Person Detection in Low-Light Environments: Evaluation of an Annotation-Free Approach

Isabelli P. Gomes, and Flávio L. de Mello

***Abstract*—This work investigates how to modify an existing dataset to enhance the capability of a predictive model' to detect objects in low-light environments. From an initial set of images captured under daylight conditions, synthetic darkened versions were generated. Subsequently, various training scenarios were organized to evaluate the impact of image transformation techniques and data augmentation strategies on predictive performance. The experiments performed demonstrate that the combination of original images and artificially darkened versions, coupled with data augmentation and neutral background images, results in more stable and accurate models for low-light detection, with negligible performance degradation in visible-light scenarios. The results highlight that it is possible to improve detector performance in real-world low-light settings without the need for new image acquisitions or additional annotation.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/10779

***Index Terms*—Low Light Environments, Dataset Construction, Data Augmentation, Image Enhancement, Object Detection, Deep Learning**

I. Introduction

THE current paper addresses the generation of synthetic images for training convolutional neural networks (CNNs) [1] dedicated to object detection within the field of computer vision. The focus lies on developing a procedure to transform daytime images into nighttime versions, enabling dataset expansion and enhancing network performance for applications in environments with low or no illumination.

The research concentrates on images obtained via conventional cameras. The developed method is not intended for systems that already utilize thermal cameras or any other specialized form of sensing since they produce very different images from the ones obtained in the visible spectrum. Furthermore, this study focuses on improving the predictive process without promoting changes to the underlying deep learning architecture. In the current work, primarily for the ease of image acquisition and to streamline the study, person detection in videos was chosen.

The artificial intelligence market is currently highly active, driven by the development of new applications across various sectors. A prominent field is the use of AI techniques for object identification in images—a process widely employed in security and monitoring. However, such image identification systems often suffer from reduced performance when using nighttime images because low light creates high photon noise, low contrast, and color distortion. These factors obscure critical edges and textures that the network relies on to identify objects, leading to false detections and classification errors. In these situations, it is expected that performance will drop by a factor between 3 or 4 [2], [3].

Nevertheless, for such identification to achieve high accuracy, the existence of representative datasets is fundamental. Constructing these datasets requires a large volume of images that must capture specific details and present sufficient variety to allow object recognition under diverse conditions. Moreover, most training is conducted using daytime images, and there is a scarcity—or near absence—of datasets containing nighttime images. These images only comprise 0,07% of an image collection when merging well-known datasets like Microsoft COCO, PASCAL VOC and ImageNet [4]. Consequently, due to the lack of adequate training images, recognition and detection systems tend to demonstrate reduced performance when applied in dark environments, even when infrared illumination is activated [5]. A trivial solution would be to collect images during low-light hours or employ infrared sensing; however, this adds further costs to the dataset construction stage, which is traditionally an economic, temporal, and operational barrier to project viability [6], [7]. Acommercial service may charge from 0.02 USD to 1.00 USD per bounding box annotation, and 0.05 USD to 4.00 USD per image acquisition (for instance, see CVAT, Scale AI, LabelBox, Kili, O-Mega AI, etc for quotes).

Hence, the novelty of the current work lies in an image processing-based method to transform traditional daytime images into nighttime images, enabling the use of synthetic datasets—featuring more diverse and balanced imagery—in the training of detection neural networks. The boundary conditions are atypical for scientific research, which is generally directed toward developing increasingly refined replacements for established models in the mature state of the art that are not easily absorbed by the industry. The requirements of this research dictate the use of well-known scientific techniques and straightforward image transformations which are combined to fill a market knowledge gap, resulting not in an approach, but in a novel resolution to a problem. The contribution is to provide a very low-cost operational solution to the performance degradation problem in 24/7 image-based detection systems—a challenge that often only appears at the end of the intelligent systems production chain and yet which significantly impacts the perceived quality of the products.

The associate editor coordinating the review of this manuscript and approving it for publication was Giner Alor-Hernández (*Corresponding author: Flávio Luis de Mello*).

This project is supported in part by Coppetec Foundation DEL/Poli/UFRJ (Project Poli-19.257).

I. P. Gomes and Flávio Luis de Mello are with the Electronics and Computer Department from Polytechnic School at Federal University of Rio de Janeiro, Brazil (e-mails: isabelli.pgomes@poli.ufrj.br, and fmello@poli.ufrj.br).

II. Related Works

Ma et al. [8] and Agnew [9] show that the quality of datasets and image annotations has a direct impact on the performance of detection models. More precise annotations in conventional datasets, for instance, resulted in significant improvements during testing, even without changes to the model or hyperparameters. The Exclusively Dark Image Dataset [10] (ExDark) is a low-light environment dataset option containing 7,363 images and 12 reasonably balanced classes. This dataset has been used in many studies as a benchmark and presents satisfactory results, such as in the recent work by Ren et al. [11]. However, results are often limited to a restricted set of classes and employ complex architectures with low availability, which hinders their application in commercial environments and everyday scenarios.

One widely investigated research line involves the fusion of images captured simultaneously in the visible and infrared spectra. The FCDFusion [12] work proposes an efficient fusion technique with low color distortion, prioritizing performance and structural fidelity. Similarly, Donia et al. [13] apply convolutional neural networks to generate visual compositions with improved contrast from IR-RGB pairs. Zhu and Zhang [14] improve detail visibility and target contrast of fused infrared and visible images based on the characteristics of low and high-frequency layers, a brightness correction function, and a detail adjustment function. Another approach is taken by Kim et al. [15], who explore cross-spectral domain translation using GANs to convert near-infrared (NIR) images into color versions, while preserving semantic mapping. Ma et al. [16] also use GANs for visible-to-infrared image translation to enrich infrared data, providing a theoretical reference for image generation in tasks such as multisource object detection and data association. While these works present important advances in multispectral information representation, they all rely on data obtained from specific sensors, limiting their applicability in contexts where such resources are unavailable.

