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

Linwei Chen
Beijing Institute of Technology
chenlinwei@bit.edu.cn

Abstract

Performing object detection and instance segmentation under low-light conditions poses several challenges. e.g., *images captured in low-light environments often suffer from poor quality, leading to loss of detail, color distortion, and prominent noise. These factors significantly hinder the performance of downstream vision tasks, particularly object detection and instance segmentation. To address this challenge, the CVPR 2024 PBDL Challenge Low-light Object Detection and Instance Segmentation aims to assess and enhance object detection algorithms' robustness on images captured in low-light environmental conditions. In this report, we present our solution for tackling object detection and instance segmentation in low-light conditions. Specifically, we utilize DINO and Mask DINO as strong baseline models, along with disturbance suppression learning for training to enhance robustness against image noise, and test-time augmentation with a suitable weighted box fusion technique to further improve test accuracy. Ultimately, our solution achieved box/mask AP of 0.75 and 0.59 in the PBDL Challenge Low-light Object Detection and Instance Segmentation.*

1. Network Architecture

Object Detection. We utilized DINO as our foundational network. As shown in Figure 1, DINO is an advanced end-to-end Transformer detector that employs several innovative techniques, including contrastive denoising training, look forward twice, and mixed query selection. These techniques significantly enhance both training efficiency and detection performance. We chose DINO for our competition due to its demonstrated efficiency and robustness in handling complex detection tasks. Its high performance on benchmark datasets make it an ideal choice for achieving competitive results in the specific task of "Low-light Object Detection and Instance Segmentation" competition.

Instance segmentation. We utilized Mask DINO [4] as our foundational network. As shown in Figure 3, Mask DINO is a unified Transformer-based framework designed for both object detection and image segmentation. This network is an extension of DINO, which was originally developed for detection, and adapts it to handle segmentation tasks with minimal modifications to key components. Mask DINO stands out due to its superior performance, outperforming previous specialized models and achieving the best results in instance, panoptic, and semantic segmentation tasks among models with fewer than one billion parameters.

One of the critical advantages of Mask DINO is its ability to enable task cooperation, demonstrating that detection and segmentation can mutually enhance each other within query-based models. Additionally, Mask DINO leverages better visual representations pre-trained on large-scale detection datasets to improve semantic and panoptic segmentation. This synergistic approach not only enhances the per-

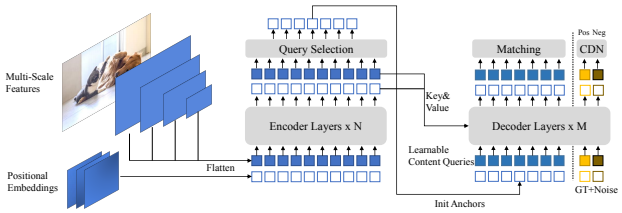

Figure 1. Framework of DINO [7].

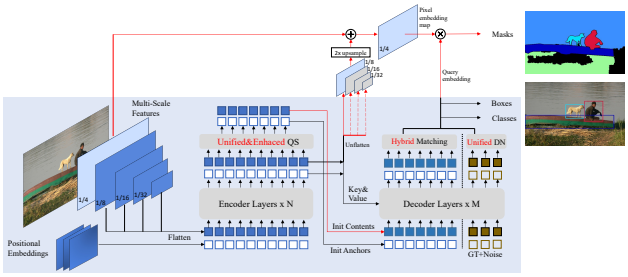

Figure 2. Framework of Mask DINO [4].

Figure 3. Example scenes in LIS dataset. Four image types (long-exposure normal-light and short-exposure low-light images in both RAW and sRGB formats) are captured for each scene.

formance but also provides a robust and versatile framework capable of handling multiple vision tasks effectively. By employing Mask DINO, we aim to leverage these strengths to achieve superior results in the "Low-light Object Detection and Instance Segmentation" competition.

Feature alignment. We integrated the Feature-aligned Pyramid Network (FaPN) [3] to enhance our network. FaPN is a simple yet effective top-down pyramidal architecture designed to generate multi-scale features for dense image prediction. FaPN comprises two key modules: a feature alignment module and a feature selection module. The feature alignment module learns transformation offsets of pixels to contextually align upsampled higher-level features, while the feature selection module emphasizes lower-level features rich in spatial details. Empirical results show that FaPN consistently and substantially improves performance over the original FPN across four dense prediction tasks and three datasets.

We chose FaPN for our competition due to its demonstrated ability to improve multi-scale feature generation. Its integration into our network aims to leverage these strengths, thereby enhancing our model's accuracy in the competition.

2. Low-light Instance Segmentation Dataset

To systematically investigate the effectiveness of the proposed method in real-world conditions, a real low-light image dataset for instance segmentation is necessary and foundamental. The challenge utilizes the Low-light Instance Segmentation (LIS) dataset, introduced by [1, 6].

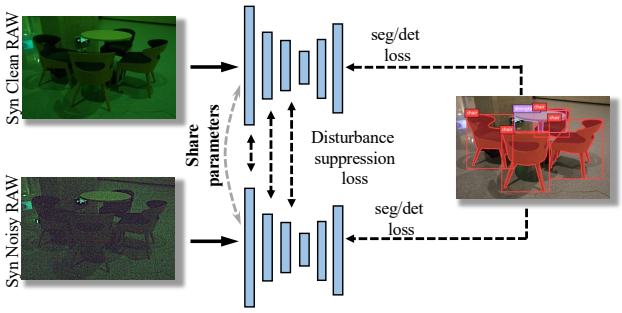

Figure 4. Framework of Disturbance Suppression Learning [1].

