

Comparative Study of YOLOv11, SSD, and Fast RNN for Nighttime Face Detection on the DarkFace Dataset

Yehui Cai

*College of Computer Science and Technology, Huaqiao University, Xiamen, China
2225102086@stu.hqu.edu.cn*

Abstract: Object detection in low-light scenarios has always been a challenge. The detector, which was been trained in a well-illuminated dataset, presents lower performance in low-light scenarios. This is mainly due to factors such as low image visibility and large noise interference that affect the clarity and distinguishability of the target. To study the performance of different detection models in the task of night face detection, this paper selects three typical object detection models, YOLOv11, Fast Region-based Convolutional Network (Faster R-CNN), and Single Shot MultiBox Detector (SSD), based on the DarkFace night face dataset for performance evaluation and comparison. In the experiment, the DarkFace dataset is used for testing, and mAP@0.5 is used as the main evaluation indicator. The results show that You Only Look Once version 101 (YOLOv11) is significantly better than Faster R-CNN and SSD models in terms of accuracy and F1-recall and has stronger night object detection capabilities, which means that YOLOv11 has better robustness and detection performance in low-light environments and can be used as an effective solution for night face recognition tasks.

Keywords: Low-light object detection, Nighttime face detection, YOLOv11, Faster R-CNN, SSD

1. Introduction

With the fast development of computer vision, object detection has become a key component in various applications such as surveillance, autonomous driving, and face recognition. But in practical scenes, especially for low-illumination (or nighttime) settings, the existing detection models pre-trained on the bright dataset degrade seriously, mainly because the brightness makes the detection model unable to acquire sufficient detection details, resulting in increased noise and lost fine-grained information [1]. Therefore, obtaining a robust object detection model that can work effectively under a low-illumination environment is widely practical and important.

To produce high-quality low-light images, many low-light enhancement schemes, such as Retinex-based, end-to-end image enhancement networks, and Transformer-based enhancements, such as Retinexformer, have been designed recently [2-4]. Although some of such techniques make visual enhancement, detection accuracy can be degraded [5]. Therefore, instead of solely focusing on the pre-processing procedure, it is still necessary to design robust detection architectures in low-light conditions.

This work tries to compare the performance of three typical object detection models: YOLOv11, Faster R-CNN, and SSD on nighttime face detection tasks. The results show that YOLOv11

significantly exceeds the performance of both Faster R-CNN and SSD in terms of mAP and recall, suggesting superior robustness in night-time usage.

2. Method

2.1. Dataset preparation

To evaluate object detection models' performance under the weakly illuminated setting, it adopts the DarkFace dataset, which contains thousands of real-world low-illumination images with the annotation of people's facial areas [6]. The dataset contains severe weak illumination, occlusion, motion blur, and a variety of city backgrounds.

To support various object detection frameworks, it prepares and labels the dataset in two commonly adopted formats: YOLO format and Pascal VOC format. YOLO format is used to train the YOLOv11 model, while the Pascal VOC format is used to train Faster R-CNN and SSD models. The two format annotation scheme ensures uniformity of data from all train sets as well as the interchangeability of various model forms.

2.2. Model architectures

It selects three representative object detection models from both one-stage and two-stage paradigms. YOLOv11, the newest member of the YOLO family, YOLOv11 involves additional state of the art designs: C3k2 (Cross Stage Partial with kernel size 2), SPPF (Spatial Pyramid Pooling – Fast) and C2PSA (Convolutional block with Parallel Spatial Attention), which benefit multi-scale feature fusion, spatial localization and inference speed [7, 8]. Faster R-CNN is a two-stage detection model that combines a Region Proposal Network (RPN) and a Fast R-CNN classifier. It predicts region proposals and uses the proposals for classification and regression of the boxes. It is of high detection precision but at a higher computational cost [9]. SSD (Single Shot MultiBox Detector), a one-stage model, can directly predict the bounding boxes and class probabilities of an image based on multi-scale feature maps and makes a trade-off between detection speed and accuracy, especially for objects with different scales [10].

2.3. Evaluation metrics

It evaluates model performance based on mean Average Precision at $\text{IoU} = 0.5$ ($\text{mAP}@0.5$) and validation loss (val loss). $\text{mAP}@0.5$ is a standard metric for the object detection evaluation, which considers the precision and recall for all images in the test set. The validation loss is computed on the validation set at the end of each epoch and serves to evaluate both the model's convergence and generalization performance. These statistics enable us to evaluate not just the final detection accuracy but also the training dynamics and training stability of each model.

3. Result

This section shows the performance comparison for YOLOv11, Faster RNN, and SSD models in the DarkFace dataset by comparing the mean Average Precision at 0.5 IoU ($\text{mAP}@0.5$) and the validation loss on training epochs. These metrics show a combined performance of final detection accuracy and training convergence behaviour.