Another relevant branch focuses on enhancing the visual quality of images captured in dark environments, seeking to restore information degraded by low exposure. The study by Guo et al. [17] presents a comprehensive review of classical and modern methods, including techniques based on adaptive tone mapping, histograms, and deep networks. Luo et al. [18] describe a low-light image enhancement method to restore images taken under poor lighting using a self-paced learning-based approach. Saifullah and Drężewski [19] describe particle swarm optimization combined with histogram equalization preprocessing for medical image segmentation, focusing on lung CT scans and chest X-ray images, which represent likely low-light scenarios. Furthermore, Sriram and Sreekumar [20] discuss the main challenges in enhancing degraded images—including underexposure, loss of contrast, and color noise—categorizing algorithms according to the type of visual distortion. Mahdizadeh et al. [21] examine the direct impact of illumination on the efficacy of object detection models, reinforcing the importance of visual preprocessing in adverse lighting scenarios. All these works typically start from real low-light images or infrared captures to transform them into visible, well-illuminated images.

Digital image processing frequently faces challenges related to intense shadows, underexposed regions, and sudden lighting variations. One of the most relevant techniques for handling these varying luminosity conditions is Shadow Recovery [22], which seeks to restore details in dark areas of an image without compromising the visual balance of well-lit regions. To achieve this, non-linear transformations are applied to the image histogram, prioritizing local contrast and edge preservation. In the current work, the concept of Shadow Recovery was adapted with an objective opposite to the conventional one: instead of recovering shadows, the technique was used to artificially create shadowed regions in originally well-lit images. Professional photographers refer to this concept as Highlight Recovery [23], however, the term carries a completely different meaning within the context of computerized image processing.

III. The Proposed Method

There are two primary approaches to improving results in an object detection system: (1) those with access to the inference phase; and (2) those with access to the predictive model training phase (see Figu. 1). In the inference phase, input images can be preprocessed to extract features that facilitate the task of the predictive model. In this case, preprocessing introduces additional latency to the inference pipeline, which may compromise real-time applications—a critical requirement in current market products. In the model training phase, one can develop new deep learning architectures or improve existing ones to obtain more satisfactory results across both bright and dark contexts. It is also possible to increase the representativeness of the dataset and submit it to an existing architecture. The first option is often unfeasible for commercial projects, as it requires specialized, expensive labor and extensive testing and refinement, typically restricting it to research environments. The second option is more viable commercially, although still costly and time-consuming, as it involves new image acquisition and manual bounding box annotation. Since the context of the current work is the performance degradation of 24/7 detection systems in commercial environments, developing or refining architectures from scratch is cost-prohibitive. Thus, the remaining viable strategies involve transforming input images, during either training or inference.

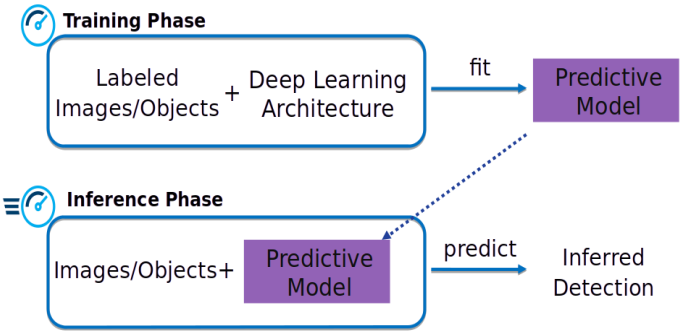

Fig. 1. Classic schematic for the construction and deployment of Deep Learning solutions.

The performance degradation of 24/7 image-based detection systems during nighttime or low-light periods is notably due to the inadequacy of the predictive model. A network's generalization capability depends directly on the variety of examples provided during training. It is uncommon for datasets to feature a balanced variety of daytime/illuminated and nighttime/dark images of the target objects. Generally, dataset images are almost exclusively from daytime or well-lit environments, while nighttime or poorly lit images are rare exceptions.

In the current work, the transformations applied to convert daytime/illuminated images into nighttime/dark versions are point operations based on pixel intensity values, thus, they do not alter pixel coordinates. In the inference phase, this is a highly desirable characteristic as these operations are easily optimized by GPUs. In the training phase, assuming that daytime images have already been annotated with bounding boxes, these annotations can be reused for the newly generated images. This results in data augmentation without a significant increase in the workload required for data preparation.

In the inference phase, variations in brightness and contrast were applied, along with a shadow recovery filter and a combination of multiple preprocessing transformations, with each action generating a distinct group of images. The values assigned to these variables are not universal, and there is no mathematical formula to derive them; the criterion used is visual and specific to the given scenario. The goal is to brighten the images to highlight objects, making their edges and, if possible, textures more evident. The transformation is then applied to the scenario's image batch. Consequently, for each new scenario, new values must be defined and applied to the corresponding batch.

For the training phase and dataset construction, the visual criterion for image transformation was maintained, as was the invariance of variable values when applying transformations within the same scenario. First, a simplistic approach was taken, producing three progressively darkened versions of a color image. In this process, the color image was converted to grayscale, and the exposure was reduced using values of 32, 16, and 8, respectively, which emulates camera shutter speed and enables artificial darkening. This is achieved by compressing input pixel values into a reduced range—i.e., values from 0 to 255 are linearly mapped to a range between 0 and the desired limit (32, 16, or 8).

Subsequently, data augmentation was performed on the color images, and background images without the target object were introduced, followed by a new application of the darkening process to the augmented images. These groups of image transformations were combined to construct new datasets which, when submitted for training, produced different predictive models. Each predictive model exhibits distinct behavior, and the goal is to identify a model that performs optimally across daytime/nighttime and bright/dark environments.

IV. Experiment Description

The deep learning architecture selected for experimentation was YOLOv5 by Ultralytics [24], specifically utilizing the yolov5s.pt weight file. This architecture was chosen due to its status as a standard, widely deployed model in commercial environments. It uses Sigmoid Linear Unit activation function in all of its hidden convolutional layers, and the detection head uses Sigmoid. This version adjusts its weights during training using backpropagation governed by the Stochastic Gradient Descent and Exponential Moving Average. During the inference phase, the network was employed as-is; for the training phase, transfer learning was performed starting from the same weights file. While YOLOv5 is capable of detecting 80 object classes, in the current work only the "person" class was used to minimize the number of variables in the experiment. All video samples employed contain only a single person in the scene. Consequently, the detection of two individuals within a single frame is interpreted as an error. This significantly simplifies the evaluation of the correctness of the generated output. The confidence threshold for object bounding box detection was set at 20%. Code, datasets and video samples are available at Github repository [25].

