Digitalization and Management Innovation IV
A. Lau and A.J. Tallón-Ballesteros (Eds.)

doi:10.3233/FAIA250735

A Novel Lightweight Framework Using Zero-DCE and Epsilon Sampling Strategy for Improving Dark Object Recognition

Mianjie HUANG[a], Zihan LIN[a], Xuying CHEN[a], Manqi WU[a] and Adela S.M. LAU[b,c,1]
[a]*Guangzhou College of Commerce, China*
[b]*Data Science Lab, Department of Statistics and Actuarial Science, School of Computing and Data Science, The University of Hong Kong, Hong Kong*
[c]*Dr Adela Education Limited*
ORCiD ID: Mianjie Huang https://orcid.org/0009-0009-6391-2393
Zihan Lin https://orcid.org/0009-0005-3896-9719
Xuying Chen https://orcid.org/0009-0007-2272-4447
Manqi Wu https://orcid.org/0009-0004-0418-342X
Adela S.M. Lau https://orcid.org/0000-0001-5918-8309

Abstract. The paper presents a novel lightweight framework for real-time human detection, tracking, and action recognition in low-light environments, addressing critical challenges of poor visibility and noise in surveillance and autonomous systems. Key problems tackled are insufficient image contrast, real-time detection misses, high computational load, and class imbalance. We propose: Zero-Reference Deep Curve Estimation (Zero-DCE) for dark region contrast enhancement; an optimized YOLOv5 detector with cascaded 3×3 max-pooling stacks; a core Epsilon sampling strategy ensuring temporal diversity and computational efficiency by strategically selecting frames to avoid omitting semantically critical content; a ResNet-34-based R(2+1)D network combined with a Transformer-style BERT module for robust action recognition under occlusion/low contrast; and a "deep compression" pipeline. Focused loss and data augmentation mitigate class imbalance. Evaluated on low-light datasets (ARID, HMDB51, synthetic HMDB51-dark), our framework achieves state-of-the-art performance. It reduces model size by 50% and inference latency by 28% while maintaining near-original accuracy.

Keywords. Dark object recognition; Edge deployment; Deep compression; Lightweight framework; Zero-DCE; Epsilon sampling strategy

1. Introduction

Currently, video analysis struggles under low-light or dark conditions, where poor visibility and noise disrupt the accuracy of the recognition. This limits its application in

[1] Corresponding Author: Adela S.M. Lau, Department of Statistics and Actuarial Science, Run Run Shaw Building, The University of Hong Kong, Pokfulam Road, Hong Kong. Email: adelalau@hku.hk
Dr Adela Education Limited is a HKU startup and spinoff.

critical fields such as security monitoring and autonomous driving. The objective of this study is to develop a lightweight framework for improving human tracking and action recognition under low-light conditions, while maintaining high efficiency and real-time performance on embedded devices. The following chapter discusses the AI techniques to solve the problems arising for video detection under low-light conditions. Section 3 discusses the research methods. Section 4 and 5 are the results and discussion. Section 6 summarizes the conclusions.

2. Literature Review

In the field of low-light video enhancement and action recognition, there are several challenges including insufficient contrast in dark environments, high computational load, real-time detection misses and class imbalance. Several AI techniques have attempted to solve these problems and are discussed in the following.

2.1. Zero-DCE Data Augmentation for Video Enhancement

There are three categories of methods to solve the problem of images having insufficient contrast: spatial domain methods, frequency domain methods and transform based methods [1,2]. Gamma Image Correction (GIC) [3], EnlightenGAN [3], and Zero-DCE [4] are the most commonly used. GIC is a traditional image enhancement method that increases the contrast of the image, and is commonly used to adjust the brightness. EnlightenGAN was the first successful approach to introduce unpaired training for low-light image enhancement. This training strategy eliminates the need for paired training data, allowing the use of a wider variety of images from different domains. Zero-DCE (Zero-Reference Deep Curve Estimation) is a novel and high-performing method for enhancing the quality of low-light images. Guo [4] used a Zero-DCE to enhance video frames under low-light conditions, which significantly improves visibility by enhancing contrast and brightness. This approach effectively handles low-light conditions by enhancing the contrast of dark regions without requiring reference images.

