Technique Report of Team GroundTruth for CVPR 2024 PBDL Challenge Low-light Object Detection and Instance Segmentation

Xiaoqiang Lu, Licheng Jiao, Fang Liu, Xu Liu, Lingling Li, Wenping Ma, Shuyuan Yang
School of Artificial Intelligence, Xidian University
xqlu@stu.xidian.edu.cn

Abstract

Low-light conditions pose a significant challenge in maintaining image quality compared to well-lit environments, often resulting in noticeable degradation such as loss of detail, color distortion, and noticeable noise. These factors are detrimental to the performance of downstream vision tasks, especially object detection and instance segmentation. To this end, CVPR 2024 PBDL Challenge Low-light Object Detection and Instance Segmentation is held to evaluate and advance object detection algorithms' robustness on images captured from low-light environmental situations. In this report, we introduce our solution for addressing object detection and instance segmentation in the low-light condition. Specifically, we tackle the challenge with three tricks: (1) Pre-trained model. The Object365 and COCO datasets are two common basic datasets for object detection, where the model pre-trained on them can provide more accurate prior knowledge for transfer learning. (2) Strong backbone and detector. FocalNet and ViT-adapter are recently the most advanced feature extractors for image representation, and DINO and HTC are recently the most advanced detectors for dense predictions. (3) Training schedule and suitable test-time augmentation. Finally, without using extra detection data, we achieved AP of 0.76 and 0.62 in PBDL Challenge Low-light Object Detection and Instance Segmentation.

1. Object Detection

1.1. Network Architecture

DINO [11] is adopted as our detector which uses a contrastive way for denoising training, a mixed query selection method for anchor initialization, and a look forward twice scheme for box prediction in an end-to-end manner, as shown in Fig. 1. The most advanced and robust backbone FocalNet-Large [10] is utilized to extrack informative features, which introduce focal attention to additionally aggregate summarized visual tokens far away to capture coarse-grained and long-range visual dependencies, as shown in Fig. 2. In order to increase the receptive field of each roi feature, we exploit the roi pooling on the feature map of the corresponding level to get the global context feature, which is used to enhance the roi feature of the corresponding level by adding them. We also add SyncBN to each box head to make the training process more stable.

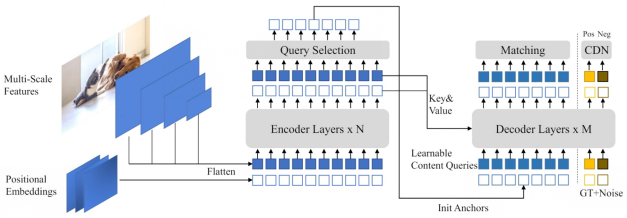

Figure 1. Framework of DINO [11].

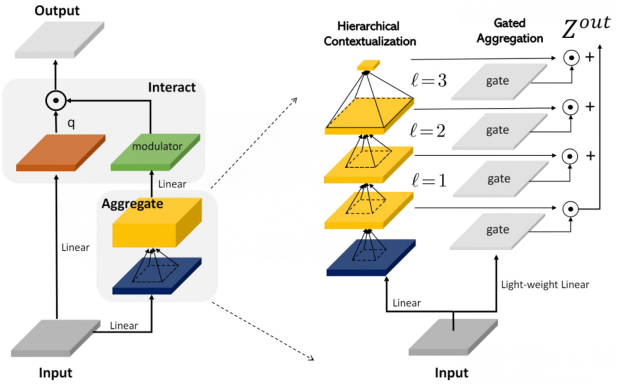

Figure 2. Framework of FocalNet [10].