For video acquisition, an Intelbras VIPC 1230 GB G2 Bullet 2MP PoE camera was used. The RTSP stream was captured using ffmpeg 4.4.2, generating video files with a 1920x1080 resolution using the MPEG-H Part 2/HEVC (H.265) codec. This setup represents an off-the-shelf camera and a standard MPEG video. The camera also supports infrared (IR) capture, which produces grayscale recordings.

All produced videos featured only one person in the scene, and there were 17 different scenes, each with its own recording takes. In total,12 scenes were used for training, and 5 scenes for validation. Each take was recorded three times: first with visible lighting (color video), second with infrared illumination (grayscale video), and third with no lighting. To ensure consistency between takes, a script was followed to dictate the actions, timing, and movement within the scene. Actors rehearsed until they could replicate the performance almost identically. In the current research, it is assumed that these minor variations do not compromise the comparative analysis. Concurrently, it is acknowledged that small spatial and temporal variations can influence detection confidence. Although this dynamic is neither readily demonstrable nor easily validated through direct experimentation, explicitly addressing this assumption is critical for contextualizing the evaluation boundaries.

A. Inference Phase

During this phase, several scenes were filmed in an indoor nighttime environment. Fig. 2a shows the color image illuminated by overhead fluorescent lamps; Fig. 2b corresponds to the image captured with the lamps off and infrared illumination active; and Fig. 2c represents the configuration with neither fluorescent nor infrared lighting. Thus, the first and second images serve as control scenarios, while the third represents the polar opposite case. In this phase, the predictive model cannot be altered; therefore, only preprocessing of the input images can be performed. Consequently, the images in Figs. 2d, 2e, and 2f were derived from Fig. 2c.

Fig. 2d corresponds to a pre-processing stage where brightness and contrast variables were adjusted. The maximum pixel value for the grayscale Look-Up Table (LUT) was modified to 150, with brightness set at 45% and contrast at 53%. In Fig. 2e, shadow recovery was applied with an exposure factor of 3.87. Finally, for Fig. 2f, the following operations were applied sequentially: shadow recovery with an exposure factor of 3.87; 53% contrast; histogram equalization; a bilateral filter with a 9x9 mask, color space sigma of 75, and coordinate space sigma of 75; and a Gaussian blur with a 5x5 kernel.

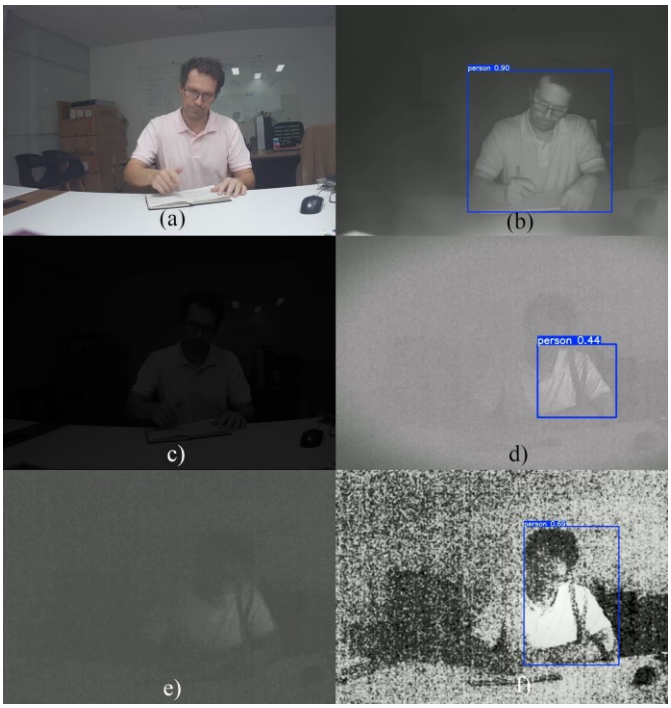

Fig. 2. Example of images acquired under various configurations: (a) color and illuminated by fluorescent lamp; (b) grayscale with infrared; (c) dark without any illumination; (d) with brightness and contrast adjustments; (e) with shadow recovery; (f) with a combination of transformations.

B. Training Phase

In this stage of the experiment, only the generation of the predictive model was varied based on the different datasets used for training. Furthermore, as a 24/7 detection system would typically encounter input videos similar to the images in Figs. 2a and 2b, these served as the primary objects of experimentation.

Model training was conducted using the default Ultralytics script [24], using the yolov5s.pt initial weight file for transfer learning. The configuration included an image size of 640x640, a batch size of 16, no frozen layers, 500 epochs, and an early stopping set to 8 epochs without improvement to prevent overfitting.

The images comprising the initial training dataset were extracted from the color videos illuminated by fluorescent lamps used in experiment A (Inference Phase), totaling 414 samples. For validation, a new set of videos was recorded—also featuring fluorescent lighting and dark environments with infrared illumination—in a different location from the first set (see Fig. 3). These images were never included in the training dataset. Image annotation was performed using LabelImg.

The datasets constructed for this study are listed in Table I.

The "color" dataset is a collection of original color images under fluorescent lighting, representing the images typically found in publicly available online datasets or those constructed from daytime acquisitions.

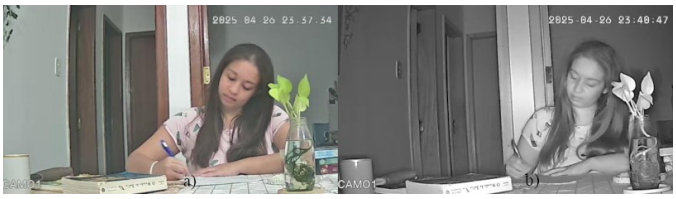

Fig. 3. Example of the image used for the validation of the new predictive models: (a) color and illuminated by fluorescent lamp; (b) grayscale with infrared.