3.1. $\text{mAP}@0.5$ performance

Table 1 shows that detectability by the three models is very different. YOLOv11 performs the best, reaching mAP at 0.2056 first, which quickly rises throughout the epochs, finally getting the stable

and high mAP at 0.4204 after the training is completed, indicating the robustness of low-light face detection. The curve of Fast RNN is increasing at a small rate, and it finally reaches a final mAP of 0.1011. Its overall accuracy is lower than SSD but better than the performance of SSD. SSD displays a fast early rise in mAP but plateaus quickly, ending with a final mAP of 0.0593, indicating a weak ability to learn discriminative features in nighttime conditions. This illustrates that YOLOv11 can achieve better detection results with inadequate illumination.

Table 1: Final mAP@0.5 of YOLOv11, Fast RNN and SSD

Model	Final mAP@0.5
YOLOv11	0.4204
Fast RNN	0.1011
SSD	0.0593

3.2. Validation loss analysis

Validation loss illustrates whether the network is likely to overfit the training data. The validation trend graphs in Figure 1 confirm the mAP trends. The loss of YOLOv11 rapidly declines to under 1.0 in 30 epochs, and the gradual decline line shows good convergence and stable model optimization. Fast RNN starts at a loss of 0.77 and descends continuously to about 0.45. As for this figure, it can say that it's rather stable but also slow so it is not a particularly good learning. SSD begins with a steep validation loss (9.63) and decreases smoothly but then becomes very noisy with oscillations until it flattens around 4.5. Moreover, those trends of losses validate that YOLOv11 learns quickly and consistently and also gets a high accuracy, thus demonstrating superior robustness in training using the low-light scenario.

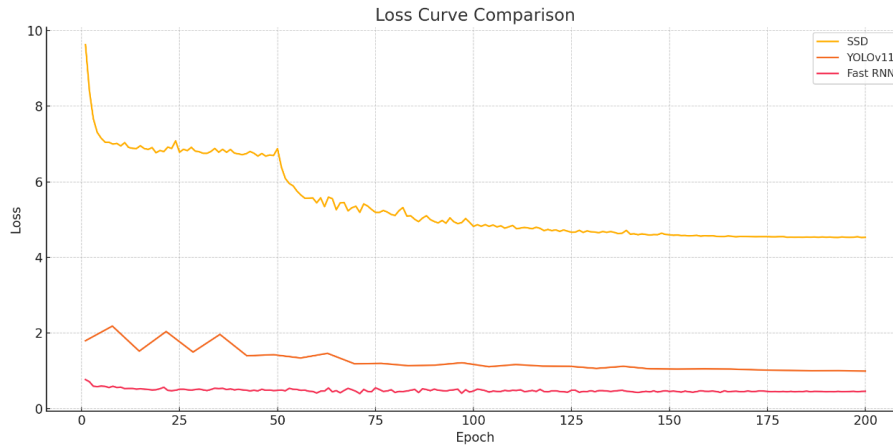


Figure 1: Loss curve of YOLOv11, Fast RNN, and SSD (photo/picture credit: original)

Together, the comparison between mAP@0.5 and the validation loss indicates that YOLOv11 gives the best overall result, which achieves great detection performance with fast convergence of training. The learning performance of the Fast RNN is not bad, but the accuracy is not good. The SSD's performance in the aspect of accuracy and loss stability is the worst.

4. Discussion

4.1. Interpretation of results

YOLOv11 trained by the proposed method outperforms SSD and Faster RNN by a large margin on both test accuracy and convergence training curve. YOLOv11 training with 5 class mAP@0.5 achieves 0.4204 compared with 0.1011 of Fast RNN and 0.0593 of SSD. The loss of YOLOv11 will converge faster than Faster RNN and SSD during training, and its validation loss is less than 1.0 in 30 epochs.

The excellent YOLOv11 is mainly due to these architecture-level design improvements. Following Khanam et al. (2024) [7], YOLOv11 proposed the C3k2 block, SPPF block, and C2PSA spatial attention modules, which bring benefit to feature extracting and spatial location ability - two key aspects of the LL detection task [8].

However, SSD performs real-time detection with a simple architecture, and its detection performance in low-visibility images does not reach a good level [10]. Due to SSD using the fixed default boxes and not having adaptive attention, it does not perceive sufficient context, especially when it is applied to small and occluded images at night. This can be partly explained by the earlier plateau in the early-mAP and the high validation loss during training, whereas the fast training early phase.

Faster RNN maintains relatively constant training and accuracy but is less effective than SSD in challenging dark scenes due to the model's limited spatial representation as well as insufficient hierarchy context fusion. Also, it loses a bunch of architecture optimization, such as YOLOv11.