1.2. Challenge Dataset

The challenge uses the Low-light Instance Segmentation (LIS) dataset, introduced by [3], which contains 892 labeled images as the training set and 669 images as the testing set. The LIS dataset comprises paired images collected across various scenes, encompassing both indoor and outdoor en-

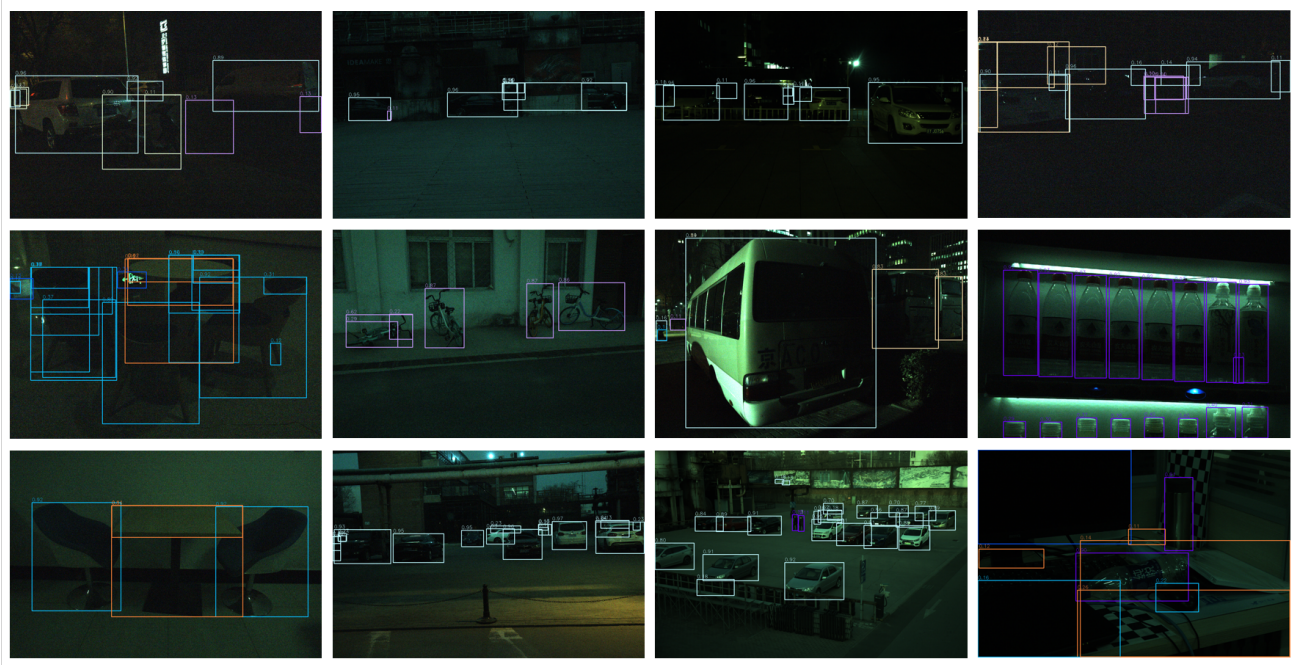

Figure 3. Visual results of our method on the testing set.

vironments. We utilize all labeled data for training and do not perform online evaluations during training. After training, we directly use the last checkpoint to predict the testing data.

1.3. Training and Testing Details

During training, we take the model pre-trained on the Object365 dataset and finetuned on the COCO dataset as the pre-trained model. Specifically, our model is trained on 8 NVIDIA Tesla V100-32G with a total batch size of 8, numbers of queries of 900, and numbers of proposals of 100. Since the training set is small, we train the detector using the AdamW optimizer with an initial learning rate of 0.0001 and weight decay of 0.0001, to alleviate overfitting. We employ the standard 1x schedule to train the model, and random horizontal flipping with a probability of 0.5 and random resize-crop-resize are introduced as weak augmentation.

During testing, simple test-time augmentation like horizontal flipping and multi-scale testing are exploited, in which the scales include ×1.0, ×1.125, ×1.25, ×1.375, and ×1.5. The NMS is not adopted and the detector directly outputs 100 box predictions end to end. Specifically, the initial test image size is 1333x800, and horizontal flipping is adopted to boost model performance. After obtaining ten predictions with different scale augmentation, we further use weighted boxed fusion (WBF) [9] to ensemble them as our final submission, which achieves an AP of 0.76 in the test phase.

In addition, we attempt to introduce some advanced low-light image enhancement methods, such as CIDNet [6], GlobalDiff [7], and Retinexformer [1], to enhance the challenge data, and perform detection algorithm on the enhanced images. Unfortunately, the performance has not been improved or even decreased. We argue that since the challenge dataset does not have pairs of low-light and normal scene images, this leads us to use these image enhancement methods for cross-domain inference, which corrupts the distributional information in the data itself, and ultimately leads to a degradation of detection performance.