The "dark*" datasets consist of artificially darkened versions, where the pixel intensity limits are set to 32, 16, or 8. Subsequently, the "color" dataset is combined with the "dark*" versions to provide the predictive model with examples of the investigated object class in both daytime/bright and nighttime/dark scenarios.

TABLE I
COMPOSITION OF THE DATASETS USED FOR TRAINING IN THE EXPERIMENT

Dataset Name	*Description*	*Figure*
color	Original set of images captured in a daytime environment (baseline)	5
ExDark	Low-light images dataset [10]	6
dark32	Artificially darkened version with pixel intensity limit set to 32	-
dark16	Artificially darkened version with pixel intensity limit set to 16	-
dark8	Artificially darkened version with pixel intensity limit set to 8	-
color+dark32	Union of color and dark32	7
color+dark16	Union of color and dark16	-
color+dark8	Union of color and dark8	-
color+dark32+dark16	Union of color, dark32, and dark16	-
color+dark32+dark8	Union of color, dark32, and dark8	-
color+dark16+dark8	Union of color, dark16, and dark8	-
color+dark32+dark16 +dark8	Union of color, dark32, dark16 and dark 8	-
color+aug+bg	"color" dataset with data augmentation and background (negative) samples	8
(color+aug+bg)_dark 32	Augmented/background set with a global intensity limit of 32 applied	9

The "color+aug+bg" dataset represents an effort to mitigate potential overfitting and the occurrence of false-positive detections, given that the "color" dataset contains a relatively small number of images compared to the datasets used by the original model providers (e.g., YOLOv5, which uses over 330,000 images). Data augmentation was achieved by applying horizontal flips, random rotations between 10° and 15°, subtle blur, and slight vignetting to the "color" images to simulate variations in focus and lighting. For flips and rotations, the

corresponding transformations were also applied to the annotation files. In addition to direct transformations of the original images, background samples—images of the scene without the presence of the "person" class—were added. This was intended to train the model to correctly identify the absence of targets, thereby reducing false detections in environments with high visual clutter. Finally, the "(color+aug+bg)_dark32" dataset consists of the "color+aug+bg" set supplemented by the artificial darkening of all its images using LUT compression to a value of 32.

V. Results

A. Inference Phase

Fig. 4 presents the confidence values for all video frames across each preprocessing effort submitted to the pre-trained predictive model. Fig. 4a displays the video recorded with infrared illumination (referencing Fig. 2b), showing confidence values of around 90% for many frames. While this may initially be considered a good result, it is subject to a caveat discussed below. The video of the scene without conventional or infrared lighting (Fig. 2c) allows for almost no detections, and the few that occur exhibit very low confidence, around 20%. The video processed with shadow recovery (Fig. 2e) also yields poor results. In contrast, the video with brightness and contrast adjustments (Fig. 2d) outperforms the subsequent two videos, achieving more frequent detections with an average confidence of approximately 30%; while this is a marked improvement, it is still not considered a satisfactory result.

Fig. 4b presents the confidence levels for the video under fluorescent lighting (Fig. 2a), which represents the ideal condition for which the predictive model was designed, thus serving as the control result. Comparing the infrared video to this control, similar values are observed, except at the beginning and end of the plot, where values drop. These lower values are associated with the actor entering and exiting the scene, which is expected in a video of this nature. Infrared illumination in commercial cameras tends to be concentrated at the center of the frame, while the edges remain darker—a phenomenon known as vignetting. Consequently, when the object of interest is located in these darkened regions, detection quality is expected to decrease. Furthermore, certain detections at the end of the video are noteworthy because, according to the control results, they should not exist; these are cases of false positives. Finally, the video generated by the transformation script (Fig. 2f) shows a profile similar to the brightness/contrast adjustment but with slightly higher confidence values, although still far below the control results.

These results suggest that the preprocessing efforts implemented in this article do not contribute significantly to improving outcomes in 24/7 object detection systems when the predictive model is kept static. While a script with more elaborate transformations might further boost results, it would require more processing power, leading to higher latency between camera capture and detection. In real-world conditions, such latency is undesirable. It is also noted that simple brightness and contrast adjustments helped the detector to identify objects and may be useful options for minor improvements. Finally, the emergence of false positives in the infrared video is a critical issue that must be addressed.

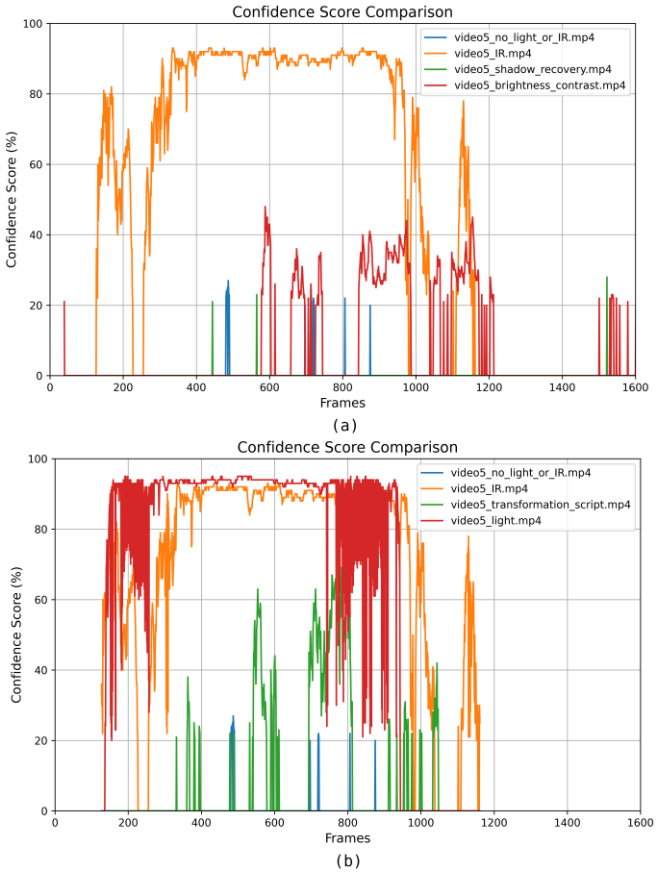

Fig. 4. Confidence results for all video frames obtained during image preprocessing attempts prior to model inference.