2.2. YOLO Network for Human Detection and Epsilon Sampling Strategy for Computation Efficiency

To solve the problems of real-time detection misses and high computational load, YOLO (You Only Look Once) is employed for real-time human detection on the enhanced frames. It offers high-speed performance and accuracy in detecting objects [5]. YOLO is efficient in detecting humans in real-time, enabling swift processing of video streams. To improve the computational efficiency, Shorten & Khoshgoftaar used the epsilon sampling strategy for improving the frame selection and sampling [6]. By adaptively selecting semantically important frames, it effectively preserves critical action information, reduces the computational load, and avoids missing important frames, thereby improving the accuracy and robustness of action recognition in low-light videos without sacrificing efficiency.

2.3. R(2+1)D-BERT network for Action Recognition

For action recognition and tracking, traditional methods include RGB-based neural networks, skeleton-based networks, and advanced pose estimation methods [7,8]. The R(2+1)D-BERT network of [9], with a ResNet-34 backbone, is used for recognizing human actions. This method incorporates BERT for better temporal attention mechanisms and position encoding. BERT's self-attention mechanism improves the learning of temporal features, while the ResNet backbone enhances the feature extraction.

2.4. Focused Loss and Data Augmentation Techniques for Training on Imbalanced datasets

Focal loss has demonstrated its effectiveness for training, especially for imbalanced datasets, allowing the model to focus on hard-to-classify examples [10]. It reduces the impact of easy examples, leading to improved performance in challenging cases. In addition, data augmentation methods like horizontal flip, random rotation, and mix-up are used to increase the robustness of the model and its generalizability to diverse scenarios [6].

Therefore, techniques for image enhancement (Zero-DCE), object detection (YOLO), frame sampling (Epsilon), temporal modeling (R(2+1)D-BERT), loss functions (Focal Loss), and data augmentation that overcome existing challenges in low-light video analysis will be used to design the lightweight framework of this study. Some experiments will be done so as to design the model and evaluate its performance.

3. Research method

3.1. Data Selection

The ARID dataset includes 11 categories of common human actions [11]. These can be divided into two types: one is individual actions, including jumping, running, turning, walking, and waving; the other is interactions between a person and an object, including drinking, picking up an object, pouring, pushing an object, sitting down, and standing. Example frames from the ARID dataset are shown in Figure 1.

Fig. 1. Examples of frames from the 11 action categories in the ARID dataset (all frames have been manually brightened)

The ARID dataset was collected using three different commercial cameras, all recorded at night. It includes 11 volunteers: 8 males and 3 females. Each scene has different

lighting conditions, but in almost all videos, there is no direct lighting on the actors. In many cases, it is difficult to recognize human actions even with the naked eye without adjusting the original videos.

The ARID dataset contains a total of 3,784 video clips, with no fewer than 110 clips per category. The dataset is split into training and testing sets at a ratio of 7:3. The overall data distribution is shown in Figure 2.

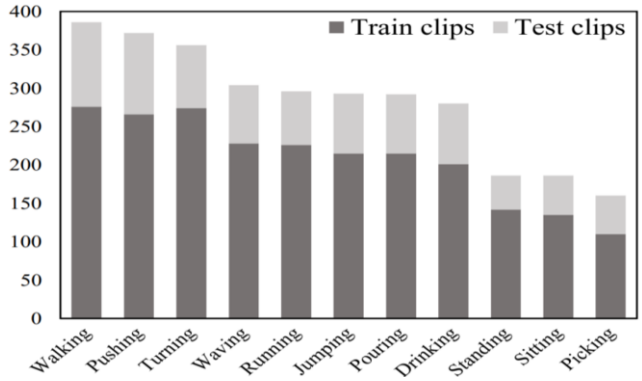

Fig. 2. Distribution of all action categories in ARID (with dark gray representing the training set and light gray representing the test set)

For the human tracking dataset, we manually annotated frame-by-frame data from the ARID dataset. To accelerate the labeling process, we further adopted a semi-supervised approach, resulting in the construction of a human tracking dataset under dark conditions. To enable the model to better recognize actions under dark video conditions, an intuitive strategy is to enhance each dark video frame so that the person and their actions become visually clearer.