4.2. Limitations

While the results are encouraging, there are still some constraints to this paper. The evaluation is limited to face detection. Other object categories (e.g., vehicles, animals) under low-light conditions may yield different performance trends. It performs all experiments only on the DarkFace dataset. It is very comprehensive, but it might not capture all low-light variations (e.g., backlight, rain, rural areas). YOLOv11 claims high inference efficiency; however, the real-time FPS (over different hardware platforms, such as mobile and embedded systems) of YOLOv11 cannot be proven in this paper. It did not include any low-light preprocessing or enhancement models, which could further influence detection accuracy. The models were trained with typical settings. No fine-tuning and other complex methods (e.g., learning rate warmup, data balancing) were attempted.

As a follow-up implementation to increase the detection accuracy under low lighting environment, it proposes using Image enhancement-based model preprocessing (e.g., RetinexNet, Retinexformer) to increase the contrast and lightness for input images before detection [2-4]. According to the result from Cai Y, Retinexformer got the PSNR of 25.16 dB on the dataset of LOL-v1, which was 12.2% higher than the Restormer, which is Transformer-based. This improvement shows advantages in target detection tasks in edge computing and resource-sensitive scenarios and can effectively solve the color shift and structural blur problems of existing non-Retinex methods in low-light environments [4].

5. Conclusion

Through this paper, three popular object detection models, YOLOv11, SSD, and Fast RNN, were compared and analysed on the DarkFace dataset for face detection at night time. The results indicate that YOLOv11 exceeds performance over SSD and Fast RNN on DarkFace in terms of detection accuracy (mAP@0.5) and training stability, thus, YOLOv11 could be a more suitable model for nighttime face detection tasks. Features of YOLOv11's architecture, such as C3k2, SPPF, and C2PSA,

made a great contribution to its excellent performance under the situation of low-light conditions. SSD has excellent performance in terms of speed of inference, but the face accuracy in terms of mAP@0.5 and convergent performance in terms of IOU were not ideal in low-light conditions. Thus, SSD does not perform well in critical detection tasks. Furthermore, although the performance of Fast RNN is improved incrementally, it does not have the fast speed of inferencing, and the final performance is weaker than YOLOv11, which implies that the model could not still handle this task with little help from recurrent features alone.

Based on the results described in this section, it concludes that a practical usage for a deployment of YOLOv11 is low light (such as in surveillance and automatic monitoring systems). Extensions of the current work in the direction of multimodal inputs (RGB and infrared) or the use of enhancement techniques for images could improve YOLOv11 robustness in low/high/ and extreme light scenarios. Also, extensions of YOLOv11 into more lightweight models could lead to deployments on the side edge.

In general, YOLOv11 is a very efficient solution for low-light face detection, and more work may be done on the model optimization and hybrid methods to enhance the variety of its application on more realistic scenes.

References

- [1] Chen, C., Chen, Q., Xu, J., & Koltun, V. (2018). *Learning to see in the dark*. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3291-3300).
- [2] Ma, L., Ma, T., Liu, R., Fan, X., & Luo, Z. (2022). *Toward fast, flexible, and robust low-light image enhancement*. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5637-5646).
- [3] Wu, Y., Pan, C., Wang, G., Yang, Y., Wei, J., Li, C., & Shen, H. T. (2023). *Learning semantic-aware knowledge guidance for low-light image enhancement*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1662-1671).
- [4] Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., & Zhang, Y. (2023). *Retinexformer: One-stage retinex-based transformer for low-light image enhancement*. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12504-12513).
- [5] Yang, W., Yuan, Y., Ren, W., Liu, J., Scheirer, W. J., Wang, Z., ... & Qin, L. (2020). *Advancing image understanding in poor visibility environments: A collective benchmark study*. *IEEE Transactions on Image Processing*, 29, 5737-5752.
- [6] Jin, Y., Yang, W., & Tan, R. T. (2022, October). *Unsupervised night image enhancement: When layer decomposition meets light-effects suppression*. In *European Conference on Computer Vision* (pp. 404-421). Cham: Springer Nature Switzerland.
- [7] Khanam, R., & Hussain, M. (2024). *Yolov11: An overview of the key architectural enhancements*. *arXiv preprint arXiv:2410.17725*.
- [8] Jegham, N., Koh, C. Y., Abdelatti, M., & Hendawi, A. *YOLO Evolution: A Comprehensive Benchmark and Architectural Review of YOLOv12, YOLO11, and Their Previous Versions. Yolo11, and Their Previous Versions*.
- [9] Ren, S., He, K., Girshick, R., & Sun, J. (2015). *Faster r-cnn: Towards real-time object detection with region proposal networks*. *Advances in neural information processing systems*, 28.
- [10] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). *Ssd: Single shot multibox detector*. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21-37). Springer International Publishing.