B. Training Phase

The datasets described in Table I were used for training, each generating a predictive model that was subsequently evaluated. It needs to be remembered that the intended contribution of this paper is to provide a low-cost operational solution to the performance degradation problem in surveillance and monitoring systems. Although false positive predictions are undesirable—as they draw unnecessary attention—it is understood that achieving high accuracy in truly positive person detections takes precedence. Furthermore, while obtaining elevated precision and recall values is desirable, within the context of this study, the latter is of greater importance than the former. Table II demonstrates that datasets consisting of very dark images slightly degrade precision. Conversely, recall improves when dark-image datasets are utilized, yielding results superior to the exclusive use of the color dataset. These two metrics also suggest that the "dark8" dataset hinders the detection process, likely because its images are extremely dark and exhibit almost no distinct variance among themselves. Another critical point is that the "ExDark" dataset underperforms compared to "color+aug+bg", and is unequivocally outperformed by the "(color+aug+bg)_dark32" configuration.

The "color" dataset (see Fig. 5) revealed that the model trained with original color images maintained performance similar to the YOLOv5 base model when processing videos under adequate lighting, showing consistent and reliable detections in bright environments. However, a significant reduction in detection rates was observed for infrared images, along with a total absence of predictions in darkened videos—including those manipulated with the illumination correction script developed in the Inference Phase. This result supports the hypothesis that a model trained exclusively on high-luminosity data does not generalize well to low-light conditions, reinforcing the need to include nighttime/dark image samples during training.

TABLE II
TRAINING METRICS

Dataset Name	*Precision*	*Recall*	*mAP@ 0.5*	*mAP@ 0.5:0.95*
color	0.99	0.69	0.995	0.930
ExDark	0.93	0.76	0.985	0.919
dark32	0.99	0.75	0.901	0.825
dark16	0.97	0.73	0.814	0.733
dark8	0.93	0.70	0.763	0.684
color+dark32	0.99	0.71	0.994	0.928
color+dark16	0.98	0.71	0.992	0.925
color+dark8	0.97	0.70	0.989	0.917
color+dark32+dark16	0.98	0.71	0.983	0.912
color+dark32+dark8	0.97	0.70	0.929	0.880
color+dark16+dark8	0.96	0.70	0.840	0.860
color+dark32+dark16+ dark8	0.97	0.71	0.932	0.898
color+aug+bg	1.00	0.75	0.995	0.934
(color+aug+bg)_dark3 2	0.99	0.88	0.999	0.947

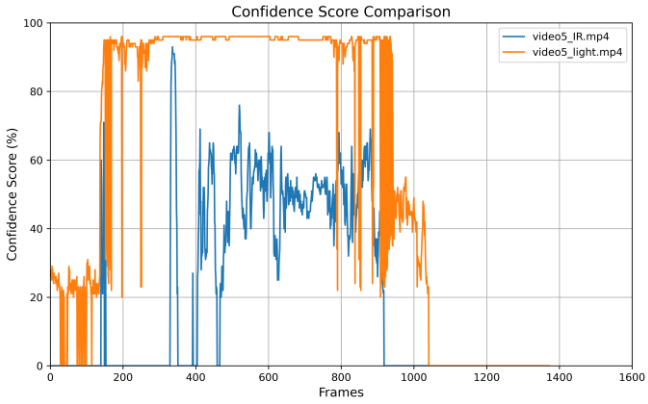

Fig.5. Testing the model trained with the "color" dataset under different lighting conditions.

The "ExDark" dataset (see Fig. 6) induced a significant reduction in detection confidence within illuminated environments. For infrared imagery, the performance closely mirrored that observed with the "color" dataset, albeit exhibiting slightly lower variance. Furthermore, no detections occurred in dark images, suggesting that nighttime imagery is not equivalent to dark imagery, as exemplified by nighttime imaging without infrared support.

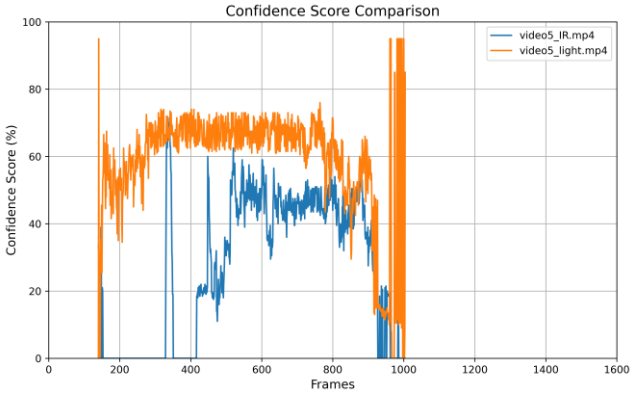

Fig. 6. Testing the model trained with the "ExDark" dataset under different lighting conditions.

A gain in detection quality for nighttime/dark images was expected when starting training from a stabilized network (yolov5s.pt) and adding only images of situations unknown to the predictive model (datasets dark8, dark16, and dark32). However, this not only yielded poor results in the nighttime/dark scenario but also significantly compromised the daytime/illuminated scenario. The dark32 dataset achieved the best performance among these, producing some detections, but with confidence levels below 10%. This indicates that while the dataset can sensitize the detector, it fails to provide a sufficient increase in nominal confidence values.

Datasets that combine sharp color images with dark images (color+dark*) aim to present to the detector both a daytime/bright scene and its homologous nighttime/dark variants, directing the learning process to capture features for both situations. Of these combinations, only the color+dark32 warrants reporting; its confidence values are shown in Fig. 6. A more stable confidence behavior is noted across luminosity variations compared to the color dataset in both lighting scenarios. However, the confidence value drops for videos with infrared illumination. This result reinforces the hypothesis that gradual illumination adjustments can benefit the adaptation of computer vision models to nighttime scenarios, provided visual coherence is maintained with dark images where the scene remains visible, even under intense darkening. When darkening precludes visibility, the effect is reversed, leading to a significant loss in detection capability.