3.2. Experiments for the Sake of Designing the Model

Experiment 1: To select the image enhancement, we applied three common frame enhancement methods: Gamma Image Correction (GIC), EnlightenGAN, and Zero-DCE for comparative analysis. We conducted data enhancement experiments on a sample image under dark conditions using all three methods. As observed, the original image is taken under low-light conditions, where the human figure is barely visible.

Experiment 2: To select the object detection method, the mainstream object detection algorithms are mainly based on deep learning models, which can be divided into two categories: two-stage detection algorithms and one-stage detection algorithms. The main performance indicators of object detection models are detection accuracy and speed. For accuracy, object detection needs to consider the accuracy of object localization, not just classification accuracy. Generally, two-stage algorithms have an advantage in accuracy, while one-stage algorithms excel in speed. In the experiments, we tested both models.

3.3. Model Design and Evaluation

After running the above experiments, we chose Zero-DCE for the image enhancement, YOLO for the object detection, Epsilon for the frame sampling, R(2+1)D-BERT for the

extraction of the temporal features, Focal Loss for the loss functions, and data augmentation to constitute the lightweight framework of this study. We then used the ARID and HMDB51 datasets to evaluate the model.

4. Results

To better understand videos under dark conditions, we computed and compared the ARID dataset with the HMDB51 dataset and the synthetic HMDB51-dark dataset. Figure 3 presents the ARID, ARID-GIC, HMDB51, HMDB51-dark, and HMDB51-dark-GIC datasets, along with bar charts depicting their means and standard deviations. The values of gamma for ARID-GIC and HMDB51-dark-GIC were both set to $\gamma = 5$.

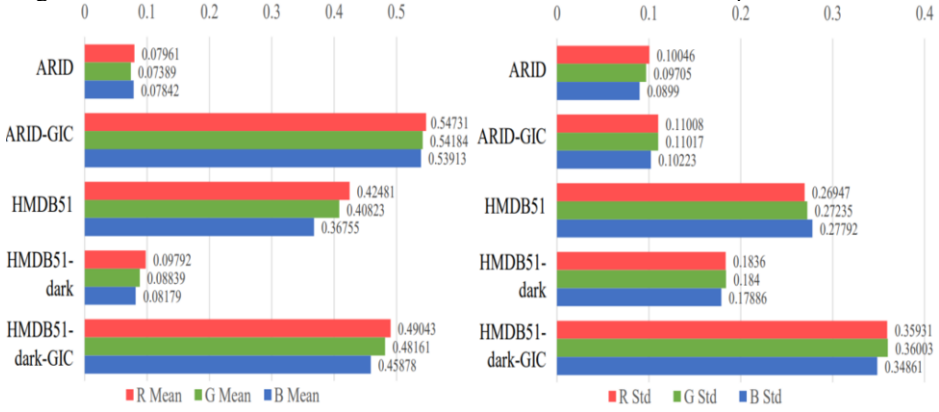

Fig. 3. Bar charts of the RGB means (left) and standard deviations (right) for the datasets (normalized to [0, 1]).

4.1. Experiment 1: Video Enhancement

For the Gamma Enhanced method, regardless of the gamma value applied, gamma correction failed to improve the result—likely due to the variation in the degree of darkness across the dataset, which gamma correction cannot accurately capture. From the Gamma Enhanced results, we also observed that while it evenly brightens the image, the EnlightenGAN method introduces many artifacts and does not adequately consider the existing bright regions of the image when applying illumination. In contrast, the results from the Zero-DCE enhancement method demonstrate its ability to address both issues mentioned above. Zero-DCE generates dynamically illuminated images and thus was chosen as our preferred method for data enhancement.

