SOTA low-light image enhancement and task-specific methods on well-lit datasets, including COCO [14] for Object Detection and CityScapes [6] for Semantic Segmentation.** All methods are trained end-to-end using the same baselines for direct comparison. Leveraging illumination-invariant learning, TorchAdapt not only preserves baseline performance but also achieves improvements on already well-lit datasets, demonstrating its adaptability and effectiveness.

Method	Object Det.		Instance Seg.	
	LL	LA	LL	LA
Zero-DCE [9]	76.2	57.6	43.5	32.1
Zero-DCE ‡	79.1	62.9	52.7	35.7
As *Torch*	79.2	63.2	52.9	36.1
SCI [18]	75.5	57.9	43.5	32.1
SCI ‡	79.2	63.0	51.0	34.9
As *Torch*	79.3	63.3	52.1	35.6

Table II. **Making LLIE methods (Zero-DCE [9] and SCI [18]) better and light-agnostic, by plugging them as a Torch in our TorchAdapt framework.** These results affirm the model-agnostic generality of the TorchAdapt framework.

Light-Agnostic Training	Object Det.		Instance Seg.	
	LL	LA	LL	LA
×	79.9	62.9	52.9	35.8
✓	**80.1**	**64.0**	**53.6**	**36.6**

Table III. **Ablating Light-Agnostic training in the TorchAdapt and its impact on object detection and instance segmentation tasks in both low-light and light-agnostic settings.** Light-agnostic training brings stronger gains, specifically in the LA setting. However, it is worth mentioning that even without light-agnostic training, TorchAdapt produces reasonable gains when compared with prior LLIE and task-related methods.

References

[1] 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. 1, 3

[2] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 1, 3, 4, 5

[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)*, 2018. 3

[4] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 3

[5] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking, 2020. 4

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3, 5, 7, 9

[7] 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, 2021. 1, 2

[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. 1, 3

[9] 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. 2, 7, 9

[10] Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, and Muhammad Zeshan Afzal. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6725–6735, 2023. 1, 2, 3, 4, 9

[11] Khurram Azeem Hashmi, Alain Pagani, Didier Stricker, and Muhammad Zeshan Afzal. Boxmask: Revisiting bounding box supervision for video object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2030–2040, 2023.

[12] Khurram Azeem Hashmi, Talha Uddin Sheikh, Didier Stricker, and Muhammad Zeshan Afzal. Beyond boxes: Mask-guided spatio-temporal feature aggregation for video object detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8122–8133, 2025. 4

[13] 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. 1, 3

[14] 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. 1, 4, 7, 9

[15] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 1, 4

[16] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *CoRR*, abs/1805.11227, 2018. 1, 2, 4

[17] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022. 1

[18] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5637–5646, 2022. 2, 7, 9

[19] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 1

[20] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 1, 2

[21] Sebastian Ruder. An overview of gradient descent optimization algorithms, 2016. 1, 3

[22] 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. 4, 5

[23] 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, 2021. 3, 5

[24] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020. 1

[25] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *CoRR*, abs/1808.04560, 2018. 1

[26] 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)*, 2019. 4

[27] 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*, page 2154–2162, New York, NY, USA, 2022. Association for Computing Machinery. 3

[28] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 1, 4, 5

[29] 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. 1, 4, 5

[30] 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. 4, 5, 6