It is collected using a Canon EOS 5D Mark IV camera. Figure 3 shows examples of annotated images from LIS dataset. The LIS dataset exhibits the following characteristics:

- **Paired samples:** The LIS dataset includes images in both sRGB-JPEG (typical camera output) and RAW formats. Each format consists of paired short-exposure low-light and corresponding long-exposure normal-light images. We term these four types of images *sRGB-dark, sRGB-normal, RAW-dark, and RAW-normal*. To ensure pixel-wise alignment, we mounted the camera on a sturdy tripod and used remote control via a mobile app to avoid vibrations.
- **Diverse scenes:** The LIS dataset consists of 2230 image pairs collected in various indoor and outdoor scenes. To increase the diversity of low-light conditions, we used a series of ISO levels (*e.g.*, 800, 1600, 3200, 6400) to capture long-exposure reference images and deliberately

decreased the exposure time by various low-light factors (*e.g.*, 10, 20, 30, 40, 50, 100) to capture short-exposure images, simulating very low-light conditions.

- **Instance-level pixel-wise labels:** For each image pair, we provide precise instance-level pixel-wise labels annotated by professional annotators. This results in 10,504 labeled instances across eight common object classes: bicycle, car, motorcycle, bus, bottle, chair, dining table, and TV.

The LIS dataset includes images captured in different scenes (indoor and outdoor) and under varying illumination conditions. As shown in Figure 3, object occlusion and densely distributed objects add to the challenges presented by the low-light conditions.

3. Training and Testing Details

Training details. During training, we use a model pretrained on the Object365 dataset and fine-tuned on the COCO dataset as our base. Our training setup includes 8 RTX 3090 GPUs, with a total batch size of 8. All other settings are kept the same as in the original paper. We follow the standard $1\times$ training schedule and apply weak data augmentation techniques, including random horizontal flipping with a probability of 0.5 and random resize-crop-resize.

Disturbance Suppression Learning. When fine-tuned on COCO, we utilize the low-light RAW synthetic pipeline from [1], which consists of two steps, namely, unprocessing and noise injection, to obtain synthetic low-light clean/noisy RAW images. We adopt disturbance suppression learning from previous work [1]. Ideally, a robust network should extract similar features whether the input image is corrupted by noise or not. To achieve this, we introduce disturbance suppression learning, which encourages the network to learn disturbance-invariant features during training. This approach is independent of architectural considerations.

The total loss for learning is defined as:

$$L(\theta) = L_{\text{IS}}(x;\theta) + \alpha L_{\text{IS}}(x';\theta) + \beta L_{\text{DS}}(x,x';\theta), \quad (1)$$

where x is the clean synthetic RAW image, x' is its noisy version, and α and β are the weights of the respective losses. We empirically set $\alpha = 1$ and $\beta = 0.01$.

The loss L_{IS} is the task loss, *e.g.*, instance segmentation loss, which consists of classification loss, bounding box regression loss, and segmentation (per-pixel classification) loss. The specific formula for L_{IS} is related to the model, we employ the same loss as the origianl model. This loss is applied to both the clean image x and the noisy image x' to ensure the model performs consistently regardless of noise.

The loss L_{DS} is the feature disturbance suppression loss, defined as:

$$L_{\text{DS}}(x,x';\theta) = \sum_{i=1}^{n} \|f^{(i)}(x;\theta) - f^{(i)}(x';\theta)\|_2^2, \quad (2)$$

where $f^{(i)}(x;\theta)$ represents the i-th stage of feature maps of the model. By minimizing the Euclidean distance between the clean features $f^{(i)}(x;\theta)$ and the noisy features $f^{(i)}(x';\theta)$, the disturbance suppression loss encourages the model to learn disturbance-invariant features. This reduces feature disturbance caused by image noise and improves the model's robustness to corrupted low-light images.

Unlike perceptual loss [2], our approach does not require pretraining a teacher model, making our training process simpler and faster. With $L_{\text{IS}}(x;\theta)$ and $L_{\text{IS}}(x';\theta)$, our model can learn discriminative features from both clean and noisy images, maintaining stable accuracy regardless of noise. In contrast, the "student" model in perceptual loss [2] only sees noisy images, which can degrade performance on clean images and limit robustness. Additionally, the domain gap between the feature distributions of the teacher and student models can harm the learning process. By minimizing the distance between clean and noisy features predicted by the same model, we avoid this problem.

Testing details. During testing, we employ simple test-time augmentation techniques such as horizontal flipping and multi-scale testing. The multi-scale testing involves resizing the shorter side of the image to various sizes: 400, 500, 600, 700, 800, 900, 1000, 1100, and 1200 pixels. Horizontal flipping is also used to enhance model performance. For detection, after obtaining ten predictions with different scale augmentations, we use Weighted Box Fusion (WBF) [5] to ensemble them for our final submission.

References

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

[2] Abhiram Gnanasambandam and Stanley H Chan. Image classification in the dark using quanta image sensors. In *ECCV*, pages 484–501, 2020. 3

[3] Shihua Huang, Zhichao Lu, Ran Cheng, and Cheng He. Fapn: Feature-aligned pyramid network for dense image prediction. In *ICCV*, pages 864–873, 2021. 2

[4] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 1

[5] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021. 3

[6] Hong Yang, Wei Kaixuan, Chen Linwei, and Fu Ying. Crafting object detection in very low light. In *BMVC*, pages 1–15, 2021. 2

[7] 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