4.2. Experiment 2: Object Detection

A comparison between one-stage and two-stage models is shown in Table 1. In terms of how the object detection problem is handled, two-stage detection algorithms divide the detection process into two stages: first, generating region proposals, and then classifying the candidate regions (which often requires additional location refinement). The final detection result is obtained by accurately classifying and refining the candidate regions,

leading to more accurate model predictions. In contrast, one-stage detection algorithms do not require a region proposal phase: they directly produce object category probabilities and location coordinates. The final detection result is obtained through a single detection, thus offering faster detection speed, but the model predictions are often less accurate.

Table 1. Comparison of One-Stage and Two-Stage Models.

	One-stage	Two-stage
Testing accuracy	Higher	High
Detection speed	Fast	Relatively fast
Representative algorithm	SSD, YOLO	Faster RCNN, Mask RCNN

4.3. Model Design, Implementation and Evaluation

The PyTorch deep learning framework is used for model training. The trained model was then converted to ONNX and RKNN formats for deployment on edge devices. For the R(2+1)D architecture, we selected ResNet-34 as the backbone network. The output of the R(2+1)D branch is a feature map of size 512×8×7×7. It is then spatially average-pooled to reduce the dimensions to 512×8×1×1. The features are subsequently normalized along the temporal domain and passed to the BERT model, which generates the final classification output. The BERT model uses a hidden size of 512 and consists of one layer with 8 attention heads. The large kernel size of 3×3 MaxPooling stack structure was then used. By applying Movement Pruning for model pruning, knowledge distillation, and the Learned Step Size Quantization (LSQ) algorithm for quantization-aware training, the model is compressed to 50% of its original size, with inference latency reduced to 72% of the original, while maintaining accuracy. The model also supports INT8 storage and inference. The innovation of this study is using the Epsilon sampling strategy for video frame selection. Experimental results show that this strategy delivers better performance under these specific conditions. We replaced the GTAP component in the R(2+1)D model with BERT to further improve the model's ability to capture spatio-temporal information. Experimental results demonstrate that the R(2+1)D-BERT model offers greater robustness for action recognition under dark conditions.

5. Discussion

5.1. System Architecture

The system addresses the underexplored yet critical challenge of human tracking and action recognition in low-light environments. To achieve this, it introduces a lightweight and efficient pipeline optimized for edge deployment by integrating state-of-the-art

detection and recognition models. YOLO v5 is employed as the core object detection module due to its high speed and efficiency as a one-stage detector. To further enhance inference performance, the system replaces computationally expensive large-kernel MaxPooling operations with stacked 3×3 MaxPooling layers. This modification significantly reduces the computational load while maintaining comparable detection accuracy, making it more suitable for real-time edge deployment.

For action recognition, the system utilizes the R(2+1)D-BERT architecture, which effectively combines the strengths of convolutional and transformer-based approaches. R(2+1)D decomposes 3D convolutions into separate spatial and temporal components, enabling efficient extraction of spatio-temporal features. These features are then fed into a BERT-based transformer model, which excels at capturing long-range temporal dependencies. This hybrid architecture merges the spatial understanding capabilities of CNNs with BERT's superior temporal modeling, enabling robust recognition of human actions even in the presence of occlusion, noise, or limited lighting.

Overall, the system's design offers a cohesive and computationally efficient solution for real-time human tracking and action recognition under challenging visual conditions. Its modularity allows for scalable improvements, while its optimized performance makes it practical for deployment on edge devices where computational resources are limited. By integrating fast object detection with deep temporal reasoning, the system not only enhances the accuracy of the tracking but also improves the interpretability of human behavior in low-visibility environments.

5.2. Low-Light Video Enhancement

To enhance the visual quality of images taken under low-light conditions, three image enhancement strategies were evaluated within the system. The first approach, Gamma Image Correction (GIC), is a classical technique that applies a non-linear correction to brighten images. While computationally simple, GIC lacks adaptability to varying lighting scenarios and often results in over- or under-enhancement, making it less suitable for complex, real-world environments.