Some visual results of our method on the testing set are shown in Fig. 3.

2. Instance Segmentation

2.1. Network Architecture

HTC [2] is adopted as our detector which can learn more discriminative features progressively while integrating complementary features together in each stage, as shown in Fig. 4. To simplify its use, we directly employ the original masks of objects as semantic maps. The most advanced and robust backbone ViT-adapter [4] is utilized to introduce the image-related inductive bias to a plain ViT [5], which allows plain ViT to achieve comparable performance to vision-specific transformers, as shown in Fig. 5. In order to increase the receptive field of each roi feature, we exploit the roi pooling on the feature map of the corresponding level to get the global context feature, which is used to enhance

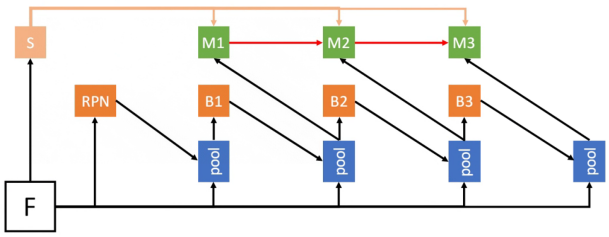

Figure 4. Framework of HTC [2].

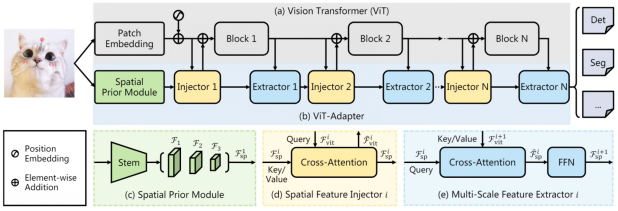

Figure 5. Framework of ViT-adapter [4].

the roi feature of the corresponding level by adding them. We also add SyncBN to each box head to make the training process more stable.

2.2. Challenge Dataset

The challenge uses the Low-light Instance Segmentation (LIS) dataset, introduced by [3], which contains 892 labeled images as the training set and 669 images as the testing set. The LIS dataset comprises paired images collected across various scenes, encompassing both indoor and outdoor environments. We utilize all labeled data for training and do not perform online evaluations during training. After training, we directly use the last checkpoint to predict the testing data.

2.3. Training and Testing Details

During training, we take the model trained on the COCO [8] datasets as the pre-trained model. MMdetection is used to implement our method. Specifically, our model is trained on 8 NVIDIA Tesla V100-32G with a total batch size of 8. Since the training set is small, we train the detector using the AdamW optimizer with an initial learning rate of 0.0001 and weight decay of 0.05, to alleviate overfitting. We employ the standard 1x schedule to train the model, and random horizontal flipping with a probability of 0.5 and random resize-crop-resize are introduced as weak augmentation.

During testing, simple test-time augmentation like horizontal flipping and multi-scale testing are exploited, in which the scales include ×1.0, ×1.125, ×1.25, ×1.375, and ×1.5. The NMS is applied with a score of 0.001 and iou threshold of 0.5. Specifically, the initial test image size is 1200x800, and horizontal flipping is adopted to boost model performance. After obtaining five predictions with different scale augmentation, we further use weighted boxed fusion (WBF) with hard-voted masks to ensemble them as our final submission, which achieves an AP of 0.62 in the test phase.

We also conduct low-light image enhancement experiments and obtain similar results with object detection.

References

[1] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12504–12513, 2023. 2

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, pages 4974–4983, 2019. 2, 3

[3] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *IJCV*, 131(8):2198–2218, 2023. 1, 3

[4] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *ICLR*, 2023. 2, 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[6] Yixu Feng, Cheng Zhang, Pei Wang, Peng Wu, Qingsen Yan, and Yanning Zhang. You only need one color space: An efficient network for low-light image enhancement. *arXiv preprint arXiv:2402.05809*, 2024. 2

[7] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[8] 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 *ECCV*, pages 740–755. Springer, 2014. 3

[9] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021. 2

[10] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *NeurIPS*, 35:4203–4217, 2022. 1

[11] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023. 1