Observing Figs. 5 and 6, adding color images alongside their darkened counterparts produced less noisy confidence variations but hindered infrared video detection and introduced false positives. This suggests that the transfer learning dataset required expansion; thus, the experiment proceeded toward data augmentation combined with the inclusion of scenes without detectable objects. Fig. 7 illustrates the behavior of the model produced from the color+aug+bg dataset, indicating its ability to maintain continuous and reliable detections in illuminated environments, without significant loss of precision between frames. This behavior reinforces the effectiveness of data augmentation and the inclusion of background images in

improving generalization and prediction stability in daytime contexts. It also confirms that the lack of dark samples negatively affects detection in infrared video.

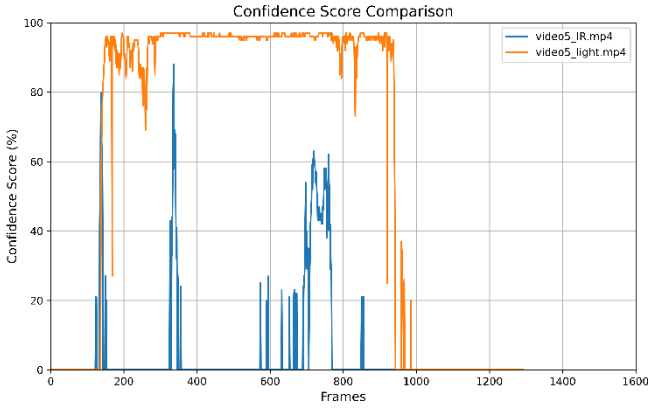

Fig. 7. Testing the model trained with the "color+dark32" dataset under different lighting conditions.

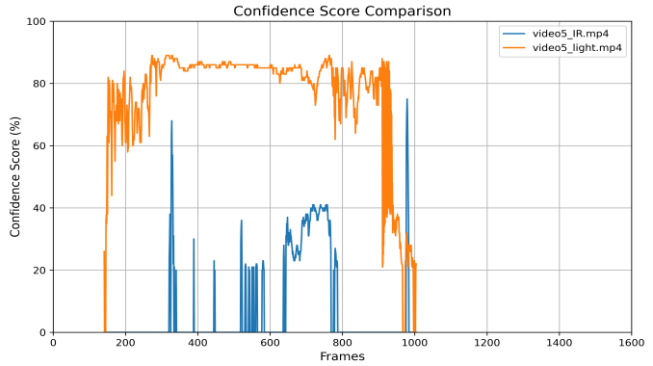

Fig. 8. Testing the model trained with the "color+aug+bg" dataset under different lighting conditions.

Consequently, it is logical to combine the dataset construction—comprising data augmentation and the inclusion of background images—with their homologous darkened images. Fig. 8 presents the confidence behavior in this context, showing that the model achieved more balanced and consistent confidence values, even in low-light situations—a result not observed in previous experiments. This represents a breakthrough in terms of consistency and generalization.

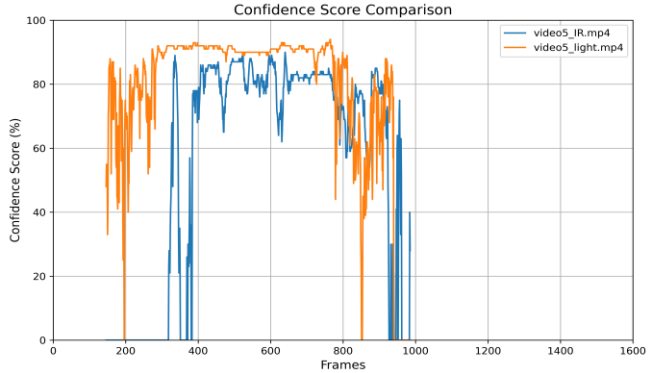

Fig. 9. Testing the model trained with the "(color+aug+bg)_dark32" dataset under different lighting conditions.

The primary problem identified throughout the experiments was the instability of detections under low-light conditions, especially in scenes with reflections, sharp shadows, and loss of contrast. As previously observed, the base model trained only with the color dataset (Fig. 5) performed well in well-lit environments but significantly lost reliability in darker contexts. This limitation showed that the model relied heavily on the luminosity features of the original dataset, hindering its generalization to nighttime/dark situations. To mitigate this, several dataset variations were tested. Fig. 8 shows the best results, based on the integration of the original dataset with data augmentation, background images, and darkened versions.

The model generated by (color+aug+bg)_dark32 outperforms the original. In the visible-light video plot (orange), a small reduction in the maximum confidence value is observed—the upper plateau is slightly lower—but this loss is offset by a noticeable increase in confidence in the peripheral regions where light oscillation was previously higher.

In the infrared video plot (blue), the improvement is even more evident: the originally noisy behavior, centered around a confidence of approximately 50%, is replaced by a more stable plateau in the 80% range, indicating greater detection consistency. Furthermore, the edges of this plateau show a significant increase, rising from values near 0% in the base model to substantially higher levels in the final model. Additionally, the blue and orange curves cease detection at very similar timestamps, evidencing greater alignment between the two lighting scenarios and reinforcing the robustness acquired by the model.

While the original predictive model exhibits relevant interruptions and fluctuations across frames, the enhanced model maintains more stable predictions, including in partially shadowed segments. These results confirm that the combined use of color and darkened images, alongside data augmentation and neutral backgrounds, constitutes the most effective solution for the initial problem of poor low-light performance. Furthermore, it demonstrates that compressing input pixel values to a reduced intensity range of 32 is an appropriate parameter choice.