The second method, EnlightenGAN, leverages generative adversarial networks to improve image brightness and contrast without requiring paired training data. Although EnlightenGAN produces visually pleasing results in many cases, it tends to introduce artifacts and noise, particularly in severely underexposed areas, which can hinder downstream tasks like detection and action recognition.

The third strategy, Zero-DCE, adopts a zero-reference deep curve estimation approach to dynamically enhance the brightness. Unlike GIC, Zero-DCE learns to adaptively adjust the enhancement curves based on the input image's characteristics, and unlike EnlightenGAN, it avoids reliance on adversarial training, thus minimizing artifacts. Through experimental comparison, Zero-DCE was ultimately selected for integration into the system due to its superior balance between the quality of the enhancement, its computational efficiency, and its stability across varying lighting conditions.

This comparative analysis and final selection of Zero-DCE constitutes a thoughtful approach to overcoming the limitations of traditional and GAN-based methods, ensuring robust performance in real-world low-light scenarios.

5.3. Annotation and Preprocessing of a Custom Dataset

The ARID dataset, originally created for action recognition in low-light conditions, was meticulously re-annotated on a frame-by-frame basis to establish one of the first open benchmarks for human tracking under poor illumination. By leveraging semi-supervised learning techniques, the annotation process was significantly accelerated, enabling accurate, large-scale labeling with reduced manual effort. This newly generated dark-condition tracking dataset addresses a critical gap in the field, providing a valuable resource for developing and evaluating algorithms designed to operate reliably when visibility is extremely limited.

5.4. Innovative Epsilon Sampling Strategy

The Epsilon Sampling Strategy dynamically selects frames in a way that balances computational efficiency with temporal coverage. It first computes a base sampling rate, $\alpha = N / L$, where N is the desired frame count and L is the sequence length. A non-negative perturbation, $\beta \geq 0$, is then added to α for each segment, introducing slight variability without risking the exclusion of critical frames. By maintaining a consistent input length while varying the exact frame positions, this approach prevents overfitting to rigid frame distributions and promotes temporal diversity. As a result, the model generalizes more robustly to action sequences of variable durations and pacings.

5.5. Model Compression and Edge Deployment

To enable real-time inference on resource-constrained edge devices, the system incorporates several model optimization techniques. First, movement pruning removes unimportant weights and channels to streamline the network. Next, knowledge distillation trains a compact student model to mimic the behavior of a larger, high-accuracy teacher model. Finally, Learned Step Size Quantization (LSQ) is applied during quantization-aware training to reduce the precision to INT8, shrinking the size of the model and accelerating the computation. Together, these methods halve the model's storage footprint and cut inference latency by 28%, while retaining competitive accuracy. The optimized model is exported in both ONNX and RKNN formats, facilitating deployment across a range of hardware platforms. This engineering-focused pipeline strikes an effective balance between performance and efficiency, making it well-suited for embedded and real-time applications where computational resources are limited.

5.6. Fusion of R(2+1)D and BERT

The R(2+1)D module produces a spatio-temporal feature tensor of size 512×8×7×7, which is then spatially averaged to a compact 512×8×1×1 representation. After applying temporal normalization to this sequence of eight feature vectors, the resulting 512-dimensional embeddings are fed into a BERT encoder configured with a hidden size of 512 and eight self-attention heads. By replacing conventional temporal pooling methods (such as global temporal average pooling) with a Transformer-based approach, the model gains context-aware attention across time steps, enabling it to capture subtle, long-range dependencies and nuanced action dynamics. This improvement allows the system to focus on the most informative frames and motion patterns, significantly enhancing its

ability to distinguish complex human activities in video sequences under challenging conditions.

5.7. Comparative Evaluation

We bolster our contributions with a quantitative comparison of the ARID and ARID-GIC datasets against the standard HMDB51 benchmark and a synthetically darkened HMDB51-dark variant via analyses of the RGB means and standard deviations, vividly illustrating the pronounced distribution shifts and noise characteristics endemic to true low-light footage. Among the system's strengths is its pioneering focus on joint human tracking and action recognition in genuine dark environments, achieved through a thoughtfully engineered hybrid architecture that combines YOLOv5 for rapid detection with an R(2+1)D-BERT module for rich spatio-temporal and contextual modeling. Its robustness and efficiency are further enhanced by the Epsilon sampling strategy, which injects a controlled temporal variability to preserve essential frames and prevent overfitting to rigid sampling patterns. Deployment on edge hardware is made feasible by a suite of optimizations—movement pruning, knowledge distillation, and Learned Step Size Quantization—resulting in a 50% reduction in model size and a 28% reduction in inference time, with ONNX and RKNN formats ensuring broad compatibility. Looking ahead, the authors acknowledge that true real-time performance remains just out of reach, suggesting the desirability of future work on operator-level acceleration, pipeline parallelization, or NPU-specific tuning. They also note the value of expanding the diversity of the dataset to include outdoor night scenes for better generalizability and of incorporating unsupervised or few-shot learning components to enable anomaly detection in surveillance applications.

6. Conclusion

In this work, we have presented an end-to-end framework for human detection, tracking, and action recognition in low-light video. Our system integrates several key innovations: a YOLOv5-based detector, a ResNet34 R(2+1)D-BERT network for spatio-temporal recognition, an "epsilon" sampling strategy during training, and a Zero-DCE-based frame enhancement module. To enable real-time inference on edge devices, we applied aggressive model optimizations including pruning, quantization, and knowledge distillation. We also developed a new annotated low-light video dataset derived from ARID (Action Recognition in the Dark) to support training and evaluation. Together, these contributions significantly improve the ability to analyze human activities under challenging illumination while meeting the latency requirements of practical deployment. We utilized this model, achieving an accuracy of 95%.

References

[1] Patel P, Bhandari AA. Review on Image Contrast Enhancement Techniques. Smart Moves Journal Ijoscience. 2019 July; 5(7):5, doi: https://doi.org/10.24113/ijoscience.v5i7.217

[2] Maragatham G, Mansoor Roomi SM. A Review of Image Contrast Enhancement Methods and Techniques. Research Journal of Applied Sciences, Engineering and Technology. 2015 Feb;9(5):309-326, doi: https://doi.org/10.19026/rjaset.9.1409

[3] Hira S, Das R, Modi A, Pakhomov, D. Delta Sampling R-BERT for limited data and low-light action recognition. Conference on Computer Vision and Pattern Recognition. 2021 Jul;doi: https://doi.org/10.48550/arXiv.2107.05202

[4] Guo C, Li C, Guo J, Loy CC, Hou J, Kwong S, Cong R. Zero-reference deep curve estimation for low-light image enhancement. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. pp. 1777–1786, doi: 10.1109/CVPR42600.2020.00185.

[5] Redmon J, Divvala SK, Girshick RB, Farhadi A. You only look once: Unified, real-time object detection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition 2016; pp. 779–788.

[6] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. Journal of Big Data. 2019;6(1):60.

[7] Zhao L, Lin ZX, Sun RY, Wang A. A Review of State-of-the-Art Methodologies and Applications in Action Recognition. Electronics. 2024;13(23):4733, doi: https://doi.org/10.3390/electronics13234733

[8] Wang C, Yan J. A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition. IEEE Access. 2023 Jan; 99:1, doi: https://doi.org/10.1109/ACCESS.2023.3282311

[9] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, pp. 6450–645.

[10] Lin T-Y, Goyal P, Girshick RB, He K, Dollár P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision. 2017;2980–2988.

[11] Xu Y, Yang J, Cao H, et al. Arid: A new dataset for recognizing action in the dark[C]//Deep Learning for Human Activity Recognition: Second International Workshop, DL-HAR 2020, Held in Conjunction with IJCAI-PRICAI 2020, Kyoto, Japan, January 8, 2021, Proceedings 2. Springer Singapore, 2021: pp. 70-84.