VI. Discussion

Loh and Chan [4] report that the ExDark dataset [10] comprises ten distinct types of low-light conditions. Among these, the configuration that most closely resembles the images in our experiment is 'Low', which consists of images characterized by very low illumination and barely visible details. The authors study utilize 12 classes, including a 'people' class that corresponds to the 'person' class in this paper. They also report a detection confidence of approximately 50% when the dataset is evaluated using a ResNet-50 trained on MS COCO. Indeed, this baseline can be inferred from Fig. 6, which strengthens the validity of the results presented in Fig. 9, where our proposed (color+aug+bg)_dark32 dataset achieves a confidence level of approximately 80%. Additionally, Li et al. [26] provide a comprehensive survey on low-light image enhancement methodologies aimed at improving detection rates

within deep learning architectures. Their work also addresses the ExDark dataset, suggesting that across the 14 evaluated techniques, confidence values exceeding 65% can be considered robust results.

However, more critical than the absolute confidence values achieved is the operational complexity required to produce such outcomes. On one hand, the 14 methodologies analyzed by Li et al. [26] operate at the frontier of the state of the art in digital image processing—solutions that commercial engineering teams managing surveillance and monitoring systems typically lack the resources to reproduce. Conversely, the transformations proposed in the current paper can be readily executed using standard commercial tools, such as Photoshop and GIMP, making them highly accessible to practical engineering workforces. On the other hand, engineering a custom dataset such as ExDark for a specific client is often economically unviable, as discussed in Section I. The transformations suggested herein generate a more robust dataset tailored for 24/7 operations without imposing prohibitive financial constraints on the project.

VII. Conclusion

A. Final Remarks

The current work investigated the impact of various visual conditions on person detection in low-light environments using the YOLOv5 model. The experimental results support the conclusion that conventional and straightforward preprocessing images to emphasize features for a pre-trained predictive model is not an approach that yields significant benefits. Conversely, the study evaluated how lighting and visual diversity—known drivers of the training process—constitute a superior approach for improving model generalization. In this regard, the recommended approach is to generate a new predictive model from an existing one, utilizing new images of the target object combined with data augmentation and the inclusion of neutral backgrounds, all synthetically darkened by reducing image exposure to a limit intensity of 32 in the LUT. Thus, this study concludes that visual darkening strategies, coupled with data augmentation and the introduction of varied background scenarios, render the model more capable of handling low-light environments, representing a promising path for surveillance systems, monitoring, and real-world applications.

B. Future Works

Despite the progress achieved, several opportunities exist to deepen the current research. A primary direction consists of substantially expanding the low-light video database by applying the darkening procedure developed in this work prior to model training. In scenarios with greater diversity and data volume, a performance gain under low-light conditions is naturally expected.

Furthermore, there is potential to explore advanced methods of domain adaptation or style transfer specifically aimed at simulating infrared (IR) imagery. Methods based on Generative Adversarial Networks (GANs) could pave the way for generating versions that more closely resemble the actual output of IR sensors, moving beyond the subjective criteria of simulated exposure reduction.

It is important to note that this work did not utilize features intrinsic to the chosen deep learning architecture, therefore, the experiment suggests that the conclusions remain valid for other architectures. However, this does not eliminate the need to apply the methodology to different frameworks to evaluate sensitivity to such variations and provide statistical support for this claim.

While the increase in dataset images improved model behavior, the effects of applying dropout, weight regularization, and layer freezing during transfer learning for new model generation remain to be explored.

Considering the specific use case, where a model is deployed exclusively in a predefined location—i.e., a single scene with distinct objects—one could "bias" the predictive model with images from the exact site of deployment. While scientifically considered a shortcut to boost results, in a production environment under these specific conditions, it serves as a viable alternative to further enhance detections by sacrificing unnecessary model generality.

Finally, future work could incorporate temporal information from video sequences, exploring inter-frame coherence to reduce the incidence of false positives. The analysis of temporal and color neighborhoods can assist in distinguishing between static environmental objects and targets of interest, especially in low-light scenarios where visual noise and excessive contrast tend to compromise single-frame detection. Approaches integrating temporal smoothing, tracking, or multi-frame information fusion represent promising future directions.

Acknowledgments

We thank Dr. Erik Russell Wild for providing a native English speaker revision of the manuscript, and Robin Hambly for a second native English speaker revision. We would like to pay special regards to AI2Biz Lab, an Intel AI Builder Partner, for providing a rich environment for discussions.

References

[1] Krizhevsky, A.; Sutskever, I.; Hinton, G.E., Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105, 2012, doi: doi.org/10.1145/3065386

[2] Romera, Eduardo; Bergasa, Luis M.; Yang, Kailun; Alvarez, Jose M.; Barea1, Rafael. Bridging the Day and Night Domain Gap for Semantic Segmentation. In: IEEE Intelligent Vehicles Symposium, 2019, doi: 10.1109/IVS.2019.8813888

[3] Xueqing Deng; Peng Wang; Xiaochen Lian; Shawn Newsam. NightLab: A Dual-level Architecture with Hardness Detection for Segmentation at Night. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.16917-16927, 2022, doi: 10.48550/arXiv.2204.05538.

[4] Loh, Yuen Peng; Chan, Chee Seng. Getting to know low-light images with the Exclusively Dark dataset. In: Computer Vision and Image Understanding, v. 178, pp.30–42, 2019, doi: 10.1016/j.cviu.2018.10.010

[5] Zongjiang Gao; Yingjun Zhang; Yuankui Li. Extracting features from infrared images using convolutional neural networks and

transfer learning, In: Infrared Physics & Technology, v. 105, pp.103237, 2020, doi: 10.1016/j.infrared.2020.103237.
[6] A. Kapoor; R. Greiner, Learning and classifying under hard budgets, In: Proceedings of the 16th European Conference on Machine Learning, Springer-Verlag, Berlin, Heidelberg, pp. 170–181, 2005, doi: 10.1007/11564096_20.
[7] P. Melville; M. Saar-Tsechansky; F. Provost; R. Mooney, An Expected utility approach to active feature-value acquisition, In: Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM '05), IEEE Computer Society. Washington, DC, USA, pp. 745–748, 2005, doi: 10.1109/ICDM.2005.23.
[8] Ma, Xiang; Xia, Lei; Zhou, Bolei; et al. The Effect of Improving Annotation Quality on Object Detection Datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022, doi: 10.1109/CVPRW56347.2022.00532.
[9] C. Agnew, A. Scanlan, P. Denny, E. M. Grua, P. van de Ven and C. Eising. Annotation Quality Versus Quantity for Object Detection and Instance Segmentation, IEEE Access, v. 12, pp. 140958-140977, 2024, doi: 10.1109/ACCESS.2024.3467008
[10] Loh, Yuen Peng and Chan, Chee Seng. Getting to Know Low-light Images with The Exclusively Dark Dataset. Computer Vision and Image Understanding, v. 178, pp. 30-42, 2019, doi: 10.1016/j.cviu.2018.10.010.
[11] Ren, B., Xu, Z., Zhao, J. et al. Complex dark environment-oriented object detection method based on YOLO-AS. Sci Rep 15, 21873, 2025, doi: 10.1038/s41598-025-07348-0.
[12] Li H, Fu Y. FCDFusion: A fast, low color deviation method for fusing visible and infrared image pairs. Computational Visual Media,11(1): pp.195-211, 2025, doi: 10.26599/CVM.2025.9450330.
[13] Donia, Eman; El-Rabaie, El-Sayed; El-Samie, Fathi; Faragallah, Osama; A.El-Hag, Noha. Infrared image fusion for quality enhancement. Journal of Optics. 52. Pp.1-7, 2023, doi: 10.1007/s12596-022-01018-4.
[14] Haoran Zhu; Wenying Zhang. Infrared image fusion for quality enhancement. IEEE Access, v. 13, pp. 61862-61875, 2025, doi: 10.1109/ACCESS.2025.3557799.
[15] H. Kim, J. Kim and J. Kim. Image-to-Image Translation for Near-Infrared Image Colorization, 2022 International Conference on Electronics, Information, and Communication (ICEIC), Jeju, Korea, Republic of, 2022, pp. 1-4, 2023, doi: 10.1109/ICEIC54506.2022.9748773.
[16] D. Ma, S. Li, J. Su, Y. Xian and T. Zhang, "Visible-to-Infrared Image Translation for Matching Tasks," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 18199-18213, 2024, doi: 10.1109/JSTARS.2024.3468456.
[17] Jiawei Guo; Jieming Ma; Ángel F. García-Fernández; Yungang Zhang; Haining Liang. A Survey on Image Enhancement for Low-Light Images. Heliyon, v. 9, n. 4, 2023, doi: 10.1016/j.heliyon.2023.e14558.
[18] Y. Luo, X. Chen, J. Ling, C. Huang, W. Zhou and G. Yue. Unsupervised Low-Light Image Enhancement With Self-Paced Learning, in IEEE Transactions on Multimedia, vol. 27, pp. 1808-1820, 2025, doi: 10.1109/TMM.2024.3521752.
[19] Saifullah S, Drežewski R. Advanced Medical Image Segmentation Enhancement: A Particle-Swarm-Optimization-Based Histogram Equalization Approach, Applied Sciences, v.14, n. 2, pp. 923-949, 2024, doi: 10.3390/app14020923
[20] R. Maini and H. Aggarwel, "A Comprehensive Review of Image Enhancement Techniques," Journal of Computing, Vol. 2, No. 3, 2010, pp. 8-13. Available at: https://arxiv.org/pdf/1003.4053. Accessed: 15 jul. 2025.
[21] Mohammad Mahdizadeh, Shijie Chen, Peng Ye. Illuminating the night: A light source-aware and exposure-balanced low-light enhancement approach for real nighttime scenes, Digital Signal Processing, v. 159, pp. 104999, 2025, doi: 10.1016/j.dsp.2025.104999.
[22] Gonzalez, Rafael C.; Woods, Richard E. Digital Image Processing. 3. ed. Pearson, ISBN 978-0131687288, 2018.
[23] Syl Arena. Speedliter's Handbook: Learning to Craft Light with Canon Speedlites, ed.2, Peachpit Press, ISBN 978-0-134-00791-5, 2016.
[24] ULTRALYTICS. YOLOv5 by Ultralytics. 2020. Available at : https://docs.ultralytics.com/models/yolov5/. Accessed: 13 abr. 2025.
[25] Gomes, Isabelli Pinto. Github Repository: YOLOv5 by Ultralytics. 2020. Available at: https://github.com/IsabelliGomes/low-light-person-detection-yolo. Accessed: 01 jun. 2026.
[26] Chongyi Li; Chunle Guo; Linghao Han; Jun Jiang; Ming-Ming Cheng; Jinwei Gu. Low-Light Image and Video Enhancement Using Deep Learning: A Survey. In: IEEE Transactions on Pattern Analysis and Machine Intelligence , v.44, n.12, pp. 9396-9416, 2022, doi: 10.1109/TPAMI.2021.3126387.

Isabelli Pinto Gomes received her technical degree in Electronics from ETE Henrique Lage (FAETEC), Rio de Janeiro, Brazil, in 2015. She is completing the Bachelor's degree in Electronic and Computer Engineering at the Federal University of Rio de Janeiro (UFRJ), Brazil, with expected graduation in 2026. Since 2021, she has been working at UOL Group, where she currently holds the position of Software Engineer. Her work focuses on full-stack web development, performance optimization, and scalable software systems, contributing to the development and improvement of large-scale digital platforms. Her research interests include computer vision, machine learning, and the application of artificial intelligence techniques to real-world software systems.

Flávio Luis de Mello received his DSc. in Theory of Computation and Image Processing from the Federal University of Rio de Janeiro - UFRJ (2006), MSc. in Computer Graphics from the Federal University of Rio de Janeiro - UFRJ (2003), Undergraduate degree in Systems Engineering from the Military Institute of Engineering - IME (1998). He developed command and control systems and implemented military messages interchange applications during twelve years as a Brazilian Army officer. He was responsible for developing software applications based on machine learning and knowledge reasoning from Mentor Group. For seven years, he was the technical leader of Intel's Center of Excellence on Artificial Intelligence in Brazil. Dr Mello currently is Full Professor at the Electronic and Computer Engineering Department (DEL) of Polytechnic School (Poli) at the Federal University of Rio de Janeiro (UFRJ). He is head of the Machine Intelligence and Computing Models Laboratory (IM^{2}C).